WILEY

encyclopedia of automotive engineering



1

Editors-in-Chief

David Crolla | David E. Foster | Toshio Kobayashi | Nicholas Vaughan

FISITA

Possibilities of Coil Springs and Fiber-Reinforced Suspension Parts

Joerg Neubrand

Chassis Technologies—Mubea Fahrwerkstechnologien GmbH, Attendorn, Germany

1 Introduction 1
2 Calculation of Coil Springs 1
3 Side Load Springs 3
4 New Generation Coil Springs 4
5 Materials and Robustness 7
6 Composite Springs 8
References 11

1 INTRODUCTION

As early as over 400 years ago, the first helical compression springs (coil springs) were already used for wheel suspensions of a wagon body, and at the very latest since the invention of the wheel suspension strut (McPherson), they have represented the best spring design for the vertical dynamics of passenger cars. They took over suspension functions and because of their advantages regarding weight and installation space, replaced leaf springs almost completely. The necessary suspension arms to carry the coil springs may be regarded as a disadvantage, but their design enables exceptionally good road holding and safety.

Springs will be deformed elastically and during this process, they take up potential energy, which will be released when relieved. They will have to live up to such duties also during repeated and dynamic loads and considerable deforming processes. For this reason also, steel continues to be an ideal material to make springs.

Over the past few years, resource-saving weight reduction has assumed growing importance. Assuming that the weight of a smaller middle class car was kept constant, the weight of a chassis support spring was reduced by about 55% since 1992 (Figure 1). Such weight reductions were brought about by higher loads, underpinned by optimized manufacturing technologies and new types of steel, without any reductions in robustness.

One example of the developments in manufacturing technology is the high performance process (HPP) developed by the Mubea company, as a result of which the load-bearing strength of helical compression springs was increased by more than 10%. This method was first introduced in 2003, and it has been used as the global benchmark since 2004.

Moreover, in future, there will be increasing demands by car builders as regards CO_2 reductions, lower vehicle weights, and a reduction of unsuspended sections and the robustness of springs, particularly as affected by corrosion. This means that optimum use of materials employed will play a decisive role, underpinned by efficient design. In addition to that, alternative materials, such as composite materials, may be increasingly important when used for suspension purposes.

2 CALCULATION OF COIL SPRINGS

In general, the equations of the German DIN standard 2089 can be used for an initial estimate of the dimensions and stresses of a cylindrical compression coil spring with a constant wire diameter. Figure 2 shows the definition of calculation values for one deformable coil ($n = 1$).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

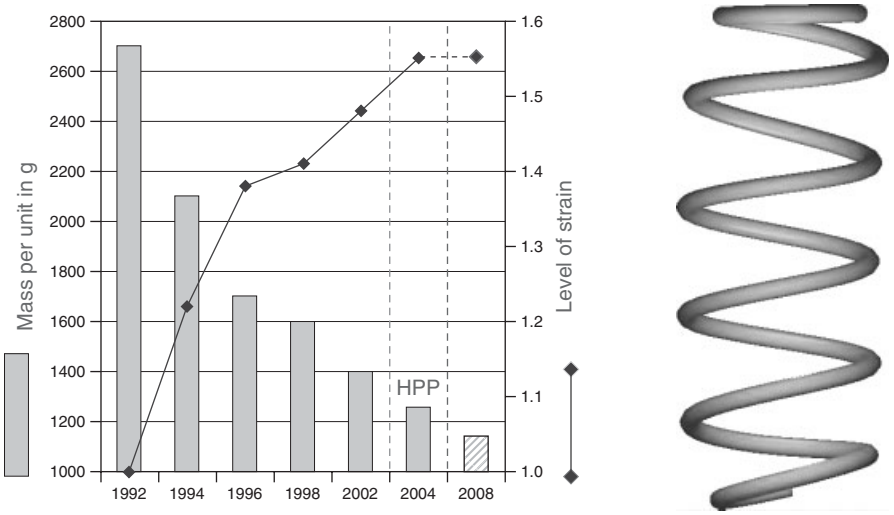


Figure 1. Evolution of weight and load stress levels of suspension springs based on a medium-sized vehicle, given constant vehicle weight. (From Neubrand *et al*., 2010. Copyright © 2010 SAE International. Reprinted with permission.)

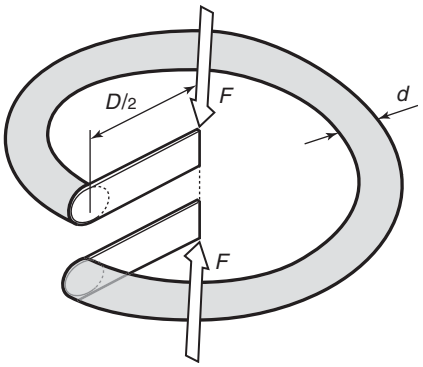


Figure 2. Compression coil spring calculation values. (Reproduced from Carlitz and Neubrand, 2008. With kind permission from Springer Science+Business Media.)

Spring work is defined as:

$$W = \frac{F \cdot s}{2} \tag{1}$$

where s is the spring stroke.

Spring force can be calculated as follows, considering G as the shear modulus:

$$F = \frac{G \cdot d^4 \cdot s}{8D^3 \cdot n} \tag{2}$$

Spring stiffness is given by

$$c = \frac{F}{s} = \frac{G \cdot d^4}{8D^3 \cdot n} \tag{3}$$

Shear stress may be calculated using the following equation:

$$\tau = \frac{G \cdot d \cdot s}{\pi \cdot D^2 \cdot n} = \frac{8F \cdot D}{\pi \cdot d^3} \tag{4}$$

Such stress equation is valid for a straight wire and torsion stress only. In coil springs, the wire is curved and the stresses on the inside are higher than those on the outside of the coil. Torsion and shear forces will have to be added. Moreover, the same force is being applied to a smaller area on the inside of the coil (Figure 3).

Therefore, stress has to be corrected by the factor k, depending on w, the ratio of the coil diameter to the wire diameter.

$$w = \frac{D}{d} \tag{5}$$

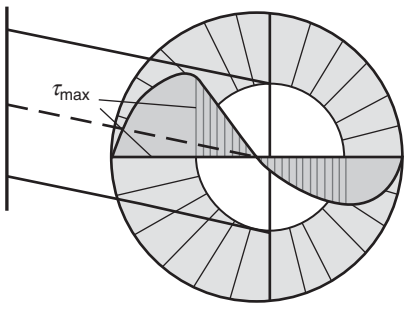


Figure 3. Stress increase on the inside of a coil. (Reproduced from Carlitz and Neubrand, 2008. With kind permission from Springer Science+Business Media.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

$$k = \frac{w + 0.5}{w - 0.75} \tag{6}$$

Corrected stress may then be calculated as follows:

$$\tau_k = k \cdot \tau \tag{7}$$

When designing a spring, the maximum stress τ_{max} is most important for minimizing the weight of a spring. Assuming a uniform deflection without any disturbing side forces, the mass of a cylindrical coil cannot be lower than a certain minimum mass m_{min} (Brandt, Kobelev, and Neubrand, 2007). This mass m_{min} is calculated according to Equation 1 with ρ as the density of the spring wire material and G as the shear modulus. The force F_{max} is the load at full jounce height. The coil spring mass m_{min} is reciprocally related to the maximum shear stress τ_{max}.

$$m_{min} = 2\rho \cdot G \frac{F_{max}^2}{c \cdot \tau_{max}^2} \tag{8}$$

In general, there are two ways to reduce coil spring weight:

1. Shape optimization for decreasing the load stress along the coil spring wire.
2. Increasing maximum allowed stress τ_{max}.

Until quite recently, it was common practice to decrease spring weight by increasing the stress level of suspension springs (Figure 1) and keeping robustness with certain measures regarding material and process optimization.

In real-life springs, load stresses are distributed nonuniformly along the coils. Therefore, it is necessary to calculate springs using a finite element analysis (FEA) as described, for example, in Georges (2000).

Owing to tremendous progress in calculation methods and corresponding production technologies, shape optimization is now possible for increasing the material usage by stress homogenization.

3 SIDE LOAD SPRINGS

The concept and design of a vehicle axle mainly defines the shape of a suspension coil spring. Moreover, the loads occurring and packaging are important factors. Two types of loads can be differentiated. First, the coil over shock, where both ends of the coil spring travel in line with each other, spring and damper form a single device. The second load application is to place one end of the coil spring on a suspension arm. Here, this end travels along a spatial curve, which leads to a nonuniform stress distribution along the spring profile and causes a distortion of the spring body (Carlitz and Neubrand, 2008).

The McPherson strut is a special case of the first kind of load. Owing to the conceptual axle design (Figure 4), unwanted moments occur, resulting in side loads at the

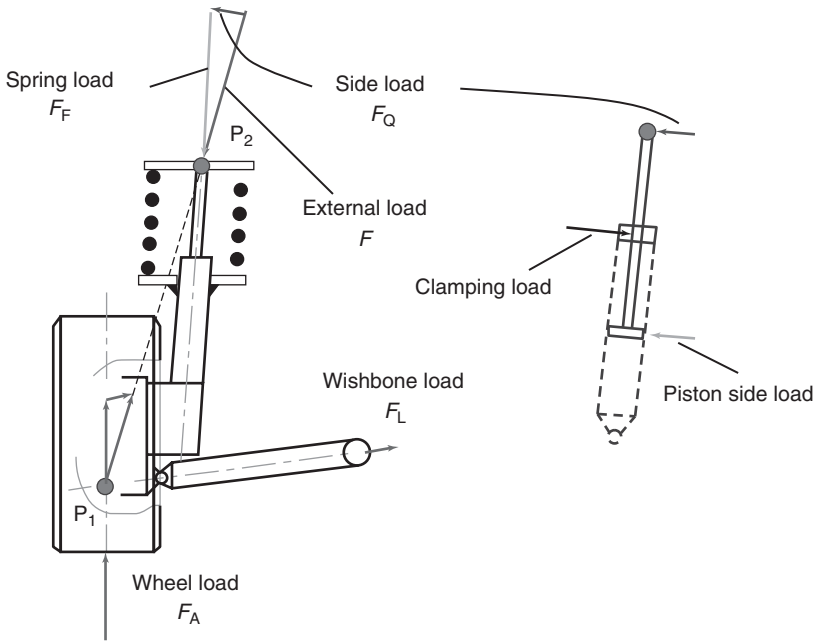


Figure 4. Principal kinematics of a McPherson strut. (From Neubrand *et al*., 2010. Copyright © 2010 SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

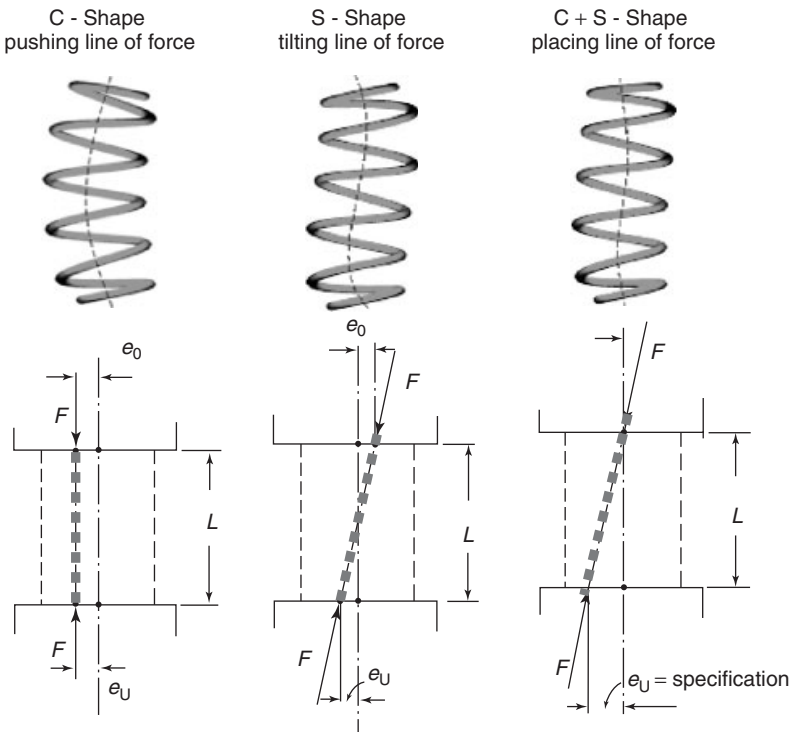


Figure 5. Adjustment of the load vector through spring shape. (From Neubrand *et al*., 2010. Copyright © 2010 SAE International. Reprinted with permission.)

damper rod and therefore, to an increased friction along the damper rod and damper piston bearings. This friction does not only generate increased wear and tear of the damper piston and seals but driving comfort may also be reduced. A slip–stick effect is noticeable especially at very small damper travels.

Until the late 1980s, this problem had been solved by tilting a huge cylindrical coil spring on a McPherson strut, which offered compensation for such unwanted moments. However, as packaging space for suspension components becomes ever smaller, this solution was soon obsolete because of the voluminous coil springs needed to allow for tilting.

For this reason, the development of side load (SL) springs had been started (Muhr and Schnaubelt, 1991; Kobelev *et al*., 2002; Brandt and Neubrand, 2001). These counteract the unwanted moments within a small package because of geometrical intelligence. These springs are showing an S-shaped centerline in unloaded condition, which leads in loaded condition to a tilted load vector and fully compensates the unwanted moments around the McPherson strut (Carlitz and Neubrand, 2008; Muhr and Schnaubelt, 1991; Kobelev *et al*., 2002; Brandt and Neubrand, 2001).

Furthermore, the load vector of these coil springs can be further attuned by adjusting spring geometry, without having to change spring seat designs. Moreover, spring seats can be minimized, because SL coil springs are mostly designed as double-pig-tail springs.

The principle of control of the load vector is shown in Figure 5, and the load vector offset can be "pushed" by a C-shape and "tilted" by a pure S-shape. A combination of these two adjustments leads to a typical SL spring, where the load vector can be "placed."

The technology of the SL spring as explained earlier triggered the development of multiple versions and modifications of spring designs and as of today side load technology is used in more than 70% of all McPherson strut applications worldwide.

Moreover, the ability for controlling the load vector of coil springs could also be used for additional functions, such as height leveling systems as described in Carlitz, Neubrand, and Hengstenberg (2005).

4 NEW GENERATION COIL SPRINGS

Nonparallel coil spring deflection leads to a nonuniform load stress distribution along a coil spring. Even though an SL coil spring is a coil over shock, the deflection is nonparallel. The load vector is inclined relative to the axis

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

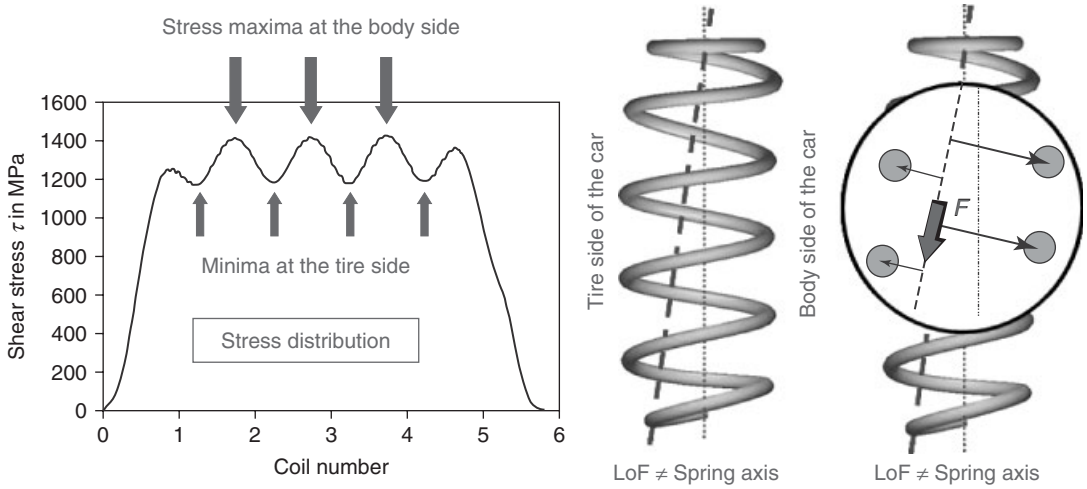


Figure 6. Stress distribution of a conventional SL spring with inclined load vector. (From Neubrand *et al*., 2010. Copyright © 2010 SAE International. Reprinted with permission.)

of the damper and this is causing a nonuniform distance of coil spring wire to the load vector. Distances from the wheel are smaller than the distances on the opposite side of the load vector (Figure 6).

The distance from the coil spring wire to the load vector also reflects the moment arm of the moment around the coil spring wire centerline. Thus, nonuniform deflection of the spring and an inhomogeneous stress distribution occur under load (Figure 6). This leads to local stress minima at the coil position toward the wheel and local maxima toward the vehicle body. Hence, material utilization is not at its best and spring weight is too high, considering maximum stress $\tau_{\max}$ as the limiting factor.

Therefore, a significant weight reduction and optimal materials utilization may be achieved by homogenizing stress distribution. Modern FEA methods as well as corresponding manufacturing processes are offering such a chance.

The following equations are used to describe the true shape of a suspension spring (Figure 7; Neubrand *et al*., 2010).

$$\underline{r_{\mathrm{g}}} = \begin{bmatrix} \cos(n) \cdot R(n) & \sin(n) \cdot R(n) & z(n) \end{bmatrix}^T \tag{9}$$

$$\underline{r_{\mathrm{m}}} = \begin{bmatrix} x(z) & y(z) & z(n) \end{bmatrix}^T \tag{10}$$

$$\underline{r} = \underline{r_{\mathrm{m}}} + \underline{\underline{T}} \cdot \underline{r_{\mathrm{g}}} \tag{11}$$

These equations can be used for a numerical determination of local stresses of the whole spring wire during deflection. Latest FEA software offers an optimized coil spring shape respecting all restrictions such as package, load vector, rate, and maximum stresses and the shape is automatically translated into required computer-aided design (CAD) formats. Automated pre- and postprocessing as well as parametrical definition of the spring shape allow for systematic adjustment of coil spring parameters while solely focusing on the FEA output.

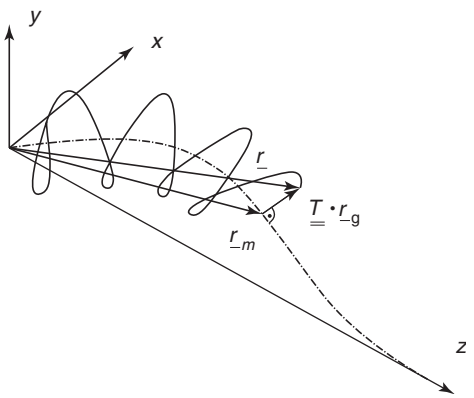


Figure 7. Parametric description of an SL coil spring. (Reproduced from Neubrand *et al*., 2001. © R. Brandt and J. Neubrand.)

Using this powerful method, it is possible to automatically generate the uniform load stress distribution of a new generation spring (Figure 8). Coil spring geometry is optimized in a way that preserves the defining spring characteristics, while load stress is uniformly distributed along the coil spring profile. Hence, the lowest possible spring mass $m_{\min}$ has been numerically calculated.

This new generation spring still offers all advantages of a traditional SL spring regarding side load compensation within small package space and furthermore, a huge weight reduction due to perfect material utilization.

This method of design may also be applied to other kinds of coil springs. The following example shows a rear

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

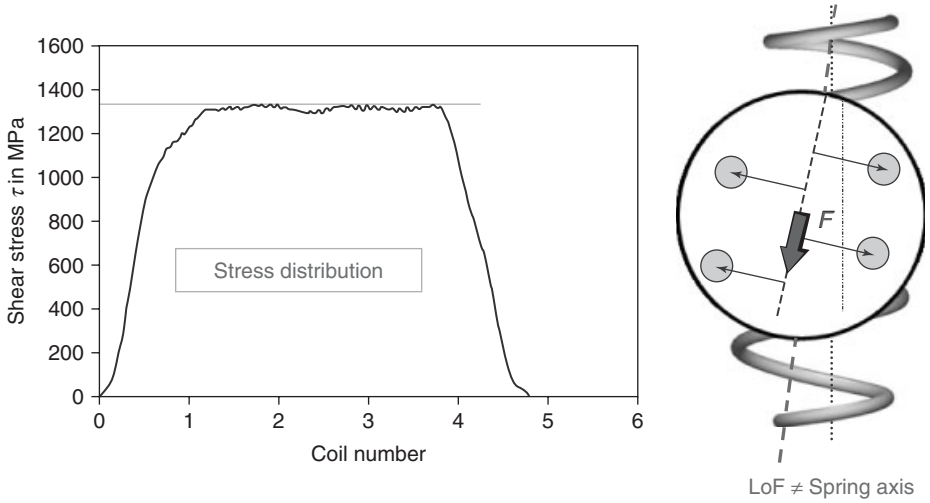


Figure 8. Uniform load stress distribution with Mubea SL new generation spring. (From Neubrand *et al*., 2010. Copyright © 2010 SAE International. Reprinted with permission.)

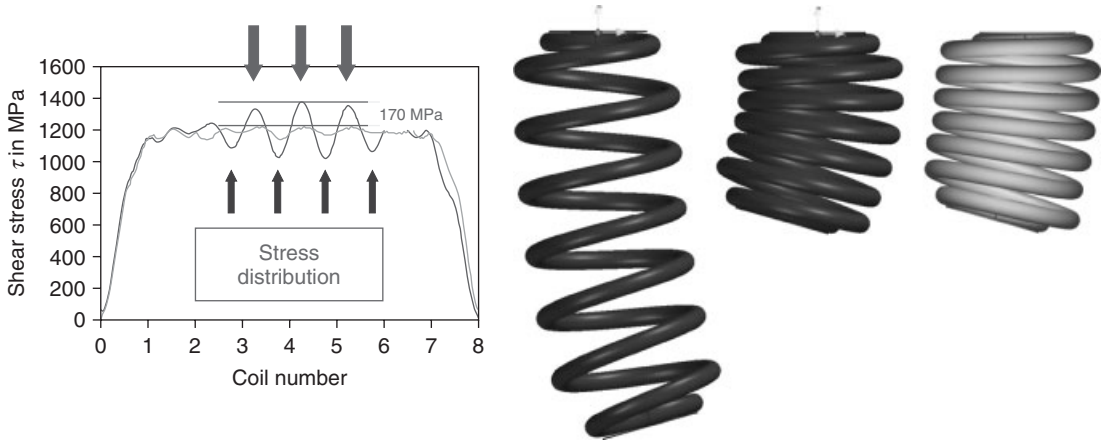


Figure 9. Mubea new generation suspension spring for axles with spatial curve travel. (From Neubrand *et al*., 2010. Copyright © 2010 SAE International. Reprinted with permission.)

axle coil spring, where the spring seats deflect along a spatial curve (Figure 9). The light gray springs show a conventional cylindrical spring at rebound and full jounce condition, whereas the dark gray spring shows a new generation spring at full jounce condition.

Spring deflection in such a system is highly nonuniform; deflection extremes even change from one direction (at rebound height) to the other side (at jounce height). The result of this nonuniform deformation once more is an inhomogeneous load stress distribution along the coil spring profile (Figure 9). Furthermore, coil clearance issues particularly at jounce height may occur.

Applying the same methods employed for new generation spring design and relying on the same principles as were employed for strut application, uniform deflection, and load stress distribution can be achieved. Just like for strut applications, significantly better steel utilization results in welcome weight reduction while load stress limits have not been increased.

The idea of stress homogenization by new generation design and, therefore, better material utilization offers a weight reduction of more than 10% without increasing maximum stress and/or better robustness.

Coil springs are being produced by hot or cold forming. In recent years, cold forming won significant market shares because of the high flexibility in forming complex shapes. The latest production processes can transfer the FEA design into the forming operation and plastic as well as elastic deformations during manufacturing may be anticipated.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Coiling programming systems have been developed to efficiently produce most complex coil spring designs with the highest possible dimensional capability.

5 MATERIALS AND ROBUSTNESS

Spring materials must have superior properties regarding storing and releasing huge amounts of elastic energy. Therefore, materials with high shear and elasticity modulus are showing the best properties for springs. Furthermore, materials should have high strength and high elasticity to withstand high loads without any plastic deformation (Neubrand, 2004). Properties such as corrosion resistance, sag loss resistance, and dynamic fatigue strength are very important; in addition, superior toughness is also important, which is increasing resistance against notches and cracks.

In general, steel is still fulfilling most of these required properties as best in class. Several alloy elements can be added for improving the steel. Until the late 1980s, CrV steels were commonly used for coil springs. Nowadays, SiCr steels represent a major part of vehicle suspension springs because of their very good properties, such as high toughness at great strength and high settling strength, even at higher temperatures.

Until the late 1980s, material impurities such as inclusions represented a major problem for springs regarding durability, especially in case of high cycle fatigue. Super clean steels were developed as the superior material for valve spring applications, avoiding critical inclusions as far as possible (Neubrand, 2004) by creation of deformable, low melting inclusions.

Continuously decreasing packages, increasing vehicle weights and higher demands regarding component lightweight design caused a significant increase of tensile strength and hardness. The inductive tempering represents a milestone of improvement in this direction, because it provided an excellent fracture toughness for tensile strength higher than $R_m = 2000\,\text{MPa}$. In addition to that, corrosion robustness is assuming ever greater importance. This is the result of combining higher stresses with high tensile strength material under corrosive environmental conditions, when the topic of stress corrosion cracking is assuming ever more relevance. In principle, a distinction is made between two corrosion mechanisms: corrosion in the presence of oxygen (rust), where corrosion scars will be formed by a local dissolving of iron and corrosion under the influence of acids, which is better known as *hydrogen embrittlement*, that is to say, the material will get brittle when cathodic hydrogen will be formed along the grain boundaries (Carlitz and Neubrand, 2008).

These two processes may be counteracted by a number of steps:

- turning the layer of rust into a covering layer by alloying metals such as Ni, Mo, or Cu, thus reducing the speed of corrosion;
- forming hydrogen traps with the help of V (vanadium) or Ti (titanium), where hydrogen will be held up and where there will be less hydrogen embrittlement as a result.
- bringing up fracture toughness by fineness of grain formers such as V, so that tension peaks at corrosion scars can be more easily reduced as a result of plastic distortion.

Using such modifications of materials, such steels as the so-called HPM190 were developed by changing chemical composition on the basis of classical 54SiCr6. Especially in Japan, other alloys were also developed. However, they always limited the strength of steel, so that it was not possible to use the weight reduction potential to the full. Over and above, such modifications of materials using alloy elements are very expensive and they very often result in "special steels." These can only be used for individual spring or car factories and they will not be made in sufficient quantities to gain a sustained foothold in the market.

Over the past few years, a number of very interesting techniques were brought to series production that increase toughness at high strength. Thermomechanical treatment offers one case in point. Here, material will be formed during the austenitic stage in such a way that a very fine grain structure will be produced during the martensitic state that results in very good mechanical properties.

Classical tempering in various types offers additional methods to bring about very small grain sizes. The temperature for austenitization is quite important for the grain size evolution. An adequate method to create a fine microstructure in the rim of a wire could be a heating at high temperatures for hardening the whole wire section, followed by an austenitization at low temperatures just within the rim to create a fine microstructure (Liu, 2011). Other methods such as intensive mechanical working by multiple drawing processes as well as subsequent simple or multiple tempering were investigated. Here, a multitude of nuclei generated by mechanical working will be used to produce small grains. Shaping and tempering parameters need to be very finely balanced for all these methods. Surface layer modification is a procedure of its own (Neubrand and Hartwig, 2009), in that it combines a tough lower strength marginal layer with the hard core of the wire. Giving it some thought, this method can be integrated into inductive wire tempering processes, which makes it very economical (Figure 10).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

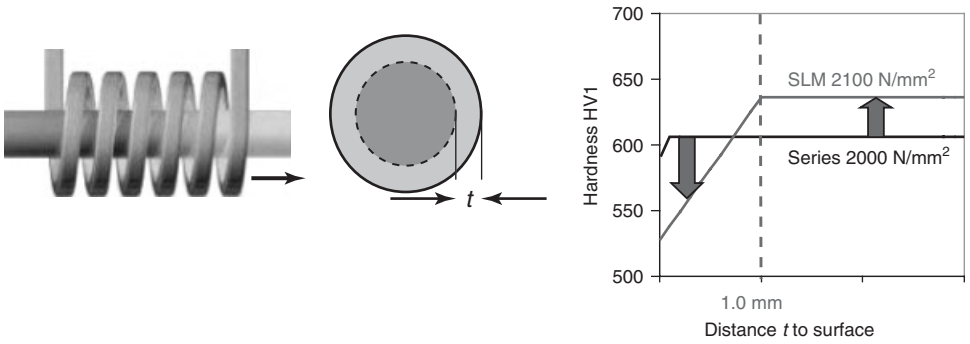


Figure 10. Mubea surface layer modification.

Over the past decade, alternative materials have been investigated to a large extent but could not really move from applications in racing and exotic cars to volume car manufacture. Some particular mass production applications occurred on and off, such as titanium springs for the VW Lupo FSi (Schauerte *et al*., 2001) and composite leaf springs for the Daimler Sprinter and VW Crafter, but the main reason against greater usage in volume cars can be defined as poor cost competitiveness compared to steel springs. Nevertheless, steel still has some further potential for improvement, but the properties are going to reach a certain limit and therefore, new material will become of greater importance, also in the spring world.

6 COMPOSITE SPRINGS

From the materials point of view, composite materials harbor specific potential. Of practical relevance are combinations of glass and carbon fibers with pressure setting plastics material. Thermoplastics materials are also gaining importance as matrix materials; however, because of technological restrictions, they have been unable so far to meet suspension challenges at high dynamic loads. Glass fibers represent a better material for spring application because their lower elasticity module compared to carbon fibers is favorable regarding high strokes and deforming requirements. Given their high specific strength and the stiffness of composite materials, it becomes in principle possible to achieve weight reductions ranging from 30% to 70%. While reducing unsuspended masses it is also possible to reduce driving dynamics and also as regards noise, vibration, and harshness (NVH) behavior, composite materials may offer advantages when compared to steel, because their material amplitudes are very much higher. In view of high corrosion resistance and resistance against the influence of other environments, surface protection mostly is something that can be discarded, but special protection will be needed when it comes to possible damage by rocks. Studies into the comprehensive energy input for making compound components have shown that their CO_2 footprint is larger than that for steel making (Geuder, 2004). However, this may be compensated for partly or wholly in view of the considerable reduction in quantities required during operational cycles.

However, the potential of this group of materials is also limited by serious disadvantages, which have so far prevented the use of fiber composite materials in large quantities. The load transmission is demanding special designs. Typically, quite frequently high loads come up cross-wise to the main load direction so that the material is not taking up a load in an ideal way and following the direction of the fiber, so that only medium loads may be imposed on the matrix. In addition, allowances will have to be made when it comes to large series production and available manufacturing processes when comparing these with units made of steel. At present, these represent the focal point of research and development efforts all over the world.

Continuous reinforced compound fiber materials as they are used for structural elements in car-making reveal strong anisotropic, that is to say, direction-dependent properties. Fibers employed will be oriented as to the loads occurring so that they take up tension and pressure, if possible. Typical fiber volume percentages for structural elements reach about 60%.

When it comes to suspension, leaf springs made of composite fibers represent an application offering high potential for glass-fiber-reinforced plastics in the chassis section that may bring about industrialization and large quantity manufacturing (Figure 11).

Predominant bending requirements for leaf springs offer a possibility to align fibers in one direction only. When compounded, their elasticity modulus may reach about 45,000 MPa. Compared to conventional multi-leaf springs made of steel, it is possible to score considerable reductions

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 11. Tension leaf spring made from glass-fiber-reinforced plastics (Schürmann and Keller, 2010).

in weight, amounting up to 70%. The number of leaves in a multi-leaf spring may be reduced; sometimes springs may be replaced by a one-leaf spring only. Tension leaf springs (Schürmann and Keller, 2010) represent a specific development, which might replace multi-leaf springs. Over and above reduced weight, it might also be possible to set up progressive spring characteristics.

Using fiber compound materials, it may also be possible to improve damping properties. Contrary to steel leaf springs, spring eyes will not be shaped, they have to be linked to the interfaces with a wedge. A number of different types emerged for steel leaf springs, and they come in differing width and breadth. Such structural shapes may also be offered in glass-reinforced fibers. The simplest shape is a rectangular spring with a constant cross section. Other shapes of cross sections lead to (approximate) constant bending stress, which means a constant use of material employed. Hyperbola-shaped springs may easily be manufactured of glass-fiber-reinforced plastics. In this type of spring, there is a linear reduction of width from the center, whereas the cross-sectional area remains constant. For springs of this type of manufacture, no cut fibers need to be used and it is quite easy to make a blank, or a preform, of various sections. There is a constant width for parabolic springs and their width will be adapted to the load course. Given their varying cross sections, different quantities of fibers will be needed over the course of the spring. This is equivalent to more input in complex manufacture; on the other hand, parabolic springs are weight-optimized structural types, and constant width is in keeping with the installation space situation in a car (Franke, 2004). It is also possible to bring about a compound-fiber compromise between a rectangular and a parabolic spring, by designing a smaller change in width than for the parabolic spring, which may be achieved by a variable fiber volume percentage (Götte, 1989), so that fibers will not have to be cut. Variations in the range 45–65% in the fiber volume would be possible. In practice, however, such low fiber volume percentages may lead to unequal distribution of fibers over the entire cross section.

There is a large number of manufacturing methods for the production of fiber-plastics compounds. Because of the resulting good properties at dynamic loads, the prepreg method stood the test of time for leaf springs. Prepreg processes produce optimum bond strength (Schürmann, 2007).

Component manufacture here first involves impregnating fibers with resin so that subsequent parts of the process may work with strip sections easy to handle (Figure 12).

An impregnated semifinished product represents the first step (preimpregnated fibers). To make it, the resin system and the fiber will be brought together on backing paper while adding some heat. The cross-linking reaction of the resin begins when a prepreg is made. This process, however, may be interrupted by cooling, which means that a weakly cross-linked stage will be reached, and the resulting material may be stored when cooled down. Afterward, prepreg strips will be cut and a blank will be prepared. The blank will then be put into a compression mold and cured at high pressure and temperatures between 110°C and 170°C. Springs will then be subject to mechanical postprocessing.

It is also possible to manufacture leaf springs using a resin injection process. For this resin injection process,

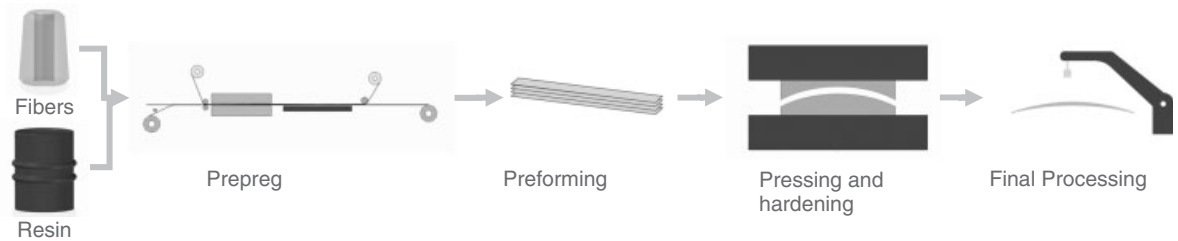


Figure 12. Sequence of operations during spring leaf manufacture, using the prepreg process. (From Müller, 2012. Reproduced by permission of David Müller, Mubea © Mubea.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a fiber structure will first be manufactured of the dry reinforcing fibers, which follows the component geometry required. Structural cohesion may be achieved as needed using textile methods, such as sewing processes or involving binder, gluing fibers together. Such fiber structures are known as *preforms*. This preform will then be inserted into a pressing tool, which may be used to inject the resin system. The tool has the jets required and the resin system will be fed into this unit by a mixing device. Resin injection comes in a large number of varieties, distinguished primarily by the type of injection used. Pressure, vacuum, and combined processes are used to achieve the desired results. There are also variations of closing the tool and applying press pressures. The term *resin transfer molding* (*RTM*) is used for the group of pressure-supported resin injection methods (Mitschang and Neitzel, 2004). Compared to prepreg methods, resin injection processes use resins with considerably lower viscosity, as fibers will have to undergo a wetting process as complete as possible within a very short time frame. This is decisive for resultant compound properties and at the same time, this is also the decisive disadvantage for this particular process when manufacturing thick-walled components.

Conventional suspension springs, used as helical suspension springs, offer another field of application for fiber compound materials. Such helical compression springs are mainly made of steel today. Figure 13 shows different types of suspension springs, which are currently investigated regarding feasibility in fiber composites. Specific properties of fiber compound materials are integrated into their design in various types of manufacturing (Müller, 2010). In the case of a mere materials substitution when comparing a steel spring and a helical compression spring (Sardou, 2002), the direction of the fibers must be chosen in such a way that torsion loads may be borne by the spring in an appropriate manner. A classical orientation is to position fiber layers interchangeably in $+45^\circ$ and -45° directions. This corresponds to the direction of the main normal tension in cases of pure torsion.

For spring cross sections, both a circular cross section and a circular ring cross section similar to a tube have been proposed. The circular ring cross section not only offers additional weight-saving potential but it also involves more installation space for the finished product. A significantly reduced shear modulus of compound fiber materials using a $\pm 45^\circ$ layering when compared to steel results in the fact that the spring geometry of a steel spring cannot be transferred, but that it must be rethought. Composite fiber materials have a low compressive resistance. Therefore, the spring interfaces must differ from those for steel springs, and the design of these must consider the reduction of surface stresses.

A bellows-type spring structure offers one possibility of materials-adapted design (Marquar *et al*., 2010a).

As opposed to a conventional helical spring, this spring element bears loads by bending and membrane tension. Thus, it is possible to make full use of fiber strength.

A third type of design is the meander-type spring (Kobelev *et al*., 2008). It is almost exclusively bearing bending loads, which means that it may also be used effectively for compound fiber applications. Subdivided into two elements, this type of spring may also be placed around a shock absorber.

First concepts are also available for designing wheel-bearing units and shock absorbers to be made of compound fiber material. Because of high stiffness demands, carbon-reinforced plastics are given preference here (Marquar *et al*., 2010b). Compound fiber material is very appropriate for such complex components, because the number of elements may be brought down and additional functions

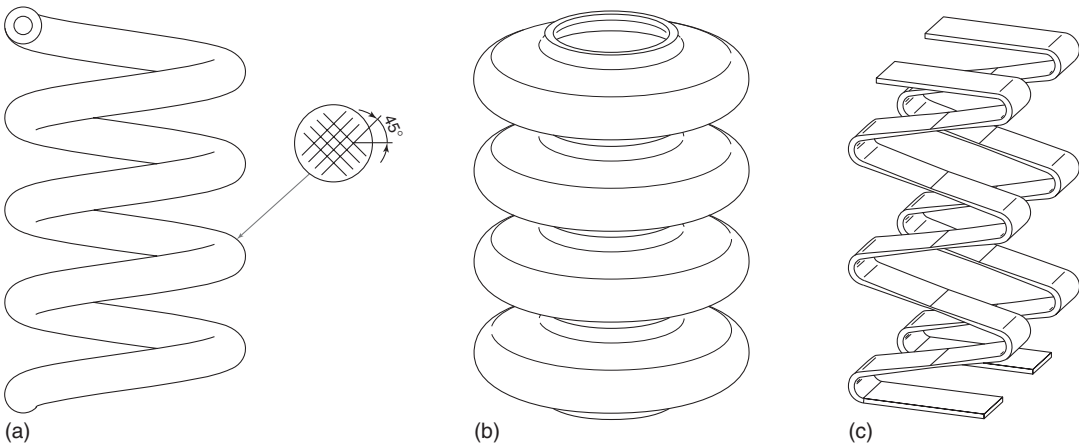


Figure 13. Different types of composite springs: (a) coil spring, (b) bellows-type spring, and (c) meander-type spring.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

can be integrated. As already known from air transport, it is for instance possible to integrate sensors directly into such material, so as to protect them against external stress.

A significant disadvantage of compound fiber materials is to be seen in that manufacturing technologies lack sufficient flexibility to represent various forces and constants as taken up by springs. Given a large platform, a great diversity of brands, engine types, and configurations must be implemented for the same installation space with similar connection concepts. This means that quite readily one may need over one hundred different spring force/spring constant combinations for one platform and for one axle at each platform. For the present compound fiber concepts, this would need a new tool for every new component number to produce the exact spring geometry. In case of steel, an isotropic material, this can be achieved fairly easily by varying wire diameters, the number of active windings, or the diameters of coil windings.

REFERENCES

Brandt, R. Kobelev, V., and Neubrand, J. (2007) Simulation und FE-Analyse der Fahrzeugfedern. Presented at the Seminar Fahrzeugfedern, Technische Akademie Esslingen, March 14.

Brandt, R. and Neubrand, J. (2001) Kaltformtechnik für PKW-Tragfedern, Kontrolle der Kraftwirkungslinie und ihr Einfluss auf die Dämpferreibung unterschiedlicher Achssysteme. Presented at Fahrwerke, ihre Komponenten und Systeme, Haus der Technik Essen, February 2.

Carlitz, A. and Neubrand, J. (2008) Federn und Stabilisatoren in Fahrwerkhandbuch (eds B. Heißing and M. Ersoy), Vieweg Verlag, Wiesbaden. ISBN: 978-3-8348-0444-0

Carlitz, A. Neubrand, J., and Hengstenberg, R. (2005) Radaufhängung mit Federverstellung für Kraftfahrzeuge. Patent EP05025854.0-1264, Muhr und Bender, November 26.

Franke, O. (2004) Federlenker aus Glasfaser-Kunststoff-Verbund—Spannungs- und Festigkeitsanalyse zur Optimierung eines hochbelasteten Bauteils, Shaker Verlag, Aachen.

Georges, Th. (2000) Zur Gewichtsoptimierung von Fahrwerkstragfedern unter besonderer Beachtung des schwingfestigkeitsmindernden Einflusses bruchauslösender Fehlstellen im Halbzeug Federdraht. Dissertation. VDI Verlag, Düsseldorf.

Geuder, M. (2004) Energetische Bewertung von Windkraftanlagen. Diplomarbeit FH Würzburg-Schweinfurt.

Götte, T. (1989) Zur Gestaltung und Dimensionierung von Lkw-Blattfedern aus Glasfaser-Kunststoff, VDI-Verlag GmbH, Düsseldorf.

Kobelev, V. Neubrand, J. Brandt, R., and Lebioda, M. (2002) Radaufhängung. Patent DE10125503C1, December 12.

Kobelev, V., Westerhoff, K., Neubrand, J., Brandt, R., and Brecht, J.D. (2008) Patentschrift EP 2 082 903 B1; May 21.

Liu, Y. (2011) The analysis of the influence in grain refinement through increased dislocation density by cold drawing and heat treatment of Si-Cr spring steel, depending on the degree of deformation. Masterarbeit. Universität Siegen/Muhr und Bender.

Marquar, H., Schuler, M., Renn, J., *et al*. (2010a) Offenlegungsschrift DE 102010040142 A1, Anmeldetag September 2.

Marquar, H., Schuler, M., Renn, J., *et al*. (2010b) Offenlegungsschrift EP 2295826 A2, Anmeldetag September 7.

Mitschang, P. and Neitzel, M. (2004) *Handbuch Verbundwerkstoffe—Werkstoffe*, Verarbeitung, Anwendung, Hanser Verlag, München Wien.

Muhr, K.-H. and Schnaubelt, L. (1991) Radaufhängung mit einem radführenden Federbein. Patentschrift DE3743450C2, March 28.

Müller, D. (2010) Entwicklung eines Prozesses zur Fähigkeitsbewertung von dynamisch beanspruchten Federelementen aus FVW. Masterarbeit. Universität Siegen/Muhr und Bender.

Neubrand, J. (2004) *Entwicklungstendenzen bei Werkstoffen für Fahrwerksfedern, Beitrag zur Tagung Federung und Dämpfung*, Car Training Institute, Düsseldorf.

Neubrand, J. Brandt, R. Junker, C., and Lindner, A. (2010) Light weight suspension coil springs by advanced manufacturing techniques and innovative design definition methods. SAE World Congress.

Neubrand, J. and Hartwig, M. (2009) Gehärteter Federstahl, Federelement und Verfahren zur Herstellung eines Federelements. Patent EP 2 192 201 A1, Muhr und Bender, November 23.

Sardou, M. (2002) Patent FR 000002837250 B1, Anmeldetag March 18.

Schauerte, O., Metzner, D., Krafzig, R., *et al*. (2001) Fahrzeugfedern federleicht. *Automobiltechnische Zeitschrift*, **103**, 654 ff, Wiesbaden.

Schürmann, H. (2007) *Konstruieren mit Faser-Kunststoff-Verbunden*, 2.Auflg., Springer Verlag, Berlin.

Schürmann, H. and Keller, T. (2010) GFK-Blattfeder für Transporter Hinterachse mit progressivem Verlauf der Federkennlinie ohne Federaugen. Patentanmeldung 102010015951.4, TU Darmstadt, March 12.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto001
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Suspension Arms, Steel versus Aluminum, Where are the Benefits?

Dirk Adamczyk, Markus Fischer, and Klaus Schüller

ZF Friedrichshafen AG, Friedrichshafen, Germany

1 Introduction 1
2 Key Factors for The Design of Suspension Arms 3
3 Suspension Arm Types 9
4 Suspension Arm Technologies 13
5 Design Matrix (Judgment of Technologies for Different Suspension Arm Types) 20
6 Summary and Outlook 23
References 23

1 INTRODUCTION

Suspension arms or links connect the wheel or the wheel carrier with the vehicle body, usually via a subframe, and allow with respect to the vehicle chassis the controlled motion of the wheel. They have to transmit the loads from braking, cornering, vehicle acceleration, and impacts from the road surface. There are even strong crash requirements for the suspension arms. Crash or controlled buckling and durability are key targets for suspension arms. In special axle types, the suspension arm has to transmit the loads from vertical acceleration and "carry" the vehicle load. In this case, the spring is directly or indirectly assembled to the arm (Heißing, 2011).

Owing to the fact that the control arm is assembled to the wheel carrier, the mass of the arm is at least partially unsprung mass. On the basis of this, lightweight is one of the most important design targets for a suspension arm as it supports two design targets on vehicle level:

1. General low vehicle weight to reduce CO_2 emissions and
2. To improve the vehicle performance, for example, improved vehicle acceleration and improved vehicle safety and comfort by reducing the unsprung mass (Adler, 1991).

Suspension arms are highly integrated in the overall vehicle package. Main drivers are the required wheel or tire clearance and the ground clearance. As on the one hand, wheels are getting bigger and bigger with every new vehicle model, and on the other hand, engines and especially hybrid engines require more and more package space, the allowable space to package suspension arms gets more and more limited.

Beside all functional and packaging requirements, there is a very strong focus on costs and the manufacturing/assembly process. As suspension arms have a critical function for the vehicle, it is essential to have a strong focus on quality and robust production processes that have to be globally available for worldwide vehicle platforms.

1.1 Where are the benefits?

We have to take aluminum to get the best weight and steel to get the best cost.

Of course, it is not as simple as this. Although for many applications, the sentence is true.

Considering the material data (Table 1), it is easy to imagine that there is no simple answer to the question: Steel versus aluminum – where are the benefits?

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Table 1. Mechanical properties of steel and aluminum.

	Steel	Aluminum
Young's modulus	210 Gpa	70 GPa
Density (ρ)	7850 kg/m^3	2700 kg/m^3
Tensile strength (typical high end)	800 MPa	Up to 380 MPa

Some thoughts about ratios between the material properties of aluminum and steel:

$$\frac{\text{density (steel)}}{\text{density (alu)}} = \frac{7.85\frac{\text{g}}{\text{cm}^3}}{2.70\frac{\text{g}}{\text{cm}^3}} = 2.91 \quad (1)$$

$$\frac{\text{young's modulus (steel)}}{\text{young's modulus (alu)}} = \frac{210\,\text{Gpa}}{70\,\text{Gpa}} = 3.00 \quad (2)$$

$$\frac{\text{Tensile strength (steel)}}{\text{Tensile strength (alu)}} = \frac{800\,\text{MPa}}{380\,\text{MPa}} = 2.11 \quad (3)$$

One of the main drivers for using aluminum in automotive business is its weight reduction potential. The density of aluminum is 2.91 times lower than for steel (1). At the same time, the tensile strength is lower as well, but only 2.11 times lower (3). This means, if tensile strength is essential for the design, it is still possible to achieve a major weight reduction.

An example that considers previous material facts is calculated (Figure 1), based on a simple tension rod. A tension load of 40 kN was chosen, as it is an average force in today's mid-sized passenger cars.

This simple assumption of a rod tension loaded by a one-axis static load shows the aluminum weight-saving potential of 27.7%. In both cases, the tensile strength of the respective material is on the high level of the forging alloys that are typically used in the automotive industry for chassis components (380 vs 800 MPA).

If the stiffness is essential for the design, with the Young's modulus being the relevant key parameter, then there are some slight advantages for steel regarding weight, as the Young's modulus of aluminum is even 3.00 times lower than the Young's modulus of steel (3).

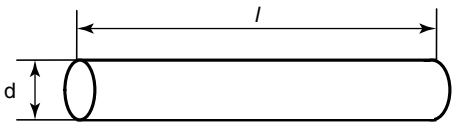
For most of the control arms, the required buckling load is important for design and weight. A basic factor for the buckling load is as well the Young's modulus. Taking a simple straight control arm, the following ideal example shows that the steel version is even lighter than the aluminum version – at least the middle section (Figure 2).

For the buckling calculation, there are different buckling conditions (Figure 3), but for typical suspension arms, the mode (2) has to be used.

For the calculation in the example, the following extended equation will be used. The equation includes bending/offset (Figure 4) to cover, for example, tolerances.

$F_{\text{Buckling}} = \frac{\delta_{\text{steel/alu}}}{\frac{e+f}{W}+\frac{1}{A}}$; based on offset [$e$]; additional displacement/bending [f] under load [F]; maximum stress for steel or alu [δ]; resistant torque [W]; cross section [A].

F_{max} = 40 kN (Load applied to the rod)
l = 300 mm (Total length of the rod)
$R_{\text{m steel}}$ = 800 MPa (Ultimate tensile strength)
$R_{\text{m alu}}$ = 380 MPa (Ultimate tensile strength)



$$\sigma = \frac{F}{A} = \frac{F}{\frac{\pi d^2}{4}} \quad \text{(Stress)}$$

$$d = 2\sqrt{\frac{F}{\sigma\pi}} \quad \text{(Required rod diameter)}$$

$$m = \frac{d^2\pi}{4}\, l\, \varrho \quad \text{(Mass of the rod)}$$

$$d_{\text{steel}} = 2 \cdot \sqrt{\frac{40\ \text{kN}}{800\ \text{MPa} \cdot \pi}} = 8.0\ \text{mm};\ m_{\text{steel}} = \frac{(0.8\ \text{cm})^2 \cdot \pi}{4}\ 30\ \text{cm} \cdot 7.85\ \frac{\text{g}}{\text{cm}^3} = 118.4\ \text{g}$$

$$d_{\text{alu}} = 2\sqrt{\frac{40\ \text{kN}}{380\ \text{MPa} \cdot \pi}} = 11.6\ \text{mm};\ m_{\text{alu}} = \frac{(1.16\ \text{cm})^2\,\pi}{4}\ 30\ \text{cm} \cdot 2.70\ \frac{\text{g}}{\text{cm}^3} = 85.6\ \text{g}$$

Figure 1. Example: a mass comparison of a steel rod and an aluminum rod under the same load requirement.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

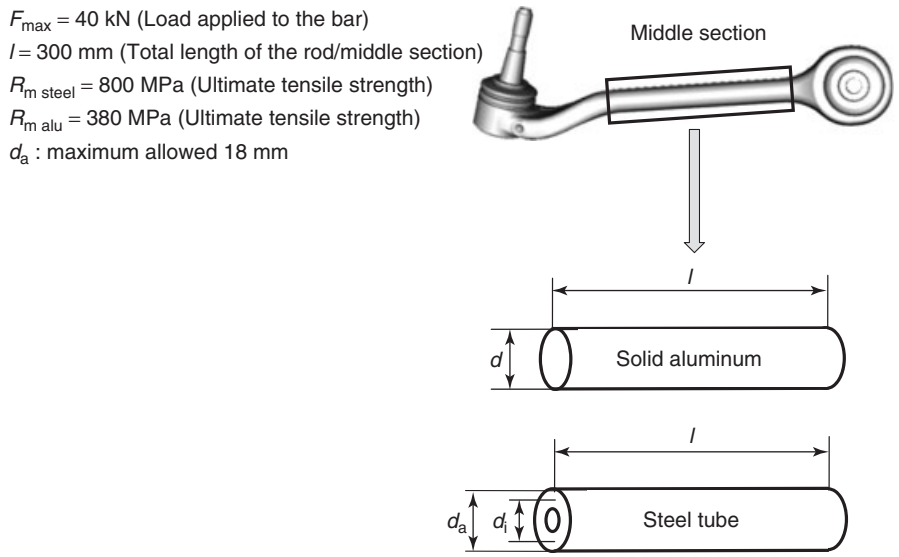
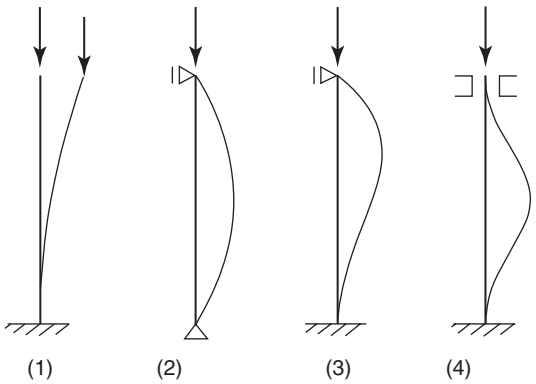


Figure 2. Example: middle section of a straight control arm.



(1) Rigid clamping/free, (2) Trivalent joint/bivalent joint,
(3) Rigid clamping /bivalent joint, (4) Rigid clamping /univalent joint

Figure 3. The four Euler buckling modes. (Reproduced from Beitz and Grote, 2001. © Springer Science+Business Media.)

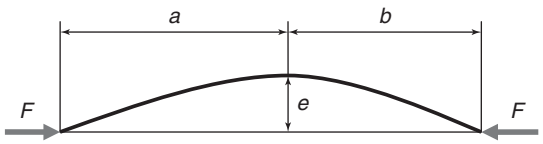


Figure 4. Buckling including offset.

$$f_{\text{buckling}} = -\left[\frac{e+W\left(\frac{1}{A}-x\right)}{2}\right] \pm \sqrt{\left(\frac{e+W\left(\frac{1}{A}-x\right)}{2}\right)^2 + W \cdot e \cdot x};$$

displacement [f] under buckling load [F]; $x_{\text{buckling}} = \frac{\delta_{\text{steel/alu}} a \cdot b}{3 \cdot E \cdot I}$; distance [$a$]/[$b$] (Table 2).

The simple comparison (Figure 5) shows that aluminum does not necessarily lead to the best weight solution. In this case, the steel design is approximately 11% lighter than the aluminum design. This example is also a very simplified case based on only one requirement – in reality, further boundary conditions such as interfaces and packaging requirements between suspension arm, subframe, and knuckle have to be considered. Considering all the packaging requirements, the picture will likely change and a weight advantage for the aluminum design can be expected.

There are additional aspects that further influence weight (and costs). In typical sheet metal applications, there are weld seams that reduce the material properties significantly. For forging and casting parts, the tolerances of the raw part (e.g., thickness) have to be considered.

The existing applications show a typical weight-saving potential of 10–50% by the use of aluminum. The more complex a part is and the more it is effected by fatigue, the higher the weight-saving potential is. For the cost side, it is more or less the opposite.

The following sections show the typical influencing factors for the design and the choice of the best material and the manufacturing process for suspension arms.

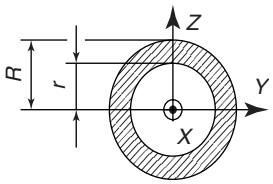
2 KEY FACTORS FOR THE DESIGN OF SUSPENSION ARMS

2.1 Weight and environmental aspects in suspension design

Looking at today's vehicle fleet, vehicle simulations indicate that fuel consumption is reduced by 0.25 l/100 km or

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Annulus.



Annulus

Cross section	$A = \pi \cdot (R^2 - r^2)$
Moment of inertia	$I_y = I_z = \frac{\pi}{4} \cdot (R^4 - r^4) = \frac{A}{4} \cdot (R^2 + r^2)$
Resistant torque	$W_y = W_z = \frac{\pi}{4 \cdot R}(R^4 - r^4)$
	A circle is defined as special case of the annulus with r = 0

Reproduced from Beitz and Grote, 2001. © Springer Science+Business Media.

Given geometry data:
s length = 300 mm
$a = b$ = 150 mm
e = 0.5 mm
d_{max} = 18 mm (R_{max} = 9 mm)

Given data for a steel tube:	Given data for an aluminum rod:
R_{steel} = 9 mm	R_{alu} = 9 mm
r_{steel} = 7.5 mm	r_{alu} = 0 mm
A_{steel} = 77.7 mm^2	A_{alu} = 254.5 mm²
I_{steel} = 2668 mm^4	I_{alu} = 5153 mm^4
W_{steel} = 593 mm^3	W_{alu} = 1145 mm³
$R_{e\ steel}$ = 680 MPa (δ_{steel})	$R_{e\ alu}$ = 330 MPa (δ_{alu})

Results steel tube:	Results aluminum rod:
$x_{steel} = 0.0091 \frac{1}{mm^2}$	$x_{alu} = 0{,}0069 \frac{1}{mm^2}$
f_{steel} = 0.77 mm	f_{alu} = 3,87 mm
F_{steel} = 45 kN	F_{alu} = 43 kN

Conclusion:	
The mass comparison for aluminum and steel version: ($F_{buckling}$: $F_{alu} \approx F_{steel}$)	m_{steel} = 183 g m_{alu} = 206 g $\Delta m \approx$ 11%

Figure 5. Calculation including offset.

even up to 0.4 l/100 km and 0.5 l/100 km (an improvement of ca 1 mile per gallon) for light trucks per every 100-kg weight reduction (Cheah, Heywood, and Kirchain, 2010; Goede, 2007). This figure shows the increasing need for the reduction of weight and in consequence fuel for all vehicle types – vehicles with combustion engine, vehicles with hybrid engine, or electric vehicles. This requirement of weight reduction puts a lot of pressure on new designs and challenges current production processes regarding improvements and optimizations.

The goal of weight reduction nowadays can be realized by the utilization of high strength steel, standard and high strength aluminum, higher strength ductile iron (DI), and the advancement of plastics. Nevertheless, the most common material for weight-optimized links is still aluminum. There are solutions for nearly all imaginable control arm applications, using forging, casting, and performance casting processes as well as cast–forge with various alloys. Only strongly limited package spaces in combination with the need to maintain a certain maximum load level

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

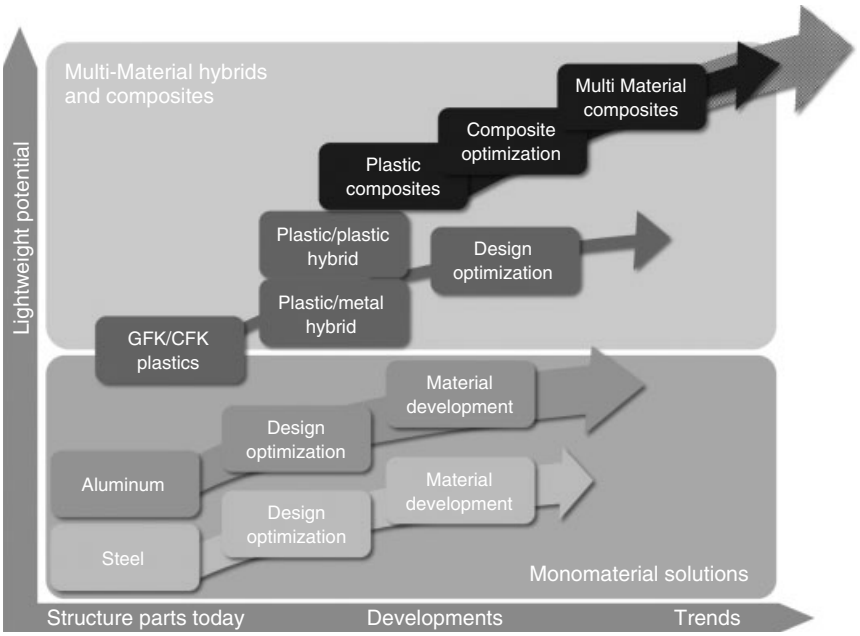


Figure 6. Material mixture as trend for lightweight concepts. (Reproduced with permission from ZF.)

would not allow using the light metal, from a functional point of view.

On the other hand, the energy balance producing aluminum generally speaks against this alloy: the amount of energy used – throughout the process from the raw material to the final product – is higher compared to that used for the processing of steel products.

Nowadays, not only the amount of energy used to produce an alloy is considered but also, seeing the global warming, a more holistic approach is chosen: more and more attention has to be paid to the overall CO_2 emission – not only during the usage of the vehicle but also under cradle-to-grave aspects. A study on a front end of a 2007 Cadillac CTS, for instance, came to the result that the aluminum design achieves the break-even distance from energy use and GHG (greenhouse gas) emission perspectives earlier within the vehicle lifetime (Dubreuil *et al*., 2010). This achievement in overall environmental friendliness has further potential for improvement. It can be reached by considering the capability of aluminum to be recycled indefinitely, which requires only 5% of the original energy to be put back into use (Heidi and Thomas 2011).

Apart from the well-known disadvantages of steel arms compared to aluminum – such as higher weight and the need for corrosion prevention – the steel producers and manufacturers push toward high strength material and smart material mixes. This is to improve weight and fight the losses caused by aluminum and upcoming composite applications. Steel control arms using sheet metal, forging and cast iron alloys, and processes are still being used widely; especially in mid- and low-sized vehicles with standard engines, they show an increasing market share. Reason for this is the focus on cost in this market segment as main driver for design selection, and other considerations have lower priorities.

Figure 6 shows an overview about trends in lightweight solutions similar to the trend we see for the vehicle body (Goede *et al*., 2008). Beside momo-material solutions, multi-material solutions will offer further lightweight potential for chassis.

2.2 General design aspects

Control arms are typically designed for two load conditions – fatigue load cycles and misuse load cases: load duty cycles (fatigue load or "dynamic" load): this is a combination of a number of single load events or block cycles. These loads are generated during "normal" driving conditions.

Single load events (the so-called quasi-static):

(a) Loads that appear frequently: these can be taken out of a load spectrum, as the load level is below the maximum, which means there are no special events or maximum loads included. The criteria are no crack or fracture, no or only minimal plastic deformation.
(b) Loads that appear just a few times over the vehicle life, such as special events and misuse. These load

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

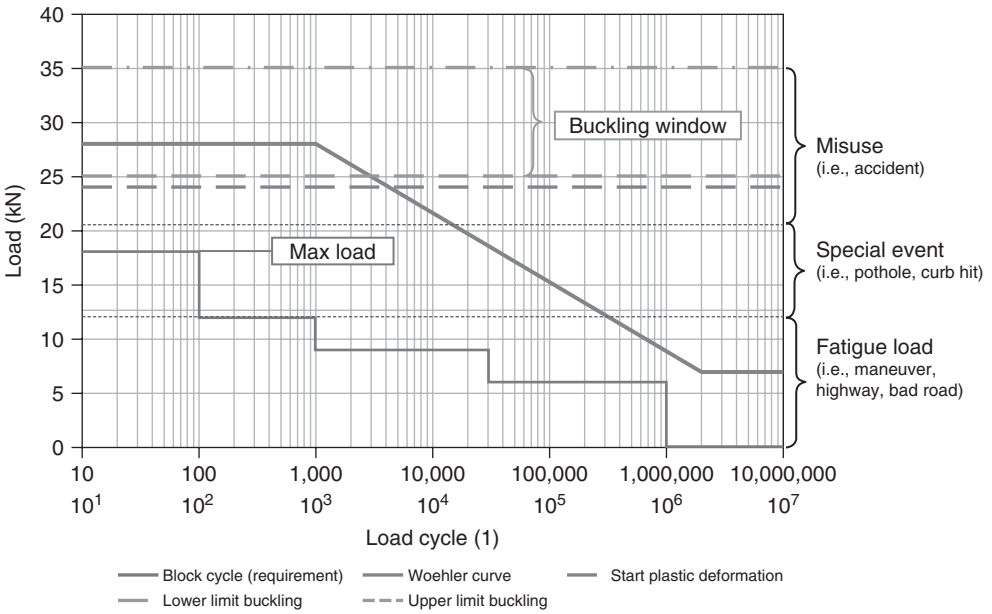


Figure 7. Relation between the load cases fatigue load, special event, and misuse.

cases are usually given through "Nastran-load-decks" (an input data format for finite element calculations) or "Adams-loads" (a software to simulate the kinematic and kinetic vehicle/component function) if not available out of real road load data measurements with test vehicles. They cover various driving maneuvers according to certain load philosophy rules such as "brake over railway tracks," "obstacle overdrive," and "washboard braking" just to name a few possible events. These loads are generated or simulated/calculated through pothole braking, curb impacts, crash (like through an accident) events, and some more. The load level reaches up to the maximum load expected from all relevant events. Depending on customer and design targets, it might be expected that after a special event full fatigue life is still ensured.

The relationship between the different load cases is shown in Figure 7.

Simulation tools check the dimensional layout of control arms before physical prototypes are built. As a standard, linear and nonlinear finite element analysis (FEA) calculations are used for the "quasi-static" cases and fatigue life FEA if needed. In most cases, it is not required to perform a fatigue life FEA because design rules for maximum strength will cover it. Structural or surface-related iteration tools such as "Optistruct" as one example support optimizations of complex structural parts considering the load requirement optimizations.

At the beginning of the design process, the following pieces of information are typically available.

The material selection or –request (material or weight target), load data and calculation requirements (i.e., for instance "Abaqus Skin"). FEA parameters such as bearing constraints (i.e., "RBE" = rigid body elements). Assessment requirements such as a maximum plastic elongation under a permissible load, maximum damage rate for fatigue life or stiffness and buckling requirements.

For suspension arm design, the following aspects are to be considered if applicable:

- Structural related: concept model constraints, special tolerances, target weight, specific section modulus or cross-sectional constraints, and stiffness. A control arm package is sometimes given as a CAD model with all applicable component contacts in worst condition, by minimum distance faces or a kinematics model (Figure 8). The kinematics model contains the kinematics points of all components and distances, as well as parameters for jounce, rebound, and steering, including stiffkinematic (structural components without elasticity content) and elastokinematic (elasticity added, such as from rubber bushings). The kinematics analysis should also consider interim movement positions to catch all possible positions that could touch minimum package requirements. Some defined steering overtravel might be required as well. In addition, it is closely to be judged and decided whether full jounce and rebound of the strut should be reduced to a reasonable amount.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

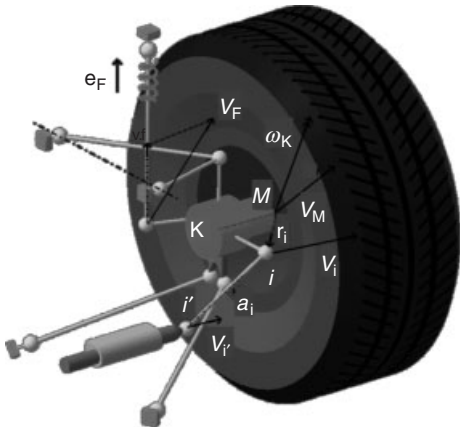


Figure 8. Kinematics model. (Reproduced with permission from ZF.)

Special care should be given to not create any stress raiser notches. Not only radii of main structures can be designed too sharp but also tight transitions to bushing bosses, mold part sections, mold part markings, and surface imperfection due to manufacturing process constraints. This is especially important not only to fulfill the fatigue lifetime requirements but also at fastener connections and for the application of coatings. In addition, pockets that can collect dirt or water have to be avoided or drainage solutions have to be found. Material raw part properties impact the function of a control arm through their mechanical, chemical, and grain conditions; material tolerance ranges; work hardening stiffening; and fracture behavior, which is different from push to pull direction. Furthermore, the geometrical tolerance chain from the raw part over the machining contour and the assembly components has to be considered in the common locator scheme at all production stages to avoid unnecessary machining or assembly effort or conflicts.

Smooth and "flowing" blended transitions are intended to achieve a consistent stress distribution under load. Special attention has to be given to avoid flat, straight, and even control arm structures, which would provide an unintended so-called close-to-Euler buckling case that reacts sensitively – small changes in typical tolerance windows of manufacturing processes would lead to large buckling load variations. Especially, this design rule contradicts the target of minimum weight because "leaving" the direct kinematics line will add weight. As a compromise, structural pinching dedicated to certain load directions can be a solution for cases that cannot allow for "buckling friendly" bending.

In some rare cases, it might be required to design a minimum natural frequency (nuisance avoidance) or minimum energy consumption (crash characteristic).

Table 3. Examples for aluminum.

	High Performance Aluminum Casting	Aluminum Forging Design
Material	SAE A 356-T6	SAE J 454-6082-T6
Yield strength	220 MPa min	240 MPa min
Ultimate strength	290 MPa min	290 MPa min
Elongation	8% min	10% min

Table 4. Examples for steel or iron.

	Ductile Iron Casting	Steel Forging Design
	Example Middle Range	Typical
Material	EN-GJS-500-7	30MnVs6 + P
Yield strength	320 MPa min	560 MPa min
Ultimate strength	500 MPa min	820 MPa min
Elongation	7% min	14% min

- Load related: maximum allowable stress for specific load cases, misuse requirements such as lateral and longitudinal buckling windows (minimum and maximum permissible buckling load) considering push and pull direction. This can be in conjunction with a required defined minimum plastic deformation limit and the direction to ensure the detection of a misuse event by a steering misalignment for instance or a defined travel of control arms relative to other components. In the case of a front impact, the requirement can be to provide a certain limited wheel travel, which must not block the doors for instance. Consequences of plastic deformations in case of hits or accidents may apply to the ball joints as well. It is to be considered that even a defined control arm bend would not separate the ball joint – in order to provide permanent wheel and/or steering control.

Example: Typical mechanical material properties of different manufacturing processes (Tables 3 and 4).

The expression "high performance casting" contains various processes such as counterpressure casting (cpc), vacuum riser-less casting (VRC), pressure riser-less casting (PRC), and cast forge (Cobapress) with high mechanical properties.

2.2.1 Optimization of suspension arms

Figure 9 shows the typical design process for forged and cast control arms starting with package model and the specification (step 1), initial optimization, for example, with

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

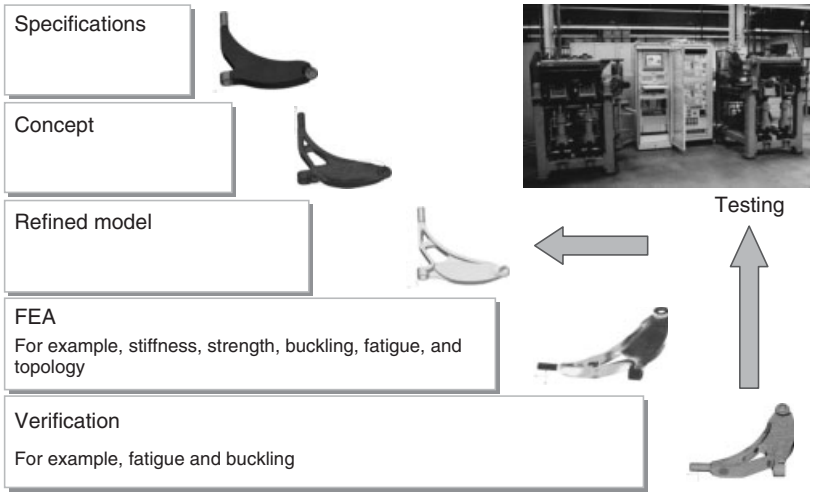


Figure 9. Typical development optimization process via simulation.

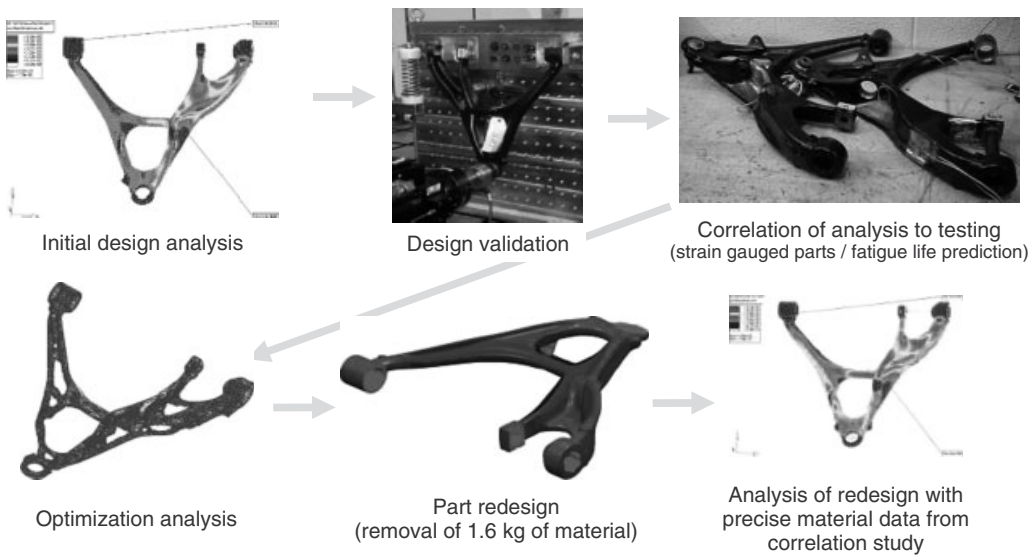


Figure 10. Development optimization including correlation study.

Optistruct® (step 2), optimized design (step 3), FEA (step 4), verification by fatigue simulation (step 5), and finally verification by testing (step 6). In some cases, a shape optimization will be done on top to further optimize some design details. In Figure 10, there is a special process shown to even consider design- and process-specific material behavior based on initial verification tests.

The above-mentioned process takes the optimization process even one step further. Real part material properties depend not only on the alloy but also on the raw part manufacturer and their production processes and variation of mechanical properties. On the basis of this knowledge, parts have been produced, strain gauged and real stress in critical areas of the part have been measured. These data have been used in FEA to further optimize the design based on a selected foundry process and tool design. This approach could reduce the weight by another 1.6 kg over the original FEA with typical material properties for the shown example of a front lower control arm (ZF).

2.3 Buckling and failure chain

Compact and light axle arrangements lead more and more to extraordinary challenges for the design to cover multiple

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

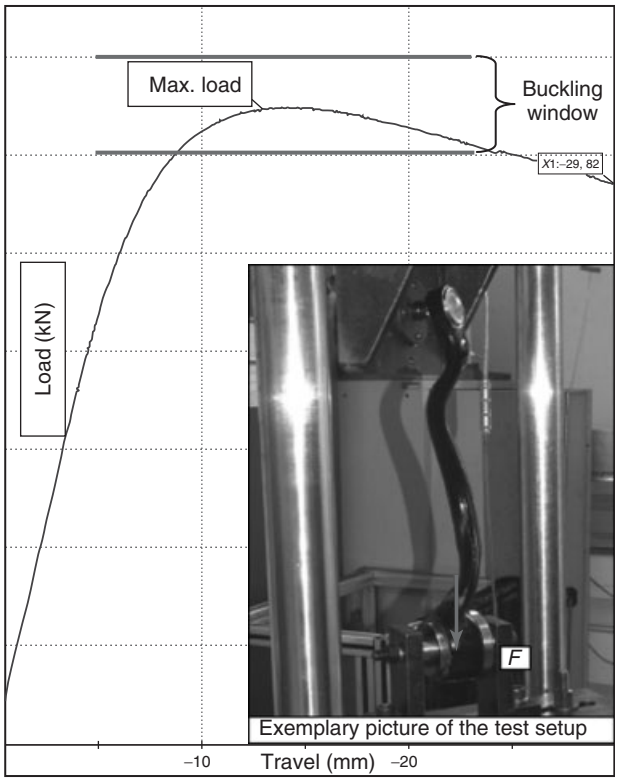


Figure 11. Typical buckling curve measured in compression direction.

load requirements for various load directions. On top of that, the so-called damage chain is defined for the vehicle. The damage chain describes the cascade of misuse load through the chain of components linked to each other, leading to a recognizable failure of single defined components. This follows the principle "leak before break" and means that a driver for instance recognizes a wheel misalignment after hitting a curb with too high speed.

The damage chain would be cascaded down from the major vehicle weight and ability requirements down to each component. For a side impact at the front end, the tie rod needs to deform first to protect the steering gear, and typically a control arm needs to buckle before the knuckle or subframe is damaged. The suspension arms are usually relatively easy to be replaced and therefore lead to low repair costs in case of minor misuse cases. However, this depends also on the manufacturing damage chain philosophy. These components are therefore the so-called sacrificial or victim parts, which act like a mechanical fuse. The damage chain defines the maximum permissible load for structural parts and for link attachments and screw connections. Figure 11 shows an example for a control arm. The buckling window describes the range for the maximum permissible load.

3 SUSPENSION ARM TYPES

3.1 Front axles

Requirements toward a chassis are high and complex and have to be a compromise among but not limited to driving dynamics, ride comfort, safety, weight, cost, reliability, durability, allowable space, crash behavior, and environmental friendliness. Owing to these requirements, the following suspension principles and modifications within those are typical for modern vehicles and SUV as front suspension (Figure 12; Elbers *et al.*, 2004)

3.1.1 Suspension arms for McPherson axles

The McPherson suspension type (Figure 13) is the preferred solution on passenger cars up to the luxury segment – thanks to the simplicity of design, associated cost effectiveness, and cross-car installation space. Nevertheless, this type of suspension has still limitations due to the fact that the shock has to transfer additional loads besides the vertical ones, which can add friction to the system and therefore influences the performance of the suspension. In today's cars and crossover vehicles, the typical design is a rear or front facing L-arm, manufactured out of stamped steel (single and double shell), DI,

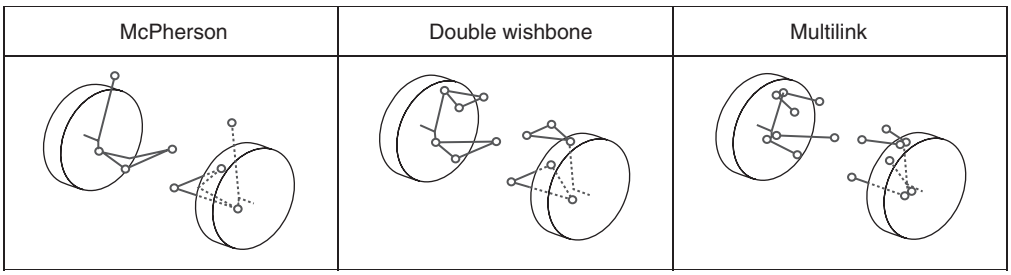


Figure 12. Typical front axle concepts.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

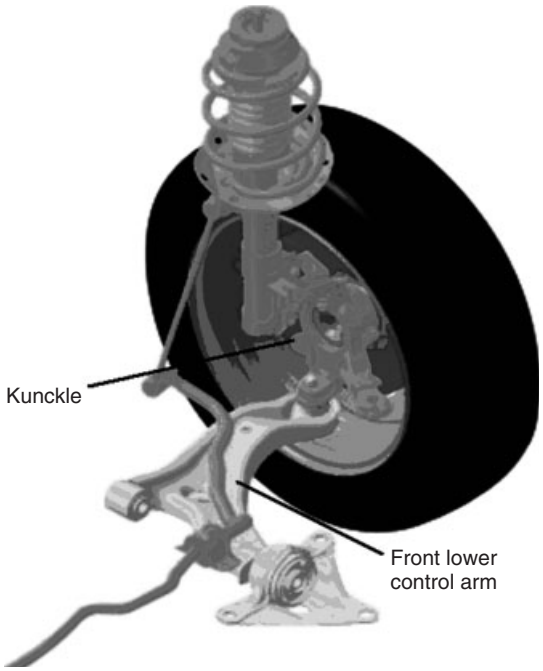


Figure 13. McPherson axle.

aluminum forging, and aluminum casting (Figure 14). From the design perspective, vertical loads are not too high, but the fore-aft and cross-car loads plus the crash performance have to be considered during the design as well as the space requirement for tire envelope, ground clearance, and stabilizer bar.



Figure 15. Strut front suspension with two lower suspension arms.

The figures above show examples for the different technologies for L-shape suspension arms. Application in the vehicle depends on requirements such as loads, packaging space, and cost. Therefore, all of the above-mentioned samples are feasible and production intend.

3.1.2 Two separated arms

To improve the freedom to position the lower hard point (virtual hard point) and therefore also to move the upper shock mount further out, the lower control arm can be replaced by two links (Figure 15 – links are marked red), typically called *tension* and *lateral* or *compression link*. These arms are typically designed out of forged steel or forged aluminum, but, if clearance and deformation requirements allow casting or tubular, weld designs are also feasible (Figure 16).

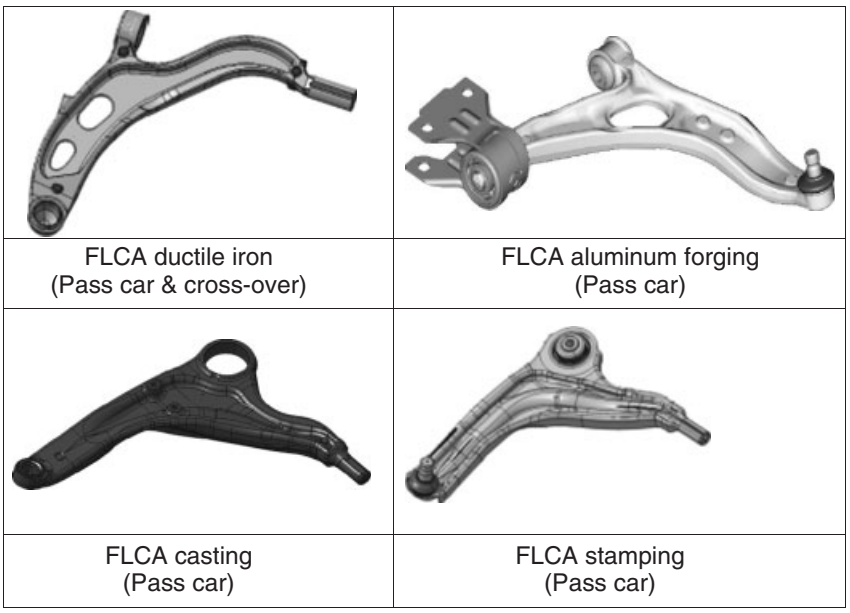


Figure 14. Examples: front lower controls arms.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

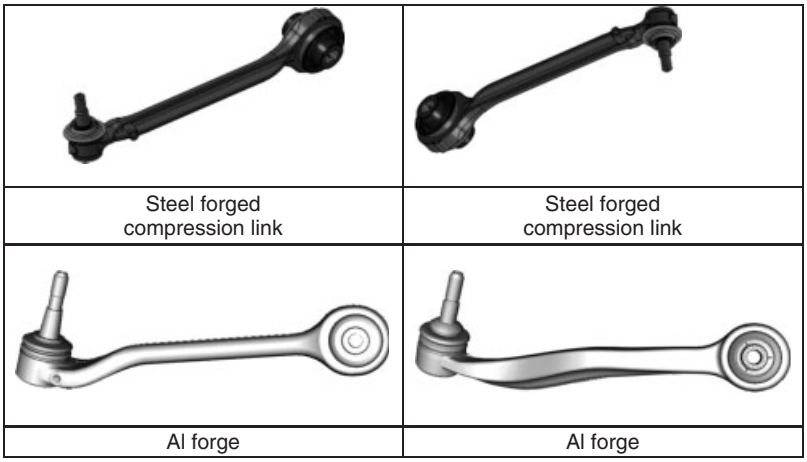


Figure 16. Examples: front lower compression and tension links.



Figure 17. McPherson suspension with revolute joint.

A special modification of the McPherson suspension is the revolute joint (Figure 17), which improves the scrub radius and the king pin offset; it therefore can be used for high powered front-wheel-drive vehicles.

3.1.3 Suspension arms for double wishbone

The double wishbone suspension or SLA (short long arm) is typically used on SUVs, pickup trucks, and luxury passenger cars. The big advantage of this suspension is that the shock absorber is decoupled from the wheel guiding function. The lower control arm carries the vertical forces in addition to driving loads and the upper control arm on an SLA suspension has to provide clearance for the strut mount.

3.1.4 Lower triangle arm, A-arm, and V-arm

Lower arms can be made out of DI, steel forging, aluminum casting, or complex multipiece stamping. All of those have to provide attachments for shocks and springs, which can be torsion bars or coil springs (Figure 18).

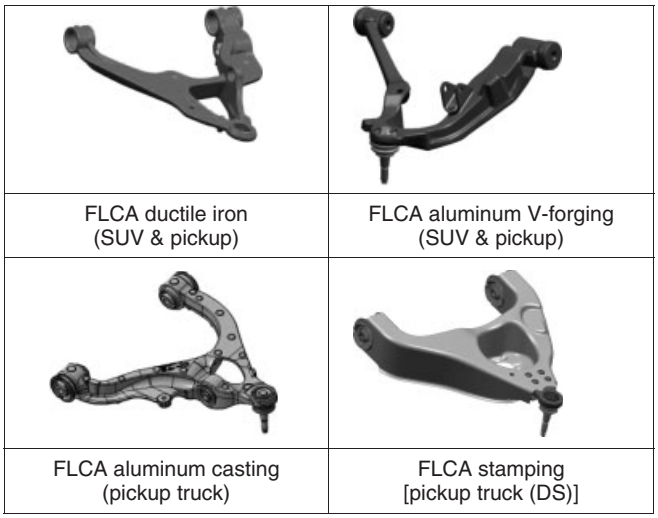


Figure 18. Example: SUV front lower control arms of double wishbone suspension.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.1.5 U-shape arm

The upper control arm of a double wishbone suspension (Figure 19) has a typical "U-shape" to provide clearance for the strut assembly, which makes it sometimes difficult to achieve the required stiffness values. Owing to the position of the arm, the vehicle loads are low, but the arm has to be built short to allow space for the engine compartment. Therefore, the ball joint and the bushings have high requirements regarding working angles especially on SUV and pickup trucks.

3.1.6 Separated arms (preloaded arm and wheel-guiding arm)

Similar to the strut suspension with two arms, a double-wishbone axle can be designed with two arms on the lower and/or upper level for the same reason (Figure 20).

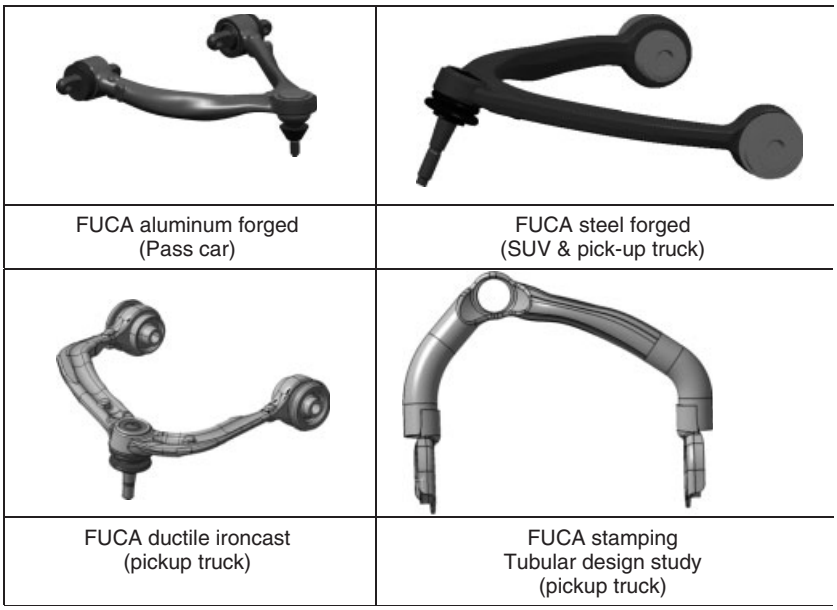


Figure 19. Example: front upper control arms.

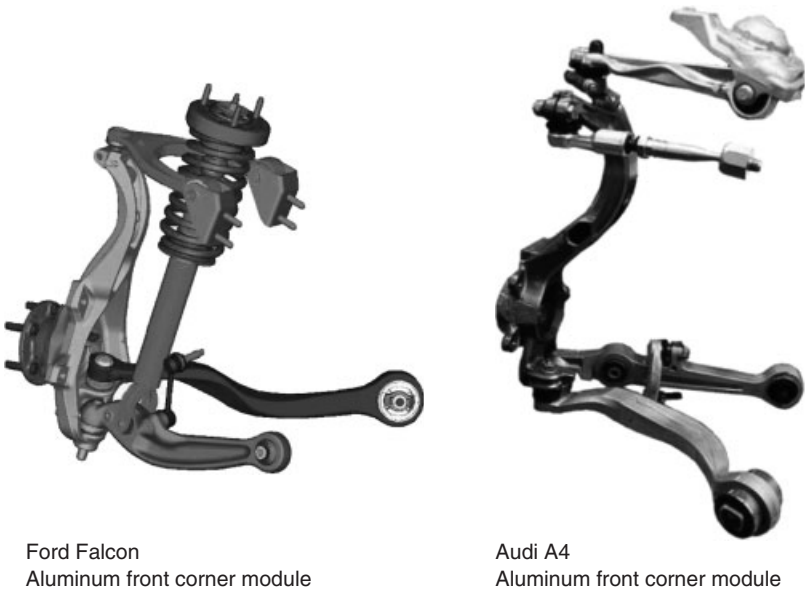


Figure 20. Examples: multilink front axles.

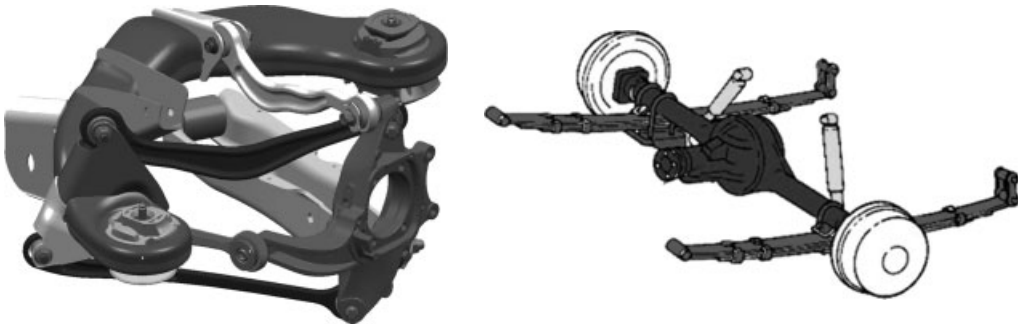
Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



BMW 5series
Aluminum forged front lower control arm with integrated ball joint

Rolls Royce Ghost
Aluminum forged front lower control arm with three integrated ball joints

Figure 21. Examples: preloaded suspension arms.



Multilink suspension

Solid pickup axle

Figure 22. Rear axle types.

3.1.7 Preloaded suspension arms

For multilink axles, one of the control arms has to carry the load of the spring/strut as a permanent static preload (Figure 21).

3.2 Rear axles

The typical modern suspension can be a multilink rear suspension, for smaller vehicles a twist beam axle and for pickup trucks – due to high payload – a rigid rear axle (Figure 22).

3.2.1 4-point suspension arm

Especially, for premium rear axles with higher loads, we can often find suspensions with a 4-point suspension arm. Owing to the size of the arm, there are a lot of aluminum solutions available for this kind of control arms. Owing to high stiffness requirements and complex load conditions, the hollow cast solution offers the best weight-saving potential (Figure 23).

4 SUSPENSION ARM TECHNOLOGIES

Today's independent passenger car and truck suspensions use either ferrous material or aluminum. Within those materials, forging, casting, and stamping are the prime choices in ferrous design; casting and forging, potentially extrusion, are the choices in aluminum design. Castings, especially new riser-less technologies, provide freedom to design sections on the suspension with a minimum of unused material (sprue and flash). Forging and sheet metal can provide higher strength materials, but typically the rate of nonutilized material is higher than with casting technologies (flash, trim-off, and raw part net weight vs weight of the forged part).

The different technologies stay in constant competition; the technologically and commercially best solution has to be found for each application, based on vehicle goal and target requirements (Figure 24).

4.1 Integrated ball joint

What is special about suspension "die-based" technologies such as forging and casting is the possibility to integrate the ball joint directly into the structural component element. Although this is a common technology, ZF (Figure 25)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 23. Four-point suspension arm. (Reproduced with permission from ZF.)

	Weight	Costs	Package
Forged aluminum	+++	O	+
Cast aluminum, LPPM	++	+	O
Special cast aluminum processes	++	++	+
Forged steel	O	++	++
Ductile iron (*)	O/–	+++	++
Sheet metal integrated control arm (SMICA)	+	++	+
Sheet metal CA + press-in ball joint	+	++	O
Sheet metal CA + flange ball joint	+	+	+

(*) Crossover, SUV, pickup truck / passcar

Figure 24. Technology matrix.

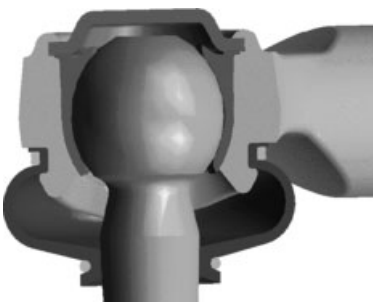


Figure 25. Integrated ball joints (ZF Friedrichshafen AG). (Reproduced with permission from ZF.)

has improved the functional performance of such joints to implement low or, if needed, high motion torques with small tolerance range on the one hand and lifetime wear resistance in each driving condition on the other hand. Furthermore, the position of the hard points to each other is kept extremely tight with minimum tolerances. Special ball joint types cover different needs such as suspension- or compression-loaded variations.

In general, this is different to sheet metal, which requires a separate bolted, riveted, or pressed-in ball joint solution. It is obvious that an integrated ball joint has advantages compared to others. Pressed-in ball joints can be just right from a functional performance point of view but, when assembled, the press fit can have negative impacts on the motion torque of the ball joint. This is caused by the design of the press fit with the respective tolerances and the selected stiffness of the hub. A press fit has to be chosen to maintain a minimum of press-out to withstand vehicle loads, which can as a result double or triple the motion torque of the ball joint.

To provide the advantages of integrated joints to sheet metal control arms as well, ZF has developed and patented

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

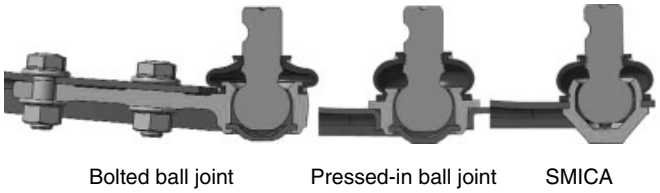


Figure 26. Ball joint types for sheet metal control arms.

the so-called sheet-metal-integrated control arm (SMICA) solution. The unique solution with the robustness of a conventional integrated joint contains a complete ball joint with a steel housing, welded by laser as a cartridge into the arm. The advantages of this technique are a smaller package space to the knuckle, less parts, and less weight compared to all other solutions; this brings the steel solutions a little closer to the weight saving that light metal solutions can offer (Figure 26).

4.2 Ferrous material

4.2.1 Castings

Like all casting technologies, DI casting gives a lot of freedom in design (shape, change of cross sections, and potentially coring). The freedom to distribute material where it is needed in a shape that is most efficient to support loads or the stiffness requirement is the big advantage of the casting process. All this is possible without the need of putting too much material in flash that has to be trimmed, which results in a good material usage. Furthermore, small draft angles can be realized and portions can be cored to reduce the raw material usage.

The typical DI alloy used in suspension components is DBC 450 (Rm = 450 MPA), optionally DBC 550 (Rm = 550 MPA), which both offer a compromise between strength and ductility.

SiboDur® is a material for use in chassis components subject to high levels of stress as typically seen in control arms and steering knuckles. SiboDur has been developed by Georg Fischer Automotive on the basis of spherical graphite cast iron. The name SiboDur is derived from the two materials silicon and boron and from the English word durability. SiboDur makes it possible to produce lightweight cast iron parts.

Tempered ductile iron (ADI) is a heat-treated form of cast DI. It improves the strength of the DI and maintains the ductility to a certain degree. Sizes of the parts and packaging densities in the heat treatment furnace have to be considered to develop a cost-effective part.

Overall, with its flexibility in design, its potential to core certain areas, and considering its mechanical properties, DI offers a high potential to design chassis components such as control arms and knuckles efficiently and cost effectively.

4.2.2 Forgings

4.2.2.1 General. Steel forging had been the most common production process for control arms for decades until the weight factor became more and more demanding. Even today, it still can be efficiently used in axles with specifically tight package areas and high load requirements leading to mostly locally concentrated yield strengths way above 360 MPa or ultimate 400 MPa and up.

The damper and spring as well as the spring tower for the upper arm and the lower arm, especially with separated arms on the outbound ends of the control arms – because of the additional ball joint and the special kinematics with a virtual center for the steering axis, limit the package on double wishbone corners. At the inbound end of a control arm, it is sometimes tight for the bushing bosses in combination with large bushing diameters driven by desired stiffness and allowable space, limited by available space to the subframe. Further examples are on lower L-arms with a spike pin bushing attachment; the transition from the pin to the structure of the arm is a weak point for bending stress. This can be handled more easily with forgings. In general, it is a big benefit if the packaging and the loading are specified correctly with little potential of change during the program: so a decision for a particular material and process selection can be made with high confidence.

Even though – from a functional standpoint – quenched and tempered material is generally more desirable for a control arm design because of its high ductility and more preferable impact behavior, it has been nearly completely eliminated from the control arm market. As the 1990s, microalloy steel has been the only option in use. Inventions on these specific material types were driven in the 1970s already, especially because of their cost efficiency for crankshafts. The cost efficiency lies in the forge–temper–rework process chain: the final raw part reaches its mechanical properties just by controlled air cooling on a conveyor after forging and trimming with least thermal. Additional advantages are a delayed tendency to crack under fatigue life, a far better machinability because of its specific grain structure and a smaller border-to-core hardness decrease compared to quenched and tempered material. However, the disadvantages are a faster crack progress, once a crack is induced, and a worse notched-bar impact works (Wegener, 1998). So, it is possible to meet the demanded combination of best function and least possible cost.

The most popular alloy in Europe is 30MnVS6 +P, in the typical range 560-MPa yield and 800-MPa ultimate

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 27. Two-point link in steel forging technology for BMW (ZF Friedrichshafen AG). (Reproduced with permission from ZF.)

strength. The alloy is also available on the Asian and American market or can be purchased with very close similar chemical properties such as 15V24.

The process and the cost with some major aspects drive the forging design. Two-point arms with straight and easy routing as well as small cross sections with easy geometry are the most desirable ones. The more complex the geometry gets and the wider the projected area in forging direction is, the less likely it is to find a cost-effective design (Figure 27). If steel forging is needed in particular sections of the arm, then large arms will be hybrids out of sheet metal or aluminum-plus-steel combinations.

The material consumption on a steel-forging arm surely depends on the arm's size and complexity, but there is a general disadvantage with this process. The flash extension required to fill the cavity of the mold properly is large compared to that required in casting processes. It is extraordinary on flat arms with a wide extent but significantly lower on straight thin arms. A rate factor larger than 2 for the operational weight in proportion to the final part weight is borderline in terms of cost efficiency. In addition, the larger and more complex the arms are, the bigger is the height and mold shift tolerance impact with its effect on the function of the part, like buckling. Aluminum forgings of a similar size and complexity are easier to control, compared to steel (Figure 28).

4.2.3 Sheet metal (1-shell, 2-shell)

Sheet metal is widely used in vehicle suspensions and on all types of vehicles from passenger cars over SUVs up to pickup trucks (Figure 31). Depending on load case, the stampings can be single shell, double shell, or fabricated and welded together out of multiple different sheet

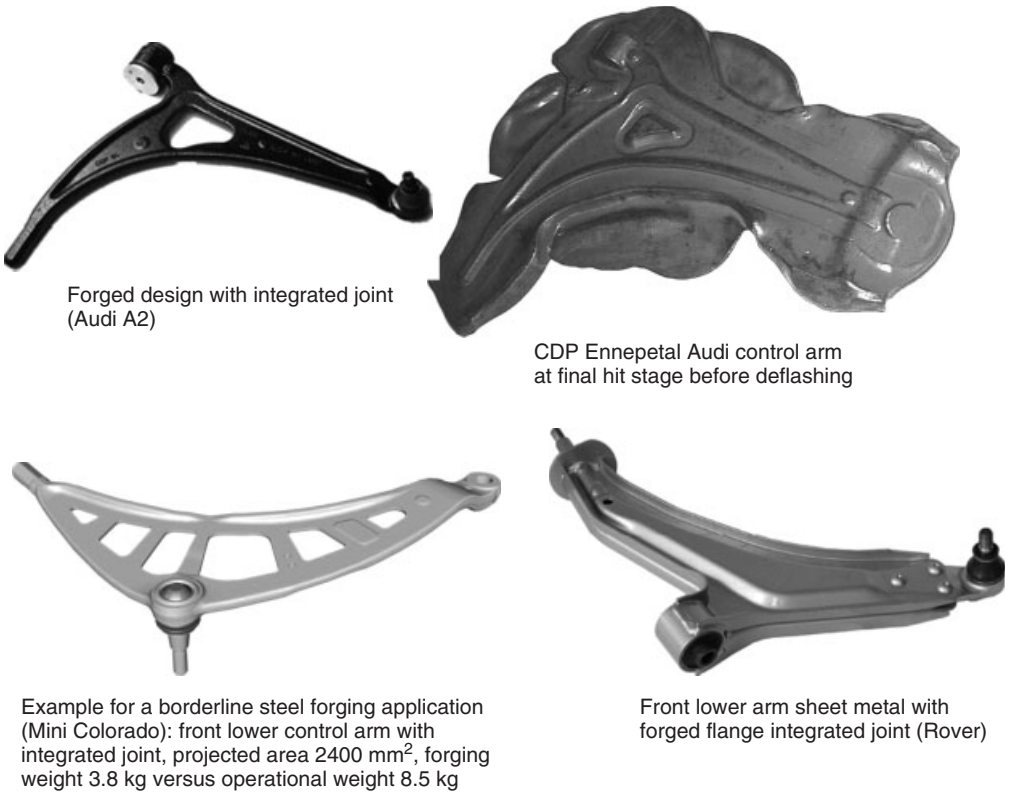
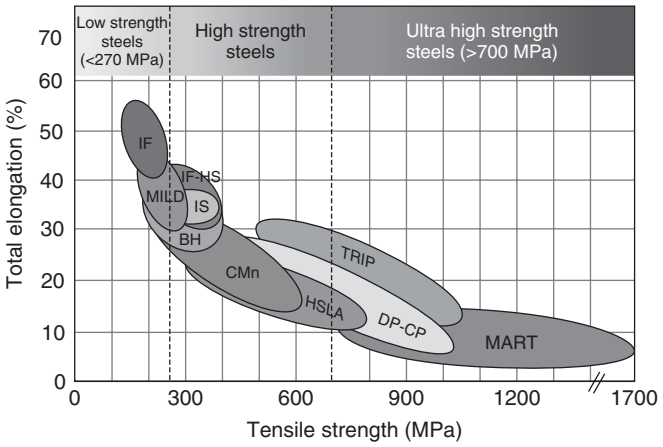


Figure 28. Technology range for suspension arms.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Courtesy - American iron and steel institute

IF	Interstitial free
MILD	low carbon steel, for example, 1008 and 1010
IF-HS	High strength interstitial free
BH	Bake hardenable
CMn	Carbon manganese
HSLA	High strength low alloy
TRIP	Transformation-induced plasticity
DP	Dual phase
CP	Complex phase
MART	Martensitic

Figure 29. Sheet metal technology range. (Adapted from Shaw, 2003. Reproduced by permission of the American Iron and Steel Institute.)

metal parts and reinforcements. The more complex the stamping gets, the more efficient it might be to substitute it by a casting. During the past years, new higher strength sheet metal alloys with good formability have been developed to enable lightweight stampings (Figure 29). This development was feasible because of dual and multiphase steel alloys that provide a soft ferrite grain matrix with a bainitic or martensitic second-phase deposit at the grain borders, which work hardens through the plasticization of the grain (ThyssenKrupp Steel AG, 2008). Although such applications are in production, there are still some huge disadvantages: high steel cost, lack of steel market volume, and missing availability on the global market. In view of increasing global platforms, it appears to be more advisable to accommodate widely used, common alloys of the range up to 600-MPa UTS.

ZF investigated the possibility to produce a single-shell L-arm in a high strength steel grade for vehicles in the C/D class. The aim of this study was to avoid any welds with the exception of welds needed for the ball joint integration. The study showed that single-shell control arms can be utilized up to typical loads for C-segment vehicles; formability of

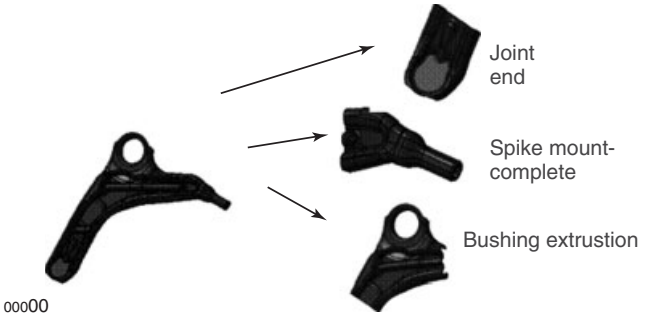


Figure 30. Study of a single-shell L-arm.

the boss hole for the vertical bushing plus the forming of the pin for the rear facing bushing could be realized. Furthermore, the production process could be optimized so that the part can be coated after stamping and then all additional processing can be done in the production line (ball joint integration and bushing assembly). The welding method to integrate the ball joint was developed, so that the ball joint can be integrated at precise position without any negative effects on the ball joint torque and performance and with minimum space requirements to the brake disk and knuckle.

Figure 30 demonstrates the steps of the study: to assure that the single-shell L-arm was feasible, the most critical areas regarding the degree of deformation such as ball joint end, spike mount, and bushing extrusion have been developed as sections. After those successful trials, the whole control arm was formed to demonstrate design feasibility (Figure 31).

4.3 Aluminum

4.3.1 Aluminum casting

Despite intense research and in contrast to sheet metal technology, the alloy used for aluminum suspension parts has not changed over the past 10 years and is out of the 3xx.x series with the elements silicon, copper, and/or magnesium; commonly A 356 is used with the T6 heat treatment (solution heat treated and artificially aged). Even though the casting alloy for control arm applications has not changed, a development process took place on the casting production process side: a variety of performance casting processes has been developed such as squeeze casting, VRC/PRC, Cobapress™, Vacural®, CPC, and others. These performance castings could be used in the chassis application as they provide low porosity parts with good mechanical properties and sufficient elongation values (T6 heat treatment).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

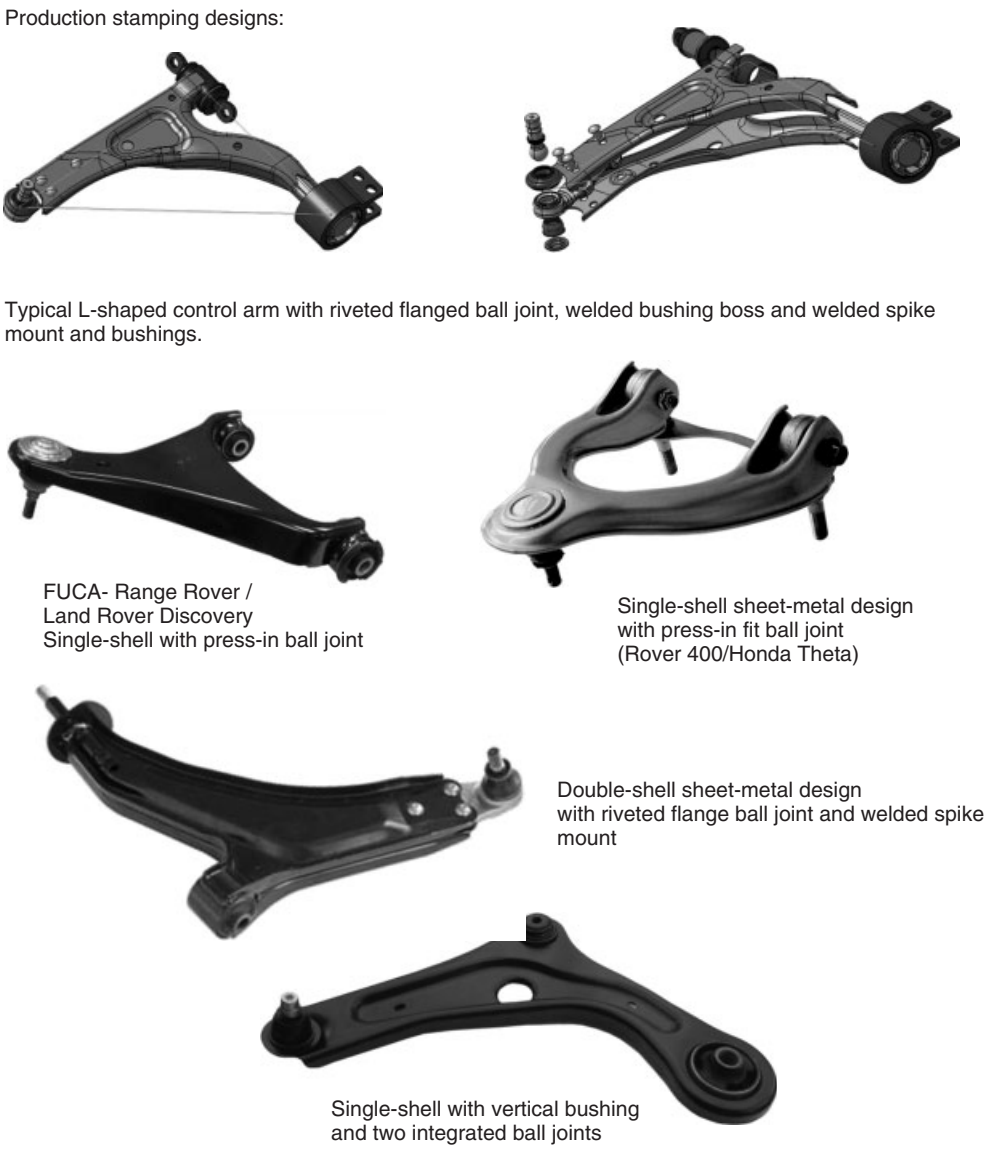


Figure 31. Examples for sheet metal control arms.

Owing to the need to further reduce weight, an increasing focus on the potential for hollow cast is seen; currently, different chassis parts such as control arms and knuckles (Figures 32 and 33) are under development. The usage of cores allows a box section in the part that gives an ideal shape for stiffness (Figure 34); in addition, local reinforcements can be casted into the part. Furthermore, reduced heat transfer to the core may allow thinner walls and may reduce mold complexity as no mold pulls are required; designs may be possible that cannot be realized with pulls. In addition, a core can be designed without a draft angle or even undercut, which would create mold lock in a typical casting process.

4.3.2 Aluminum forging

4.3.2.1 General. Aluminum forgings do not share the fate with steel forging, which is more and more decreasing. As weight has become more important and therefore strength and ductility, aluminum is in favor. The market share of aluminum forgings is still increasing; an aluminum forging is the best functional solution with the best process capability. Despite the fact that aluminum forgings come with relatively high costs, they are still very popular at OEMs especially for front axle applications, as casting processes require expensive quality controls such as X-ray checks.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

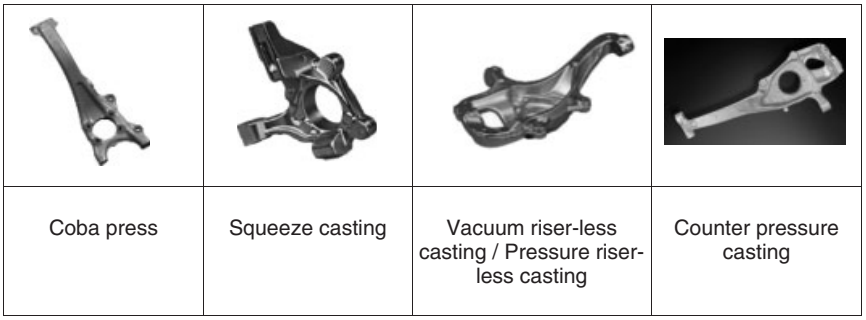


Figure 32. Example Knuckles in Al casting.



Figure 33. Examples: suspension arms in cast technology.



Figure 34. Example: BMW X5 RLCA: VRC/PRC with sand core (hollow cast).

If the design constraints described in the steel chapter allow for softer material, then aluminum forge will be the best choice of all, if the comfort ability of best fit and function can be afforded. What has been said about steel forging is also valid for aluminum: the production process is very stable and reliable; high end forging is for instance controllable by fully automated high speed lines.

Compared to steel, aluminum material provides some additional advantages beside weight reduction, which support the decision to use it. Very obvious is the optical appearance, a dull or even shiny silver like surface with smooth roughness. The functional aspect of the smooth surface is the lack of roughness notches acting as stress raiser on fatigue life. However, the disliked effect with aluminum is that it does not show a constant long life fatigue strength level. Therefore, assumptions need to be made like setting a minimum cycle number of 2×10^6 for long life criteria.

Similar to casting materials, forging material ranges of the alloy SAE J 454–6061 (yield 260 MPa, UTS 300MPA, and elongation 10%) and SAE J 454–6082 (yield 310 MPa, UTS 340MPA, and elongation 10%). Depending on the specific company alloy composition and heat treatment, a level of yield 390 MPa with UTS 410 MPA is possible especially for two-point links.

It is very important for aluminum arms to maintain the rule of providing as little coarse grain as possible.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

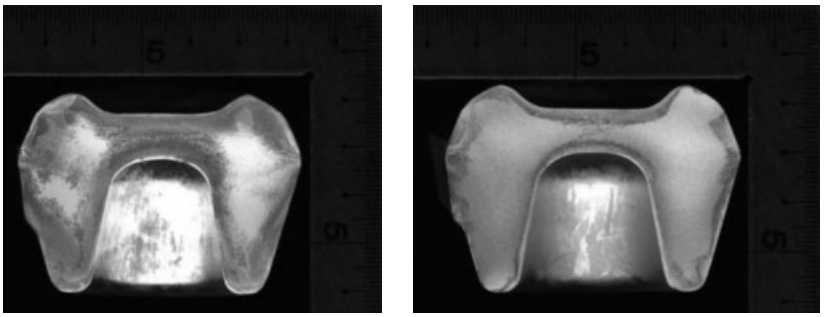


Figure 35. Cross section of forged aluminum suspension arms (grain structure).

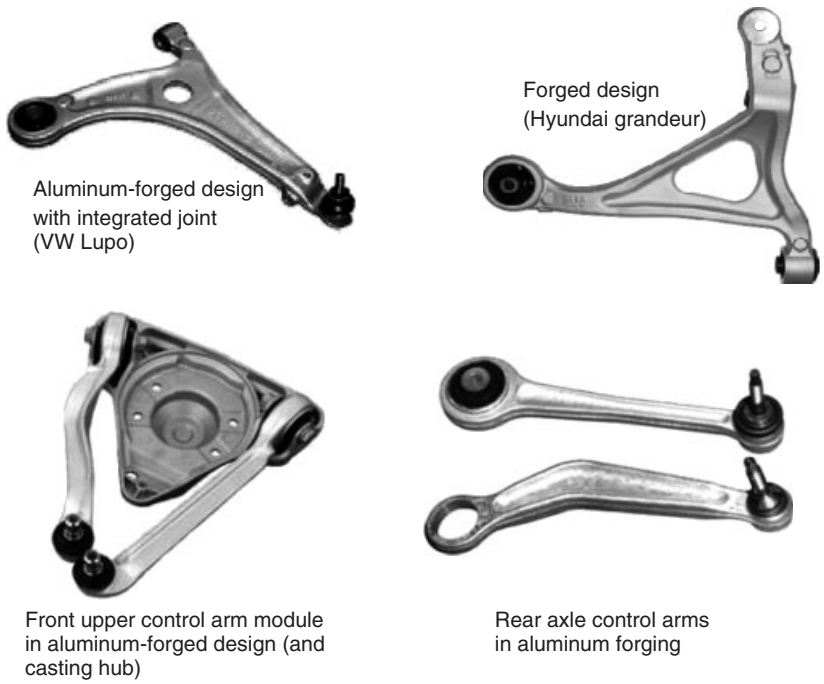


Figure 36. Examples: aluminum forging.

On typical arms, only a thin layer of coarse grain is allowed at the skin to enable the desired deforming and fatigue life behavior as well as the formability after machining – allowing the integration of ball joints (Figure 35).

In addition to the description in the design chapter, the aluminum forging process generally provides the same restrictions as the steel forging process. The advantages versus steel are hidden in details such as smaller draft angles or sharper radii. The loss of material in the flash extension is less (thinner), and, as a rule, the scrap material is completely used for company-internal recycling, as mentioned in chapter 2.1.

5 DESIGN MATRIX (JUDGMENT OF TECHNOLOGIES FOR DIFFERENT SUSPENSION ARM TYPES)

5.1 L-shape front lower control arm for McPherson suspension

L-shape control arms are currently the preferred FLCA (front lower control arm) for McPherson suspensions; they are typically designed to lateral and traversal stiffness, ground clearance, crashworthiness, clearance for tire envelope and drive shaft, stabilizers, and so on – Owing to those requirements with regard to space and performance,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

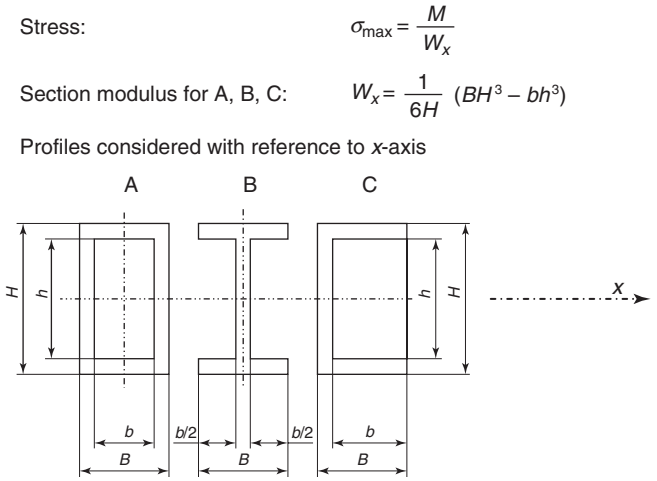


Figure 37. Shape (cross section) of control arms.

not only one material or process is applied. L-arms are rather produced in aluminum and steel forged, stamped, or casted parts.

The following arm was designed as double-shell control arm during early development stages with regard to fore-and-aft stiffness and lateral stiffness. Considering the fore-and-aft stiffness in a simplified way, the design engineer, while keeping the bending stress in mind, has to find the best section modulus and therefore the best cross-sectional shape suitable for the intended manufacturing process.

The different cross sections in Figure 37 have the same sectional modulus (based on reference to the *x*-axis) and therefore the same maximum stress and the same weight. This means that the H profile, typical for casting processes, and the box profile, typical for double-shell stampings, can be designed competitively for the relevant processes.

The following front lower control arm (Figure 38) shows a production design for a high volume application. The numbers below show the ranking – "1" for best in this category down to "4" worst in this category. The design was in competition with a double-shell stamping. Before deciding for this aluminum performance casting, the following other options have been considered by the vehicle manufacturer.

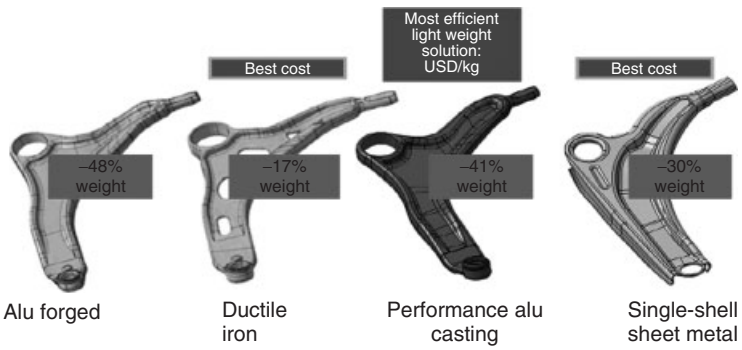
Fulfilling customer requirements for stiffness, packaging, and durability, the above table shows the performance of the different materials. In this case, a performance aluminum casting process (vacuum riser-less casting/pressure riser-less casting) was chosen for series production, as it offered the customer a weight advantage at a reasonable cost.

5.2 Front upper control arms

Front upper control arms for a double wishbone suspension with a long knuckle (goose neck) are currently the typical design, because of space and kinematic advantages over the McPherson strut suspension. The upper arm is typically designed to allow space for the strut module assembly; it provides short length to limit cross car space and for stiffness. Furthermore, the arm has to allow for jounce and rebound travel and tire clearance. The following example is taken from a design competition against a steel forging.

For the new vehicle, requirements have been updated regarding lateral and fore-and-aft stiffness; different designs in stamping, DI, and aluminum have been analyzed including a ranking (see above) (Figure 39).

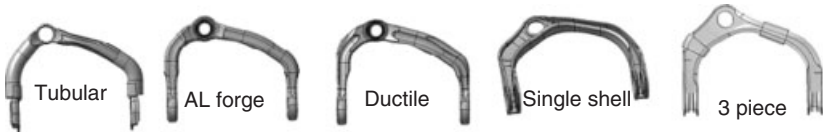
In this case, none of the investigated parts could fulfill the upgraded stiffness requirements – even not the incumbent



Weight	1	4	2	3
Stiffness	2	3	1	4
Cost	4	2	3	1

Figure 38. Technology range for FLCAs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Weight	2	1	3	2	4
Stiffness	2	3	1	4	2
Cost	4	2	1	3	5

Figure 39. Technology range for FUCAs.

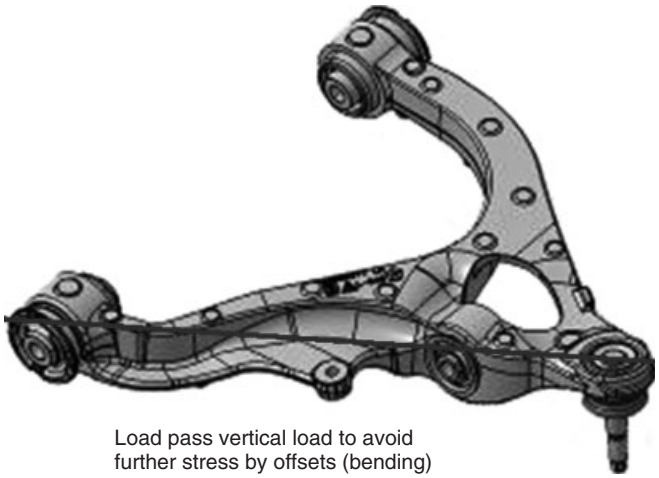


Figure 40. Example: A-shape FLCA.

forged steel design. With reduced stiffness, it appears that a stamping or an aluminum cast could provide the best compromise between cost and desired performance.

5.3 A-shape front lower control arms (FLCA) for double wishbone suspension

These are in the market in all different materials and processes. Besides all the requirements previously discussed for the L-arm, FLCAs of a double wishbone suspension have to additionally carry the vertical force. The following suspension arm (Figure 40) was developed for a pickup truck. The first generation had a multiple piece double-shell stamping design, which was – for cost reasons – replaced by a DI control arm with torsion bar attachment. The next model changed from a torsion spring to a normal strut-mounted coil spring (due to crash performance), and the design was carried out in DI as the previous arm. During the last refresher, weight and fuel economy came into focus – a lighter design had to be developed for nearly identical requirements as to loads, clearance to tires, and drive shaft. As the DI cast part had already been optimized for maximum allowable packaging space and clearance, the aluminum casting could only be designed in the same space. Moreover, material could

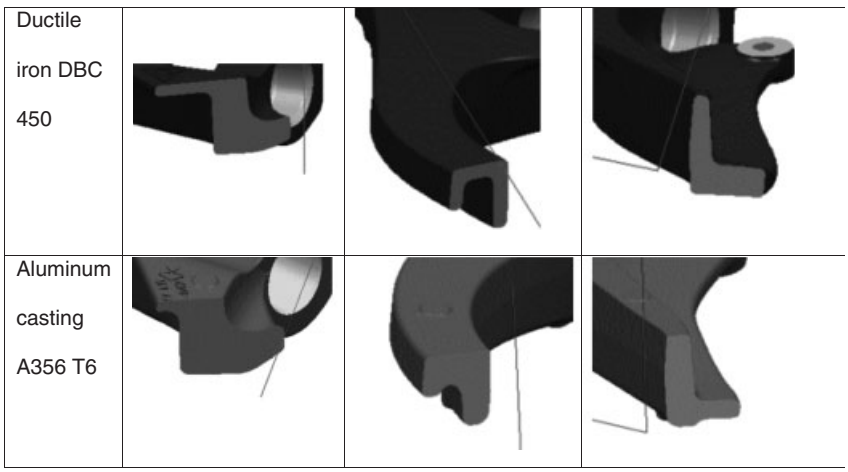


Figure 41. Cross-sectional variation possibility in casting process – ductile iron versus aluminum casting.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

only be added in areas that are not as efficient to create force-carrying capability (Figure 41).

The development history shows that for similar applications different processes and materials have been in production, based on overall vehicle goals, which resulted in aluminum casting with the main focus on weight at reasonable costs. With the replacement of DI – which had been the cost effective solution – by aluminum casting, a weight saving of 5.3-kg per control arm was achieved, resulting in an overall weight saving of ca. 40% for the complete assembly.

6 SUMMARY AND OUTLOOK

Suspension arms, steel versus aluminum, where are the benefits? Coming back to the initial question, this chapter shows the overall tendency that aluminum applications offer the best weight whereas steel offers best cost. Owing to the fact that designs for suspension arms are often driven by stiffness and/or buckling requirements, the individual solution might lead to different results – so in the end, it depends on the application. For other chassis parts, for example, knuckles, the direction is more constant.

Sheet metal solutions already have a big market share and, based on ultrahigh strength steel, there is further weight reduction potential. Key for a good sheet metal design is the design of the weld seem. Or, the other way round, the best material properties do not support a lightweight design if we have to weld too much in critical areas. Owing to cost advantages and the possibility of weld reduction, one-shell solutions (if feasible) seem to be most interesting for sheet metal solutions.

Steel forging has more and more become a niche process for volume products. For heavy and small packages and high loads (heavy vehicles) it will remain an important technology for suspension arms.

DI solutions are often seen as old-fashioned heavy solutions. This chapter has shown some examples where DI is the best option or compromise among stiffness, package, and cost. There are alloys on the market that even offer very good elongation properties to fulfill the buckling/bending requirements of automotive industries. Depending on the market situation, DI can lead to the best-cost solution at a competitive weight.

Aluminum forging seems to offer low weight but high cost. Competitive costs can be reached for designs with a good material utilization of the aluminum preform to avoid as much machining as possible and as much aluminum scrap due to flash as possible. The machining of aluminum is cheaper than the machining of steel, and in some chases, this more or less compensates for the higher material costs for aluminum. There are materials on the market with much higher strength than standard steel.

High performance aluminum cast solutions and combined cast forge solutions are becoming more and more interesting. The market offers an increasing number of special production technologies that support a good stable quality of the material and material properties with high strength. This leads to best-cost lightweight solutions. Depending on the axle concept, we have to pay special attention to the overall crash concept because of limited elongation properties of cast solutions.

Owing to the upcoming endless range of fiber-reinforced plastic applications especially for vehicle bodies, the interest in plastic applications for chassis components has increased, too. There is still some skeptics regarding plastics for chassis application, but it becomes a more open discussion. For some chassis parts, there are already solutions in series production, for example, the leaf spring for GM Corvette or Volvo 940 and the anti-roll bar links for several BMW and GM vehicles. The complex structure of suspension arms with their high load and crash requirements make it much more difficult to substitute traditional suspension arms. So to overcome hurdles and find different axle layouts that better support plastic applications will take some time.

In other words, the question of "Suspension arms, steel versus aluminum, where are the benefits?" will survive some more years.

REFERENCES

Adler, U. (1991) *Robert Bosch, Kraftfahrtechnisches Taschenbuch*, VDI-Verlag, Düsseldorf.

Beitz, W. and Grote, K.-H. (2001) *Dubbel, Taschenbuch für den Maschinenbau, C28, C43*, Springer, Berlin, Germany.

Cheah, L., Heywood, J. and Kirchain, R. (2010) *The Energy Impact of U.S. Passenger Vehicle Fuel Economy Standards*. http://web.mit.edu/sloan-auto-lab/research/beforeh2/files/IEEE-ISSST-cheah.pdf (accessed 08 January 2014).

Dubreuil, A., Bushi, L., Das, T., *et al*. (2010) A comparative life cycle assessment of magnesium front end autoparts. SAE Paper 2010-01-0275. Society of Automotive Engineers: PA, USA.

Elbers, C., Ersoy, M., Hausfeld, G., *et al*. (2004) *Automotive Chassis Technology*, Vmi, Landsberg am Lech, Germany.

Göde, M., Stehlin, M., Rafflenbeul, L., *et al*. (2008) published online by European Conference of Transport Research Institutes (ECTRI).

Goede, M. (2007) *Contribution of light weight car body design to CO_2 reduction (Karosserieleichtbau als Baustein einer CO2-Reduzierungsstrategie)*. Presentation at 16th Aachener Kolloquium Fahrzeug- und Motorentechnik 2007, 9th October, Aachen, Germany.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Heidi B. and Thomas B. (2011) *Annual Report – The Aluminum Association*, http://www.aluminum.org (accessed 08 January 2014).

Heißing, E. (2011) *Chassis Handbook*, Springer Science+Business Media, Berling, Germany.

Shaw, J. (2003) *ULSAB-Advanced Vehicle Concepts ULSAB-AVC*, presentation to the American Iron and Steel Institute. http://www.autosteel.org/~/media/Files/Autosteel/Great%20Designs%20in%20Steel/GDIS%202003/02%20-%20ULSAB%20Advanced%20Vehicle%20Concepts.pdf (accessed 19 February 2003).

ThyssenKrupp Steel AG (2008) Dualphasen-Stähle DP-W® und DP-K®: Für die Herstellung komplexer hochfester Strukurelemente. http://www.thyssenkrupp-steel-europe.com/upload/binarydata_tksteel05d4cms/39/88/78/02/00/00/2788839/Dualphasen_Staehle_de.pdf (accessed 08 January 2014).

Wegener, K.W. (1998) Werkstoffentwicklung für Schmiedeteile im Automobilbau, *ATZ Automobiltechnische Zeitschrift*, **100** (12), 918–927.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto002
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Designing Twist-Beam Axles

Veit Held and Rüdiger Hiemenz

Adam Opel AG, Rüsselsheim, Germany

1 Introduction 1
2 The Functional Principle of a Twist-Beam Axle 1
3 The Process of Designing a Twist-Beam Axle 4
4 Further Advancements of the Axle Concept 9
5 Summary 11
Related Articles 11
References 12
Further Reading 12

1 INTRODUCTION

When the twist-beam axle was introduced in 1974 in the Audi 50, in the VW Golf, and later in the Scirocco, it became a rapid success. Historically, this concept can be seen as a further development of the trailing arm axles that were widely used in front-wheel-driven vehicles. The immediate forerunner was the non-driven rear suspension that was launched in the DKW Junior, 1959–1962 (Figure 1). While this axle connected the trailing arms with a slotted axle tube, which did already work as a stabilizer, the longitudinal control arms were still flexible and a Panhard rod was needed to take the lateral forces.

In the Audi 50, we find all the typical elements of a twist-beam axle that work in the way, as we know it today (Figure 2).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.

This article is © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Today, twist-beam axles or, as they are also referred to, compound crank rear suspensions, have by far the largest volume on the global passenger car market of all rear suspension types (Table 1). It can be expected that, with the growth of the small and compact car segment, the demand for the twist-beam axle type will increase.

Twist-beam rear suspensions dominate the mini to lower medium passenger car segments (Table 2). Typical models in the mini category are the Toyota IQ and Peugeot 107. In the small car segment, we find the Opel Corsa, the Ford Fiesta, and the VW Polo, and in the lower medium category, we find vehicles such as the Honda Civic (EU) and the Renault Megane.

The reasons for this widespread usage have been summarized in numerous textbooks, for example, in Heißing, Metin Ersoy, and Gies (2007). The main advantages and disadvantages of this simple, but very effective axle are listed in Table 3.

2 THE FUNCTIONAL PRINCIPLE OF A TWIST-BEAM AXLE

Apart from the usual design elements that apply to any axle (such as the positions of dampers and springs), the twist-beam axle has the following characteristic components that define the specific properties of this concept (Figure 3):

1. the torsion beam
2. the front bushing or, as it will be called in the remainder of this chapter, the “A-bushing”
3. the left and right trailing arms

This highly integrated design takes all forces and moment which are applied to the tire and wheel during driving. Its mechanical parameters allow tuning both ride and handling

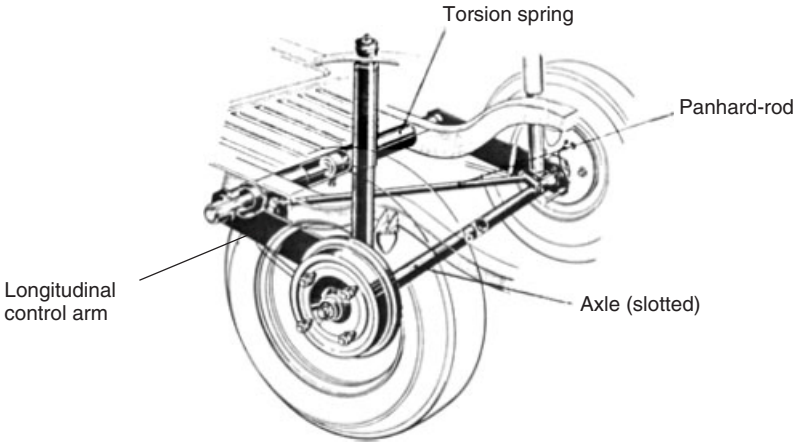


Figure 1. The forerunner of the twist-beam axle was a non-driven rear suspension with elastic longitudinal control arms. (DKW Junior, 1959–1962. Reproduced from Audi company archives, with permission from Audi AG.)

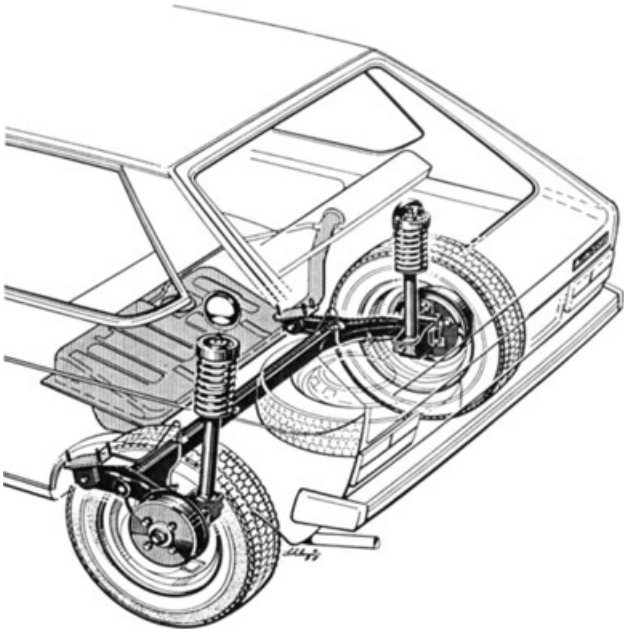


Figure 2. Non-driven rear axle with stiff longitudinal control arms and torsional elastic beam, Audi 50, 1974–1978. (Reproduced from Audi company archives, with permission from Audi AG.)

Table 1. Rear suspensions, worldwide light passenger vehicles around 2007 (%).

	RWD	FWD	4WD	Total
Twist-beam	0.0	31.9	—	31.9
Semi-trailing	0.5	12.1	—	12.6
Solid beam	10.9	10.8	2.1	23.8
Double wishbone	2.2	3.7	0.2	6.1
Multilink	4.0	18.1	2.2	24.3
Strut and others	0.3	0.9	—	1.2
Total	17.9	77.5	4.5	99.9

The automotive suspension systems report (2013), Data Source: ZF Friedrichshafen.

Table 2. Twist-beam penetration by segment, worldwide light passenger vehicles (%).

	Twist-Beam
Mini	2.8
Small	11.9
Lower medium	9.3
Medium	4.5
Upper medium	0.9
Luxury and sport	0.1
Off-road	0.1
MPV	2.2
Transport	0.2
Pickup	<0.1

The automotive suspension systems report (2013), Data Source: ZF Friedrichshafen.

behavior as well as the noise isolation of this suspension. The small number of components requires each of them to serve several functions, unlike the links in a multilink suspension that have dedicated functions and dominating load directions.

The torsion beam (1) is the key element that differentiates this axle from all other concepts. All side forces that act on the wheels try to turn the whole body of the axle around a vertical axis. This creates internal stresses in the structure, particularly in the crossbeam, which has to provide a stiff coupling of the wheels. The higher the bending stiffness of the profile, the better is the control of the pivoting moment by tuning the elasticity of the A-bushings. With respect to the roll response during cornering, the torsion profile has exactly the same role as the stabilizer bar in independent suspensions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Advantages and disadvantages of twist-beam axles.

Advantages of the Twist-Beam	Disadvantages
• Very simple design—essentially, it takes one body plus two rubber bushings • Excellent packaging, that is, needs little space and has a very flat shape (exceptions: housing the dampers and accommodating the vertical movement of the crossbeam during suspension movements) • Easy to assemble • Stabilizer function included • Small unsprung mass • Enables damper ratios around one (damper can be placed vertically at the wheel center) • Roll steering independent of the load • Good compensation of pitching during braking • Small changes of track width under load	• Load peaks at the joints between the rigid trailing arms and the crosslink • Lateral force steer needs toe-correcting bushings • Limited side force stiffness due to deflecting moments in torsion beam and control arms. Compensating measures (e.g., toe-correcting bushings) affect road harshness • Difficult to enable all wheel drives • High axle loads difficult to achieve because of undue stress on welds • Conflicting targets of ride comfort, noise isolation, and handling are difficult to meet because all forces have to be taken by the front bushings

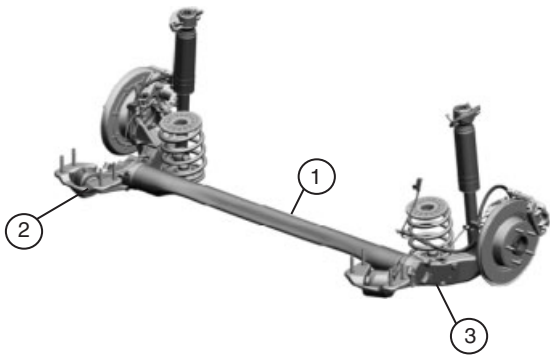


Figure 3. Key elements of the twist-beam axle. (Reproduced by permission of Adam Opel AG, Germany.)

The A-bushing (2) takes all lateral and longitudinal forces. The orientation and the spring rates of the bushing have a huge impact on the steering behavior during lateral loading.

The trailing arms (3) carry the brake force, the side loads during cornering, and the resulting moments, which are mainly the bending moments around the lateral and the vertical axis plus the twisting moment around the roll axis. Therefore, the stiffness of the trailing arms in all directions and around all axes controls the deflection and orientation of the wheel and tire under loads. Stiff welded sheet metal structures or cast profiles are established design solutions.

These three components will be discussed in more detail in the following sections.

2.1 The torsion beam

Understanding the twist-beam axle means understanding the torsion beam. In the case of most other concepts, the movement of the wheel is defined by the geometry of the bushings and the links. In the case of the twist-beam axle, the roll center, the camber angle, and the toe-in are largely defined by the twisting behavior of the beam.

When driving over a wide obstacle such that the left and right wheels move in parallel, the twist-beam axle works exactly like a rigid axle. The situation is more complex under lateral loads or when the car is rolling. If one wheel is raised while the other one is lowered, then the beam gets twisted around its lateral axis, which induces shear forces in the profile. As a consequence of these internal forces, the beam will twist around the shear center of the profile (Figure 4).

If the left and right wheels are moved vertically in opposite directions, then the shear centerline will pivot around its center point in the middle of the car (Figure 5). This means that the intersection of the shear centerline with the middle of the car defines one pivoting point of the left and right half-axles; the second one is the respective A-bushing. In Figure 5, the dashed line marks the axis around which the left wheel turns upward. Apparently, a twist-beam axle behaves like a semi-trailing arm axle with the same A-bushing as the twist-beam and a central bushing at the intersection of the shear centerline with the middle of the vehicle. Figure 5 depicts this relationship in the top view.

The designer can position this "virtual center bushing" by moving the beam backwards or forwards. This "virtual bushing" can also be moved up or down by choosing the profile and the orientation of the beam such that the shear center is located at the desired height. Moving the shear center above the A-bushing will result in a roll understeer behavior, moving it below will result in roll oversteer behavior (Figure 6).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

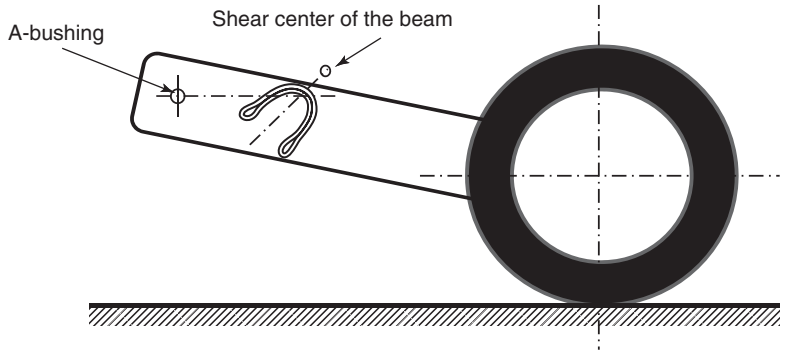


Figure 4. The torsion of the beam is defined by the shear center of the profile.

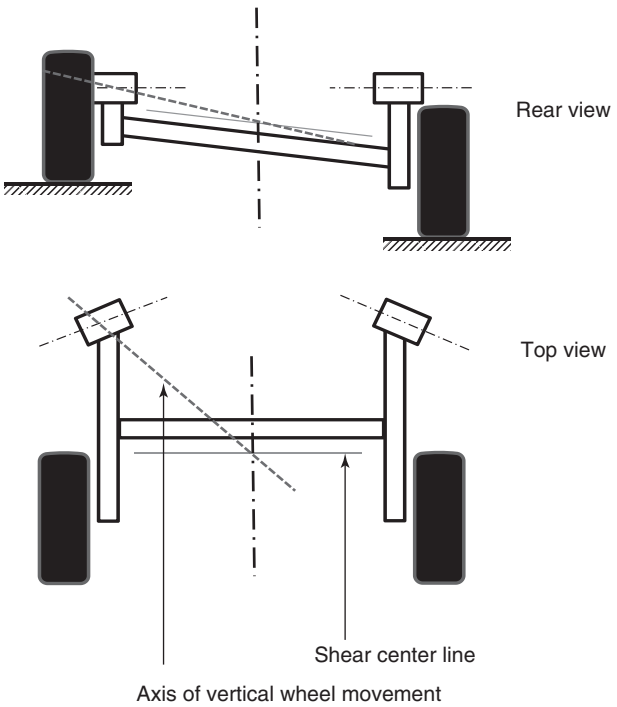


Figure 5. Wheel movements when the vehicle is rolling.

To define both the torsional rate and the position of the shear center, the designer can select the material and the shape of the profile (Table 4).

2.2 The A-Bushing

The A-bushing of the twist-beam has to resist longitudinal and lateral forces and the resulting moments. There are multiple conflicting design goals that make the design of the A-bushing critical. On one hand, the bushing should be soft to absorb the noise and vibrations from the road and to reduce the impacts from obstacles. On the other hand,

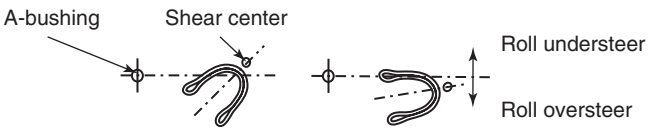


Figure 6. Positioning the shear center will define roll under- and oversteer.

the bushing needs to be stiff to create a precise handling behavior and determine the desired over- or understeer behavior due to side forces or rolling movements. To meet these requirements, the designer needs to define the vertical and horizontal spring rates of the bushing and the orientation.

2.3 The trailing arms

Because all side forces have to be transferred to the A-bushing, these arms have to be very rigid. Most commonly, sheet metal arms are chosen to save cost and mass. Another option is to use cast aluminum, which allows for topology optimization. One of the biggest challenges is to design the link between the trailing arm and the torsion beam, in particular, if different materials have to be combined.

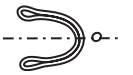
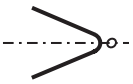
3 THE PROCESS OF DESIGNING A TWIST-BEAM AXLE

In the previous chapter, the main design elements of the twist-beam axle were introduced. In the following chapter, the main steps will be summarized to develop a specific twist-beam axle for a particular carline.

The first step in the design of any rear axle is to do the "concept selection", which means to find the right axle type for a given vehicle. In this phase, all alternative concepts

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Commonly used profiles to define torsional stiffness and shear centerline.

Shape	Design	Property
	Closed section in U- or V-design, made from tube	High torsional/roll stiffness
	Open section in U- or V-design, made of sheet metal	Low torsional/roll stiffness
	Open section in U- or V-design, made of sheet metal	Additional stabilizer bar for increased torsional/roll stiffness

will be evaluated. Apart from the traditional chassis requirements, there are many other boundary conditions to be met. Some of the primary drivers are the packaging constraints, which are imposed by the required trunk space and fuel tank volume, and the ground clearance. All these requirements have to be evaluated against the allowable system cost.

3.1 Essential design parameters and system characteristics

The twist-beam axle has a specific set of design parameters that can be selected or tuned to achieve the desired handling behavior, ride comfort, and noise isolation. These parameters are marked in Figure 7 and the impact on the targeted properties is listed in Table 5.

Table 5. Design parameters of twist-beam axles.

No.	Parameter	Impact
(1)	Position of the crossbeam Torsional stiffness	Roll center height, Stabilizer rate
(2)	Position of the shear center	Roll center height, Roll steer
(3)	Angle of the A-bushing	Lateral force steer
(4)	Compliances in *x*- and *z*-direction, Compliance in *y*-direction	Impact harshness, Lateral force steer, Road noise, Lateral force compliance
(5)	Side view swing arm angle	Wheelbase change and hence impact in harshness Anti-dive angle

3.2 Understeer contribution

The performance of a twist-beam axle depends on how well the conflicting requirements for the overall steering behavior of the axle and the ride and noise refinement can be balanced within the given package constraints. A typical design iteration process starts with the understeer contribution that the rear axle has to deliver in the context of the desired overall vehicle handling trim.

The rear axle understeer contribution is the sum of roll steer, lateral force compliance steer, and vehicle roll rate fore/aft distribution. Each element relates to a design factor with its own package constraints that limits the amount of total steer that can be derived from the single element. Consequently, the design process of a twist-beam

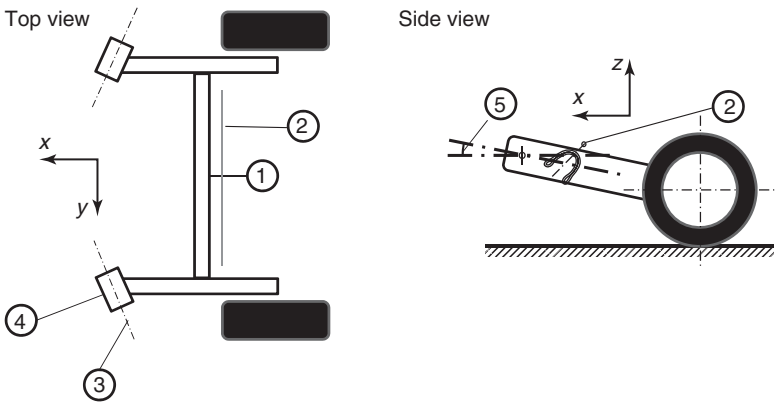


Figure 7. The design parameters of the twist-beam axle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

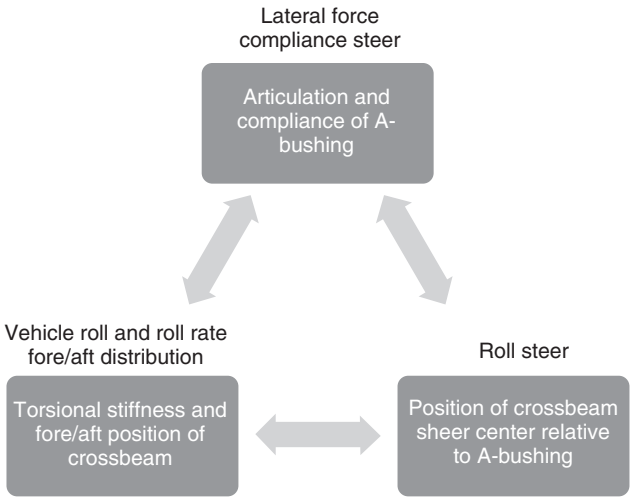


Figure 8. Main elements that contribute to the total axle steer.

axle makes best use from all three factors in combination (Figure 8).

The design evolution starts with the definition of the total vehicle understeer and its distribution to the front axle and rear axle.

3.2.1 Lateral force compliance steer

By its nature, the twist-beam axle goes into toe-out (oversteer) during cornering. Oversteer is caused by the moment that is generated by the side forces acting on the tire contact point longitudinally behind the A-bushings. The moment twists the axle into the direction of the acting side force, that is, oversteer (Figure 9a).

Mainly, three factors control the amount of inherent oversteer. To reduce the amount by which the axle is turned into toe-out by cornering loads, one can shorten the distance between the tire contact point and the A-bushing, choose a wider track width, or increase the fore/aft stiffness of the A-bushing. Of the three main factors, one can assume track width to be a program-given constraint that cannot be used as a design variable to control steer.

The control arm length can be chosen more freely. However, if the arm becomes too short, then the twist-beam is too close to the tire contact point in the fore/aft direction. As a result, the twist-beam is predominantly bent by the loads and not twisted anymore. In this case, the axle behaves more like a rigid axle—with all the known disadvantages. The fore/aft position of the twist-beam is furthermore restrained by the underfloor package where tank, muffler, rails, and foot space leave a small window for the suspension.

The fore/aft stiffness remains the main free parameter that can be tuned to control oversteer. Unfortunately, the fore/aft stiffness not only controls side force oversteer but also affects the noise isolation. Hence, the fore/aft stiffness is a compromise to meet both requirements. The acceptable fore/aft rates alone do not allow to provide sufficient oversteer control.

For this reason, a kinematic trick has become standard in modern twist-beam axles. The pivot line of the A-bushing is not perpendicular to the longitudinal axis of the car, but is angled towards the front.

If the bushing has sufficient axial compliance, then the cornering force pushes the axles sideward up the ramp on the side with the higher side force and down the ramp

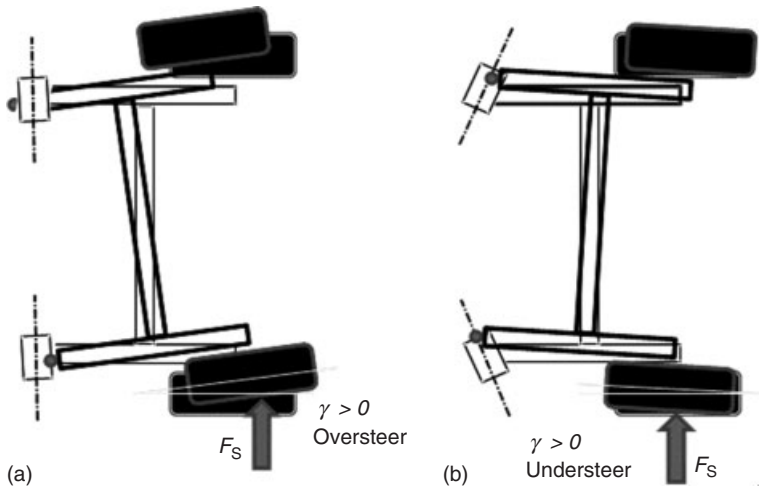


Figure 9. (a) Typical side force oversteer effect caused by the rotation of the axle within the fore/aft constraint of the bushing. (b) Partial compensation by twisting the A-bushing such that the axle slides along the ramping angle, thus inducing a turn.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

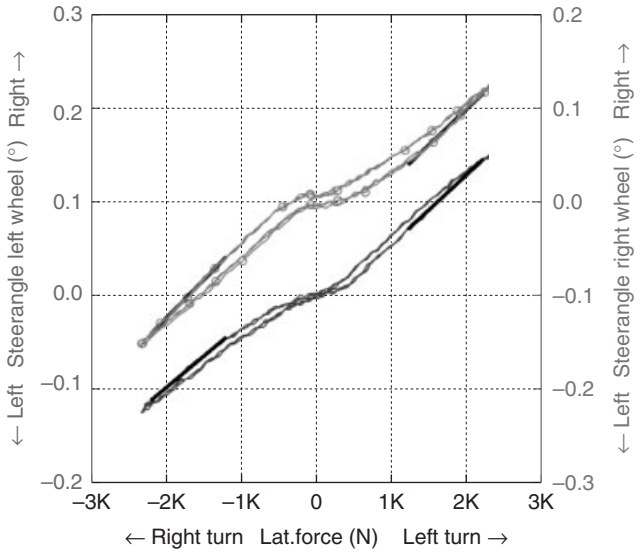


Figure 10. Typical side force steer curves.

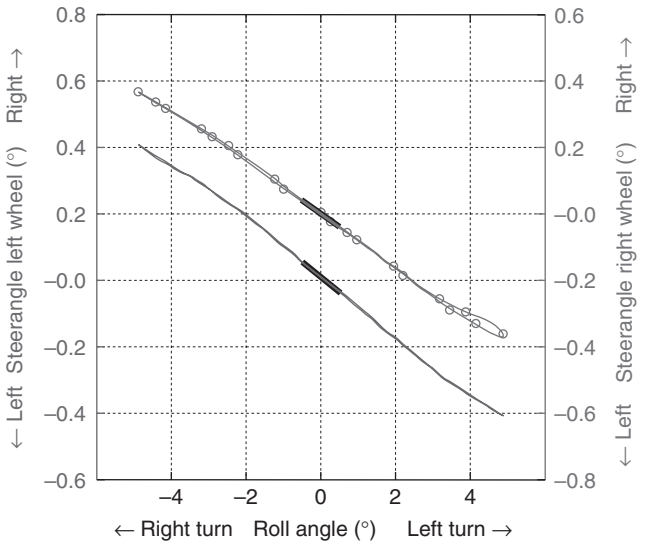


Figure 11. Typical roll steer curve.

on the other side. Effectively, the whole axle is turned into understeer (Figure 9b). The exploration of this effect is limited by the need for directional control of the rear axle that is compromised by the lateral compliance in the bushings. Figure 10 shows a typical curve for the steering angle of the rear axle as a function of the lateral force.

3.2.2 Roll steer

As shown in the last chapter, a particular characteristic of the twist-beam axle is the dependency of roll steer from the position of the shear center of the torsion profile relative to the pivot line of the axle. To further compensate for the remaining compliance oversteer tendency, most passenger cars require a roll understeer contribution from the rear axle (Figure 11).

Maximum roll understeer is gained if the open sections of U- or V-shaped profiles point downwards such that the shear center is above the pivot line of the axle. The optimal vertical position of the A-bushing would be low for understeer; however, for impact smoothness a high position is desirable. The amount of understeer to be gained is limited mainly by packaging restrictions and conflicting performance requirements.

The more the A-bushing is raised, the higher the torsion profile must be situated to provide understeer. Naturally, the underfloor height and the exhaust routing limit the vertical packaging space for the torsion beam. Hence, roll steer will always be a compromise with the influencing factor "harshness".

Twist-beam axles with a tubular interface to the trailing arm offer a marvelous opportunity. With identical crossbeams and trailing arms, the location of the shear centerline can be positioned simply by turning the beam and welding the joint at different angles to the crossbeam (Figure 12). This feature allows tuning the roll steer of an axle to the mass and the mass distribution of the vehicle.

3.2.3 Vehicle roll and roll rate fore/aft distribution

The roll rate of a twist-beam rear axle is defined by the crossbeam torsion stiffness, possibly by the twist rate of an additional stabilizer bar and by the suspension spring rate. The roll rate distribution front to rear together with the axle load distribution front to rear affects the overall steering behavior of the vehicle. The effective roll rate of the crossbeam is a product of its diameter and wall thickness and the fore/aft position of the beam.

By selecting an appropriate effective torsion stiffness as a function of the axle load distribution, the overall steer behavior of the vehicle can be tuned to meet the specification. Furthermore, the steering behavior can be maintained constant for all model variants despite different axle loadings.

The concept of the crossbeam connecting both rear wheels has earned the twist-beam axle the label "semi-independent". In respect to its "copying behavior" of excitations from one side to the other, this notion is misleading. The copying effect is no different to that of an independent suspension that utilizes a stabilizer bar to create roll stiffness. This stabilizer bar does exactly the same thing

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

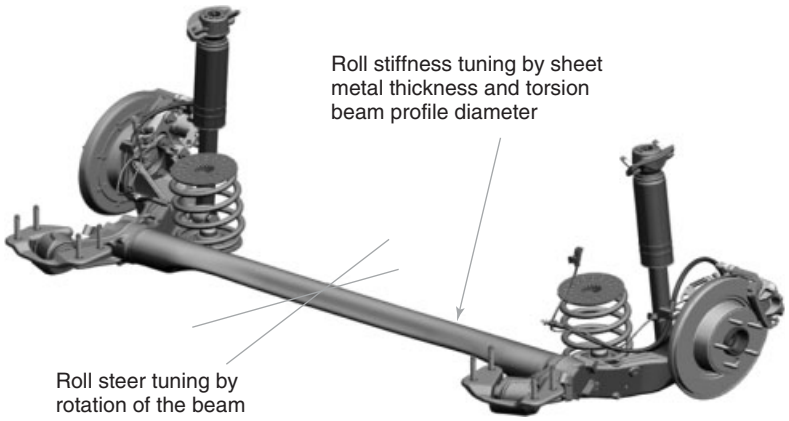


Figure 12. Tuning flexibility of a twist-beam axle with tubular crossbeam.



Figure 13. Main elements that contribute to the impact compliance.

as does the crossbeam of a twist-beam axle: it transfers loads and excitations from one side to the other regardless whether intended (to control body roll) or unintended (if caused by road irregularities). A certain fore/aft transfer of road excitations, however, is inherent to the twist-beam, yet far less apparent than the mutual vertical interference that is common to suspension with a high roll stiffness.

3.3 Impact compliance

The main factors that influence the ability of a wheel to elude road irregularities have already been mentioned in connection with understeer control (Figure 13).

3.3.1 Brake force compliance

Forces that are induced by braking have the same effect as any other forces that are generated from an uneven road. In both cases, the tire/wheel moves backwards against the effective compliance of the A-bushing. High fore/aft compliance of the A-bushing allows the wheel to swing back rather freely while contacting a road input such as a crack or a bump.

Although high fore/aft compliance of the A-bushing is desirable for impact isolation, it is undesirable for side force oversteer as shown in the previous section.

3.3.2 Side view swing arm angle (SVSAA)

Side view swing arm angle (SVSAA) is the angle between the horizontal and the line through the wheel center and the A-bushing. The influence of the SVSAA can be easily envisaged by picturing the overcoming of a step with a wheelbarrow. Pushing the wheelbarrow forward into the step results in a high impact, whereas pulling the wheelbarrow instead, while holding the handles high, results in a smaller impact and a more smooth negotiation of the step. This is because the wheel is kinematically guided away from the step, instead of being pushed into it.

Likewise, the A-bushing should be above the wheel center (positive SVSAA) for minimal impact sensitivity. The optimal solution regarding impact compliance, however, is very often not feasible due to interference with the body rails or the foot space in front of the rear seats. In this case, minimal negative SVSAA should be considered.

Since fore/aft compliance is limited by side force oversteer, and SVSAA is limited by packaging constraints, the achievable impact comfort is a compromise. This drawback is inherent to conventional twist-beam axle concepts.

3.4 Roll camber/camber stiffness

Common multilink suspensions have the property that the upper point of the wheel moves inboard when the wheel gets pushed upwards. This implies that the camber angle deteriorates when the vehicle rolls in a corner and the tire can take less side forces. This effect is almost negligible in the case of twist-beam axles.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

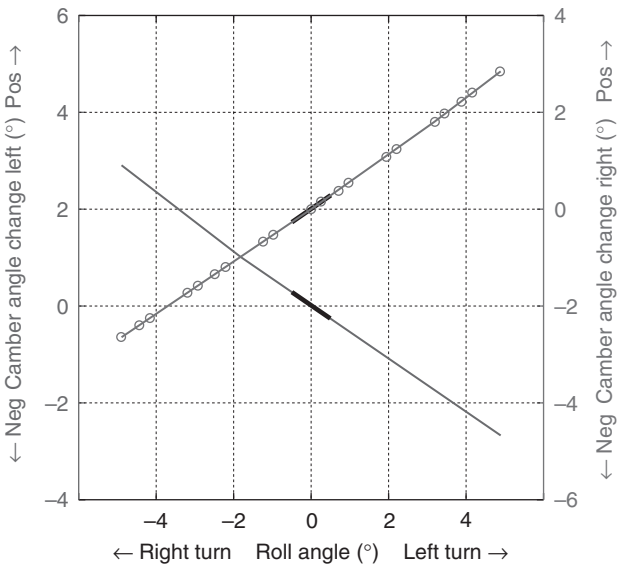


Figure 14. Typical roll camber curve.

Kinematic camber loss is minimal due to the nature of the connection of the axle to the body by the A-bushing that constrains the axle in roll and heave like a hinge with one rotational degree of freedom. Figure 14 shows a typical curve of the camber angle as a function of the roll angle.

Compared to multilink suspensions that suffer from higher kinematic camber loss, the twist-beam axle has significantly better camber compensation.

The second relevant aspect of camber is side-force-induced camber loss. The side force during cornering creates a moment that tilts the wheel in the direction of positive camber. The amount of this effect is predominantly controlled by the torsional stiffness of the trailing arm. Modern cast iron or box-type sheet metal designs provide excellent torsional resistance; hence, twist-beam axles have minimal side-force-induced camber loss.

Sufficient camber compensation and high camber stiffness lead to high residual camber during cornering. This helps grip and rear axle stability.

3.5 Summary: typical design parameters of a twist-beam axle

The said mutual interdependencies and restrictions of the design parameter of twist-beam axle as well as the packaging constraints define a solution space. Most twist-beam axles in the compact and small car segment fall into the bandwidth shown in Table 6.

4 FURTHER ADVANCEMENTS OF THE AXLE CONCEPT

Lately, the usage of the twist-beam axle has been extended to applications that go beyond the conventional performance expectations or use cases of this suspension. New materials and further enhancements to its mechanics show that the limits of the twist-beam axle have not yet been fully explored.

4.1 Twist-Beam with Watts linkage

It has been shown that the twist-beam axle has big advantages to other axle types with respect to camber stiffness, camber compensation, and jounce hysteresis. Lateral force steer and fore/aft compliance, however, are known weaknesses. The A-bushing cannot control the latter three

Table 6. Typical performance attributes of existing twist-beam axles.

Performance Attribute	Phys. unit	From	To
Swing arm length	mm	375	460
Swing arm angle	deg	−6	+10
Ride rate	N/mm	17	35
Suspension rate	N/mm	18	40
Ride steer	deg/m	0.0	10.0
Ride camber	deg/m	0.6	4.6
Total roll rate with tire	Nm/deg	350	1150
Roll steer with tire	%	4.0	13.0
Roll camber with tire	deg/deg	−0.7	−0.4
Roll center height	mm	100	250
Lateral force steer	deg/kN	−0.15	0.03
Lateral force camber	deg/kN	−0.45	−0.16
Lateral compliance at wheel center	mm/kN	0.3	3.0
Brake force steer	deg/kN	−0.0	−2.0
Fore/aft compliance	mm/kN	0.07	2.50

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

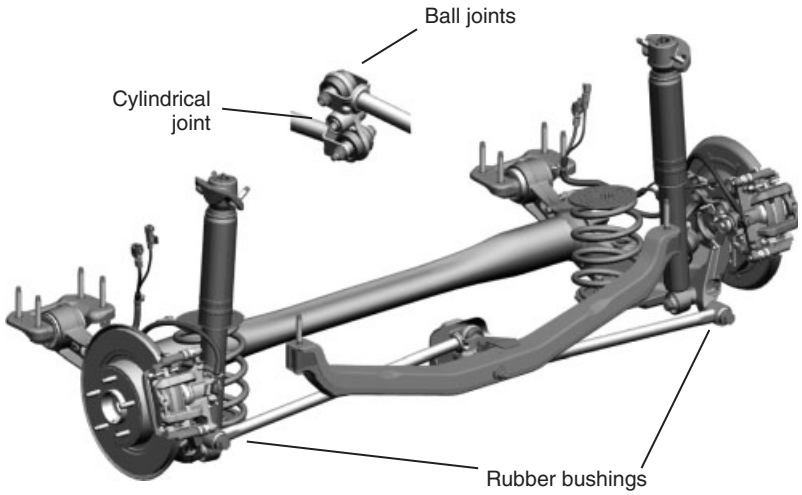


Figure 15. Rear view of a twist-beam with Watt linkage.

characteristics in an optimal way. The compromise lies in moderate, hence safe side force oversteer for the price of relatively harsh impact response and a certain lateral softness.

In 2009, Opel presented the Astra J with a twist-beam axle that was upgraded by two lateral links with a third "differential" link (Harder and Ohligschläger, 2010; Hiemenz and Harder, 2010). This mechanism is called *Watt linkage*. The Watt linkage serves the purpose to maintain the inherent advantages of the twist-beam axle while effectively solving the contradicting requirements for lateral force and steer and fore/aft compliance.

Watt link suspensions for normal front-wheel drive cars have been used before, particularly, in racing together with live axles or with pure torsion beam axles (torsion beam positioned between the wheels). Compared to the historical designs, the attachment for the Watt linkage in the environment of a twist-beam axle has to be reversed. The differential link is attached to the body through an additional subframe and the two lateral links are attached to the axle (Figure 15).

The Watt linkage essentially allows separating the longitudinal support from the lateral support. The lateral forces are reacted by the Watt linkage with minimal lateral displacement. However, if the complete side forces are being compensated by the Watt linkage, then there is still a remaining moment around the vertical axle that will turn the axle into oversteer while cornering. Lateral compliance and lateral force steer are negligible with the Watt linkage. Since the A-bushing is relieved from the support function against side loads and resulting twisting moments, it can be made very compliant in fore/aft direction. This results in an excellent impact behavior and low road noise.

Table 7. Comparison of twist-beam with and without Watt linkage.

	Compound Crank	CC w Watt Link
		
Lateral force steer	0	+
Lateral force deflection	0	++
Camber stiffness	0	0
Camber compensation	0	0
For/aft compliance	0	+
Side view swing arm angle	0	0
Jounce hysteresis	0	—
Tuning flexibility	0	0
Tolerance sensitivity	0	0
Architecture integration	0	0
Weight	0	—
Development risk	0	—

Table 7 summarizes the advantages of the twist-beam axle with Watt linkage in comparison to the conventional twist-beam.

4.2 Special solutions for four-wheel drive

Fiat Sedici, Suzuki SX4, and Opel Mokka are examples of the increasingly popular "sub-compact sports utility vehicles" that are commonly available with both front-wheel

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 8. Typical performance attributes of existing bended twist-beam axles.

Performance Attribute	Phys. unit	From	To
Swing arm length	mm	400	500
Swing arm angle	deg	−7	0
Suspension wheel rate	N/mm	22	30
Suspension ride rate	N/mm	20	24
Ride steer	deg/100 mm	0.00	0.20
Ride camber	deg/100 mm	0.30	0.40
Total roll rate with tire	Nm/deg	600	800
Roll steer with tire	%	13	16
Roll camber with tire	%	−50	−60
Roll center height	mm	200	250
Lateral force steer	deg/kN	0.00	−0.10
Lateral force camber	deg/kN	−0.30	−0.40
Lateral compliance at wheel center	mm/kN	1.00	1.50
Brake force steer	deg/kN	−0.10	−0.20
Fore/aft compliance	mm/kN	1.00	1.50



Figure 16. Four-wheel-drive-capable twist-beam axle. (Reproduced by permission of Adam Opel AG, Germany.)

drive and four-wheel drive. Vehicle size, performance specifications, and cost targets suggest the usage of a twist-beam axle in this segment. Packaging, however, is a severe issue as the crossbeam of the twist-beam normally blocks the space needed for the rear drive module and the propeller shaft.

A bended twist-beam has been found to be a performance-compliant and cost-effective solution that allows the application of a twist-beam axle with both front-wheel drive and four-wheel drive (Figure 16).

In this case, the twisted beam is formed like an arc over the rear drive module. The design does not dictate the static suspension metrics as shown in Table 8. It must be noted, however, that roll steer tuning by rotation of the beam around the *y*-axis is limited to a few degrees because of the interference of the beam with the driveline.

5 SUMMARY

Starting with the historical designs, the development of the twist-beam axle and its main design parameters that control roll stiffness and roll steer are explained. Chassis designers can take this chapter as a guideline for their own work. Several data samples and additional design hints are offered to help designing such an axle.

The latest developments of twist-beam axles show that some of the intrinsic limitations could be overcome and a modern twist-beam axle is able to have a defined steering behavior and is even able to be used as a driven axle. This will enable the chassis designer to use this lightweight and inexpensive axle for future applications.

RELATED ARTICLES

Possibilities of Coil Springs and Fibre Reinforced Suspension Parts
Suspension Arms, Steel Versus Aluminium, where are the Benefits?
ULSAS – Suspensions, a comparison of rear suspension design
The harshness of air springs in passenger cars
Customer oriented Evaluation of Vehicle Handling Characteristics

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

REFERENCES

Harder, M. and Ohligschläger, S. (2010) The New Opel Astra Rear Axle. *Proceedings of the Chassis.Tech Plus Conference*, P 281ff, München, June 8–9.

Heißing, B., Metin Ersoy, M., and Gies, S. (2007) *Fahrwerkhandbuch*, Friedrich Vieweg & Sohn Verlag, Braunschweig.

Hiemenz, R. and Harder, M. (2010) Evolution der Opel Verbundlenkerachse für den neuen Astra. Tag des Fahrwerks, Aachen, October 4.

Hill, A. *et al*. (2013) The Automotive Suspension Systems Report. Stamford, Supplier Business, IHS Global Limited.

FURTHER READING

Arkenbosch, M., Mom, G.P., and Neuwied, J. (1992) *Das Auto und sein Fahrwerk*, Band 1, Motorbuch, Stuttgart.

Bleck, U.N. (2004) Fahrzeugeigenschaften, Fahrdynamik und Fahrkomfort in *ATZ/AMZ* Sonderausgabe März 2004, Vieweg, Wiesbaden, pp. 76–78.

Bostow, D., Howard, G., and Whitehead, J.P. (2004) Car Suspension and Handling. SAE International, Warrendale: SAE 2004.

Braess, S. (2001) *Handbuch Kraftfahrzeugtechnik*, Vieweg, Wiesbaden.

Heißing, B. (2002) Grundlagen der Fahrdynamik. Seminar, Haus der Technik, Berlin.

Piepereit (2003) Fahrwerk und Fahrsicherheit. Vorlesungsumdruck, FH Osnabrück.

Preukschaeid, A. (1988) *Fahrwerktechnik: Antriebsarten*, Vogel, Würzburg.

Wallentowitz, H. (1998) Quer- und Vertikaldynamik von Fahrzeugen. Vorlesungsumdruck Kraftfahrzeuge 1, IKA Aachen, FKA-Verlag.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto003
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Air Suspension Systems—What Advanced Applications May Be Possible?

Friedrich Wolf-Monheim
Ford Motor Company, Aachen, Germany

1 Introduction 1
2 State of the Art—Four-Corner Air Suspension Systems 3
3 Continuously Variable Air Suspension Systems 4
4 Air Spring Damper Systems 5
5 Interlinked Air Suspension Systems 6
6 Active Air Suspension Systems 10
7 Summary 11
Endnote 12
References 12

1 INTRODUCTION

Air suspension systems offer various advantages compared to conventional suspension systems, featuring usual coil or leaf spring elements, leading toward improvements in terms of the ride comfort, the vehicle dynamics performance, and the driving safety. The advantages of air suspension systems in general relate mainly to the practicability of a vehicle body leveling functionality, the technical feasibility of a constant eigenfrequency independent of the loading condition of a vehicle, and the flexibility in terms of the tuning potential of the force versus travel characteristics of the air spring elements by shaping the rolling pistons.

Using air suspension systems, the distance between the vehicle body and the road surface can be varied within the limitations given by the available wheel travels with the aid of the defined supply and exhaustion of pressurized air to and from the air spring elements. On the one hand, the vehicle body can be raised compared to the design height to generate a greater ground clearance between the vehicle underbody and the road surface under severe off-road driving conditions. On the other hand, the vehicle body can be lowered compared to the design height to allow for an unimpeded ingress and egress for all vehicle passengers when the vehicle is standing still. Similarly, the lowering of the vehicle body can enable the reduction of the aerodynamic driving resistance at higher vehicle speeds, for example during highway driving. At the same time, the body lowering improves the road holding abilities of the vehicle, the vehicle handling stability and finally the brake stability of the vehicle. A well-known possibility to reduce fuel consumption and to maintain stability at high speeds is changing the vehicle body pitch angle (lowering the vehicle body height at the front axle only). This functionality is currently used in the Bentley Continental GT.

In the same way, the vehicle body height can be maintained constant independent of the loading condition of the vehicle by varying the air pressures in the individual air spring elements. As a result, also the body eigenfrequency remains unchanged as the air spring stiffnesses can be adapted to the loading conditions. This effect improves the ride comfort for the vehicle passengers.

Furthermore, the geometric design of the rolling pistons can be used to adapt the force versus travel characteristics of the air spring elements to the specific tuning needs over a wide range. Compared to conventional coil or leaf springs, highly nonlinear curve characteristics can be accommodated.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Basically, all air suspension systems consist of air spring modules as primary suspension force elements, a pneumatic compressor in order to ensure the supply of pressurized air, an electronic control unit to handle the control tasks, one or more sensors to detect the ride height and the tilting of the vehicle body as well as pneumatic and electrical connection lines. Some air suspension systems also include a compressed air reservoir in addition to the pneumatic compressor in order to supply the air spring modules with high air mass flow rates if required.

In contrast to the commercial vehicle sector, almost exclusively rolling lobe air spring modules are used for passenger cars. With regard to the ride comfort of air suspended vehicles, the initial dynamic response of the individual air spring modules is particularly important (Wallentowitz, 2005). The contribution of the air spring module stiffnesses with decreasing excitation amplitudes to the total wheel suspension vertical stiffnesses plays a major role in this context. Increasing the wheel suspension vertical stiffnesses results in higher vertical body accelerations with a negative impact on the ride comfort for the vehicle passengers. In the field of automotive engineering, the increase of the wheel suspension vertical rates with decreasing excitation amplitudes is usually referred to as the *harshness effect* (Puff, 2009). Figure 1 illustrates the dynamic air bellow[1] stiffness differences between comfort bellows (one axial fiber layer) and conventional bellows (two crossed fiber layers) with respect to the excitation amplitude.

The frequency dependency starts already slightly above 5 Hz. Thus, the passengers can feel the stiffening of the air bellows. This reduces the driving comfort. Several car companies therefore decided to use comfort bellows and designed air spring systems with externally guided bellows. These differences are explained in more detail in the following: essentially rolling lobe air spring modules can be divided into those that are externally guided and those that are not. From a technical point of view, air spring modules without external guiding can be best executed with cross-ply lobes. Cross-ply lobes normally consist of two or more layers of rubber-coated rayon or nylon cord laid in a cross-ply manner in an angular arrangement with respect to the direction of motion of the air spring unit with an outside layer of abrasion resistant rubber and sometimes an additional internal layer of impermeable rubber to minimize the loss of air (Heisler, 2002). The cross-ply design enables the stability of the geometrical outer shape of the air spring lobe, even under high internal air pressure, without an external guiding tube, as it allows the lobe to carry not only the axial acting force components but also the radial acting force components of the internal air spring pressure. Naturally, cross-ply lobes show greater wall thicknesses compared to axial lobes with external outer guiding tubes. This is the reason for the higher bellow stiffness in Figure 1. In addition, the hysteresis of these bellows during jounce and rebound motions of the suspension is larger compared to the hysteresis of comfort bellows. It is well known that the more friction and hysteresis there is in a suspension system, the less the system deflects for small road irregularities, giving rise to the vehicle riding on its tires rather than on its suspension; this leads to a "busier" secondary ride and to what the Americans call "boulevard jerk". The behavior of former diagonal tires compared to the behavior of radial tires of today is in some way comparable to the behavior of air bellows, which is discussed in this chapter.

The lobes of air spring modules with external outer guiding tubes can be designed with lower wall thicknesses compared to cross-ply lobes without external outer guiding tubes, as the external outer guiding tube carries the circumferential forces. Usually, axial lobes are made from one

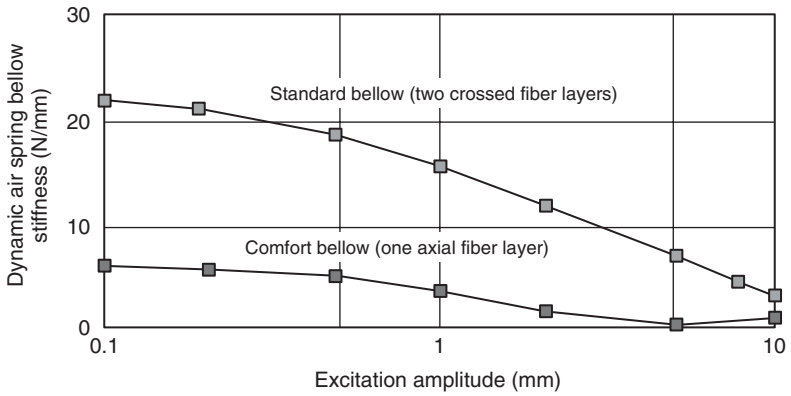


Figure 1. Differences in the dynamic stiffness of one-layer axial air spring bellows and two-layer cross-ply air spring bellows, indicating harshness differences.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

layer of rubber-coated rayon or nylon cord in a parallel arrangement with respect to the direction of motion of the air spring unit with an outside layer of abrasion resistant rubber and sometimes an additional internal layer of impermeable rubber to minimize the loss of air (Heisler, 2002). Owing to the lower wall thicknesses of air spring lobes with external guiding tubes, the harshness effect is usually significantly lower compared to air spring lobes without external guiding.

Generally, air spring modules with outer guiding are more limited in terms of the allowable cardanic movements between the upper and lower ends within a given suspension architecture compared to air spring units without outer guiding. This disadvantage can be compensated by alternative guiding concepts of the air spring lobe, for example, the use of pivoted rolling pistons. Next to the design of air spring modules, also the history of the modules (aging) influences the harshness characteristics. In terms of the air spring module design not only the external guiding elements of the air spring lobes but also the upper and lower connecting joints (ball joints, bearings, bushings, etc.) as well as the dimensions of the modules themselves influence the harshness performance. In the area of air spring lobe design, the influencing factors are the fiber orientation of the different layers, the fiber material, the fiber diameter, the elastomeric material and the air spring lobe wall thickness. These are additional contributors to friction and hysteresis. With regard to the aging of air spring lobes, various environmental influencing factors such as ozone in the ambient air, the combination of ultraviolet light and oxygen, reactive metals such as copper, oils and fats play a key role in terms of the harshness effect. In addition, the air spring module loading history (temperature, force and pressure) has an effect on the harshness characteristics (Puff, 2009).

2 STATE OF THE ART—FOUR-CORNER AIR SUSPENSION SYSTEMS

Vehicles equipped with a four-corner air suspension system are characterized by the fact, that all suspension springs are air spring modules. Four-corner air suspension systems can be divided into two distinct classes: Those without switchable additional volumes and those with switchable additional volumes. Already in the 1960s, vehicles as for example the Borgward P100 and the Mercedes 600 were equipped with four-corner air suspension systems. Even today, four-corner air suspension systems are mainly offered in the premium passenger car segment of the upper and middle classes because of the high cost of such systems. The air spring modules of today's state-of-the-art four-corner air suspension systems are usually equipped with air spring bellows with small wall thicknesses and outer guiding tubes to optimize the ride comfort for the vehicle occupants. Generally, the systems incorporate comprehensive functionality with regard to automatic leveling. This functionality comprises for example the speed-dependent lowering of the vehicle body to reduce the aerodynamic driving resistance or to increase the dynamic driving stability and the hoisting of an SUV body to improve the off-road capabilities. The functions are available to the driver either fully automatic or on demand of the driver.

Figure 2 shows the Jaguar XJ front suspension with the air spring modules of the four-corner air suspension system integrated into a double-wishbone-type front suspension as an example of application. Driven by the front suspension architecture of the Jaguar XJ, air spring modules with outer guiding tubes can be used because no significant bending moments need to be carried by the air spring and damper assemblies. For air sprung strut type front suspension architectures, the bending moments on the damper are usually compensated with the aid of asymmetric shaped air spring bellows without outer guiding elements. Alternatively, it is possible to use rotationally symmetrical air spring bellows that are angularly arranged with respect to the dampers to achieve a side force compensation. Four-corner air suspension systems as in the Jaguar XJ are often combined with adaptive damping systems. The switchable damper units are usually centrally arranged inside the air spring modules to optimize the vehicle package.

Four-corner air suspension systems with switchable additional volumes are semiactive systems. By connecting switchable additional volumes to the individual air spring modules, the air spring rates can be varied depending on the driving condition without changing the pressure levels inside the air spring modules. By opening the pneumatic connections between the individual air spring modules and



Figure 2. Four-corner air suspension system—Jaguar XJ front suspension. (Reproduced by permission of Jaguar Land Rover.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

their corresponding additional volumes, softer air spring rates can be achieved because of the volume increases of the different air spring modules. This results in an improved ride comfort for the vehicle occupants. In the case where the pneumatic interconnection lines between the air springs and their additional volumes are closed, the driving stability and therefore the driving safety during lateral dynamic driving maneuvers are increased because of the resulting higher air spring rates. Therefore, air suspension systems with switchable additional volumes are a potential enabler to defuse the conflict of goals between the driving safety and the driving comfort. Depending on the specific geometric design of the pneumatic interconnection lines, including the electromechanical valves between the air springs and the additional volumes, four-corner air suspension systems with switchable additional volumes can generate damping forces caused by the airflow processes between the air volumes through the interconnection lines, as energy is dissipated by the system. For very high suspension travel velocities and/or unfavorably designed pneumatic interconnection lines between the air springs and the additional volumes, the additional volumes can uncouple dynamically from the air springs. The dynamic uncoupling is driven by the fact that the pressure equalization between the volumes significantly deteriorates because of the dynamics of the suspension excitation and/or the pneumatic interconnection line designs. These decoupling effects can be avoided by a sufficient system design with respect to the fluidic system layout. The general design rule is to use short interconnection lines with large cross-sectional areas. However, interconnection lines with large cross-sectional areas need to be combined with suitable electromagnetic valves with similar cross-sectional areas and adequate air tightness. From a technical point of view, pneumatic valves, which are suitable for use in high volume production, are limited in terms of the maximum switchable cross-sectional area. In addition, the weight and the cost of these valves increase with increasing switchable cross-sectional areas of the interconnection lines. The integration of the individual additional volumes into the air spring modules offers the opportunity to reduce cost and weight of the entire system. At the same time, the lengths of the interconnection lines between the air spring volumes and the additional volumes can be minimized and the airflow between the volumes can be improved accordingly.

In Figure 3, the front suspension air spring module with integrated additional volume of the Porsche Panamera is shown. Air spring bellows with a small wall thickness and an outer guiding tube are used to improve the harshness characteristics of the air spring module. In the upper part of the module, the integrated switchable additional volume (1) is depicted. In the comfort setting of the air suspension system, the additional volume can be added to the base volume (2) by operating a pneumatic valve (3). The base volume is arranged in the lower part of the air spring module. The adaptive damping system with switchable valve unit (4) is coaxially integrated into the air spring module.

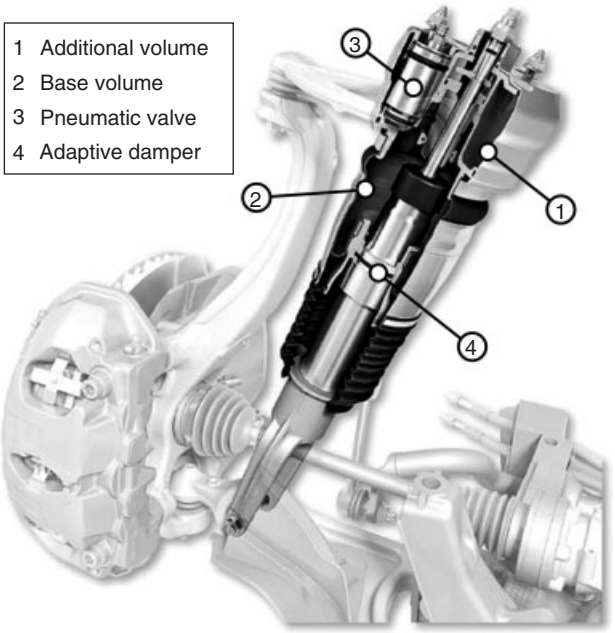


Figure 3. Semiactive four-corner air suspension system—Porsche Panamera front suspension air spring module with integrated switchable additional volume. (Reproduced by permission of Porsche Aktiengesellschaft.)

3 CONTINUOUSLY VARIABLE AIR SUSPENSION SYSTEMS

Usually, four-corner air suspension systems with one switchable additional volume for each air spring module allow only the implementation of two discrete force versus travel characteristics per module. In contrast to these systems, continuously variable air suspension systems enable the individual, seamless transition of the air spring module stiffnesses over a large adjustment range without changing the amount of air in the different air spring modules. In general, two different design approaches for the air spring modules of continuously variable air suspension systems are known. On the one hand, the air spring volumes or the additional volumes can be actively varied to generate continuously changing air spring rates. On the other hand, the effective areas of the air spring modules can be actively increased and decreased over specific roll

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

areas of the air spring bellows over the rolling pistons to generate seamless changing rates of the air spring modules. The interrelationship between the effective area A and the stiffness c of an air spring is shown in Equation 1

$$c = \frac{n \cdot p_i \cdot A^2}{V} \quad (1)$$

On the one hand, the polytropic exponent n and the internal pressure p_i cannot be influenced by the air spring design. On the other hand, the volume V and the effective area A can be influenced by the air spring design within the given package constraints.

As an example for a continuously variable air spring module, Figure 4 shows the Active Air Suspension module from Meritor, Inc. The design of the Active Air Suspension module with coaxially integrated adaptive damper unit (3) is similar to the design of a conventional air spring module. In addition to the main air spring bellow (1) of the module, an additional smaller bellow (2) is mounted directly to the rolling piston to enable the continuous variation of the air spring module rate. This second air spring piston bellow has hermetically sealed interfaces with the rolling piston on the upper and lower ends.

The filling of the air spring piston bellow with compressed air is independent from the filling of the main air volume of the air spring module. By varying the amount of air in the air spring piston bellow, the shape of the rolling piston in the roll area of the air spring bellow changes. Owing to the diameter variation of the air spring piston bellow mounted to the rolling piston with the pressure change inside the air spring piston bellow, the characteristic of the effective area as a function of the air spring module travel can be continuously varied. As a result of the variation of the effective area versus travel characteristic, the air spring module rate can be changed continuously. Because of the high sensitivity of the air spring module rate to the effective area, already relatively small variations of the air volume of the air spring piston bellow can generate large air spring module rate changes. As a result of this correlation not only quasi-static but also fast dynamic air spring module rate variations are possible. For example, at a static air pressure of 9 bars inside the main air spring bellow, the air spring module rate can be varied from 16 to 36 N/mm by changing the air pressure of the bellow mounted to the rolling piston from 1 to 14.4 bars.

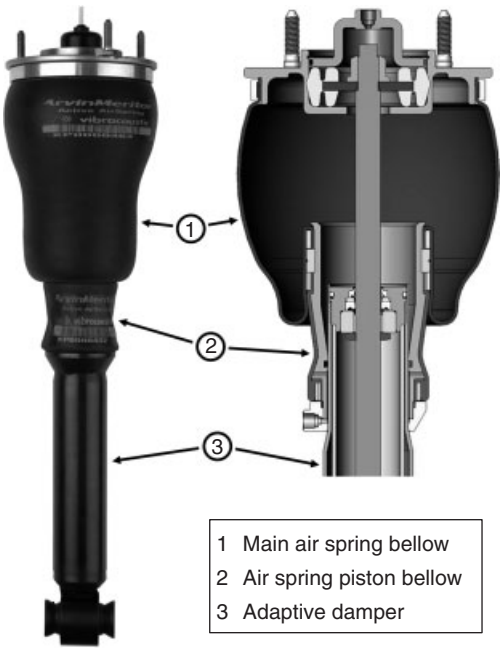


Figure 4. Continuously variable air spring module. (Reproduced by permission of Meritor, Inc.)

4 AIR SPRING DAMPER SYSTEMS

The particularity of air spring damper systems is, that not only the spring forces but also the damping forces are generated with the aid of air as the working fluid. The spring and the damper functionalities are combined in one single functional component. In this manner, the use of conventional hydraulic damper units can be dispensed with. Conventional hydraulic damper modules, used within four-corner air suspension systems with switchable additional volumes, are usually semiactive adaptive units to enable the variation of the damping force versus velocity characteristics of the damper units, in addition to the spring force versus travel characteristics of the air spring modules with switchable additional volumes, depending on the driving situation. However, adaptive hydraulic damper modules are technically more complex in comparison to conventional hydraulic damper units as they require switchable valve elements. In case of air spring damper modules, the damping functionality is achieved with the integration of throttle elements. During the bump and rebound motions of the air spring damper modules, the air is forced to flow through these throttle elements. This leads to the generation of damping forces. Fundamentally, air spring damper units can be distinguished into air spring damper units with two air volumes and air spring damper units with three air volumes. For air spring damper modules with two air volumes, only one maximum of the damping work occurs in the frequency domain that can be aligned with the vertical eigenfrequency

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of the vehicle body. With the aid of air spring damper modules with three air volumes, two separate maxima of the damping work can be implemented in the frequency range. To achieve the optimum system performance for an air spring damper system, the first maximum of the damping work can be adjusted to the vertical eigenfrequency of the sprung mass respectively to the vehicle body mass and the second maximum of the damping work to the vertical eigenfrequencies of the unsprung masses. Theoretical and practical investigations with regard to air spring damper modules and systems are documented in Kranz (1934) and Gold (1973).

Currently, in the fields of passenger cars, no applications of air spring damper systems are known for series production. In the fields of motorbikes, a series production application of an air spring damper system is known on the single arm rear suspension of the BMW HP2 Enduro. This air spring damper module is supplied by Continental and shown in Figure 5 (Müller *et al*., 2005).

The module is based on a compact aluminum part with integrated guiding for a hollow aluminum piston rod, which is shown in the upper part of Figure 5. A damper piston with integrated throttle elements is rigidly mounted to the piston rod. In the lower part of Figure 5, a fabric reinforced rolling lobe is shown, which can roll up and down in direct contact to a cone-shaped rolling piston as the air spring damper unit moves in bump or rebound direction.

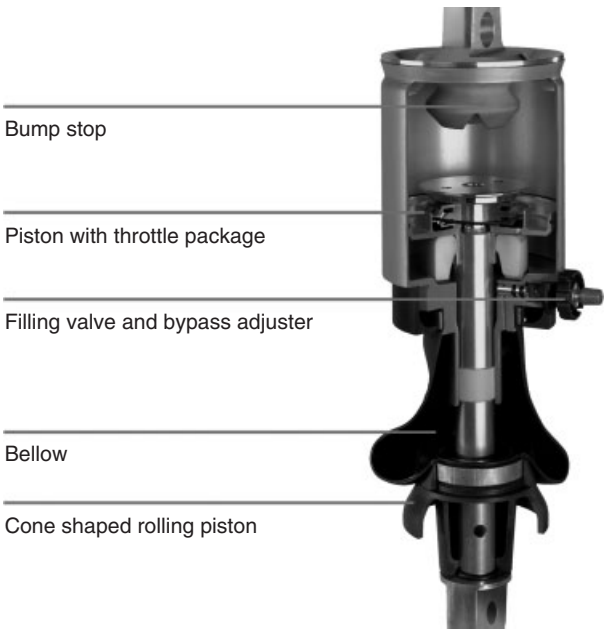


Figure 5. Cutaway view of the BMW HP2 Enduro air spring damper module. (Reproduced by permission of Continental.)

One bump and one rebound stop made of polyurethane rubber provide the necessary progressivity of the air spring damper module in the extreme ends of the available module travel. Conventional asymmetric shim stacks are used as throttle elements in the rebound and compression motion directions. The shim stacks generate annular gaps as a function of the pressure difference over the damper piston. With the aid of the bypass adjuster, which is integrated into the filling valve, two different settings can be applied to the module to avoid extreme damping force peaks. This damping force limitation enables an improvement of the driving comfort.

During the assembly process of an air spring damper module, the mounting tolerances between the base part and the dividing piston do not need to be as narrow as with conventional hydraulic damping units, because the leakages through the seals are not very sensitive to the generated damping forces. This means, that the leakages on the seals do not immediately lead to a reduction of the generated damping forces. This is based on the fact, that the transported air volumes of air spring damper modules are much higher compared to the transported oil volumes of conventional hydraulic damper units. As a result, the outer diameters of air spring damper modules are usually larger compared to the outer diameters of conventional hydraulic damper units. This leads to an increased package space demand of air spring damper modules, that needs to be considered as a part of the full vehicle layout process.

5 INTERLINKED AIR SUSPENSION SYSTEMS

In Wolf-Monheim (2011), a semiactive interlinked four-corner air suspension system, which is based on a conventional four-corner air suspension system with switchable additional volumes, is analyzed. The system enables an improved ride comfort for the vehicle occupants without neglecting the driving safety aspects. The semiactive interlinked four-corner air suspension system is currently not in series production. To analyze the system performance in terms of the ride comfort improvement for the vehicle occupants, the system was installed into a Volvo XC90 prototype vehicle. The semiactive interlinked four-corner air suspension prototype system consists of four air spring modules with switchable additional volumes as well as two pneumatic pipes routed from the left side to the right side of the vehicle to interconnect the two front suspension air spring modules with one another and the two rear suspension air spring modules with one another. Depending on the current driving condition, the pneumatic interconnection lines between the air spring modules can be switched

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

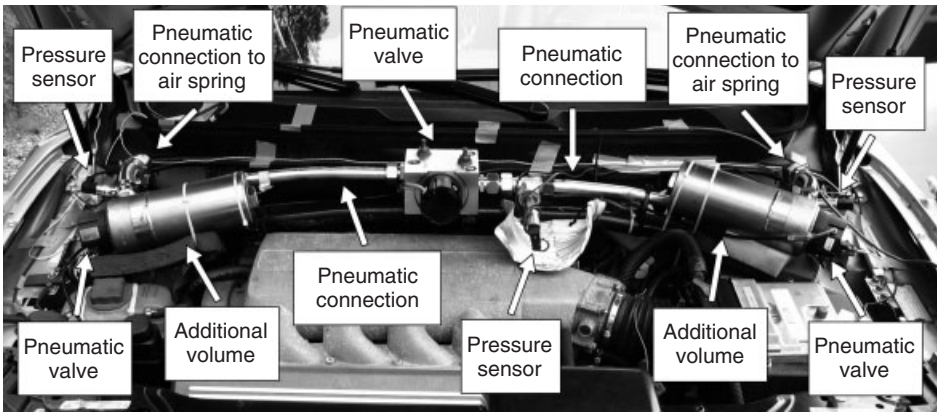


Figure 6. Interlinked air suspension system—Volvo XC90 front suspension. (Reproduced by permission of Wolf-Monheim, 2011. © Wolf-Monheim.)

on and off with the aid of electromagnetic pneumatic valves.

A pneumatic connection between the front and rear suspension air spring modules is not shown in Wolf-Monheim (2011). Within semiactive four-corner air suspension systems, a front to rear connection can be used to reduce the pitch angles of the vehicle body while driving over single impacts and to reduce the pitch angle oscillation accelerations while driving on uneven road surfaces. Within active four-corner air suspension systems, a pneumatic connection between the front and rear suspension air spring modules can be used to allow for an active anti-lift and/or anti-dive pitch angle compensation during acceleration and braking maneuvers.

The achievable improvement in terms of the ride comfort for the vehicle passengers of the semiactive interlinked four-corner air suspension system shown in Wolf-Monheim (2011) results from the fact, that the roll stiffness of the vehicle can be adapted to the individual driving conditions. In this manner, both high roll stiffness during high lateral acceleration driving maneuvers, such as while driving around bends, and low roll stiffness to improve the driving comfort for the vehicle passengers while driving straight ahead on uneven road surfaces can be achieved. Furthermore, semiactive interlinked four-corner air suspension systems can specifically generate roll damping by the appropriate design of the pneumatic interconnection lines and the electromagnetic pneumatic valve geometries as well as by the controlled variation of the cross-sectional areas of the electromagnetic pneumatic valves. Specifically, in combination with switchable additional volumes and continuously variable hydraulic damper modules, the interlinked semiactive four-corner air suspension system can independently influence the roll, pitch and bounce oscillation accelerations of the vehicle body.

In Figure 6, the semiactive interlinked air suspension system with two switchable additional volumes of the prototype vehicle front suspension is shown. Figure 7 shows the equivalent semiactive interlinked air suspension system with two switchable additional volumes installed into the rear suspension of the prototype vehicle. The flows of the pressurized air through the pneumatic interconnection lines between the left and right air spring modules of the front axle and those of the rear axle, respectively, can be interrupted with the aid of electromagnetic pneumatic valves. The additional volumes are connected to the air spring modules with the smallest possible distances and can also be switched on and off separately by means of electromagnetic pneumatic valves. The lengths of the pneumatic interconnection lines of the front and rear suspensions are approximately 1 m. Owing to packaging restrictions, the pneumatic interconnection lines on the front and rear suspensions have some curvatures.

The semiactive interlinked four-corner air suspension system enables a significant improvement in terms of the ride comfort for the vehicle passengers compared to a conventional passive four-corner air suspension system. Compared to a semiactive four-corner air suspension system with switchable additional volumes the ride comfort improvement of the interlinked air suspension system without switchable additional volumes is on a similar level. However, the required package space of the interlinked four-corner air suspension system is significantly lower in comparison to the four-corner air suspension system with switchable additional volumes. Compared to the four-corner air suspension system with switchable additional volumes, the interlinked four-corner air suspension system enables a significantly greater reduction of the body roll oscillation accelerations. The body roll oscillation accelerations are particularly uncomfortable for the vehicle passengers

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

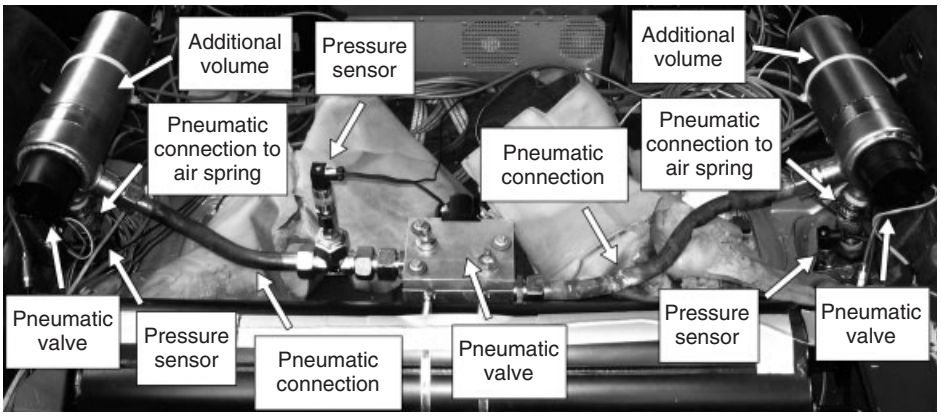


Figure 7. Interlinked air suspension system—Volvo XC90 rear suspension. (Reproduced by permission of Wolf-Monheim, 2011. © Wolf-Monheim.)

compared to the other translational and rotational oscillation accelerations of the vehicle body (Ilgmann, 1979; Kenneth and Griffin, 1978; Parsons, Whitham, and Griffin, 1979).

The vibration intensities introduced to the vehicle body can be further reduced by the combined use of the pneumatic interconnection lines and the additional volumes. Finally, the biggest advantages of the semiactive interlinked air suspension system with switchable additional volumes can be shown in combination with a semiactive adaptive damping system. As mentioned earlier, up to now the semiactive interlinked air suspension system was realized as a prototype system installed into a prototype vehicle only. For future applications of the system under series production conditions, further development work is needed. When the semiactive interlinked four-corner air suspension system with switchable additional volumes is applied to a particular vehicle, it is important that the specific characteristics of the system are aligned with the specific characteristics of the other chassis and suspension systems and components, especially with regard to the driving dynamics. Especially in this context, the tuning of the anti roll bars of the front and rear suspensions should tend toward softer rates whereas the rates of the air spring modules of the front and rear suspensions should tend toward higher rates, as the semiactive interlinked air suspension system allows for a variation of the body roll stiffness. In addition, the air volumes of the pneumatic interconnection lines on the front and rear suspensions should be considered when the air spring modules are designed.

In Wolf-Monheim (2011) various driving test measurement and simulation results on full vehicle bases are presented and analyzed using a prototype vehicle with semiactive interlinked air suspension systems with switchable additional volumes installed on the front and rear suspensions. Initially, simulation and measurement results are analyzed while driving on a ride comfort test track with firstly a sinusoidal profile on one vehicle track only and then secondly with sinusoidal profiles on both driving tracks of the vehicle. During the test drives with only one driving track of the vehicle on the ride comfort test track with the sinusoidal road profile, the other driving track of the vehicle runs on a flat road surface. The results of the test runs with a single track driving on the sinusoidal test track road profile show that the resulting dynamic air spring module stiffnesses can be significantly influenced by the diameters of the pneumatic interconnection lines and by the vehicle driving velocity. Furthermore, the resulting body roll oscillation accelerations and the resulting seat rail vertical oscillation accelerations in relation to the excitation frequencies are studied. The influence of various system designs on the respective effective accelerations is analyzed. With the aid of suitable analysis models to describe the ride comfort as perceived by the vehicle passengers, weighting functions are selected and applied to the simulation as well as the measurement results. In this context, parameters not only for the individual vibration severities but also for the total vibration severities are determined for the various designs of the semiactive interlinked four-corner air suspension system with switchable additional volumes. For the vehicle with opened pneumatic interconnection lines with diameters of 19 mm compared to the prototype vehicle with a reference system design with closed pneumatic interconnection lines and disconnected additional volumes driving with one single track on the ride comfort test track with sinusoidal road profile, the vibration severities of the vehicle body roll oscillation accelerations and the seat rail vertical oscillation accelerations sensed by the vehicle passengers can be reduced by 24%. On various ride comfort test tracks with a wide range of different stochastic unevenness distributions, the potential of the semiactive interlinked air

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

suspension system with switchable additional volumes to reduce the vibration intensities introduced to the vehicle body is also analyzed. In the first place, the potential for variations with regard to the vehicle body control with opened pneumatic interconnection lines of different diameters on the front and rear suspensions, respectively, and with different additional volume sizes is analyzed in the frequency range up to 10 Hz in comparison to the reference vehicle configuration as described earlier.

In this context it becomes evident, that the body roll motions can be influenced by opening the pneumatic interconnection lines to a far greater degree compared to the connection of the additional volumes. The vehicle body roll motions are perceived more clearly by the vehicle passengers compared to other vehicle body motions, such as the vehicle body pitch motions, the vehicle body yaw motions or the vehicle body bounce motions. The reason for this observation can be explained as follows. The road profile excitation content of the left and right driving tracks, primarily leading to vehicle body roll oscillations, results in less dynamic wheel load fluctuations between the two front and two rear wheels, respectively, for the vehicle configuration with opened pneumatic interconnection lines on the front and rear suspensions and disconnected additional volumes in comparison to the vehicle configuration with closed pneumatic interconnection lines on the front and rear suspensions and connected additional volumes.

This is due to the fact that in the case of opened pneumatic interconnection lines on the front and rear suspensions and antiphase travel excitations of the wheels of each axle, the pressure differences between the two front and two rear suspension air spring modules, respectively, can be equalized because of the air mass exchanges through the pneumatic interconnection lines. This results in almost constant forces of the air spring modules of one axle.

In contrast to the aforementioned, the level of the bounce oscillations of the driver's seat as well as the level of the pitch oscillations of the vehicle body can be better varied by connecting or disconnecting the additional volumes compared to switching the pneumatic interconnection lines between the left and right sides of the vehicle. The greater efficiency of the switchable additional volumes, with regard to the bounce and pitch angle degrees of freedom, follows from the fact, that the road profile excitation content of the left and right driving tracks, primarily leading to bounce and pitch angle oscillations of the vehicle body, can only influence the forces introduced into the vehicle body and therefore the vehicle body bounce and pitch angle oscillations in the case, when the additional volumes on the front and rear suspensions are varied respectively switched with the resulting variations of the base rates of the individual air spring modules.

The pneumatic interconnection lines between the air spring modules of the front and rear axles, respectively, have only a minor influence on the resulting air spring module forces, when the road profile excitation content of the left and right driving tracks primarily leads to bounce and pitch angle oscillations of the vehicle body. This can be explained by the fact, that the base rates of the air spring modules are equal for opened and closed pneumatic interconnection lines between the left and right sides of the vehicle for almost identical road profiles of the left and right tracks.

In a similar way, the oscillation acceleration reduction potential in the enhanced frequency range of up to 100 Hz is analyzed for different design variants of the semiactive interlinked air suspension system with the switchable additional volumes, in comparison to the reference vehicle configuration with closed pneumatic interconnection lines and disconnected additional volumes. These tests are based on acceleration measurements on the outer seat rail of the driver's seat and on the steering wheel. In addition to the analyses of the influence of the pneumatic interconnection lines and the switchable additional volumes on the ride comfort for the vehicle passengers, also the influence of a semiactive adaptive damping system is analyzed. The results are evaluated in various individual frequency bands. For all tests, the percentage reductions of the root mean square values of the power spectral densities (PSDs) are determined.

For opened pneumatic interconnection lines, the PSDs of the seat rail vertical oscillation accelerations can be reduced in comparison to the reference vehicle configuration by up to 9.6%, depending on the individual frequency range. The PSDs of the steering wheel oscillation accelerations can also be reduced significantly (by 15.6%) with opened pneumatic interconnection lines in comparison to the reference vehicle configuration. In case the additional volumes are connected, instead of opening the pneumatic interconnection lines, the PSDs of the seat rail vertical oscillation accelerations as well as the PSDs of the steering wheel oscillation accelerations can be reduced by a similar magnitude. On the seat rail of the driver's seat, the PSDs can be decreased by up to 11.3% and on the steering wheel by up to 12.7%, depending on the frequency range.

The combined use of the pneumatic interconnection lines and the additional volumes enables the further increase of the ride comfort perceived by the vehicle passengers. In this case, the PSDs of the seat rail vertical oscillation accelerations can be diminished by up to 15.4%, depending on the frequency band. The PSDs of the translational steering wheel oscillation accelerations can be reduced by up to 19.8%.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

A further improvement of the vibration comfort felt by the vehicle passengers is possible if a semiactive adaptive damping system is used in addition to the opened pneumatic interconnection lines between the air spring modules of the front and rear axles, respectively, and the connected additional volumes.

In comparison to the reference vehicle configuration with closed pneumatic interconnection lines between the left and right sides of the vehicle, disconnected additional volumes, and inactive adaptive damping system, this system configuration enables the reduction of the PSDs of the outer seat rail vertical oscillation accelerations measured at the driver's seat of up to 24.6%. In addition, it is possible to decrease the PSDs of the translational steering wheel oscillation accelerations by up to 29.5%.

Subsequent publications referring to semiactive interlinked four-corner air suspension systems are Wolf-Monheim *et al*. (2008a), Wolf-Monheim *et al*. (2008b), Wolf-Monheim *et al*. (2008c), and Wolf-Monheim *et al*. (2009).

6 ACTIVE AIR SUSPENSION SYSTEMS

From a physical point of view, active air suspension systems can be based on either an active air mass change or an active air volume change in the individual air spring modules. Next to the ride height control functionality of conventional semiactive four-corner air suspension systems with and without additional volumes, active four-corner air suspension systems also enable additional functionalities, for example, active roll and/or pitch angle compensations. On the one hand, an active air suspension system based on dynamic air mass change can be realized, for example, by means of high pressure air reservoirs and air compressors. If required, the air masses can be shifted actively from the high pressure air reservoirs to the air spring modules. On the other hand, an active air suspension system based on dynamic air volume change can be made, for example, with the aid of hydraulic, pneumatic, or electromechanical actuators. To minimize the energy consumption of the actuators used within the system layout, the static loads generated by the sprung mass can be carried partly or fully with the help of passive spring elements. This means, that only the requested dynamic force variations are demanded from the actuators.

In Zhang (2006), three different active air suspension systems based on dynamic air volume change are presented and compared with one another with the aid of complex numerical simulation models. All the three systems documented in Zhang (2006) are based on a driven double-acting cylinder, which moves two additional connected mechanical two-chamber cylinders. These mechanical two-chamber cylinders are used to supply pressurized air to one air spring module and remove pressurized air from another. With the help of appropriate valve technology, the active air suspension system can be switched between roll angle compensation mode for lateral acceleration maneuvers and anti-lift and anti-dive pitch angle compensation mode for acceleration and braking maneuvers, respectively, depending on the driving situation. For the fully pneumatic system (Figure 8), a pneumatically operated double-acting drive cylinder is used, whereas a hydraulically operated double-acting drive cylinder is utilized within the hydropneumatic system (Figure 9). The third system is referred to as the *fully hydraulic system* (Figure 10), as in this case, a hydraulically operated double-acting drive cylinder moves two additional connected hydropneumatic two-chamber cylinders. In this system, four additional passive cylinders with one hydraulic and one pneumatic chamber are needed to realize the active air volume change of the individual air spring modules.

For the fully pneumatic active air suspension system, two key advantages in comparison to the other two systems can

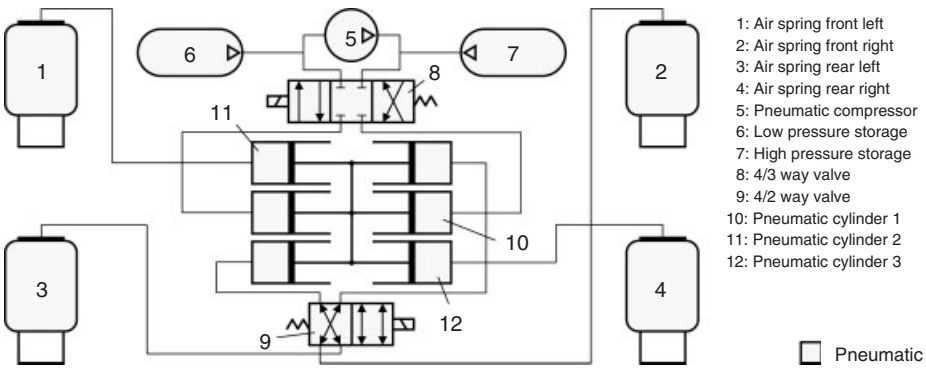


Figure 8. Active air suspension system—fully pneumatic system. (Reproduced by permission of Zhang, 2006. © Zhang.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

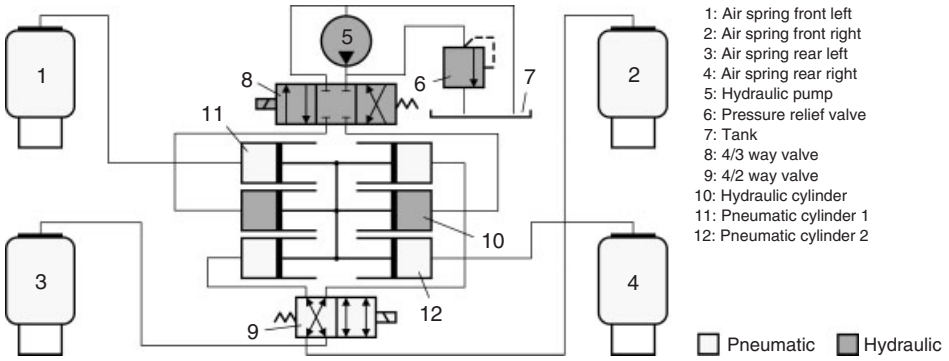


Figure 9. Active air suspension system—hydropneumatic system. (Reproduced by permission of Zhang, 2006. © Zhang.)

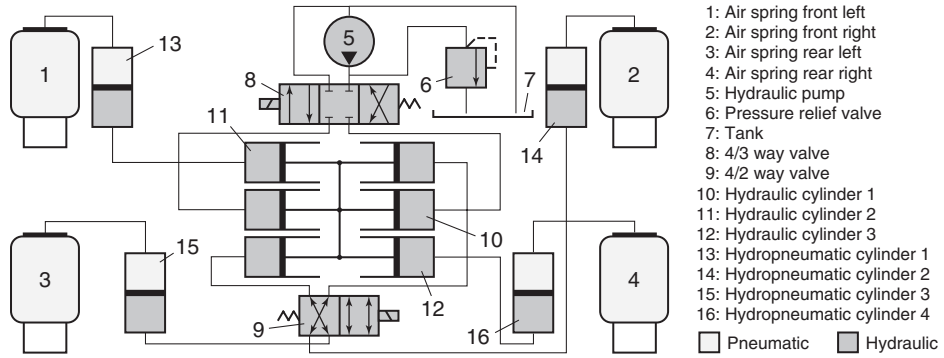


Figure 10. Active air suspension system—fully hydraulic system. (Reproduced by permission of Zhang, 2006. © Zhang.)

be mentioned. On the one hand, only pressurized air is used in this case as the working fluid and hydraulic oil can be avoided to improve the environmental friendliness of the system. On the other hand, the actual design and development of the system is relatively easy. However, the disadvantage of the fully pneumatic active air suspension system is that it tends to oscillate because of the compressibility of the pressurized air as the working fluid. This results in difficulties in terms of the precise position control of the individual actuators.

In contrast to the fully pneumatic active air suspension system, the fully hydraulic active air suspension system can be controlled in a stable way as it features only a small amount of volumes filled with pressurized air. However, it is technically more complex and demands more packaging space compared to the fully pneumatic and the hydropneumatic active air suspension systems. In addition, it is also more expensive compared to the other two systems. The hydropneumatic active air suspension system is implemented and tested in a prototype vehicle.

A further active air suspension system is presented in Kranen (1996). In the system presented here, a volume exchanger is utilized as the actuator to shift the pressurized air between the air spring modules. The power steering device supplies the energy needed to actuate the system.

7 SUMMARY

This chapter gives an overview of advanced air suspension technologies and applications based on current state-of-the-art four-corner air suspension systems. In the introduction, the main advantages and base functionalities of air suspension systems as well as the available design variants of air spring modules are presented.

Section 2 deals with the two key distinct categories of the state-of-the-art four-corner air suspension systems. The first category covers air suspension systems without switchable additional volumes and the second category systems with switchable additional volumes.

Continuously variable air suspension systems presented in Section 3 offer the advantage to vary the force versus air spring module travel characteristics of the individual air spring modules and therefore also the air spring module stiffnesses in a continuous way. For these systems, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

stiffness variations can be generated by varying either the effective areas or the air spring module respectively additional volumes.

Air spring damper systems as discussed in Section 4 do not only provide a springing functionality but also a damping functionality based on pressurized air as the working fluid. The air damping is achieved with the aid of pneumatic throttle elements integrated into the air spring damper modules where the pressurized air is forced to flow through while dissipating energy. Compared to conventional hydraulic damping elements, air spring damper units avoid the use of hydraulic oil as a working fluid and therefore improve the environmental friendliness.

In Section 5, interlinked air suspension systems, characterized by pneumatic interconnection lines between the individual air spring modules, are described. Semiactive interlinked four-corner air suspension systems offer enhanced possibilities to tune the vehicle body bounce and roll stiffness properties independently from one another, depending on the current driving condition. In addition, the vehicle body pitch angles can be influenced while driving over single impacts or on uneven road surfaces.

Finally, active four-corner air suspension systems are introduced in Section 6. Active four-corner air suspension systems enable additional functionalities, for example active vehicle body roll angle or anti-lift and/or anti-dive vehicle body pitch angle compensations during lateral acceleration or longitudinal acceleration and braking maneuvers.

ENDNOTE

1. The use of "bellow" should be understood as a flexible bladder or bag containing pressurised air.

REFERENCES

Gold, H. (1973) Über das Dämpfungsverhalten von Kraftfahrzeug-Gasfedern. PhD thesis. RWTH Aachen University, Aachen, Germany.

Heisler, H. (2002) *Advanced Vehicle Technology*, 2nd edition, Elsevier Science, Amsterdam, The Netherlands.

Ilgmann, W. (1979) *Ergonomische Untersuchungen über die Einwirkung rotatorischer Schwingungen, Forschungsbericht aus der Wehrtechnik*, Bundesministerium der Verteidigung, Dokumentationszentrum der Bundeswehr, Bonn, Germany.

Kenneth, P. and Griffin, M. (1978) The effect of rotational vibration in roll and pitch axes on discomfort of seated subjects. *Ergonomics*, **21** (8) pp. 615–625. Taylor & Francis, London, England.

Kranen, H. (1996) *Auslegung und Aufbau einer Steuereinheit für ein aktives pneumatisches Wankausgleichssystem*. Diploma thesis. Institute of Automotive Engineering (ika), RWTH Aachen University, Aachen, Germany.

Kranz, M. (1934) *Luftfederung für Kraftfahrzeuge*. PhD thesis. Technical University of Stuttgart, Stuttgart, Germany.

Müller, P., Reichl, H., Heyl, G., *et al*. (2005) Das neue "Air Damping System" der BMW HP2 Enduro. *ATZ Automobiltechnische Zeitschrift*, **107** (10) pp. 848–857. Vieweg+Teubner Verlag / GWV Fachverlage GmbH, Wiesbaden, Germany.

Parsons, P., Whitham, E., and Griffin, M. (1979) Six axis vehicle vibration and its effects on comfort. *Ergonomics*, **22** (2) pp. 211–225. Taylor & Francis, London, England.

Puff, M. (2009) Entwicklung einer Prüfspezifikation zur Charakterisierung von Luftfedern. Abschlussbericht zu einem vom VDA FAT AK20 geförderten Projekt. Projektbericht Fluidsystemtechnik, Technische Universität Darmstadt, Verband der Automobilindustrie (VDA), Berlin, Germany.

Wallentowitz, H. (2005) *Vertikal-/Querdynamik von Kraftfahrzeugen, Vorlesungsumdruck Fahrzeugtechnik II*, Forschungsgesellschaft Kraftfahrwesen Aachen mbH, Aachen, Germany.

Wolf-Monheim, F., Frantzen, M., Seemann, M., and Wilmes, M. (2008a) Modeling, Testing and Correlation of Interlinked Air Suspension Systems for Premium Vehicle Platforms. *Proceedings of 32nd FISITA Congress, F2008-SC-040, FISITA*, London, England.

Wolf-Monheim, F., Seemann, M., Schommer, M., and Wilmes, M. (2008b) Fahrkomfortoptimierung durch den Einsatz vernetzter Luftfederungssysteme. 17. Aachener Kolloquium Fahrzeug- und Motorentechnik 2008, Forschungsgesellschaft Kraftfahrwesen Aachen mbH, Aachen, Germany.

Wolf-Monheim, F., Frantzen, M., Seemann, M., and Wilmes, M. (2008c) Das Potenzial gekoppelter Luftfederungssysteme zur Verbesserung des Fahrkomforts. 2. Fachtagung Federn und Dämpfungssysteme im Fahrwerk, mic—management information center, Munich, Germany.

Wolf-Monheim, F., Schumacher, M., Frantzen, M., *et al*. (2009) Interlinked air suspension systems—the influence on ride comfort in testing and simulation. *ATZautotechnology*, **9** (3) pp. 58–61. Vieweg+Teubner Verlag / GWV Fachverlage GmbH, Wiesbaden, Germany.

Wolf-Monheim, F. (2011) Fahrkomfortoptimierung durch den Einsatz vernetzter Luftfedersysteme. PhD thesis. RWTH Aachen University, Aachen, Germany.

Zhang, J. (2006) Aktives Luftfedersystem für einen PKW. PhD thesis. RWTH Aachen University, Aachen, Germany.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto004
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ULSAS—Suspension, A Comparison of Rear Suspension Design

Henning Wallentowitz

RWTH Aachen University, Aachen, Germany

1 Introduction 1
2 Starting Situation with Rear Axle Suspension Systems 1
3 Rear Axles Developed by Lotus 4
4 Conclusion 10
References 12

1 INTRODUCTION

Chassis engineering basically refers to the designing of axles and chassis suspension systems and their vehicle integration. A company unique driving performance shall be achieved. Chassis suspension systems, especially those of passenger vehicles, can be categorized by dividing into rigid axles, semirigid axles, and independent suspensions. Figure 1 shows this classification.

Ever since the intensive use of aluminum in vehicle construction, there have been widespread discussions about materials used in the chassis suspension systems. Those discussions are still on-going. Owing to the differentiation between sprung masses (body mass) and unsprung masses (axle parts), the lightweight layout of the unsprung masses with regard to driving comfort, weight, as well as kinematics, is considered in the design of suspension parts. Axles should be as light as possible, hence higher grade steels, lightweight metals, or even synthetics, for example, fiber composites, provide possible solutions, at least on first sight.

Ten years ago, the international steel industry gave an assignment to Lotus engineering to do a survey on lightweight chassis suspension systems. The weight-based competitiveness of steel should be shown for these applied usages. The design proposal, which became known as the *ULSAS-study* (*Ultra light Steel Auto Suspension*) (Lotus, 2000; World Auto Steel, 2012; Stahl-Informations-Zentrum, 2012), focused on rear axles and came up with interesting solutions. This study will be partly reviewed in the following article. For front axles, this volume of the Encyclopedia contains the detailed report "Lightweight front suspension, a comparison" (see Light Weight Front Suspensions, a Comparison) by Mrs Dr. Chang (2006).

2 STARTING SITUATION WITH REAR AXLE SUSPENSION SYSTEMS

The engineers at Lotus started working on the ULSAS study by analyzing existing car chassis suspension systems. They focused on the four most important axles. Figure 2a–d summarizes those axles.

Figure 2a shows the so-called Double Wishbone axle (often known as *Short-Long Arm* suspension in the United States) with a body-connected longitudinal control arm, which is made from a deep-drawn steel plate. The two lower transverse control arms are mounted to a small rear suspension subframe and they are made of forged steel. The front lower transverse control arm carries a suspension strut that is directly held up at the body at its upper end. The knuckle is forged as well and is shaped to follow the contour of the tire until it reaches a ball joint, to which

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

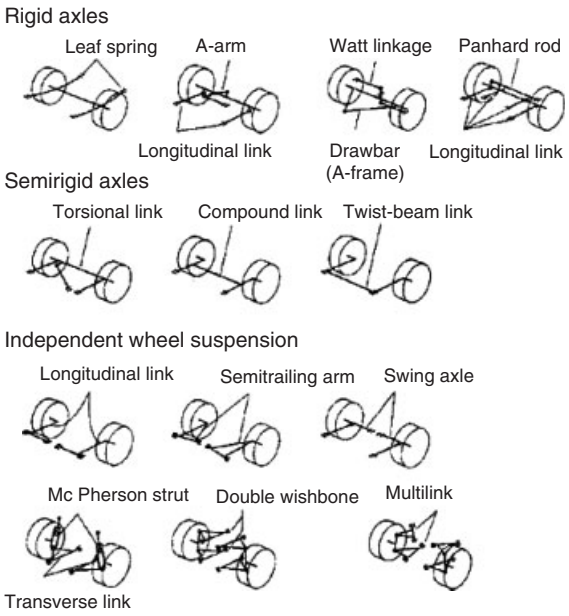


Figure 1. Systematic classification of chassis suspension of passenger vehicles. (Reproduced with permission from ika RWTH Aachen University. © ika Aachen.)

the upper transverse control arm is mounted. This upper control arm is also forged and is directly attached to the body at its other end with an elastomeric bush. According to the ULSAS study, the mass of the axle is 39.8 kg. This nondriven suspension assembly is fitted to the rear of the Honda Accord.

Figure 2b shows a driven rear axle of BMW, the so-called integral axle. The large suspension subframe is made of hydroformed aluminum. Welded consoles, also consisting of aluminum, are the connections to the upper transverse control arms. These control arms are made from forged aluminum. The lower control arm consists of a large box section, which is made up of metal sheet, and is mounted to the suspension subframe with two bearing connections and with one ball joint to the wheel carrier. The so-called integral connecting rod connects the wheel carrier with the large box section transverse control arm (Matschinsky, 1998). The "integral rod" reduces elastic rotational deformations during acceleration and braking. This reduces the so-called wind-up of the wheel carrier that otherwise may lead to wheel hop or "tramping" [reasons are changes in friction (slip) between the tire and the road because of the wind-up of the wheel carrier].

Suspension is done through suspension struts, which are not shown in Figure 2b. Those struts are directly mounted to the wheel carrier. The large suspension subframe enables the connection of the axle in the area of the vehicle's stiff longitudinal chassis beam. This suppresses the transmission of noises into the body, created by the tires or the rear axle differential. The differential is supported in the suspension subframe by elastomeric bushes, as is the suspension subframe in the body. That is why this is called a *double-elastic bearing*. The mass of this BMW integral axle is stated at 46.67 kg in the ULSAS study (aluminum axle).

Axle 2c is a lightweight suspension strut rear axle. It was taken from a Ford Mondeo. Via a tension strut

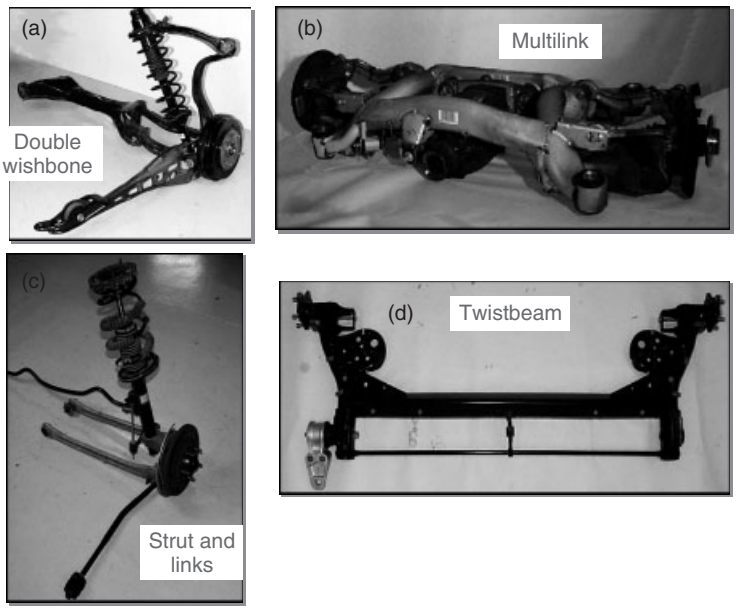


Figure 2. (a–d) Most important axles. (Reproduced from World Auto Steel. © World Auto Steel.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and connected rubber bearings, the longitudinal forces are transferred from the wheel carrier into the car body. The transverse forces are transferred to the body through two lower transverse control arms (forged aluminum parts) and the upper suspension strut bearing. This upper suspension strut bearing additionally transmits the vertical forces from spring and damper.

Owing to the upper suspension strut bearing having to brace side forces, the piston in the damper and the piston rod bearing have to handle lateral forces. The induced torque by the vertical forces can be compensated by the inclined function line of the coil spring within the design position of the wheel suspension. This "trick" to compensate torque is used commonly on front suspension McPherson struts, but it can also be applied in rear axles. To achieve a high stabilizer conversion ratio, the stabilizer (or antiroll bar) is mounted to the suspension strut using a rod.

To provide adequate clearance between the tire and the damper strut, the bottom mount of the strut into the hub carrier (often clamp design or welded) is offset laterally inward. To carry the torque between wheel hub and damper, the hub carrier is designed as a cast or forged part to accommodate this. This suspension strut rear axle is designed as a nondriven axle. The ULSAS study quotes a mass of 39.7 kg.

The twist-beam axle that is shown in Figure 2d, and which originates from the Volkswagen Golf, is made from steel plates to achieve the required stiffness. Longitudinal control arms, connected via the twist beam, carry the wheel carrier and the brakes (which are not shown in Figure 2d). The longitudinal arms are mounted directly to the body via rubber bearings. The springs and dampers are then connected to the longitudinal control arms via brackets. As the connections of the brackets, as well as the connections between the longitudinal control arms and the twist beam, are usually achieved by welding, the stiffness is limited through the welding seams. Additionally of disadvantage is the fact that, during cornering, the lateral forces tend to cause the whole axle assembly to turn about a vertical axis, thus the sideslip angles are reduced. The wheels are steering into the turn; this generates an oversteering tendency. The mass of this twist-beam axle is 33.4 kg (without brakes) according to the ULSAS study.

For illustrating the results of the survey, the Lotus engineers used the so-called spider charts. The performance ratio of the four axles in view of certain demand aspects is evaluated. This ratio is then marked on the chart. The higher the ratio, the better is the performance of the axle. Figure 3 shows the spider chart for the four analyzed axles.

The greatest differences can be observed in the categories of cost, manufacturing, and the space demands each individual axle poses to the vehicle structure. The multilink axle is superior to all other axles in respect of "ride and handling" and "refinement (NVH)" performances (NVH, noise–vibration–harshness). The better ride and handling performance is due to the better wheel control afforded by the multilink design, and the better NVH performance is a result of the double-elastic bearing as mentioned earlier. The implemented mass is rated the same with all four axles.

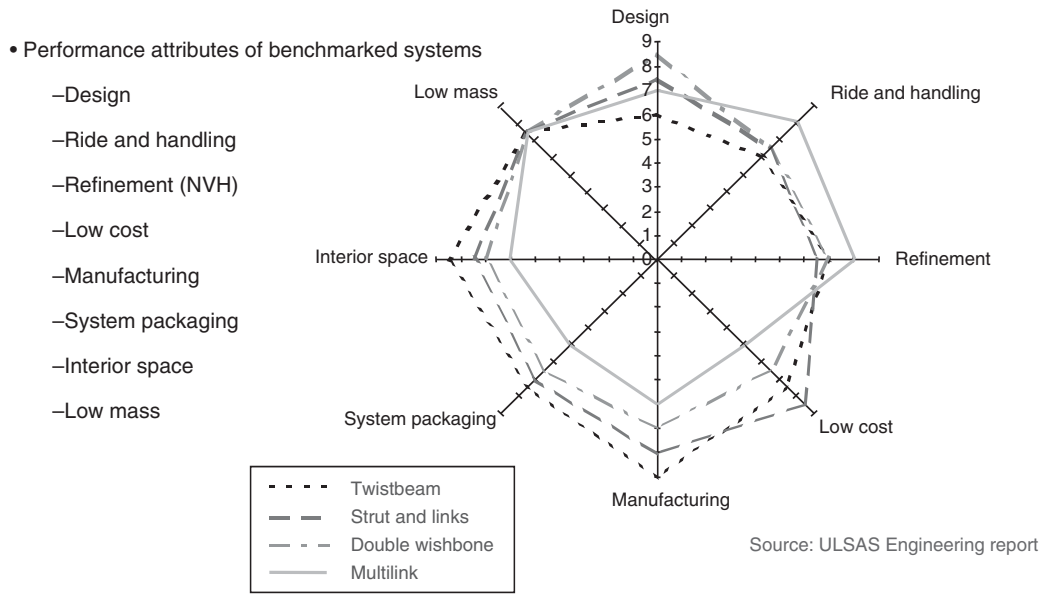


Figure 3. Spider-chart. (Reproduced from World Auto Steel. © World Auto Steel.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

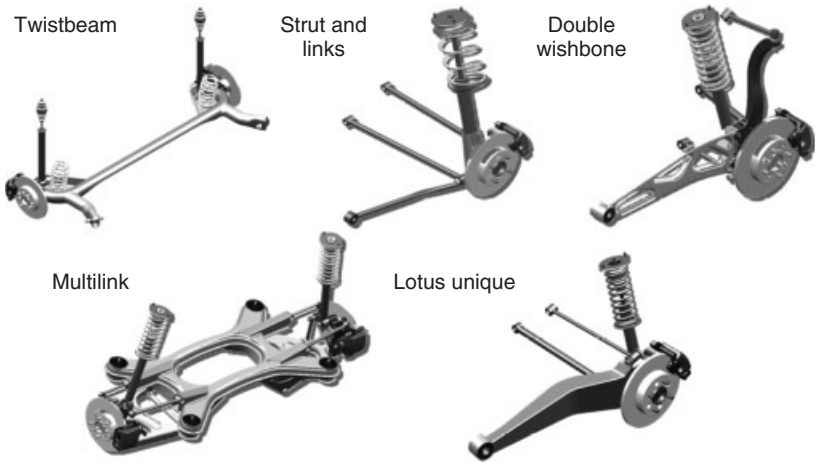


Figure 4. Steel axles as proposed by Lotus engineering. (Reproduced from World Auto Steel. © World Auto Steel.)

3 REAR AXLES DEVELOPED BY LOTUS

Within the development done by Lotus, the engineers proposed rear axles made of just steel. However, different grades of steel were used. The development target was to

1. decrease the mass of the steel-made axles by at least 20% without increasing the costs.
2. achieve the same mass as the aluminum-made axles but using steel instead, and at the same time decrease the costs by at least 20%.

To accomplish these targets, five new axles were developed. They were worked on in design engineering as well as using simulation technology (finite element analysis). Figure 4 shows a compilation of those proposals. In addition to the axles used in practice, there is also another solution shown, called *Lotus Unique*.

Furthermore, they were able to achieve the set goals of reducing the mass, as shown in the comparison (Table 1), illustrating the masses of current axles and the new developments.

Table 1. Mass comparison of the analyzed and the newly developed axles.

	Realized (kg)	ULSAS (kg)	b.	Car
Double wishbone	39.80	32.88	17%	Honda Accord
Multilink	46.67	48.00	3%	BMW 5er
Strut and links	39.70	29.87	25%	Ford Mondeo Taurus
Twistbeam	33.40	27.30	18%	VW Golf
Lotus unique	39.80	26.49	34%	Honda Accord

The differences in cost and the production studies performed by the ULSAS study are not examined any further in this report because of them being rated differently today. However, axle developers are advised to take a closer look into the study, as it is full of interesting suggestions.

3.1 Double wishbone axle

This axle matches the analyzed axle's basic buildup. However, there are differences in some parts. Figure 7 shows these parts; especially the longitudinal chassis beam that is explicitly shown as pressed steel plate made of high grade steel.

The housings for the rubber bearings are tube sections, which are welded; the transverse control arms are also steel tubes with rubber bearings on the end. This enables a connection to the wheel carrier on the one end and to the body on the other end. Whether there is supposed to be a suspension subframe in between the body and the transverse control arm or not is left unanswered in Figure 5. The wheel carrier itself consists of very high grade forged steel. A ball joint connects the upper transverse control arm to the knuckle. The control arm itself consists of only a tube. The suspension strut is directly screwed to the wheel hub and the connection to the body is done conventionally using suspension strut mounts.

The overall rating of this axle was entered into a spider chart as shown in Figure 6. It shows the rating as compared to the other, currently used axles, which were set out as benchmarks. The indications ULSAS D&P refer to are the vehicle classes the axles are intended to be used within. Vehicle class E will be looked at later (Figure 10).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

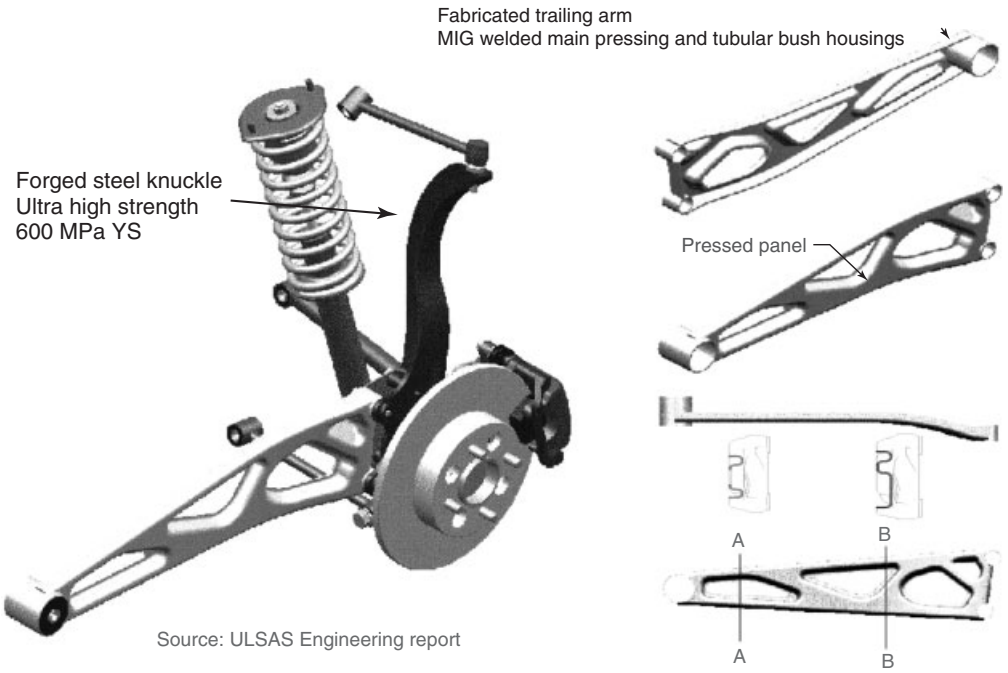


Figure 5. Lightweight construction double wishbone axle. (Reproduced from World Auto Steel. © World Auto Steel.)

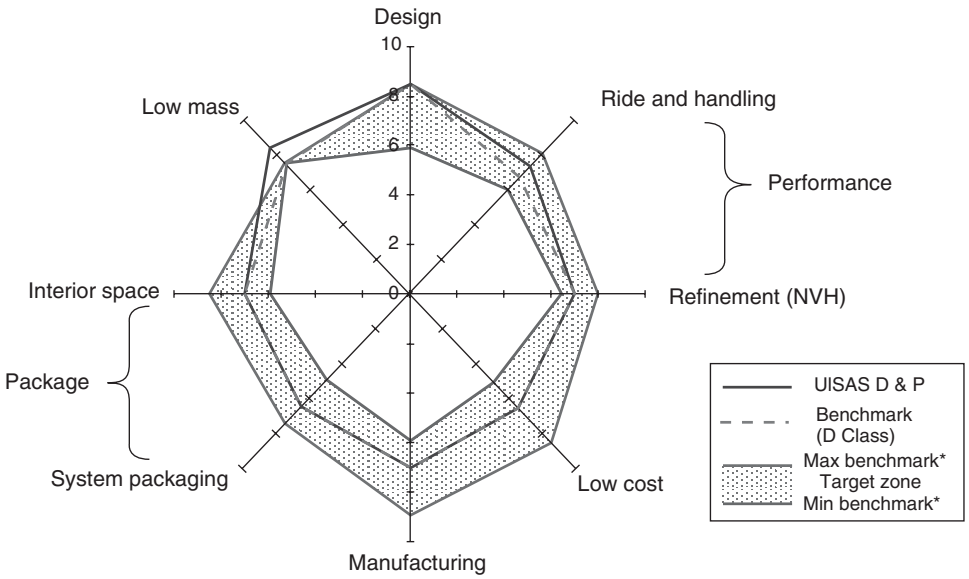


Figure 6. Classification of the double wishbone axle into the benchmark zones as determined before. (Reproduced from World Auto Steel. © World Auto Steel.)

As a result of this classification, it shows that especially the mass could be reduced without creating additional costs. The behavior of the axle (ride, handling, and NVH) competes very well within the competitor class, as well as the other characteristics also staying within the benchmark parameters.

3.2 Multilink rear axle

The multilink rear axle, which in the analyzed integral axle version was equipped with an aluminum suspension subframe and taken from an upscale vehicle class, will also feature a suspension subframe in its steel version.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 7. Multilink rear axle. (Reproduced from World Auto Steel. © World Auto Steel.)



Figure 9. View of a multilink rear axle. (Reproduced from World Auto Steel. © World Auto Steel.)

This suspension subframe will, however, not be created by hydroforming but by pressing two sheet metals into frames that are then welded together. The study suggests using 250–500 MPa YS steel. The connectors for the transverse control arms are welded onto the sheet metals of the suspension subframe. These connectors are simply created through angled sheet metal. Figure 7 shows this axle in detail.

Further elements made of steel plates are the lower transverse control arms. They were also designed in a clamshell way. Their production is recommended using 250 MPa YS steel. Figure 8 contains special production instructions for these transverse control arms. Especially important is the flange depleted welding of the two half shells.

The tubular wishbones can be seen clearly in Figure 9. The vehicle-based side has rubber bearings; the wheel-based side has a ball joint. Figure 9 also features the integral link as a connection between the wheel carrier and the lower wishbone.

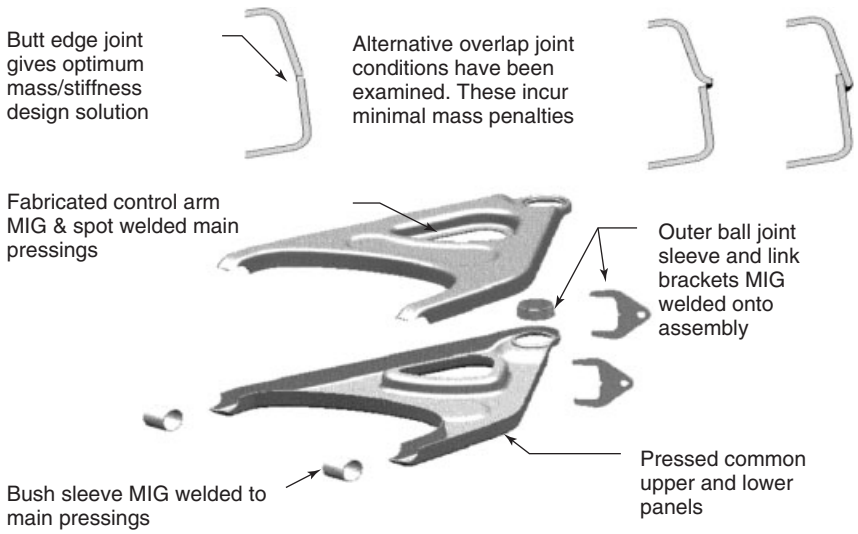


Figure 8. Production instructions for the lower transverse control arms of the multilink rear axle. (Reproduced from World Auto Steel. © World Auto Steel.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

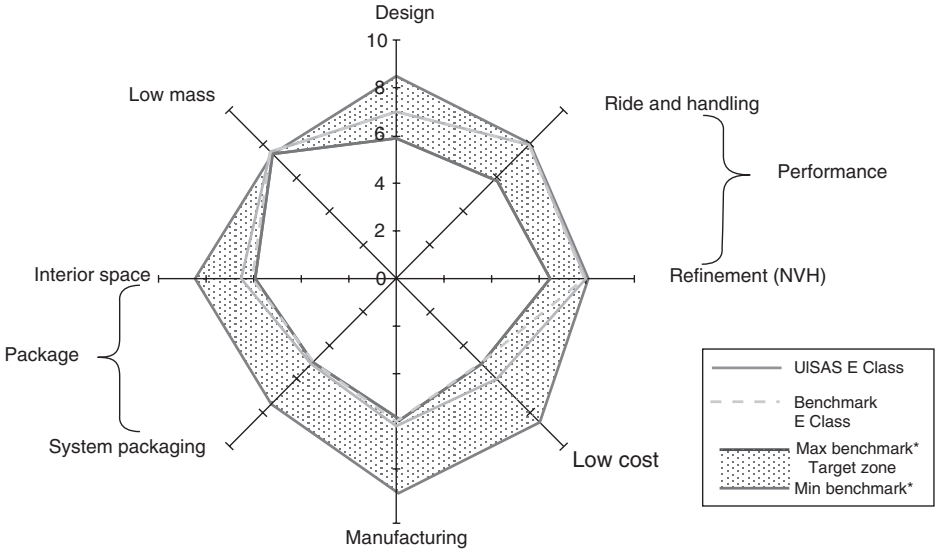


Figure 10. Spider chart of the multilink rear axle. (Reproduced from World Auto Steel. © World Auto Steel.)

The wheel carrier is forged of 750 MPa YS steel; the wishbone tubes consist of 250 MPa YS steel. The engineers at Lotus put the performance data of the axle into a spider chart once again. Figure 10 shows this chart.

It shows that only a minor weight reduction could be achieved, but a significant 30% cost reduction was realized. Performance and packaging match the evaluation criteria of the axles.

3.3 Strut and links rear axle

This axle type was designed for light- and heavyweight vehicles. Figure 11 shows the suggestion for the lighter vehicle. The applied constructions only differentiate in the connection of the tubular longitudinal support beam to the wheel carrier. Instead of connecting to the frontal wishbone, the solution for heavyweight vehicles applies the connection

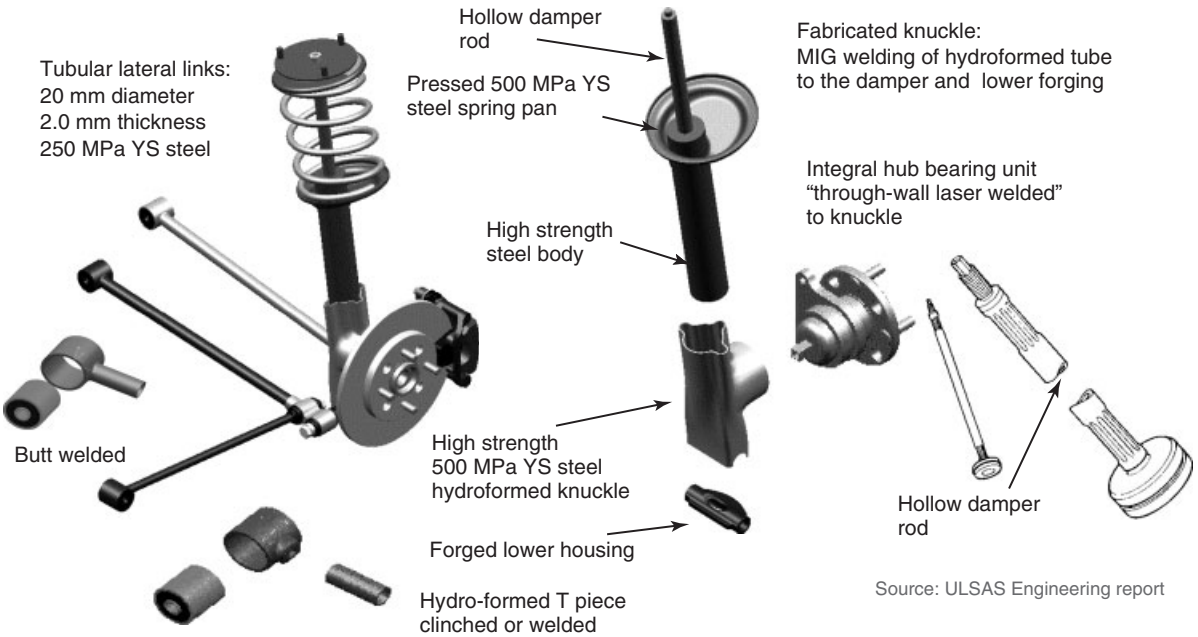


Figure 11. Overview over the suggestions of the strut and links rear axle. (Reproduced from World Auto Steel. © World Auto Steel.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of the longitudinal support beam to the knuckle. Then, the "forged lower housing" part is altered.

Essential for utilizing the lightweight potential of this axle is the creation of the wheel carrier, which is produced from a hydroformed tube. The hub bearing units, as "finished parts," are welded into the metal sheet knuckle. A special "through penetration welding by laser process" is applied as the laser beam bonds the hydroformed part directly to the lower casing of the wheel hub unit. Accordingly, the steel part, which represents the lower end of the hydroformed wheel carrier, is welded into place as well. This part is welded onto the shock absorber tube. The wheel hub unit also carries the brake caliper mounts. The wishbones are produced as tubes onto which are welded smaller tubular parts that serve as housings for the rubber bearings.

The piston rod for the damper is suggested to be a hollow pipe, which, for stiffness reasons, is manufactured with longitudinal grooves but not in the areas that cross the sealing.

The mass comparison, as given in Table 1, indicates a weight reduction of 25% as compared to contemporary state of technology.

The estimation of performance of the axle for both the light- and heavyweight vehicles is shown in Figure 12. It shows that the mass, but not the costs, could be reduced. Especially ride and handling experience a positive influence with this type of axle construction. Otherwise, the axle is within the benchmark parameters.

3.4 Twist-beam rear axle

With lower and middle class vehicles, which usually have front wheel drive, the rear axle of choice is typically the twist-beam axle. The analysis showed that its welding seams limit the capacity of this axle. Hence, the Lotus engineers put the task on themselves to reduce those seams as much as possible. Figure 13 shows the suggested design and the applied steel grades.

Figure 13 underlines that both the longitudinal control arms and the twist beam are made from one piece of pipe. To achieve the torsion softness of the twist beam, the pipe is cut open with a plasma cutter and the edges are bent a little. This enables only the trailing arm toward the bearing, the spring base, and the mount for the wheel bearing to be welded. The bolted wheel bearing construction then enables the connection of the brakes to the axle. The angling of the rubber bearing in plan view, between the axle and the body that enables some compensation of side force oversteering, can be achieved by simply twisting the trailing arm. This trailing arm is produced by hydroforming. The housing in which the rubber bearing is mounted to the trailing arm can be manufactured by rearranging the trailing arm tube.

In addition, the position of the shear center, which together with the bearings of the longitudinal control arms determines the kinematic data of the axle, can be chosen freely (within a certain range) with this type of design. It only matters as to which side the cross profile is cut open,

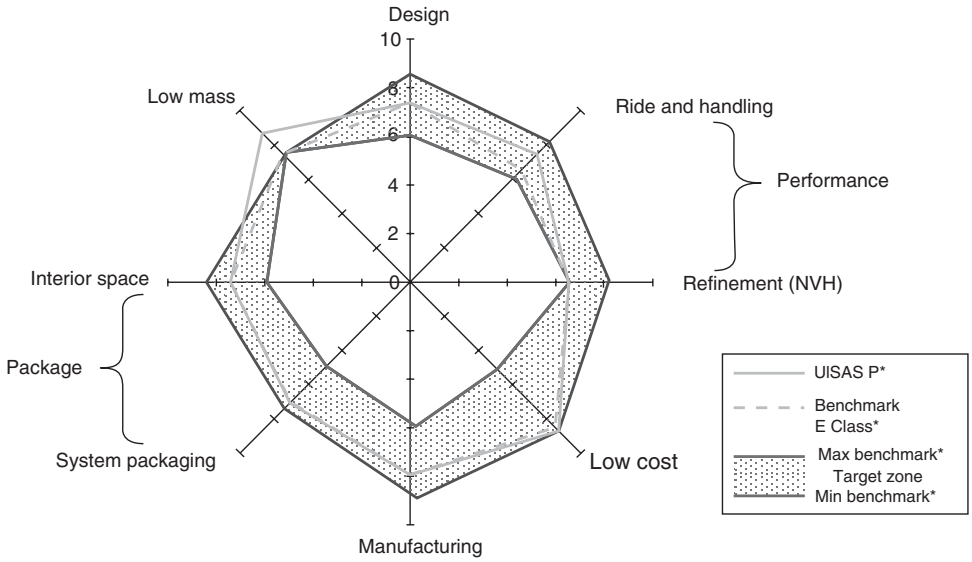


Figure 12. Evaluation of the analyzed strut and links rear axles into the benchmark areas. (Reproduced from World Auto Steel. © World Auto Steel.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

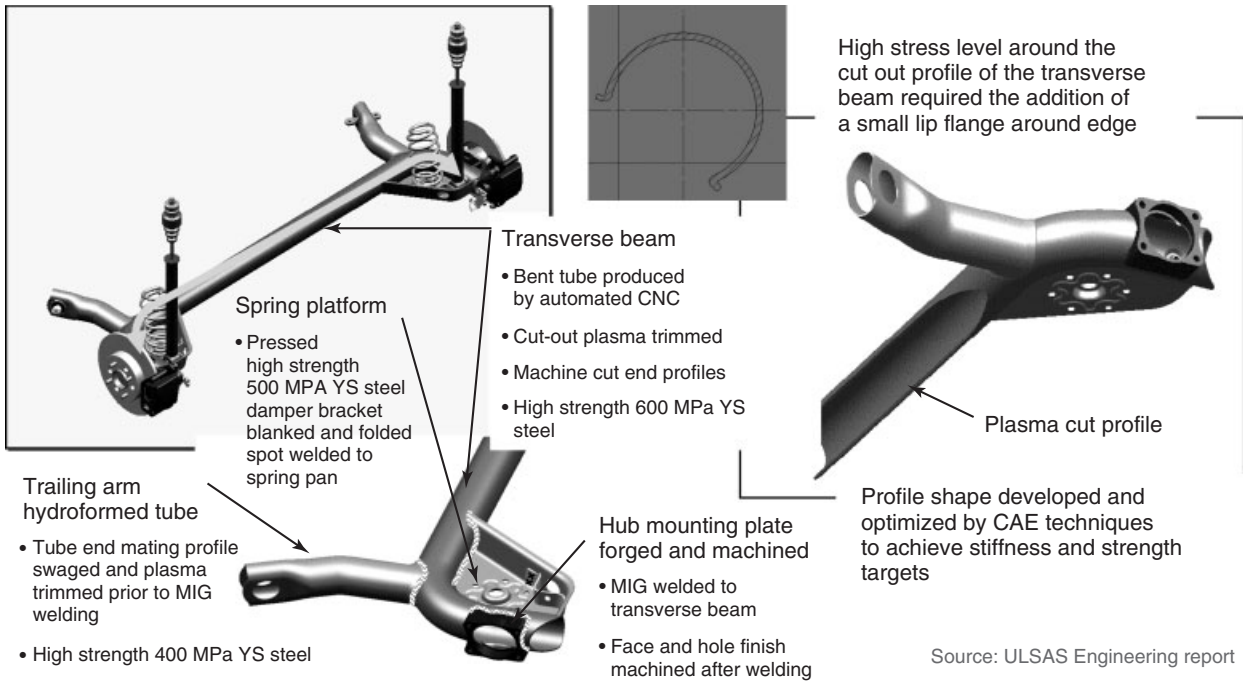


Figure 13. Twist-beam axle as lightweight construction axle. (Reproduced from World Auto Steel. © World Auto Steel.)

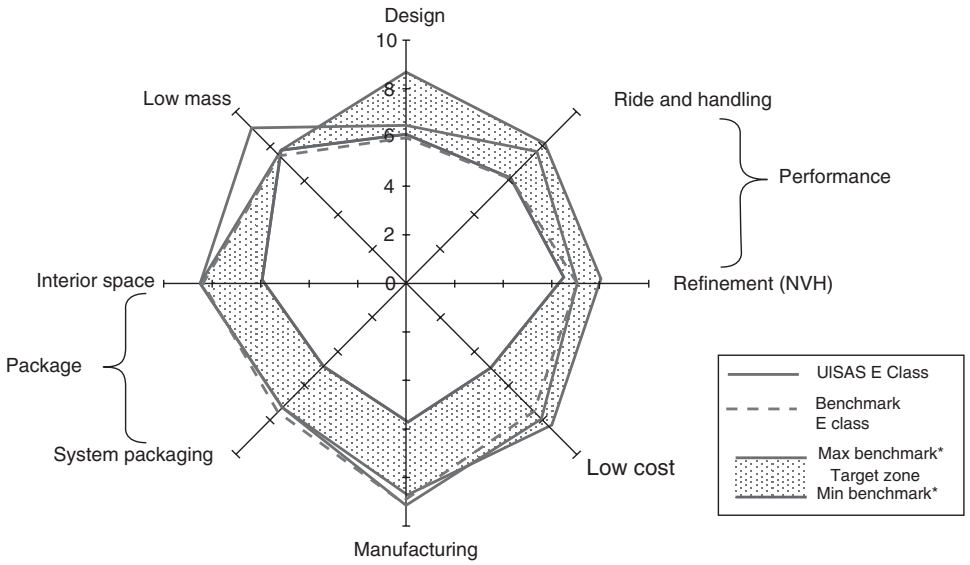


Figure 14. Evaluation of the twist beam axle into the benchmark areas. (Reproduced from World Auto Steel. © World Auto Steel.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

or rather which lengthwise angle is used for welding these longitudinal control arms to the twist-beam tube.

The functionality of the axle is shown in Figure 14.

Especially remarkable are the weight reductions; the cost reductions are less impressive. Design and behavior is even better than the benchmark, whereas production and packaging both meet the benchmark standards.

3.5 Lotus unique rear axle

In addition to the various types of axles considered, the engineers at Lotus also presented their own axle suggestion. This axle can be used as both a driven and a non-driven axle.

A massive body-trailing arm is connected to two tubular wishbones. Thus, the mechanism is statically defined and the spring and damper combination does not have to take any shear forces.

To a certain degree, the different lengths of the wishbones enable an influence on the wheel kinematics (for example, there is no track width alteration in the area of design position of the car). The distinctiveness in the construction of the trailing arms lies in the usage of "Tailor Welded Blanks." Steel plates of different thickness and grade are welded together and are then deep-drawn. Using the strain data available that the specific parts have to withstand enables a strain-related construction. Figure 15 shows these modifications. The trailing arm is welded together from two panels, each of which was manufactured individually. These measures help to achieve the aim of reducing the weight as only as much material as needed is used for production.

These Tailor Welded Blanks can also be used for other axle constructions. By now, there are production techniques that allow rolling metal blanks with varied thickness. The varying stiffness within one part can be achieved by tempering.

The evaluation of the Lotus unique axle by the engineers can be seen in Figure 16. This construction achieved significant reductions in both cost and weight. The packaging demands are evaluated to be adequate; production and assembly is easy. The performance is perceived to be good as well.

The evaluations in Figure 16 show that this axle is positioned in the middle to the upper level of the benchmark. It is therefore advisable to have a closer look at this construction.

4 CONCLUSION

Entering the evaluation figures of all five axles into a spider chart, as done in Figure 17, the advantages of the axles become clearly visible.

This chart exemplifies that the aims to reduce cost and mass were achieved. The multilink rear axle is superior to all others in terms of driving behavior and NVH. Since it is a steel-made axle, it also features a cost advantage in comparison to the axle produced out of aluminum.

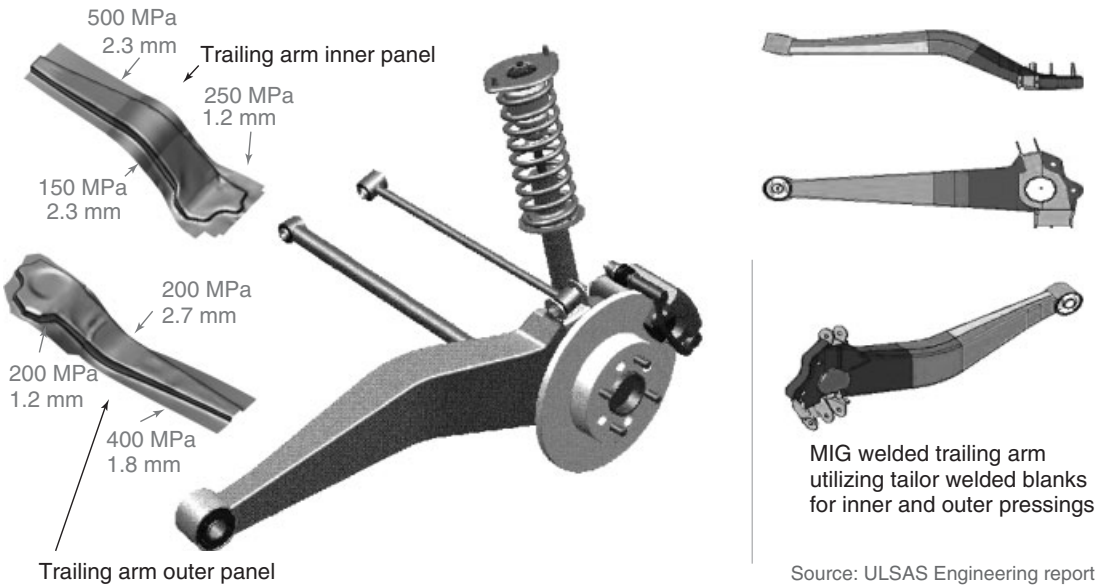


Figure 15. Lotus unique rear axle. (Reproduced from World Auto Steel. © World Auto Steel.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

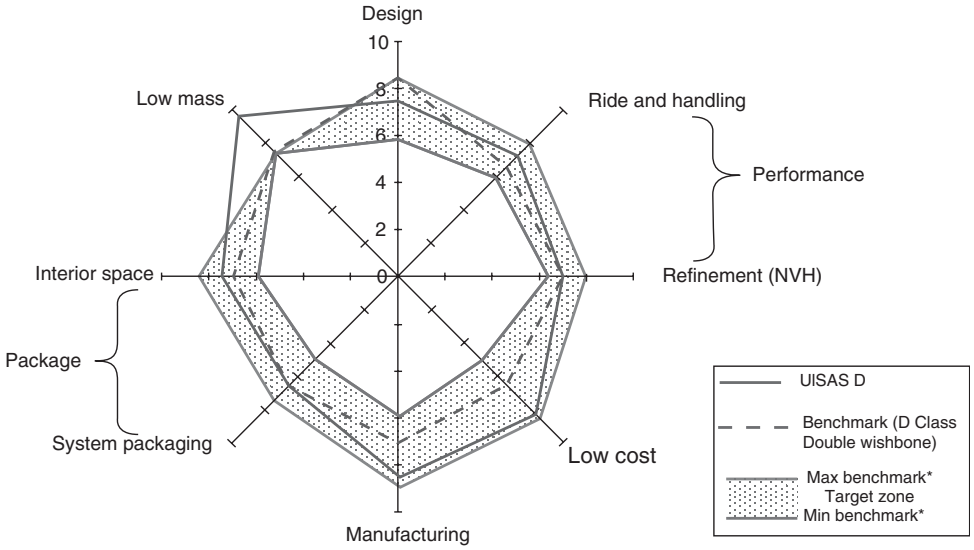


Figure 16. Evaluation of the Lotus unique axle. (Reproduced from World Auto Steel. © World Auto Steel.)

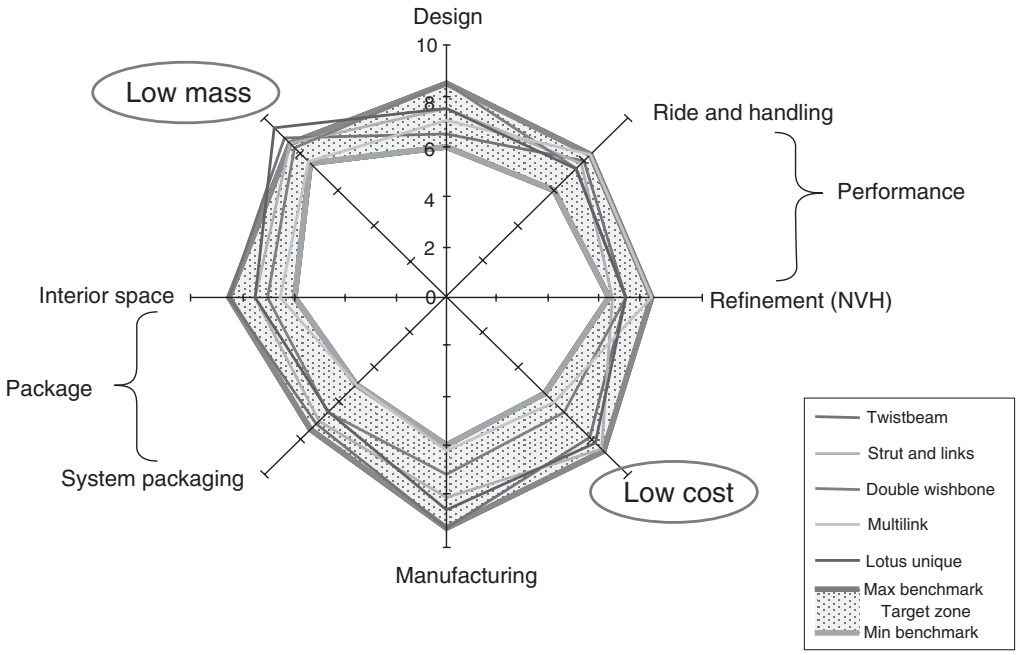


Figure 17. Summary of the different axles. (Reproduced from World Auto Steel. © World Auto Steel.)

Concluding the ULSAS-study, it can be determined that using the suggested steel-axles a mass reduction of up to 34% could be achieved without creating any additional cost. With steel-axles, matching the mass of aluminum-axles there were also no further costs created.

None of the axles compromised the driving performance.

Improved qualities of steel made this progress possible. Since these advancements in steel continued over the last 10 years, a revision of the suggested constructions with contemporary steel qualities and production methods, like flexible rolling and further improved joining technology, is highly suggested. For chassis-developers, there are still

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

some very interesting and innovative ideas left in the ULSAS-study that are still awaiting implementation.

REFERENCES

Chang, M.-Y. (2006) Leichtbau-Vorderachsbauweisen unter Berücksichtigung verschiedener Fahrzeugklassen. Dissertation. RWTH Aachen University, Germany.

Lotus (2000) Ultra Light Steel Auto Suspension, Corus Group.

Matschinsky, W. (1998) Die Radführung der Strassenfahrzeuge Kinematik, Elastokinematik und Konstruktion, Berlin, Germany.

Stahl-Informations-Zentrum (2012) http://www.stahlinfo.de/stahl_im_automobil/ultraleicht_stahlkonzepte/ulsas/ulsas_detailinfo.htm (accessed 12 June).

World Auto Steel (2012) www.worldautosteel.org/project/ulsas (accessed 12 June).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto006
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Active Front Steering for Passenger Cars

Matthias Wiedmann
AUDI AG, Pfaffenhofen, Germany

1 Introduction 1
2 History 1
3 General Functional Principle 1
4 Construction 2
5 Customer Functions 5
6 Steering Stabilization 7
7 Safety 8
8 Summary 10
Endnotes 11
References 11

1 INTRODUCTION

For manyyears now, power steering has been an indispensable feature of our automobiles and has almost completely replaced mechanical steering. One of the difficulties when selecting power steering settings is deciding on the correct steering ratio, so that the driver enjoys a comfortable driving feel and a sense of safety in all driving situations, from parking to high speed driving. When parking, the driver expects low effort at the steering wheel and easy maneuverability at low speeds, without losing a sense of complete control at high speeds.

This classic conflict of objectives can be solved by a superimposed steering system. A controlled additional angle is superimposed on the steering wheel angle chosen by the driver. In this way, both a variable steering ratio and other functions, such as steering stabilization, can be incorporated.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 HISTORY

The first patent applications for superimposed steering systems were filed in the early 1970s (Pilon, Sattavara, and Schechter, 1972). Practical implementation was unsuccessful then because such complex mechatronics were not available. The patent already contained all the components of a modern superimposed steering system, for instance secure mechanical through-drive with provision for steering angle superimposition. In this case, the superimposition itself was obtained from a double planetary gear set with the ring gear turned by an electric motor through a worm gear drive.

This patent and the further development (Karnopp, 1990) reached series production in 2003 (Köhn *et al*., 2002), but a year earlier an alternative concept (Musser, 1955) reached series production in which the angle superimposition was obtained by means of a harmonic drive. The functionality will be described in the following chapters.

3 GENERAL FUNCTIONAL PRINCIPLE

All the systems for active front steering in series production are based on superimposing an additional angle on the steering wheel angle chosen by the driver. Other functional principles can be envisaged, for example displacement of the steering system in relation to the vehicle or altering the length of the tie rods, but these technologies have not so far progressed beyond the patent application stage (Mouri, 1992).

Different to pure steer-by-wire systems, superimposed steering systems have a through mechanical drive to

the vehicle's road wheels. As the steering system now possesses a further degree of freedom, an additional motor angle δ_M can be superimposed on the driver's chosen steering angle δ_H. The total angle present at the steering system input δ_S can then be calculated with the simple formula

$$\delta_S = \delta_H + \delta_M \quad (1)$$

As torque equilibrium is present in the entire system, the motor must withstand all the torques that occur during vehicle operation. This is done by the high gear ratio (1 : 50) between the electric motor and the gearbox housing. In addition, this design satisfies package and weight requirements.

In the case of vehicle maneuvers with no active contribution from the driver by turning the wheel, for example side wind compensation or vehicle stabilization by electronic stability control (ESC), the driver always has to hold up or compensate for the steering torque that arises from the angle superimposition system.

4 CONSTRUCTION

Three systems differing in their gear-set technology and installed position on the vehicle are currently in use for active front steering. In principle, superimposed steering needs the following components:

(a) superimposition gears;
(b) electric motor;
(c) motor angle sensors; and
(d) a lock to restore the through mechanical connection from the steering wheel to the steering gear in the event of a fault or loss of electric power from the system.

Two types of gear drives have been established on the market:

1. the double planetary gear set as used by BMW/ZFLS[1] and
2. the harmonic drive gear set (in series production at Audi/ZFLS and in a slightly modified form at Lexus/JTEKT[2]).

The main gear-set requirement is a high ratio, so that minimum electric motor torque is needed to withstand steering torque, and the motor can therefore be of small size. In addition, the gear set should (if possible) operate without slack to ensure precise initial steering response. Acoustical considerations also have an increasingly important role to play, when determining the gear set's design ratings, as these steering systems are mostly installed on high quality vehicles that are correspondingly quiet.

In view of the high demands, they must satisfy in respect of control quality and performance, the electric motors are in all cases of the permanent magnet synchronous type (brushless electrically commutated DC motors). One of the main factors to be considered when determining the motor rating is high power density, so that the desired torques are reached with a satisfactorily dynamic action and torque ripple is as low as possible, in order to avoid feedback that can be felt at the steering wheel.

With the aid of a rotary position sensor, the actual position of the rotor in the electric motor is measured as a means of determining the actuating signals and phase currents for the individual motor phases. The closed-loop control circuit thus eliminates deviations between the actual and the desired motor positions.

The locking device is a solenoid that, when its power supply is interrupted, locks the system by means of a preloaded spring. This lock must satisfy the most stringent safety requirements, as it is the system's mechanical fail-safe device. Rapid, reliable locking in all environmental conditions and extremely high failure protection are needed. Among the measures adopted to ensure this are redundant springs to actuate the locking pin.

4.1 Types of system

4.1.1 Hollow-shaft harmonic drive

The actuator is notable for its hollow-shaft concept. The permanent magnet synchronous motor is located concentrically around the input shaft. The other components such as the gear set and motor position sensors are also mounted on a joint shaft. The gearbox technology consists of a harmonic drive gear set with a ratio of 1 : 50. This type of gear set needs very little space and provides a high gear ratio with complete freedom from play (without special measures having to be taken) and excellent torsional rigidity. The gear set is also suitable for high torques, which makes it capable of withstanding misuse situations (parking close to the curbstone and turning the steering wheel with full torque) reliably.

The harmonic drive (Figure 1) operates by means of a flexible thin-race ball bearing (a flex-bearing), which is pressed onto the elliptical motor shaft and thus acquires an elliptical shape as well and acts as the wave generator (WG). This elliptical ball bearing changes the shape of the gearwheel connected to the steering wheel (flexible spline, FS). Thanks to its elliptical shape, the FS (which has 100 teeth) is able to mesh with the circular spline (CS) or ring gear, which has 102 teeth. The output shaft leading

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

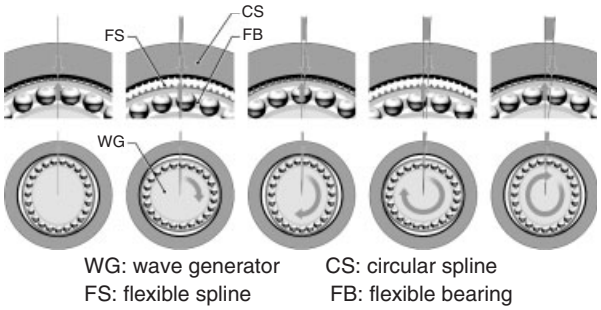


Figure 1. Function of harmonic drive.

to the steering system itself is attached to the circular spline.

When the motor shaft revolves through 360°, the 100 teeth of the FS rotate within the 102 teeth of the circular spline and thus generate a gear ratio of 1 : 50 and therefore a superimposition of 7.2° per revolution of the motor. This is added to the angle reaching the circular spline by way of the input shaft and FS.

When the vehicle is parked or if a system fault occurs, the motor shaft is locked by way of the lock ring and locking pin; this inhibits one of the degrees of freedom of the gear set and ensures that the through mechanical drive to the wheels remains operational.

This compact design can be installed in the upper part of the steering column and therefore in the car's interior, which has advantages above all in meeting the environmental requirements that the system must satisfy (temperature and humidity), whereas this might be more difficult if the system is installed in the engine compartment. At the same time, however, acoustical requirements and also questions of crash protection have to be considered when the actuator is located in the passenger compartment.

The superimposition steering system, which is used in the Audi A6 and which is installed in the steering column, is shown in Figure 2. Explanations are given in Schwarz and Dick (2007), whereas another design can be found in Rothmund *et al*. (2006).

4.1.2 Double planetary gear set

A characteristic of this system is that the actuator is mounted directly on the steering system. The superimposition function is performed by a double planetary gear set (Wallbrecher, Schuster, and Herold, 2008). This position on the steering system behind the torque sensor/hydraulic valve has the advantage that effects such as friction and feedback because of the motor's mass moment of inertia can be compensated for by the power steering assistance, thus reducing possibly unwanted feedback at the steering wheel. As an additional reduction gear stage to the electric motor, which is also an electronically commutated, permanent magnet synchronous motor, the planetary gear set has a worm gear. This provides a very high overall reduction ratio and also a degree of self-locking. This self-locking produces a certain degree of redundancy in the locking action. It enhances the safety concept. In addition, the electric motor is prevented from rotating by a locking device (falling pin) when the system is inactive and in the event of a fault.

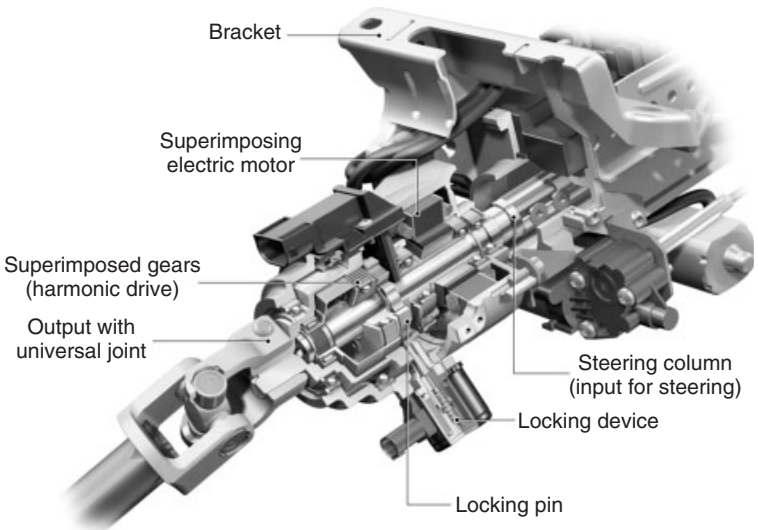


Figure 2. Construction of Audi dynamic steering system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The actual superimposition is obtained from two planetary gear stages with different ratios. The first of these ($i_1 = 15/12$) has an input shaft (ring gear) with 15 teeth and planet wheels with 12 teeth. The second stage ($i_2 = 13/14$) has an output shaft (ring gear) with 14 teeth and planet wheels with 13 teeth. The two sets of planet wheels are linked by a planet wheel carrier (spider) and therefore always rotate at the same speed. If the electric motor causes the planet wheel carrier to perform a single revolution, the ratio of output to input shaft rotation is 1.35 : 1 ($i_{ges} = \frac{15/12}{13/14}$). As manufacturing tolerances mean that the teeth in these gear sets can never be entirely free from play, the individual planet wheels must make a sprung connection in order to prevent play in the steering. This measure, however, can lead to increased friction in the system.

4.1.3 Harmonic drive with rotating housing

In the same way as the system described in Section 4.1.1, this system is based on a harmonic drive assembly, but unlike that system, the housing also turns with the steering. In addition, there are slight differences in the gear set itself, for instance the circular spline is divided into two rings. One ring has 100 teeth and is connected to the steering wheel and the other has 102 teeth and acts as the output to the steering system. The two rings are linked together by the FS, which also has 100 teeth and is coupled to the electric motor. Here too this is a permanent magnet synchronous motor. In this system, when the electric motor performs one revolution, there is a difference of two teeth between the steering wheel and the steering system. The gear ratio is therefore once again 1 : 50. As the housing also rotates, it is not necessary for components such as the motor and the locking device to be constructed with a hollow shaft. To make the assembly as compact as possible radially, the locking device is integrated into the housing and connected by a pivoting lever. This concept can have the disadvantage of requiring more installation space and also that electric power has to be transmitted by a volute spring.

As this system too is installed ahead of the steering system, additional measures are needed to improve the acoustics and minimize unwanted feedback that would be felt at the steering wheel. A ‘Hardy’ flexible-joint disk therefore decouples the system.

4.2 Adaptations to the vehicle

Using Audi Dynamic Steering as an example, it can be seen which additional components on the vehicle need to be adapted in order for the superimposed steering to operate safely and conveniently to its full extent.

4.2.1 Steering system

As superimposition increases steering angle speeds at the steering system, especially when performing stabilizing movements, dynamic steering requirements are higher. In the case of hydraulic systems, for example, pumps with increased displacement are needed to ensure that the necessary volumetric flow is reached. In view of the increased energy consumption and therefore the higher hydraulic fluid temperature, electronically regulated pumps are often used, with an electrically energized valve to regulate the flow volume. A power steering oil cooler of larger capacity may also be needed.

Similar requirements apply to vehicles with electromechanical power steering (EPS). If axle loads and therefore the forces acting on the tie rods are high, the electric motor rating and possibly even the supply voltage may have to be modified.

The demands imposed by superimposed steering must also be considered when choosing the steering ratio. In combination with superimposed steering, it is best to adopt the most direct possible basic gear ratio (for example 1 : 13.5), so that assistance from the actuator, especially when parking or driving slowly, is kept low (1 : 12). This improves particularly the acoustics and the amount of feedback felt by the driver in this critical range. However, this approach is subject to limits, as the driver must be capable of supplying the necessary level of passive fail-safe effort in the event of a defect. This applies to the overall steering ratio and also to the sudden jump in the effective ratio if a fault happens.

4.2.2 Acoustic requirements

In particular when parking and driving slowly, the superimposition must not cause any noise that could be regarded as unpleasant. However, a mechatronic system with gears and an electric motor can never be entirely noiseless, and additional measures must therefore be taken. When deciding-g on these measures, a distinction has to be made between structure-borne noise and airborne noise. Airborne noise can be reduced by suitable encapsulation. To prevent the propagation of structure-borne noise, damping elements must be installed. One that has already been mentioned is the “Hardy” flexible disk between the actuator and the steering wheel, but damping elements can also be installed between the actuator and the steering column if the system uses steering-column actuators. A disadvantage of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

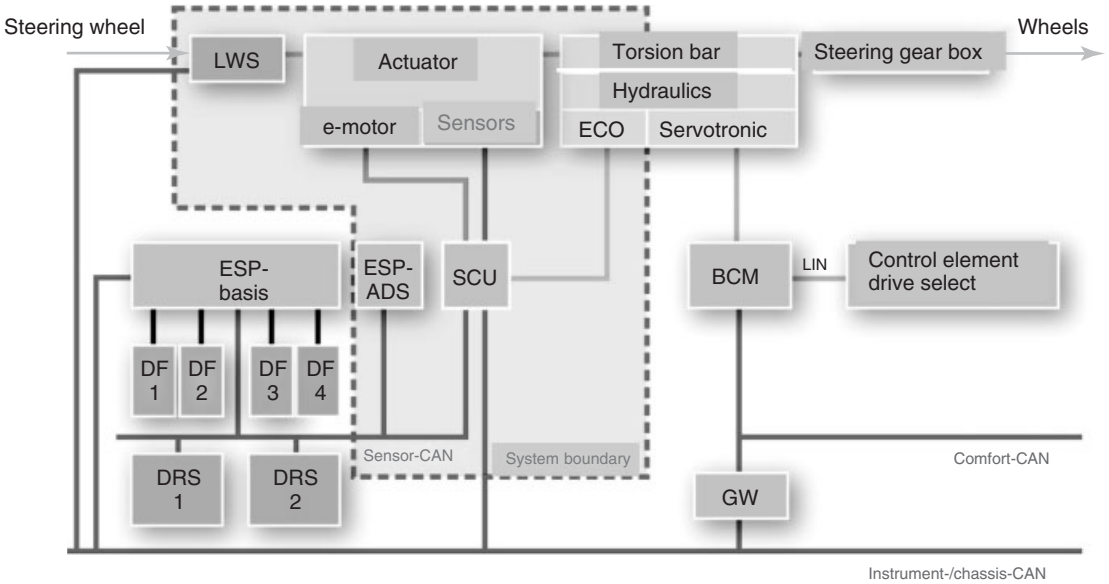


Figure 3. In-car system networking.

these additional measures is that they reduce rigidity. This influences steering response, for instance, and the ability to achieve the required eigenfrequency in the steering column.

It is therefore advisable for the acoustic requirements not to be disregarded when the concept is chosen. Major factors in the selection process are the type of gear set chosen and the gear ratio, though classic conflicts of objective arise here, as the gear ratio also affects the size of the motor and the maximum torque that can be withstood.

4.2.3 Networking adaptations

As superimposed steering effectively communicates with all the vehicle's systems that need steering angle information, networking is accordingly complex. Additional systems supply the input values for the system, for example a driving program switch for the selection of different vehicle dynamics characteristics. If in particular there is also a vehicle stabilization function, this will call for intensive data exchange with the ESC. The ESC itself must satisfy more stringent demands, for example yaw rate-sensing redundancy. If EPS is installed, there will also be a communication with this system, so that further functions such as steering performance limiting can be realized.

Figure 3 shows this signal networking. Networking in such systems is also dealt with in Knoop, Flehming, and Hauler (2008).

5 CUSTOMER FUNCTIONS

5.1 Variable steering ratio

The variable steering ratio is the basic function of a superimposed steering system. This function makes it possible to avoid the compromise that is always present in conventional steering systems between stability at high speeds and avoidance of excessive steering-wheel effort at low speeds. This can be achieved by way of the steering angle (in a similar manner to a variable-ratio steering rack, but with a greater spread of ratios possible) or by way of road speed. The steering ratio can be varied over a wide range between a more direct ratio (for optimal vehicle agility and maneuverability) and a less direct ratio (for high vehicle stability).

The various driving situations can be approximately classified under three headings:

1. Low speed driving (city traffic/parking),
2. Medium speed driving (regular out-of-town roads), and
3. High speed driving (freeways, etc.).

When selecting settings in these three categories, it is important to maintain a harmonious steering ratio characteristic so that the driver does not encounter any unexpected vehicle reactions.

Figure 4 shows the AUDI realized gear ratio function for different vehicle dynamics variants in dependence of the driving velocity.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

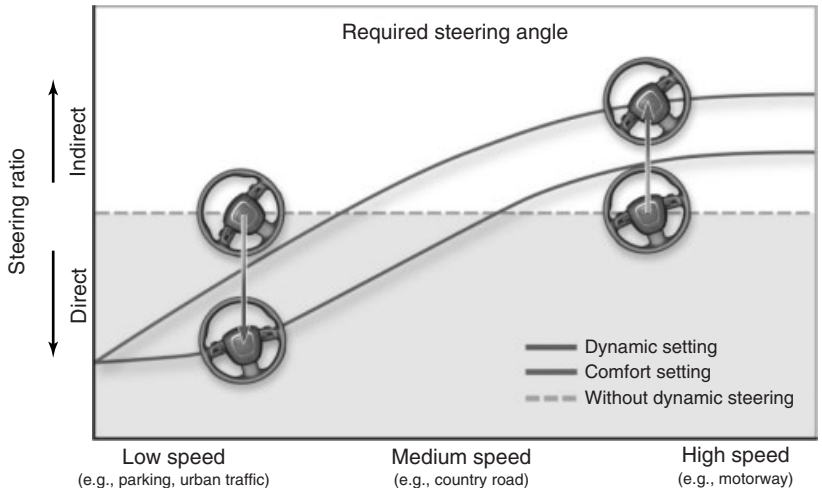


Figure 4. Steering ratio characteristic.

5.1.1 Low speeds

At low speeds, the aim must be to keep driver effort at the steering wheel to a minimum. The steering is normally rated to permit the driver to control the vehicle's direction without having to reposition his or her hands on the steering wheel. Possible factors limiting the degree of directness could be the high degree of superimposition needed, the associated noise, and the driver's ability to adjust to what is felt at the steering wheel. When parking, a superimposed steering system permits the effort required at the steering wheel to be reduced to between one and a half and two turns from lock to lock.

5.1.2 Medium speeds

In this speed range, the vehicle's agility and ease of handling are the decisive criteria. Agility can be described by the increase in the vehicle's yaw rate. The aim in AUDI is to choose a more direct steering ratio than the standard value, one that builds up yaw rate gain plotted against the vehicle's road speed more rapidly (Figure 5). A deeper description can be found in ATZ-Automobiltechnische Zeitschrift (2008).

5.1.3 Road behavior at high speeds

At high speed, a less direct steering ratio is chosen so that the vehicle can be driven in a relaxed manner at high speed. It reacts calmly and less nervously to steering wheel movements. In the BMW superimposed steering even countersteering is installed. At high speed, the drivers steering input is reduced by the electric motor to make the car more stable.

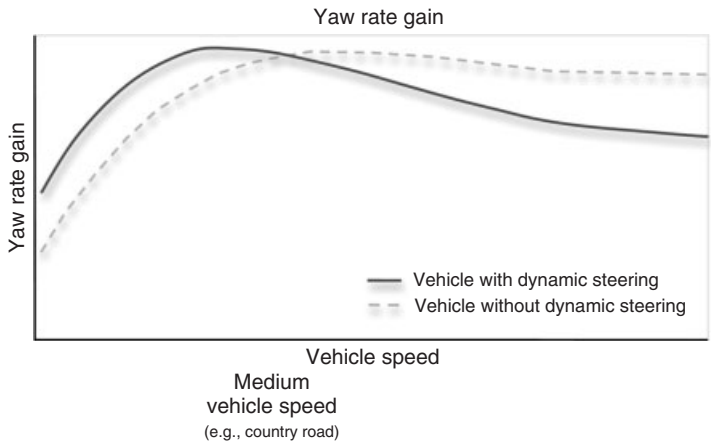


Figure 5. Yaw rate gain plotted against road speed.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5.1.4 Change between characteristics

As the steering ratio can be tuned, the characteristic of the steering gear ratio can be coupled to a driving program switch if this is installed in the vehicle. This gives the driver another means of influencing the steering in addition to the manual torque exerted at the steering wheel. Furthermore, with the aid of a superimposed steering system, it is relatively simple to provide entirely different steering and therefore vehicle characteristics, for example, on a sport model within a model line. This is difficult to achieve with conventional steering systems, as a separate version has to be developed for a limited number of vehicles. The arising costs are prohibiting this. The superimposed steering makes this differentiation possible.

6 STEERING STABILIZATION

As a means of moving the road wheels independently of the steering wheel is now available, it is possible to stabilize the vehicle in critical situations by turning the steering. Until now, vehicle stabilization was only possible with ESC, by applying individual wheel brakes, but as any form of brake application naturally slows the vehicle down, the subjective sense of sportiness is reduced.

At high road speeds, the superimposed steering system demonstrates its advantages most clearly in the form of reaction times that are distinctly faster than any brake application and they are much more comfortable. Steering correction stabilizes the vehicle, whereas the brake application reduces its dynamic performance in an equivalent situation.

At slower speeds or in less critical situations, it may in certain circumstances even be possible to dispense entirely with brake applications. In view of this, a vehicle with steering stabilization maintains its progress more smoothly than one that is only stabilized by means of the brakes. This is especially obvious when friction in the form of tire grip is low. If brake applications are still necessary in such situations, they can be less powerful or will only be needed at a later stage in the critical situation.

Stabilization is effective to a differing degree according to steering performance, which is limited by the actuator. A parameter for this is the steering angle gradient for the stabilizing action. However, the maximum gradient that can be adjusted not only depends on the actuator's performance but also on how it is networked on the vehicle. It is essential to ensure that a superimposition angle that has been called for incorrectly does not lead to a critical safety situation. The latency time of the bus system, for example, also has a decisive part to play.

The decision as to which form of intervention will bring the desired result—brake application, steering stabilization, or a combination of both—is also an important factor for stabilization quality. Comprehensive evaluations of stabilization procedures are reported in Baumgarten *et al*. (2004) and Holle (2003).

6.1 Oversteer

If a vehicle oversteers, an experienced driver will sense its reaction and apply an opposite steering input at the steering wheel. To make this situation easier for less experienced drivers as well, the superimposed steering can apply the opposite steering actively if oversteer occurs, and in this way achieve the optimal front-wheel steering angle. Thus, undesirably high vehicle yaw reaction can be reduced in effect or entirely canceled out. At high vehicle speeds, in particular, the vehicle's fast reaction to the steering angle allows the brake application to take place later and more harmoniously.

Figure 6 demonstrates this behavior of a car. In case of Figure 6a, the brakes are applied only. If the active front wheel steering is added to the stabilization procedure of the car, the brake intervention can be reduced. This is shown in Figure 6b. In addition, it becomes obvious that the total steering effort for the driver is reduced as soon as the superimposed steering is applied.

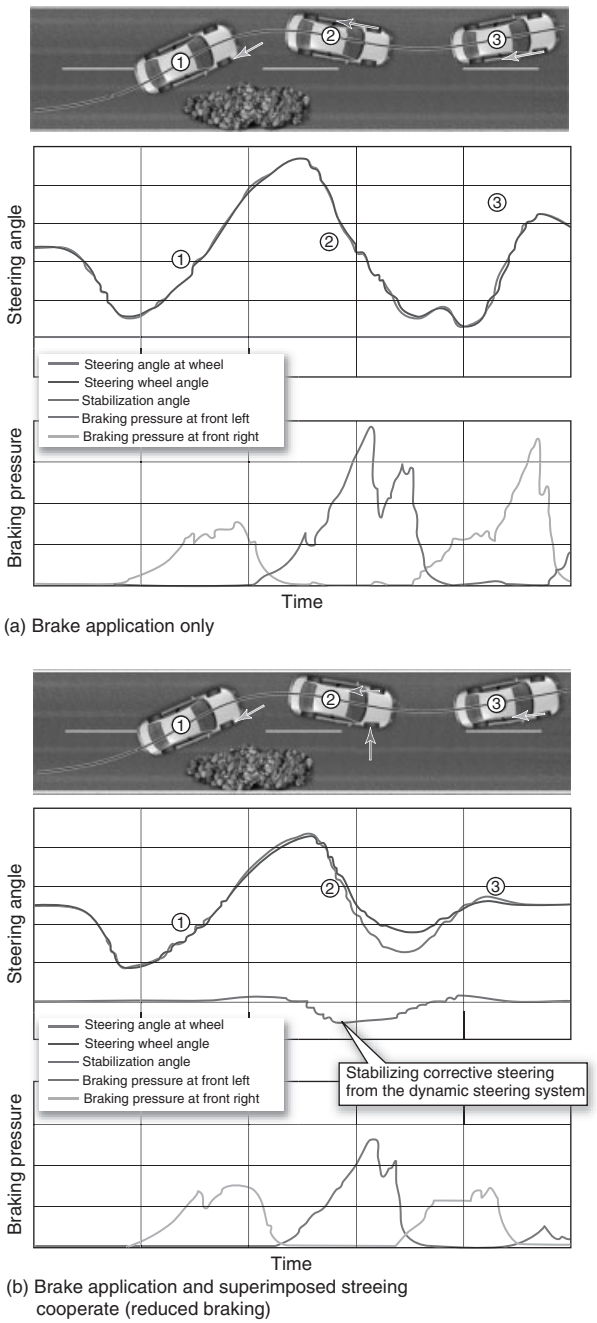
6.2 Understeer

A vehicle that is understeering is driving more straight ahead at the front wheels, because the maximum force that the tire contact patches on the front axle can transmit has been exceeded (Figure 7). This limits the effect of any action taken through the front steering. The aim must therefore be to provide the driver with access to the maximum adhesion point for as long as possible. When action is taken in this situation, the steering ratio is made less direct so that the steering does not pass the point of maximum friction too quickly.

6.3 μ-Split

μ-Split is the term used to describe a situation in which the amount of grip on one side of the vehicle is noticeably different from the other side. This occurs in the autumn or winter, for example, when one side of the road is covered with wet leaves and the other is dry. When the brakes are applied in this situation, the difference in effective braking force generates a yaw moment that causes the vehicle to turn toward the side with the higher grip. To

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) Brake application only

(b) Brake application and superimposed streeing cooperate (reduced braking)

Figure 6. (a,b) Oversteer intervention.

keep the vehicle moving in the chosen direction, the driver has to steer toward where tire grip is lower. To give the driver sufficient time to react, braking pressure is built up relatively slowly.

With superimposed steering, it is possible to select the correction angle at the steering automatically. All the driver has to do is turn the steering wheel in the correct direction, and as the superimposed steering system reacts quickly,

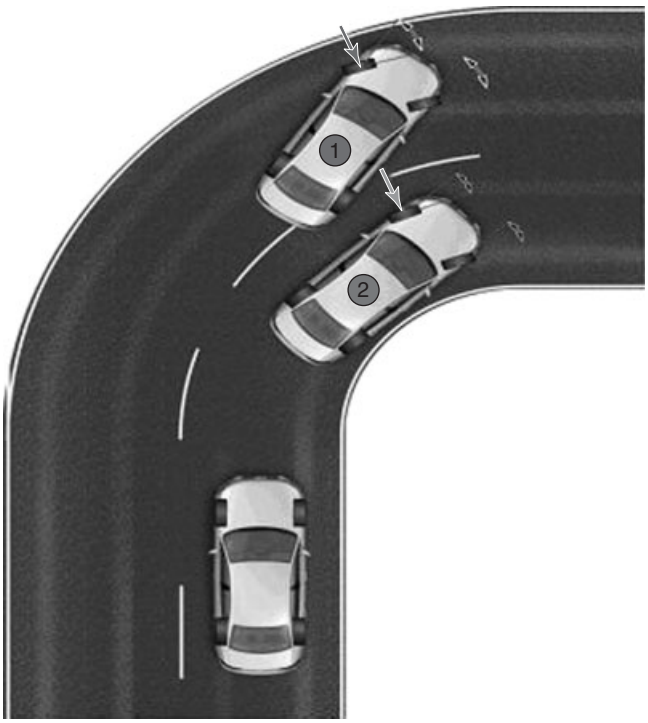


Figure 7. Understeer intervention.

braking pressure can be built up more rapidly as well, so that the braking distance can be shortened. The yaw movement will be reduced or even avoided. Figure 8 gives an insight into the brake forces and the steering angles. The brake pressure on the high friction side can be higher than on the low μ side. The front wheel on the high μ side produces side forces to counteract to the yaw moment, which is generated by the different brake forces. The superimposed steering system generates these side forces.

7 SAFETY

Like other mechatronic steering systems, the superimposed steering system is a product with safety relevance. The primary development objective must therefore be to avoid potentially critical situations. These can include

- Preventing reversible or irreversible errors that could be caused by the control unit, the electric motor, or the motor position sensor.
- Monitoring externally computed stabilizing interventions and initiating suitable measures to prevent maximum permissible positioning errors from being exceeded.
- Ensuring that in the event of an error the maximum tolerable gear ratio jump is not exceeded.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

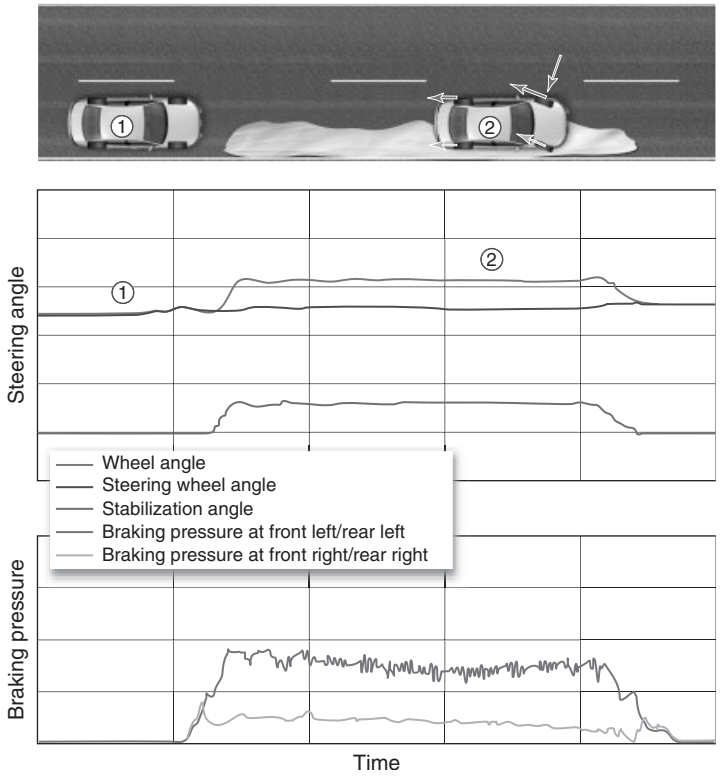


Figure 8. μ-Split braking with and without superimposed steering.

- Preventing any situation in which steering movement is uncontrolled.

Figure 9 shows the three-level steering control unit (SCU) safety concept adopted for the Audi dynamic steering system. All the functionally necessary software modules are integrated into layer 1, including signal plausibility checks and error strategy. All critical paths that could lead to a malfunction are calculated in a diverse manner in layer 2. This ensures that system errors (for example, programming errors) or sporadic RAM defects cannot result in a malfunction. Diverse means that the computer hardware and the used software in layer 2 should be different from layer 1. If possible, different persons should do even the engineering. Layer 3 ensures the correct program sequence, for example, and checks by way of question–answer communication that the set of commands has been carried out correctly (watchdog function).

The difficulty that arises with computing in a diverse manner is achieving the same result in the diverse path with other algorithms as in the main path. Two principal measures enable this:

1. Under error-free conditions, it is possible to achieve the same results with the main function and the diverse function with no relevant time delay. The reason is that the variable steering ratio functions follow feed forward control rather than feedback control.
2. For position control, including sensor evaluation, both paths are monitored in layer 2 by read-back and checking of the motor position signal in relation to its desired angle, as shown in Figure 10.

To ensure high availability, AUDI follows a stepwise degrading of system functions, depending on the error that has occurred:

- Selection of a constant steering ratio (as on vehicles with standard steering) if vehicle velocity information is lacking.
- Inhibition of external stabilizing interventions if reduced performance is to be anticipated, for example, because of vehicle power supply fluctuations.
- If an error is suspected, deactivation of the system, just when the steering wheel angle is zero, in order to prevent the steering wheel from having a constant angle deviation. Other companies are using this constant steering wheel error as a hint of the failed system. The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

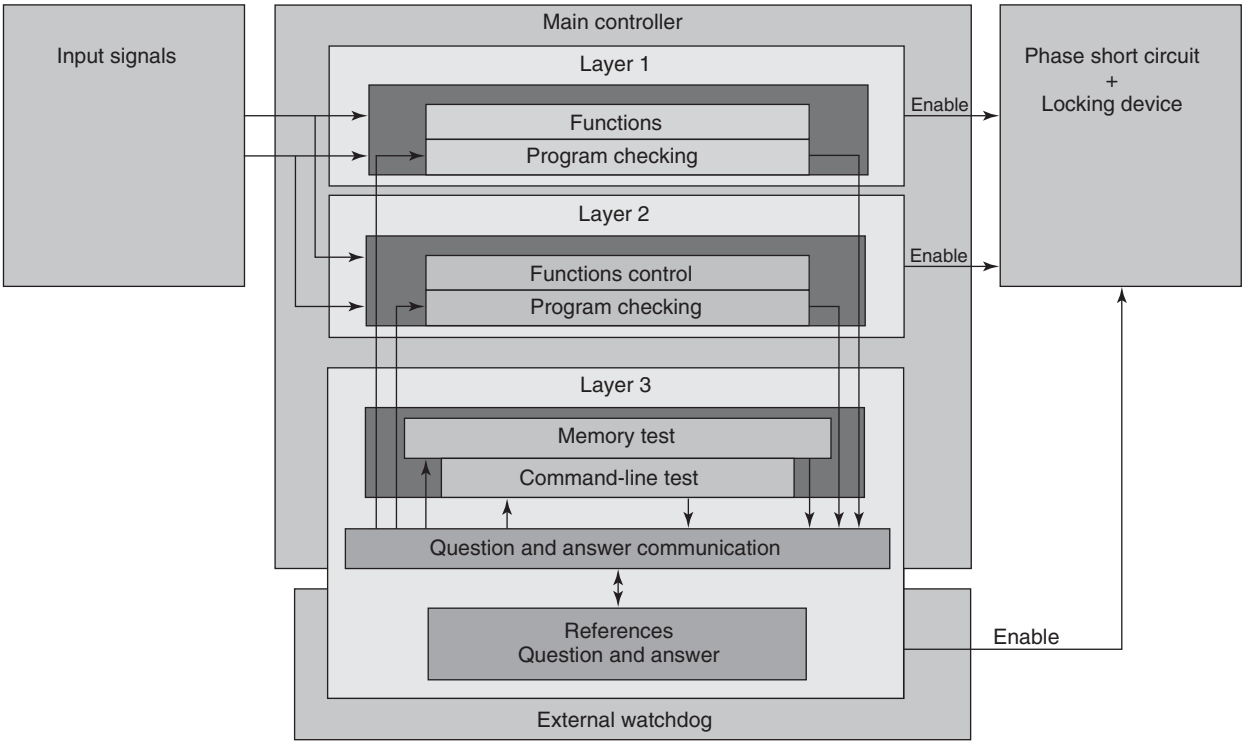


Figure 9. Monitoring levels.

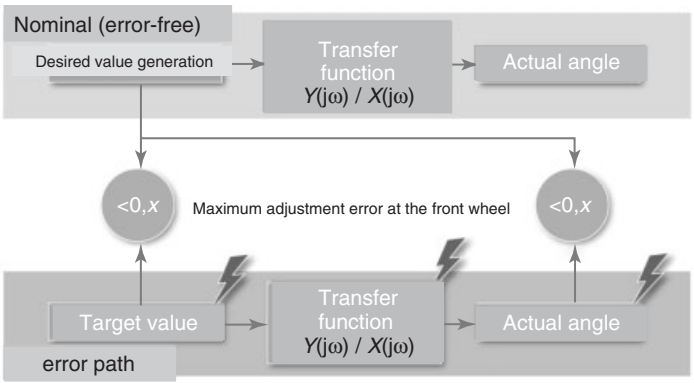


Figure 10. Position control error path.

customer should be advised to go to a workshop and to take the car for repair.

- Immediate, complete deactivation in the event of severe faults.

In the AUDI system, the integrated integral index sensor makes it possible to restart the system without having to take the car to a workshop even if a severe error has occurred, for instance, the loss of electric power. For this purpose, the motor angle is reset in an initializing phase.

The SCU output signals must also satisfy the relevant safety criteria, as other vehicle systems with safety relevance, for example, ESC, react to them. In Jakobi (2011), there are intensive discussions about fallback behaviors of the steering system.

8 SUMMARY

One of the latest mechatronic systems in cars is Active Front Steering. This means that in addition to the driver's

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

steering input an electric motor can steer the front wheels of a car. Driver and the electronical system are normally working in cooperation. The amount of automated steering depends on the car speed, a selected control program, and the driver's steering input.

At present, there are two technical solutions used, the double planetary gear set and the harmonic drive gear set. The systems are described and the different positioning in the steering shaft is mentioned. In addition, the requirements in acoustics and in the oil supply are mentioned. The Active Front Steering is connected to the safety network of the car.

For normal driving, it is possible to use different steering ratio characteristics. The individual driver's wish can be fulfilled. Some examples are explained in this contribution and the basic support for vehicle dynamics is mentioned.

Special requirements for such a system are demanded by the system's safety. The SCU and the used sensors must be diverse. This means that hardware and software must function in different ways. This makes the system more complex. On the other hand, there is so much benefit for the drivers that this effort is justified.

ENDNOTES

1. ZFLS: ZF Lenksysteme GmbH, a Joint Venture of Robert Bosch GmbH and ZF Friedrichshafen AG.
2. JTEKT Corporation Japan (Toyota is one of the shareholders of this company).

REFERENCES

ATZ-Automobiltechnische Zeitschrift (2008) *Dynamiklenkung im AUDI Q5, ATZ Sonderheft Ausgabe Nr.: 2008-02*, Springer Automotive Media, Wiesbaden.

Baumgarten, G., Hofmann, M, Lohninger, R, *et al*. (2004) Die Entwicklung der Stabilisierungsfunktion für die Aktivlenkung, ATZ 106 Heft 9.

Holle, M. (2003) Fahrdynamikoptimierung und Lenkmomentenrückwirkung durch, Überlagerungslenkung, PhD Thesis, RWTH Aachen University, Germany.

Jakobi, F.R. (2011) *Eigensichere Überlagerungslenkung mit elektrischer und mechanischer Rückfallebene, Schriftenreihe des Instituts für Verbrennungsmotoren und Kraftfahrwesen der Universität Stuttgart, Band 57*, Expert Verlag.

Karnopp, D. (1990) Motorbetriebenes Servolenksystem, Deutsches Patent DE 4031316 C2, filed 1990, Robert Bosch GmbH.

Knoop, M., Flehming, F., and Hauler, F. (2008) Improvement of vehicle dynamics by networking of ESP with active steering and torque vectoring. 8th Stuttgart International Symposium, Stuttgart.

Köhn, Ph. *et al*. (2002) Die Aktivlenkung—Das neue fahrdynamische Lenksystem von BMW Aachener Kolloquium Fahrzeug- und Motorentechnik, Aachen.

Mouri, T. (1992) Variable Ratio Steering Gear for a Motor Vehicle, British Patent GB 22 59 062 A, Fuji Yokogyo Kabushiki Kaisha, filed 1992.

Musser, W. (1955) Strain Wave Gearing, US Patent 2906143, United Shoe Machinery Corporation, Felmington N.J., filed.

Pilon, H.M., Sattavara, S.W., and Schechter M.M. (1972) Power Steering Gear Actuator, US Patent 383170. Ford Motor Company, filed.

Rothmund, M., Schwarzhaupt, A., Sulzmann, A., *et al*. (2006) Lenksystem mit Stelleinrichtung und Harmonic-Drive-Getriebe, Deutsches Patent DE 10 2006 055 774 A1, Daimler AG, filed.

Schwarz, R., Dick, W. (2007) Die neue AUDI Dynamiklenkung, Tagung Reifen, Fahrwerk, Fahrbahn, 11. Internationale Tagung Hannover, 23/24. 10. 2007, Hannover.

Wallbrecher, M., Schuster, M., and Herold, P. (2008) Das neue Lenksystem von BMW—Die Integral Aktivlenkung. Eine Synthese aus Agilität und Souveränität. 17th Aachener Kolloquium Fahrzeug- und Motorentechnik, Aachen.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto007
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

New Electrical Power Steering Systems

Mathias Würges
NSK Deutschland GmbH, Ratingen, Germany

1 Introduction 1
2 Electric-Power-Assisted Steering 1
3 System Components 5
4 Steering Functions 11
5 Electric Motors for EPAS Systems 13
6 Functional Safety in EPAS Systems 14
7 Summary 15
Further Reading 16

1 INTRODUCTION

Following the introduction of the first steering systems with an electromechanical servo unit (electric-power-assisted steering, EPAS) at the end of the 1980s, they have become more and more widespread in recent years. This development is driven by the necessity to economize on energy and thus reduce CO_2 emissions. Depending on vehicle type and driving style, EPAS systems contribute to a reduction in fuel consumption of between 0.3 and 0.5 L/100 km.

The global EPAS market in 2010 already totaled 26 million units, and it is expected to almost double by 2015. This trend stems from the rapidly increasing technological development of electrical and electronic components and safety concepts, which can be used in small-range to top-of-the-range vehicles, and from the expansion of EPAS technologies in high growth markets such as China and Brazil.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 ELECTRIC-POWER-ASSISTED STEERING

Nowadays, there are different EPAS systems on the market, which are used according to the vehicles' boundary conditions and the vehicle manufacturers' technological philosophy. Significant technical factors in the selection of suitable systems are the necessary steering rack force and the steering ratio, that is, the ratio between steering wheel angle and the steering rack stroke or the corresponding front wheel steer angle (Figure 1).

In spite of the different designs of the various EPAS systems, they all share the most important functional requirements:

- Safe operation in all driving situations and a very high level of availability
- Highly dynamic response characteristics in the most varied driving situations
- A sufficient level of steering assist for the driver in the case of intensive actuation forces, for example, parking maneuvers
- Minimal noise during all steering maneuvers. As for this vehicle functions, acoustic feedback is not desirable
- High quality steering characteristics in line with the philosophy of the vehicle brand
- More and more steering functions are being integrated into modern EPAS system, which improve safety or comfort for the driver and can be correspondingly marketed by the vehicle manufacturers.

However, the introduction of EPAS systems—and with it the substitution of hydraulic steering systems—was primarily driven by the reduction in fuel consumption. As it is crucial for the steering assist to respond highly dynamically in every driving situation, the oil pressure

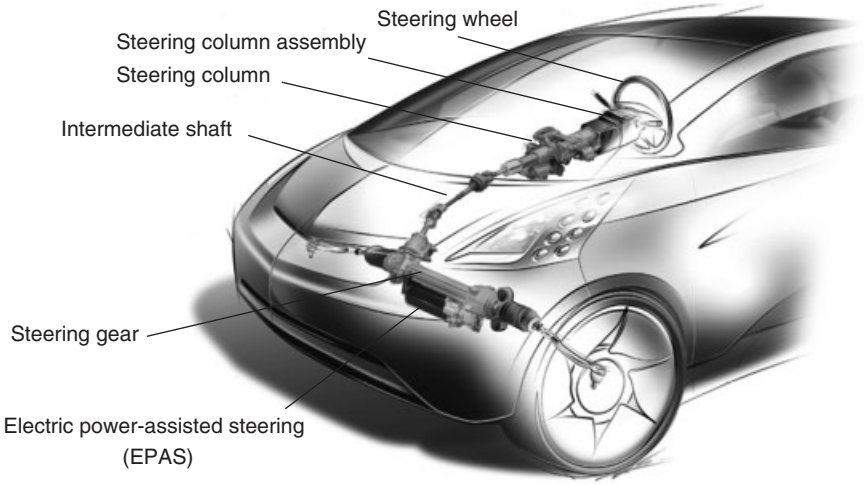


Figure 1. Electric-power-assisted steering (EPAS) system.

in hydraulic systems must be maintained at all times. This results in power dissipation through the continuous operation of the pump and therefore to a high demand for energy. Meanwhile, EPAS systems convert electrical energy from the vehicle power supply, drawing only the amount needed for the length of time required by the respective steering requirement. A further advantage of EPAS is its simple installation and dismantling, because the EPAS actuator system can be connected simply by attaching the power and signal plug connectors. There is no time-consuming handling of hydraulic fluids.

Further advantages of electromechanical steering are its good control capability, the incorporation of the EPAS control unit into the vehicle communication network, the system's highly dynamic properties, and its most temperature-independent characteristics. These properties are also used to introduce new (steering) functions, which increase safety and comfort levels when steering.

The exacting safety demands made of a steering system require the development and validation of extensive safety functions, as well as the application of standardized quality processes. The universal IEC 61508 safety standards for electronic safety-related systems originally applied to the application of safety concepts. However, the new safety standard ISO 26262 specifically for safety-related electrical and electronic systems in road vehicles, which came into force in November 2011, will be the standard for future developments.

2.1 General functions of EPAS systems

The driver applies a manual steering torque to the steering wheel. This is detected by a torque sensor and is transmitted as an analog or digital signal to the electronic control unit (ECU) of the steering system. The ECU calculates the necessary assist torque, considering the driving situation. The drive status is determined using system-internal and system-external information, such as the vehicle speed. The ECU controls the electric motor correspondingly via the power electronics. The steering torque accumulated from the manual torque and assist torque is converted into an actuation force by a pinion on the steering rack and transmitted to the wheel unit via the tie rod (Figure 2).

To minimize loading of the vehicle electrics—and thus the demand for energy—the steering system must operate

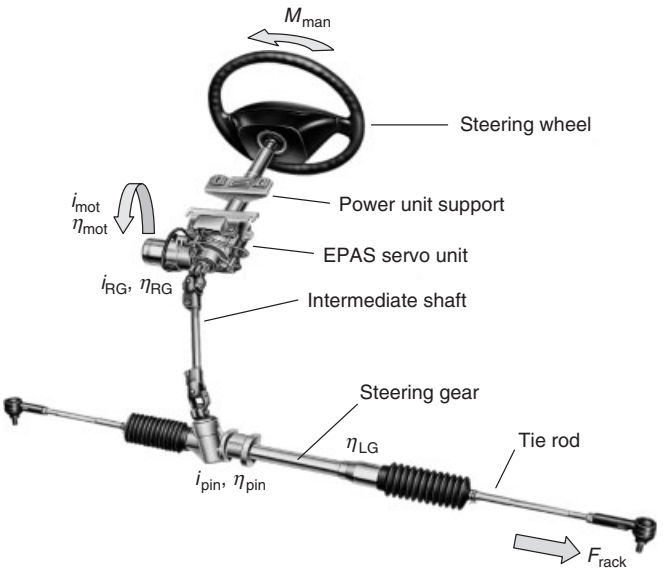


Figure 2. The equilibrium of forces in steering column EPAS.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

as efficiently as possible. The system design makes a significant contribution to this with low mechanical friction losses as well as efficient motors and transmission trains. The following equation shows the definitive mechanical influence variables in an example of a steering column electric-power-assisted steering (C EPAS) system

$$F_{\text{rack}} = i_{\text{pin}} \times \eta_{\text{pin}} \times [M_{\text{man}} + (\eta_{\text{mot}} \times M_{\text{mot}}) \times i_{\text{RG}} \times \eta_{\text{RG}}] \times \eta_{\text{LG}}$$

F_{rack}	Actuation force on the tie rod
i_{pin}	Gear ratio of the pinion on the steering column to the steering rack of the steering gear
η_{pin}	Efficiency of the rack-and-pinion gear
M_{man}	Manual steering torque applied by the driver
η_{mot}	Mechanical efficiency of the motor
M_{mot}	Mechanical torque of the motor
i_{RG}	Gear ratio of the reduction gear
η_{RG}	Efficiency of the reduction gear
η_{LG}	Efficiency of the steering gear

This equilibrium of forces clearly illustrates that the manual steering torque is influenced directly by changes to the controlling torque of the motor. Speed-dependent regressive steering assist draws on this principle. At low vehicle speeds, such as when parking, the EPAS motor is activated with a relatively high current (assist current), so that smooth steering is achieved. At high speeds, only small steering movements are generally carried out and the driver requires precise tactile feedback; therefore, a smaller level of steering assist is provided by the EPAS system in this case (Figure 3).

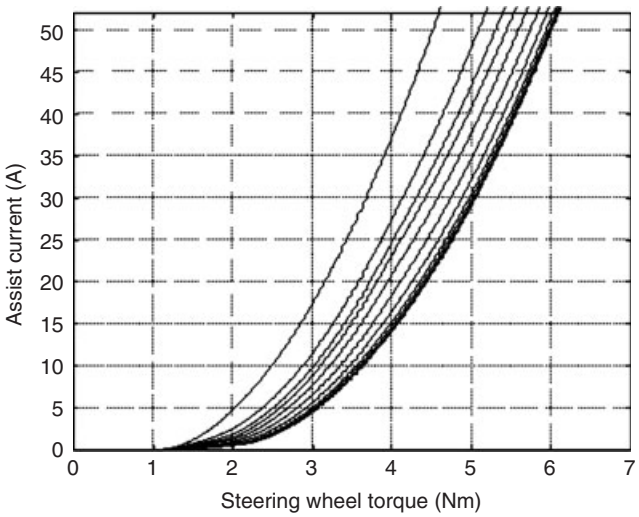


Figure 3. Regressive steering support.

Highly dynamic drives are used in EPAS systems, so that the necessary assist torque can also be provided when completing fast steering movements, for example, when parking or conducting quick evasion maneuvers. Generally, EPAS systems are capable of providing full assist torque, even at angular velocities of 360°/s.

2.2 EPAS technologies

2.2.1 Steering column EPAS

Steering C EPAS is used primarily in small and compact vehicles. Thanks to further enhancements to the technology used for the electric motor, the (power) electronics, the component stiffness, the reduction gears, and the control software; however, modern steering C EPAS is also used in all classes of vehicle, small-range to top-of-the-range models.

In case of steering C EPAS, the motor/control unit and the torque sensors are integrated into the steering column. By means of a reduction gear, the motor transmits a supporting torque to the steering column. As the technology is incorporated into the vehicle interior, this type of EPAS can be configured for relatively moderate ambient conditions.

As the servo unit is integrated directly into the steering column, modern EPAS systems must be small enough to adapt flexibly to the vehicle installation space. Owing to its close proximity to the driver, the servo unit must be very quiet at all steering speeds. The high demands made of the steering feeling are satisfied by the stiff mechanics of the steering column and the intermediate shaft (I-shift), as well as by corresponding control algorithms (Figure 4).

A look at the product portfolio of the globally active Japanese steering manufacturer NSK clearly shows the wide

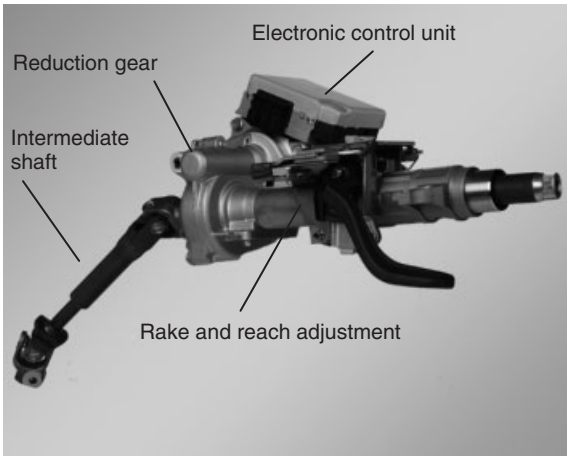


Figure 4. Steering column EPAS. (Reproduced by permission of NSK Deutschland GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

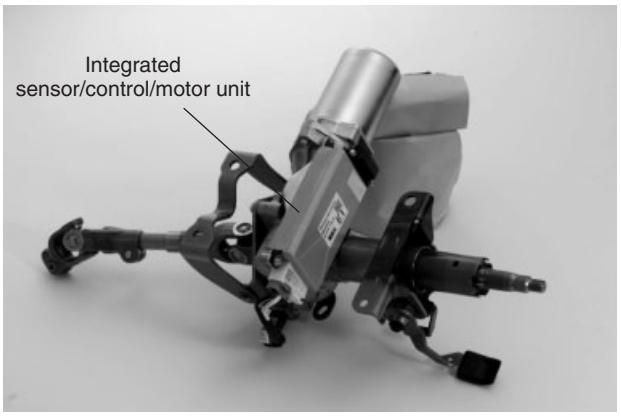


Figure 5. Small package Toyota iQ steering column EPAS. (Reproduced by permission of NSK Deutschland GmbH.)

range of possible applications for steering C EPAS systems. Figure 5 shows the smallest system currently available on the market, which is used in the Toyota iQ. With its steering system for the new Toyota Sienna, NSK sets the standard worldwide for the performance and power density of C EPAS systems. This system provides a steering assist of 12.5 kN rack load and is thus a classic example of an alternative to hydraulic steering assist that can be used for various vehicle classes, from small vehicles to cars with over 3-L engines. At 8.5 kN, the highest performance C EPAS on the European market was also developed by NSK. It is used in models including the Renault Mégane Scénic.

2.2.2 *Single-pinion EPAS*

In case of the single-pinion EPAS, the servo unit is positioned directly on the steering pinion. Integrating the torque sensor, servo unit and reduction gear into the steering pinion housing results in a compact EPAS system, which, however, has a relatively inflexible layout. Single-pinion EPAS is used in small- and middle-range vehicles.

As they are located in the engine compartment, single-pinion EPAS and the following systems are exposed to higher ambient temperatures than C EPAS systems (up to 135°C) and must be designed accordingly, especially as regards the electronic components. Its positioning also means that the EPAS system is exposed to dirt particles and moisture; therefore, the casing construction must be sealed (Figure 6).

2.2.3 *Double-pinion EPAS*

As the name suggests, this type of EPAS system features a second pinion. This second unit contains the electric motor, the controls, and the reduction gear as well as the pinion. Because of the principle used, the torque sensor is integrated into the pinion unit on the steering column, similar to the single-pinion EPAS, so that a separate signaling line is routed to the electrical control unit. Compared to the single-pinion concept, double-pinion EPAS offers a greater level of flexibility in the arrangement of the servo unit and thus its placement within the limited space available in the engine compartment. As it is independent of the steering pinion ratio, the gear ratio of the reduction gear can be optimized for applications with somewhat higher controlling torque (Figure 7).

2.2.4 *Axially parallel EPAS*

Axially parallel drive systems are used in vehicles with medium to high axle loads and are presently enjoying strong growth, as hydraulic systems are now being replaced with electric servo systems in this vehicle class too as EPAS systems becoming increasingly mature.

The unit, which is composed of an electric motor and an ECU, is arranged axially parallel to the steering gear. A pinion on the motor shaft drives a toothed belt, which transfers the torque to the nut of a ball screw drive, whose spindle is on the steering rack. Owing to the effective transmission stages, toothed belt, and ball screw drive, this

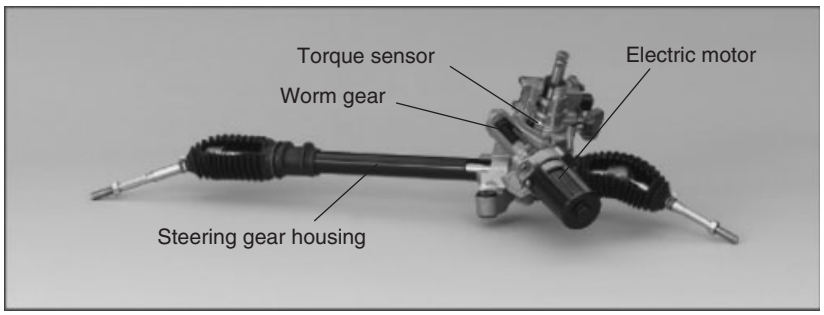


Figure 6. Single-pinion EPAS (with casing open). (Reproduced by permission of NSK Deutschland GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

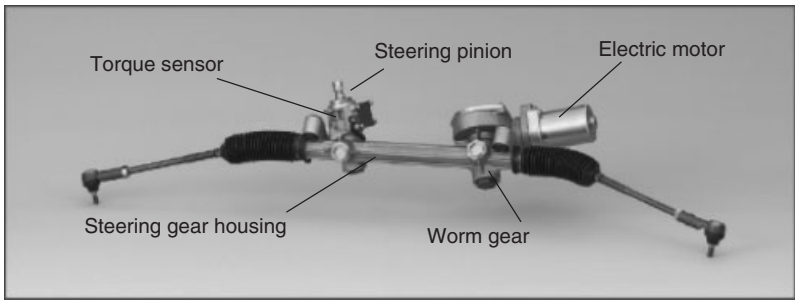


Figure 7. Double-pinion EPAS. (Reproduced by permission of NSK Deutschland GmbH.)

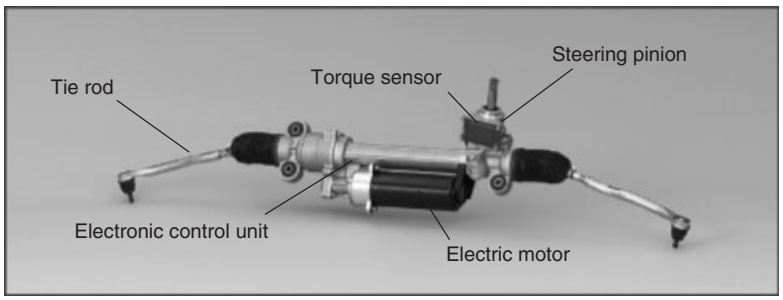


Figure 8. Axially parallel EPAS. (Reproduced by permission of NSK Deutschland GmbH.)

form of EPAS has a high overall efficiency, which enables it to control large steering forces. The torque sensor is in turn integrated into the steering pinion unit and connected with the servo unit via a signaling line (Figure 8).

2.2.5 Rack-concentric EPAS

Rack-concentric EPAS systems are used in vehicles with high axle loads and a correspondingly high actuation force requirement. However, they are still not very widespread today.

With this system, the rotor of the electric motor is seated directly on the ball screw nut. The torque of the electric motor is converted into an actuation—force acting on the steering rack by the ball screw. As the system includes only one transmission stage, the electric motor must provide a very high torque (Figure 9).

3 SYSTEM COMPONENTS

3.1 Mechanics of the steering column in column EPAS

The steering column mechanics have the primary task of absorbing the steering torque applied by the driver and

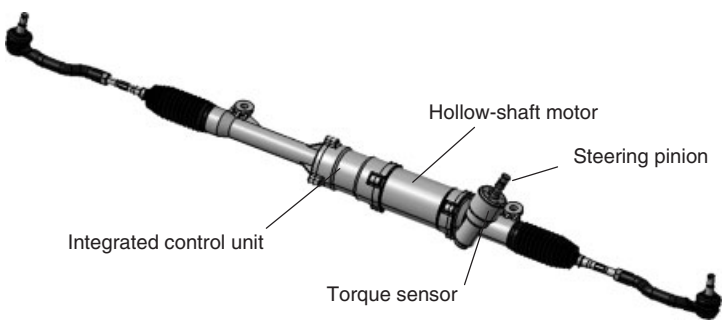


Figure 9. Rack-concentric EPAS.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

transferring it to the I-shift with the lowest possible friction losses. As the steering must always be available as a top-priority safety system, the design of the steering mechanics is correspondingly robust and subject to intensive tests. In the case of a crash, the steering column also has the task of absorbing the impact energy by means of a controlled collapse rate should the driver be thrown against the steering wheel.

The steering column has the following mechanical interfaces to the driver/vehicle system (Figure 1):

- Steering wheel (torque input)
- I-shaft (torque transfer)
- Cross car beam (fixing)
- Instrument panel (package interface).

The mechanics of a steering C EPAS system include a divided steering shaft, whose two parts are connected by a torsion bar. Depending on the torsion bar stiffness, typical values are between 2.0 and 25 Nm/°, this design ensures a rotation angle difference of the input and output shafts relative to the manual steering torque. This angular difference is used to detect the steering torque (Section 3.3.1; Torque sensor). The accumulated torque from the manual steering torque and the servo assist torque is transmitted, via an I-shaft, to the steering pinion and translated into the resulting steering rack force.

A distinction is made between the following steering column types based on their adjustability:

- A steering column that cannot be adjusted individually to the driver position is known as a *fixed steering column*.
- For enhanced comfort, the height of some steering columns can be adjusted. In this case, the steering column is released using a lever and set to the right height for the driver's position.
- The third type of steering column can be adjusted for height and reach. This allows the steering wheel position to be moved up or down and in or out.

The adjustable systems feature a divided steering shaft, which is compatible with a clamp system to fix the steering column in position. A special version features electrically adjustment for height and reach. This is mainly found in premium vehicles. Such electrical adjustment systems are made using an electromechanical system separate from the steering actuator system (Sections 3.1.1–3.1.3).

3.1.1 Manual height adjustment

When vehicles are developed, the steering wheel position is designed for a driver of a certain height. Depending on the actual anatomy of the driver and his or her individual seat position, a height-adjustable steering column can increase the level of comfort. Height-adjustable steering columns pivot at their lowest mounting suspension point. The steering column is guided along a slot on the upper fixing, which determines the possible tilt of the steering column (approximately ±20–25 mm is typical). A clamping system guarantees that the steering column is fixed securely in the selected position, even when it is exposed to loads, such as the driver pulling himself or herself up, using the steering wheel. Both friction-based mechanical systems and positive locking are established options.

3.1.2 Manual reach adjustment

A further adaptation of the steering wheel position can be achieved by adjusting it for reach, that is, changing the distance between the driver and the steering wheel. There are two types of reach adjustment, which are used, based on the philosophy of the vehicle manufacturer and the space available.

Type 1: During reach adjustment, only the inner steering column tube is moved, leaving the upper cardan joint of the I-shaft in position.

Type 2: The entire upper steering column is moved, including the upper cardan joint. A sliding mechanism in the I-shaft compensates for this movement in the system.

For both types, reach adjustment is guided by a slot, which defines the possible adjustment (approximately ±20 mm is typical). The steering column is fixed in position using friction clamping or positive locking. The system must be designed to ensure that the forces applied by strong drivers do not accidently alter the steering wheel position. This comfort-enhancing option is generally combined with height adjustment, so that the driver can be optimally adjusting the system.

3.1.3 Steering column with electric height and reach adjustment

Some premium vehicles have electrically adjustable systems. These systems have one electromechanical actuator system for height adjustment and one for reach adjustment, allowing the driver to adjust the steering wheel position within the permitted range by activating the corresponding switch. The mechanism is controlled by a separate system, independently of the EPAS control unit.

These systems are not only used to adjust the steering wheel position when driving, but they also perform additional functions. For example, when the driver exits the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

vehicle, the steering wheel can be lifted automatically to make getting out easier. After the driver gets in, the steering wheel is automatically returned to the specific position stored for the respective driver.

3.1.4 Passenger safety in the case of a collision

As part of the passive safety system, the steering system includes mechanisms for the protection of the driver in the case of a vehicle collision. If the driver is thrown against the steering wheel, the steering column should absorb energy and thus decrease the forces acting on the driver. Basically, a vehicle collision—in particular a head-on collision—can be divided into two phases, the primary and secondary collisions.

During the primary collision, the vehicle hits the obstacle and the engine compartment is deformed. Because of this, the engine is pushed back, which would move the steering wheel toward the driver. This relative movement between the steering gear and the steering wheel is compensated by a sliding mechanism in the I-shaft—the same as that used for reach adjustment (Figure 11).

The secondary collision refers to the rapid deceleration of the vehicle and its passengers and the reaction of the steering system. The air bag is released and the driver is flung against the steering wheel. The steering system now protects the driver by absorbing largely the impact energy. Defined counterforces are generated here, whose characteristics can be represented in three phases.

As already mentioned in the description of reach adjustment (Section 3.1.2), the steering wheel can also be subject to high loads exerted in the direction of the steering shaft during operation. For example, the steering wheel is often used for support when the driver moves or changes the position of his or her seat. In these cases, the steering column must not collapse, that is, high initial forces acting parallel to the axis of the longitudinal column must be absorbed (typically 1.5–2.5 kN) in order to initiate the first phase of the absorption system. In constructive terms, this breakaway force is represented by crash elements, which activate the collapse of the steering column at a defined force. The crash elements are connecting elements between the carrier bracket and the steering column, which allow movement parallel to the axis above a predefined force threshold. This function can be realized by means of a friction force-oriented design. However, this is generally dependent on the tightening force of the bolts during vehicle assembly. Alternatively, a capsule can be used, whose predetermined breaking points allow the steering column to collapse if the predefined force threshold is exceeded. The properties of this capsule are independent of the vehicle assembly.

In the second phase, the steering column is slid together against a relatively constant force level. One way to absorb the force is to use a design, which generates a defined friction between the internal and external steering column tubes. For instance, ribs can be pressed into the internal steering column tube, which generates almost a constant friction force between the tubes. This phase can be overlaid by additional elements, which allow the counterforce characteristic to be set individually (tuning). This option is particularly interesting in the case of platform applications for steering columns, where it may be necessary to adapt the mechanism to the various models in the design series.

In the third phase of the defined collapse of the steering column, the system reaches its limit, meaning that the residual kinetic energy is absorbed by elasticity in the steering system and in the vehicle. The counterforce rises sharply as a result.

3.1.5 Reduction gear

The reduction gear has the task of converting the fast rotating movements of the electric motor's drive shaft into the rotation of the steering column shaft. As well as the purely functional requirement of converting this power, the EPAS gear has a significant influence on the acoustic and tactile characters of the power steering.

Steering C EPAS systems typically use worm gears as reduction gears (Figure 10). The worm interlocks with the motor shaft and the casing contains roller bearings. The bearing arrangement is elastic, so that the system is constantly subject to a defined preload. The preload, combined with the precise manufacturing process used for the friction partners (the worm and worm wheel), leads to a quiet, uniform transmission of energy. Worm gears are used in steering C EPAS, single-pinion EPAS, and double-pinion EPAS systems.

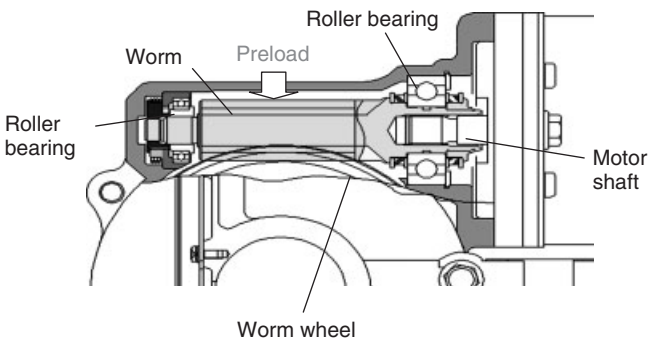


Figure 10. Worm wheel gear in steering column EPAS.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

As the reduction gear acts directly in the torque flow of the steering mechanics, any jamming or mechanical damage must not block the steering. A low coefficient of friction in the reduction gear ensures that the driver is also still able to steer the vehicle manually if the steering assist is switched off. As the mechanics are permanently in action, the construction must also prevent any blocking of the motor (Section 5).

3.1.6 Intermediate shaft

The I-shaft uses a positive fitting design to transfer the steering torque from the upper steering column to the pinion of the steering gear.

The dimensions of the I-shaft depend on the steering torque to be transferred and the respective load profile of the steering system. In the case of steering C EPAS, both the manual steering torque applied by the driver and the assist torque of the servo unit must be transferred. The I-shaft of EPAS systems that are integrated into the pinion gear or into the steering gear only need to transfer manual steering torque. On the basis of their application, I-shafts must be designed to meet demanding vehicle-specific or platform-specific load profiles throughout their service life and to cope with the maximum force requirements. As well as safety requirements, stiffness and play tolerance values must be considered in the construction of the I-shaft, depending on the vehicle-specific requirements for steering feeling.

Figure 11 shows an intermediate shaft for an EPAS system. Cardan joints are used at the interfaces to the steering column shaft and to the pinion on the steering gear side, in order to balance the angle and the axis offset between these component parts. An equalizing mechanism compensates for relative movements between the upper and lower cardan joints as they occur, for example during vehicle assembly, in the case of a crash, when adjusting the steering wheel reach, or when driving as a result of elasticity.

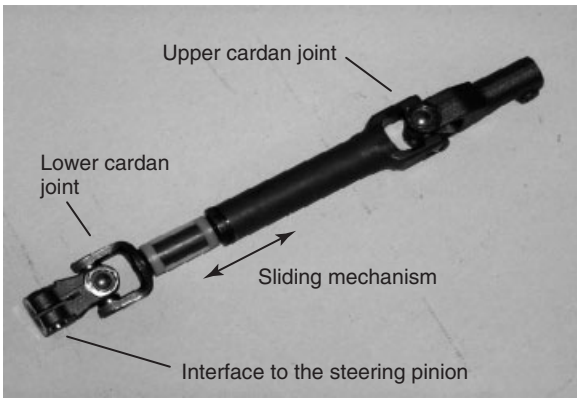


Figure 11. Intermediate shaft. (Reproduced by permission of NSK Deutschland GmbH.)

3.2 Electromechanical system

The main task of the electromechanical system for EPAS is to identify the driver's steering requirements and reliably provide assist torque. Various analyses and secondary functions are integrated into the software to perform this task conveniently and reliably. Additional steering functions are increasingly being incorporated into EPAS systems, which are not perceived directly by the driver, as are functions, which offer more obvious benefits, such as parking assistance. The most important functions are described in Section 4.1.

The electrical components of a modern steering C EPAS system are shown in Figure 12. The torque sensor, ECU, and electric motor are the core components of all forms of EPAS. To identify the driver's steering requirements, the manual steering torque is registered by a torque sensor and transmitted in the form of an electric signal to the ECU. As well as this internal signal from the EPAS system, vehicle data—such as the vehicle's speed or the rotational speed of the combustion engine—can be supplied to the ECU by means of a CAN or FlexRay bus. From this input data,

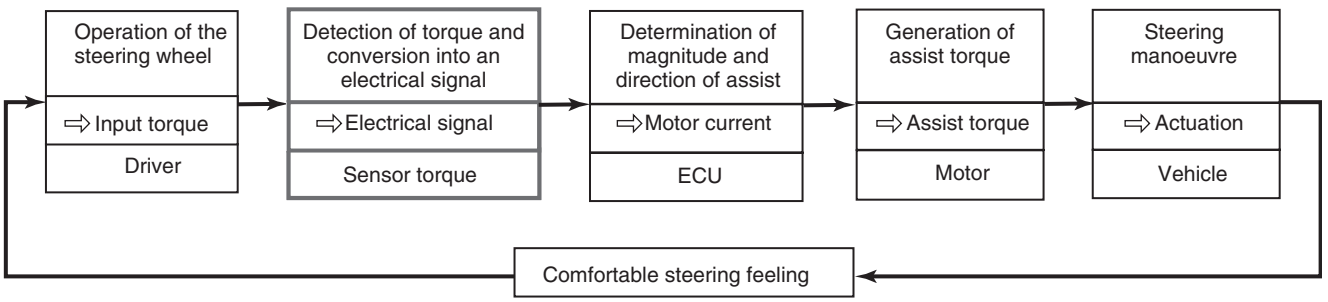


Figure 12. Workflow of the electrical system of an EPAS.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the ECU's processor unit determines the necessary assist torque, and the phase currents are computed for the electric motor and injected into the phases via the power electronics. The engine torque provided is converted by the reduction gear and supplements the driver's manual steering torque.

Knowledge of the absolute position of the steering wheel angle is necessary for different steering functions. The necessary signal can be registered by an external steering angle sensor and made available to the EPAS system via the interface to the vehicle communication. In modern EPAS systems, the absolute steering angle can be calculated using the existing wheel speed signals by means of an integrated electrical hardware system or a complex algorithm. The absolute steering angle position is used not only for different steering functions but also for other vehicle systems such as ESP. As both the steering and the ESP system must satisfy the highest safety demands, the signals are developed according to standardized safety norms (see also Section 6 on functional safety).

The subsystems, which make up the electromechanical system, are described in the following sections.

3.3 Sensor technology

3.3.1 Torque sensor

The steering torque exerted by the driver is a crucial input variable for the calculation of the electrical steering assist. High demands are therefore placed on the torque sensors for EPAS systems. These exceptional requirements are described in the following list:

Exact—In order to be capable of calculating the assist torque required by the driver, a torque sensor with high resolution of about 0.1° at high angular speeds up to $2500–3200\,^\circ$/s and good signal quality is used.

Safety—The safety of the steering system relies on the magnitude and direction of the steering movement being recorded and the correct signals being transmitted. The construction of the system must safeguard this. Sending incorrect information to the control unit can lead to unwanted steering assist or countersteering. These are major safety-related faults. Such faults must be prevented and/or identified using reliable error detection. Suitable action must be taken to avoid situations, which could lead to accidents.

Reliable—The technology used and the design in question must have a high level of reliability, that is, very low failure rates, even when subject to external influences such as vibrations, high temperatures or fluctuations of temperature, and under the influence of electromagnetic radiation.

Low friction—As modern EPAS mechanics are constructed using a very low friction design, this requirement must also apply for the subsystems, such as the torque sensor. Therefore, low friction systems or systems without contact are used.

Figure 12 shows a simplified control circuit, which illustrates the function and importance of the torque sensor. The sensor acts as the EPAS system's interface to the driver by registering the steering requirements and the steering torque applied and converting these into an electric signal. On the basis of the signal, the necessary assist power or current is calculated in the control unit, considering the driving conditions. The generated actuation force itself initiates a vehicle reaction, which in turn causes a steering action by the driver.

A mechanical system, which generates a physically measurable variable dependent on the torque, forms the basis for the detection of the steering torque. To achieve this, the steering shaft is divided into an input shaft and an output shaft. The mechanical connection for the force transfer is realized via a torsion bar with a defined stiffness. The manual torque input twists the torsion bar, changing the angle between the input and output shafts. This angle is used to calculate the steering torque. Similar arrangements are used in hydraulic power-steering systems.

Figure 13 shows the torque sensor module assembly of a steering C EPAS system. The torque from the input shaft is transferred to the torsion bar via a toothed profile. In order to enable relative movements to be absorbed, for example, in the case of reach adjustment, the input shaft can be displaced axially with respect to the torsion bar. While the assist torque is transmitted to the output shaft directly by the EPAS, the manual torque is transmitted by the torsion bar to the output shaft via a pressed-in pin. A secure mechanical

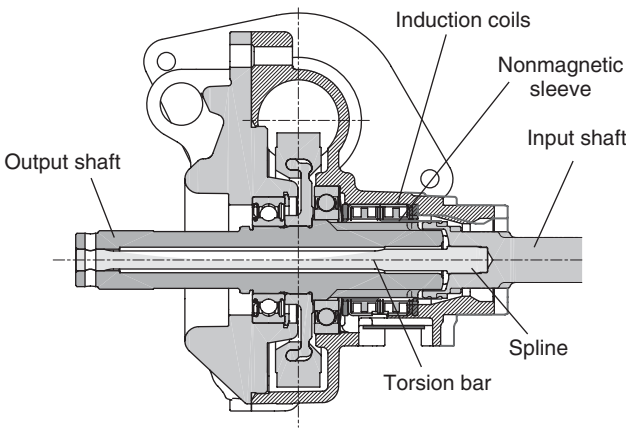


Figure 13. Integrated torque sensor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

layout and intensive validation ensure that the system is extremely reliable. However, even if the torsion bar or its connections fail, a mechanical stop between the input and drive output shaft ensures that the vehicle can be steered manually.

In the system shown, a nonmagnetic sleeve is fixed to the input shaft. The sleeve has windows, which are rotationally offset with constant phase angles. The relative movement of the output shaft's tooth profile in relation to the sleeve changes the magnetic flux in the coils, which causes a change in the coil impedance. This change in the coil impedance is converted by the sensor electronics into an electric signal and transmitted to the ECU. This electric signal is a measure of the required torque assist.

3.4 Motor angle sensor

For the brushless motors used in many systems, as well as for induction generators, the respective motor angle and the motor angular speed are required as a basis for the control algorithm. The motor angle sensor has the task of converting the relative angle position into an electric signal and transmitting it to the ECU. In order to avoid any friction, contact-free principles, such as magnetoresistive (MR) technologies or resolvers, are used in modern EPAS systems.

3.5 Steering angle sensor

The control unit, based on the information supplied by the motor angle sensor, considering its resolution, the mechanical gear ratios, and the mechanical tolerance stack can calculate the relative steering angle. As the motor angle sensors are not multiturn capable, that is, the position can only be determined within one revolution (360°) of the motor shaft, the absolute position of the steering shaft cannot be determined. However, some steering functions require the absolute steering angle, such as the parking assistant.

One way to determine the absolute steering angle without any additional electrical hardware is using an algorithm, which can calculate the straight-line progress using the wheels' rotational speeds. This learning algorithm uses the wheels' rotational speeds determined by the anti-lock brake system (ABS) after the engine is started, in order to index the position of the straight-line progress. By means of the EPAS system's motor angular speed signal, the absolute steering angle can be calculated on the basis of this information. In principle, this process requires a certain driving distance, which is dependent on the steering profile and the road. For most steering functions, determining the absolute steering angle position in this way is sufficient.

However, ESP systems in particular require high precision, rapidly available signals. The so-called true power-on systems provide the absolute steering angle position as soon as the steering system control unit boots up when the engine is started.

The ECU is the distribution center of the electromechanical steering system. All signals arrive here and the suitable assist torques and power flows are computed for the respective driving situation then outputted to the electric motor. The system states and the active processes are monitored by the ECU in a complex safety structure, and appropriate action is taken if a fault is detected (Figure 14).

The input circuits of the signal electronics convert the torque sensor signals from the vehicle power supply and the angle position of the motor into electric signals, which can be processed by the ECU. To generate suitable steering assist, the central processor unit (CPU) uses this information to calculate the current to be injected into the individual motor phases. To activate the motor phases, the CPU sends a control current to the gate driver, which in turn activates the field-effect transistors (FETs) in the power output stage, which is arranged in an H-bridge. As well as the resistive losses, switching losses also occur in the power electronics, which lead to a high level of heating in the component parts. For this reason, the power electronics are placed directly on the housing of the control unit for optimal heat dissipation and the construction ensures that heat is effectively transferred to the steering gear housing.

The monitoring of the CPU is a key component of the safety concept for an EPAS system, because incorrect outputs can lead to fatal faults. The calculated results and/or the computing capability of the processor and the control algorithms must be checked independently. For instance, important calculated results are checked by independent parallel algorithms within the CPU, which have to considerably differentiate from the basic control algorithms. Furthermore, the system diagnostics monitor the software and the electrical hardware of the control unit by running calculations in two independent computer units. Some of the monitoring is completed in the central processor, whose results are checked by a further processor (sub-CPU). If, in the comparison calculations, differences are diagnosed outside of a stipulated tolerance range, which can lead to faults with safety implications, the system goes into safe mode.

Depending on the type of fault identified, the system can revert to replacement actions, which only reduce the level of steering comfort or the degree of steering assist. If a more severe fault is identified, power-assisted steering is switched off and the vehicle must be steered manually. As modern EPAS systems are extremely reliable, these interventions by safety functions are extremely rare occurrences, however.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

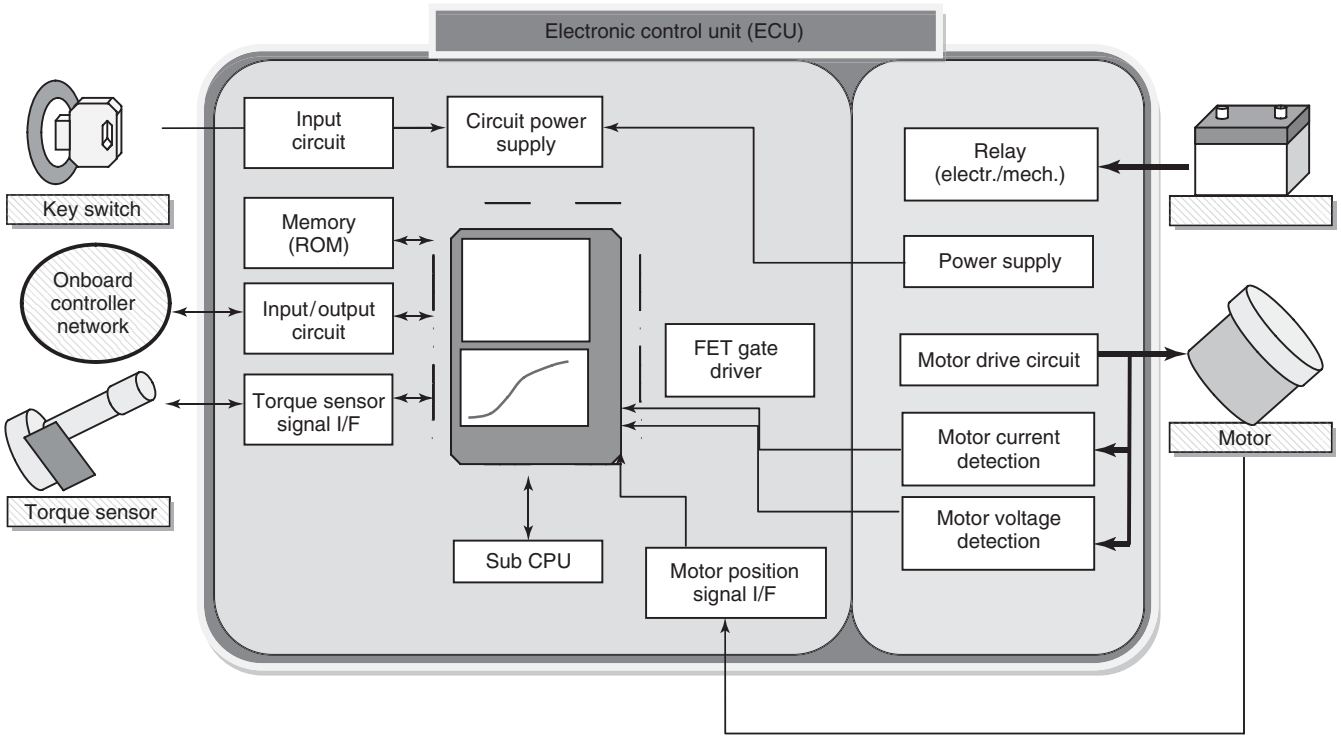


Figure 14. Simplified block diagram of an EPAS electronic control unit.

4 STEERING FUNCTIONS

All functions, which are provided by means of software algorithms within an EPAS system, are referred to as *steering functions*. A distinction can be made between functions, which can be incorporated into the EPAS system and functions, which are supported by the EPAS but are activated by other control units in the vehicle. Generally, the functions, which can be implemented via the EPAS system, serve to support the driver and keep providing him or her with appropriate feedback on the driving situation. However, the driver remains responsible for decision-making regarding the steering. Tactile feedback from the steering system should also place the driver in a position to identify faults in the vehicle, such as a badly balanced wheel or tire pressure loss.

4.1 Basic steering functions

4.1.1 Steering assist

The steering assist provided by the servo motor is the elementary function of an EPAS system. As described in Section 3.3.1, the steering assist is outputted primarily based on the manual steering torque applied by the driver, steering requirements, and the vehicle's speed. The characteristics of the steering assist are extensively coordinated in test runs by vehicle manufacturers and steering suppliers, and these characteristics determine the character of the steering in the respective range of vehicles. The philosophy of the engineers, responsible for coordinating steering in a certain manufacturer's vehicle, reflects the driving feeling that the driver associates with a brand, regardless of the platform in question.

4.1.2 Damping

During operation, the steering system is subject to various influences, which tend to affect the stability of the steering system in a negative manner. In the ideal case of a straight-line drive, no steering intervention and thus no servo assistance would be necessary. In reality, however, acceleration is continuously present from external disturbance variables, such as the stimulation from the road surface or the small steering movements made by the driver, so that these influences do not lead to instabilities in the steering activation, these disturbance variables are detected by the system and dampened in a speed-dependent manner. The character of the damping can be defined by corresponding parameterization in the vehicle coordination phase.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.1.3 Friction compensation

In the electromechanical steering system, friction losses arise at different points, which have a corresponding influence on the steering characteristics. A supplementary assist torque and a corresponding compensation current are calculated to compensate for the friction and added to the basic steering assist.

4.1.4 Inertia compensation

Steering a vehicle is a dynamic process, which—considering the transmission ratios typical of EPAS systems—is disturbed by the inertia of the relatively large masses moved. Without further measures, the driver would have to continuously steer against the resulting forces. Inertia compensation means that the driver no longer has to do this and gives him or her the sense that the steering responds immediately and exactly to his or her actions.

4.1.5 Reduction of the assist power

If the driver steers into an obstacle, such as a kerb, the wheels are blocked, so that no further steering movement is possible. As long as the driver continues to exert a steering torque, however, the EPAS system will provide a powerful assist torque, which will generate very high power dissipation, leading to heating of the system. In order to reduce this loading and to protect the system, an algorithm is implemented that identifies any blocking and reduces the assist power.

4.2 EPAS steering functions

A great advantage of EPAS systems is the option of going beyond basic steering features by integrating further steering functions via software algorithms without the need for extra hardware. Section 4.2.1 describes functions, which can be realized by the EPAS system using existing information in the vehicle network.

4.2.1 Active return

In a high quality steering system, the driver expects the steering wheel to return clearly to the center position depending on the vehicle speed. This means that, when cornering, he or she expects a uniformly increasing counterforce before the steering is reset on the next straight stretch. Furthermore, the neutral position of the steering wheel should be unambiguously reached and maintained as long as the driver exerts no steering torque. On the whole, modern chassis kinematics are no longer designed with this focus in mind as the axle construction considers the option of active return via the EPAS system and the mechanics are optimized with regard to other aspects.

The active return function calculates the necessary active return torque through the EPAS system, considering the steering wheel torque, the vehicle speed, the absolute steering angle, and the steering angle speed. Without steering torque input by the driver, in relation to the vehicle speed, the vehicle automatically goes back to traveling in a straight line after cornering.

4.2.2 Soft end stop

Very high rack forces and steering speeds can occur in an EPAS system, depending on the steering requirements. If no other measures are implemented, the steering gear can reach its end stop and be exposed to very high mechanical loading as a result. The mechanics must therefore be designed robustly, which drives up weight and costs.

To optimize the mechanical system, a function can be introduced that can reduce the EPAS assist before it reaches its mechanical end stop, or even oppose the driver's residual manual steering torque. For the realization of this function, knowledge of the absolute position of the steering is necessary, calculated from the absolute angle position of the steering wheel. As well as the mechanical design optimization possibilities offered by this function, the driver also perceives steering toward the end stop as significantly softer, which further enhances comfort.

4.3 Vehicle functions

The technical possibilities of EPAS systems enable different comfort-enhancing functions to be introduced at vehicle level. Other control units in the vehicle manage overall functionality, and the necessary function of the EPAS system is called up via the vehicle communications. However, responding to requirements remains the responsibility of the EPAS system. Many of these additional driver assistance functions lead to an immediately perceptible increase in tactile comfort for the driver, and they can be marketed correspondingly. It is expected that greater use will be made of such assistance functions in the coming years. Some of them are described in the following sections.

4.3.1 Lane keeping assistance

If a vehicle changes or leaves its lane following an action such as use of the indicators or an unambiguous steering movement, it is clearly the driver's intention to do so. However, if the vehicle slowly drifts out of its lane, this

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

could be caused by inattentiveness on the part of the driver and could lead to an accident.

Unintentional departure from a lane is generally detected by means of camera systems and feedback about the driver's actions. In order to make the driver aware of this situation, the EPAS system can send tactile feedback to the driver, for example, by making the steering wheel vibrate briefly. However, the driver is still responsible for taking any action to correct the vehicle's direction. This function is also known as *lane departure warning*.

Before active support can be provided to avoid unintentional lane departure, several safety criteria must be met. The driver must be able to override the active steering intervention at all times and the system must always be able to identify whether the driver has his or her hands on the steering wheel by evaluating the manual steering torque. If these prerequisites are met, any active correction by the EPAS system using information from the camera system can be activated.

EPAS can also actively support safety by giving the driver a steering torque recommendation. If oversteering or understeering is registered at vehicle level, or the vehicle skids after braking on μ-split condition, the driver can be encouraged to counteract this by means of short torque pulses to the steering wheel. Here too, the driver must retain ultimate control over the steering, that is, it must also be possible to override these steering pulses.

4.3.2 Parking assistance

Parking assistance helps drivers with parallel and reverse parking. Different types of parking assistance are available, which differ in the degree to which the vehicle's acceleration is automated (accelerating and braking). Steering is always taken over by the vehicle. Autonomous steering places special demands on the steering system in that different steering requirements are communicated by an external signal via the vehicle communication, instead of via the system's internal torque signal. As the driver still has to keep control of the procedure, any possible intervention on the steering wheel is identified by the EPAS system and is responded to appropriately, for example, by terminating the parking procedure.

5 ELECTRIC MOTORS FOR EPAS SYSTEMS

The electric motor in an electromechanical steering system primarily has the task of converting electrical energy into the required mechanical assist torque. Its effect on the character of the EPAS system becomes clear when the most important requirements are considered in more detail:

Performance: From small steering angle speeds to speeds of one steering wheel revolution per second, an EPAS system must be able to make the maximum nominal torque available. Load profiles, which consider the frequency and scales of the assistance provided, are used to design the layout of the motor. For instance, the maximum steering torques only ever have to be made available on a short-term basis and a low level of performance is all that is usually required during operation.

Dynamics: The size and direction of the steering assist changes highly dynamically during steering operations. The motor is therefore required to have the lowest possible rotor inertia and very dynamic response characteristics.

Efficiency: As electric steering uses a comparatively large amount of power, the system must be highly efficient in order to make optimum use of the limited energy available from the vehicle power supply. A high level of efficiency, in combination with a high power density, is also necessary to enable EPAS systems to be positioned in the confined spaces available in the vehicle.

Acoustics: The electric motor for an EPAS system should be as quiet as possible. Any residual noises should be "pleasant," that is, in a relatively low frequency band. Low levels of mechanical friction, the avoidance of resonance, and a good electric design, as well as corresponding motor control, contribute to achieving these objectives.

Tactile properties: A uniformly running rotor with low torque fluctuation results in a high quality steering feeling.

Safety: When considering the functional safety of EPAS systems (Section 6), the focus is on self-steering and blocking in particular. In principle, any blocking of the rotor by a foreign body in the motor can lead to blocked steering. The system is therefore constructed so as to exclude foreign bodies during manufacturing and/or prevent the release of particles during operation. Electric short circuits must also be avoided by constructive means, because these can lead to inadmissible braking torques.

Robust against environmental impacts: EPAS motors are exposed to vibrations and a large ambient temperature range, as well as to different media depending on where they are installed. The motor module assembly must be designed robustly to withstand these influences and provide the required performance reliably throughout the vehicle's service life.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Costs: Costly materials are used in electric motors, such as copper fillings and various magnets. Ongoing improvements to the design and materials utilized for electric motor development reduce motor costs and thus contribute to maintaining the overall costs of EPAS systems at a competitive level.

5.1 Motor types

In EPAS systems, DC motors or brush motors (brushed DC) are used for a mechanical power in the region of 500 W, asynchronous motors (ASMs) are used for 300–500 W, and brushless motors (brushless DC or BLDC) are used up to approximately 900 W.

5.2 Structure of the brushless DC motor

Even though brush motors continue to have their areas of application, BLDC motors are preferred for modern EPAS systems because of their low inertia, long service life, high efficiency, and high power density. The structure of this motor type, which is also designated as a permanent magnet synchronous motor (PMSM) or electronically commutated (EC) motor, is described later.

In order to achieve a high power density, high energy magnets (approximately 360 kJ/m^3) made of neodymium iron boron (NeFeB)—the so-called rare-earth magnets—are used for BLDC motors. During manufacturing, the magnetic powder is pressed, sintered, and then cut into the respective application shape with diamond tools. Ring or segment magnets are used in EPAS motors, which are fixed on the surface of a laminated core. In order to prevent the motor from being dangerously blocked by particles released from the brittle magnetic material, the magnets are bandaged. Rectangular pocket magnets, which are inserted into corresponding pockets in the laminated core, are another magnet construction design. The laminated core with the magnets is pressed onto the motor shaft. This combination forms the rotor, which is supported by roller bearings in the motor housing and in the housing cover.

EPAS motors of the newer generation are generally structured with technology using individual tooth winding. This enables the spatial separation of the phase windings, in line with the safety objective of preventing short circuits between the phases. The running properties of the motor are determined by the combination of the number of slots and the number of poles. 9/6 (9 slots/6 poles), 12/8, and 12/10 motors are frequently used in EPAS systems. As in the case of electrical machines with slotted stators, an undesired cogging torque occurs due to the varying magnetic resistance associated with the change from stator tooth to stator slot. This effect is countered constructively by skewing the rotor or the stator.

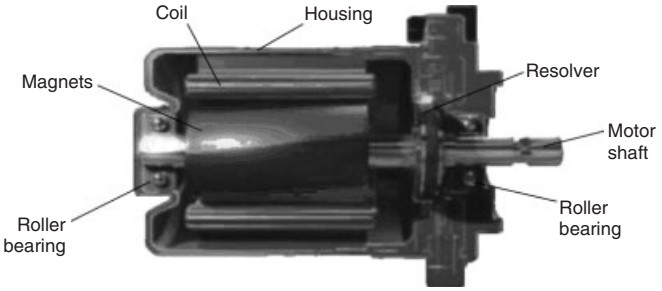


Figure 15. Motor structure of a synchronous motor.

BLDC motors in EPAS systems are commutated electronically in relation to their position. Resolvers integrated into the motor module assembly, or MR measurement methods, are used to determine the rotor position. In line with the rotor position, a stator current is injected via the power electronics of the control unit to create the stator magnetic field (Figure 15).

6 FUNCTIONAL SAFETY IN EPAS SYSTEMS

To put it simply, the steering system should always follow the requirements of the driver, that is, the steering torque should be exerted as requested in terms of both magnitude and direction and the requested steering angle should be set. Similarly, it is important that no steering assist is provided when it is not requested and that the system is always steerable.

In order to ensure product safety, the systems are based on standards, which cover the entire life cycle of safety-related electrical and electronic systems. Systematic development in defined processes forms an elementary component of these standards.

The IEC 61508 standard "Functional security of safety-related electrical/electronic/programmable electronic systems" originally dates from 1998 and is used for the development and evaluation of EPAS systems. It covers "all safety-related systems that include electrical, electronic, or programmable electronic components (E/E/PES)." Depending on the risk to persons and the environment, the systems considered are classified into Safety Integrity Levels SIL1 to SIL4. A steering system is classified as SIL3, which corresponds to the highest safety level in the automotive sector and requires the system to be approved by independent third parties.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

While IEC 61508 for safety-related electrical/electronic systems covers a range of power systems including EPAS, a specific standard (ISO 262629) has now been developed for automotive electrical/electronic systems on the basis of IEC 61508. This has been officially valid since the end of 2011. This standard will replace IEC 61508 in the development of electrical steering systems, which will affect the product generation process and the product design.

Corresponding to the IEC norm, ISO 26262 also classifies systems using Automotive Safety Integrity Levels from ASIL A to ASIL D. On the basis of an analysis of the system risk, EPAS systems are classified in the highest category, ASIL D. As mechanical aspects are not considered in the standard, fault-free mechanical systems are always assumed in case of the electromechanical systems analysis.

6.1 Identification of the safety hazard

In the risk analysis, the possible safety-related effects of faults are identified in the overall system considered. For an EPAS system, the following safety-related faults are considered:

- *Self-steering*: The servo unit assists in magnitude or direction, which does not correspond to the driver's requirements.
  Self-steering is caused by incorrect torque assist from the electric motor, which is incorrectly activated, for example, by an incorrect signal from the torque sensor, a fault in the electrical circuit or a software error.
- *Steering blocking*: A distinction is made between two scenarios.
  (a) The steering is blocked mechanically. As only the electrical system is considered in IEC and ISO, this case is reduced to a blockage of the electric motor.
  (b) The steering is so stiff that the vehicle can no longer be safely steered. This fault can be caused by a phase short circuit in the motor or in the motor control circuit. In this case, the motor is operated as a generator, so that a corresponding electromagnetic braking torque acts against the steering movement.

6.2 Safety objectives

ISO 26262 requires the application of a redundancy and diagnostics concept, which also guarantees that the system enters safe mode in the case of an error. Measurable target values are defined using special metrics, which are divided into three categories:

- Isometrics give the target value for the maximum acceptable probability of occurrence for safety-related faults. This target value is defined for the occurrence of a fault with reference to the operating time (failure in time, FIT);
- Single point of failure (SPF) refers to errors that lead directly to a safety-related fault;
- Latent faults are undetected faults that, in combination with a further fault, lead to a violation of the safety objective.

In order to ensure that these objectives are met, EPAS systems are developed strictly according to the automotive SPICE model, in combination with further measures from the ISO standard.

6.3 Functional safety in use

The consideration of functional safety includes all components of the electrical system, which have an influence on the actuation force and are capable of causing blocking. The sensor technology, the electrical hardware, the software, and the motor must be considered. Depending on their impact on functional safety, external information is also verified again in the EPAS system.

The top priority is to develop components with a very high level of reliability, that is, to avoid faults in operation. Nevertheless, as there are chances for a fault occurring in operation, a fault identification mechanism must be in place, which analyzes the fault and classifies it according to criticality.

If a fault with a possible impact on safety is identified by the diagnostics, a system response is triggered. Depending on the type of fault, substitute reactions are initiated that, as far as possible, are not perceptible to the driver or are only perceptible to a small degree. In some cases, a gradual reduction of the assist power is necessary to make the vehicle safe. With every fault detection and the corresponding system response, a defining factor is that the driver remains control of the vehicle at all times.

For development test drives and homologation (certification for road use), the results are finally evaluated from the process of developing the electric subsystem and also from the approval tests of the electromechanical overall system before approval.

7 SUMMARY

Modern EPAS systems will continue to become much more widespread in the coming years, with the different systems being used in top-of-the-range vehicles and in small- and middle-range models. While basic functions,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

reliable operation, and safety were the focus when developing the first generation of EPAS, modern EPAS systems will feature new steering functions, reduced weight, and a more compact design. However, the introduction of the new ISO 26262 standard for functional safety will also lead to further development of the electronic concepts. In order to optimally exhaust the possibilities for further development, it is crucial to develop the system in an integrated way, ensuring compatibility between mechanics, electrical hardware, and software. The global steering manufacturer NSK is one company, which combines a deep knowledge of steering column mechanics with an expansion of the development of steering-specific control units. Such kind of system competence leads to modern safety architectures and steering functions to meet future needs for steering systems.

FURTHER READING

Braess, H.-H. and Seiffert, U. (2001) *Vieweg Handbuch Kraftfahrzeugtechnik* 2. Auflage, Friedrich Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden.

Heißing, B., Ersoy, M. and Gies, S. (2011) *Fahrwerkhandbuch: Grundlagen, Fahrdynamik, Komponenten, Systeme, Mechatronik, Perspektiven (ATZ/MTZ-Fachbuch)*, Vieweg + Teubner, Wiesbaden.

Isermann, R. (2007) *Mechatronische Systeme: Grundlagen*, Springer, Berlin, Heidelberg.

Pfeffer, P. and Harrer, M. (2011) *Lenkungshandbuch (2011)*, Vieweg + Teubner Verlag, Wiesbaden.

Wallentowitz, H. and Reif, K. (2011) *Handbuch Kraftfahrzeugelektronik*, Vieweg + Teubner Verlag, Wiesbaden.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto008
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Steer-by-Wire, Potential, and Challenges

Lutz Eckstein, Lars Hesse, and Michael Klein
RWTH Aachen University, Aachen, Germany

1 Introduction 1
2 History of Steer-By-Wire Systems 2
3 State of the Art 4
4 Potential 7
5 Challenges 10
6 Summary 12
References 13

1 INTRODUCTION

Since the invention of the automobile in 1886 by Carl Benz, vehicle technology has rapidly evolved multiplying performance, efficiency, and safety of today's vehicles compared to those of more than 125 years ago. Nevertheless, some technical solutions seem to be immune to technological progress: after a few attempts to steer a vehicle by a crank-like device, the steering wheel commonly, and almost exclusively, became the operating device used to influence the vehicle's direction. The first step to improve the complex vehicle operation back then was made by Alfred Vacheron in a redesigned 1893 Panhard 4 hp driven in the race Paris-Rouen in 1894 (Alexandre, 1894; Dick, 2004). The main reason for the steering wheel's success was the steering gear, which offered an ideal force transmission realizing an effortless set and hold to the vehicle's course.

Regarding the steering system, which translates the driver's steering command into a change of the vehicle's direction and—at the same time—provides an adequate feedback in order to support the stability of the driver–vehicle control loop, the predominant solution in street legal passenger and commercial vehicles is still based on a mechanical transfer of forces and torques. Numerous inventions created an evolution from purely mechanical linkages via power-assisted (see Active front steering for passenger cars; New Electrical Power Steering Systems) steering to electromechanical steering systems offering additional but limited functionalities regarding vehicle stabilization and driver assistance.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

1.1 Motivation

The motivation for introducing steer-by-wire systems can be attributed to three aspects:

- functionality,
- vehicle architecture, including benefits in production and costs, and
- human factors.

The three-level model for driver assistance systems classifies the vehicle operation into the levels of navigation, guidance, and stabilization. While today's electromechanical steering systems enable functionalities such as speed-dependent power assistance on the level of vehicle stabilization and lane-keeping assistance regarding vehicle guidance, the mechanical part of the system restricts the functionality because of the transfer of torque between steering wheel and steering system. In case of a fixed steering ratio, an operation of the steering actuator yields both a change in steering wheel angle and angle of the steered wheels.

The idea of superimposing an additional angle to the steering wheel angle while having a mechanically safe operating mode was patented for the first time in 1972 (Pilon *et al*., 1972). However, after another 30 years, Toyoda Machinery Works and Lexus, as well as ZF-Lenksysteme and BMW, have introduced a superimposed steering system into the market (Köhn *et al*., 2002; NN, 2003). Active front steering introduced by BMW superimposes an electronically commanded steering angle in addition to the steering wheel angle using a planetary gearing. This principle clearly enables additional functionality such as a variable steering ratio and improves driving stability but is still limited by the fact that the driver needs to compensate the change in steering torque resulting from the (see New Electrical Power Steering Systems) actuation of the active front steering (Köhn *et al*., 2002).

The second motivation for steer-by-wire is given by the fact that the mechanical integration of the steering system has a large effect on the vehicle package, as the steering column needs to connect the steering wheel with the steering gear in right- and left-hand-drive vehicles. The geometry of a conventional steering system also limits the vehicle variants that may be derived on a modular basis, as the ergonomically required position and angle of the steering wheel largely changes with the different postures of the driver, for example, in a limousine versus a sports utility vehicle. Introducing steer-by-wire yields a robust package of the steering system and engine compartment and maximum spreading of vehicle variants on one technical platform.

The third motivation also plays a major role in the history of aviation: with increasing aircraft size and performance, the pilot's ability to control the aircraft without significant auxiliary forces and flight control systems rapidly decreased. Moreover, the classical control column in front of the pilot consumed a large amount of valuable space and surfaces in the cockpit that were needed to integrate additional controls and instrumentation. Consequently, the entire cockpit layout has changed with the introduction of fly-by-wire incorporating compact joysticks instead of large column sticks. In addition, in a road vehicle, the cockpit layout can be significantly improved by introducing by-wire controls in terms of human factors, design, active safety, and passive safety (Eckstein, 2001).

1.2 Definition

A steer-by-wire system uses an electronic communication replacing the mechanical linkage between the operating device (e.g., steering wheel) and the steering system, as stated by the term *by-wire*.

Extended definitions of steer-by-wire systems found in literature (Binfet-Kull, 2001) also include other nonmechanical connections between the steering control and the steering actuator(s), for example, hydraulic or pneumatic systems. This broad definition is not adopted here, as these systems significantly differ in terms of technology and functionality.

Consequently, a steer-by-wire system comprises at least one sensor in order to sense the input, an electronic control unit calculating a steering command and one actuator influencing the angle of the steering system.

The exact technical solution depends on the individual requirements concerning operating device, functionality, steering actuation, steering feedback, and system integrity. In order to keep up the steering functionality in case of a system fault, a fallback mode is required that clearly distinguishes pure by-wire systems with an electrical/electronical redundancy from those systems incorporating a hydraulic, pneumatic, or mechanical backup.

2 HISTORY OF STEER-BY-WIRE SYSTEMS

Braess (2001) points out that since the 1950s, the demands on steering systems have increased because of increasingly high engine powers, improved road surfaces, and hence higher vehicle velocities. Consequently, the design of conventional steering systems was largely improved by, for example, introducing hydraulic power-assisted steering, electrohydraulic (see New Electrical Power Steering Systems) power-assisted steering, electric power steering (EPS), or superimposed steering. (see Active front steering for passenger cars Similarly to the introduction of fly-by-wire systems, the next step would be the introduction of steer-by-wire systems with a for example, mechanical, hydraulic, electric, or integrated fallback (Figure 1).

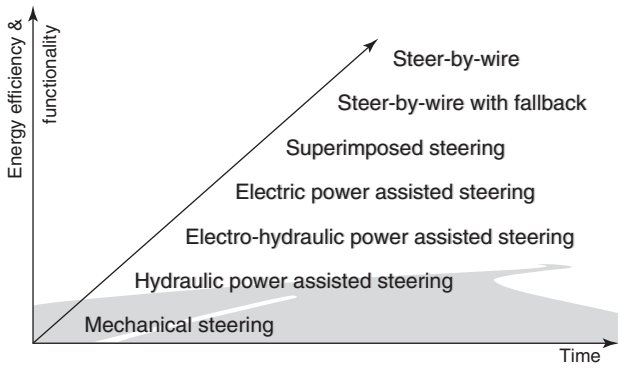


Figure 1. Timeline steering systems.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The first series production of hydraulic power-assisted steering systems started in the Chrysler models New Yorker and Imperial in 1951. The principle itself had already been filed in a patent by F.W. Davis in 1920s and was ready to go into serial production of GM's Cadillac by 1933 (Davis, 1927). However, owing to the world economy crisis, GM planned to sell only 15,000 vehicles, leading to cost-intensive machine tools in production. Finally, the introduction of the system into the market was stopped. During World War II, American and British army forces equipped military vehicles with Bendix–Davis hydraulic power-assisted steering systems and the market for heavy-duty vehicles started to develop. Series production started in 1951; by 1956 already every forth vehicle in the US market was equipped with a hydraulic-assisted power steering system. Right from the beginning, the servohydraulic steering gear has been supplied by a hydraulic pump, which has been directly driven via a belt drive by the combustion engine. Sporadically, steering systems have been equipped with electric motors driving the hydraulic pumps due to packaging reasons. The automotive R&D departments have always been focusing on designing energy-efficient steering systems. At the end of the 1990s, electrohydraulic power steering systems have been introduced into series production. The development of electric motors and pumps brought forth improved degrees of efficiency and the control of the pump rotational speed decreased the total power demand (Pfeffer, 2011; Rixmann, 1962). Following hydraulic and electrohydraulic power-assisted steering systems, EPS systems have decreased energy consumption and broadened the horizon of steering functionality. The patent from Bayle and Ecquevilly (1972) presents an early example for the idea of an electromechanical steering system. This solution shows an electric motor integrated into the steering column. The first vehicle having an EPS system was the Suzuki "Cervo" in 1988. This system developed by Koyo (Japan) has an electric motor as a component integrated in the steering column. As the front axle wheel load was small, the electric motor only needed an input power of 240 W and was connected to the steering shaft via a worm drive with a gear ratio of 1 : 16 (Stoll, 1992). The next innovation in steering design was the superimposed steering system, which has been patented in 1972 (Pilon *et al*., 1972). In 2002, ZF Lenksysteme and BMW, as well as Toyoda Machinery Works and Lexus, presented a superimposed steering in series production (Köhn *et al*., 2002; NN, 2003).

Regarding the functional motivation, *automation* was one key driver in research on steer-by-wire systems in close analogy to the aviation industry. Since the 1960s, research activities concentrated on the idea of autonomous driving. At the same time, the Concorde was the first civil airplane to perform, in 1969, its first flight with analog fly-by-wire technology. Three years later, NASA's Vought F-8 Crusader flew with a digital fly-by-wire system (Brockhaus, Alles, and Luckner, 2011; Henke, 2010).

The Eureka project PROMETHEUS (PROgraMme for a European Traffic of Highest Efficiency and Unprecedented Safety) and its successive project MOTIV, initiated by German car manufacturers, bundled the activities of many European companies and Universities starting from 1987 (see Automated Driving). In the same year, the Airbus A320 had the first completely digitalized fly-by-wire system. The electronic system was designed with redundancies. A mechanical pitch elevator and yaw rudder pedals served as mechanical backup (Brockhaus, Alles and Luckner, 2011; Henke, 2010). In addition to the PROMETHEUS project, similar projects have been initiated (e.g., NAHSC, IVI, and AHSRA) (see Automated Driving) (Stiller, 2007; Sun, Bebis, and Miller, 2006).

The conceptual idea of intelligent vehicle highway systems (IVHSs) started in the General Motors Pavilion at the 1939 World's Fair. The automotive future was presented by relaxing drivers in self-driving cars (Fenton, 1994). In recent years, the US Defense Advanced Research Projects Agency (DARPA) organized the Grand and Urban Challenges, which were intended to boost US research activities (see Automated Driving). Broggi *et al*. (2010) presented the VisLab Intercontinental Autonomous Challenge (VIAC), where four electric vehicles aimed to drive autonomously along a 13,000-km trip from Italy to China (see Automated Driving).

Apart from these competitions and races, many car manufacturers investigate the potential of driver assistance systems, which not only support the driver on the longitudinal driving task, such as adaptive cruise control, but also provide steering support. Such a system would offer autonomous driving in defined scenarios, such as traffic on motorways, and constitute an important step toward full autonomous driving.

While there is little official available information on the second key motivation for steer-by-wire, namely benefits regarding vehicle architecture, modularity, and production, the third key driver of research on steer-by-wire, *safety and human factors*, was also inspired by progress in aviation technology. Before the invention of the airbag in 1951 (Linderer, 1951) and its final breakthrough in the Mercedes-Benz W126 in 1980 (Patzelt, Schiesterl, and Seybold, 1971; Kramer, 2009), the steering wheel and column often caused severe injuries and many deaths. Already in 1959, General Motors presented a research vehicle based on a Chevrolet Impala, which had a door-integrated joystick to control the longitudinal and lateral vehicle dynamics (Bidwell and Cataldo, 1958). General Motors designed the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

vehicle stick displacement to be proportional to lateral acceleration at velocities >10 or 15 mph. As there have been no mechanical connections between stick and the controlled vehicle, the freedom in designing the most suitable steering response characteristics was one principal challenge. The joystick feeds back the car's motion to the driver. According to Bidwell (1958), the vehicle's path stability in critical situations was improved while the control knob acted as an accelerometer and also the sensitivity to wind gusts was decreased. The GM Firebird III was presented at GM's Motorama in the same year. This vehicle concept, which was inspired by the space age, had an integrated single joystick like that of the GM Impala replacing steering wheel, gas, and brake pedal (Bidwell, 1959; Davis 2004).

In recent two decades, numerous research prototypes and concept cars with steer-by-wire systems have been built and as such, only a few examples can be mentioned in this chapter. A concept car, the Saab 9000, having an active joystick with force feedback control was presented in 1991. The sidestick concept, originally developed for a military aircraft, was moveable in lateral direction, whereas a passive spring–damper combination and an electric motor applied the feedback force (Bränneby *et al*., 1991). The DaimlerChrysler F200 Imagination, presented in 1996 at the Paris Motor Show, was a research vehicle featuring a steer-by-wire system with two hydraulic sidesticks instead of a steering wheel. Eckstein (2000) carried out substantial research on control algorithms and human factors resulting in a Mercedes-Benz SL 500 (R129) prototype with two active joysticks (Figure 2). The control commands for steering, throttle, and brake were based on force transducers, whereas the displacement of the sidesticks provided an active feedback in lateral direction; in longitudinal direction, the joysticks were isometric. In 2002, General Motors presented the GM Hywire, a steer-by-wire concept where steering was achieved by gliding up or down the steering device handgrip creating a feeling similar to that of a conventional steering wheel. In 2003, DaimlerChrysler published a steer-by-wire concept in the F500 Mind on the Tokyo Motor Show.

3 STATE OF THE ART

3.1 Classification of steering systems

The oldest steering system is the turntable steering, which still can be found nowadays in truck trailers. Here, the drawbar connects the rigid axle and the turntable. The articulated-frame steering is another concept in the heavy equipment sector. In this case, the vehicle bends around a joint in the middle. Both designs are not applicable for passenger cars driving at high speeds because of high steering forces, the performance of the steering kinematics, and the required packaging space (Stoll, 1992).

Single-wheel steering and axle-pivot steering are known as *Ackermann steering systems*, which can be distinguished by the way the steering force is generated (Stoll, 1992):

- *Manual Steering System*. The driver has to apply the steering-wheel force. Only the steering gear reduces the force to a manageable level.
- *Power Steering System*. The driver is assisted in the steering task by additional components reducing the required steering wheel torque.



(a) (b)

Figure 2. Steering with joysticks. (a) Cabin of a driving simulator to compare joystick driving with steering wheel driving and (b) joystick driving in a real car.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- *Full-Power Steering Equipment*. The required steering forces are provided solely by one or more energy supplies ECE Reg. 79 (2005).

One potential of steer-by-wire systems is the freedom in designing both the steering torque assistance and the steering ratio. Steering systems can also be classified by the type of force transmission between steering control and steered wheels (Stoll, 1992) as

- mechanical steering systems,
- hydraulic steering systems,
- pneumatic steering systems, and
- electrical steering systems.

Steering systems can also be distinguished by the type of steering gear:

- translational movement (rack-and-pinion steering gear),
- rotational movement (cam-and-roller (Gemmer) or ball-and-nut steering gear), and
- wheel individual steering actuators.

The rack-and-pinion steering gear converts the rotational movement of the steering wheel angle into a translational motion of the tie rod. In contrast, the second type of steering gear transmits the rotational movement of the steering column to a rotation of the pitman arm. Wheel individual steering actuators are often used in specialized on-road and off-road vehicles on the basis of a hydraulic system.

3.2 Steer-by-wire system architectures

In recent decade, many steer-by-wire system architectures have been proposed. This chapter gives an overview of the system architectures without making a claim to be exhaustive. In addition, means to improve system safety are presented.

Steer-by-wire systems do not have a mechanical linkage between the steering control (e.g., steering wheels and joysticks) and the steering system. A basic system configuration comprises a sensor for detecting the driver steering input, an electronic control unit, and an actuator for the operation of the wheels and adequate means for communication between the system elements. Integrating an additional actuator at the control element enables feedback information to the driver, depending, for example, on the steering angle of the wheels or the lateral acceleration of the vehicle. The force feedback actuator thus generates a haptic feedback, which aims at easing vehicle control. The steering actuator is linked to or replaces the conventional steering gear and operates the wheels (Wallentowitz and Reif, 2006). This system is able to provide almost all the functions described in Chapter 4 Potentials in Steer by Wire, Potential and Challenges. Splitting the steering tie rod and adding single-wheel steering actuators increases functionality again. However, in case of a fault in the E&E architecture, this extended basic configuration does not achieve the reliability of conventional steering systems. Of course, also the electric power source has to be considered in the respective analysis for system integrity. As steering systems are considered to be safety critical, the system safety has to be guaranteed in all operating conditions and over vehicle lifetime. As a one-channel electronic control unit and mechatronic subsystems do not have a failure occurrence of $<10^{-7}$ per hour (Wallentowitz and Reif, 2006), a system without redundancy would not be sufficient for safe operation. In aviation industry for example, the Joint Aviation Authorities (JAAs) have specified that failures in primary control systems, such as fly-by-wire systems of large airplanes, should have a failure occurrence of $<10^{-9}$ per hour to be considered as improbable (Joint Aviation Authorities Committee, 1989; Reichel, 2004). Thus, in order to make a steer-by-wire system as safe as a conventional steering system, the system architecture has to allow single electric and electronic faults in each of its mechatronic subsystems without leading to loss of control by the driver (Wallentowitz and Reif, 2006).

Hayama *et al*. (2008) show the development of a basic steer-by-wire to a fault-tolerant architecture without considering the electric power source and required sensors. The baseline steer-by-wire architecture (Figure 3) consists of one steering wheel angle sensor, one reaction torque actuator, one steering actuator, one tire angle sensor, two controllers, and one battery. Considering the state

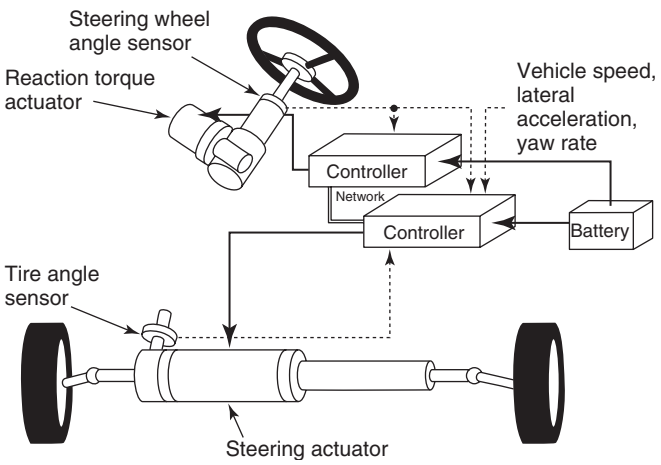


Figure 3. Baseline steer-by-wire architecture. (From Hayama *et al*., 2008. Copyright © 2008 SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

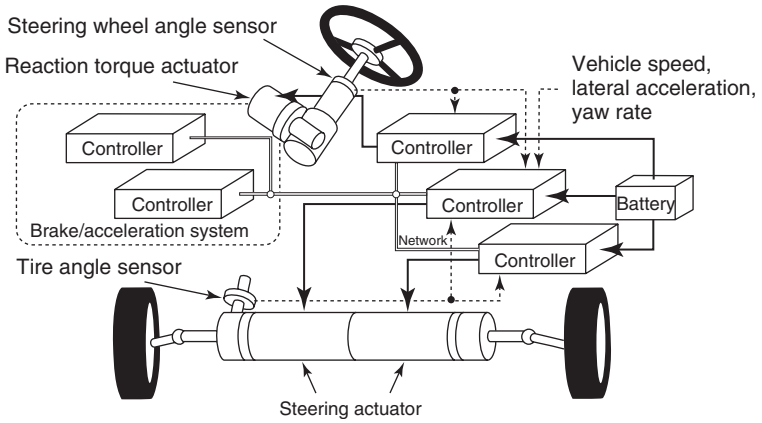


Figure 4. Integrated control steer-by-wire architecture. (From Hayama *et al*., 2008. Copyright © 2008 SAE International. Reprinted with permission.)

transition of different steer-by-wire architectures, Hayama *et al*. (2008) suggest an integrated steer-by-wire architecture (Figure 4). Compared to the baseline structure, the integrated control architecture has two redundant steering actuators and a controller. In addition, the steering system has an interface to the braking and accelerating system in order to generate a yaw momentum by an appropriate wheel torque distribution, thus supporting the driver's steering intention.

Wallentowitz and Reif (2006) have presented another fault-tolerant architecture that can be used for steer-by-wire systems. The idea is to combine two fail silent units (FSUs) to one fault-tolerant unit (FTU) in order to increase system integrity. (Fail safe means that a system is in a safe state, even in case of a failure; fail silent means that the system does not generate adverse effects, by still being active). The superior fault-tolerant architecture (FTA) consists of one sensor, controller, and actuator FTU. Moreover, the power supply is redundant. The strategy to guarantee system integrity is based on local redundancies for E/E subsystems. Figure 5 shows the E/E architecture of the proposed steer-by-wire architecture. The hand wheel and front axle actuator are designed as FTU with two motors, three sensors on one single shaft, and two electronic control units. Each control unit processes the data of two different sensors and controls one actuator. Two electronic control units representing the central electronic control unit as FTU are connected to the actuator FTUs via a fault-tolerant real-time data bus system. This bus transfers the data of the sensor signals, the actuator set point commands, and the system status.

These two examples show that there is no general solution for a steer-by-wire architecture. The design depends on many factors such as system integration, steering functionality, and vehicle properties but will always comprise some degree of redundancy, which leads to additional costs.

3.3 Current steering systems relevant for steer-by-wire

This chapter gives an overview of steering systems that provide steer-by-wire functionality to some degree and that have already been introduced to the market. So far, no true steer-by-wire system has been presented beyond concept vehicles and prototypes; thus, the presented steering systems only fulfill the extended definition of steer-by-wire, for example, with a hydraulic or electrohydraulic linkage between steering control and steering system.

In the heavy equipment sector, for example, agriculture, construction, and forestry, vehicles with steer-by-wire functionality have already penetrated the market because legal regulations allowed such steering systems for vehicles with low maximum speeds at an early time. The main technical driver for the introduction of hydraulic steering systems without mechanical fallback in this sector have been huge vehicle dimensions resulting in high steering torques also due to heavy-duty tires. Moreover, autonomous driving plays an important role in some off-road applications.

In the agriculture sector, John Deere has presented the ActiveCommand Steering (ACS™) system. The ACS steering system offers a wheel-offset control, a variable steering ratio, and torque and eliminates steering wheel drift. The system consists of a power supply, a feedback unit, two controllers, control valves, and an electric-driven backup pump. In case of a lack of oil, the latter one supplies the oil to the steering system and brakes. The feedback unit applies light feedback in field use, slightly heavier feedback in transport mode and cornering at high speed. The system is able to correct minor tire angles, for example, when forward driving, to reduce necessary driver corrections. Considering the sensor technology, the system comes with a gyroscope to measure the tractor yaw

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

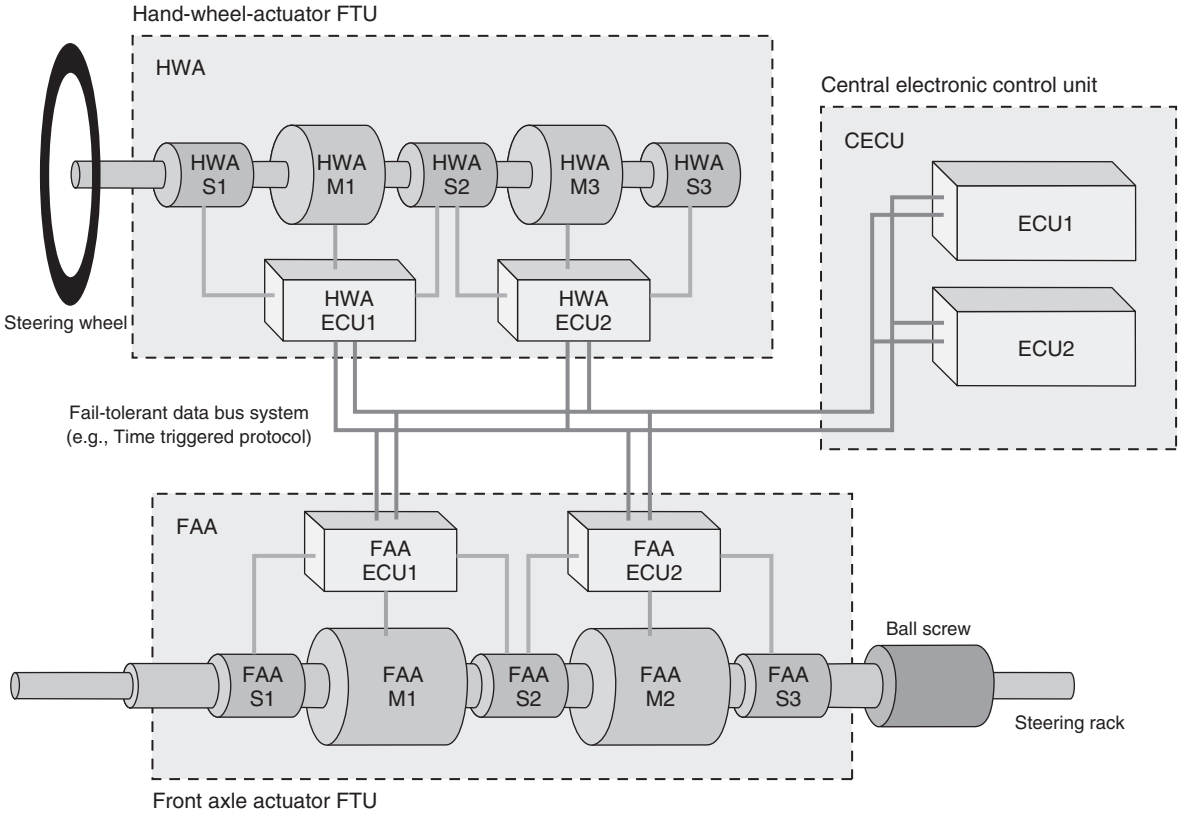


Figure 5. Fault-tolerant steer-by-wire architecture. (Reproduced from Wallentowitz and Reif, 2006. With kind permission of Springer Science+Business Media.)

rate, wheel angle sensors, and steering wheel angle sensors. The ACS is designed to fail operational. Basically, every function is backed up by another component. Thus, the primary controller is replaced by the second one in case of a fault. If the alternator power fails, the battery takes over. The electric-driven backup pump resumes control and supplies hydraulic oil to the system if the engine quits running (NN, 2012).

Other well-known examples are specialized heavy-duty vehicles for transporting, for example, production plants, which comprise several actively steered axles. In these applications, the angle and vertical travel of each wheel is controlled individually and actuated hydraulically.

In aviation technology, fly-by-wire systems are already state of the art. Besides the reduction of weight and volume, fly-by-wire systems offer the possibility of a variable computer-aided flight control. Binfet-Kull (2001) describes, among other advantages, the ability to handle the steering input feeling and the electrical transmission of pilot input signals along short to long distances within a short period of time at minimal energy. In addition, saving time-consuming adjustments of the mechanical steering system reduces assembling and maintenance costs. Binfet-Kull (2001) points out that the demands on steer-by-wire systems compared to fly-by-wire systems differ in major aspects. Examples are the danger potential, the highly trained operation personnel, and the maintenance intervals. First, if a fly-by-wire system fails, this breakdown could directly affect hundreds of human lives. Second, in comparison with normal car drivers, pilots are highly trained personnel who learned in many flight simulator hours how to react if technical faults cause worst-case scenarios. Finally, airplanes are subject to numerous inspection and maintenance intervals, the so-called letter checks, which secure its proper operation.

4 POTENTIAL

Steer-by-wire systems are able to resolve many compromises that occur during the conventional layout process of a steering system. These potentials are widely discussed throughout literature. Binfet-Kull (2001) provides a complete list of the potentials unlocked by steer-by-wire systems:

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- Free configuration of steering ratio and steering assistance
- Improvement of active and passive safeties
- Maximum potential of functionality
- Suspension design without regarding steering feedback
- Improved package situation
- Reduction of variants
- Simplification of assembly
- Introduction of new human–machine interfaces (HMIs).

These aspects can be categorized into functionality potentials, package and production potentials, and HMI potentials. These categories are discussed in detail in the following sections.

4.1 Functionality potentials

Steer-by-wire inherently allows a simple implementation of any kind of steering function. Additional hardware is not necessary and communication with other systems can improve overall vehicle functionality beyond the steering system. Several related functions of these categories such as basic steering, vehicle dynamics, advanced driver assistance, and autonomous driving are introduced in the following sections.

4.1.1 Basic steering functions

The basic functionality of a steering system can be divided into two different aspects. On the one hand, the driver needs to set the trajectory of the vehicle. On the other hand, a feedback of the current state of the steering system or the vehicle dynamics can be provided to the driver.

As in a steer-by-wire system no mechanical connection between the steering wheel and the wheels exists, traditional conflicts in the layout process can be resolved. Transmission ratio and operating force can be chosen independently and individually. Thus, the functions of a conventional steering system "variable steering assistance" and "variable steering ratio" are inherently available in a steer-by-wire system. This allows, for example, a freely designed steering feedback, which could inform the driver of a significant change of the operating force when vehicle dynamics limits are about to be reached. As the correlation of steering wheel and wheel movement depends on software only, it can be adapted depending on the actual driving situation or personal preference of the driver.

4.1.2 Vehicle dynamics control

On the basis of a steer-by-wire system, yaw-rate and sideslip angle control can be realized. As no brake forces are utilized, as would be the case in conventional electronic stability control (ESC) systems, control intervention is possible in less critical driving states and is regarded less invasive by drivers. Yih (2005) derives that, in an oversteering situation, control intervention by means of the steering system is more efficient. On the one hand, this is due to the fact that the resulting lever arms from the tire contact patch to the center of gravity are longer, thus influencing the yaw behavior of the vehicle more than a brake force. On the other hand, both tires of the front axle contribute to the stabilization when the invention is realized by the steering system.

Breuer and Bill (2006) show the potential of an active steering system when braking on a μ-split surface. In this situation, the driver of a conventional vehicle is required to compensate the yaw momentum by operating the steering wheel to stabilize the vehicle. The standard driver will likely lose control of the vehicle. To reduce the yaw momentum, the brake force on the μ-high side can be reduced by the antilock braking system (ABS) leading to a longer braking distance. By controlling the steering angle, the yaw momentum can be compensated and maximum possible traction can be used to obtain the minimum braking distance. Results from the EU-project PEIT (powertrain equipped with intelligent technology) (Maisch *et al.*, 2005) show that these advantages can also be transferred to commercial vehicles.

In order to fully unlock the vehicle dynamics potential of a steer-by-wire system, it is necessary to integrate the steering control system into a global chassis control (GCC) system (Semmler and Rieth, 2004; Krüger, Pruckner, and Knobel, 2010). With this holistic approach, all degrees of freedom of the integrated chassis control systems can be used to obtain maximum driving dynamics performance. Consequently, the steer-by-wire system can provide a change in yaw momentum in any driving situation in combination with other chassis systems, not only in critical driving situations, but also during normal driving, for example, in order to compensate crosswind or road disturbances. Installing a more complex steer-by-wire system, for example, with individual wheel steering actuation, it is possible to extend the limits even further, as the friction potential at every wheel can be used in an optimal manner (Eckstein, 2012).

4.1.3 Advanced driver assistance systems (ADAS)

Steer-by-wire systems provide the best possible prerequisites to integrate the steering system into advanced driver assistance system (ADAS). Possible applications include functions such as lane-keeping and lane-change assistance and reverse driving trailer assistance. Another

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

possible feature is (partly) autonomous parking, allowing to smoothly maneuvering the equipped vehicle hands-free into a parking space. Compared to conventional approaches to introduce these functions using an electromechanical steering or a superimposed steering system, steer-by-wire systems allow a realization without steering wheel movement or torques that need to be supported by the driver. Thus, also autonomous driving as demonstrated by Deutschle (2006) can be implemented. Steer-by-wire systems are able to completely fulfill any requirement of autonomous driving as the steering actuation is designed to perform any steering maneuver.

4.2 Package and production potentials

Two main package advantages result from the absence of a mechanical connection between steering wheel and the wheels. First of all, the absence of a steering column allows much better space utilization in the engine compartment as no specific packages have to be considered for right- or left-hand drive. In addition, during a frontal crash, there is less likelihood that the impact will force the steering wheel to intrude into the driver's survival space (Yih, 2005). In case of individually actuated steering of the wheels, the space usually occupied by the lateral connection, for example, by the rack and pinion steering gear is vacant as well.

Secondly, the entire steering mechanism can be designed and installed as a modular unit, thus leading to higher production volumes and easier assembly. Especially for commercial vehicles, the assembly process can be revolutionized with the introduction of steer-by-wire. Today, the chassis and the driver's cab are integrated at a very early stage because of the difficult assembly of the multipart steering column. Steer-by-wire would permit to assemble the chassis and the driver's cab independently, joining them at the very end of the production line. This would save significant space in the production plant and allow a more efficient task sharing.

Thirdly, steer-by-wire would allow an integrated cockpit concept including the steering device. Apart from advantages in production, this concept allows a significantly larger spread of vehicle derivatives especially with respect to the seating position and dashboard height. This is due to the fact that the mechanical steering column determines the angle of the steering wheel, which has a large effect on the driver's posture and is thus an important limiting factor.

4.3 HMI potentials

As, in a steer-by-wire system, operating forces are not directly transferred from the operating element to the wheel, the steering ratio between interface and wheel and the amplification of steering torque can be designed without any restrictions. In literature, this feature is sometimes referred to as *full steer-by-wire functionality*.

The benefit of freely designing the steering transmission ratio is the possibility to reduce the steering effort at low speeds and tight turns or while parking as well as to ensure a precise and stable control at high velocities. In contrast to a superimposed steering system, the spread in steering ratio is not limited by the fact that, in case of a failure, the ratio step must not exceed a value of about two points (e.g., from 17 to 19) in order to maintain sufficient controllability (Freitag *et al*., 2001). In addition, without a direct mechanical connection between the steering wheel and the road wheels, noise, vibration, and harshness (NVH) from the road no longer have a direct path to the driver's hands and arms through the steering wheel (Yih, 2005), thus increasing driving comfort.

Owing to the lack of necessity to apply the complete steering forces manually in case of a disabled power steering system, it is possible to introduce new and innovative steering controls to steer the vehicle. Winner and Heuss (2005) conclude that the use of a steering wheel is caused by the historical development of motor vehicles and does not necessarily need to be continued when steer-by-wire systems are introduced. Among many options, Winner believes that the most likely successful option would be an active joystick, that is, a force actuated stick that is located at both sides of the driver and provides an appropriate steering feedback. As early as 2001, Eckstein shows that this operating concept is suitable for controlling a passenger car (Figure 2). His investigations in a motion-based driving simulator show that novice drivers are able to learn to operate a vehicle with sidesticks just as well as a vehicle with a conventional steering wheel and pedals.

Another option, which arises, is the possibility to allow vehicle control from several positions inside or even remote control from outside a vehicle. This is especially of interest in special-purpose vehicles and heavy-duty trucks and already state of the art.

Winner and Hakuli (2006) introduce a consistent advancement of the steer-by-wire concept. Their paradigm of conduct-by-wire removes the driver from the vehicle stability control loop leaving the responsibility to command the vehicle on the guidance level. The vehicle will be controlled by passing maneuver commands to the system. The actuation of the individual vehicle variables, that is, driving and braking torques as well as steering angles, is determined by a centralized controller and executed autonomously. The responsibility to monitor proper operation of the vehicle remains with the driver. To implement this innovative driving paradigm, a complete x-by-wire

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

architecture of the vehicle is required as well as a very complex sensor concept to capture the environment. The conduct-by-wire approach merges today's ADASs and vehicle stabilization systems into one complex vehicle conducting system.

5 CHALLENGES

Among the powerful potentials that can be realized by deployment of a steer-by-wire system, many challenges originate from it as well. Some of them are discussed in the following sections.

Two main aspects that have to be mentioned are reliability and safety. Besides these technical challenges, perception and acceptance by customers also play an outstanding role for possible success of this technology on the market. Acceptance also depends very much on costs.

In addition to technological and economic challenges, *legal aspects* also need to be addressed. While the relationship among the authorities, the manufacturers, and the driver is not affected in the first place, a clear definition of responsibilities between the vehicle manufacturer and the supplier of a steer-by-wire system is decisive and will also affect the development process and the system layout itself. The legal aspect is discussed first in the following section.

5.1 Legal requirements for steer-by-wire systems

Legal requirements such as the ECE reg. 79 (2005) define rules regarding the approval of steering equipment for vehicles. Revision 2 of the ECE reg. 79 (2005) from 2005 considers the advancements in steering technology and the advantages of steer-by-wire systems compared to steering systems having a mechanical link. Thus, it is nowadays possible to approve steer-by-wire systems having no mechanical connection between wheels and steering control.

According to the definition of ECE reg. 79 (2005), steer-by-wire systems belong to the steering equipment classification of full-power steering equipment. Solely one or more energy supplies provide, that is, the steering forces.

The ECE reg. 79 (2005) provides additional requirements that full-power steering systems need to fulfill. Regarding the system design, the regulation specifies different failure provisions:

- A failure in the transmission that is not purely mechanical has to be indicated to the vehicle driver via predefined warning signals. In this failure mode, a change in the average steering ratio is allowed only if the steering effort is not exceeding a predefined maximum effort value of 300 or 450 N over a period of time of 4 or 6 s, both values depending on the vehicle category.
- In the event of a failure in the energy source of the control transmission, the vehicle having an energy storage level at which the failure was signaled to the driver should be able to drive at least 24 "figure of eight" maneuvers with a loop diameter of 40 m at a velocity of 10 km/h with the same performance level given for an intact system.
- If a failure in the energy transmission occurs, with the exception of the parts not liable to breakage (e.g., steering column), no immediate change in the steering angle is permissible. A vehicle, still able to drive faster than 10 km/h, has to fulfill all test provisions after having finished at least 25 "figure of eight" maneuvers with a 40 m diameter at 10 km/h minimum speed. The energy storage level at the beginning of the test maneuver has to be the same as in the failure of the energy source of the control transmission.

The development process of steer-by-wire systems has to be in accordance with relevant standards, such as the ISO 26262, which address the functional safety of road vehicles. According to the ECE reg. 79 (2005), the manufacturer of complex electronic vehicle control systems has to accomplish specific requirements for documentation, fault strategy, and verification with respect to the safety aspects. Concerning the fault strategy, the design process has to guarantee the safe operation of the vehicle fail-safe procedures, required redundancies or necessary driver warning systems. ECE reg. 79 (2005) gives examples for design provisions for a system failure, which are for example

- fallback to operation using a partial system,
- change over to a separate back-up system, and
- removal of high level functions.

In summary, for approval of such a safety-critical system, the manufacturer has to provide a complete documentation of the transparent design process and on the system means to guarantee its safe operation.

5.2 Technical challenges

From legal requirements, customer demands, and OEM (Original Equipment Manufacturer) design goals, a complex list of technical challenges results. Accordingly, some of the most important challenges include the provision of functional safety, acceptable steering feel, and reducing costs and weight.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5.2.1 Safety

While mechanical systems are considered inherently safe, as a mechanical failure can be detected by inspection and is unlikely to fail in a sudden manner when properly designed and used within its specification boundaries, a mechatronic system can fail without prior signs of damage or wear. Thus, great care has to be taken designing a system that ensures a safe operation as well as acceptable reliability.

To achieve a safe state of a vehicle when a system failure occurs, it is obviously sufficient if the vehicle is stopped at a suitable location, for example, the emergency lane of a highway. To reach this safe state, it is important that any fault does not lead to an instable reaction of the vehicle and that the vehicle remains controllable by the driver during the emergency maneuver. This means that

1. the initial effect of a fault must be limited,
2. the cause of the fault is detected and addressed within a finite time (typically some milliseconds), and
3. the resulting vehicle behavior can be controlled safely by the driver.

Customer requirements by far exceed the demand for a safe system according to the earlier definition. System reliability plays an important role in terms of availability of all system and vehicle functions, as it greatly affects customer acceptance and warranty costs for the OEM. Depending on the likeliness of a fault occurrence and the presumed consequence, the developer needs to decide, which countermeasures need to be taken to reach a system reliability that guarantees sufficient availability.

A comparison to current solutions in the aeronautical industry shows that the general approach is to maintain control over the vehicle with no or only manageable interruptions. This can be achieved by either introducing redundancies that are able to fulfill the original functionality without limitations or time limited, thus providing a fail-operational behavior. Examples for redundancies are duplicated bus systems to ensure data transmission or additional batteries that allow further operation of the system for a limited time in case the main power supply should fail.

Besides the deployment of redundancies, a restricted functionality resulting from a failure of a subsystem or a combination of different faults is often commonly used in aeronautics. For example, it is possible to operate an airplane using only some of the available control surfaces. The plane is still controllable although some maneuvers might not be possible. This state of the system with reduced functionality is referred to as a *degraded state*.

Within the automobile industry, the concept of degraded states is used as well. ESC systems will be deactivated in case of a detected fault, whereas the basic hydraulic brake system is still in operation. Similarly, steering assistance by a power steering system is deactivated when a fault occurs. Both breakdown situations are accepted as a safe state although functionality and controllability are considerably reduced.

Proposals for steer-by-wire system architectures including suggestions for redundancies and degraded states can be found in literature (Binfet-Kull, 2001; Freitag *et al*., 2001; Heitzer, 2003; Wallentowitz and Reif, 2006; Hayama *et al*., 2008). To achieve a robust vehicle behavior, fault and event detection is a key factor. In this respect, the analogy with aviation may not be valid, as the required system reaction time in a vehicle driving on the road, for example, to prevent an unintended lane departure, is usually far shorter than those of an airplane.

5.2.2 Steering feel

As for any subjective assessment of a technical system when evaluating steering systems, different approaches and criteria are used depending on brand, vehicle class, or regional preferences. When new technologies are introduced, usually their impact and acceptance are discussed controversially. When, for example, EPS was first implemented in passenger cars, the fact that the steering feel differs from conventional hydraulic power steering systems led to criticism of the innovative system for its synthetic steering feel. As EPS[1] offers several additional innovative functions such as lane keeping, automatical[2] parking, side wind compensation, and even more, it was not surprising that EPS systems significantly penetrated the market. On the one hand, this was certainly due to an improved parameterization of the system functions and a reduction in fuel consumption; on the other hand, it was also successful due to a certain habituation effect by the drivers.

Regarding steer-by-wire systems, it is evident that steering feel needs to be provided completely on a synthetic basis, which may compose of an electromechanical steering feedback actuator as well as passive elements such as springs and dampers. The passive elements can assure the stability of the dynamic behavior of the steering wheel system itself in case of a fault of the feedback actuator. Using all degrees of freedom in designing steering feel dependent on, for example, vehicle speed, lateral acceleration, road friction, or lane keeping, a considerable amount of effort is necessary to achieve an adequate steering feel in all relevant driving states (Koch, 2010). On the other hand, the freedom in designing steering feel must be regarded as a big advantage of a steer-by-wire system, as a specific steering feel can easily be provided for different vehicle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

models and markets including parameters for individualization and customization.

5.2.3 *Costs and weight*

From an OEM perspective, using the same steer-by-wire system across many vehicle models and markets yields cost savings because of economies of scale and reduction of variants. At least part of these cost savings need to be invested in compensating higher system costs. These can be expected to be higher even compared to a superimposed steering system, as additional components are needed to provide steering feel and redundancy of the steering actuator. The precise effect of this change in technology on costs, weight, and efficiency largely depends on the specific system layout and thus cannot be quantified in general. Aspects such as the number, location and performance of actuators, and the type of driver interface will be most influential. Barthenheier (2002) expects that it will be possible to reduce cost and weight because of the fact that many parts can be removed from the system. Other sources (Fleck, 2003; Grell, 2003) state that no positive effects are to be expected as the effort to achieve a safe system with redundancies will overcompensate the advantages. In the end, the success of *steer-by-wire systems* will largely depend on whether or not the customer has to expect additional costz for the system. This refers to purchase as well as to operating and maintenance costs.

Regarding the development cost for steer-by-wire systems, as for any complex technical innovation, the initial research investment is very high. If steer-by-wire systems are introduced in a small segment such as luxury vehicles, development costs are hardly acceptable (Winner *et al*., 2004). If, on the other hand, steer-by-wire would be widely introduced in mass production, the risk from an economic and marketing point of view is high if a large number of vehicles need to be recalled in case of quality problems.

5.3 Customer acceptance

When steer-by-wire systems are introduced to the automotive market, they will have to compete with current modern steering systems. As described earlier, steer-by-wire systems do have significant advantages regarding possible steering functions, although the advantages might not be immediately evident for the average driver. The main marketing challenge for the successful introduction of steer-by-wire systems will be to clearly communicate the advantages of the new and innovative system.

Whether this will be possible is not assessable from today's perspective, as customer acceptance is dependent on many influences and is hardly ascertainable in an early stage of the development process as Binfet-Kull (2001) states. Daniels (2003) expects that a major challenge will be skepticism toward the safety of a steering system lacking a mechanical connection between the steering wheel and wheels. The mechanical system has proved its reliability in recent decades and in general customer perception it is regarded as safe, whereas many drivers have already experienced failures of electrical systems—whether in their car or using consumer electronics.

The first steer-by-wire systems thus need to comprise a very convincing safety concept, which is highly effective and reliable. Communication should focus on evident benefits of such a system and ideally avoid the term *by-wire* in order to generate a positive perception of this promising innovation.

6 SUMMARY

While substantial research on steer-by-wire systems has been carried out over the past decades, this innovative approach has not made its way into volume production. Active steering systems superimposing an additional steering angle to the driver's command can be regarded as forerunner of steer-by-wire, as their performance would be sufficient to influence lateral vehicle dynamics under normal driving conditions while holding the steering wheel straight. In contrast to steer-by-wire, the torque provided by the steering actuator needs to be counterbalanced by the driver, and in case of a system malfunction, the mechanical steering column represents a well-known fallback.

While active steering and electric-powered steering systems already enable innovative functionality regarding vehicle stabilization and advanced driver assistance, full steer-by-wire systems offer three areas of potential: firstly, 100% freedom in designing the functional relationship between the operating device and the steered wheels, secondly more flexibility in package and production yielding time and cost savings, and finally maximum design freedom regarding the driver's working place.

On the other hand, the market introduction of steer-by-wire systems faces some major challenges: apart from legal requirements, the main challenge concerns the provision of functional safety while at the same time limiting the costs due to resulting requirements, for example, on redundancies of sensors, communication, and actuators. Finally, customer acceptance is the prerequisite for market success—the benefit of the introduction of steer-by-wire has to be understood immediately providing a unique driving experience.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

REFERENCES

Alexandre, H. (1894) Voitures Automobiles. L'ingénieur Civil, September 15, 1894.

Barthenheier, T. (2002). *Steer-by-Wire-Systeme – Stand und Entwicklungsaussichten*. 47. Internationales Wissenschaftliches Kolloquium, X-by-Wire-Workshop, September 2002, Ilmenau.

Bayle, R. and Ecquevilly, Y. (1972) *Servomechanismus*, Patent DE 2237166

Bidwell, J.B. and Cataldo, R.S. (1958) Single Stick Member for Controlling the Operation of Motor Vehicles. US patent 3 022 850.

Bidwell, J.B. (1959) *Vehicles and drivers—1980*. Presentation at the SAE Annual Meeting, Sheraton-Cadillac and Statler Hotels, Detroit, Michigan, January 12–16.

Binfet-Kull, M. (2001) *Entwicklung einer Steer-by-Wire Architektur nach zuverlässigkeits- und sicherheitstechnischen Vorgaben*, Verlag Mainz, Mainz.

Bränneby, P., Palmgren, B., Isaksson, A., *et al*. (1991). Improved Active and Passive Safety by Using Active Lateral Dynamic Control and an Unconventional Steering Unit. *13th International Technical Conference on Experimental Safety Vehicles, Proceedings 1*, Paris, pp. 224–230.

Braess, H.-H. (2001) Lenkung und Lenkverhalten von Personenkraftwagen. Was haben die letzten 50 Jahre gebracht, was kann und muss noch getan werden? *VDI-Berichte*, **1632**, 13–55.

Breuer, B. and Bill, K.H. (2006) *Bremsenhandbuch*, Friedr. Vieweg & Sohn Verlag/GWV Fachverlage GmbH, Wiesbaden.

Brockhaus, R., Alles, W., and Luckner, R. (2011) *Flugregelung*, Springer-Verlag, Berlin Heidelberg.

Broggi, A., Bombini, L., and Cattani, S., *et al*. (2010) *Sensing requirements for a 13,000 km intercontinental autonomous drive*. 2010 IEEE Intelligent Vehicles Symposium, University of California, San Diego, CA, June 21–24, 2010.

Daniels, J. (2003) Hidden wires *European Automotive Design*, **2003**, 29–31.

Davis, F.W. (1927) *Hydraulic Steering Mechanism*, Patent US 1 790 620.

Davis, M.W.R. (2004) *General Motors a Photographic History*, Arcadia Publishing, Charleston SC.

Deutschle, S. (2006) The KONVOI Project—Development and Evaluation of Electronic Truck Platoons on Highways. *Proceedings of 15th Aachen Colloquium "Automobile and Engine Technology" 2006*, Aachen.

Dick, R. (2004) *Mercedes and Auto Racing in the Belle Epoque, 1895–1915*, Mcfarland & Co Inc, Jefferson, NC.

ECE R79 (2005) Regulation No. 79—Rev. 2—Steering Equipment, http://live.unece.org/fileadmin/DAM/trans/main/wp29/wp29regs/r079r2e.pdf (accessed 25 November 2007).

Eckstein, L. (2000) Sidesticks im Kraftfahrzeug – ein alternatives Bedienkonzept oder Spielerei? Ergonomie und Verkehrssicherheit. *GfA Konferenzbeiträge der Herbstkonferenz 2000*, 12.-13. Oktober 2000 an der Technischen Universität München, Herbert Utz Verlag, München.

Eckstein, L. (2001) *Entwicklung und Überprüfung eines Bedienkonzepts und von Algorithmen zum Fahren eines Kraftfahrzeugs mit aktiven Sidesticks*, VDI Verlag GmbH, Düsseldorf.

Eckstein, L. (2012) *Vertical and Lateral Dynamics of Vehicles*, Forschungsgesellschaft Kraftfahrwesen Aachen mbH, Aachen.

Fenton, R.E. (1994) IVHS/AHS: driving into the future *Control Systems, IEEE*, **14** (6), 13–20.

Fleck, R. (2003) Systematic Development of Mechatronic Steering Systems with Steer-by-Wire Functionality. *Proceedings of fahrwerk.tech 2003*, München.

Freitag, R., Moser, M., Hartl, M., *et al*. (2001) Safety Concept Requirements of Steering Systems with Steer-by-Wire Functionality. *Proceedings of* "*Internationale Tagung Elektronik im Kraftfahrzeug*", VDI-Berichte Nr. 1646, 2001. VDI Verlag GmbH, Düsseldorf.

Grell, D. (2003) *Innovationslawine in der Autotechnik. C't – Magazin für Computertechnik, 14/2003*, Heise Zeitschriften Verlag GmbH & Co. KG, Hannover.

Hayama, R., Higashi, M., Kawahara, S., *et al*. (2008) Fault-tolerant architecture of yaw moment management with steer-by-wire, active braking and driving-torque distribution integrated control. SAE World Congress & Exhibition, April 2008, Detroit, MI. Session: Safety-Critical Systems (Part 2 of 3).

Heitzer, H. (2003) Entwicklung eines fehlertoleranten Steer-by-Wire Lenksystems. *Tagung* "*PKW-Lenksysteme*", Haus der Technik, Essen.

Henke, R. (2010) Vorlesung Flugzeugbau II (SS2010) Version 2.0. Thema 8: Fly-by-Wire.

Joint Aviation Authorities Committee (1989) Joint Aviation Requirements, JAR 25, Large Aeroplanes, ACJ No. 1 to JAR 25.1309, http://www.jaa.nl/publications/crd/jar-25-change13.pdf (accessed 25 November 2007).

Kramer, F. (2009) *Passive Sicherheit von Kraftfahrzeugen*, Vieweg + Teubner Verlag, Wiesbaden.

Krüger, J., Pruckner A., and Knobel C. (2010) Control Allocation for Road Vehicles—A System-Independent Approach for Integrated Vehicle Control. *Proceedings of 19th Aachen Colloquium "Automobile and Engine Technology" 2010*, Aachen.

Koch, T. (2010) Untersuchungen zum Lenkgefühl von Steer-by-Wire Lenksystemen. Dissertation. TU München, München.

Köhn, P., Baumgarten, G., Richter, T., *et al*. (2002) Active Steering—The BMW Approach to Modern Steering Technology. *Proceedings of 11th Aachen Colloquium "Automobile and Engine Technology", Band 2*, Aachen, 7.-9. Okt.

Linderer, W. (1951) *Einrichtung zum Schutze von in Fahrzeugen befindlichen Personen gegen Verletzungen bei Zusammenstößen*, Patent DE 896 312.

Maisch, A., Kandar, T. and Arpad, M. *et al*. (2005) PEIT: powertrain equipped with intelligent technologies. Deliverable D15, Final Report.

NN (2003) *Annual Report 2003 Toyoda Machine Works, LTD*, http://www.jtekt.co.jp/ir/pdf/ar_2003.pdf (accessed 25 November 2007).

NN (2012). *John Deere Homepage—Presentation of Active Command Steering*, http://www.deere.com/wps/dcom/en_US/products/equipment/tractors/row_crop_tractors/8r_8rt_series/8260r/8260r.page (accessed 25 November 2007).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Patzelt, H., Schiesterl, G. and Seybold, A. (1971) *Schutzvorrichtung, insbesondere für die Insassen von Kraftfahrzeugen*, Patent DE 2152902.

Pfeffer, P. (2011) *Lenkungshandbuch*, Vieweg + Teubner Verlag, Wiesbaden.

Pilon, H. M., Park, A., Sattavara, S. W., *et al*. (1972) *Power Steering Gear Actuator* Patent US 3831701.

Reichel, R., Armbruster, M. and Bäuerle, K. (2004) Gemeinsame Systemstrukturen im Flugzeug-und Automotiv-Bereich für sicherheitskritische Steuerfunktionen *5. Braunschweiger Symposium Automatisierungs- und Assistenzsysteme für Transportmittel*, Braunschweig, 200–220.

Rixmann, W. (1962) Die Daimler-Benz Servolenkung *Automobiltechnische Zeitschrift*, Jahrgang 64, Heft 1, 1–9.

Semmler, S.J. and Rieth, P.E. (2004) Global Chassis Control—The Networked Chassis. *Proceedings of 13th Aachen Colloquium "Automobile and Engine Technology" 2004*, Aachen.

Stiller, C. (2007) Autonome Mobile Systeme *Informatik aktuell*, Part 6, 163–170.

Sun, Z., Bebis, G., and Miller, R. (2006) On-road vehicle detection: a review *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28** (5), 694–711.

Stoll, H. (1992) *Fahrwerktechnik: Lenkanlagen und Hilfskraftlenkungen*, Vogel Verlag, Würzburg.

Wallentowitz, H. and Reif, K. (2006) *Handbuch Kraftfahrzeugelektronik. Grundlagen, Komponenten, Systeme, Anwendungen*, Friedr. Vieweg & Sohn Verlag/GWV Fachverlage GmbH, Wiesbaden.

Winner, H., Isermann, R., Hanselka, H., and Schürr, A. (2004). When Does By-Wire Arrives Brakes and Steering? *Proceedings of AUTOREG 2004*, VDI Berichte 1828, VDI Verlag GmbH, Düsseldorf.

Winner, H.; Heuss, O. (2005). X-by-Wire Betätigungselemente – Überblick und Ausblick. *Darmstädter Kolloquium Mensch & Fahrzeug*, Darmstadt. 08.+09. März, ISBN:3-935089-83-X.

Winner, H. and Hakuli, S. (2006). Conduct-by-Wire – Following a New Paradigm for Driving into the Future. *Proceedings of FISITA World Automotive Congress*, 22.-27. Oktober 2006 in Yokohama, Japan.

Yih, P. (2005) Steer-by-wire: implications for vehicle handling and safety. Dissertation. Stanford University, Stanford.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto010
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Performance Target Conflicts in Normal Tires and Ultrahigh Performance Tires

Reinhard Mundl and Burkhard Wies
Technische Universität Wien, Vienna, Austria
Continental, Hannover, Germany

1 Introduction 1
2 Product Development Strategy by Enhancing Target Conflicts 1
3 Systematic Identification of Target Conflicts by Multiple Performance Analysis 2
4 Design Potential of Main Tire Domains and Related Target Conflicts 3
5 Target Conflicts of UHP Tires 8
6 Outlook for Future Improvements of UHP Tires 10
References 10

1 INTRODUCTION

Few people are aware of the fact that we are surrounded in our daily life almost everywhere by tires. Although they are often perceived as commodity, we all rely strongly on the high quality of their performance. Beside the very basic safety-related tire properties of durability and robustness, there is a bundle of the so-called additional tire performances such as grip on all surfaces, mechanical and acoustic comfort, safe handling properties, and, last but not least, a long lifetime. In recent times—also driven by legislation such as the new EU tire label—additionally, low rolling resistance has become important as a means to reduce fuel costs and climate change.

Unfortunately, these entire tire properties cannot be improved simultaneously. Each change in the tire's design that leads to an improvement in one specific tire property often leads to a trade-off in another one. It is now the demanding job of the tire engineer to come up with a well-balanced optimum for all the tire performances to satisfy the needs of the customers. Intensive research in tire technology and a deep know-how is needed to identify innovations to shift conflicting tire properties to a higher level. This chapter describes in the following these conflicts and provides the underlying rubber physics and tire mechanics.

2 PRODUCT DEVELOPMENT STRATEGY BY ENHANCING TARGET CONFLICTS

At the very beginning of the development of a new tire line stands the definition of a performance requirement book, which is based on competition analysis, marketing needs, and customer—more specifically, original equipment—requirements. This "book" covers tire performances, which are in most cases listed in relation to a reference tire. The differences are quantified by the relative deviation in percentage. More than a dozen of these tire performances define the strengths and weaknesses, which represent the individual character of a tire line on the market.

As an example, see the multidimensional presentation of tire competitors in a simplified, the so-called, spider graph in Figure 1. It shows on each ray the respective relative

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

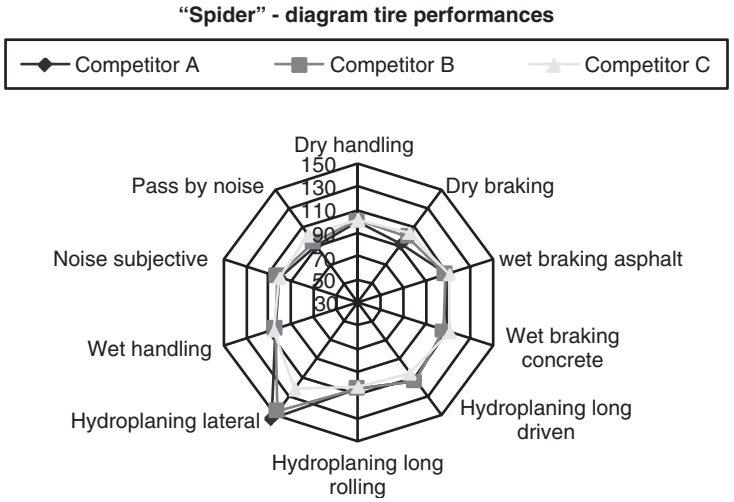


Figure 1. Multidimensional presentation of selected tire performances.

tire performance, thus delivering a specific picture of each competitor. The improvement of a single tire performance conflicts often with another one, making it very difficult for the tire developer to find a new optimum for all of them on a higher level. This sort of multitasking challenge can be easier overseen if one tries to achieve an improvement within a so-called performance conflict, neglecting the interactions to other performances. Owing to tire physics, conventional design changes of specific tire components cause only a shift along a so-called performance conflict line, thus improving only one performance but diminishing the conflicting one. This sort of adaption to changing targets is state of the art for an experienced tire designer, whereas innovative measures improve both performances at once and lead to a higher level perpendicular to the conflict line (Figure 2).

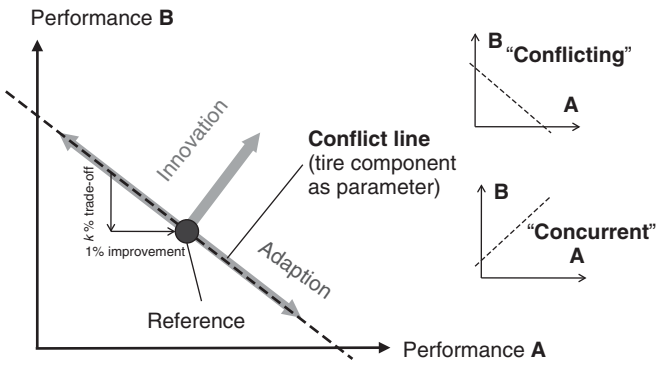


Figure 2. Schematic tire performance conflict diagram, distinguishing oriented measures of adaption and innovation and classification of dependencies.

3 SYSTEMATIC IDENTIFICATION OF TARGET CONFLICTS BY MULTIPLE PERFORMANCE ANALYSIS

From a mathematical viewpoint, if one has a number of n tire performances under investigation, a number of $n(n-1)/2$ target conflicts are possible. Regardless of the underlying physics, one can analyze the results of basic tire programs, where the influence of the change of a specific design parameter on a set of tire performances has been tested. This approach allows for rules of thumb to quantify the existing target conflicts. In particular, this leads to a ratio describing the decrease of performance B of $k\%$, when performance A is increased by 1%. In terms of mathematics, this ratio stands for the inclination k of the so-called conflict line to be seen in Figure 2. This conflict line can be extracted from the test results by applying linear regression analysis and is equivalent to the regression line.

Making these regressions, we can distinguish pairs of tire performances, which are *conflicting*, meaning that the inclination k is negative, and pairs of tire performances, which are *concurrent*, defined by a positive inclination, see again the visualization in Figure 2. It is noteworthy to mention that the described conflicting behavior has always an underlying change in a specific design parameter that can be positioned on the conflict line.

This sort of data mining has been done by Continental utilizing test results collected by the winter tire predevelopment department over almost one decade to get an overview on existing target conflicts for tread pattern variations (Figure 3). As mentioned at the beginning of this chapter, from the 13 tire performances investigated, 78 dependencies

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

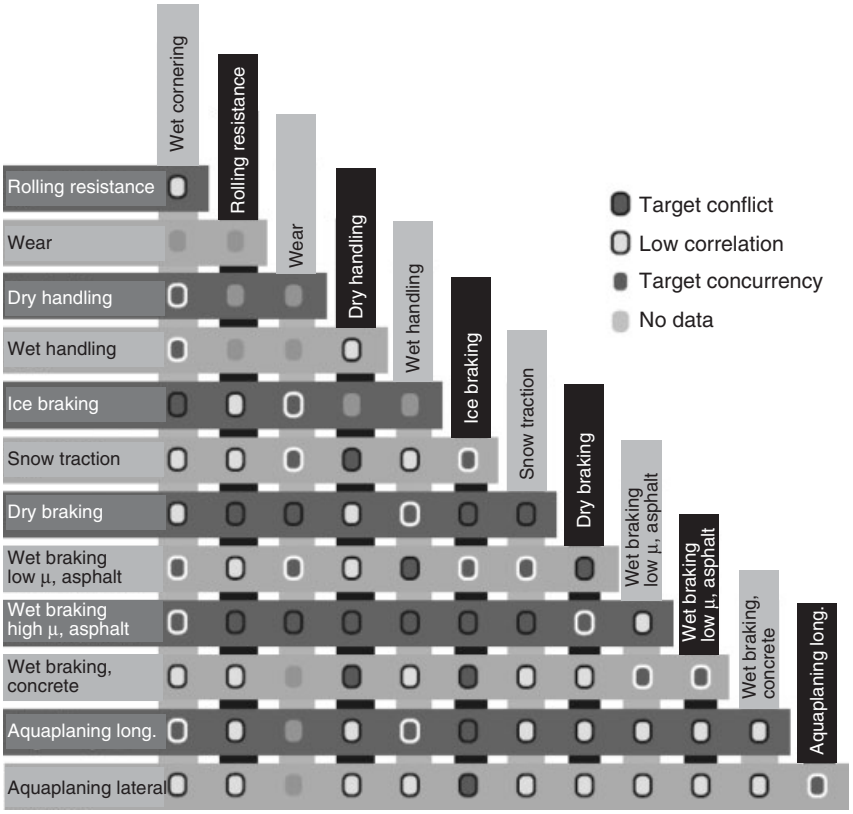


Figure 3. Overview of tire performance conflicts/concurrencies for winter tires.

were analyzed, out of which 18 are conflicting. An example for such a (pattern-inherent) performance target conflict, which can be explained by tread pattern mechanics, will be given in the following chapter.

4 DESIGN POTENTIAL OF MAIN TIRE DOMAINS AND RELATED TARGET CONFLICTS

Within a tire design, one can distinguish four main domains with significant influence on tire performances, the tread pattern, its compound, the outer shape of the cross section, named contour, and the tire carcass including the belts and cap plies (the latter domain herein summarized as construction). A qualitative estimation of the performance potential of design changes within these domains can be found in Heißing and Ersoy (2007) and is to be seen more detailed in Figure 4.

4.1 Tread pattern

From Figure 4, it can be seen that the tread pattern most influences grip on different surfaces and noise. Especially with media on the road, such as water, snow, or ice, the edges within a pattern, from blocks and tire sipes, contribute significantly to tire grip. In a large basic program, published in Doporto *et al*. (2003), a series of patterns ranging from ultrahigh performance (UHP) summer tires to highly siped Scandinavian tires have been tested in their interactions with a similar wide range of tread compounds.

From these results, we take as an example the dependency of wet and dry braking on the density of edges in the patterns structure (Figure 5). The rule of thumb for this *pattern-based conflict between wet and dry braking* consists in the same amount of trade-off for dry braking when one tries to improve wet braking by adding sipes to a pattern structure. This conflict can be explained by different frictional effects of the rubber edges: on a wet surface, the leading edges improve braking by wiping away the lubricating water film; on a dry surface, the trailing edges lift off because of higher frictional forces thus reducing the local contact area and in consequence dry braking.

For the tread *compound*, one sees a *concurrency between wet and dry braking*, where the improvements in dry braking are about only a third of the ones in wet braking. Both performances are concurrent because they are both

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

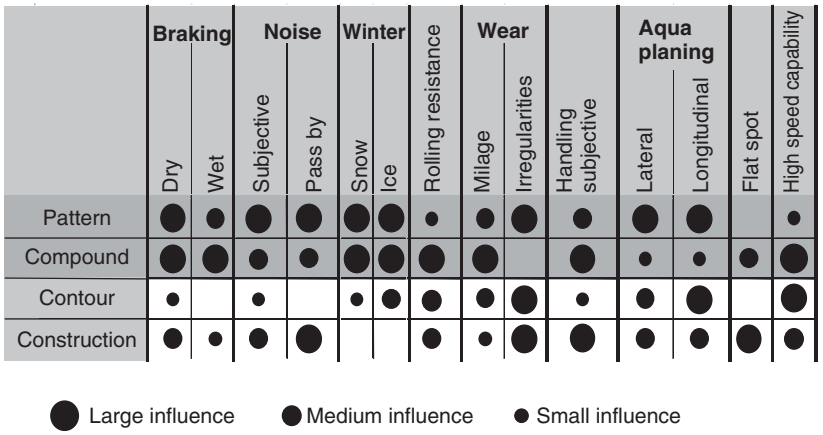


Figure 4. Design influence potential of tire domains on tire performances. (Reproduced from Heißing, 2007. With kind permission of Springer Science+Business Media.)

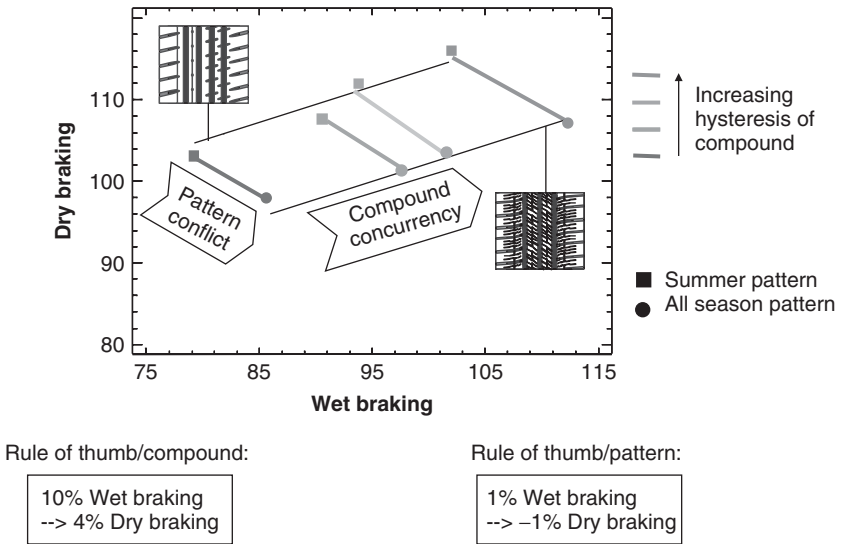


Figure 5. Example for tread induced performance conflict/concurrency and related rules of thumb.

caused by the traction physics of the compound's hysteresis. Its larger contribution to wet braking might be explained by a larger spread in hysteretic compound behavior at lower local contact temperatures because of the cooling effect of the intermediate water film.

Another extensive basic study, investigating the effect of pattern void on all relevant tire performances, was published in Mundl, Roeger, and Wies (2009). In Figure 6, one can see all major and minor conflicts/concurrencies visualized in a kind of balance: conflicting performances can be found on opposite sides of the balance and concurring performances on the same side. The lengths of the levers indicate the change in a specific property while increasing the void of the tire pattern by 1.5%. By void, one understands the empty space in a pattern realized dominantly by grooves in relation to the tread volume of the respective smooth tire. Opposing levers of a pair of properties signal a performance conflict and quantify the related rule of thumb by the levers' lengths.

The main conflict to be influenced by *void* is the one between *hydroplaning and tire noise*. Increasing of void delivers more space for water for drainage, thus increasing the critical speed at lift off of the tire's contact patch by a water wedge. By contrast, when larger grooves contact the road surface, more air volume is stimulated to vibrate and to emit noise. To a lesser extent, we encounter also the above-described conflict between braking on wet surfaces and dry ones. Less known is a conflict for winter

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5
www.it-ebooks.info

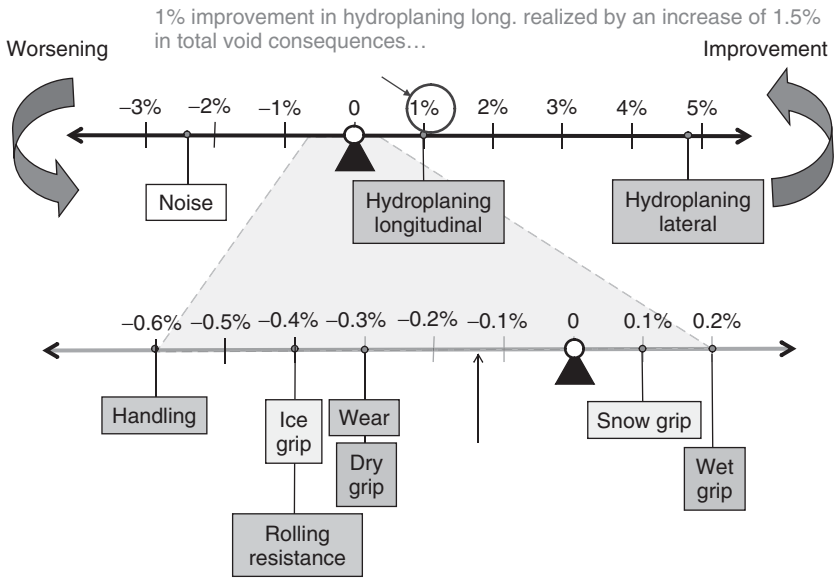


Figure 6. Pattern void based quantified conflicts/concurrencies. (Reproduced from Mundl, Roeger and Wies, 2009. © The Tire Society.)

tires, namely between traction on snow and ice. Owing to increased sinking into the snow, additional void is better for interlocking between the soft snow and the tread pattern, whereas on stiff ice, the contact area is relevant for braking. Finally, more void weakens the tread's structure and, in consequence, cornering stiffness, which is relevant for handling. Less void means also having more rubber available to wear thus increasing the rolling distance over lifetime. On the other side, the higher vertical deformations of the smaller rubber blocks under contact pressure induce higher hysteresis loss and increase rolling resistance.

It has to be mentioned that some of these results are inconsistent with those in Figure 3. The results, documented in Figure 3, are elder ones and mainly based on variations in sipe technology at constant void. Obviously, one has to be careful with general statements on performance conflicts and has to consider always in detail which influencing parameter is the dominant one in steering the conflict!

4.2 Tread compound

Figure 4 shows the dominant influence of the tread compound on braking, winter performance, rolling resistance, wear, and handling. Four of these capabilities can be found in the so-called "magic triangle," which visualizes the challenges for a rubber chemist (Figure 7). Centrally located is the wet braking behavior, being also most important from a driving safety viewpoint. It conflicts not only rolling resistance but also winter performance and wear.

Especially the conflict between *wet braking and rolling resistance*, which according to Societe de Technologie Michelin (2005) causes about 20–30% of cars fuel consumption, must be dealt with in depth. For the reason of this conflict, both performances will be made transparent to the consumer by help of a future tire label classifying these performances by characters from A to G [see the related legislation from the European Union in des Europäischen Parlaments und des Europäischen Rates (2009)]. The material property causing this conflict is the hysteresis of rubber, measurable by the so-called loss angle δ. This loss angle significantly depends on the temperature, as shown in Figure 8. As rolling resistance is 50% because of the tread compound, we have to consider its operating temperature around 50°C and its operating frequency of deformation of about 10 Hz, which is also the excitation frequency of the measurement in Figure 8. Therefore, it can be concluded that the indicated dark gray

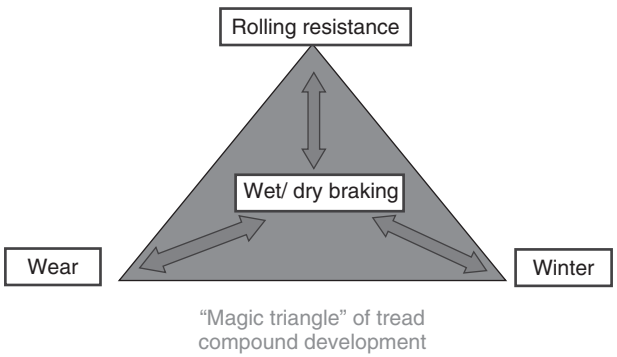


Figure 7. Schematic presentation of main tire performance conflicts influenced by tread compound.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

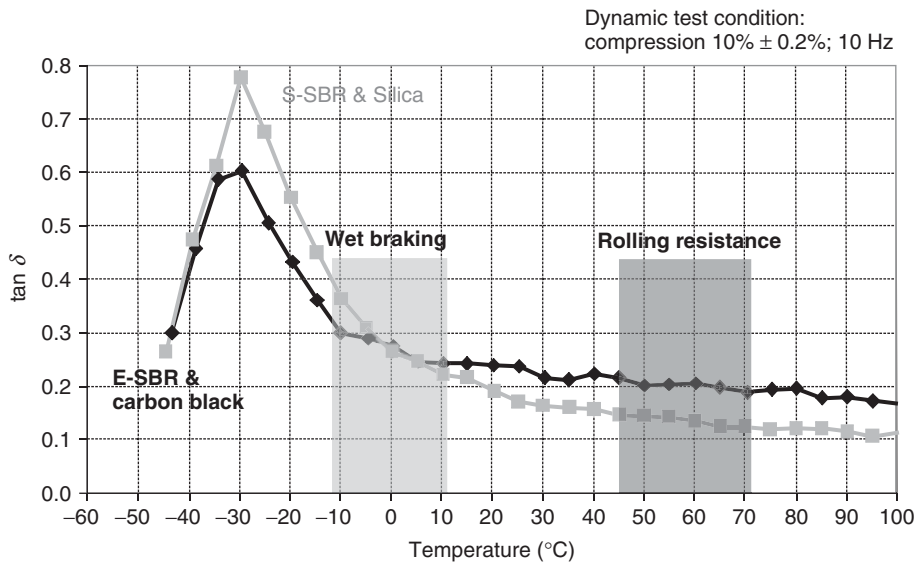


Figure 8. Loss angle as function of test temperature with correlating temperature ranges to rolling resistance and wet braking.

temperature range is the range of correlation in ranking between measured tan δ and measured rolling resistance. The light gray correlation range for wet braking is located around 0°C because of the underlying hysteresis friction mechanism, described in Kummer (1966), which operates around excitation frequencies of 10^4–10^7 Hz, depending on the roughness of the road and the sliding speed of the tread blocks. To make the conclusion complete, one has to consider the equivalence between temperature and dynamic excitation for rubber (Clark, 1982): qualitatively spoken, this stands for the phenomenon that rubber behaves equally when a colder operating temperature is compensated by a higher loading frequency and vice versa. This fact explains the shift of the correlation range for wet braking in regard to rolling resistance to lower temperatures.

One can now overcome the conflict between wet braking and rolling resistance only by changing rubber chemistry in a way that tan δ is increased at lower temperatures and decreased at higher temperatures. This happened with the introduction of silica as substitute for carbon black in the mid-1990s, shown by the light gray curves compared to the black curve in Figure 8. The tremendous progress in reducing this conflict since then, achieved by the rubber industry, is shown in Figure 9. It gives hope that, by use of new rubber materials, this most important conflict will be reduced significantly in future.

The conflict between *wet braking and winter capabilities* can be explained by a shift of the graph for the loss angle δ to negative temperatures. This shift is necessary to achieve winter capabilities by maintaining, under winter temperatures, the flexibility of rubber that is responsible for enhanced traction on snow and ice (Mundl, Wiese, and Wies, 2011). Doing so, one can deduce the lower ranking of those compounds in the temperature range for wet braking at 0°C. The conflict between *wet braking and wear* can be illustrated by the molecular model of the internal damping of rubber: it can be imagined as the energy consuming sliding of the molecular chains of polymers attached to the graphite crystal surfaces of the filler material, carbon black, by van der Waals forces. Hindering this internal sliding increases the internal cohesion and in consequence wear resistance but it lowers hysteresis necessary for wet braking as well.

4.3 Tire contour

A tire contour can be characterized by two main parameters: the radius of the tread and its width. These parameters span a two-dimensional plane qualitatively limited by the terms small/large and narrow/wide, wherein the optimal location of several tire properties are marked qualitatively (Figure 10). With round and narrow treads, the drainage of water is eased because of laterally shorter flowing distances for expressing water and a *hydrodynamic* better suited oval shape of the contact patch for plowing through the water layer. *Noise* is reduced because a narrower leading and trailing edge of the contact patch sees less impacting and snapping out of tread structures. The rounder shape of the contact patch has probably a more uniform and therefore less lateral noise radiation.

These tire performances *conflict* now with less *rolling resistance* and *wear* with a flatter curvature. This fact can

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

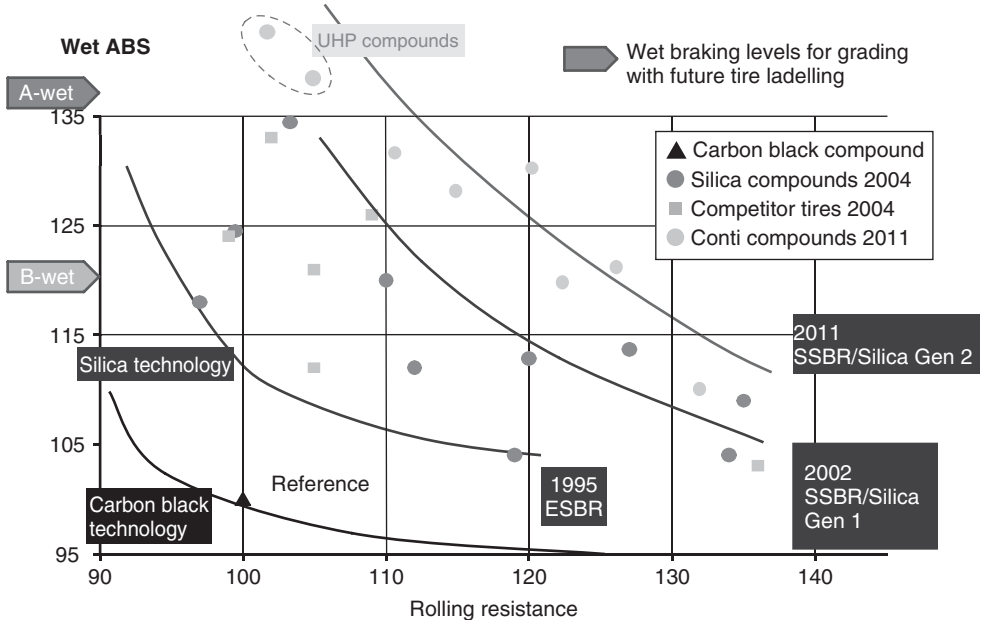


Figure 9. Evolution of the performance conflict between wet braking and rolling resistance from 1990 to 2008.

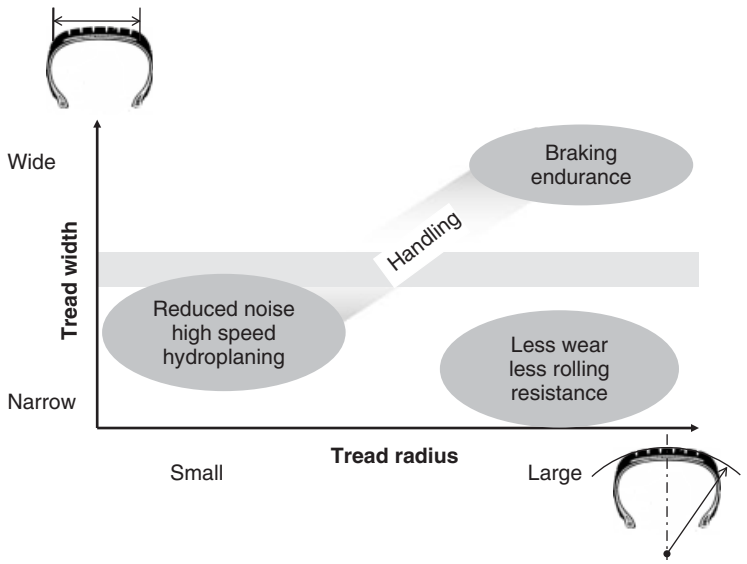


Figure 10. Target conflicts, resulting from tire contour variations.

again be explained with less energy input for flattening the tread structure resulting in a lower amount of internal damping and less frictional energy, which is responsible for abrasion.

They *conflict* also with *braking* capability, which improves with wider treads. This fact can be explained with a shorter contact length, which reduces opposed shear forces in the longitudinal direction, diminishing the overall friction potential.

4.4 Tire construction

Tire construction is first responsible for the endurance and durability of a tire. Secondly, it serves to optimize handling properties. It holds also for about 50% of rolling resistance. The dominant performance conflicts are therefore the ones between *durability/handling versus rolling resistance*. To make, for example, the tire sidewall more resistive against punctures under bad road conditions, a thicker rubber

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

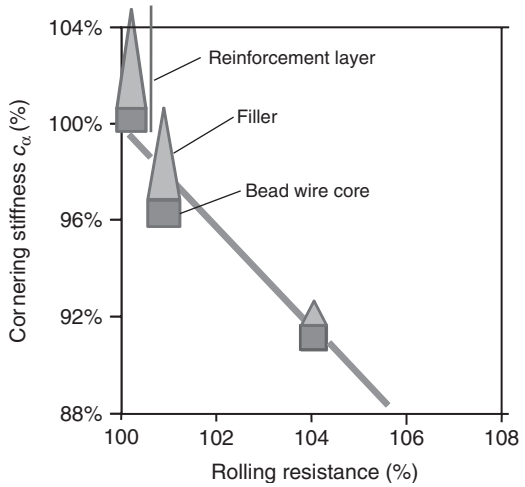


Figure 11. Conflict between handling and rolling resistance controlled by bead stiffness (various schematic bead designs).

coating and an additional ply are necessary to protect the reinforcing cords. This of course increases rolling resistance because more energy is needed to deform the thicker sidewall. For better handling, stiffening of the bead area by a higher apex often helps. Again more material is needed and increases rolling resistance, as can be seen in Figure 11. Therein the different bead constructions are indicated by symbols. The dark gray area stands for the bead wire layers, the light gray area for the apex rubber, and the gray line for the textile reinforcement layer. The textile reinforcement is only influencing the cornering stiffness, not the rolling resistance. Figure 11 shows that 2% increase in cornering stiffness leads to a change of about 1% in rolling resistance.

5 TARGET CONFLICTS OF UHP TIRES

5.1 Description of virtual development process

The development of UHP tires is dominated by the improvement of the three following performances: handling, high speed capability, and wear on race tracks. To shorten the development process and to save costly testing efforts, a major part of the development process has become virtual, which means to achieve handling requirements by simulation of virtual prototypes. Following Fischer *et al*. (2010) and visualized in Figure 12, the interface between tire evaluation and tire design is defined by tire target characteristics. In particular, a corridor in cornering stiffness for front axle (FA) and rear axle (RA) and increased lateral grip limits has to be achieved by the proposed tire variants under investigation. Both properties can be identified in the side force/sideslip angle dependency shown in Figure 13. The cornering stiffness is the inclination of above-mentioned dependency for slip angles up to 3°, whereas the grip limit is achieved at high slip angles. On the one hand, the cornering stiffness results only from stiffness properties of the tire, in particular the tread pattern stiffness and tire body stiffness connected in series. On the other hand, the grip limit is dominated

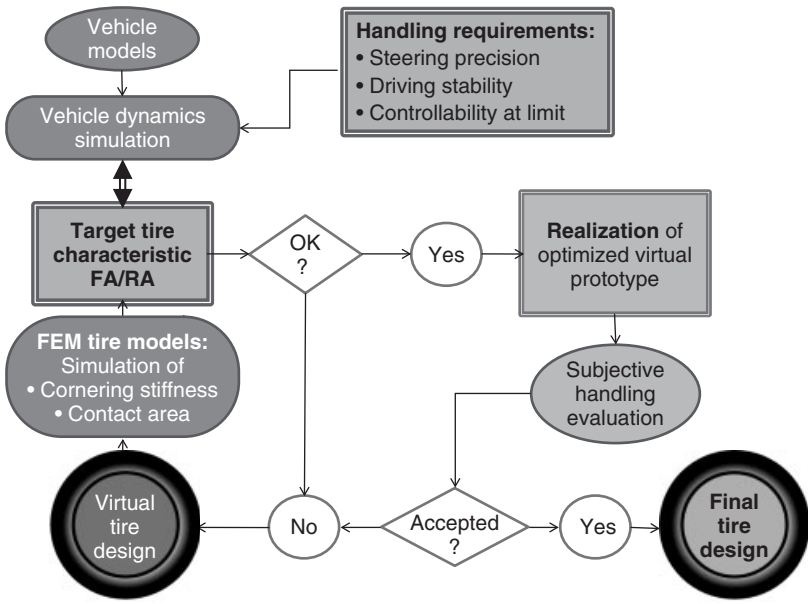


Figure 12. Virtual development process of UHP tires.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5
www.it-ebooks.info

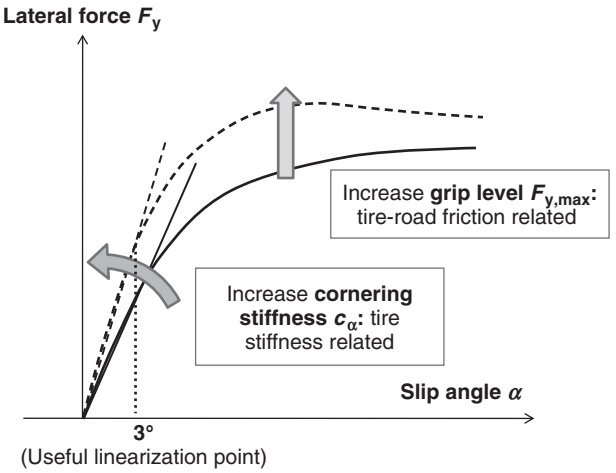


Figure 13. Main parameters of lateral force/slip angle behavior.

by the frictional properties of the sliding tread influenced by tread pattern structure and tread compound in reaction to the road surface. In a first loop of the virtual development process, the proposed tread pattern variants are simulated in their lateral stiffness; this shows the potential for improvement, the cornering stiffness being relevant for steering precision and driving stability. The related simulation tool is discussed in Mundl *et al*. (2008). The optimal tire pattern variant is realized in tire prototypes and checked against a base tire in its subjective handling performance on test tracks. These subjective evaluated results are associated with objective measurements of vehicle dynamics such as vehicle slip angle versus lateral acceleration (Figure 14).

In a second loop, proposed changes in tire body design are modeled and evaluated by finite element analysis in cornering stiffness and contact area. Optimization criteria are again maximal cornering stiffness and additionally maximum contact area under severe lateral forces with the consequence of an as low as possible contact pressure distribution for a maximum coefficient of friction. Again, a final vehicle handling evaluation is performed for the realized best variant to ensure the achieved targets in handling capability in serial production.

5.2 Main target conflict: grip limit versus cornering stiffness

As described earlier, it is essential for UHP tires to design them in a way that excellent handling on sports cars is achieved. This means that both cornering stiffness and lateral grip limit have to be increased and that in a way that the balance between FA and RA is maintained for the purpose of stable cornering. In general, driving at the limit may become also more challenging because of a smaller range of maximum grip (Figure 13). Inexperienced drivers may feel this as unpleasant. As grip is mostly affected by measures in the tread, a balance has to be found predominantly for the tread compound stiffness. The softer the tread compound is, the more the road asperities penetrate into the surface of the rubber blocks and interlock with them, thus producing enhanced grip. On the other hand, these softer blocks produce less lateral contact shear forces at a specific slip angle and therefore a lower cornering stiffness, as visualized qualitatively in

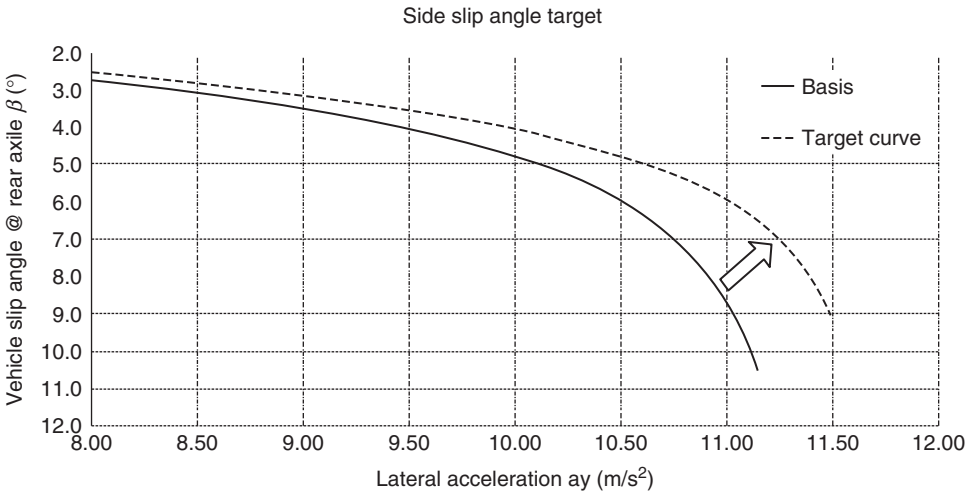


Figure 14. Slip angle at rear axle as indicator for improved grip at the limit. (Reproduced from Fischer *et al*., 2010. With permission from Aachen Colloquium Automobile and Engine Technology, 2010, M. Fischer, J. Ehlich, C. Schroeder, K. Peda and B. Wies, Continental.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

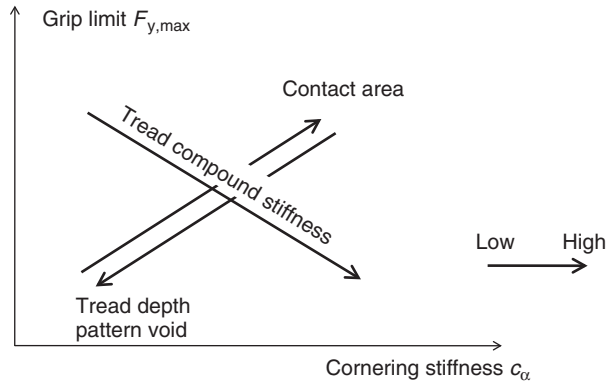


Figure 15. Qualitative influence of tire design parameters on grip/cornering stiffness conflict.

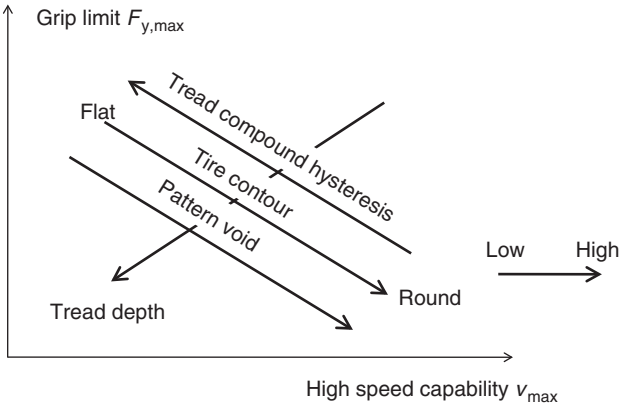


Figure 16. Qualitative influence of tire design parameters on grip/high speed capability conflict.

Figure 15. By chance, other design measures in the tread pattern, such as increased positive or a lower height of the pattern blocks, are positive for grip and cornering stiffness simultaneously (Figure 15). The same is true for a larger overall contact area of the contact patch, which can additionally be controlled by tire contour and tire body design.

5.3 Secondary target conflicts: grip limit versus high speed capability and wear

With their higher maximum speed, sports cars do need tires with a raised high speed capability. This can be in conflict to lateral grip as well (Figure 16). The hysteresis of the tread compound dominates this conflict. On the basis of the model of hysteresis rubber friction, see again the theory in Kummer (1966), a high internal material friction leads to increased grip, whereas hysteresis causes heat buildup in the rubber blocks, thus limiting maximum rolling velocity by depolymerization of the tread compound. Going back to Figure 10, we recognize therein that the contour is improving grip when it is flat but worsening high speed at the same time. The reason for this behavior is that a flat contour has a tendency to larger contact areas and better grip as mentioned earlier but causes higher pressure peaks in the shoulder area because of the dynamic growths induced by centrifugal forces and, in consequence, larger deformations and additional heat buildup. A high void in the pattern is a countermeasure for this, because it helps cooling but again negatively influences grip (Figure 16).

It has to be mentioned that resistance against wear of the tread compound conflicts also its grip capability. The explanation is that, again, higher hysteresis of the tread compound causes increased heat buildup, thus making the compound softer and less resistant to wear. The explanation is that under traction forces onto the driven axle, softer tread structures suffer from larger deformation energy in the contact area that is transformed into frictional energy at the trailing edge, resulting in increased wear.

6 OUTLOOK FOR FUTURE IMPROVEMENTS OF UHP TIRES

A conflict analysis of tire performances is helpful in giving hints where future innovations may take place. In the field of UHP tires, potential can be seen in two ways to improve the conflict between enhanced steering properties and increased grip level dominated by future compound aspects: an anisotropic compound behavior, such as higher shear stiffness and lower radial stiffness, may be achieved by oriented microfibers within the compound matrix, see also the physical reasoning in Chapter 4.2. Another vision is a deeper interaction with tire properties and advanced vehicle control systems using active steering, thus improving steering precision and handling stability even at lower cornering stiffness of the tires and leaving development space for ultrahigh grip levels.

REFERENCES

Clark, S. (1982) *Mechanics of Pneumatic Tires*, US Government Printing Office, Washington, DC, p. 24.

Doporto, M., Mundl, R., Wies, B., *et al*. (2003) Zusammenwirken zwischen Profil und Laufflächenmischung. *Automobiltechnische Zeitschrift*, **105**(3), 238–249.

Verordnung Nr. 1222/2009 des Europäischen Parlaments und des Europäischen Rates (2009), Über die Kennzeichnung von Reifen in Bezug auf die Kraftstoffeffizienz und andere wesentlichen Parameter, Brüssel, November 25.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fischer, M., Ehlich, J., Schroeder, C., et al. (2010), Virtual based development process for UHP tires with target setting for an excellent driving experience. *Aachener Kolloquium Fahrzeug- und Motorentechnik*.

Heißing, B. and Ersoy, M. (2007) *Fahrwerkhandbuch*, 1st edn, Vieweg Verlag, p. 346.

Kummer, H.W. (1966) *Unified Theory of Rubber and Tire friction*, Pennsylvania State University, 94.

Mundl, R., Fischer, M., Strache, W., *et al*. (2008) Virtual pattern optimization based on performance prediction tools. *Tire Science and Technology*, **36**(3), 192–210.

Mundl, R., Roeger, B., and Wies, B. (2009) Influence of pattern void on hydroplaning and related target conflicts. *Tire Science and Technology*, **37**, 187.

Mundl, R., Wiese, K., and Wies, B. (2011) An analytical thermodynamical approach to friction of rubber on ice. Presented at the 2011, Tire Society meeting, Akron, Ohio; also in), *Tire Science and Technology*, **40**, 124–150.

Societe de Technologie Michelin (2005) *The Tyre, Rolling Resistance and Fuel Savings*, 2nd edn, Karlsruhe, p. 35, CD-Rom, Abt. Öffentlichkeitsarbeit.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto012
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Tire Pressure Monitoring Systems

Victor Underberg[1], Thomas Roscher[1], Frank Jenne[1], Predrag Pucar[2], and Jörg Sturmhoebel[2]

[1]*AUDI AG, Munich, Germany*
[2]*NIRA Dynamics AB, Linköping, Germany*

1 Introduction 1
2 Direct Tire Pressure Monitoring Systems 3
3 Indirect Tire Pressure Monitoring Systems 8
4 Legal Environment 13
5 Future Trends and Development 16
Acknowledgments 17
Related Articles 17
Further Reading 17

1 INTRODUCTION

The tire represents the link between vehicle and road surface. As such, the tire characteristics determine the transfer of forces between the vehicle and the road. One major factor influencing the actual tire properties with respect to vehicle handling, braking, fuel consumption, and tread wear is the tire pressure. Only with the vehicle-manufacturer-recommended tire pressures, the optimum of all the above-mentioned characteristics can be guaranteed. The recommended tire pressures do depend not only on the specific vehicle/tire combination only but also on anticipated driving conditions as vehicle load (partly or fully loaded) and travel speed (low or high).

Already in 1923, Bosch patented a mechanical device—the "Bosch-Glocke" (Bosch-Bell)—which was designed to warn the driver of under-inflated tires (Figure 1). The concept behind the "Bosch-Glocke" was that the deformation of an underinflated tire running through the contact patch activates a wheel-mounted lever, which, in turn, will ring a high frequency bell to alert the driver.

Today's tire pressure monitoring systems (TPMSs) are designed to monitor the current tire pressure in all four wheels in service—in some cases also the spare wheel—and to provide a warning to the driver if the tire pressure in at least one the wheels falls below a predefined threshold level. In general, there are three different pressure loss scenarios (Figure 2). First, there can be a tire blow out, where the tire loses all its air in just seconds because of sudden and severe tire damage, for example, by hitting a pot hole (TPMSs are not designed to warn drivers in the blow out scenario, as the driver will be alerted immediately by a sudden change in vehicle response). The second scenario is an air leakage or punctures with pressure loss rates of approximately 10 kPa (equiv. to 1.45 lb/in^2) per minute up to a week, usually caused by a puncture (nail, screw, etc.), a leaking valve, or improperly mounted tires. In most cases, a tire blow out or a leakage only affects one tire at a time. However, tires in service also naturally lose air through a process called *diffusion*, also known as *permeation*. In this case, air molecules pass through the tire, as the tire inner liner does not guarantee total impermeability. This scenario affects all four tires in the same way with an approximate pressure loss rate of about 10 kPa per month and will cause significant under inflation if tire maintenance is neglected by the driver, that is, tire pressure is not checked and adjusted over a long period of time.

At first, TPMSs were introduced as a comfort feature in the late 1980s. The model year 1987, Porsche 959

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

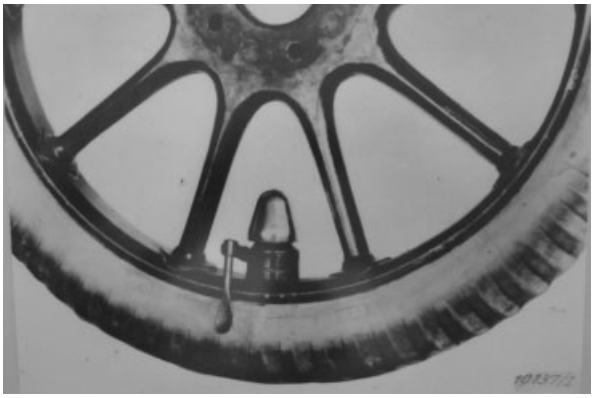


Figure 1. 1923 "Bosch-Glocke" (Bosch-Bell).



Figure 3. 1987 Porsche 959.

was the first series production car to feature a TPMS (Figure 3). The main components of the Porsche 959 TPMS, pressure switch and high frequency unit, are represented in Figure 4. After the turn of the century, having the correct tire pressure with the effect of maintaining low rolling resistance and by that increasing fuel economy shifted into focus. The first TPMSs were "direct" measuring systems, meaning tire pressures, and often temperatures, were measured within the wheel and the information was transmitted via radio frequency to a receiver inside the vehicle. The so-called indirect systems use the information embedded in the wheel speed signals of the antilock brake or electronic stability control system (ABS/ESC) to determine under inflation. Indirect measuring systems of the first generation were able to detect puncture scenarios only at one tire, whereas current state-of-the-art second-generation indirect systems are able to detect puncture and diffusion scenarios in up to all four tires. Indirect systems of the first generation became widely used in the early 2000s as a cost efficient alternative to complex and, back then, error prone direct TPMSs. The increasing market share of run-flat tires (others say "extended mobility" tires) around that time also caused the increasing market share of first generation iTPMS (indirect tire pressure monitoring system) as puncture detection systems. Especially when using run-flat tires, drivers have to be provided with underinflation information, as a flat tire will not be detected by visual inspection.

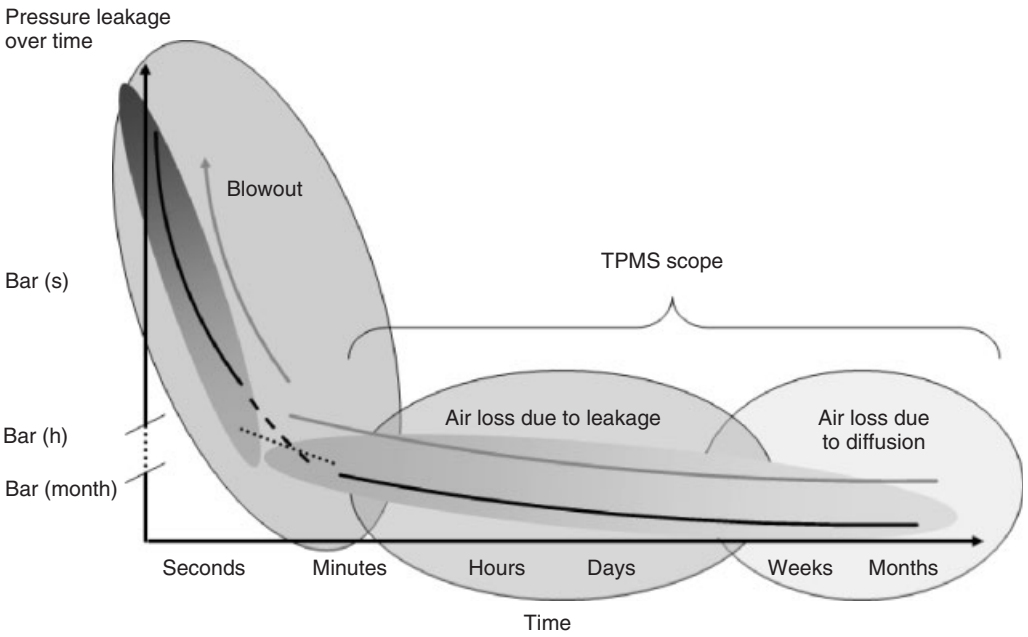
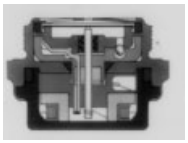
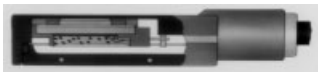


Figure 2. Pressure loss scenarios.

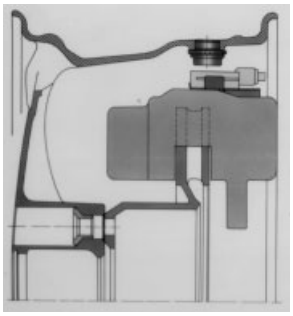
Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) Pressure switch



(b) High frequency sensor



(c) Wheel assembly

Figure 4. (a–c) 1987 Porsche 959 TPMS components.

In the year 2000, a series of fatal accidents related to tire tread separation on Firestone tires mounted on Ford Explorer sport utility vehicles triggered the US congress to pass the so-called TREAD Act (Transportation Recall Enhancement, Accountability and Documentation Act). Subsequently, in 2005, the new Federal Motor Vehicle Safety Standard No. 138 (FMVSS 138) was enacted, requiring all new passenger cars, multipurpose passenger vehicles, trucks, and buses with a gross vehicle weight rating (GVWR) of 4536 kg (10,000 pounds) to be equipped with a TPMS. Starting with a phase-in period in 2005, by September 2007, all new vehicles had to be fitted with a TPMS conforming to FMVSS 138 requirements. With that the United States was the first country to have a regulation specifying the installation and performance criteria for TPMSs. Targeting the fuel efficiency in addition to the safety aspect of correct tire pressure, in 2009, the EU parliament mandated ECE-R 64, a regulation requiring all new vehicle types to be equipped with a TPMS starting November 2012 and all new vehicles to be registered from November 2014.

2 DIRECT TIRE PRESSURE MONITORING SYSTEMS

Direct pressure monitoring systems are based on the principle of measuring tire internal pressure (and temperature) by wheel units fixed to the rim. The sensors send their data by radio frequency to vehicle body fixed antenna(s). The antenna(s) provides the data to an electronic control unit (ECU) for evaluation. The ECU communicates with the car's network and issues messages that warn the driver in case of a pressure loss via telltales and messages in displays in the instrument cluster and/or human–machine interface (HMI).

The functional principle introduces system complexity to the vehicle design. The parts required for the direct pressure monitoring functionality are the following:

1. wheel units
2. antenna(s)
3. ECU and wiring harnesses
4. optional parts and displays
5. development.

The following sections discuss the function of each system component.

2.1 Wheel unit

This component is mounted at least in each driving wheel. It measures at least tire pressure and temperature. In addition, data such as acceleration and rotational direction may be collected. The wheel unit sends the data wirelessly on standardized carrier frequencies of 433 or 315 MHz to a vehicle body fixed antenna.

Figure 5 shows a block diagram of an HUF Electronic sensor (formerly known as *BERU Electronics*). The low frequency (LF) module enables wake-up and data requests from outside the tire for diagnosis or functional purposes. The "over temperature protection circuit" allows survival for temperatures up to 150°C for several minutes. At 120°C, the wheel unit will fall into a protection mode to safeguard electronic parts at these temperatures. A lithium-based consumer battery, providing 3.0 V output, provides energy. In general, state-of-the-art sensors transmit data:

1. Continuously, when activated once (e.g., one transmission/minutes in regular intervals).
2. Roll switch activated at a certain threshold acceleration.
3. Outside requested, for example, by a diagnostic tool or TPMS trigger (initiated by an LF signal).

As the inner tire environment is hostile because of temperature variations (−40 to +140°C), humidity, and high acceleration (up to 2000 g), the plastic-housed electronics are protected by a potting compound. The opening for pressure measurements is either at the top of the housing or at membrane shielded and integrated into the potting.

For mounting the wheel unit to the rim, several different techniques are used. The most common solution is the attachment of the wheel unit to the valve as snap in or fixed by a screw to an alloy valve (Figure 6). In Figure 7, a valve-mounted wheel unit is shown in the wheel unit assembly and in a cross section. Besides that, metal bands around the rim are also in use.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

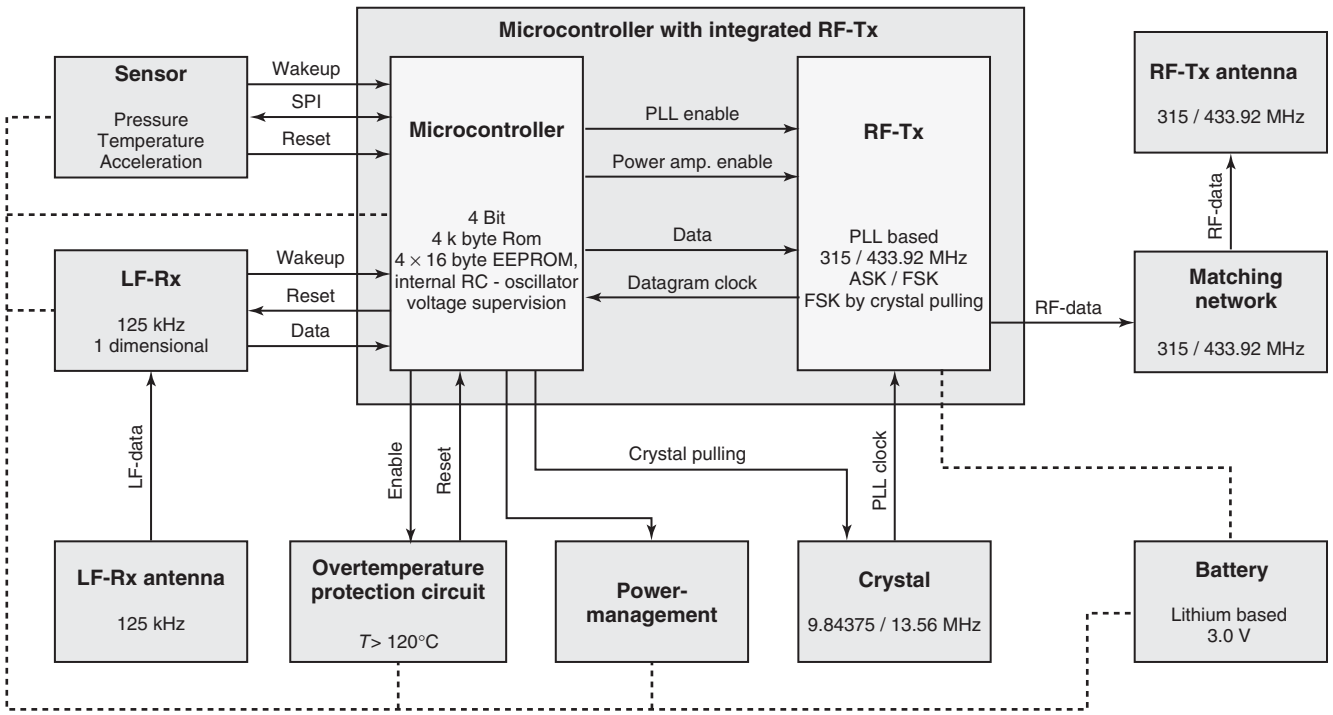


Figure 5. HUF Electronics sensor, block diagram.



(a) HUF Gen1 (b) HUF Gen2 (c) HUF Gen3

Figure 6. HUF Electronic wheel units, generation 1 (a), 2 (b), and 3 (c).

2.2 Antenna

As to the antenna, receiving data from the wheel unit, several solutions are available:

1. passive external antenna
2. active external antenna
3. integrated antenna.

Passive antennas only receive the radio signals without processing the data. In this case, wire connections transmit the analog signals to a receiver, implemented, for example, into a TPMS ECU or a body computer. During the advent of TPMSs, receiver technology was unable to filter out disturbances effectively from the use-signal. For this reason, these systems were prone to electromagnetic compatibility (EMC) issues. An example of a passive antenna implementation is shown in Figure 8. As receiver technology progressed, and being less susceptible to EMC issues, this cost-effective solution was base for the so-called integrated solutions. Integrated in this sense means sharing receiver and antenna with other vehicle systems (e.g., keyless entry) and implementing the TPMS software into the body computer.

Active antennas represent the next step for improved receiver performance. In this setup, the receiver is integrated into this antenna module. An example of an active antenna component is shown in Figure 9. It decodes the received signals and transmits them in digital format via wire to the TPMS ECU where the analysis takes place. Figure 10 shows an example of an active antenna system implementation. This antenna generation is more sensitive to radio signals but more resistant against EMC issues. It assures easier application and modular system configuration. Another technology leap is represented by the so-called intelligent antenna, which means that TPMS ECU, receiver, and antenna are combined in one housing to reduce parts (Figures 11 and 12).

2.3 Electronic control unit (ECU) and wiring harness

The TPMS ECU processes and evaluates the data received by the antenna or additional data from other ECU; an

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

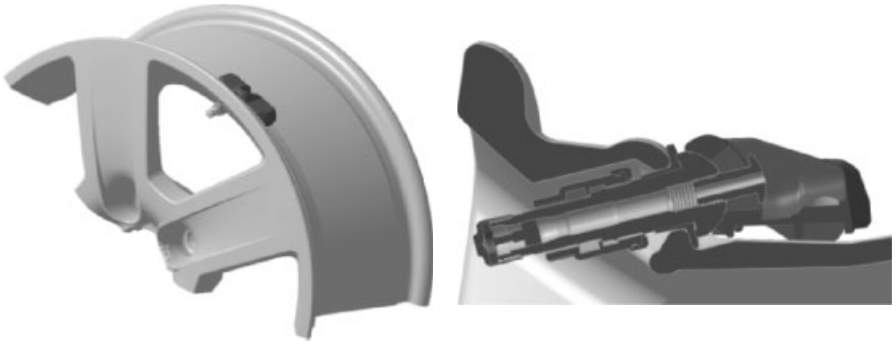


Figure 7. Wheel-mounted pressure sensor with cross-sectional view.

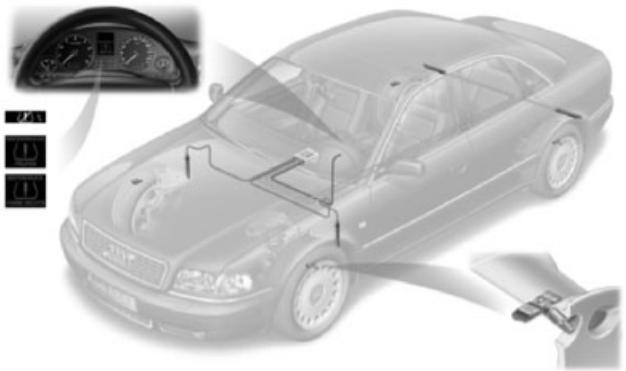


Figure 8. 2001 Audi A8 (D2) with passive antennas.



Figure 10. 2004 Audi A8 (D3) with active antennas.



Figure 9. Example of a digital antenna component.



Figure 11. 2007 Audi Q7 modular high TPMS.

example of a TPMS standalone control unit is shown in Figure 13. It manages warnings, allocation of wheel units, and internal and external system diagnoses and provides data and messages for displays and menus. The ECU is connected to the car's wiring harness and communicates as part of this network with other ECU such as gateway, dashboard, and MMI. Wiring depends on TPMS structure and function. As minimum, there is connection to controller area network (CAN) bus, power supply, and ground. Additional components may be linked and organized by local internal network (LIN) or direct linked.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 12. 2007 Audi Q7 modular low TPMS.



Figure 13. Stand-alone TPMS electronic control unit.

Learning and management of wheel units may be done via diagnostic tools at a car or tire dealer or by the TPMS software itself via auto learning and auto location. Auto learning and auto location is more or less static evaluation of information such as number of received tele messages, their signal strength, acceleration in combination with vehicle speed, and turning direction.

2.4 Optional parts and displays

Optional parts may be, for example, a 125 kHz transmitter, the so-called trigger, an example is shown in Figure 14. With this, the TPMS is able to build up a bidirectional communication. The transmitters, one placed in each wheelhouse behind the mudguard, are connected to the ECU. If the wheel unit receives the LF signal from this transmitter, it sends its data on 433/315 MHz to the TPMS antenna no



Figure 14. 125 kHz Transmitter unit, trigger.

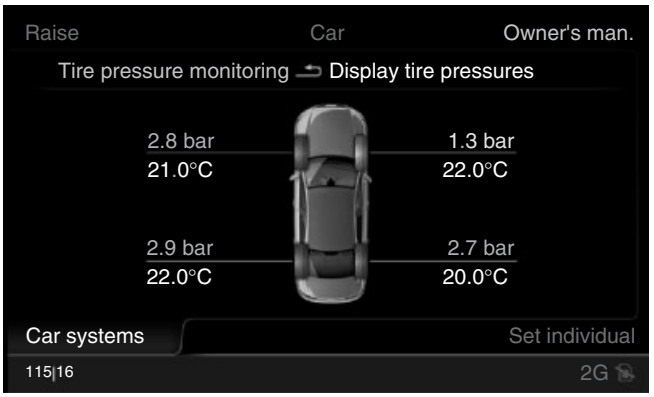


Figure 15. TPMS information in Audi MMI.

matter if the car moves or is stationary. The ECU is always able to request data, for example, when data are lost or for comfort purpose at door open as a status request. In addition, using this additional communication does allocation of sensors quickly and accurately.

Warning telltales and text messages in case of pressure loss or malfunction are displayed in a display in the dashboard. Additional information, such as pressure/temperature status, may be shown either on displays in the dashboard or suitable HMIs (Figure 15).

For operations such as resetting the system, start learning new wheel units, or storing tire pressures after adapting them, a push button connected directly to the ECU, or a menu either in the driver information system in the dashboard (DIS) or another HMI, provides this function

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to be executed by the driver. As the placard pressure may vary with mounted tire size, velocity to be driven, load and other characteristics TPMSs can provide various parameters to be chosen.

2.5 Development

Development of TPMS depends on OEM (original equipment manufacturer) functional, HMI, and legal requirements. In general, TPMS development can be grouped in the following categories:

1. hardware development
2. location/placement of hardware in the vehicle body (application)
3. software development.

2.5.1 Hardware development

All components of TPMS, except the wheel units, are subject to automotive industry standards depending on the location/placement of the different parts. For example, high temperature conditions can occur in the dashboard area. In addition, outside the passenger compartment, placed components need to be shielded against high humidity influences. In some circumstances, special protection against stone chipping is also needed. Electrical tests, for example, cover over/under voltage, voltage ramping conditions, EMC, or quiescent current topics.

As mentioned earlier, wheel units have to withstand the hostile environment inside a tire. In a joint effort, German car manufacturers (AUDI, BMW, Daimler, Porsche, and VW) and TPMS supplier DODUCO (later BERU, HUF Electronics) have developed detailed requirement specifications that describe the functional testing and validation of TPMS wheel units and attachments to the wheel by Alligator-developed alloy valves. Especially important were considerations about accuracy over component lifetime, minimal temperature dependency, resistance to high acceleration (up to 2000 g), temperatures higher than 120°C, and temperatures as low as −40°C. In addition, realizations of reliable functionality withstanding the high mechanical stresses were challenges mastered.

2.5.2 Placement of hardware at car body (application)

Positioning the receiver is crucial for TPMS functionality, as it strongly depends on the signal strength and reception rate of the wheel unit sensors at the location of the receiver. When placing LF transmitters in the wheelhouse, positioning of these components also requires that the wheel unit can be reached in all possible wheel positions (360° wheel rotation and all steering positions of wheel).

Investigations on digital mockup (DMU) adduce the first positions of components chosen based on available space and experiences from former vehicle projects. Subsequent tests on prototype vehicles give first impressions of the component functionality at the predefined position. To evaluate the receiver placement, the wheel unit sensor is set into a steady-state sending mode. The signal strength at the receiver is measured and evaluated for a full 360° rotation of each wheel. In addition, at the front axle wheels, this procedure needs to be carried out at full steering wheel angle, left and right turns. The objective is to have best possible reception and, if they cannot be avoided, only short "black spots," that is, positions where the signal strength falls below predefined threshold values during one revolution of the wheel. The same principle test needs to be performed for 125 kHz transmitter positions at each wheel position. The objective here is evaluating the flux density at the 125 kHz antenna of the wheel unit sensor during one full turn of a wheel. In addition, here, no or only minor black spots are to be accepted (Figures 16 and 17).

The reception rate of wheel unit sensor tele messages is measured while proving ground driving at different steady-state speeds, for example, 100, 150, and 200 km/h. Usually, two test conditions are analyzed and verified: confirmation of HF reception, when triggers are deactivated, and confirmation of the LF communication, when trigger mode is activated. Subsequent driving tests in all customer-relevant scenarios (high and low temperature, snow, rain, etc.) need to confirm proper functionality. As specific tire constructions may influence signal reception (signal damping effect), close cooperation with tire development is also necessary. EMC of the TPMS components/system within the vehicle environment needs to be verified as well, meaning that the TPMS is not influencing other electronic systems in the vehicle and vice versa.

2.5.3 Software development

TPMS software components can be grouped into base software, diagnostic and functional software. The base software ensures basic communication between the TPMS ECU and other car components via the vehicle network. Functional software contains warning algorithms and algorithms defining the appearance to the driver. To function within a vehicle network, data exchange via signals and protocols are standardized. Diagnostic behavior for defect and malfunction analysis in the technical service environment (e.g., workshop) also needs to be covered and defined to ensure fast and effective troubleshooting.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

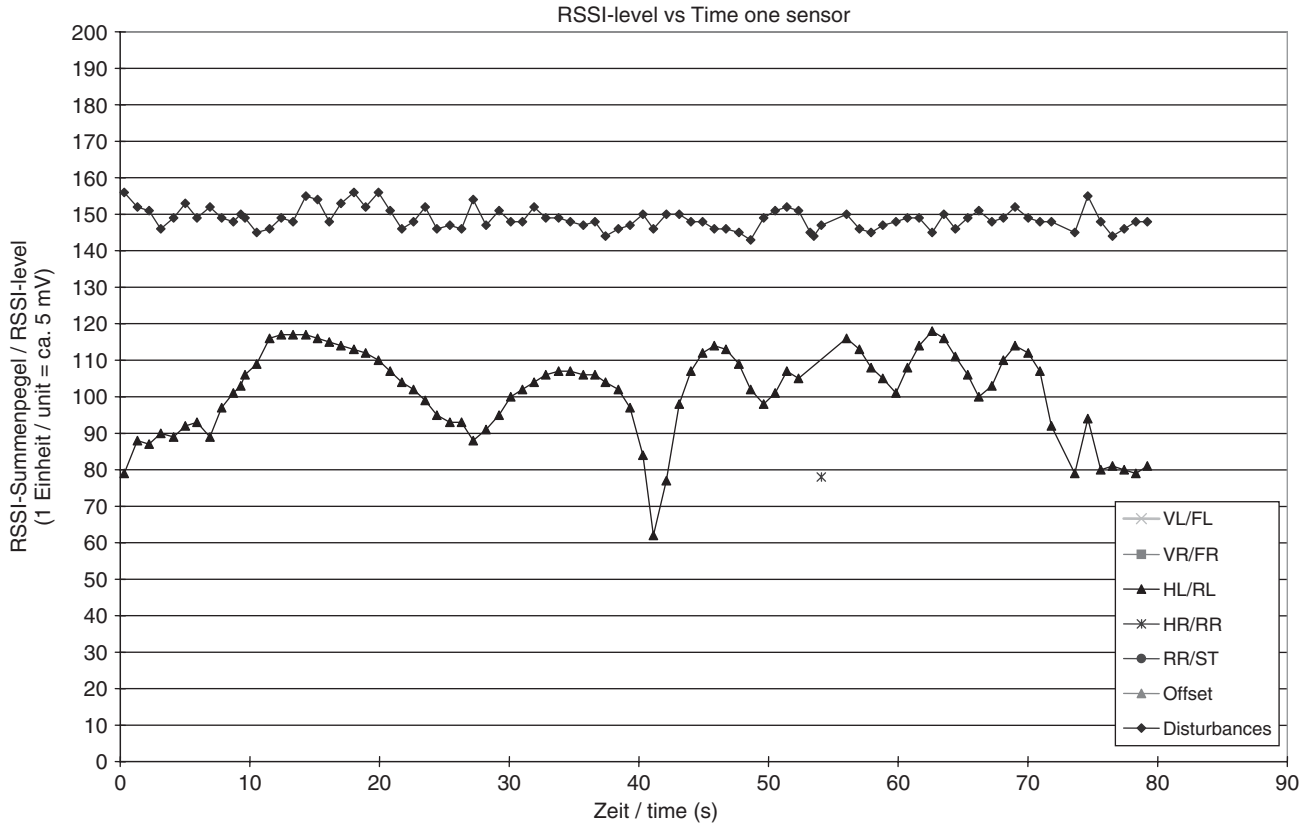


Figure 16. Signal strength diagram for a 360° rotation.

The functional software part defines warning strategies for different pressure loss scenarios, learning procedure of new sensors, learning of sensor positions, and the timing and sequence of telltale warnings and displays. Functional software setup depends on OEM requirements and philosophy as well as legal requirements for different markets. The warning strategy itself defines the pressure thresholds for signaling underinflation to the driver in case of a pressure loss because of puncture or diffusion pressure loss scenarios. As puncture usually means fast pressure loss (10–20 kPa/min), the warning needs to be displayed immediately when detected. Because of its long-term nature, detection of diffusion pressure loss need not necessarily be signaled immediately to the driver. This scenario is usually low pass filtered and often not displayed to the driver in the current ignition cycle but signaled at the next ignition on to assure cold tire pressure when the driver checks the pressures. In general—for customer acceptance purposes—it has been seen that warning thresholds too close to the recommended cold tire pressure have to be avoided. When signaling a warning at a pressures slightly below the cold recommended pressure, warning may be seen as nuisance by the customer and TPMS functionality is questioned.

Software verification is performed in hardware-in-the-loop (HIL), software-in-the-loop, and simulated network tests and in whole vehicle environment. Compliance to legal requirements and endurance is the main focus when doing approval vehicle tests.

3 INDIRECT TIRE PRESSURE MONITORING SYSTEMS

3.1 Tire as sensor

Directly measuring pressure sensors usually have components whose properties are dependent on the tire pressure and can therefore be used for measuring the pressure. Applying this basic principle in a generalized way leads to the approach of using the tire itself as a sensor. This is possible as various properties of a pneumatic tire are strongly pressure dependent and some of these properties are measurable using sensors being common and already installed for other purposes in most vehicles nowadays.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

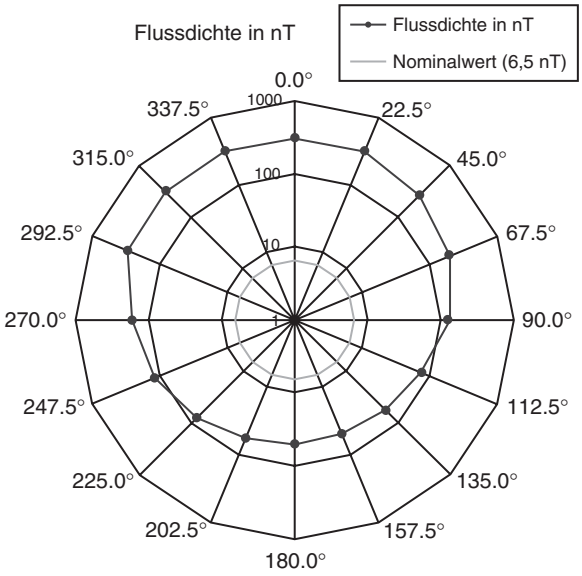


Figure 17. Flux density diagram.

The overall approach has various benefits as it avoids the installation of separate sensors and parts for the sake of monitoring tire inflation pressure. In this way, cost, weight, complexity, service, and logistics are efficiently avoided. In its most consequent application, it enables implementing a purely software-based tire pressure monitoring function into the ABS/ESC control unit as long an ABS or ESC system is fitted to the vehicle and can be utilized as a host for the TPMS software.

On the other hand, tires are complex products, which differ significantly in sizes, dimensions, materials, and technologies, as do their properties related to inflation pressure. The main challenge for iTPMS is to deliver consistent and reliable system behavior independent of the mounted tires. This variance in tire properties consequently is the main reason for iTPMS being relative by nature, meaning that they cannot monitor or display quantitative inflation pressure values but always need a reference state from which a deviation can be detected. The reference state can be viewed as an externally set reference point. The process where the user informs that a reference state is present is usually referred to as *reset* that is essential for the iTPMS. A reset is always required when inflation pressure has been adjusted, tires changed or rotated between different positions on the vehicle. In current applications, the reset is carried out actively by the driver (Figure 18). However, other options such as tire pressure gauges interacting with the vehicle could also enable other solutions in the future.

Subsequent to the reset, the iTPMS will analyze the pressure-dependent tire parameters and store the current parameters as reference data. This process is referred to as the *learning phase*. Once the learning phase is completed, the pressure-related tire parameters can be continuously monitored and compared to the stored reference data. If the current parameters deviate from the reference data in characteristics, according to predetermined patterns and/or exceeding certain thresholds, low tire inflation warnings are issued.

3.2 First-generation indirect TPMS

The first iTPMSs on the market, the so-called first generation iTPMSs, were only roll-radius or rolling-circumference-based systems. The basic principle is that the effective rolling radius of a tire is inflation pressure dependent and will decrease when the tire loses pressure. At a given vehicle speed, any inflation pressure loss can be detected by comparing and monitoring the individual rotational wheel speeds with each other (Figure 19).

The most important input signals are the individual wheel speeds. The wheel speed signals are obtained from measurements of number of teeth of the ABS rotor passing a certain point per time interval. The toothed rotors are fixed to each wheel and this way the rotational speeds used for ABS and ESC applications are calculated. This approach enables the effective and reliable monitoring of up to three out of four wheel positions but cannot detect same-rate

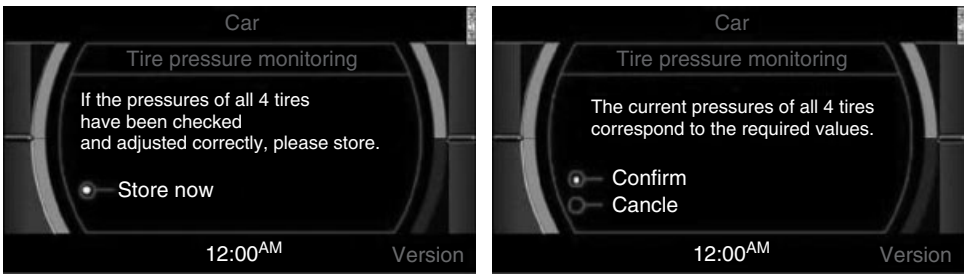


Figure 18. Guided reset procedure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

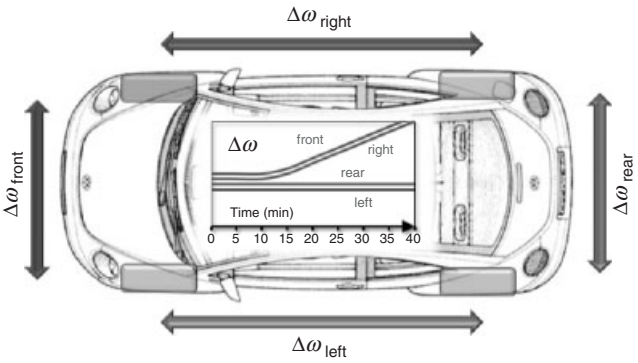


Figure 19. Four-wheel speed differences for rolling-radii-based indirect TPMS indicating a puncture front right.

inflation pressure losses on all four-wheel positions. The reason for not being able to detect the four-wheel deflation is that during the learning phase the relation between the wheel speeds is set. If all four wheels deflate in the same way, the relative difference will remain the same. Puncture detection is therefore possible with first-generation iTPMS but the diffusion/permeation detection due to slow natural air losses in all four wheels is not. Several approaches to overcome this limitation using wheel radius information only have been investigated, or are under investigation, but have not led to series applications yet.

Irrespective of the limitations, first-generation iTPMSs were mainly used and developed as the so-called run-flat-warning systems in combination with run-flat tires (see also ECE-R 64 requirements) as they are cost effective, reliable, and easy to integrate into ABS/ESC systems.

As the rolling radius is not only influenced by inflation pressure, other influencing parameters have to be considered of which the most influential one is the driving speed, which through radial forces influences the effective rolling radius and by that the wheel speed. It is common practice with iTPMS that they work in individual speed intervals for which reference values are stored. The current values are then continuously compared to the reference values stored for specific speed intervals. Other parameters that influence the rolling radius are cornering, braking and accelerating, load, and lateral or longitudinal road inclination. To eliminate the influence of these parameters, additional information such as engine torque, longitudinal and lateral acceleration, or the steering wheel angle are analyzed and used to compensate for the unwanted effects. On the basis of these auxiliary signals, the vehicle load can be estimated and compensated for. Advanced load compensation is possible using signals directly related to vehicle load, for example, air suspension pressures (when so equipped) or axle height signals commonly used with xenon front lighting systems.

3.3 Second-generation indirect TPMS

The wheel assembly is constantly excited by the road surface and oscillating in numerous vibration modes. In the course of trying to overcome the limitations of first-generation iTPMS, the oscillation properties of the wheel assembly have come into the focus of development.

Some of these vibration modes excited are highly inflation pressure dependent and some are possible to monitor with common sensor signals as the wheel speed sensors (Figure 20). The advantage of this approach is that the oscillation behavior of a wheel assembly can be monitored wheel individual and in absolute terms and not only relative to the other wheel positions as is the case for rolling radius. Realizing of monitoring the oscillation behavior therefore enabled the enhancement of iTPMS to a TPMS, comparable to direct systems, which are also capable of monitoring all four tire pressures individually. Indirect systems utilizing both the rolling radius and the oscillation behavior are called *second-generation iTPMS*. With this enhancement, the second-generation systems are able to comply with current and legal requirements that require both puncture and diffusion detections.

There are different options to utilize the vibration behavior but most common is to process the wheel speed signals and analyze them in a narrow frequency band of approximately 30–60 Hz as depicted in Figure 20 to reduce computational effort. Other approaches using vertical and/or translational oscillations have been investigated but not implemented, as additional sensors necessary are not as commonly used.



Figure 20. Spectrum behavior as function of tire pressure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

As with the rolling radius, the vibration behavior is influenced by other parameters than the inflation pressure alone. Vehicle driving speed is the most important and therefore the oscillation analysis is usually divided into different, consecutive speed intervals. The tire itself influences the oscillation behavior through their spring rates, effective mass, and inertia and so does the excitation by the road or effects of vehicle loading. One subset of these influences is compensated for through the reset procedure and learning phase, and the other subset can be actively compensated for by using additional signals such as ambient temperature or the signals also used for performing the rolling-radius analysis.

State-of-the-art second-generation iTPMSs use both rolling-radius and vibration informations and combine them by advanced signal processing to enhance the detection performance and increase robustness against disturbing factors.

3.4 Indirect tire pressure monitoring fundamentals

The so-called tire pressure indicator or TPI, an iTPMS and trademark of NIRA Dynamics, was the first iTPMS to be introduced to the US market in the 2009 model year Audi A6 to comply with regulation FMVSS 138 (Figure 21). In the following, some of its fundamentals will be discussed. Second-generation iTPMSs are also developed and marketed by Dunlop Tech (as deflation warning system or DWS) or Continental (as deflation detection system plus or DDS+).

TPI consists of two main modules; one monitoring the rolling radius called *wheel radius analysis* (*WRA*) and the other wheel spectrum analysis (WSA) is monitoring the oscillation or spectrum behavior.

The WRA and WSA modules calculate common properties or indicators, which describe the rolling-radius ratios of the wheels relative to each other as well as the wheel spectrum behavior. Having compensated for various disturbances, these properties can be used for a number of the so-called detectors, which have the purpose of monitoring certain deflation scenarios. For example, there are one-wheel-puncture detectors, two-wheel-puncture detectors, and several multi-wheel-diffusion-pressure-loss detectors. These indicators work independent from one another. Each of the indicators can, when triggered, issue a low pressure warning. Puncture detectors are typically characterized by their ability to react quickly due to the immediate safety relevance of the underlying pressure loss scenario. Diffusion detectors are typically low pass filtered to exclude influences of short-term disturbances. Detectors in TPI use both rolling-radius and spectrum informations simultaneously, such that the spectrum information is used as an attenuator or amplifier for base information obtained from the rolling-radius information. One exception is the four-wheel diffusion detector, which is purely based on spectrum information. Figure 22 shows the analyzed wheel speed differences and spectrum properties for a puncture scenario. As can be seen, the wheel radius difference and the spectrum are affected by the one-position-only pressure loss case. On the other hand, when all tires are affected at approximately the same pressure loss rate, no wheel speed

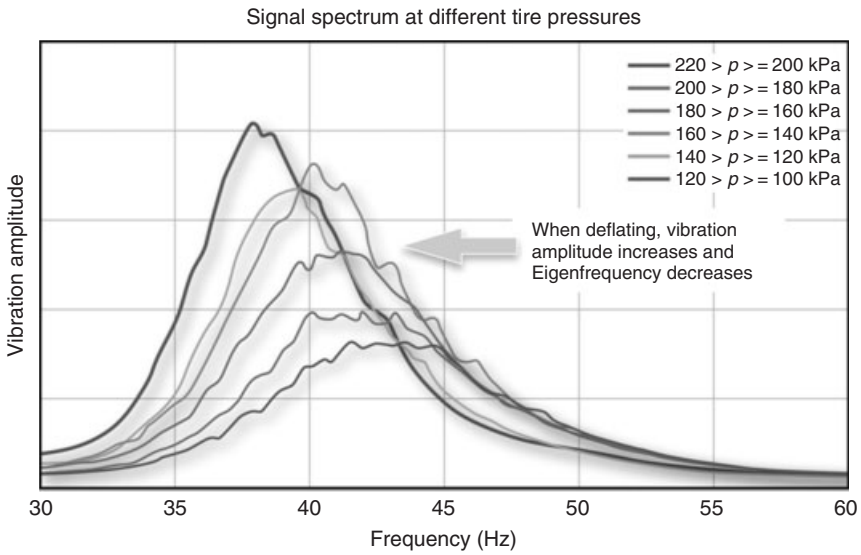


Figure 21. 2009 Audi A6.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

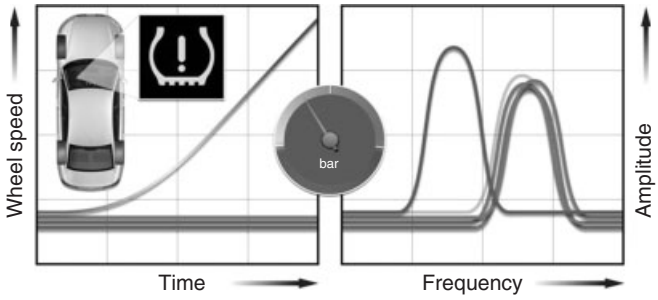


Figure 22. One-wheel puncture scenario.

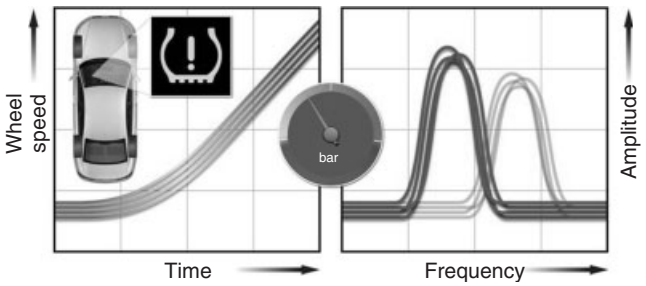


Figure 23. Four-wheel diffusion scenario.

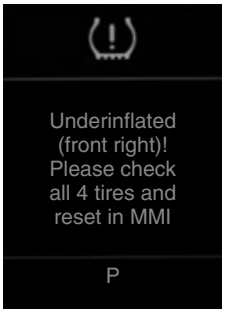


Figure 24. Under inflation warning in cluster.

difference can be seen and only wheel spectrum behavior changes detect under inflation (Figure 23).

This system design is aimed to increase robustness toward short-term disturbances and make it independent to individual tire sensitivities. The use of wheel-individual detectors also enables the so-called isolation functionality that describes the ability of a TPMS to deliver position information about affected wheels, resulting in the possibility to display this information to the driver (Figure 24). Another advantage is the possibility to parameterize and adjust—or even deactivate—detectors individually and by that adapt the system specifically to the target vehicle and its operational/display concept as well as different functional requirement specifications. Additional subfunctions cover self-diagnosis capabilities, for example, in case of invalid or erroneous input signals.

3.5 Software integration versus standalone solution

Indirect tire pressure monitoring does not need additional hardware as it only consists of a software module, which is to be executed in the vehicle. Early applications of iTPMS required one additional stand-alone control unit. Examples for that setup are the current 2007 Audi TT or 2008 Audi A6. As this setup does not support one of the most important advantages of iTPMS—no additional physical parts—the integrated solution is preferred. The setup of an indirect system as software module integrated in the ESC software is shown in Figure 25.



Figure 25. System configuration for second-generation indirect TPMS in an 2012 Audi A6.

As the wheel speed signals are essential for any iTPMS, the selection of suitable host systems is limited to systems, which supply the wheel speed signals in sufficient quality

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and reliability. Wheel slip control systems therefore represent the natural choice to host iTPMS as they are also highly reliable and provide sufficient hardware resources, software structures, and interfaces. Adoption of standards such as AUTOSAR (automotive open system architecture) is becoming common for both iTPMS and slip control system and simplifies the software integration process. The iTPMS software module is normally embedded or wrapped into a so-called middleware with the main task to control iTPMS execution and transfer input and output signals to and from the iTPMS module. It also controls the use of shared resources such as RAM (random access memory), nonvolatile memory (EEPROM, and electrically erasable programmable read only memory), and diagnostic functions such as system state handling, malfunctions, and diagnostic trouble codes (DTCs).

The integration of an iTPMS is closely linked to the HMI, operational and display concepts, and requirements of the target vehicle. For example, the system integration needs to consider whether position information of underinflated tire is signaled to the driver via additional text messages or if the reset procedure is conducted via button or interactive interface.

3.6 Reset and learning phase

iTPMS as relative measuring systems require the external setting of a reference point, usually by the driver. As soon as relevant parameters such as tire pressure, mounting positions, balancing, or tire types have changed the indirect system needs to be reset to function properly. To prevent the reset function from being executed unintentionally, physical switches (buttons) are usually placed in locations where they cannot be activated without driver awareness (e.g., glove compartment). Additional measures such as a confirmation procedure via interactive menu guidance or plausibility checks, with a minimum vehicle standstill to allow resetting, are common practice.

Once the reset request is accepted by the iTPMS, the existing reference data are erased and the iTPMS starts collecting reference data in the speed intervals visited while the vehicle is driving. Collected reference data are stored in nonvolatile memory and continued learning takes place over several driving cycles if necessary until the learning phase is completed. Different strategies on when to issue full iTPMS functionality, depending on the learning phase completion, are common. Most iTPMS gradually become active depending on the amount of collected reference data and the current speed interval. Full warning sensitivity will be reached after approximately 20 min of driving. Complete learning, covering all speed ranges usually takes about 1 h driving, assuming all speed intervals have been covered. Learning progress also depends on driving style and external road conditions. Additional functionalities, for example, speed extra- and interpolation, allow detection capabilities in speed intervals not yet visited. The reset of the iTPMS is necessary for proper performance and also eliminates numerous disturbing influences such as tire wear, tire aging, seasonal temperature changes, and/or the use of different tire types.

3.7 Application of indirect TPMS

As iTPMS use the tire as a sensor and the fact that tires are not standardized regarding their sensitivity to underinflation, covering a wide range of different tire characteristics, one iTPMS setting is the major application task. Current research activities are under way to take tire sensitivity to under inflation into consideration already when designing tires. This will generate uniform responses over a certain tire program and thus considerably reduce application effort of iTPMS. In addition, chassis or driveline influences can have to be considered when doing iTPMS application. It is therefore important to adjust and adapt the iTPMS in the target vehicle toward the target tire program. As iTPMSs are following the open-loop concept, they can effectively be tested using vehicle test recorded data files covering all input signals. These data files can be replayed off-line using different iTPMS software versions and parameter settings without repeating the actual vehicle test (Figure 26). The application of an iTPMS normally consists of one or more structured data collection campaigns, where relevant target vehicle model variants and tires are tested under various conditions and the input data being collected (detection and robustness tests, winter- and high temperature testing, road surface screenings, etc.). The data files are then replayed off-line and complemented with further vehicle tests on demand. As soon the database status makes it possible, algorithms and parameter settings are iteratively optimized until the test results meet the OEM requirement specifications. Using automatic test environments drastically reduces optimization loop time.

4 LEGAL ENVIRONMENT

After a series of fatal accidents, which were related to tire tread separation on certain Firestone tires mounted to the Ford Explorer in 2000, the US Congress enacted the so-called TREAD act and subsequently a new FMVSS 138, requiring new light vehicles [passenger cars, multipurpose passenger vehicles, trucks, and buses with a maximum

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

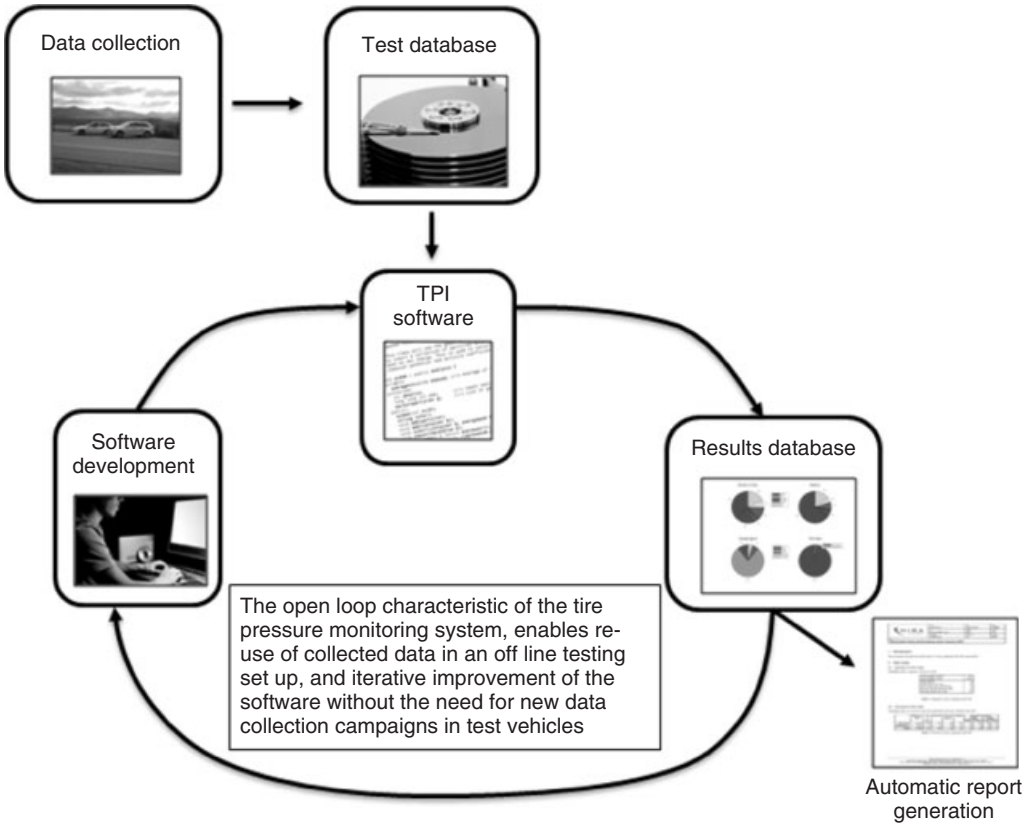


Figure 26. Representation of open-loop concept for iTPMS application.

gross vehicle mass rating of 4536 kg (10,000 pounds) to be equipped with a TPMS]. After a phase-in period starting in 2005, from September 2007, all new vehicles sold in the United States need to be equipped with TPMS conforming to FMVSS 138.

The safety goal of this regulation was to reduce accidents caused by tire failures that in turn were caused by significant underinflation. FMVSS 138 requires a warning to the driver, when the tire pressure falls below 25% of the cold recommended pressure level in one up to all four tires. In general, FMVSS 138 is a technologically neutral drafted performance standard, that is, it does not prescribe specific technology for fulfillment. Besides underinflation detection, FMVSS 138 also requires TPMS malfunctions to be signaled to the driver. A low tire pressure condition will be displayed by illumination of the standardized low tire pressure telltale (Figure 27). A number of OEMs chose to amend the minimum telltale requirements of FMVSS 138 with additional text messages to inform the driver about the situation and point him or her toward corrective actions.



Figure 27. TPMS telltale in cluster.

In addition to the safety aspect of tire pressure monitoring covered in FMVSS 138, in 2009, the EU parliament decided mainly with focus on reducing CO_2 emissions to require TPMS for all new vehicles with the new regulation ECE-R 64. The regulation applies to M1 and N1 vehicles (passenger vehicles and vehicles designed to carry goods

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

with a maximum mass not >3.500 kg). From November 2012, all new types and, from November 2014, all newly registered vehicles need to comply with ECE-R 64. ECE-R 64 requires a low tire pressure warning at 20% pressure loss with respect to the "warm tire pressure." Depending on the pressure loss scenario, ECE-R 64 requires different warning times, 10 min for a puncture on one wheel and 60 min for simulating the diffusion on all four wheels. The telltale and display requirements are mainly carried over from FMVSS 138, and the malfunction detection time has been reduced to 10 min.

Compared to US regulation FMVSS 138, the European regulation ECE-R 64 represents a substantial reducing of warning thresholds, as not only the warning level has been decreased to 20% but also the pressure reference has changed from the recommended cold tire pressure to the in-service, operational or the so-called warm tire pressure. Actually, ECE-R 64 almost cuts the warning threshold in half compared to FMVSS 138. Figure 28 shows the pressure trace for FMVSS 138 and ECE-R 64 test procedures. Once the recommended tire pressure is adjusted, the vehicles need to be driven to allow the TPMS to learn current tire/pressure characteristics (calibration or learning phase). During driving the calibration phase, the tire will heat up and the pressure will be above the initially adjusted cold inflation pressure. Pressure build up is dependent on different influences such as driving style and axle load but usually reaches values of about 10–15% of the cold tire pressure. After the calibration phase, the pressure will be adjusted to the warning level (25% for FMVSS 138 and 20% for ECE-R 64), whereas the reference pressure for adjusting is the cold tire pressure (FMVSS 138) and the in-service pressure (ECE-R 64). Subsequently, the vehicle is driven in the detection phase, where the detection time requirements vary depending on the exact scenario tested. In FMVSS 138, the warning must be issued at the latest 20 min after deflation independent of the pressure loss scenario (deflation in one up to all four wheels), ECE-R 64 requires detection in 10 min for the one-wheel puncture and 60 min for the four-wheel diffusion scenario. Another important difference between US and ECE requirements is that the United States realizes a self-certifying regulation, whereas ECE-R 64 calls for a type approval/homologation process done by the OEM to allow for registration of vehicle.

All tire pressure-monitoring systems—except first generation indirect systems—have been shown to be able to fulfill current US and ECE regulations. A performance comparison between direct TPMS and iTPMS is shown in Figure 29.

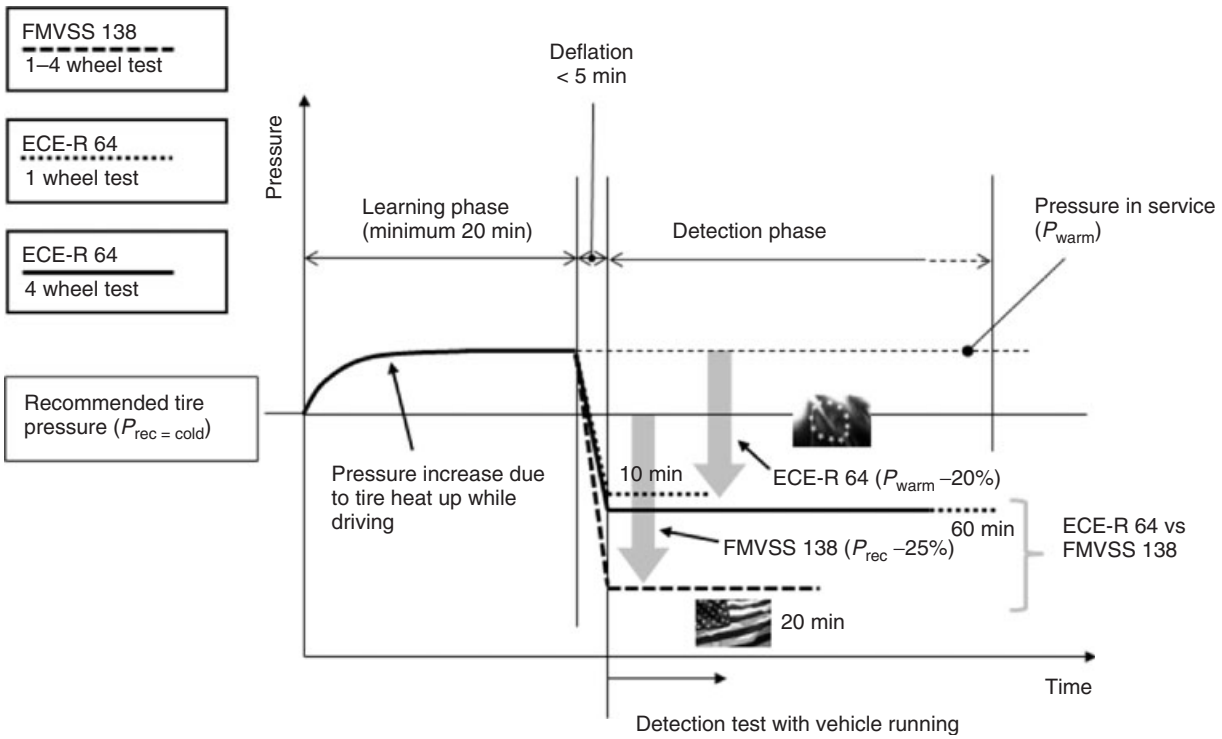


Figure 28. FMVSS 138/ECE-R 64 test procedure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

	Direct measuring TPMS	Indirect measuring TPMS	
		1st generation	2nd generation
Puncture detection	✓	✓	✓
Diffusion detection	✓	×	✓
Display of tire pressure	✓	×	×
Identifying puncture position	✓	×	✓
FMVSS 138 (U.S.)	✓	×	✓
ECE-R 64 (E.U.)	✓	×	✓

Figure 29. Performance matrix of direct/indirect TPMS.

5 FUTURE TRENDS AND DEVELOPMENT

Enacting legal requirements on tire pressure monitoring has and will further increase the market penetration of TPMS. After 100% mandating TPMS in the US market via FMVSS 138 since 2007, in 2014, the EU market will require 100% compliance with ECE-R 64.

On the side of direct systems, there is a clear trend toward reducing the number of components, complexity, and costs, most important for the sake of customer acceptance. US market experience has shown that, especially, the aftermarket has severe challenges to serve to the vast variety of direct systems currently available. Therefore, there is a clear trend toward standardization of hardware components and data transfer protocols, so systems from different OEMs/TPMS suppliers can interact in the future. It has also been understood that the issue of sensor replacement for direct systems over the lifetime of the vehicle poses a serious customer acceptance concern: market research has shown that only one-third of customers will replace sensors with empty batteries. Extensive research in the field of energy harvesting is on the way to eliminate the need for batteries in the pressure sensors of direct TPMS. Currently, the most promising approach for energy harvesting in the TPMS environment seems to be the deployment of the piezo-electric effect, but as of today, no energy harvesting system is in series production.

The topic of energy harvesting goes strongly connected to the implementation of the tire-integrated sensor (Figure 30). By attaching the sensor to the tire via gluing or curing, more functionality besides tire pressure monitoring could be realized. If permanently fixed to the tire, tire-specific data such as tire type (summer, winter, and run-flat) or speed and load index could be stored to the tire sensor, allowing for additional benefits: speed warnings could be issued when exceeding the speed index specified speed, or driver information could be provided if the tire type does not fit current weather conditions (summer tire running in winter conditions). In addition, the tire-integrated sensor could determine the size of the contact patch between road and tire, information that can be utilized when optimizing rolling resistance or determining wheel loads. Nevertheless, with regard to the tire-integrated sensors, not all issues are resolved yet. For example, the permanent sensor attachment necessary for realizing benefits beyond tire pressure monitoring has to address the issue of tire replacement when a tire is damaged or the tire life is reached because the tread has worn out. In this case, not only the tire but also the sensor will be disposed, which economically can only be done when sensor prices decrease substantially. On the other hand, making the tire-integrated sensor replaceable will not allow for realizing all possible benefits, because the stored tire information may not be correct because of tire changes. It is



Figure 30. Integrated tire sensor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

expected that the tire-integrated sensor will start its market introduction in the high price sports car segment, where the tire program is usually limited and special customer needs justify the special effort and price. At the moment, no tire-integrated sensor solution is currently available in series production.

As it has been proved that indirect systems fulfill legal requirements FMVSS 138 and upcoming ECE-R 64 legislation, the market share of iTPMS is expected to grow substantially over the coming years. The benefits of providing a robust tire pressure monitoring functionality with reduced hardware and complexity in combination with no follow-up costs over the vehicle lifetime will appeal to mid-segment, cost-conscious, and premium-segment OEMs alike. For example, Audi has implemented the indirect system into almost all the vehicles in its model line up.

ACKNOWLEDGMENTS

We would like to thank Ralf Kessler and his team of HUF Electronics Bretten for significant contributions to the segment of direct systems. Another special thanks goes to Volker Rühr of Dr. Ing. h.c.Porsche AG and Dr. Andreas Köbe of Continental Engineering Services GmbH who supported on the topic of tire pressure monitoring system introduction and history.

RELATED ARTICLES

μ-identification by Tyre Measurement (Apollo and Friction)
Performance Target Conflicts in Normal Tires and Ultra High Performance (UHP) Tires
Brake Systems, an Overview
Tyre Modelling
Batteries
Interfaces between Sensors and ECUs
Various Types of Sensors
Applications of Radio Wave Technologies to Vehicles
Chassis ECU (Vehicle dynamics, ABS)

FURTHER READING

Basbantu, G., Pellicciari, M., and Andrisano, A. (2004) On the tire monitoring systems temperature compensation. SAE-paper, 2004-01-110.

Bosch Kraftfahrttechnisches Taschenbuch (2007) *Reifendruckkontrollsystem*, S. 810/811, 26. Auflage, Friedr. Vieweg & Sohn Verlag, Wiesbaden, Germany.

Fischer, M. (2003) *Tire Pressure Monitoring, Die Bibliothek der Technik*, vol. **243**, Verlag moderne Industrie, Landsberg.

Folger, J., Riedl, H., and Wallentowitz, H. (1989) Electronic Tire Pressure Control. *International EAEC Conference*, Strasbourg.

Greenly, C. and Beverly, J. (2005) Concerns related to FMVSS No. 138: tyre pressure monitoring systems and potential implementation of a similar standard on commercial vehicles. SAE 2005-01-3517.

Kowalewski, M. (2004) Monitoring and managing tire pressure, *IEEE Potentials*, **23** (3), 8–10.

Marshek, K. and Cudermann, J. (2002) Performance of anti-lock braking systems equipped passenger vehicles – past 111: braking as a function of tyre inflation pressure. SAE 2002-01-0306.

Minf, K. (2001) A smart tire pressure monitoring system, *Sensors*, **18** (11), 40–46.

NHTSA (2005) Federal Motor Vehicle Safety Standards—Tyre Pressure Monitoring Systems, FMVSS No. 138, Final Regulatory Impact Analysis.

Paine, M., Griffiths, M., and Magedara, N. (2007) The Role of Tyre Pressure in Vehicle Safety, Injury and Environment, Road Safety Solutions, Caringhah, NSW, Australia.

Persson, N. and Gustafsson, F. (2002) Indirect tyre pressure monitoring using sensor fusion. SAE 2002-01-1250.

Pohen, F.-H. (2009) Entwicklung einer radgebundenen Reifendruckregelanlage für landwirtschaftliche Fahrzeuge. Dissertation RWTH Aachen, 2009 Forschungsbericht Agrartechnik Nr. 482, Shaker Verlag, Aachen.

Umeno, T., Asano, K., Okashi, H., *et al*. (2001) Observer based estimation of parameter variations and its application to tyre pressure diagnosis, *Control Engineering Practice*, **9**, 639–645.

Wallentowitz, H. and Reif, K. (eds) (Hrsg.) (2006) *Handbuch Kraftfahrzeugelektronik*, S. 513 f, Vieweg Verlag, Wiesbaden, Germany.

Williams, R. (1992) DWS – ein neues Druckverlust-Warnsystem für Automobilreifen, *Automobiltechnische Zeitschrift (ATZ)*, **94**, 336–340.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto014
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The Harshness of Air Springs in Passenger Cars

Andreas Kind and Andreas Rohde
Continental Teves AG & Co. oHG, Hannover, Germany

1 Introduction 1
2 Harshness of Air Springs 1
3 Harshness of Different Types of Air Suspension 14
4 Outlook 17
5 Summary 17
References 17
Further Reading 17

1 INTRODUCTION

Air suspension systems have found widespread application in the luxury automobile segment in recent years. The advantages of air suspension systems in terms of load leveling, variable spring rates, and comfort have been greeted by a high degree of acceptance among consumers.

Customers in the luxury segment demand and expect constant refinements to driving comfort more than anything else and so the goal is to reduce vibration and noise to a level that is acceptable to passengers. This has resulted in permanent refinement and improvement of chassis components.

The comfort of air suspension can be determined mainly by the curve of the spring rate in relation to the frequency and amplitude of excitation. Harshness in this context refers to the undesirable stiffening of air suspensions in response to small, rapid vibration. This effect is primarily due to the characteristics of the bellows employed. Steps to improve comfort can flow into the development process as a function of the design of the air suspension system, the geometry of the components, and the materials that have been employed.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 HARSHNESS OF AIR SPRINGS

2.1 Ride comfort

The demand for riding comfort in the passenger car segment is becoming ever more pronounced. While in the past, luxury cars were synonymous with excellent ride comfort, nowadays, middle-class automobiles, SUVs, and even sports cars come with expectations of comparable comfort. Most modern electronic chassis systems now offer the choice between comfort and sport modes. Changes in the suspension and dampers occur, depending on the input received by the wheels. These changes then influence the vibratory characteristics and, as a desired result, the impression of comfort. The result is an ideal balance between vehicle dynamics and ride comfort (Figure 1).

Undesired vibrations nevertheless do occur, depending on the type and severity of input to the wheels. The jolts travel from the pavement through the wheel and the chassis into the body. Vibrations and structure-borne noise are thus transmitted into the interior and to the occupants. Ride comfort is a reflection of the occupants' sense of well-being as a function of all vibratory influences. Depending on its frequency, vibration can be experienced as palpable vibration, noise, or a combination of both. The term *NVH* is used to refer to vibrations specific to automobiles. NVH stands for noise vibration and harshness and encompasses acoustical and mechanical vibration as well as people's subjective perception of it (Ersoy and Heißing, 2008).

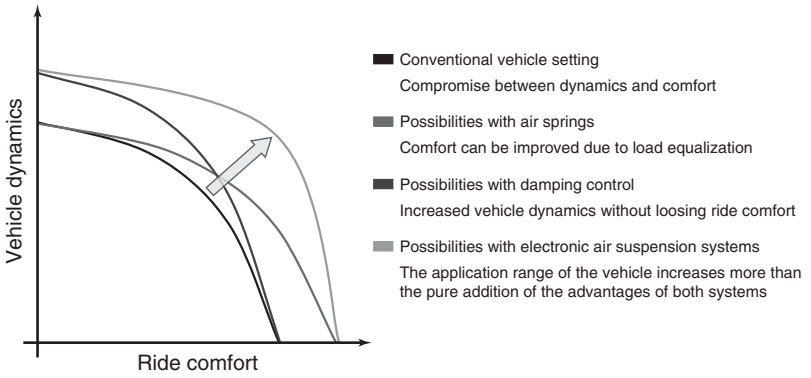


Figure 1. Increase of the chassis application range. (Reproduced by permission of Continental AG.)

2.2 A general definition of harshness

The term *harshness* refers to the intersection of vibrations and noise in a typical range of frequencies between 20 and 100 Hz. Harshness can also be interpreted as a kind of roughness. This roughness is elicited by an undesirable increase in the stiffness of the axle, which may lead to increased acceleration of the bodywork. The harshness effect is magnified at low amplitudes and high frequency excitation. Harshness is an unpleasant experience subjectively because the visual perception of the road often does not coincide with the rough ride the occupants are experiencing. For example, increased harshness can cause a ride over a road with an apparently smooth surface and not even the slightest unevenness visible to the human eye to become an uncomfortable experience because vibration and noise reign in the passenger compartment.

Axle components exert the main influence on harshness. The springs, the dampers, and all other components such as bushings and axle joints are responsible for transmitting vibration. The layout, and the resultant transmission characteristics of the components, thus exerts a direct influence on a vehicle's vibration. In addition, numerous elastomers are used because of the elastic-kinetic properties of the axle. These elastomers play a very important role because of their usually quasi-static and dynamic transmission characteristics and their nonlinear properties. Aside from the material properties, the design of the component is critical. Unevenness and abrupt transitions in the force–travel ratio lead to a marked deterioration of the impression of comfort.

Friction effects are still another leading cause of diminished comfort. On the one hand, there is material friction within the components themselves. On the other hand, friction also arises because of the relative movement of components in contact.

2.3 The basics of air suspension

The actual air suspension element is the air spring bellows. In passenger cars, it appears exclusively as a rolling sleeve-type. Figure 2 depicts a rolling sleeve-type. Metal rings clamp the bellows between the cap and the piston. Under pressure, the bellows roll over the rolling lobe on the outer piston geometry during axial motion.

Load capacity F of the air spring is a function of the pressure p_i inside the air spring and the effective area A_W (Equation 1).

$$F = p_i A_W \tag{1}$$

The effective area A_W is proportionate to the effective diameter (Equation 2):

$$F = \frac{\pi D_W^2}{4} \tag{2}$$

The spring rate c is derived from the load capacity F according to the distance s traveled by the air spring (Equation 3).

$$c = \frac{dF}{ds} \tag{3}$$

The rolling sleeve-type consists of thin layers of elastomer between which reinforcement materials in the form of individual threads or mats of fabric have been embedded. Depending on the application, reinforcement material can cross in two layers or can be arranged in one layer along an axis. One is known as a *cross-layered bellows* and the other is known as an *axial bellows*. The reinforcement material can generally only absorb the forces generated in the bellows in the direction of the threads. Cross-layered bellows can therefore also absorb circumferential forces as a function of the angle α between the strands and hence can assume a defined diameter on the application of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

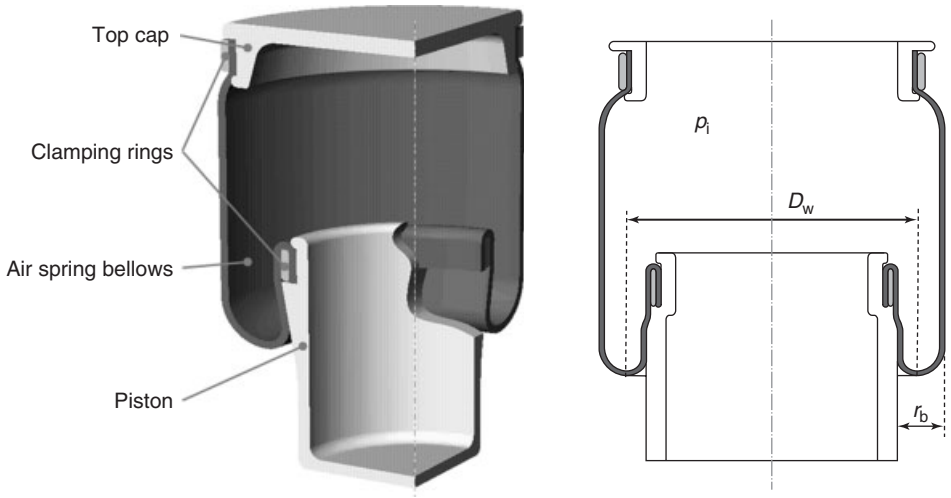


Figure 2. Construction of an air spring bellows suspension unit with pertinent parameters.

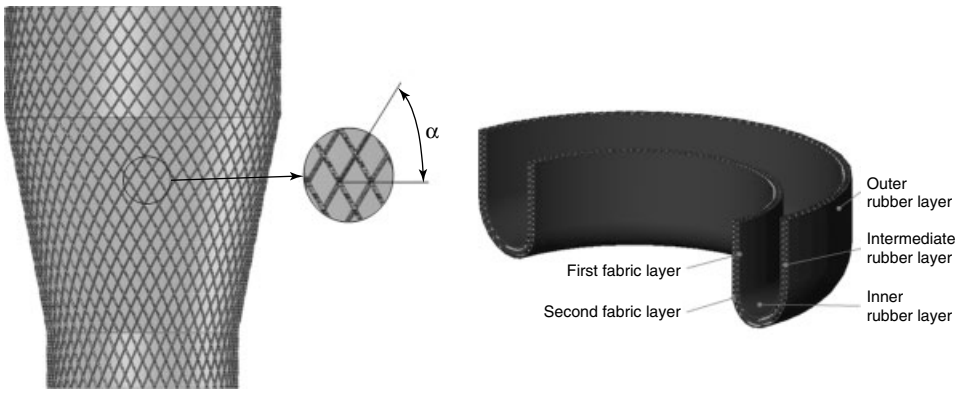


Figure 3. Cross-layered bellows and rolling-lobe shape. (Reproduced by permission of Continental AG.)

pressure. Figure 3 is a schematic diagram of a cross-layered bellows. The five-layer design consists of three elastomer layers—inner, intermediate, and outer rubber layers—and two fabric layers embedded in between. Cross layering results in an angle-dependent rhomboid pattern.

As opposed to axial bellows with a single fabric layer of 90°, the angle of the strands when manufacturing the blank for the cross-layered bellows generally lies somewhere between 45° and 70°, depending on the application. The original angle diminishes as manufacturing proceeds, first because of the bellows widening during vulcanization and shaping and then during pressurization in the static operating state. Here, an equilibrium angle and, as a function thereof, the exterior bellows diameter adjust themselves. The fabric strands stretch under pressure because of their elastic character. During operation, the outer diameter is subject to fluctuations depending on the type of the material used in the fabric. This is due to the magnitude of pressurization. These fluctuations must be considered when designing the component.

During spring motion, the bellows wall does not remain rigid; it fluctuates between the small piston diameter and the large exterior diameter. As the rolling lobe passes through, constant deformations of the bellows wall occur, as do modifications of the strand angle. The pressure inside the air spring causes the bellows wall to roll smoothly down the piston (Voss, 2002).

Alternatively, the air spring can be guided with a fixed external cylinder, which takes up the circumferential forces and diminishes stretching of the bellows wall. This is known as an *exterior-guided air spring*. Both cross-layered bellows and bellows with axial fabric layer find application in guided air springs. The advantages of guided air springs are that there is a lot less wear and tear on the bellows

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

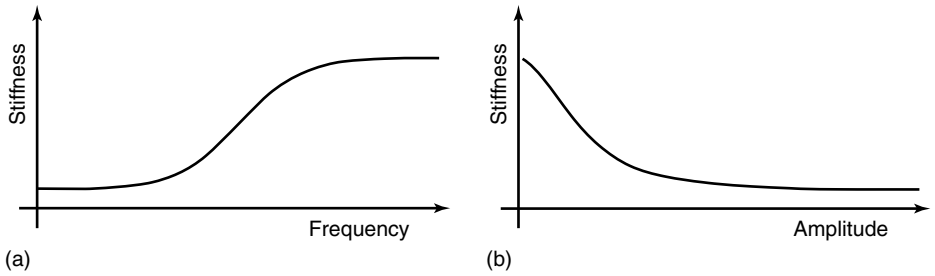


Figure 4. General stiffness curve on an air spring as a function of frequency (a) and amplitude (b).

and they guarantee a fixed diameter for design purposes. The bellows, primarily the reinforcement material and the rubber layers, can be designed much more thinly and delicately.

2.4 The harshness of air suspension

Harshness in connection with air suspension means that spring characteristics stiffen with specific axial movements. Increased stiffness is generally based on two main effects:

- *Frequency-dependent increase in stiffness*. In the direction of high frequencies due to thermodynamic changes in the air.
- *Amplitude-dependent increase in stiffness*. In the direction of low amplitudes due to the rolling resistance of the bellows.

Figure 4 illustrates the general stiffness curves of both effects.

The amplitude-dependent increase in stiffness plays the main part in the actual undesirable harshness of air suspensions. It is predominantly frequency independent and thus significant with every input from the road. The frequency-dependent increase in stiffness, on the other hand, occurs predominantly at frequencies that lie below (i.e., at frequencies below 0.1 Hz) those that are relevant for normal driving conditions (from 0.1 to 2 Hz). Therefore, the stiffness has already increased when the driving relevant frequencies are reached. What follows is a brief explanation of this behavior.

2.4.1 *Frequency-dependent increase in stiffness (due to thermodynamic changes)*

One can distinguish between static (isotherm) and dynamic (adiabatic) behaviors due to thermodynamic changes in the air. Depending on the rapidity of these changes, the following static equation applies (Equation 4),

$$pV = \text{constant} \tag{4}$$

The conditional equation (5) for the adiabatic case with $\kappa = 1.4$:

$$pV^{\kappa} = \text{constant} \tag{5}$$

The frequency-dependent increase in stiffness (Figure 4a) is marked by a transition from isotherm to adiabatic air behavior. It occurs at what is known as the *cutoff frequency* f' starting at approximately 0.1 Hz. The hysteresis of air suspension assumes a maximum value here. Force–travel curves as a function of different frequencies illustrate the marked degree of hysteresis in Figure 5.

For purposes of identifying the parameters of air suspension stiffness, test specifications thus call for sufficient distance from cutoff frequency f', mostly with test frequencies of 0.01 Hz (static) and 1 Hz (dynamic). These frequencies do not quite attain purely isotherm or adiabatic behavior but have nevertheless proved themselves to be of value for technical reasons.

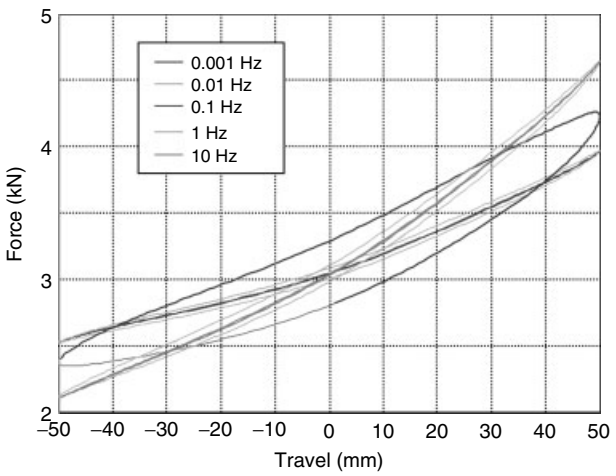


Figure 5. Force–travel curves of air suspension as a function of frequency. (Reproduced from Ilias and Sorge, 2001. Reproduced by permission of Continental AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

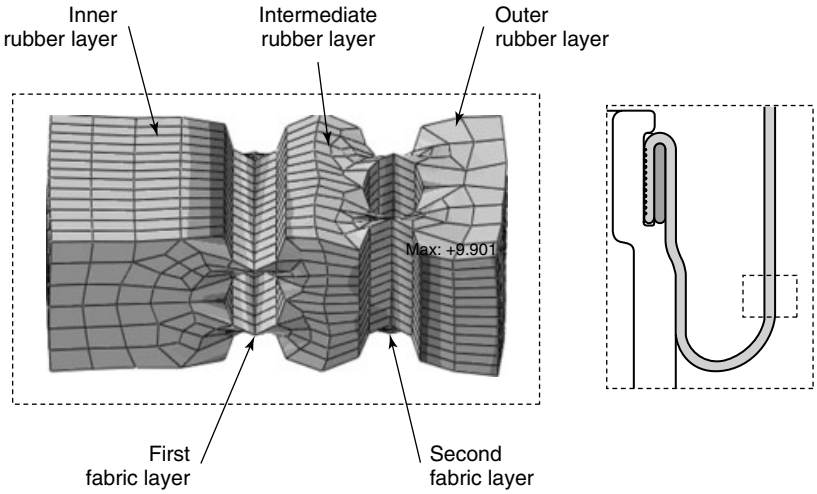


Figure 6. Extended bellows (submodel rolling lobe outside). (Reproduced by permission of Continental AG.)

Incorporating thermodynamic behavior, the value c_{spring} can be calculated as follows in Equation 6;

$$c_{\text{spring}} = p_{\text{i}} \frac{\mathrm{d}A_{\text{W}}}{\mathrm{d}s} + \frac{n(p_{\text{a}} + p_{\text{i}})(A_{\text{W}}^2)}{V} \tag{6}$$

The first term represents the spring value as a function of the change in the effective diameter. The second term in the equation reflects the thermodynamic portion of the spring value. For purely static spring action, $n = 1$ applies, whereas n to the isentropic coefficient $\kappa = 1.4$ results for purely adiabatic behavior. It is thus apparent that the consequence of dynamic spring action is increased spring rate as a function of the spring volume V.

Dynamic spring action is the main factor in a car body's natural frequency and hence in riding comfort. Static spring action, on the other hand, is responsible for pitch and roll. Static spring values are therefore also important for good handling because they have an influence on the car's stability in curves. It is thus worth trying to keep the difference between static and dynamic spring values as small as possible.

2.4.2 Amplitude-dependent increase in stiffness (due to the rolling resistance of the bellows)

The actual reduction in comfort that harshness causes results from low amplitude vibration. A classic example of this is rough, uneven asphalt on leaving the highway. However, driving over larger obstacles such as road joints, edges, or asphalt patches may result in an undesirable increase in stiffness. In addition to the harshness being affected by the bellows, often the harshness manifests itself on apparently smooth roads that have perturbations so small that the excitation cannot break through the suspension friction (including the rubber bushings of the suspension arms), resulting in the vehicle bouncing only on the tires in the frequency range of approximately 4–8 Hz and giving what Americans call "boulevard jerk." In the air spring system itself, the reason for the harshness is that the bellows pits a certain rolling resistance against axial motion once it leaves a static resting position. The flexible wall of the bellows must deform during expansion and contraction. As it passes through the rolling lobe, a complex deformation of the elastomer compound and the embedded reinforcement material occurs (Voss, 2002). Figure 6 shows how the rubber matrix stretches at the outer rolling lobe.

Added to the basic stiffness of the air suspension c_{spring} is consequently the portion of stiffness Δc_{harsh} from the rolling resistance of the rolling lobe. The result is the total amplitude-dependent stiffness of the air suspension c^*_{spring} (Equation 7):

$$c^*_{\text{spring}} = c_{\text{spring}} + \Delta c_{\text{harsh}} \tag{7}$$

2.5 Measuring harshness

Objective test methods do a good job of characterizing the harshness of air suspensions. To do this, they record the force–travel curves of an air spring at various amplitudes under harmonic sinusoidal stimulation at a frequency of mostly 1 Hz (Figure 7). The corresponding rates of stiffness are then derived from the observed data via the attendant slopes passing through zero (Figure 8). Amplitudes fall typically in the range 0.1–25 mm.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

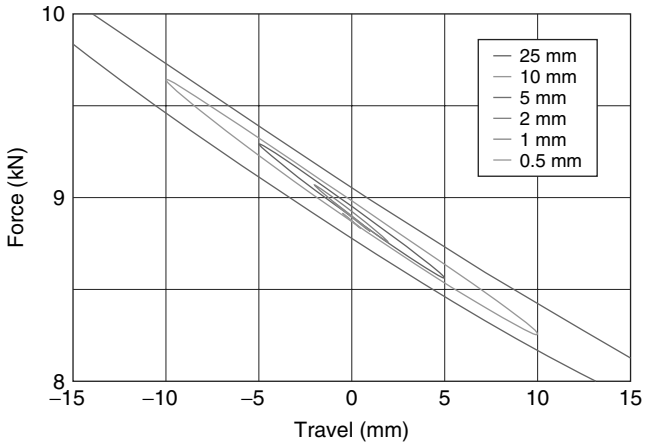


Figure 7. Force–travel curves as a function of amplitude at $f = 1$ Hz.

While the basic design stiffness of air springs regulates at high amplitudes of 25 mm, the spring rate increases in the direction of lower amplitudes. The hysteresis curves rise more steeply as they characterize the harshness effect along the stiffness curve. The harshness effect occurs largely frequency independent (Figure 8).

Measurement using actual air springs, as mentioned earlier, always includes thermodynamic changes. An alternative means of measurement is to attach the metering device to what is known as a *double rolling-lobe arrangement* in order to measure the pure resistance to rolling of the bellows during axial movement.

The device consists of two opposing, geometrically identical air springs with a cylindrical roll-off contour (Figure 9). The volumes of both are linked to each other. The double rolling-lobe arrangement is outwardly force free and allows direct measurement of the rolling resistance of both bellows. A direct, subjective impression "by hand" of harshness or the hysteresis forces is also possible. This arrangement thus permits unadulterated assessment of the components without intrusion by factors influencing the entire module.

Figure 10 illustrates the curve of forces and stiffness on the double rolling-lobe arrangement. While the hysteresis forces diminish as the amplitude declines, the force curve rises more steeply, and stiffness increases.

For the purpose of comparison and obtaining test results that can be reproduced, it is essential to precondition the components extensively and directly before recording any

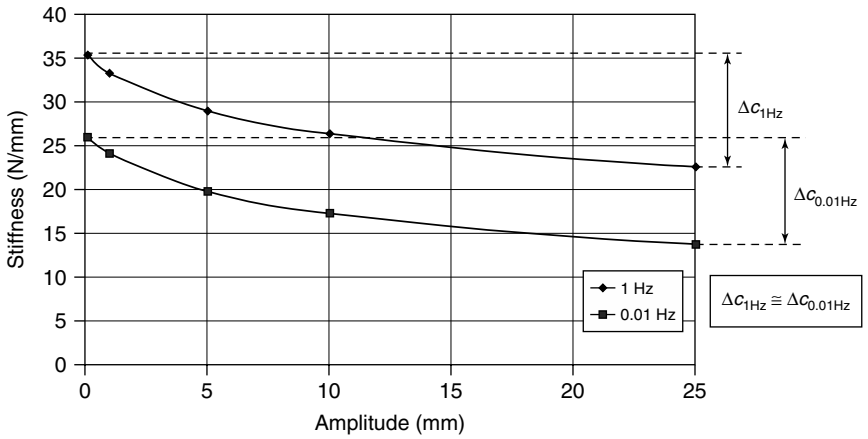


Figure 8. Stiffness as a function of amplitude at $f = 0.01$ and 1 Hz.

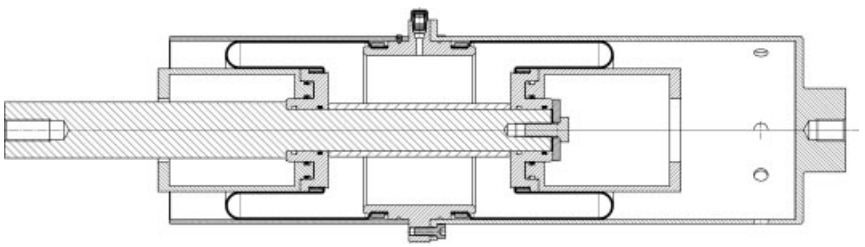


Figure 9. Design for a double rolling-lobe test bench. (Reproduced by permission of Continental AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

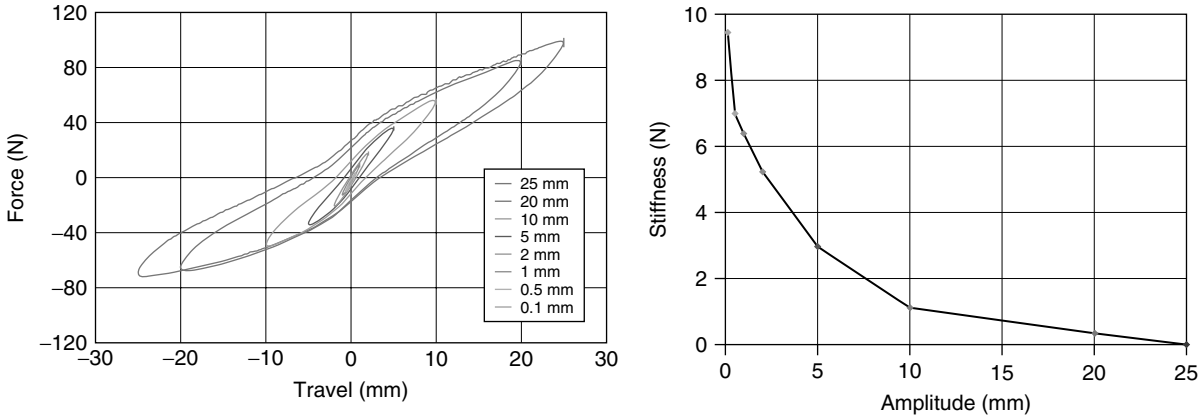


Figure 10. Hysteresis curves with deduced stiffness progress measured on a double rolling-lobe arrangement.

signals. In the case of air springs, springs are usually preflexed to the maximum for a large number of cycles.

2.6 Assessment of harshness

2.6.1 Assessment using harshness coefficients

A coefficient is often used to assess harshness. On the basis of the amplitude response (see Chapter 2.5 "Measuring harshness"), the harshness coefficient $k_{\text{harshness}}$ in Equation 8 sets the relation of increase in the spring rate in relation to the basic spring rate $c_{25\text{mm}}$ of the air spring.

$$k_{\text{harshness}} = \frac{c_{0.1\text{mm}} - c_{25\text{mm}}}{c_{25\text{mm}}} \cdot 100\% \qquad (8)$$

Use of the coefficient, however, insufficiently reflects the actual typical behavior of the component. For example, an air suspension with a high degree of harshness associated with a high basic spring rate would presumably rate better under this equation than a highly comfortable air suspension with an associated very low basic spring rate that might be brought into play by a direct final axle ratio.

The harshness coefficient $k_{\text{harshness}}$ separated from the actual curve of extreme stiffness increase only makes sense in conjunction with the assessment of a specific vehicle or a specific axle.

2.6.2 Assessment in view of the stiffness curve

Simple observation of the curve of amplitude-dependent stiffness leads meanwhile to an optimized, independent assessment of harshness. Two criteria in particular merit our attention:

1. The increase in spring rate.
2. The deflection of the curve.

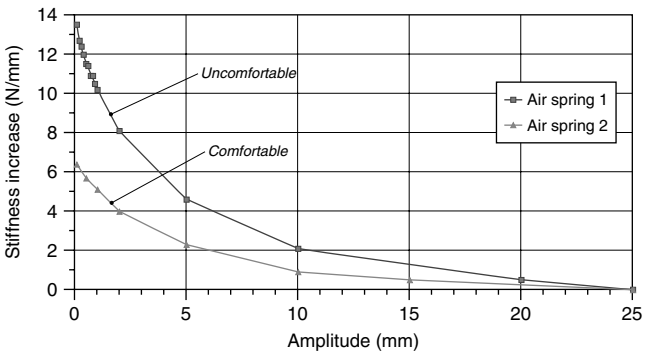


Figure 11. Illustration of amplitude-dependent increase in spring rate Δc_{harsh}.

The increase in spring rate and/or the leptokurtosis of stiffness Δc_{harsh} normalizes the amplitude-dependent spring rate by the value of maximum observed amplitude.

This sets the increase in spring rate at maximum amplitude to zero. This also considers the simple increase in spring rate. This depiction makes it easy to compare air suspensions with varying basic spring rates and facilitates discussion of their curves. Figure 11 illustrates the qualitative progression of a comfortable and an uncomfortable air suspension.

Another important assessment factor is the deflection of the curve in the form of a bulge. A slight increase in spring rate associated with a prominent bulge, that is, a rapid reaching of basic stiffness, is necessary for the comfort of air suspensions.

2.6.3 Correlation to subjective driving impressions

To what extent, there is a correlation to the increase in spring rate observed and its associated curve, in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

comparison to the subjective driving impression, depends greatly on the total automobile handling. There is thus no single specific answer to this question. Assessment of harshness on the test bench may reliably quantify the behavior of the component on the one hand. On the other hand, though, it can only provide a qualitative indication of how comfortable the vehicle is. The preconditioning performed on the components with large amplitudes (to ensure that the results can be reproduced and compared) does not correspond to the way the component actually behaves in the vehicle. What is missing is permanent high amplitude flexing. Uniform harmonic excitations also occur extremely rarely. It is therefore vital to constantly compare the observed, objective component characteristics with subjective driving impressions in order to accurately assess the comfort of air suspensions. Objective measurement on the vehicle can proceed in parallel with the aid of applied metrology and vibration profiles that can be reproduced. One example would be acceleration measurements on selected test tracks. This testing is expensive, however.

2.7 Causes and influencing factors

2.7.1 Causes of harshness

Characteristic of the increase in spring rate at low amplitudes is the *rolling resistance of the rolling lobe*.

This resistance to rolling that the bellows pits against axial motion results at the macroscopic level from the stiffness of the bellows wall at the rolling lobe. The bellows wall, acting as a stressed membrane under internal pressure, exerts a certain stiffness that influences the characteristic of the force–travel hysteresis. This portion of the bellows wall that stiffens the spring rate is the harshness effect. Among the causes is the mobility of the molecular chain, a characteristic of polymers (Voss, 2002).

To activate the spring, energy is necessary to deform the bellows wall in the rolling lobe. Various force components that occur while passing through the rolling lobe determine the amount of energy needed. Bending/flexing forces (as described by Thurow, 1995) that remain constant throughout the stroke are characteristic on the one hand. On the other hand, elastic and retracting forces occur that determine, as a function of amplitude, how much the material in the rolling lobe stretches. Viewed through a microscope, the rubber mainly deforms between and within the fabric layers during a rolling-lobe cycle. Figure 12 illustrates simulation of a bellows section of an unguided air spring transitioning from the interior to the exterior of the rolling lobe. The rubber stretches much more, particularly near the fabric layer.

Sufficiently large travel amplitudes cause a nearly constant resistance to rolling in the bellows while hysteresis remains constant. As amplitude diminishes, only a partial stroke of the rolling lobe occurs. The associated hysteresis force decreases. By contrast, the hysteresis curve rises more steeply, causing the spring rate to rise. The reason is that diminishing deformations within the elastomer matrix (due to its material properties) elicit

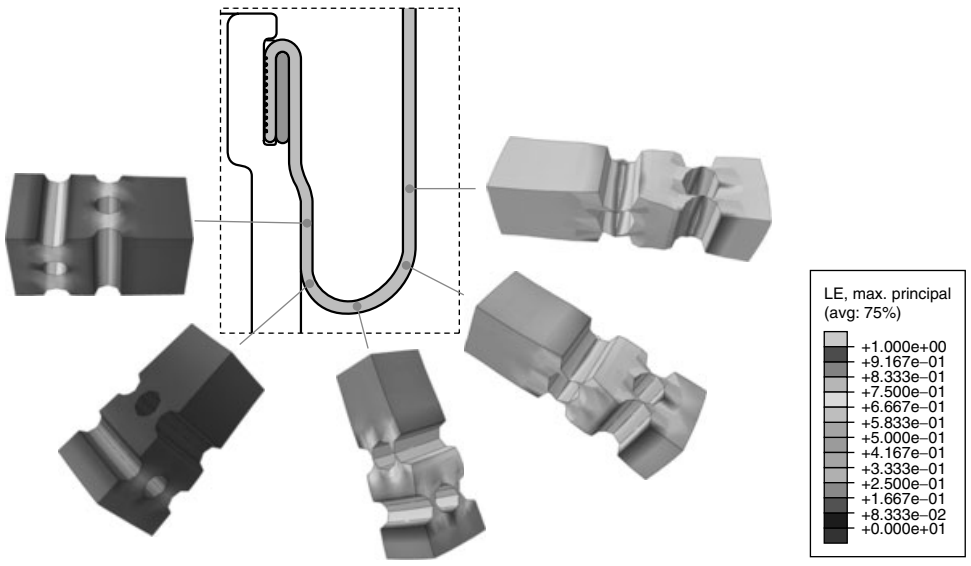


Figure 12. Stretching in the rubber matrix of a bellows submodel during a rolling-lobe cycle. (Reproduced by permission of Continental AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

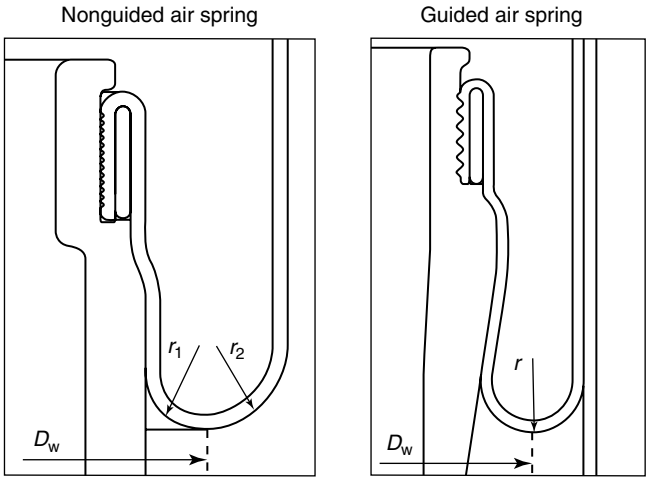


Figure 13. Geometry of the rolling lobe in an unguided and guided bellows.

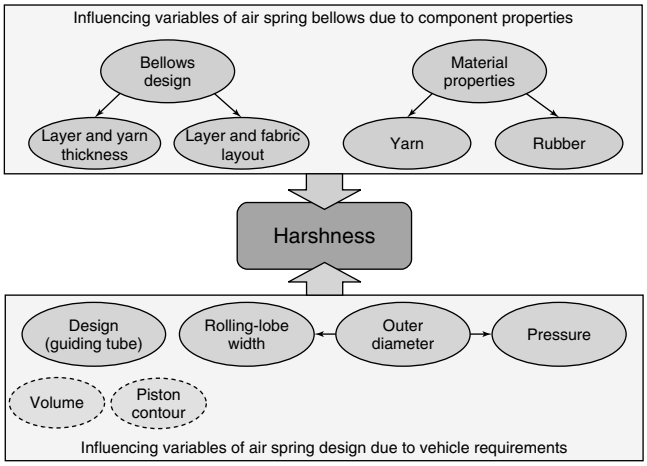


Figure 14. Factors influencing harshness.

the rising, nonlinear changes in force. Moreover, the resistance against the downward rolling motion results in deformation of the rolling lobe's geometry. A shift in the effective diameter may occur, independent of the piston geometry, depending on the stability of the rolling lobe. The result of this effect is an accompanying change in force, which plays a major part, primarily in unguided air springs. Figure 13 provides a detailed view of the rolling lobe's geometry with both a guided and an unguided air spring. The fold factor f_f describes the position of the lowest point of the rolling lobe and thus the effective diameter. With unguided air springs, the fold factor is subject to some variation because of the bellows design selected in conjunction with the material properties of the reinforcement material. The curvatures in the rolling lobe, simply described by the radii r_1 and r_2, do not act constant when rolling motions commence.

With guided air springs, the rigid outer guides largely prevent a change in the effective diameter. The rolling lobe remains more stable in its geometric form, and the characteristic geometry is more capable of retaining a constant curvature with a constant fold factor.

The bellows' resistance to rolling can be accompanied by a certain learning effect after it has stood still for a while. The persistence of the rolling lobe in one position when pressurized, as a function of time, increases resistance to any new motions. The elastomer's material memory "stores" a default position because of deformation in the matrix with the reinforcement material. Age can reinforce this effect. Added to this is the geometric formation of the bellows during vulcanization. A certain number of sufficiently large amplitudes are necessary to break down the resistance of the remembered default position.

Determining the effects of resistance to rolling demonstrated here are mainly the material properties of elastomer and yarn, plus the geometric relation of the whole assembly in conjunction with the kinematics of the rhomboid network in the rolling lobe. As a consequence, it is not impossible to cite a single specific factor for purposes of assessment. A multitude of interactive factors are responsible for the comfort of air suspensions. Figure 14 illustrates the main factors that influence comfort.

As Figure 14 demonstrates, influences from the specific properties of the bellows on the one hand, and influences from the air spring design due to the specific vehicle requirements are chiefly responsible for resultant harshness. The following discussion treats both influence groups in detail.

2.7.2 Influence of specific component properties of the bellows

2.7.2.1 Influence of the material properties of elastomer. Influencing the dynamic deformation behavior of an elastomer at low amplitudes are the fillers used in creating it. Fillers such as carbon black or silicic acid greatly improve physical properties of rubber compounds such as strength and durability.

Aside from the basic elastomer properties, the addition of fillers also leads to increased viscoelastic behavior on the part of the material. If sinusoidal deformation occurs, there will be a phase shift between stretching/shearing and tension. The complex shear module G^* can be subdivided into a memory module G' and a dissipation module G'' (Equation 9). The memory module G' is a measure of the stored elastic energy that is regained within a deformation cycle. The dissipation module G'' is the measure of energy

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

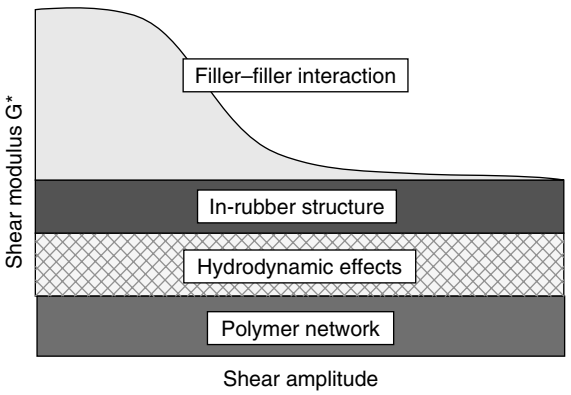


Figure 15. Payne effect. (Reproduced by permission of Joachim Fröhlich.)

dissipated as heat (Klüppel, 2007).

$$G^* = G' + iG'' \tag{9}$$

The ratio of both factors is known as the *dissipation factor* δ (Equation 10).

$$\tan\delta = \frac{G''}{G'} \tag{10}$$

The Payne effect describes the slope of the dynamic shear module G^* of a filled, vulcanized sample at an increasing amplitude of deformation (Figure 15).

The Payne effect adduces the breakup of the interactive filler compounds as the reason. According to Payne, other interactions between fillers and rubber, hydrodynamic effects of the filler particles or the properties of the pure polymer are largely amplitude independent. When the deformation retracts, these filler compounds rapidly reform. Figure 16 shows the breakdown of the interactive filler compounds using the example of a silica-filled elastomer.

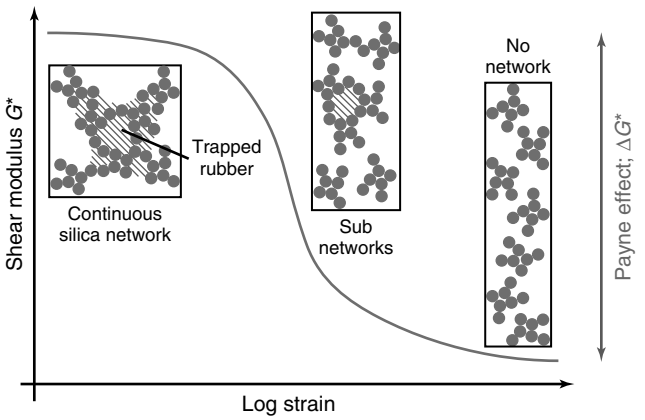


Figure 16. Payne effect regarding silica network. (Reproduced by permission of Joachim Fröhlich.)

The Payne effect and the characteristic viscoelastic properties become much more pronounced as the amount of filler increases. Figure 17 illustrates the decline of the memory module G' toward large deformations, plus the bulge of the dissipation module G'' as a function of the silica content.

There have been different interpretations of the Payne effect and its causes. The basic explanation of the material behavior of the elastomer ingredients nevertheless follows Payne's original core idea. Interactions occur not only within the filler material but also in the compound of the filler with the polymer matrix (Boehm, 2001). The literature generally speaks also of internal material friction based on a type of breakaway effect analogous to contact–friction phenomena.

Filled elastomers find application mainly in automobile air suspensions. The use of CR (chloroprene rubber), a chlorinated elastomer, is widespread. With the help of suitable ingredients, its properties render it capable of satisfying many of the requirements made on air suspension components in automobiles. Carbon black and silica are the main fillers used in various mixtures. The Payne effect is quite apparent here as well and explains the amplitude-dependent behavior of the increasing spring rate.

Alternative rubber compounds also come under consideration for improving comfort if one is willing to compromise regarding requirements for diverse automotive components. Natural rubber (NR), for example, offers great potential. A combination of various elastomer compounds for the bellows is also possible, depending on the function and the location in the bellows.

2.7.2.2 Influence of the material properties of the reinforcement material. Polyamide (PA) and, to a lesser extent, polyester (PES) and aramid find application as reinforcement materials in the form of individual strands of yarn or rolled fabric mats in automotive air suspensions. PA adheres very well to rubber and is inexpensive but exhibits rather low strength when highly stretched. The strength of PES is comparable to that of PA while its expansion properties are superior, especially with regard to temperature and durability. Aramid, on the other hand, is extremely strong and rather brittle, which somewhat limits its application because of its sensitivity to compression.

The material of the reinforcement plays a rather insignificant role in harshness. This is especially true as the configuration of the bellows leaves little freedom to choose materials because of the strength demands placed on the component. However, a strong fabric made of strong yarn

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

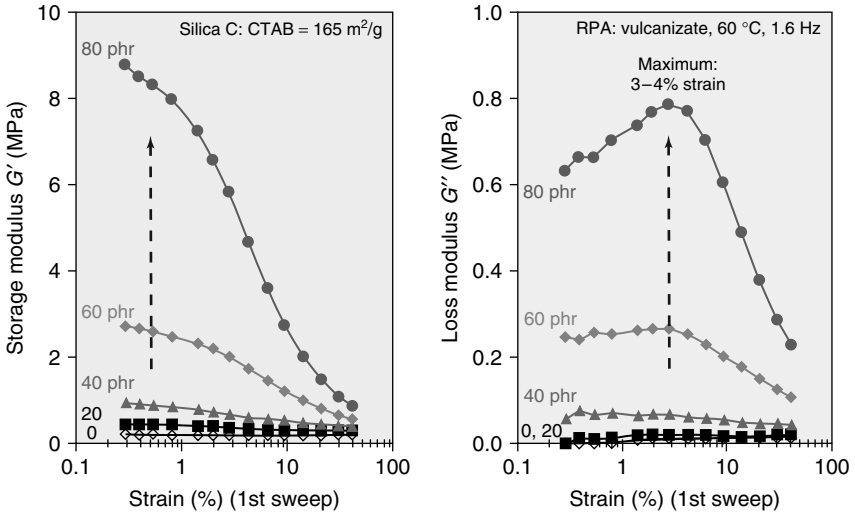


Figure 17. Memory- and dissipation-module curves. (Reproduced by permission of Joachim Fröhlich.)

and a high yarn density guarantee a stable rolling lobe with regard to less harshness. By contrast, however, the bellows wall may lack flexibility. Material properties may also have an influence on acoustical transmission.

2.7.2.3 Influence of the design of the bellows (wall thickness, yarn angle, yarn density, and thread count). The resistance to deformation of the bellows matrix not only depends on the materials used in the elastomer and the reinforcement but depends also on the design of the bellows. The main components here are the thickness of the elastomer layers, the diameter of the yarn, the yarn angle, and the thread count and/or the separation between strands. The combination of these parameters gives rise to a linked system that must undergo deformation. In the case of cross-layered bellows, a rhomboid pattern arises that determines the geometry and the deformation of the elastomer within the rhomboid. Figure 18 illustrates the stretch of the material within the rubber matrix as the rolling lobe passes through from the piston to the outer diameter. The example shows a bellows submodel of an unguided spring. The amount of material strain varies as a function of the bellows design.

To reduce harshness, material deformation must be kept low and the geometry of the rolling lobe must be kept stable. An advantageous design in this regard exhibits thin bellows with thin reinforcement material, little material, and slight deformations in the rhomboid pattern, thanks to optimized packing of yarn in conjunction with an optimized yarn angle. Such a high strength networked system, however, on the other hand, may slightly elevate the rigidity of the bellows wall during the rolling-lobe cycle.

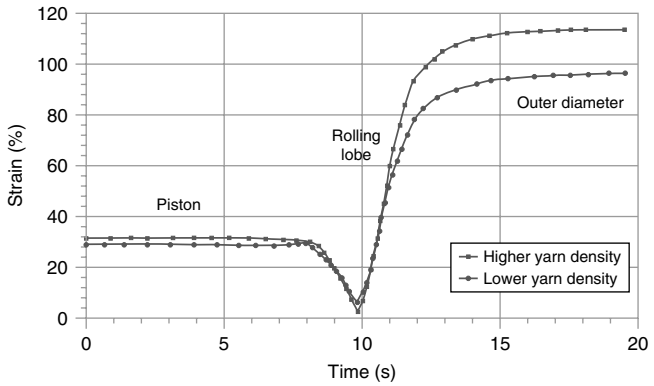


Figure 18. Depiction of the principal tensile strain in the rolling-lobe's cycle as a function of the bellows design.

Reducing the intermediate and outer rubber layer improves harshness markedly with regard to wall thickness. While yarn angle and amount in unguided springs depend essentially on the space available and requirements of engineering strength, the yarn parameters of an externally guided air spring can be put into effect for purposes of comfort under certain conditions. There is thus an ideal yarn angle for every bellows design.

Axial bellows do not exhibit any deformation with the rhomboid pattern because of the single layer of purely vertical yarn. This is the reason why axial bellows enjoyed the reputation until recently as being good for comfort. The entire arrangement, however, also provides less stability against deformation when the rolling lobe is in motion than the cross-layered arrangement. Moreover, thicker yarn must be combined with low rolling-lobe widths in an axial

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

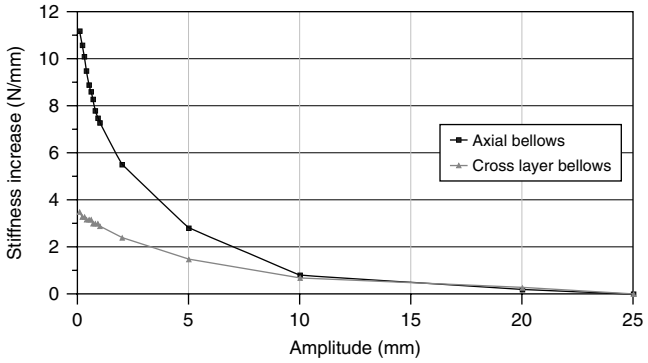


Figure 19. Harshness comparison between cross-layered bellows and axial bellows.

bellows for reasons of engineering strength. In addition, the operating pressure is capable of compressing the rubber material more strongly between the strands of yarn and deforming it even further. The rubber in the rolling lobe consequently experiences very severe deformation during movement. Bellows design is thus subject to severe limitations when it comes to the comfort of axial bellows.

By contrast, thin, guided cross-layered bellows offer the greatest possible leeway in shaping the parameters for improving harshness. Current developmental work has indicated that it is possible to greatly reduce harshness by employing very thin PA fibers and thin elastomer layers of CR in conjunction with the optimized air spring application. Figure 19 illustrates the great potential of a comfortable cross-layered bellows compared to an axial bellows used in the same air suspension of a luxury car. The excessive stiffness of the cross-layered bellows was reduced to approximately 30% of the level of the axial bellows. Its harshness of less than 4 N/mm has been reduced to a minimum.

2.7.3 Influence of the air spring design due to specific vehicle properties

The actual configuration of the air suspension exerts a major influence on harshness in conjunction with the properties of the bellows' components. Worthy of mention here are parameters such as air spring diameter with the accompanying pressure level, plus the opportunity to integrate exterior guidance, all the while considering such vehicle requirements as space for installation, load, and the intended spring rates.

2.7.3.1 Influence of air suspension diameter (pressure level and width of rolling lobe). Exploitation of the maximum air spring diameter is reflected in the reduction of the static pressure level. The internal pressure in the air spring stretches the bellows membrane and consequently puts an end to material expansion and the associated deformation (Voss, 2002). Reducing the pressure thus directly mitigates material deformation. The ideal amount for a minimum level of pressure, which would just guarantee the optimum forming of the rolling lobe in every situation, is generally not achievable in passenger car applications because of space limitations. In reality, the basic pressure $p_{\text{stat.rel.}}$ of approximately 8 bar is usual.

If there is sufficient space and clearance for the air springs, there is also a way to optimize the width of the rolling lobe r_b. Rolling lobes that are too small can quickly result in an increase of harshness because of a high rate of material deformation in the bellows wall. Moreover, they have a negative impact on engineering strength. By contrast, rolling lobes that are too wide necessitate large amplitudes for their initial pass-through, associated with a greater change of the yarn angle. The effective diameter (as a function of the rolling-lobe width) changes in parallel with the resultant operating pressure. The effect on harshness is, however, rather negligible.

Therefore, there will always be an ideal rolling lobe for each type of air suspension. Steps for reducing pressure and designing the ideal rolling-lobe width can make a major contribution toward reducing harshness. Figure 20 illustrates the influence of the rolling-lobe width using the example of a guided air spring.

2.7.3.2 The influence of exterior guidance. The expansion of the bellows reduces greatly when exterior guides are used. The rigid outside guide now absorbs the circumferential forces. This permits setting the yarn angle of the bellows generally between 60° and 90°, which then approximates the actual equilibrium angle during operation.

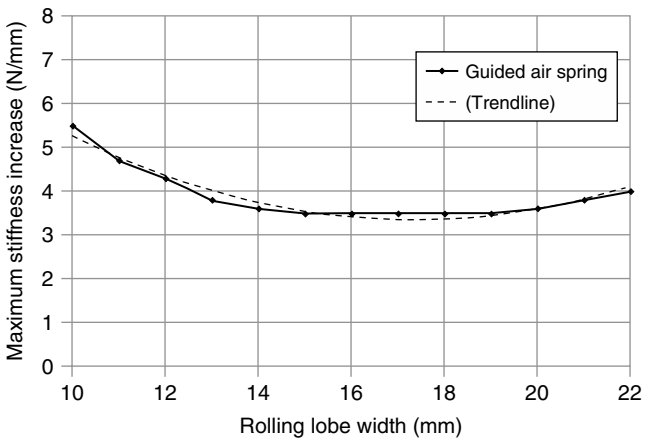


Figure 20. Influence of the rolling-lobe width on harshness.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Reinforcement material and elastomers are thus subject to much less load and deformation. This, in turn, allows for very thin-walled bellows. Above and beyond that, this protects the bellows from environmental influences. This makes it possible to keep the thickness of the outer cap to a minimum. Embedding thinner threads of yarn also contributes to a thinner bellows.

The rolling-lobe geometry remains stable because of the rigid outside guides and reduced deformation.

For purposes of comfort, air suspensions on passenger cars cannot do without exterior guidance. Most externally guided air springs are designed in the form of struts. It is somewhat more problematic to incorporate rigid guides in the case of freestanding air springs.

In contrast to the influences described earlier, the parameters such as air spring volumes and piston-roll geometry have only a negligible effect on harshness. Both generally affect only the basic spring rate.

2.7.3.3 The influence of air spring volume. The change in the air spring's nominal basic stiffness does not result in any increase in stiffness at low amplitudes. Studies of air springs with adjustable additional volume have confirmed this. While the basic spring rate increases as volume reduces, the resulting harshness only changes insignificantly (Figure 21).

On the other hand, the hysteresis forces of the air springs change as a function of the spring volume.

2.7.3.4 The influence of piston contour. The contour of the piston is an important parameter that influences spring characteristics, making it possible to adapt spring characteristics to the requirements of the car. When optimum rolling-lobe widths are considered and the angle of the piston is kept within certain limits, the effect on harshness is negligible. Of course, the piston contour exerts a strong influence on the geometry of the rolling lobe and its deformation, so that it is impossible to exclude an influence, especially for unguided air springs and very large changes in angle such as rises $>20^{\circ}$. As long as the rolling-lobe geometry can still "follow" the piston contour (i.e., the effective diameter is proportional to the piston diameter), its influence will be negligible.

Significant interactions among the influencing factors, depicted in Figure 14, will always occur. Each individual factor can change the harshness of air springs but it is necessary to find the optimum balance of all parameters for an individual application. Air spring manufacturers have specific configuration rules depending on the type of spring and bellows design employed. As a practical matter, it would be difficult to quantify values precisely with the aid of simulators and calculations because of the complex behavior of the material in the bellows structure. Values obtained empirically generally serve the development process. Table 1 summarizes the influencing factors and their importance.

It will be necessary to transfer all the influential factors shown here to the design and configuration of any air suspension system during the development process. The vehicle's axle design will determine the type of air spring. Consequently, certain limits already apply when attempting

Table 1. Factors influencing harshness.

Variable			Influence
Bellow design	Rubber	Material	Strong
		Thicknesss	Strong
	Yarn	Material	Neutral/medium
		Thicknesss	Medium
	Matrix layout		Strong
Air spring design	Guiding tube		Strong
	Pressure		Medium
	Rolling lobe width		Strong
	Volume		Neutral
	Piston contour		Neutral

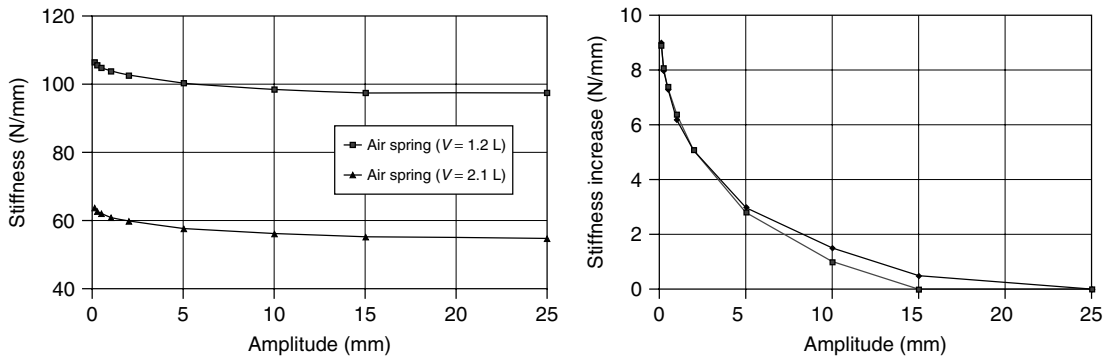


Figure 21. Harshness curves as a function of air spring volume.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to improve harshness due to the type of air suspension chosen.

3 HARSHNESS OF DIFFERENT TYPES OF AIR SUSPENSION

3.1 An overview

Air springs may be configured either as a strut or as a freestanding spring, depending on the type of axle employed. Struts see use in luxury cars predominantly on both front and rear axles while freestanding springs are found mostly on rear axles in middle-class and economy cars. Nevertheless, struts offer the greatest potential for diminishing the effects of harshness (Figure 22). The following spring–rate curves Δc_{harsh} are attainable in air suspensions being mass-produced currently:

- air strut, externally guided: $\Delta c_{\text{harsh}} = 3–6\,\text{N/mm}$;
- air spring, freestanding, externally guided: $\Delta c_{\text{harsh}} = 6–10\,\text{N/mm}$;
- air strut or air spring, unguided: $\Delta c_{\text{harsh}} = 10–15\,\text{N/mm}$.

The differences make it quite apparent that the combination of an externally guided strut on the front axle with a freestanding, unguided air spring on the rear axle elicits a pronounced mismatch of the respective harshness. While the front axle probably provides good comfort, the rear axle may lag noticeably behind. Experience has shown that it is not possible to eliminate this discrepancy. The trend toward modular construction in the automobile industry is leading to uniformity in the axle concepts of middle-class and luxury cars. Consequently, there are efforts afoot to establish the freestanding spring for the rear axle also in the luxury segment.

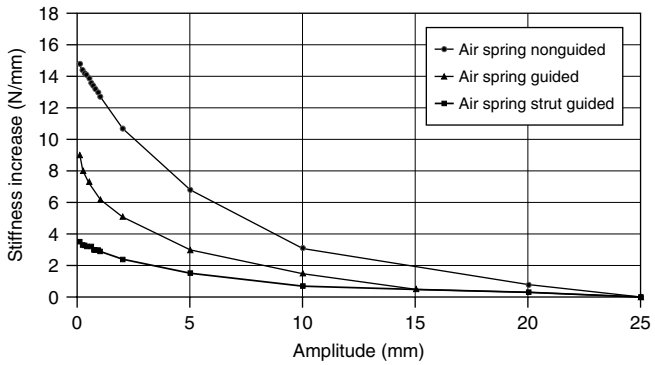


Figure 22. Harshness curves of different type of air suspensions.

3.2 Air struts

In strut architecture, the air spring is arranged concentrically around the damper, just as with a steel spring. The strut contains all the elasticity necessary to compensate for movements from the axle kinematics. Different configurations of air suspensions are possible, depending on the requirements of the vehicle and manufacturer. Figure 23 illustrates three possible configurations. Aside from simple, inexpensive unguided air springs, there are the more commonly used guided designs. The restricted motion of the bellows due to the rigid guide, however, necessitates integration of cardan compensation. Cardan means in this context that the air bellow needs to have certain flexibility into all directions. As will be shown soon, the strongly defined damper movement, defined by the suspension arm, makes the additional movement of the air spring itself necessary, as this is also connected to the top mount which is with the car body.

3.2.1 Externally guided struts

External guidance is indispensable if one is to diminish harshness appreciably. Assuming ideal parameters (see Section 2.7), it is possible to use cross-layered bellows with thin walls between 1.4 and 1.8 mm. Ultrathin PA yarn measuring 235 dtex × 1 serves as reinforcement. The yarn

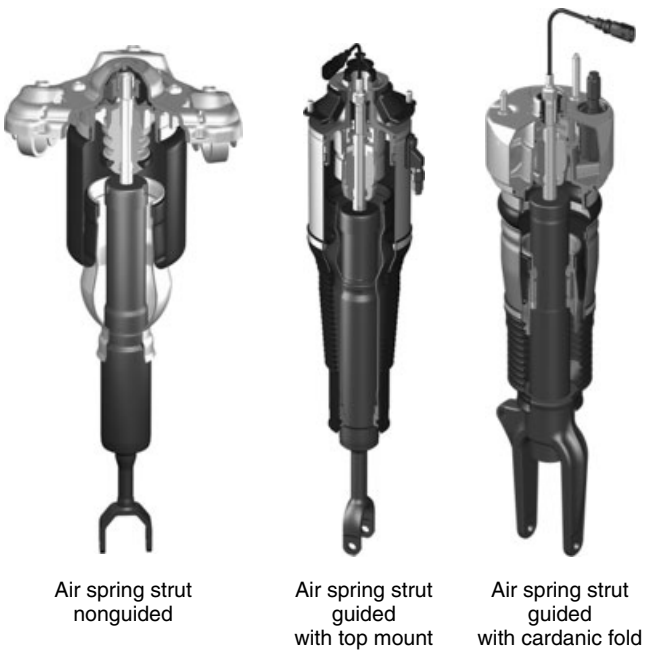


Figure 23. Selected strut configurations. (Reproduced by permission of Continental AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

has a diameter of just 0.19 mm. This makes it possible to achieve a minimal harshness Δc_{harsh} of up to 3 N/mm, which is hardly noticeable. The potential here has surely been largely exhausted. Any future improvements will require much effort in manufacturing the bellows when processing the fine yarn and the thin rubber layers. A major focus, moreover, will have to be on improving the material properties of the elastomer.

3.2.1.1 Cardan compensation. It appears much more important to focus on parallel effects that diminish comfort. The architecture of the air suspension can influence these two effects. Aside from the actual harshness of the air springs, the contact friction of the damper-piston rod plays a decisive part. Owing to the cardan motion of the strut in the axle, bending (flexural) moments are transmitted to the damper. They generate lateral forces on the piston rod, which increases friction as a direct consequence. As the strut compresses, it has to overcome a breakaway effect. This breakaway effect diminishes comfort appreciably (particularly at low amplitudes) and masks the harshness of the air springs noticeably (Ilias and Sorge, 2001).

The design of the struts determines the properties for taking up cardan movement within the axle's kinematics. The elasticity in the strut determines the quality of the cardan compensation and the transmission of bending (flexural) moments. Besides the elasticity of the air springs, this also includes the elasticity of the bearings used. A suitable air suspension design is thus capable of supporting optimum cardan compensation. Figure 24 illustrates that an additional elastomer bearing or a cardan fold could find application here. Elastomer bearings in the form of rubber and metal components can also have a decoupling effect because of their transmission behavior, and this also adds to the impression of higher comfort. Simply put, the cardan fold is an exposed, unguided bellows section that is reinforced and possesses a great deal of elasticity. It represents a simple, inexpensive solution.

3.2.2 Unguided air struts

Struts with unguided air springs play a somewhat subordinate role when it comes to comfort. The harshness effect is more marked because of their more robust structure and exceeds that of a guided spring many times over. The higher demands placed on the bellows make a thicker yarn necessary (at least PA 940 dtex × 1, yarn diameter approximately 0.33 mm) and a thicker outer cap. The total bellows wall thickness then lies in the range of around 2.0–2.4 mm. An optimum configuration of the remaining influential parameters can confine the spring rate to an acceptable range here. On the other hand, the good cardan compensation, mostly in the form of bellows elasticity, is advantageous.

3.2.2.1 Wheel-locating air struts. There is one promising way to improve comfort for unguided air springs with suspension struts such as the MacPherson strut. By locating the wheel, they transfer higher bending (flexural) movements to the strut, resulting in a greater lateral force with accompanying friction of the piston rod of the damper. An optimum air suspension design and configuration can nearly compensate for the transverse forces that occur. The reduced lateral forces consequently reduce friction of the piston rod and also provide a way

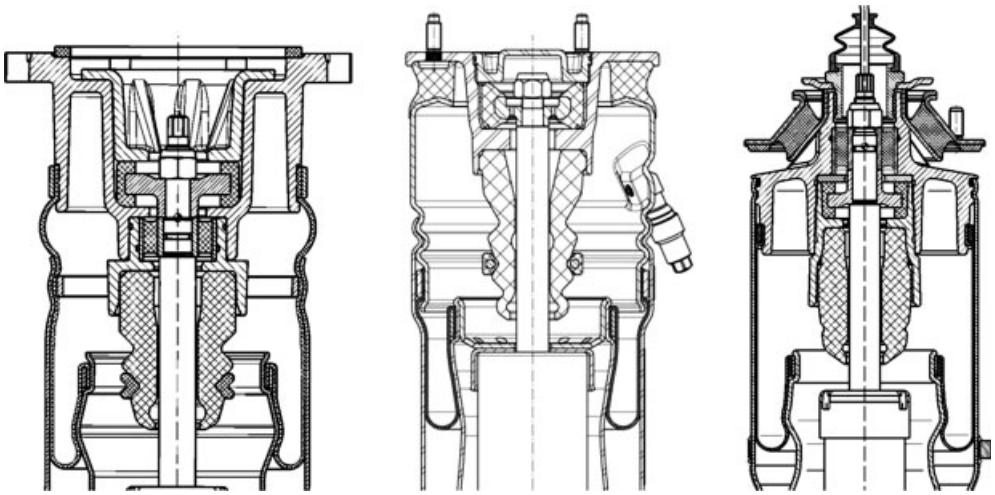


Figure 24. Cardan compensation of guided air struts. (Reproduced by permission of Continental AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to improve the design of the seals and the guides for the piston rod. However, these air suspension concepts are complicated and costly. What is more, the asymmetric arrangement of the bellows places greater demands on engineering strength.

3.3 Freestanding air springs

Rear-axle designs for automobiles call mostly for a spatial separation between dampers and springs. The freestanding springs are located more inboard. Their lower design height allows for a wider loading space in the cargo area. However, the springs' inboard position does result in limited space for installation. It also results in a relatively offset position for the air springs' pistons and cap because of spring motion. The rolling lobe may become more constrained as a consequence. Furthermore, the high spring ratio (up to $i = 0.5$) increases the load-carrying capacity and, consequently, the static pressure level. The high potential for additional load on the rear axle also adds to operating pressure. There is, therefore, a need for air suspension architecture that is capable of accepting cardan movements very well on the one hand and can offer the necessary space for spring travel with very little clearance on the other. The parameters mentioned earlier are counterproductive if one wants to optimize comfort. Using rigid, external guidance is problematic because of cardan motion and the permissible space. Parallel to this, the higher level of pressure requires robust bellows. Various types of freestanding air springs may offer a solution as illustrated in Figure 25. In addition, it is possible to make the lobe rolling upward. A guided version can also be seen in Figure 25.

The differences among the shown versions and the conventional model are the upside-down arrangement, the use of a double rolling lobe, and a flying external guide.

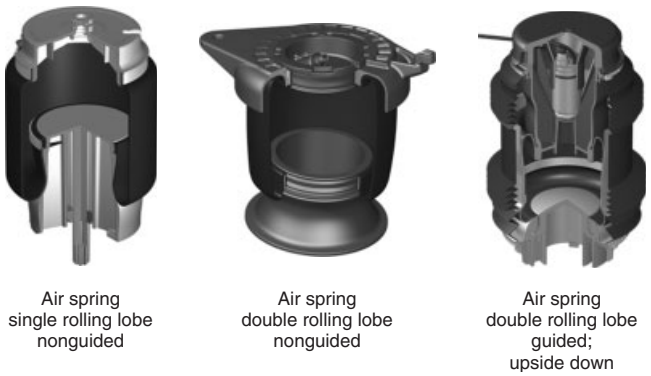


Figure 25. Selected versions of freestanding air springs and a guided one with a double rolling lobe. (Reproduced by permission of Continental AG.)

3.3.1 Externally guided, freestanding air springs

Compared to the front axle, acceptable values for harshness at the rear axle are achievable only with a guided design. To ensure cardan compensation and guarantee freedom of movement, experience has shown that it is necessary to have two rolling lobes and an upside-down arrangement. Inclusion of external guides requires a very compact flying design. This means that the external guide covers only a portion of the bellows and leaves the second rolling lobe largely exposed. All in all, the bellows wall can be made thinner although it must still be robust in the vicinity of the second rolling lobe. An extremely thin wall such as the strut is currently not possible for this reason. As a practical matter, though, it has already been possible to achieve a spring rate that is less than 10 N/mm in a luxury sports car.

There is potential for still more development. The second rolling lobe must be geometrically configured so that the stress sinks, on the one hand, and the harshness effect diminishes, on the other. In view of the above, it would be conceivable to achieve maximum stiffness rates of approximately 5 N/mm and thus to approximate the comfort level of the front axle.

3.3.2 Unguided, freestanding air springs

The use of freestanding unguided air springs in middle-class and light trucks is widespread due in part to cost and space limitations. Comfort is not the top priority in this segment, especially not in the case of rear-axle systems with load leveling. Combined with guided front air springs, the increased harshness occurs noticeably which means that it is necessary to pay attention to this area. As there is only limited freedom in configuring the bellows, any design must be functional and geometric.

An upside-down arrangement in conjunction with a second rolling lobe improves the kinematic movement of the springs and reduces the design height needed. It is then possible to design the geometry of the rolling lobe better and reduce the constriction of the rolling lobe.

What is more, a concept with a second rolling lobe permits placing it at the height of the effective diameter so that the spring rate of the rolling lobe activates in line and consequently sinks. This principle does entail limited freedom of configuration though. Optionally, an additional antiharshness layer in the form of an elastomer or polyurethane (PUR) layer can greatly improve transmission properties, particularly when it comes to acoustics. A simple elastomer layer is generally all that is necessary where the spring joins the bodywork.

The acceptable comfort range for unguided freestanding air springs is approximately 10–15 N/mm of spring rate.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

With proper measures, it should be possible to drop the rate to under 10 N/mm in the future.

These harshness numbers do not present a real alternative when it comes to comfort, however. Guided air springs are still indispensable here, ideally as a strut. Added to this is optimum cardan compensation to minimize any additional friction effects.

4 OUTLOOK

Air suspension's market penetration will continue. Air springs bestow a noticeable improvement in comfort, primarily in luxury and middle-class passenger cars. Small cars (which are coming to the market next time) need air springs most, but the cost is (up to now) the largest obstacle. The focus for future work is on improving the characteristics of air suspensions. The harshness effect is a fundamentally undesirable property of air spring bellows and is also an important, quantifiable gage of comfort for automotive components. Objective test results, however, must always be brought into line with subjective assessment.

In view of the various factors influencing harshness, it is possible to demonstrate extremely slight increases in spring rate at low amplitudes using targeted air suspension designs. Externally guided air springs with thin-walled cross-layered bellows and fine yarns of PA are capable of reducing the bellows' resistance to rolling to a minimum. These bellows find use predominantly in strut designs. More development work will be necessary to use thin bellows in freestanding spring designs in order to attain a comparable level of comfort. In addition to empirical experience with comfort-based air suspension, simulation processes will increasingly be used to predict harshness properties.

Beyond that, it will be necessary to analyze and reduce other factors that diminish automotive comfort. Besides the air springs themselves, it is worth mentioning the properties of elastomers where the suspension joins to the body and the response of hydraulic dampers as factors that can generate undesirable rubbing effects. However, the structural stiffness of the bodywork also plays an important part in transmitting undesirable vibrations and noises to the car and to the passengers.

5 SUMMARY

Air spring systems are introduced into the market, but for passenger cars, these systems suffer from harshness. This contribution gives a detailed explanation on the effects of harshness and the reasons for the existence. In addition, there are descriptions for the measurement, the simulation, and the reduction of the harshness effects to modern air spring systems. Detailed information will enable the automotive chassis designer to reduce the harshness effect in the future.

REFERENCES

Boehm, J. (2001) *Der Payneeffekt: Interpretation und Anwendung in einem neuen Materialgesetz für Elastomere*, Dissertation, University of Regensburg, Regensburg.

Ersoy, M. and Heißing, B. (2008) *Fahrwerkhandbuch*, Vieweg Verlag, Wiesbaden.

Froehlich, J. (2007) *Verstärkung durch Moderne Füllstoffe—Teil II*, Degussa GmbH, DIK, Hannover.

Ilias, H. and Sorge, K. (2001) *Grundsatzuntersuchung zum Thema Harshness, Interner Entwicklungsbericht*, Continental AG, Hannover.

Klüppel, M. (2007) *Hochfrequenzeigenschaften und Reibung von Elastomeren*, DIK, Hannover.

Thurow, G. (1995) *Lebensdauerergebnisse und Federverhalten von Geführten Luftfedern*, ContiTech Luftfedersysteme, Hannover.

Voss, H. (2002) *Die Luftfederung, eine Regelbare Federung für Straßen- und Schienenfahrzeuge*, ContiTech Luftfedersysteme GmbH, Hannover.

FURTHER READING

Betzler, J. and Reimpell, J. (2005) *Fahrwerktechnik Grundlagen*, Vogel Verlag, Würzburg.

Braess, H. and Seiffert, U. (2007) *Handbuch Kraftfahrzeugtechnik*, Vieweg Verlag, Wiesbaden.

Gauterin, F. and Sorge, K. (2001) *Noise, Vibration and Harshness of Air Spring Systems*, Continental AG, Hannover.

Puff, M. (2009) *Prüfspezifikation zur Charakterisierung von Luftfedern*, Technische Universität, Darmstadt.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto015
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Air Supply for Advanced Applications

Hans Otto Becher
WABCO, Hannover, Germany

1 Introduction 1
2 System Requirements for Air Supply Units 1
3 General Function of a Compressor for Air Suspension 4
4 Compressor Packaging/Acoustic Insulation 16
5 Summary 18
Reference 19
Further Reading 19

1 INTRODUCTION

The desired characteristic and performance of air suspensions in passenger cars can only be achieved in combination with an air supply unit (ASU). The ASU itself is necessary but undesired: it consumes space, increases weight, consumes energy, creates noise and vibration, and costs money. And exactly they are the key characteristics of an ASU, which are subject to continuous optimization.

However, the ASU is mandatory to generate compressed, dry, oil-free, and clean air to pressurize the air springs. The suspension elements have to be filled with a certain pressure for normal working operations. Natural leakage through bellows, pipes, connectors, and valves has to be considered to be compensated. In many cases, there are additional demands to increase the chassis level by inflating the air springs and decrease the chassis level by deflating the air springs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

The compressed air has to be distributed into the air springs, which is mostly done with multiple solenoid valve blocks. To lower the vehicle chassis height, the air springs have to be vented. The ASU provides solenoid valves for this function.

In addition to the above-mentioned criteria, key performance indicators for ASUs are airflow, maximum pressure, duty cycle, electrical current consumption, and peak current.

2 SYSTEM REQUIREMENTS FOR AIR SUPPLY UNITS

2.1 Two-corner applications for rear axles

There are two types of two-corner air suspension systems on the market:

- Leveling systems (combination of steel springs and air springs)
- Full air suspension systems (without steel springs).

Leveling systems are designed to carry the basic vehicle rear axle load (empty condition) with a steel spring and the additional load of passengers and cargo with an air spring. There is just one nominal level requested. Quick lifting/lowering is not needed.

Consequently, the airflow performance requirement is low, typically about 10 norm dm^3/min [where norm dm^3 is the amount of air which fills one dm^3 at 0°C and 1013 mbar at 6 bar. The pressure range however is between 1 (minimum pressure to avoid bellow damage) and 12 bar in the fully loaded condition. The pressure demand is quite high as the diameter of the air spring is limited because of

packaging reasons. These systems are popular in the North American market, but the market penetration is decreasing. The system layout is simple (only one common height sensor for right and left side) and the ASU design is very much cost driven.

Full air suspension systems for rear axles, carrying the whole rear axle load (empty vehicle load + passengers + cargo), are more demanding regarding the ASU (Figure 1). While the pressure levels are similar, the airflow requirement is higher (>20 norm dm^3/min at 6 bar). The end-user's requirement is a quick, visible compensation of chassis level drop after loading. Additional functionality is for instance a switch in the trunk area for manual lowering of the chassis for easy loading operation. This lowering procedure has to be quick as well to avoid waiting time. Consequently, the air passage through all system elements has to have a sufficient diameter. Usually, these ASUs have a double solenoid valve block to control left- and right-side air spring individually. This feature allows a right/left-balanced leveling even if the axle load is not centered. In addition, the roll stiffness is increased, as the airflow between right and left air bellows is blocked. Another feature is the load transfer away from one wheel in case of a damaged "run flat" tire. Unloading the damaged tire can increase the allowed driving distance.

Duty cycle demand for ASU's in two-corner systems is low (typically 10% on-time of a 10-min period, environmental temperature 23°C), as there is no normal operation mode, which requires a long continuous compressor run. The longest sequence is given by filling the empty air spring from buffer level to normal ride height and loaded condition. Typically, this operation does not take longer than 60 s.

In some cases [transporters and sports utility vehicle (SUV)], these systems have a reservoir for quick lifting and/or compressor running noise avoidance in certain conditions. Then, the ASU performance in terms of airflow and maximum pressure has to be even higher and on a level of typical four-corner reservoir systems (Section 2.2).

2.2 Four-corner applications

Most of the four-corner systems have a reservoir to enable quick lifting by a high airflow from the reservoir into the bellows. Performance demand for ASU's used in reservoir systems is typically >25 norm dm^3/min at 6 bar, maximum pressure up to 18 bars. Duty cycle requirement is higher (15%) than for ASU's in reservoirless systems, as the compressor run time to fill up a reservoir up to 18 bar is significantly longer than to inflate the bellows directly.

Another advantage of reservoir use is that the compressor must not be operated in the vehicle standstill condition, which is very noise critical (start–stop system, engine off). In closed systems, a reservoir is mandatory, as the air is moved between air spring and reservoir. Usually, a direct bellow filling by the compressor or exhausting of the bellows is not possible.

Some four-corner systems are working without a reservoir. These systems are designed to operate without attracting attention from the end user. Chassis-level changes are expected to be "invisible," thus the demand for the ASU airflow is low. Maximum pressure is similar to two-corner systems (12 bar). The compressor has to be operated also in standstill mode of the vehicle, and so the demand regarding very low compressor running noise is high.

Many four-corner systems are using not only the benefits of air springs in normal operation mode but also the possibility to lift or lower the chassis. Reasons for this include dynamic driving behavior and reduced fuel consumption

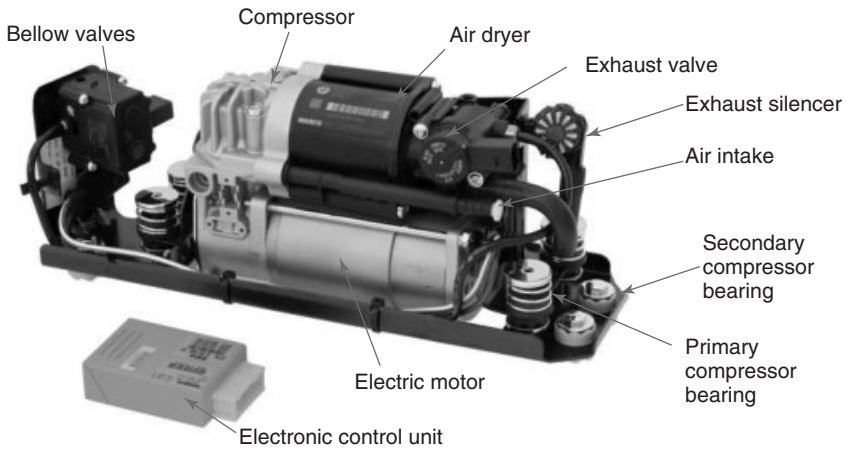


Figure 1. Air supply unit for rear axle air suspension system. (Reproduced with permission from WABCO.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

by low level at high speed, easy entry for passengers, and extended ground clearance for certain driving situations (e.g., off-road). These kinds of systems are demanding higher performance from the ASU: high airflow during lifting and lowering operations, high maximum pressure, and high duty cycle for frequent repeats of level changes. To monitor the duty cycle of a compressor and getting close to the thermal limits, the information about compressor temperature or at least environmental temperature in the area of the compressor is mandatory. Some systems have additional volumes that can be connected/disconnected to the air bellows to vary the spring stiffness. For the ASU, this means more air volume to be transported into and off the bellows. Figure 2 shows a typical example for a four-corner ASU.

2.3 Further demands for compressed air

Having an ASU installed in a passenger car increases the desire to use this energy also for other systems. As an example, a brake booster with compressed air could be packaged much more compactly than a typical vacuum brake booster. However, this concept was never realized. The only option, which was marketed so far, was a tire inflator. The compressor can be used to fill a flat tire or leisure equipment. For some special applications, there are systems on the market, which are able to control the tire pressure during driving. The main reason not using compressed air for other applications than air springs is that many features as the air suspension itself are just optional. One example is a dynamic seat contour, which uses inflatable cushions. An ASU for air suspension systems would be oversized for the seat operation alone. Therefore, these applications have own small compressors, directly installed in the seat.

2.4 General system requirements: future outlook

The load-bearing force of an air spring is as follows:

$$F = p_i \cdot A$$

The demand for smaller air bellow diameters and thus effective cross-sectional area of an air spring A because of packaging reasons in passenger cars leads to increasing bellow pressures p_i. Thus, the main future challenge for ASUs is a higher maximum pressure.

Furthermore, the vehicle manufacturers will specify size, weight, and inrush current limitation more strictly. Noise and vibration levels are also subject to improvement, because hybrid or electric drives in passenger cars allow a very low basic noise level, and the ASU must not stick out during those driving conditions.

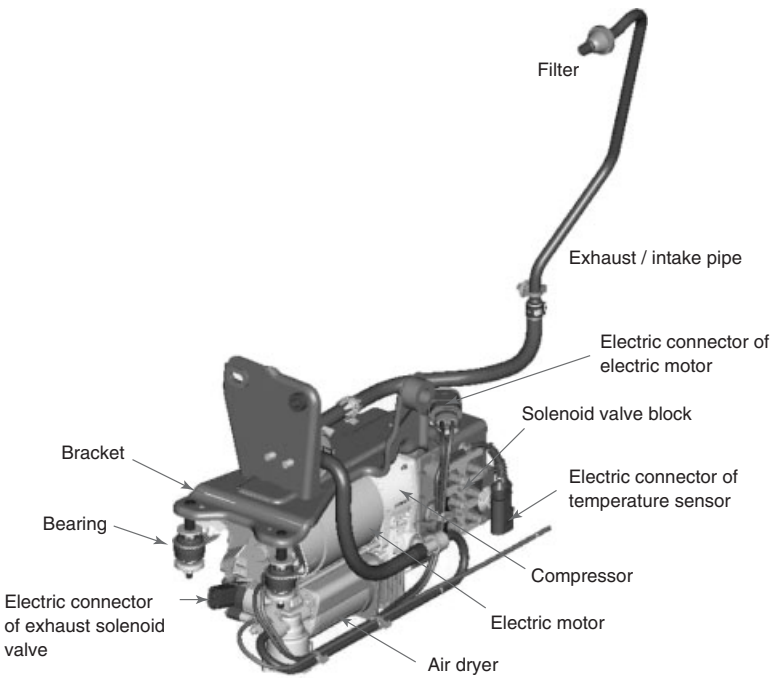


Figure 2. Four-corner air supply unit. (Reproduced with permission from WABCO.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3 GENERAL FUNCTION OF A COMPRESSOR FOR AIR SUSPENSION

3.1 Compressor

Following the system requirements in Section 2, the preferred solution for an ASU is a dry-running piston compressor driven by an electric motor. Other devices such as vane-type pumps or rotary screw compressors are not able to generate the requested pressure levels (10–18 bar) or are commercially not attractive.

3.1.1 One-stage open type

Open type means the compressor is sucking air directly from the atmosphere, compressing it to the requested level and feeding the air springs and/or reservoir. Exhausting the air from the bellows into the atmosphere lowers the chassis.

A typical 4-corner open system schematics is shown in Figure 3. By far, the most popular and common design is a one-stage piston compressor, open type (Figure 4). The air is taken in from the atmosphere through the air intake into the crankcase. If the piston is moving down,

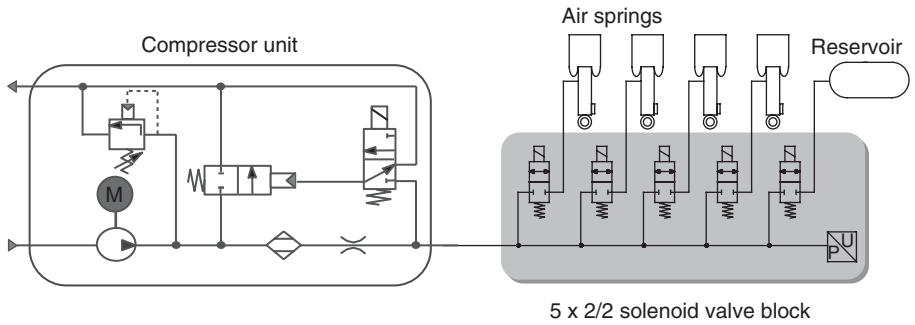


Figure 3. System schematics of a four-corner air suspension system (open system). (Reproduced with permission from WABCO Hitachi.)

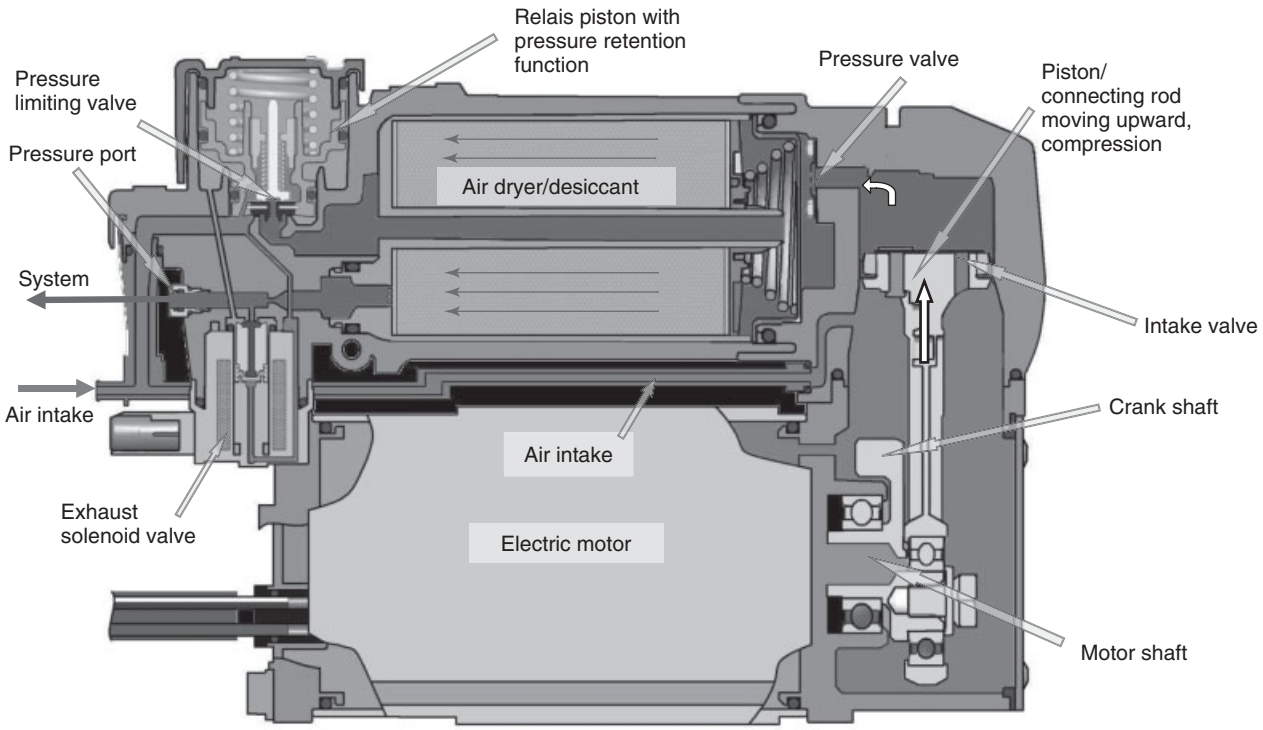


Figure 4. Sectional drawing of a one-stage open type compressor. (Reproduced with permission from WABCO.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

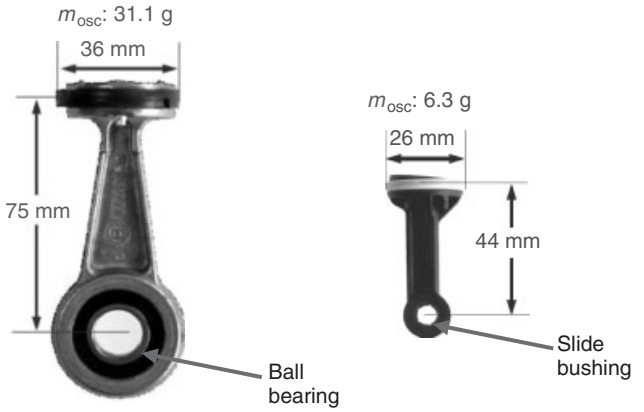


Figure 5. Wobbling pistons with ball bearing and slide bushing. (Reproduced with permission from WABCO.)

the air flows through the intake valve in the piston into the cylinder. If the piston is moving up, the air is compressed and guided through the pressure valve into the air dryer. The compressed air passes the desiccant and is then guided to the pressure port and entering the system.

All known one-stage piston compressors for air suspensions are directly driven by an electric motor. The crankshaft is fixed on the motor shaft and drives the connecting rod. The connecting rod bearing is—depending on the operation conditions and performance range—designed as a ball bearing or slide bushing (Figure 5).

In many cases, the piston and the connecting rod are designed in one piece without articulation between piston and connecting rod. This "wobbling piston" design requires special focus on the piston sealing elements, which have to tolerate an angular motion of the piston (Figure 6).

There are two possible solutions to piston sealing: piston ring or gasket (Figure 7). For both designs, the proved material is PTFE providing a sufficient heat and wear resistance without being lubricated. At the same time, the sealing has also to be able to work under low temperature down to -40°C.

The piston ring (Figure 8) is usually a machined part with a slot for assembly and for wear reasons. This design ensures a permanent defined axial contact pressure between the piston ring and the cylinder wall.

Although the piston sealing should be as tight as possible to achieve maximal volumetric efficiency, the piston ring has a natural leakage under static conditions. Thus, pressure balancing above and below the piston is given within the cylinder, which enables an easy compressor start. In comparison to the piston ring, a gasket is a more simple part, is fixed to the connection rod, and has always close contact to the cylinder. A piston ring has always a certain amount of clearance in axial direction (Figure 9) to allow

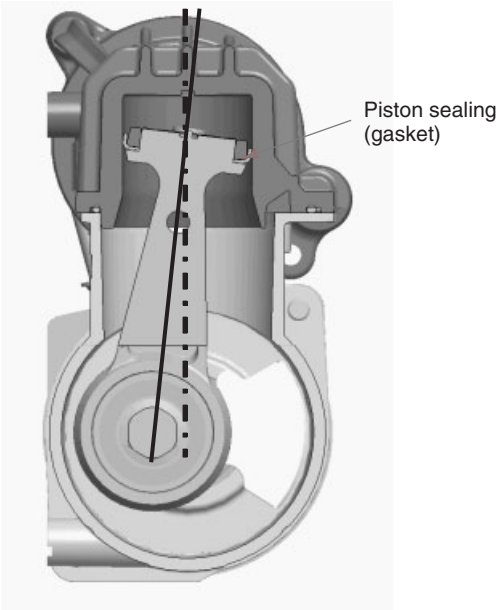


Figure 6. Wobbling piston, maximum angle position. (Reproduced with permission from WABCO.)

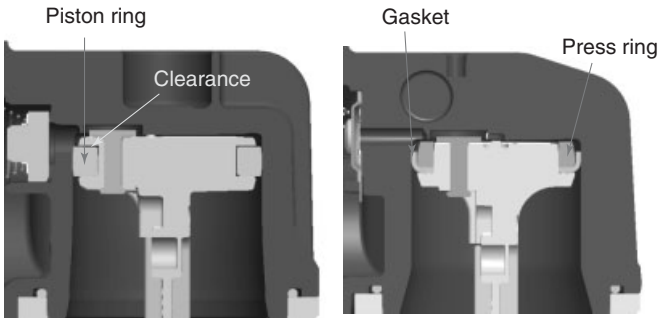


Figure 7. Two different types of piston sealing. (Reproduced with permission from WABCO.)



Figure 8. Piston ring. (Reproduced with permission from WABCO.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

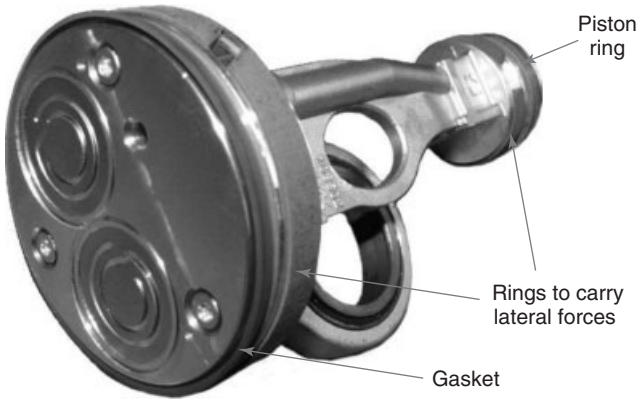


Figure 9. Piston of a two-stage compressor. compare Figure 15. (Reproduced with permission from AMK Automotive GmbH & Co. KG.)



Figure 10. Articulated compressor piston. (Reproduced with permission from WABCO.)

the ring moving. However, this clearance has a negative impact on acoustics.

Another focus lies on lateral piston forces. These have to be minimized to achieve low friction, low wear, and thus a long lifetime. In some cases, an additional ring to carry the lateral forces and unload the sealing (Figures 9 and 10) accompanies the piston sealing.

There are also compressors in the market, which have an articulation between connecting rod and piston (Figure 10). The additional articulation with another bearing and a certain clearance might have a negative impact on acoustics.

Usually, the piston material is aluminum. In the case of lower pressure ranges (and thus lower temperature), the piston can be designed from plastic (comparison in Figure 5).

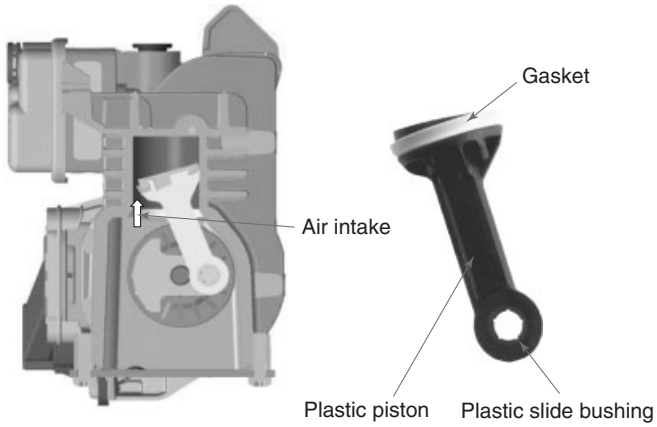


Figure 11. Low power compressor. (Reproduced with permission from WABCO.)

3.1.1.1 Air intake/suction valve. Air from the atmosphere has to be led into the cylinder. This can be done with an intake valve in the cylinder head or directly in the piston (Figure 4). The lower the stiffness of the intake valve, the higher is the volumetric efficiency. On the other hand, the design has to ensure low tension forces and sufficient strength of the part over lifetime.

In the compressor range of low pressures <10 bar, the intake valve can be omitted by opening a bypass in the lower position of the piston (Figure 11).

3.1.1.2 Pressure valve. The pressure valve is located in the hot area of the compressor, and consequently, it has to withstand high temperatures. In addition, it should have a low delta pressure (low stiffness) for efficiency and a low leakage level. Leakage would allow a flow back into the compression chamber, reducing efficiency and starting capability.

There is a trade-off between those demands: an elastomer pressure valve can achieve low leakage level, but high temperature could be an issue. Another design uses steel leaves as a check valve, which has good robustness against high temperature, but allows a certain amount of leakage. The steel leaf type has a very low mass and thus high dynamic capability. This gives advantages in terms of noise. In the market, both types of pressure valves are known (Figure 12).

3.1.1.3 Dead volume. To optimize the volumetric efficiency, the dead volume has to be minimized. This dead volume is also important for the stabilized pressure, which is achievable. Even at low environmental pressure (high geodetic heights) of about 0.6 bar, the compressor performance must not drop below the specified level. Test results are shown in Figure 13.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

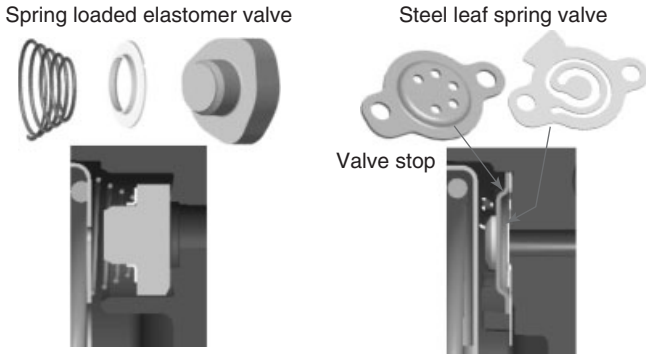


Figure 12. Two different types of pressure valve. (Reproduced with permission from WABCO.)

3.1.1.4 Compensation of oscillating and rotating masses. Figure 14 shows a typical configuration of a one-cylinder compressor. The target is a well-balanced share between rotating and oscillating masses. The force distribution in the *x*- and *y*-directions can be seen in the diagram. The overall compressor position in a car and the insulation between compressor and car chassis has to be considered as well for the layout of the balancing.

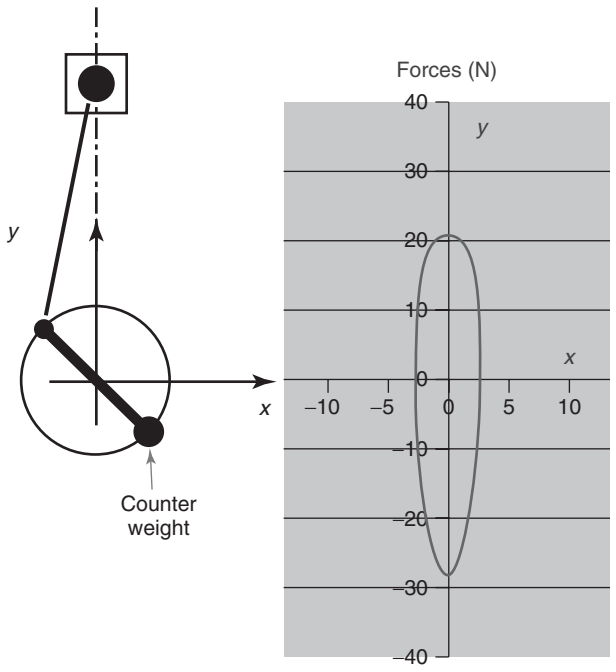


Figure 14. Trade-off between balancing of longitudinal and lateral forces. (Reproduced with permission from WABCO.)

3.1.2 Two-stage open system type

Two-stage compressors have an advantage in generating higher pressures; as a one-stage compressor is limited to maximum 18 bar, the two-stage compressor can provide much higher pressure and the efficiency in higher pressure ranges is better. Furthermore, the required

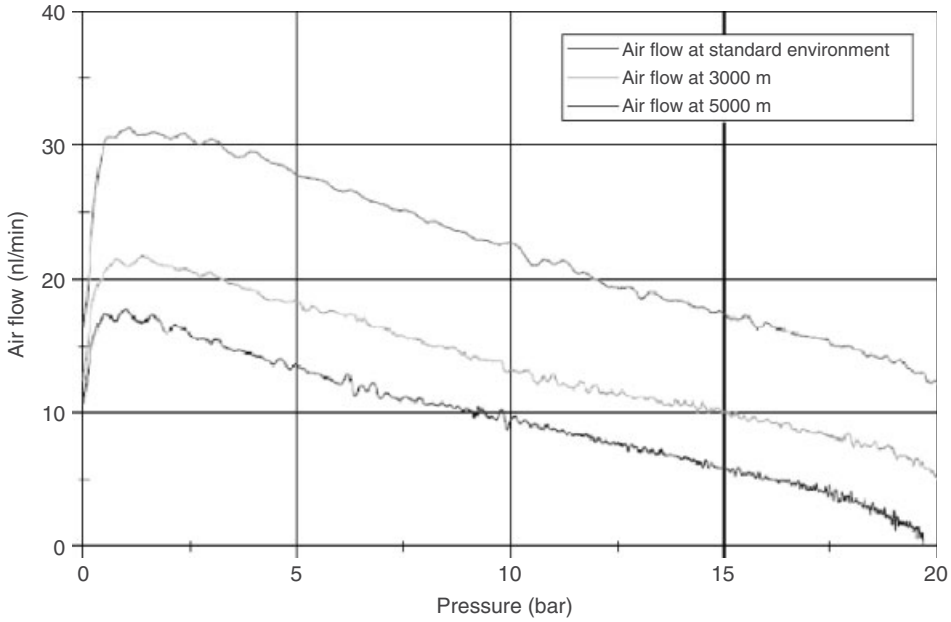


Figure 13. Compressor airflow at different geodetic heights simulated by different intake pressures. (Reproduced with permission from WABCO.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

torque to drive the compressor does not show such high peaks as with a one-stage compressor. The torque is better distributed during one cycle. On the other hand, there are more mechanical parts needed to provide this function, friction is typically higher (two-piston sealing instead of one), and, owing to that fact, the efficiency in lower pressure ranges is worse than for a one-stage compressor.

The principle of a two-stage compressor is shown in Figure 15. Air is taken into the crankcase and then compressed in stage one (at the bottom, piston moves downward). This compressed air is guided through a channel within the double piston into the second stage for compression to the final pressure.

Figure 16 shows the comparison of two medium power compressors. The efficiency advantage of a one-stage compressor in lower pressure ranges is obvious.

However, at higher pressure ranges, the two-stage compressor is showing benefits compared to a one-stage compressor. Figure 17 compares the efficiency and the flow rate of one- and two-stage high power compressors.

3.1.3 Closed system type

Currently, there is one closed system type compressor on the market (Figure 18). The system configuration is more complex than with an open system (Figure 3). The reason is the airflow control valve unit. In a four-corner system,

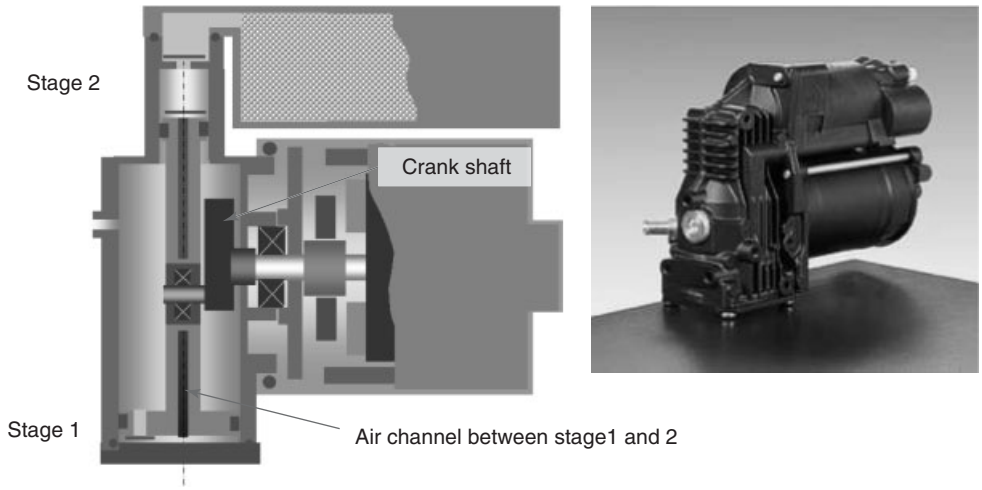


Figure 15. Two-stage compressor. (Reproduced with permission from AMK Automotive GmbH & Co. KG.)

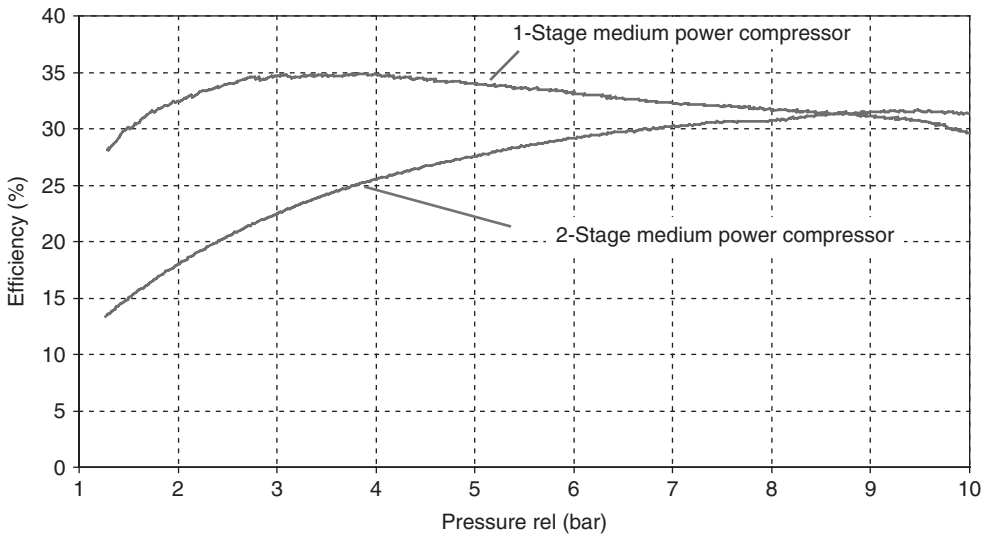


Figure 16. Comparison of a one- and a two-stage medium power compressor. (Reproduced with permission from WABCO.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

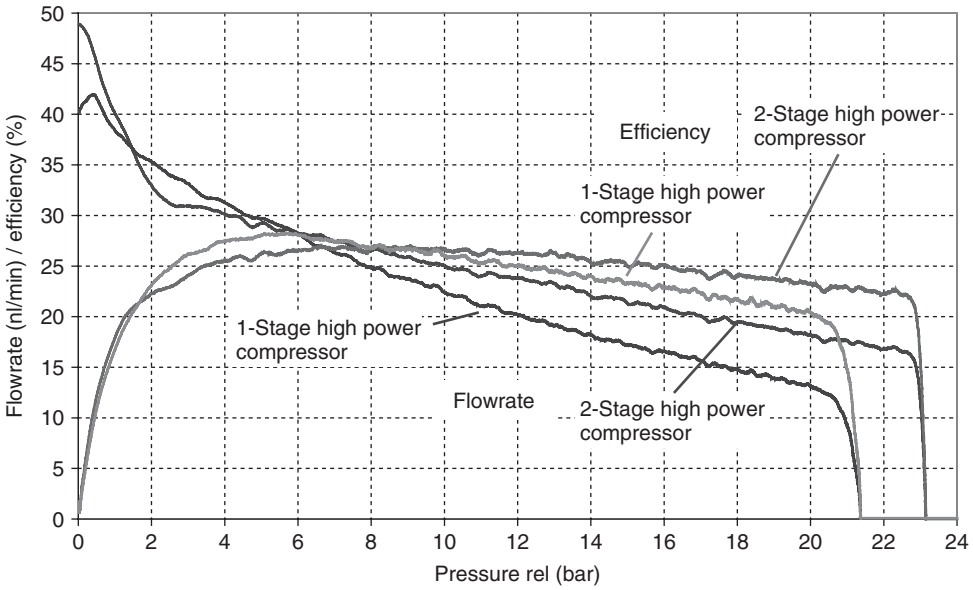


Figure 17. Comparison of a one- and a two-stage high power compressor. (Reproduced with permission from WABCO.)

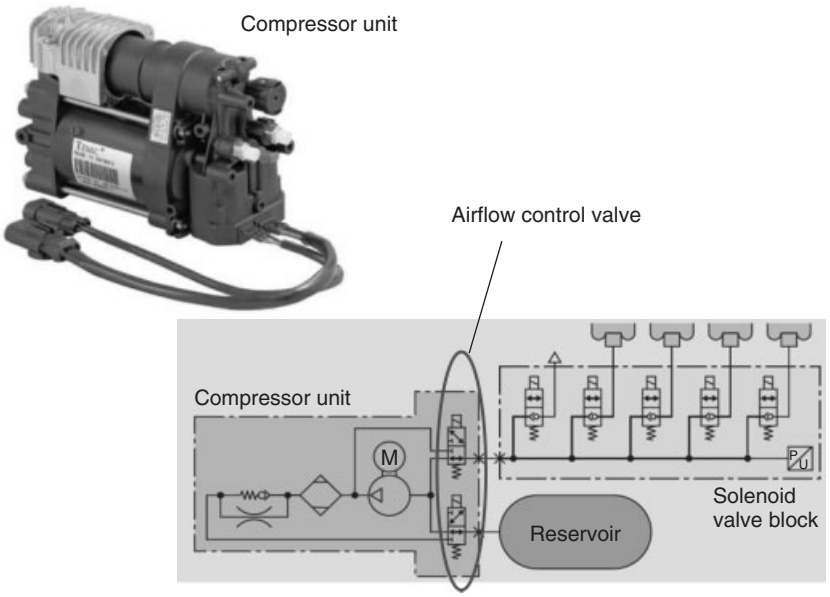


Figure 18. Closed system compressor and pneumatic layout. (Reproduced with permission from Continental AG)

a total of seven solenoid valves are necessary, whereas the open system only needs six solenoid valves.

Compared to an open system compressor, which gets always the atmospheric pressure at the air intake, the closed system type compressor can be precharged. Figure 19 compares the compression cycle of an open system type compressor and a closed system type compressor. The closed system type compressor is able to further compress a certain precharge pressure to the requested end pressure with a smaller stroke than an open type compressor. The area surrounded by the p-V line is equivalent to the work, which is needed to compress the air, or in other words, the requested energy for the compression cycle. In the shown example, the area (=work) is smaller in the closed system type compressor, and this is the reason for the energy-saving advantage of such systems.

The flow rate versus pressure of a closed system type compressor is shown in Figure 20. As a parameter, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

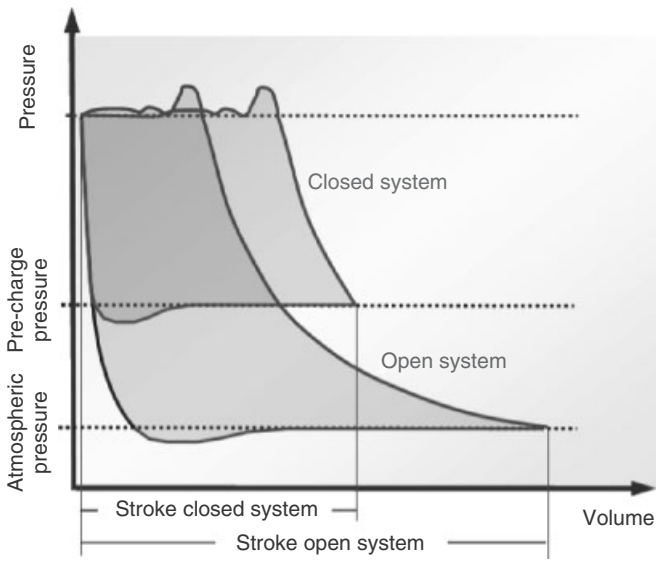


Figure 19. Compression cycle of an open system type compressor and a closed system type compressor. (Reproduced with permission from WABCO.)

precharge pressure is modified between 0 (atmospheric pressure) and 8 bar. This compressor has a smaller displacement than the open system type. Consequently, the airflow is very limited if it is operated in open system mode (e.g., system filling in workshop, see 0 bar line in Figure 20). The performance can be significantly increased if the compressor is charged with compressed air, either from the air springs or from the reservoir. The principle is to pump the air between the reservoir and the air springs or vice versa. The benefit is that the stored energy of the compressed air remains on a higher level and the compressor only has to increase the pressure by a low ratio (Figure 19). Lower energy consumption over vehicle lifetime is the positive consequence. The runtime of a closed system type compressor is roughly 30% of an open system type compressor in a four-corner application with reservoir, if just the pure lifting and lowering processes are considered. This will partly be compensated by necessary purge cycles to regenerate the air dryer: in a special operation mode, the dry air in the reservoir is used to dry the air dryer desiccant. This air loss has to be filled up from the atmosphere. Depending on environmental conditions, the share of compressor runtime for purging can be a significant amount on top of the normal compressor operation for lifting and lowering. Though, the overall compressor runtime in a closed system is about 50–60% compared to an open system.

3.2 Electric motor

Brushed direct current (DC) motors are used to drive compressors for ASU's in air suspension systems. Another option would be a compressor directly driven by the engine. This has been done in earlier times at Daimler Benz (1960s).

Advantages of electric motor for compressor drive:

- The compressor can be operated independently of engine type and vehicle.
- In the case of optional air suspension, the same vehicle engine can be used for both suspension variants.
- Compressor installation position can be anywhere in the vehicle.

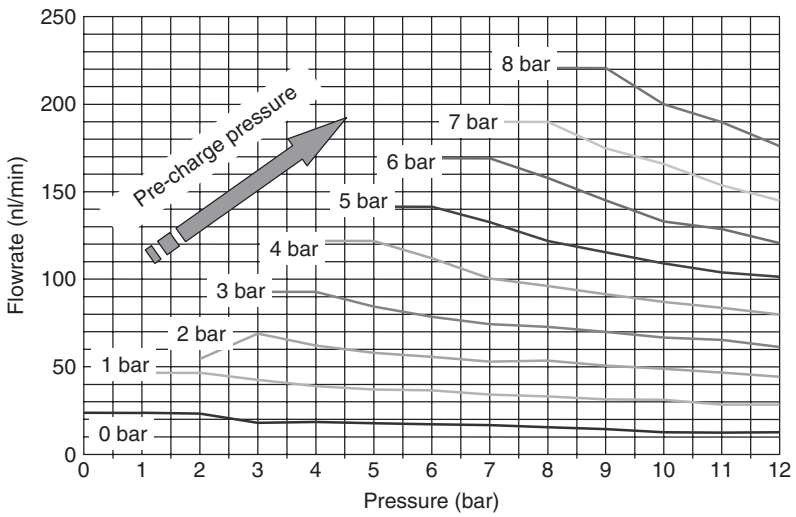


Figure 20. Flow rate versus pressure at various precharge pressures. (Reproduced with permission from WABCO.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- No issue with varying speed of the engine between idle and maximum speeds (ratio 10 : 1).
- No clutch would be required to couple/uncouple the compressor.
- In case of engine off (hybrid cars), the compressor can still be operated and thus lifting is possible even in "engine-off" condition.

Disadvantage of electric motor for compressor drive:

- Limited performance due to limited battery and generator capacity and electrical wiring dimension.
- Lower total efficiency due to double energy transformation.

In newer applications, for example hybrid cars, the inrush current of a relay-controlled DC motor is not accepted any more. To reduce the peak current and current gradient during the switch-on phase of the compressor, the relay can be replaced by a PWM (pulse-width modulation) control, which provides a soft start. The following picture (Figure 21) shows an example: the relay-controlled peak current of this compressor would be nearly 60 A. Thanks to a PWM control, the peak current can be reduced to 32 A.

In this context, there are also brushless motors considered for compressor drive. Motor life time is not the driver. Typically, the compressor run time is around 750 h in a vehicle life, which is easy to achieve with brushed motors. The main benefit of a brushless motor is the possibility for speed/current control.

The required torque at a certain speed mainly influences size and weight of the electric motor. To reduce the torque but keep the same performance, the speed has to be increased. However, the pneumatic efficiency of a compressor would decrease in case of higher speeds (valve dynamics, etc.) and thus such motors have to be transmitted to lower speeds at the compressor. Another factor, which influences weight and size, is the material of the magnets. Here is the classical trade-off between cost and weight.

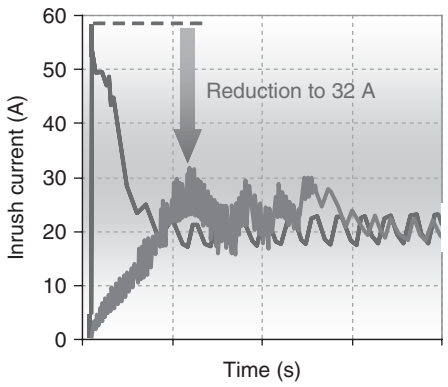


Figure 21. Comparison between relay controlled and PWM controlled compressor activation. (Reproduced with permission from WABCO.)

3.3 Compressor control

Compressors for passenger car air suspension systems are designed for intermittent operation, as a continuous run is not required and would increase weight and cost. Consequently, the compressor has to be protected against overheating. In simple applications, there is a thermal cut-off switch positioned in the electric motor. In systems with electronic control, at least the environmental temperature of the car is used to estimate the working condition. To be more accurate, a dedicated temperature sensor can be used to provide a signal to the controller. This sensor should be placed close to the hot spot of the compressor (e.g., cylinder head) to provide a realistic temperature information to the controller. In some cases, the sensor is placed somewhere in the environment of the compressor to provide at least an indication about the actual environmental condition. For economical reasons, the temperature information can be achieved using the solenoid of the exhaust valve as a temperature-dependent resistor (Bodet and Meier, 2007) (patent no. DE102005062571A1).

Another control feature is the starting and stopping of the compressor against low pressure: to avoid strong torque reactions, the gallery and the air dryer should be evacuated before the compressor is started or stopped.

3.4 Air dryer

By compressing the air taken in from the atmosphere, the natural humidity has to be considered: during compression, the water content remains constant; however, the relative humidity is increasing. This means that at a certain point water droplets are appearing as condensate.

Example: filling of a 5-L reservoir from 0 to 16 bar with air, $t = 30^{\circ}C$, and 75% relative humidity. Total mass of collected water is 1,767 g, which are about 35 droplets.

Car manufacturers expect a maintenance-free system, so a periodical manual drain of the condensate is not acceptable. Consequently, the air dryer is mandatory to avoid humidity/water ingress into the air suspension system. Without an air dryer, the amount of water transported into the air springs would be more than 5 L (considering average climate conditions in Stuttgart) and 20 L in Singapore over the vehicle lifetime with an open system. This calculation does not consider any leakage. In reality, there is always

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

leakage present. Considering a leakage rate of 5%, an open system would be able to provide dry air over the lifetime, as the air dryer regeneration does not need all air back during exhaust to achieve full regeneration.

In a closed system, the air dryer is only required in case of system filling in a workshop and in case of leakages. 5% leakage rate means 0.25 L water over the lifetime in Stuttgart or 1 L in Singapore. For this reason, even a closed system has to have an air dryer, and special purge cycles have to be fulfilled (Section 3.1.3).

As the air dryer itself has to be maintenance free, it has to provide a self-regenerating function. As desiccant in those applications, special beads of crystalline metal-aluminum silicates are used. These beads are produced synthetically and are highly porous, having a strong affinity to water. Depending on temperature, the desiccant is able to store water mass of 20% of its own weight. Figure 22 shows the principle.

Humid air flows through the desiccant cartridge (1). During this process, water is extracted from the air and stored temporarily in the desiccant beads (adsorption). The dry air leaves the cartridge via a check valve (2) into the pneumatic system.

During each lowering of the chassis level, the dry air that comes from the bellows flows through the air dryer back into the atmosphere. The 3/2 way solenoid valve (3) is activated and opens the relay valve (5). The air passes the solenoid valve and the throttle (4). In the throttle, the relative humidity of the air further decreases due to expansion. While passing the air dryer, the expanded air is regenerating the desiccant, meaning the air takes the water out of the desiccant beads and carries it through the relay valve (5) out into the atmosphere (desorption).

Key for efficient regeneration is the ratio between the nominal width at the entrance of the air dryer (dry air) and at the outlet (humid air). The orifice area at the outlet should be roughly four times bigger than at the inlet. This is to ensure a sufficient expansion of the compressed air. For further reference on this topic, see Section 3.5.

3.5 Exhaust circuits at open systems

The exhaust circuit of an ASU has to fulfill several demands:

- Controllable with electrical energy (solenoid valve).
- Providing sufficient airflow from the system into the atmosphere to achieve high chassis lowering speed.
- Expansion of the compressed air before it enters the air dryer for efficient regeneration.
- Robustness against water/ice in the "wet" area.
- Possible separation of the air dryer volume from the gallery. Advantage: air dryer volume does not need to be pressurized for a pressure measurement in the bellows or in the reservoir, thus the air loss during pressure measurement is minimized.
- Provide low noise level during exhausting.

Depending on the vehicle demands, there are several exhaust circuits on the market (see Figures 23 and 24).

3.5.1 Standard release circuit

The standard release exhaust circuit is also shown and explained in detail in Figure 22. To avoid too big and heavy solenoid valves, this circuit uses a small 3/2 solenoid as the pilot valve (3). Thus, the electrical current to operate the solenoid can be minimized. The 3/2 valve operates a relay valve (5) by air, which opens a big orifice to allow high airflow without velocity pressure out of the air dryer

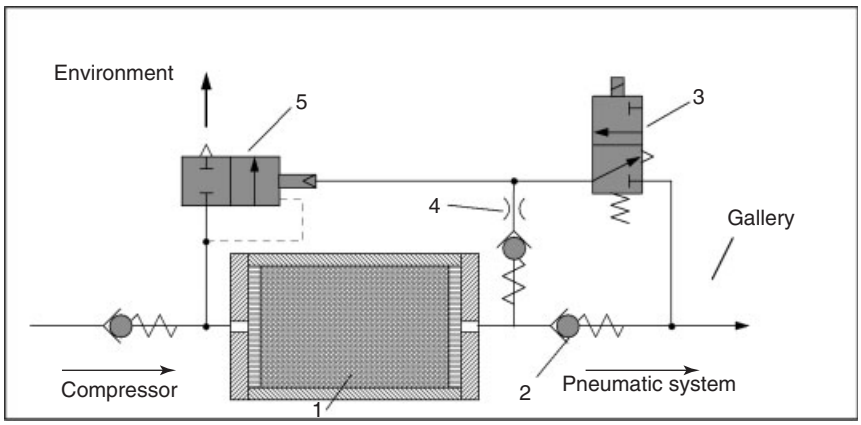


Figure 22. Standard release air dryer circuit. (Reproduced with permission from WABCO.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

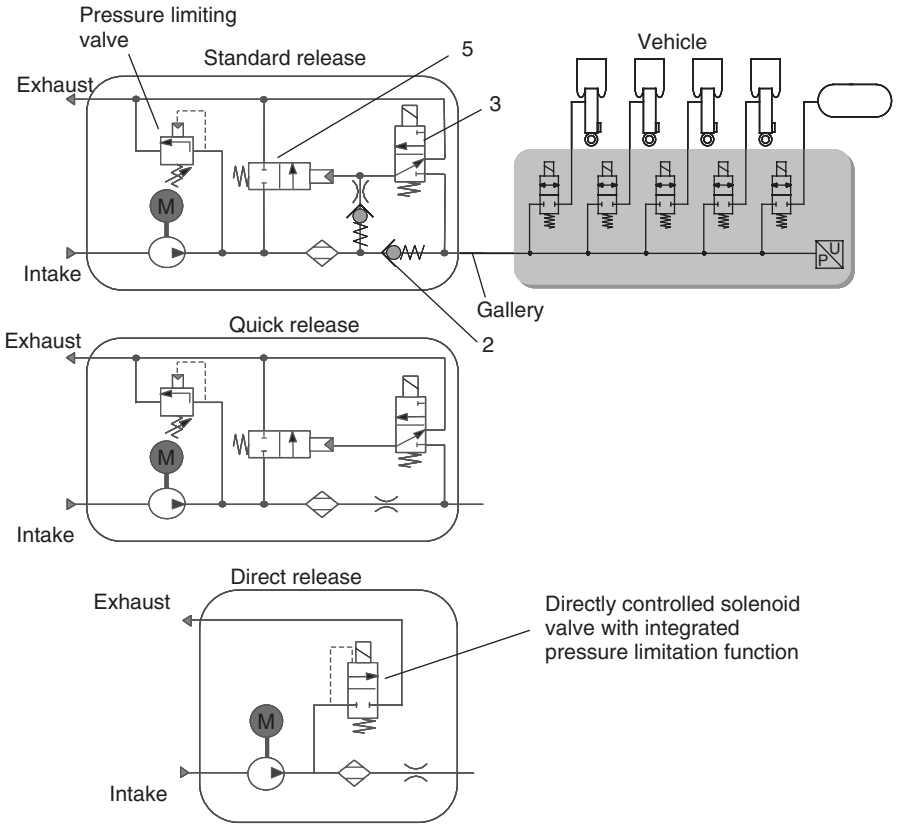


Figure 23. Different air dryer circuits in idle mode. (Reproduced with permission from WABCO.)

during exhausting. The relay valve itself, which operates in the wet area, is robust against water/ice. Separation of air dryer volume and gallery is achieved using a check valve (2). The only disadvantage of this circuit is a limited airflow, as the exhausting air has to pass the 3/2-solenoid valve (3).

3.5.2 Quick release circuit

The disadvantage of a limited orifice is not acceptable for vehicles that require quick lowering (SUV's). For this reason, the standard release circuit was modified and called a *quick release* circuit. Here, the exhausting air is not passing through the solenoid valve, and a bigger orifice is possible. To achieve this, the check valve (2) does not exist, which creates one disadvantage that the gallery and air dryer volume are always connected. In case of pressure measurements in the bellows or the reservoir, the air dryer volume has always to be inflated that consumes air and increases the chilling time for a pressure measurement.

3.5.3 Direct release circuit

For economical reasons, the relay valve/pilot solenoid valve is being replaced by one bigger solenoid valve. This circuit is called *direct release* (Figure 25). The solenoid valve is positioned behind the air dryer, and the exhausting air has to pass the solenoid valve. Special focus is on the robustness against water/ice, as the solenoid valve is located in the wet area. Pressure limiting function is integrated in the solenoid valve. Further advantages are achieved by integrating the filter/silencer into the air dryer body.

3.6 Pressure retention

In some cases, air suspension systems require a minimum pressure within the bellows, for example, if the car is lifted in a workshop. Without pressure retention, the bellows might be exhausted completely and would be damaged in case of lowering the car. The pressure retention function is provided either by dedicated valves in each bellow or by one central valve in the air dryer of the ASU. In case of standard and quick release circuit, the relay piston,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

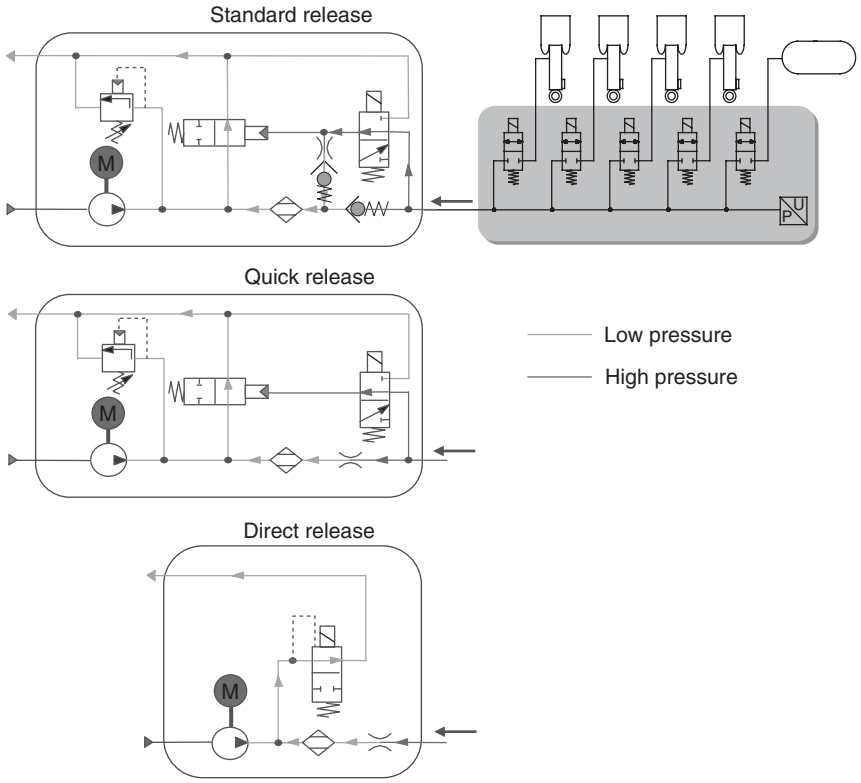


Figure 24. Different air dryer circuits in exhaust mode. (Reproduced with permission from WABCO.)

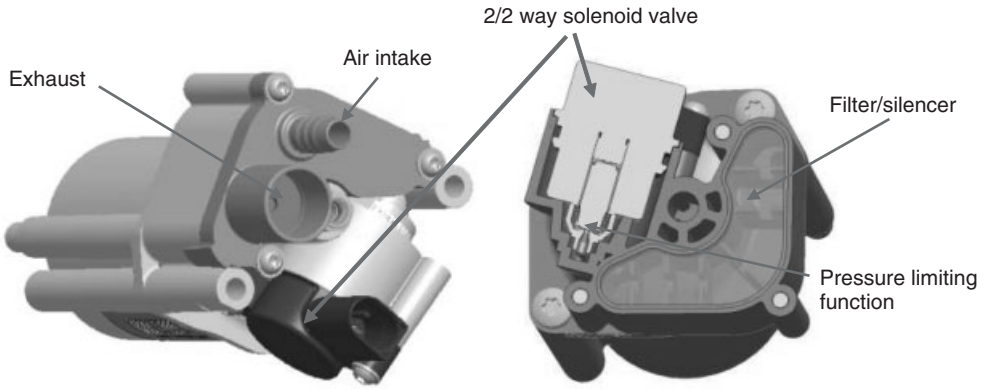


Figure 25. Direct release air dryer with integrated filter/silencer. (Reproduced with permission from WABCO.)

which closes by help of a spring, provides the pressure retention function if the system pressure drops below a specified level (Figure 4). Typically, this is between 3 and 1.5 bar.

If the air suspension system requires lower bellow pressures, this cannot be achieved with such a circuit. However, the direct release air dryer is able to exhaust the system down to atmospheric pressure.

3.7 Pressure limiting

To mitigate a risk that comes mainly from system FMEA (failure mode and effects analysis) considerations, a pressure-limiting function has to be integrated in the air suspension system. Typically, the ASU provides this function. This feature is only required if one other component has a failure. As an example, a sticking relay that

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

controls the electric motor/compressor might occur. As the compressor cannot be switched off anymore, the pressure in the compressor and the gallery would increase up to the natural stabilization pressure, depending on compressor performance and height above sea level. The desired cut-off pressure should be above the normal working pressure with sufficient margin. In reservoir systems, the maximum operating pressure is about 16–18 bar. Considering technical tolerances, the blow off valve should not open before 17.5–19.5 bar.

Pressure-limiting function is provided by a valve in the relay piston (Figure 4) or by the 2/2-way solenoid valve in the direct release air dryer (Figure 23). The bellow valves must be safe against this pressure level (ratio between gallery and bellow pressures); otherwise, the compressed air could pass through the solenoid valve and might lift the chassis unintentionally.

Although the pressure-limiting valve avoids damage in the system caused by too high pressure, the compressor might run until it overheats or empties the vehicle battery. Owing to this risk, some applications have electronic "relays" in both supply and ground line. A safe shut off of the compressor is always possible.

3.8 Air distribution

3.8.1 Air intake

The compressor has to get clean air from the atmosphere. For this reason, an air filter is installed in the air intake (Figures 2, 25, and 26). Particle sizes smaller than 10 μm are acceptable. The position where the air is ingested has to be in a dry environment (e.g., inside the car or at a place above the wade line) and where splash water has no access.

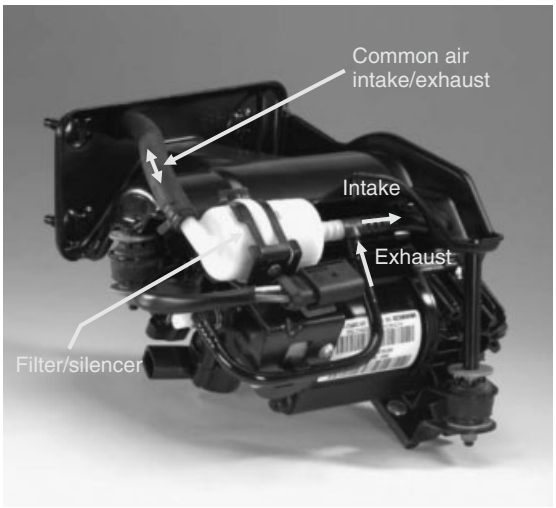


Figure 26. Air supply unit with common air intake/exhaust. (Reproduced with permission from WABCO.)

3.8.2 Exhaust

The exhaust port is noise critical. Therefore, a position outside the cab is favorable. A silencer is mandatory in many cases. Often, the air intake and exhaust of the compressor is combined into one port. Advantage is that the air filter can be used as silencer and ease of assembly. An example is shown in Figure 26.

A disadvantage of a combined air intake/exhaust is the trade-off between dry position and exhaust noise. In addition, the air-drying capability can be diminished. In the case of the exhaust process, humid air is guided through the comment pipe/filter. There might be condensation effects if the pipe temperature is lower than the exhausted air. During the next compressor run, the humid air—or even

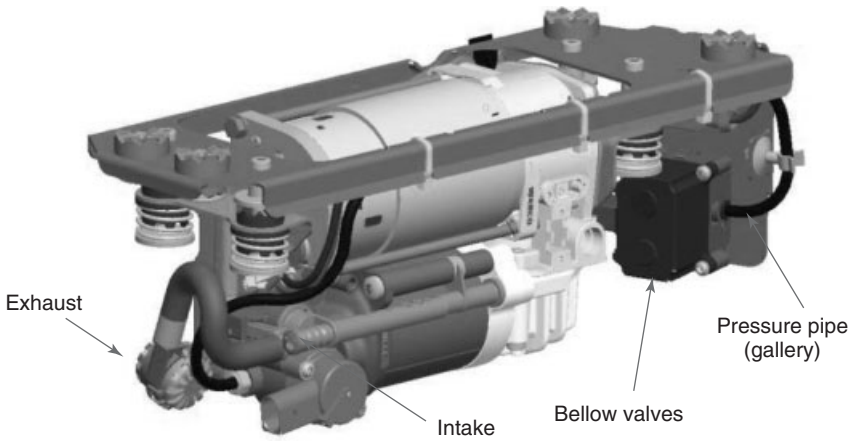


Figure 27. Air supply unit for rear axle air suspension, air intake, and exhaust separated. (Reproduced with permission from WABCO.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

water drops—in the pipe are ingested into the ASU. In case of longer air intake/exhaust pipes, a separation is recommended (Figure 27).

In many cases, the solenoid valves to control the air springs (bellow valves) and reservoir are assembled close to the compressor on one bracket (Figure 27). The connection between the multiple solenoid valve block and the compressor (gallery) should be as short as possible to reduce the volume inside this pipe. The diameter has to be selected depending on the nominal width of the bellow valves. Typically, it is a pipe 6 × 1.5 mm (outside diameter 6 mm, inside diameter 3 mm, and wall thickness 1.5 mm), in some cases, 4 × 1 mm is also used (inside diameter 2 mm).

A typical 5 by 2/2-solenoid valve block is shown in Figure 28.

A single pressure sensor is integrated. By connecting it to the bellows and the reservoir in a sequence one by one, the individual pressure can be measured. Here, the disadvantage of the quick release and direct release circuit becomes obvious, because in each case the air dryer volume has to be pressurized during a measurement. This takes a certain time and the pressure value can only be taken after a chilling time. After each measurement, the air dryer volume has to be exhausted.

Air distribution from the multiple solenoid valve block to the bellows is usually done with PA tubes 4 × 1 mm. This dimension avoids expensive preforming of the tubes, which would be required in case of bigger tubes (e.g., 6 × 1.5 mm).

Depending on the exhaust circuit, special attention is required on the pneumatic layout and dimensioning. In case of low bellow pressures and large pressure drops due to long pipes with small orifices, the relay piston in the air dryer might not receive sufficient pressure levels to stay fully open. In this case, the relay piston would swing within an intermediate position and would not be able to open the complete nominal width. This results in a bad regeneration of the air dryer, as the expansion of the compressed air cannot be provided as specified.

At rear axle systems, an integration of the bellow valves into the compressor can be advantageous. An example is shown in Figure 29.

4 COMPRESSOR PACKAGING/ACOUSTIC INSULATION

4.1 Inside mounting

Inside mounting is advantageous in terms of low demand for protection against water, corrosion, and so on. On the other hand, acoustic (air borne noise) and vibration (structure borne noise) of the ASU is very demanding. The

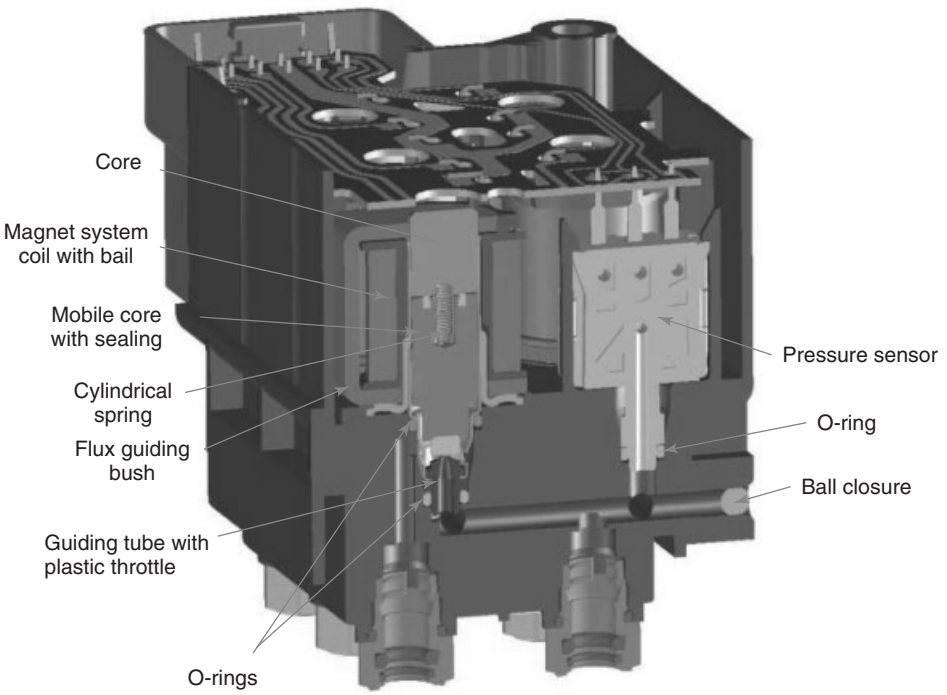


Figure 28. 5 by 2/2-solenoid valve block with integrated pressure sensor. (Reproduced with permission from RAPA GmbH)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

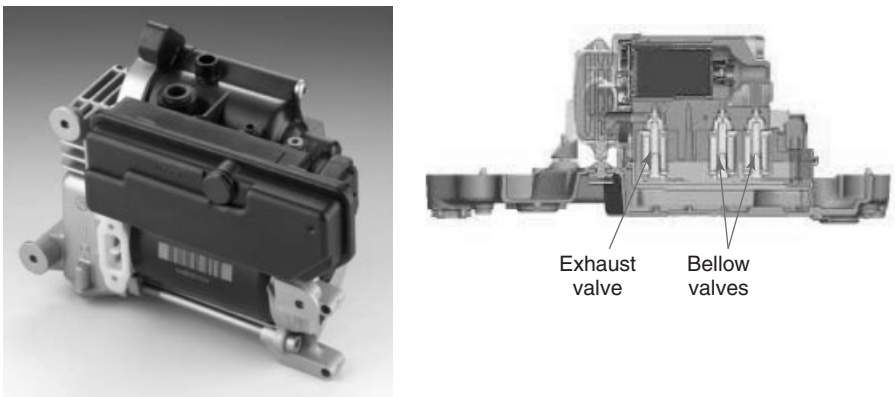


Figure 29. Air supply unit for rear axle air suspension with integrated bellow valves. (Reproduced with permission from WABCO.)

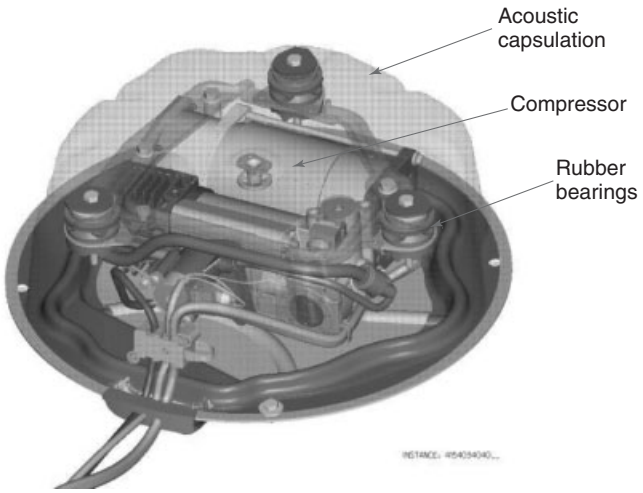


Figure 30. Capsulated air supply unit. (Reproduced with permission from WABCO.)

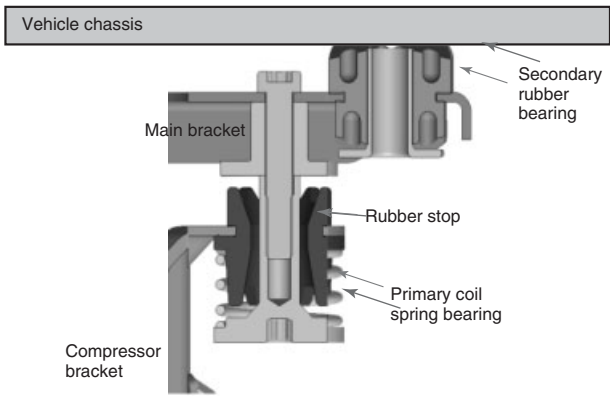


Figure 31. Typical insulation design between compressor and vehicle chassis. (Reproduced with permission from WABCO.)

compressor has to be suspended on flexible elements such as rubber bearings as shown in Figure 30.

Best insulation is achieved with coil springs (Figure 31). The tuning of the mass/spring system is typically subcritical, meaning that the eigenfrequency of the compressor/spring combination is below the first order of eigenfrequency of the compressor vibration if it is running. A well-balanced load distribution between the bearings (3 or 4) is key for NVH (noise, vibration, and harshness) reduction. As the coil springs have to be soft and damping is just done by a certain amount of friction, stroke limiters are required. To avoid the compressor structure hitting the bracket, rubber buffers are used. One reason for compressor movement is vibration caused by the vehicle on uneven roads. Another reason is the torque reaction if the compressor is switched on or off.

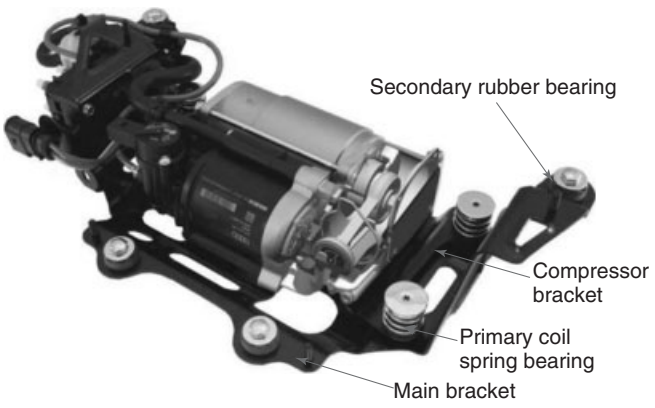


Figure 32. Air supply unit for four-corner air suspension. (Reproduced with permission from WABCO.)

4.2 Outside mounting

Corrosion resistance has to be provided following ISO 9277 NSS, typically 480 h. Protection class requirement is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

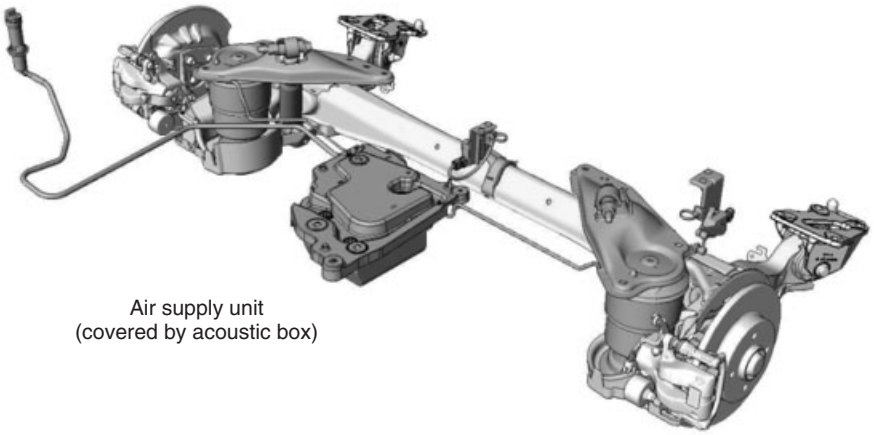


Figure 33. Rear axle installation air supply unit with acoustic box. (Reproduced with permission from WABCO.)

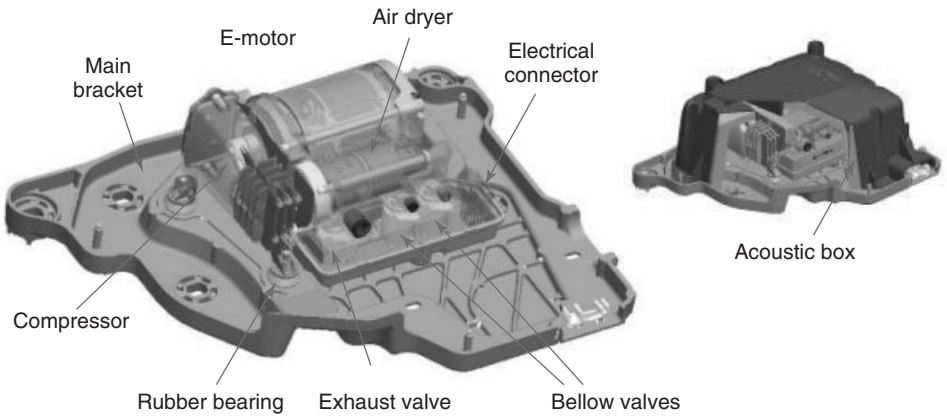


Figure 34. Details of a rear axle installation air supply unit with acoustic box. (Reproduced with permission from WABCO.)

typically IP6K6K, IP6K7, and IP6K9K according to ISO 20653.

Compared to the inside mounting, the air-borne noise is not that critical. In many cases, the ASU is not capsulated (Figure 32). A preferred installation space is the engine compartment. However, packaging and the neighborhood to hot elements might be critical. In many cases, the ASU is installed below the chassis in the rear area of the car. There might be hot elements (exhaust muffler) in the neighborhood as well.

In the following application, the ASU is covered by a case, which reduces airborne noise and protects the ASU against stone chipping and ground contact (Figure 33).

Details of the ASU for this rear axle solution are shown in Figure 34.

These examples show that—depending on the OEM (original equipment manufacturer) demand—there are various solutions how and where to place an ASU in a car. There is no standard solution, and the engineering effort to package a compressor into a car is significant both on OEM side and on supplier side.

5 SUMMARY

The purpose of ASUs in passenger cars is to feed the air springs with compressed dry air. Piston compressor technology is standard on the market, all driven by electric DC motors. System pressures vary between 7 and 20 bar, depending of the vehicle application. There are one-stage

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

compressors (for open and closed systems) and two-stage compressors (for open systems) used. Closed systems are pumping the air between the air springs and a reservoir back and forth, avoiding an exhaust of the pressurized air into the atmosphere. By keeping the pressurized air widely in the system, the overall energy balance has an advantage compared to open systems. In addition, the repeatability of lifting operations is better than with open systems. However, closed systems are more complex and a reservoir is mandatory. Furthermore, the maintenance-free air drying capability of an ASU requires purge cycles at closed systems, which worsens the energy balance. Open systems are less complex and are able to be operated without reservoirs, which also have a positive impact on overall system costs. One-stage compressors are simpler and have a better efficiency in low pressure ranges than two-stage compressor, whereas two-stage compressors have physical advantages in higher pressure ranges.

Besides those differentiators at the compressor, the major variation is packaging of compressors into the vehicle. As there is no "standard" installation space and environment, each new vehicle requires an individual design of bracket, insulation, piping, and wiring harness.

Future challenges for ASUs are weight and size limitations, acoustic performance, current consumption, peak current, and increasing pressure demand for smaller air springs.

REFERENCE

Bodet, M.-M., Meier, J. (2007) Verfahren zur Ermittlung einer Kompressorumgebungstemperatur und Kompressoranordnung zur Durchführung des Verfahrens. Offenlegungsschrift vom 28.06.2007, DE102005062571 A1.

FURTHER READING

Becher, H.O. (2004) *Steuergeräte und Kompressoren für Luftfedersysteme*. CTI Fachkonferenz Federung und Dämpfung im Fahrwerk, Stuttgart July 06, 2004.

Becher, H.O. (2008) *Rear axle air suspension for compact class and transporters*. 4th International CTI Conference Suspension & Damping Stuttgart, April 23, 2008.

Fitch, B. (2008) A new high power compressor helps to realize high-performance for air suspension systems. Vehicle Dynamics Expo North America, October 23–25, 2008.

Meier, J. (2006) Steuerung von Luftfedern in Straßenfahrzeugen. Technische Akademie Esslingen, Oktober 25, 26, 2006.

Westerkamp, H. (2007) Air suspension enters the European compact car segment. Vehicle Dynamics Expo Stuttgart, May 8, 2007.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto016
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Customer-Oriented Evaluation of Vehicle Handling Characteristics

Adrian Mihailescu, Stephan Poltersdorf, and Lutz Eckstein

RWTH Aachen University, Aachen, Germany

1 Introduction 1
2 Subjective Vehicle Handling Evaluation 1
3 Methods for Objective Evaluation 4
4 Evaluation of Subjective and Objective Data 6
5 Overview on Significantly Correlating Vehicle Handling Characteristics 7
6 Summary 15
References 16

1 INTRODUCTION

This chapter gives an overview on methods that can be used to evaluate the handling quality of passenger cars with respect to the demands of real customers.

1.1 Motivation

The increasing handling quality and driving safety of modern passenger cars makes it difficult for car manufacturers to stand out in this area of vehicle quality. The available development resources must be strictly focused on technical changes that assure improved handling evaluations by customers (and journalists). The vehicle developers, therefore, need efficient methods to evaluate the perceived handling quality in all phases of product development.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

1.2 Subjective versus objective evaluations

Classical evaluation approaches are subjective evaluations by customers or vehicle dynamics experts. Considering various rules regarding the execution (Section 2) and data analysis (Section 4), this method can generate valuable results.

The reproducibility and validity of subjective evaluations is limited because of several (human) factors, whereas the necessary effort remains high. Furthermore, the necessity for real vehicles conflicts with the increasingly virtual vehicle development process. These problems led to various approaches to support or even replace subjective evaluations by objective methods based on measurements or simulations (described in Section 3). The goal of these approaches is the identification of objective values that significantly correlate with subjective vehicle evaluations, and that can, therefore, be used to predict customer verdicts.

Successful methods for objective evaluations lead to well-defined and testable development targets. The product quality can be evaluated more often and more efficiently, thus reducing the dependence on varying or even biased subjective evaluations.

Over the last decades, a large number of useful objective measures have been identified. Section 5 gives an overview based on an extensive meta-study.

2 SUBJECTIVE VEHICLE HANDLING EVALUATION

In order to gather detailed and reliable handling evaluation based on the perception of individual drivers, several basic rules have to be considered. After a brief introduction of

the principles of subjective handling perception, these basic rules are presented in Sections 2.2 and 2.3.

2.1 Principles of subjective handling perception

Each driver stabilizes the vehicle on the chosen trajectory based on various methods of information perception and information processing. The driver's sensations and experiences in the scope of this stabilization determine the subjective evaluation of vehicle handling.

2.1.1 Information perception

The driver perceives information on the current driving situation mainly via three sensory channels: the visual, the vestibular (resulting from the organ of equilibrium), and the haptic (feeling sensors such as hands) channels (Mitschke and Niemann, 1972).

First of all, the high resolution of the visual channel enables the perception of deviations from the desired trajectory, the vehicle orientation, and the steering wheel angle. Velocities and accelerations can also be perceived based on the optical flow. Because of the widespread and precise visual information, this channel is the main information source for trajectory stabilization in the linear handling area (Schimmel, 2010). The equilibrium organs (vestibular system) enable the perception of linear and rotary accelerations. Compared to the visual channel, the vestibular sensation is less accurate but faster. The fastest information is provided by the haptic sense, which enables the indirect perception of vehicle accelerations and of the steering wheel torque across the contact surfaces between driver and vehicle.

2.1.2 Information processing

The processing of the perceived driving-state information can be described by a three-staged model (Rasmussen, 1983): skill-based driver actions constitute the first level and are very rapid and highly automated behavioral patterns not consciously modulated by the driver. On the second level, the driver makes unconscious decisions based on memorized experiences. The speed of information processing at this level is slower than on the first level. If the driver possesses no suitable skills or experiences for the current driving situation, he or she must consciously analyze the driving situation. This level of information processing is the third and the slowest stage of information processing.

With respect to subjective vehicle evaluations, one fact is crucial: because of different levels of skill and experience, different drivers handle identical driving situations using different levels of information processing and differently weighed sensory information. This results in a high variance of driver actions and vehicle evaluations among different drivers. For example, a skilled driver uses the haptic and the vestibular senses much more than an untrained driver, who mainly controls the vehicle based on visual information, which causes differences, for example, in the evaluation of steering feel.

2.2 Execution of subjective vehicle evaluations

Table 1 gives an overview on typical handling characteristics evaluated in subjective evaluation (Heißing and Brandl, 2002; Eckstein, 2011). The table shows the classes of characteristics that can be further detailed into subcharacteristics—especially in the scope of expert

Table 1. Criteria classes of subjective evaluation.

	On-center handling			Corner handling		
Steering feel	Center point feeling	Steering effort	Steering torque level and buildup	Steering torque evel and progress	Steering angle demand	Steering response
	Steering response	Self-centering	Steering friction	Stability feel	Steering precision	Feedback of road friction
		Post-pulse oscillation (hands off)		Torque and angle during parking	Steering clearance	Steering returnability
Vehicle behavior	Straight driving stability	Directional stability under braking	Response behavior	Self-steering behavior	Response behavior	Pitch and roll behavior
	Load-change behavior	General (straight) wind sensitivity	Cross wind sensitivity	Load-change behavior	Stability and lane change behavior	Corner exit behavior
	Lateral inclination sensitivity	Trailer stability	Aquaplaning behavior	Corner braking behavior	Aquaplaning behavior	

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

evaluations. Subjective characteristics that are supported by significantly correlating objective measures (Section 5) are marked in gray.

2.2.1 Choice of drivers

Usually, professional test drivers are used for vehicle evaluations because of their more accurate, differentiated, and reproducible judgment (Schimmel and Heißing, 2009; Pfeffer and Scholz, 2010; Riedel and Arbinger, 1997). In contrast, untrained drivers offer realistic representation of the skill and experience of the real customers. The lower precision and reproducibility of their judgment is sometimes accepted, for example, if questions of driving safety or controllability are investigated (Bubb, 2003). A collective of normal drivers is best generated by random sampling among large heterogeneous groups of possible customers (Breuer, 2009). Bortz (2010) and Bubb (2003) describe how the appropriate size of the driver collective can be calculated based on the questions under investigation, the desired level of significance, the used evaluation scales, and the statistical distribution of data collected in pretests. This chapter can only give rough recommendations based on these sources: tests with professional drivers only need small numbers of drivers (2–5), especially when new vehicle variants are compared to reference variants that have been evaluated by a larger number of drivers. Tests with normal drivers that are used to compare two vehicle variants need more drivers (30–50). If tests with normal drivers shall be used to understand the relationship between subjective evaluations and certain driver characteristics (gender, age, and so on), a much higher number of drivers (e.g., >1000) are needed for highly significant results.

2.2.2 Choice of driving scenario

The vehicles must be evaluated in scenarios that provide realistic representations of real driving situations. The vehicles should be driven "closed loop" on a clearly defined course, so that the task of trajectory stabilization needs to be fulfilled (Kudritzki, 1989).

The shape and velocity profile of the used course should realistically depict the typical driving range with respect to velocity, lateral acceleration, and necessary steering frequencies. This is achieved by choosing a course with a wide range of cornering radiuses (20–500 m). Corners with special characteristics (lateral inclination, increasing curvature, and so on) should also be present with different radiuses (Heißing and Brandl, 2002). The necessary steering frequencies can be influenced by more or less immediate changes in corner direction and by the curvature gradients at the beginning and the end of corners (abrupt vs continuous changes in curvatures).

The maneuver definition must clearly describe in which handling area the vehicle has to be driven (linear area or nonlinear limit area). For limit maneuvers, Kudritzki (1989) and Neukum, Krüger, and Schuller (2001) suggest driver-dependent velocities, as the significance of subjective ratings increases as soon as the evaluating driver approaches his or her individual velocity limit.

2.2.3 Execution of driving tests

The test conditions (tire condition, vehicle load, track temperature, weather, daytime, and so on) must be kept as constant as possible. Evaluations should not last longer than 1 h (Barthenheier, 2004).

All vehicles should be "blind tested," that is, no driver should know which variant he or she is currently driving (Zong, Guo, and Hsin, 2000). Ideally, the vehicle variants should only vary with respect to the evaluated characteristics in order to prevent disturbances caused by the unintended evaluation of characteristics such as design or brand.

Variants should be evaluated together with a reference vehicle that has been evaluated by a large number of drivers before. Thus, relating the evaluation results to the highly significant evaluations of the reference variant can increase the significance of the results.

It is also recommended to repeat the evaluation of some variants, for example, the reference variant, in order to detect whether the scale of individual evaluations drifts off during a test-drive (Dettki, 2005). In order to prevent the so-called transfer effect (the judgment of one variant is influenced by the variants driven before), each driver should drive the variants in a different order (Bubb, 2003).

2.3 Design of subjective evaluation questionnaires

Commonly printed questionnaires with rating scales are used for the evaluation. Sometimes, an interviewer in the car personally asks for the evaluation. Käppler (1993) and Riedel and Arbinger (1997) point out that the low reliability of many subjective evaluations is caused by an inappropriate questionnaire design. Therefore, the evaluation scales and the kind, number, and formulation of the questions should be chosen thoughtfully.

2.3.1 Absolute versus relative evaluations

Absolute vehicle evaluations try to rate the vehicle handling on an absolute and universally valid scale. Different

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

vehicles can afterward be compared and ranked based on their absolute ratings. Relative evaluations do only try to evaluate a vehicle in comparison to a second vehicle. If all vehicle variants are compared to each other relatively, a ranking can be generated based on these relative evaluations.

Even for professional drivers, it is much easier to rate a vehicle in comparison to a reference vehicle than to rate a vehicle on an absolute scale (Mummendey, 2003). If preliminary tests show that reliable and reproducible absolute evaluations are not possible for a given investigation, it is still possible to switch to a more reliable relative evaluation.

2.3.2 *Content and formulation of evaluation questions*

The questions must be adapted to the vocabulary and experience of the drivers, as a prior "teaching" of expert language to normal drivers would not significantly improve the quality of the results (Kudritzki, 2000). The evaluation becomes easier for normal drivers when they are not asked for a technical vehicle rating but for their personal sensations and experiences in the vehicle (see scale example in Figure 2).

The total number of questions should be kept below 10 (Kudritzki, 2000; Heißing and Brandl, 2002).

2.3.3 *Evaluation scales*

Usually, drivers are asked to evaluate the handling with discrete scales. The steps of a scale must be "anchored" by verbal classifiers, that is, by descriptions of the individual scale steps, in order to minimize the room for (mis)interpretations.

Depending on the scope of the investigation, two basic scale types can be used. Open "unipolar" scales are used for absolute evaluations. On these scales, the magnitude of each rating is described on an absolute scale without a reference. Closed "bipolar" scales are used for relative vehicle evaluations. In this case, each vehicle is rated in comparison to a second (reference) vehicle.

When choosing the number of scale steps, a trade-off between reliability losses because of unperceivable scale differences and detail losses because of a too coarse scale has to be found. In general, the standard deviation for the repeated evaluations of identical vehicle variants by identical drivers should stay below 2 scale steps (Käppler, 1993; Bortz, 2010).

Käppler (1993) recommends an effective scale range of 7 steps for unipolar scales and professional drivers. Harrer, Pfeffer, and Johnston (2006) recommend an effective scale range of ± 6 steps for bipolar scales and professional drivers.

2.3.4 *Visual design of scales*

According to the western reading direction, the intensity of the scale anchors should increase horizontally from left to right. The full-scale range should be shown visually, continually, and undistorted (Käppler, 1993).

Figures 1 and 2 show two scale variants that have been derived from the explained principles. The first example (Figure 1) shows an absolute scale for professional evaluators and asks for a three-stage evaluation.

The second example (Figure 2) shows a relative, bipolar scale for untrained evaluators.

3 METHODS FOR OBJECTIVE EVALUATION

Various approaches try to identify objective characteristic values that strongly determine the customer rating of a vehicle. Figure 3 provides an overview on these methods.

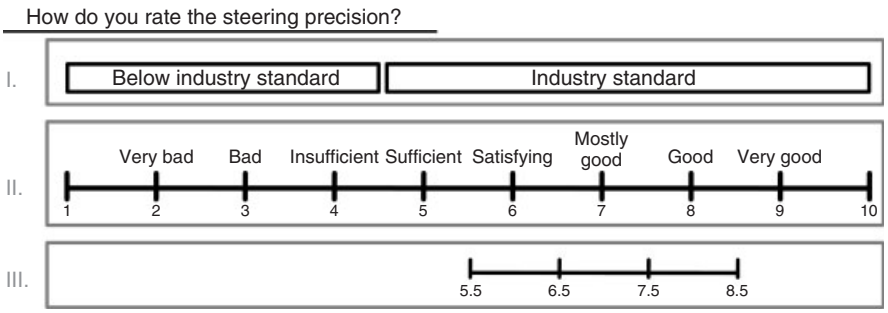


Figure 1. Unipolar scale for professional evaluators.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

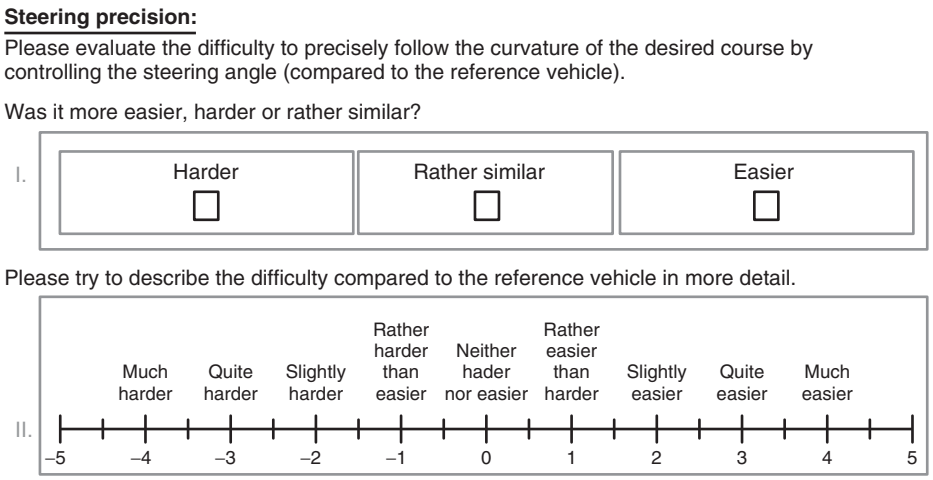


Figure 2. Bipolar scale for untrained evaluators.

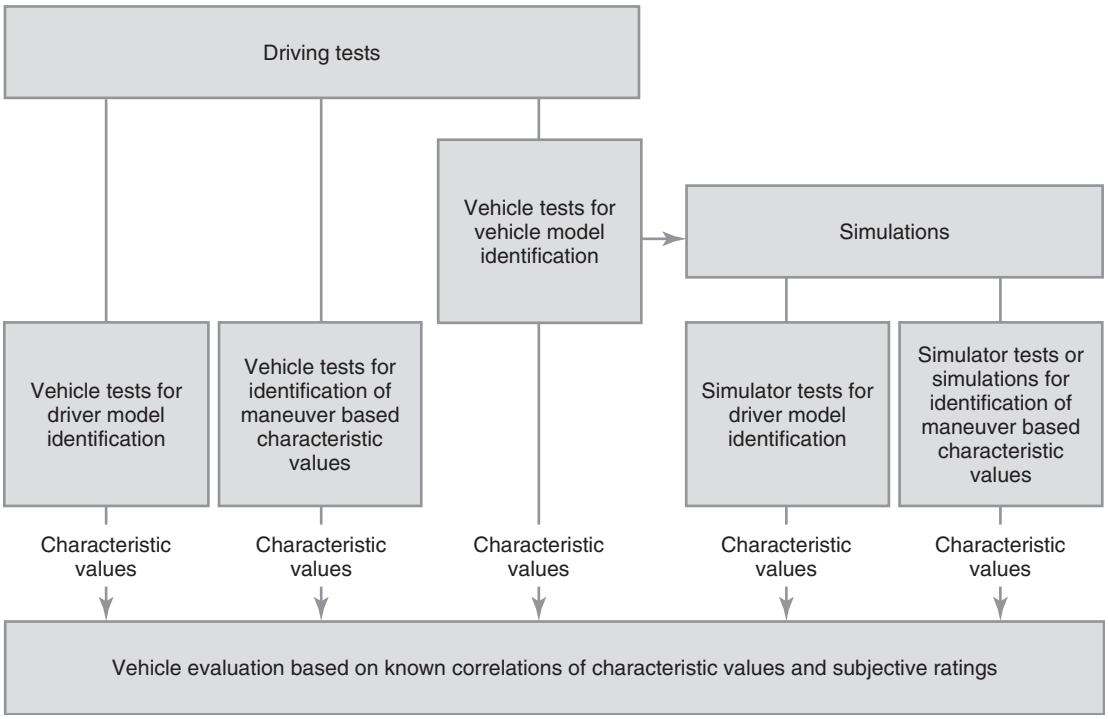


Figure 3. Objective handling evaluation methods.

3.1 Objective evaluation based on characteristic values from driving maneuvers

The classic method of objective vehicle evaluation collects characteristic values from vehicle measurements. Most of these values can also be collected efficiently in vehicle dynamic simulations. Usually, the characteristic values are collected in open-loop maneuvers, as, in this case, a high reproducibility can be achieved because of the reduced driver influence.

It is recommended that the catalog of characteristic values covers all information paths of human drivers. Visual characteristic values (e.g., lateral track offset) should be combined with haptic characteristic values (e.g., steering torque) and vestibular characteristic values (e.g., lateral acceleration) (Wagner, 2003).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Section 5 presents a large number of such characteristic values that correlate significantly with subjective evaluations. Most of these characteristic values are context-sensitive, that is, they only correlate with subjective evaluations from similar driving situations (Jürgensohn, Willumeit, and Irmscher, 1999; Pietsch, Schimmel, and Heißing, 2009a; Sagan, 2003). Therefore, a detailed evaluation of the complete handling behavior can only be achieved by a large variety of maneuvers and characteristic values.

3.2 Objective evaluation based on vehicle models

Vehicle models can be used for objective vehicle evaluations in two ways: one possibility is the determination of characteristic values from the vehicle model parameters themselves. For example, Barthenheier (2004) evaluates the steering behavior of a vehicle based on a simple steering model with the three parameters self-aligning torque, steering damping, and steering friction. So far, only a small number of significant correlations have been identified with this method.

The second and the more promising possibility is the determination of the classic characteristic values (Section 3.1) with the help of vehicle dynamic simulations. Thus, a large number of characteristic values can be gathered with high efficiency and reproducibility.

The models (e.g., extended single-track vehicle model, the dual-track vehicle model, or the steering-system models described by Zschocke, 2009, and Winner *et al*., 2003) can be parameterized automatically based on test-rig or test-drive measurements (Meyer-Tuve, 2009; Meljnikov, 2003; Pietsch, Schimmel, and Heißing, 2009a; Zschocke, 2009).

3.3 Driver-based characteristic values

A more direct way to predict subjective vehicle evaluations would be the determination of characteristic values based on the reactions of the driver in closed-loop maneuvers. A driver adapts his or her control behavior to the handling properties of a vehicle based on his or her skills and experiences (Wallentowitz, 1979). The easier this adaptation is, the higher his or her evaluation of the vehicle is going to be (Abe and Kano, 2008). This suggests that characteristic values based on the interaction of driver and vehicle highly correlate with subjective vehicle evaluations.

The central problem of all these driver-based objectivity approaches is the high variance among drivers. Using professional drivers and repeated measurements can reduce this problem.

3.3.1 Characteristic values based on driver–vehicle interaction

The results of closed-loop maneuvers can be used to determine characteristic values based on the positions of the control elements, especially the steering angle. For example, the steering angle measured in a double-lane change can be used to evaluate the vehicle stability and controllability in this maneuver. In the end, this approach is a closed-loop, driver-oriented variant of the classical maneuver-based approach.

3.3.2 Characteristic values based on driver models

This approach is based on the assumption that synthetic models can describe the control behavior of the driver. As the driver has to adapt to each vehicle, the parameters of the respective driver model must be vehicle dependent and can be utilized for objective vehicle evaluations (Abe and Kano, 2008; Henze, 2004; Decker, 2008; Dibbern, 1992). Common driver models are based on a combination of a feedforward part and a feedback part. The feedforward part estimates the steering angle necessary to follow the path ahead based on a linear single-track vehicle model. The feedback part compensates the path deviation caused by the inaccuracies of this model with a steering angle controller, for example, a PID controller. The vehicle-dependent parameters of these driver models (i.e., the feedforward and feedback amplification factors) can be automatically identified with the help of the test-drive data, either from a test track or from a driving simulator.

3.3.3 Characteristic values based on driver perception

Schimmel (2010) and Scharpe *et al*. (2012) suggest that the significance of characteristic values increases when the vehicle movement is transformed with respect to the paths of subjective handling perception. For example, characteristic values based on the lateral acceleration of vestibular organs (i.e., the driver head acceleration) or characteristic values based on the lateral pressure between driver and seat are more significant than characteristic values based on the lateral acceleration of the vehicle.

4 EVALUATION OF SUBJECTIVE AND OBJECTIVE DATA

The data gathered in subjective and objective evaluations must be correctly processed and analyzed in order to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

consider all relevant information and in order to prevent misjudgments. The description of the necessary mathematical methods lies beyond the scope of this chapter. Nevertheless, we would like to emphasize the necessity and the main goals of the data processing.

4.1 Processing of data from subjective evaluation

The subjective data must be tested for four central characteristics:

Objectivity: It must be mathematically proven that the results are independent from the persons conducting the experiments, for example, the interviewer in the car (Lienert and Raatz, 1969).

Reliability: It must be mathematically proven that the data are internally consistent and provide stable results in repeated evaluations (Magnusson, 1975; Riedel and Arbinger, 1997; Redlich, 1994; Bortz, 2010).

Validity: It has to be tested whether a scale/question is really measuring the attribute it is supposed to. It should be possible to distinguish vehicles differing in the analyzed characteristic by comparing the respective answers (Riedel and Arbinger, 1997).

Subjective–subjective correlation: It has to be tested whether there are redundancies in between questions/scales. In this case, removing these redundancies should shorten the questionnaire.

4.2 Processing of data from objective measurements

The objective data from measurements or simulations also have to be mathematically tested for reliability (internal consistency and repeatability) and validity (strong relationship to measured characteristic). An objective–objective correlation can be used to shorten the catalog of characteristic values by removing redundancies (Harrer, 2007). The data preprocessing must ensure the comparability of measurements from different vehicles with different sensor configurations, for example, by the removal of sensor delays and offsets and by the transformation of measurements to default reference points such as the center of gravity of the vehicle (Kobetz, 2004; Pfeffer and Harrer, 2011).

4.3 Correlation and regression analyses

As soon as the subjective and objective data have been processed as described earlier, the correlation and regression analyses are used to analyze the relationship between them. The correlation test determines the statistical significance (i.e., the probability of error) of relationships between objective characteristics and subjective evaluations. The absolute value of the correlation coefficient can be used as a measure for this significance (Bortz, 2010; Cohen and Cohen, 1983).

The regression analysis goes one step further by generating models that can be used to predict subjective ratings based on linear or nonlinear combinations of one or more objective values (Kudritzki, 1989; Dibbern, 1992; Harrer, 2007).

Only significantly correlating characteristics should be used for the vehicle evaluation. A certain amount of remaining probability of error (e.g., 5% for significant or 1% for highly significant correlations) is accepted.

5 OVERVIEW ON SIGNIFICANTLY CORRELATING VEHICLE HANDLING CHARACTERISTICS

Some of the characteristic values found in the literature are statistically proven, some are solely theories, and some are disapproved but often used nevertheless. In order to ensure the customer relevance, the characteristic value has to have been validated by real test-drives and regression analyses. In the further approach of this chapter, only the verified characteristic values of the 250 researched literature sources are considered.

Figure 4 shows the classification structure of the chosen vehicle handling characteristics and the section where it is elucidated.



Figure 4. Overview on the handling classification structure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In the scope of this chapter, we first subdivide the handling into two areas: *on-center handling* refers to the steering behavior on and around straight ahead driving (Farrer, 1993); the handling at higher lateral acceleration is referred to as *corner handling*. In a second classification level, vehicle handling characteristics are categorized according to the vehicle response's effect on the driver: the *steering feel* and the *vehicle behavior* are being considered. The steering feel of the driver is a central handling characteristic and is, therefore, investigated separately for both handling areas. The vehicle behavior section sums up all the literature findings on verified correlations not related to steering feel—for example, yaw behavior or roll behavior.

The follwing sections show the verified correlations and their characteristic values. For each correlation, there is a sign describing the preferable value of the objective characteristic:

⇧ stands for high values;
⇩ stands for low values;
⊙ stands for an optimal value range.

These recommendations are based on investigations done by the referenced sources and not by the authors of this chapter. In some cases, there are further notes added to the correlation. This is marked by a code and is listed in the corresponding section.

5.1 On-center handling

5.1.1 Steering feel

Correlation notes sorted by code (Table 2):

OS1: The evaluation is done between the speeds of 160 and 200 km/h.

$$K_{\mathrm{L}} = \left| \frac{\delta}{\tau \upsilon^2_{\mathrm{res}}} \right|_{f_{\mathrm{R,opt}}} \cdot f_{\mathrm{R,opt}} \tag{1}$$

$\tau \upsilon^2{}_{\mathrm{res}}$ is a value describing the intensity of the stochastic crosswind disturbance. The value τ describes the angle between the longitudinal axis of the car and the direction of the resultant wind velocity $\upsilon^2{}_{\mathrm{res}}$, which is build by vector addition of the vehicle speed and the side wind velocity. The quotient is the resonant amplification factor between crosswind and steering reactions. $f_{\mathrm{R,opt}}$ is the resonant frequency of the driver's reaction to the wind disturbance.

OS2: The evaluation is done while driving between the speeds of 160 and 200 km/h.

$$K_{\mathrm{V}} = \frac{\left| \frac{\dot{\psi}}{\tau \upsilon^2_{\mathrm{res}}} \right|_{\max}}{\left| \frac{\dot{\psi}_{\tau \upsilon^2_{\mathrm{res}}}}{\tau \upsilon^2_{\mathrm{res}}} \right|_{\max}} \tag{2}$$

$\dot{\psi}_{\tau \upsilon^2_{\mathrm{res}}}$ is the yaw rate with and $\dot{\psi}$ the yaw rate without driver intervention. $\tau \upsilon^2{}_{\mathrm{res}}$ is a value describing the intensity of the stochastic crosswind disturbance.

OS3: The evaluation is done while driving at the speed of 100 km/h. $W_{\delta M_{\mathrm{H}}}$ is the steering work.

$$W_{\delta M_{\mathrm{H}}} = \delta \cdot M_{\mathrm{H}} \tag{3}$$

δ is the steering angle demand and M_{H} the steering torque demand. $W_{\delta M_{\mathrm{H}}}$ values below 5° Nm are preferable.

OS4: The evaluation is done while driving at the speed of 100 km/h.

$$\widetilde{\delta}_{\mathrm{H}} = \sqrt{\int_a^b \Phi_{\delta_{\mathrm{H}}}(f)\mathrm{d}f} \tag{4}$$

$\Phi_{\delta \mathrm{H}}(f)$ is the power spectral density of the steering angle. $\Phi_{\delta \mathrm{H}}(f)$ is integrated for frequencies between a and b. $\widetilde{\delta}_{\mathrm{H}}$ values below 0.09° for steering frequencies between 0.2 and 1.5 Hz and values below 0.011° for steering frequencies between 1.5 and 3.0 Hz are preferable.

OS5: A delay between 0 and 0.55 ms for a sporty car and between 0.55 and 0.8 ms for a comfortable car is preferable.

OS6: A gradient less than 0.05 Nm/(°)^2 is preferable.

OS7: A quotient between 0.22 and 0.3 Nm/° for a comfortable car and between 0.3 and 0.35 Nm/° for a sporty car is preferable. The evaluation is done at the speed of 100 km/h.

OS8: An amplification factor between 0.22 and 0.28 1/s for a sporty car and between 0.18 and 0.22 1/s for a comfortable car is preferable. The evaluation is done at the speed of 100 km/h.

OS9: This correlation is dependent on the vehicle segment. The evaluation is done at the speed of 120 km/h.

OS10: The evaluation is done at the speed of 80 km/h.

OS11: The evaluation is done at the steering frequencies between 0.4 and 1 Hz.

OS12–14: The evaluation is done at the speed of 120 km/h.

OS15–17: The evaluation is done at the speed of 80 km/h and the steering frequency of 0.5 Hz.

OS18: The evaluation is done at the speed of 120 km/h and the steering frequency of 0.5 Hz. The steering torque is measured during the reduction of the lateral acceleration.

OS19: A height between 0.5 and 1.5 Nm is preferable.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Objective criteria on-center handling (steering feel).

On-center handling—Steering feel					
Subjective characteristic	Objective characteristic	Code/note		Reference	Maneuver
Steering effort	Steering angle		⊙	Farrer, 1993	Weave test
	Steering angle rate		⊙		
	Characteristic based on the amplification function between the wind disturbance and steering intervention	*OS1*	⇧	Wagner, 2003	Crosswind
	Characteristic based on the amplification functions between the wind disturbance and the resulting yaw rate with and without driver intervention	*OS2*	⇧		
	Steering work	*OS3*	⇩	Dettki, 2005	Straight driving with lateral inclination or crosswind
	Characteristic based on frequency spectrum of steering correction	*OS4*	⊙		
	Hysteresis area in phase diagram steering angle—lateral acceleration		⇩	Decker, 2008	Weave test
	Steering angle at lateral acceleration of 0 m/s^2		⇧		
	Hysteresis area in phase diagram steering angle/yaw rate		⇩		
Steering torque level and buildup	Steering torque gradient at lateral acceleration of 0 and 1 m/s^2		⊙	Pietsch and Heissing, 2009b	Sine steer
	Ratio of the steering torque/lateral acceleration hysteresis' width and height		⇩	Decker, 2008	Weave test
	Steering torque at 0°/s yaw rate in hysteresis steering torque/yaw rate diagram		⇧		
	Gradients mean value in hysteresis steering torque/yaw rate diagram		⊙		
Steering response	Phase delay between steering angle and yaw rate	*OS5*	⊙	Dettki, 2005	Weave test
	Second derivative of steering torque with respect to the steering angle	*OS6*	⇩		
	Ratio steering torque/steering angle	*OS7*	⊙		
	Yaw rate amplification factor	*OS8*	⊙	Dettki, 2005 Harrer, 2007	
	Gradient of lateral acceleration with respect to steering angle while steering away from center	*OS9*	⊙	Harrer, 2007	
	Steering torque dead band with respect to lateral acceleration		⇩	Farrer, 1993	Weave test or step steer
	Lateral acceleration dead band with respect to steering angle		⇩	Harrer, 2007	
	Lateral acceleration amplification factor	*OS10*	⊙	Harrer, 2007	Single sine
Center point feeling	Ratio steering torque/steering angle	*OS11*	⇧	Farrer, 1993	Weave test
				Schimmel, 2010	Frequency response
				Harrer, 2007	Weave test
	Ratio lateral acceleration/steering angle		⇧	Schimmel, 2010	
	Frequency at the phase minimum in the frequency response function between steering angle and steering torque		⇧		Frequency response
	Ratio lateral acceleration/steering angle		⇧		Weave test

(*continued overleaf*)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. (*Continued*)

On-center handling—Steering feel					
Subjective characteristic	Objective characteristic	Code/note		Reference	Maneuver
Center point feeling	2D rel.: char. 1: ratio steering torque/lateral acceleration; char. 2: ratio yaw velocity/steering angle	*OS12*	⊙	Harrer, 2007	Weave test
	2D rel.: char. 1: ratio steering torque/lateral acceleration; char. 2: ratio lateral acceleration/steering angle	*OS13*	⊙		
	Steering torque at lateral acceleration of 0 m/s^2	*OS14*	⊙		
	2D rel.: char 1: ratio yaw velocity/steering angle; char 2: steering friction	*OS15*	⊙		
	2D rel.: char. 1: ratio steering torque/lateral acceleration; char. 2: ratio yaw velocity/steering angle	*OS16*	⊙		
	2D rel.: char. 1: ratio steering torque/lateral acceleration; char. 2: ratio lateral acceleration/steering angle	*OS17*	⊙		
	Steering torque at lateral acceleration of 1 m/s^2	*OS18*	⊙		
	Hysteresis height steering angle/steering torque	*OS19*	⊙	Dettki, 2005	
	Hysteresis width steering angle/steering torque	*OS20*	⊙		
Self-centering	Steering torque gradient		⇧	Zschocke, 2009	Weave test
	Ratio steering torque/steering angle		⇧		
Steering friction	2D rel.: char. 1: ratio yaw rate/steering angle; char. 2: phase delay steering torque/steering angle	*OS21*	⊙	Harrer, 2007	Weave test
	2D rel.: char. 1: phase delay yaw rate/steering torque; char. 2: yaw rate response gain	*OS22*	⊙		
	2D rel.: char. 1: phase delay yaw rate/steering torque; char. 2: ratio yaw rate/steering angle	*OS23*	⊙		
	Steering torque rate at lateral acceleration of 0 m/s^2	*OS24*	⊙		
	Residual steering angle in lateral acceleration hysteresis		⊙	Zschocke, 2009	

OS20: A width between 1.5 and 5.0° is preferable.
OS21–23: The evaluation is done at the speed of 120 km/h.
OS24: The evaluation is done at the speed of 120 km/h.

5.1.2 Vehicle behavior

Correlation notes sorted by code (Table 3):

OV1: Gradients under 1/90 (°)m^2/Ns are preferable. The evaluation is done while driving straight at the speeds of 100, 140, and 180 km/h.

OV2: The evaluation is done while driving straight between the speeds of 160 and 200 km/h.

$$K_G = \left| \frac{\dot{\psi}}{\tau v_{res}^2} \right|_{max} \quad (5)$$

$\tau v^2{}_{res}$ is a value describing the intensity of the stochastic crosswind disturbance.

OV3:

$$RAT = 11.97 - 0.67 \cdot \dot{\Psi}_{max} - 7.37 \cdot T_{eq} \quad (6)$$

$\dot{\Psi}_{max}$ is the first maximum of the yaw rate. T_{eq} is the equivalent time delay (measured out of separate sine steer maneuver).

OV4: Values below 0.25 1/s are preferable. The maneuver is done at 100 km/h.

OV5: The evaluation is done at the speed of 190 km/h and a constant lateral inclination of 1.5%.

$$\tilde{\dot{\psi}} = \sqrt{\int_a^b \Phi_\psi(f)\mathrm{d}f} \quad (7)$$

$\Phi_\psi(f)$ is the power spectral density of the yaw rate. The root of the integral of $\Phi_\psi(f)$ between the steering frequencies 0.2 and 1.5 Hz shall have values below 0.1°/s.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Objective criteria on-center handling (vehicle behavior).

On-center handling—Vehicle behavior					
Subjective characteristic	Objective characteristic	Code/note		Reference	Maneuver
Cross wind sensitivity	Gradient of ratio yaw rate deviation/crosswind pressure	*OV1*	⇩	Dettki, 2005	Crosswind
	Characteristic based on the maximum of the amplification factor between the wind disturbance and the resulting yaw rate	*OV2*	⇩	Wagner, 2003	
	Characteristic based on the first yaw rate maximum and the equivalent time delay (measured out of separate sine steer maneuver)	*OV3*	⇧	Zomotor, Braess, and Rönitz, 1997	
Lateral inclination sensitivity	First derivative of the yaw rate deviation with respect to the roll angle	*OV4*	⇩	Dettki, 2005	Straight driving
	Characteristic based on the frequency spectrum of the yaw rate	*OV5*	⇩	Dettki, 2005	Straight driving with lateral inclination
Straight driving stability	Characteristic based on the steering angle in combination with the time period between two steering corrections	*OV6*	⇧	Engels, 1993	Straight driving
	Characteristic based on the hysteresis between lateral acceleration and steering angle	*OV7*	⇩	Loth, 1997	Sine steer
	Lateral acceleration at zero steering torque	*OV8*	⇧	Zschocke, 2009	Weave test
	Difference between the lateral acceleration amplification factor at 0.2 and at 0.6 Hz		⇧		ISO frequency response test
Response behavior	Steering angle dead-band with respect to yaw rate		⇩	Farrer, 1993	Weave test, transition test
	Time lag between steering angle and yaw rate		⇩		

OV6:

$$\mathrm{SU} = 11.67 - 0.47\left(\frac{\overline{\delta}_{\mathrm{Leff}}}{0.3}\right) - 1.59\left(\frac{14.3}{\overline{t}_{\mathrm{Spur}}}\right) \qquad (8)$$

$\overline{\delta}_{\mathrm{Leff}}$ is the effective steering angle of a period. $\overline{t}_{\mathrm{Spur}}$ is the time period between two steering corrections. Maneuver is done at 90 and 150 km/h.

OV7:

$$\phi = \arcsin\left(\frac{\delta_{\mathrm{LP}}}{\widehat{\delta}_{\mathrm{L}}}\right) \qquad (9)$$

δ_{LP} is half the width of the lateral acceleration/steering angle hysteresis at $0\,\mathrm{m/s^2}$ lateral acceleration. $\widehat{\delta}_{\mathrm{L}}$ is half the width of the hysteresis at maximum lateral acceleration.

OV8: The evaluation is done at the speeds of 100 and 150 km/h.

5.2 Corner handling

5.2.1 *Steering feel*

Correlation notes sorted by code (Table 4):

CS1–2: The evaluation is done at the speed of 80 km/h, the steering frequency of 0.4 Hz, and the lateral acceleration of 0.4*g*.

CS3–5: A progressive steering torque increase is preferable, that is, low torque at $1\,\mathrm{m/s^2}$, medium torque at $4\,\mathrm{m/s^2}$, and a high torque difference between 0.5 and $1\,\mathrm{m/s^2}$. The evaluation is done at the speeds of 70 and 100 km/h.

CS6–7: The evaluation is done at the speed of 100 km/h and the lateral acceleration of 0.4*g*.

CS8–10: The evaluation is done at the speed of 120 km/h and the steering frequency of 0.5 Hz with a steering amplitude of 10°.

CS11: The evaluation is done up to a lateral acceleration of 0.4*g* with a step steer maneuver at the speed of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Objective criteria corner handling (steering feel).

Corner handling—Steering feel					
Subjective characteristic	Objective characteristic	Code/note		Reference	Maneuver
Steering torque level and progress	Ratio peak yaw rate/peak steering torque	*CS1*	⊙	Harrer, 2007	Single sine
	Ratio peak steering torque/peak lateral acceleration	*CS2*	⊙		
	Steering torque at lateral acceleration of 1 m/s^2	*CS3*	⇧	Zschocke, 2009	Steering angle ramp
	Steering torque at lateral acceleration of 4 m/s^2	*CS4*	⇧		
	Difference between steering torque at lat. acc. of 0.5 m/s^2 and 1 m/s^2	*CS5*	⇧		
Steering returnability	Stationary residual steering angle (steering release after 2 s cornering at 4 m/s^2)		⇩	Zschocke, 2009	Steering return
Steering precision	Steering torque at zero steering angle		⇧	Decker, 2008 Zschocke, 2009	Sine steer/weave
	Hysteresis area in phase diagram steering angle/lateral acceleration		⇩	Decker, 2008	
	Steering angle at lateral acceleration of 0 m/s^2		⇩		
	Lateral acceleration at zero steering angle		⇧		
	Phase delay between steering angle and yaw rate at 1 Hz		⇩		
	Characteristic based on the delay time in the lateral acceleration frequency response function		⇩		
	Peak roll rate	*CS6*	⇩	Harrer, 2007	Step steer
	2D rel.: char. 1: peak roll rate; char. 2: ratio of peak roll rate and steady state value of lateral acceleration	*CS7*	⊙		
	2D rel.: char. 1: ratio yaw rate/steering angle; char. 2: steering friction	*CS8*	⊙		Sine steer/weave
	2D rel.: char. 1: ratio yaw rate/steering angle; char. 2: steering angle hysteresis in torque/angle diagram	*CS9*	⊙		
	2D rel.: char. 1: peak value of yaw rate; char. 2: steering angle hysteresis in torque/angle diagram	*CS10*	⊙		
	Ratio lateral acceleration/steering angle		⇧		
	Steering torque hysteresis in torque/angle diagram		⊙		
	Lateral acceleration hysteresis in lateral acceleration/steering angle diagram		⊙		
	Lateral acceleration at 0 Nm steering torque		⊙	Zschocke, 2009 Harrer, 2007	
	Ratio steering torque/steering angle		⇧	Schimmel, 2010	
	Steering torque phase response below 1 Hz		⇧	Zschocke, 2009	
Steering clearance	Phase lead of the steering torque		⇩	Zschocke, 2009	Sine steer/weave
Stability feel	Steering torque hysteresis at 0° steering angle (0.25–0.75 Hz)		⇧	Zschocke, 2009	Sine steer/weave

(*continued overleaf*)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. (*Continued*)

Corner handling—Steering feel					
Subjective characteristic	Objective characteristic	Code/note		Reference	Maneuver
Steering angle demand	Amplification factor between yaw rate and steering angle		⊙	Decker, 2008 Harrer, 2007 Schimmel, 2010 Zschocke, 2009	Sine steer/step steer/weave/ steady-state cornering
	Amplification factor between lateral acceleration and steering angle		⊙	Decker, 2008 Harrer, 2007 Schimmel, 2010	Sine steer/weave
	Hysteresis area in steering angle/lateral acceleration diagram		⇩	Decker, 2008	Sine steer
	Medial steering ratio		⊙	Zschocke, 2009	Steady-state cornering
Steering response	Amplification factor between yaw rate and steering angle	*CS11*	⊙	Harrer, 2007	Step steer/single sine/weave/ frequency response
	Peak value of lateral acceleration	*CS12*	⊙		Sine steer/weave
	Peak value of yaw rate	*CS13*	⊙		
	Amplification factor between lateral acceleration and steering angle	*CS14*	⊙		Lane change/ single sine/ weave
	Yaw acceleration	*CS15*	⇧	Wolf, 2008	Step steer
	Phase delay between steering angle and yaw angle	*CS16*	⇩		

80 km/h or with a single sine maneuver at the speed of 80 km/h, the steering frequency of 0.2 Hz or with a weave test at the speed of 120 km/h, and the steering frequency of 0.5 Hz with an steering amplitude of 20° or with a frequency response test at the speed of 100 km/h.

CS12–13: The evaluation is done at the speed of 120 km/h.

CS14: The evaluation is done at the speed of 80 or 100 km/h and the lateral acceleration of 0.4*g*.

CS15–16: The evaluation is done at the speed of 80 km/h.

5.2.2 *Vehicle behavior*

Correlation notes sorted by code (Table 5):

CV1: Values around 0.2–0.3 s are optimal.

CV2: The TB2 value is the product between the peak response time $T_{\dot{\psi}_{\max}}$ of the yaw rate and the maximum vehicle sideslip angle $\beta_{\max}$.

$$\mathrm{TB2} = T_{\dot{\psi}_{\max}} \cdot \beta_{\max} \tag{10}$$

The TB2 value is not valid for vehicles fitted with rear-wheel steering.

CV3: $\ddot{\Psi}_{\mathrm{QuMW}}$ is the yaw acceleration mean square value.

$$\ddot{\Psi}_{\mathrm{QuMW}} = \frac{\left|\sum_{i=0}^{t_e} \frac{\ddot{\Psi}_i}{|\ddot{\Psi}_i|}(\ddot{\Psi}_i)^2\right|}{\sum_{i=0}^{t_e} \frac{\ddot{\Psi}_i}{|\ddot{\Psi}_i|}(\ddot{\Psi}_i)^2} \frac{\sqrt{\left|\sum_{i=0}^{t_e} \frac{\ddot{\Psi}_i}{|\ddot{\Psi}_i|}(\ddot{\Psi}_i)^2\right|}}{100 t_e} \tag{11}$$

t_0 is the point in time of brake application and t_e the point in time of the first driver interaction but has a maximum of 2 s.

CV4:

$$\ddot{\Psi}_{\mathrm{relAbsMaz}} = \frac{\left|\ddot{\Psi} - \ddot{\Psi}_0\right|}{\ddot{\Psi} - \ddot{\Psi}_0} \cdot \left(\left|\ddot{\Psi} - \ddot{\Psi}_0\right|\right)_{\max_{t_0 - t_e}} \tag{12}$$

$\ddot{\Psi}_0$ is the reference yaw rate acceleration, t_0 the point in time of brake application, and t_e the point in time of the first driver interaction but has a maximum of 2 s.

CV5:

$$\ddot{\Psi}_{\mathrm{Diff}} = \left|\left(\ddot{\Psi} - \ddot{\Psi}_0\right)_{\max_{(t_0 - t_e)}}\right| - \left|\left(\ddot{\Psi} - \ddot{\Psi}_0\right)_{\min_{(t_0 - t_e)}}\right| \tag{13}$$

$\ddot{\Psi}_0$ is the reference yaw acceleration and t_0 the point in time of brake application.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 5. Objective criteria corner handling (vehicle behavior).

Corner handling — Vehicle behavior					
Subjective characteristic	Objective characteristic	Code/note		Reference	Maneuver
Self-steering behavior	Gradient of steering wheel angle with respect to lateral acceleration at 6 m/s^2		⇩	Zschocke, 2009	Steady-state cornering
Corner braking behavior	Overshoot extent of the lateral acceleration		⊙	Dreyer, 1990	Corner braking/ step steer under braking
	Response time of lateral acceleration and yaw rate	*CV1*	⊙		
	Product of response time of yaw rate and peak vehicle sideslip angle	*CV2*	⇩		
	1 s-value of the yaw angle deviation or of the vehicle sideslip angle		⇩	Zomotor, Braess, and Rönitz, 1997	Corner braking
	Peak vehicle sideslip angle		⇩		
	Characteristic based on the yaw acceleration mean square value	*CV3*	⇩	Schick and Bunz, 2002	
	Characteristic based on the peak relative yaw acceleration	*CV4*	⇩		
	Characteristic based on the difference between the minimum and maximum relative yaw acceleration	*CV5*	⇩		
Pitch and roll behavior	Characteristic based on the wheel travel	*CV6*	⇩	Kawagoe, Suma, and Watanabe, 1997	Cornering under positive longitudinal acceleration
	Characteristic based on the lateral acceleration, roll angle, rate, and acceleration	*CV7*	⇩	Botev, 2008	Lane change
	Roll rate (at 0.8 Hz)		⇩	Seyed-Ghaemi, 2005	Sine steer
	Derivative of the roll angle with respect to lateral acceleration		⇩		Steady-state cornering
	Amplification factor between roll angle and steering angle (0.5 Hz)		⇩	Zschocke, 2009	ISO frequency response
	Overshoot extent of the roll angle at 7 m/s^2		⇩		Step steer
	Characteristic based on steering angle, roll angle, roll rate, and acceleration	*CV8*	⇩	Botev, 2008	Lane change
Response behavior	Yaw Eigen-frequency		⇧	Schimmel, 2010	Frequency response
	Steering torque mean amplitude decrease		⇧		
	Yaw rate peak response time		⊙		Step steer
	Overshoot extent of yaw rate		⊙		
	Amplification factor between yaw rate and steering angle		⊙	Zomotor, Braess, and Rönitz, 1997	Step steer
	Response time of lateral acceleration and yaw rate		⇩		
	Overshoot extent of lateral acceleration and yaw rate		⊙		
	Characteristic based on the product of the yaw rate peak response time and the stationary vehicle sideslip angle	*CV9*	⇩		
	Steering angle at lateral acceleration of 0 m/s^2		⇩	Decker, 2008	Sine steer

(*continued overleaf*)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 5. (*Continued*).

Corner handling—Vehicle behavior					
Subjective characteristic	Objective characteristic	Code/note		Reference	Maneuver
Stability and lane change behavior	Vehicle sideslip angle		⇩	Riedel and Arbinger, 1997	Double lane change
	Vehicle sideslip angle rate		⇩	Zschocke, 2009	Steady-state cornering
	Roll rate		⇩	Riedel and Arbinger, 1997	Double lane change
	Time period after which the vehicle sideslip angle exceeds a threshold value		⇩		
	Vehicle sideslip angle relative to the lateral acceleration and to the steering angle of the second steer back		⇩	Botev, 2008	Double lane change
	Phase shift between lateral acceleration and yaw rate (0.5 Hz)		⇩	Huneke *et al*., 2010	Sine steer
	Characteristic based on the unbalance of the steering angle history of the second lane change and the time delay between steering angle and lateral acceleration	*CV10*	⇩	Dibbern, 1992	Double lane change
Load-change behavior	Characteristic based on the mean yaw rate deviation divided by a reaction time (0.75 s) and the mean yaw acceleration		⊙	Zomotor, Braess, and Rönitz, 1997	Load-change deceleration
	Characteristic based on the yaw rate and reference yaw rate 1.5 s after an accelerator pedal kick	*CV11*	⊙	Schweers and Röth, 1995	Load-change acceleration

CV6:

$$\mathrm{RMI} = \frac{1}{2}[(z_{\mathrm{i,f}} + z_{\mathrm{o,f}}) - (z_{\mathrm{i,r}} + z_{\mathrm{o,r}})] \tag{14}$$

z is the vertical wheel travel with the suffixes i (inner), o (outer), f (front), and r (rear). A negative RMI value is preferable.

CV7:

$$\mathrm{WI} = \frac{\ddot{\varphi}}{a_y} \cdot h_{\mathrm{k}} \cdot \frac{1}{\pi} + \frac{\dot{\varphi}}{a_y} \cdot h_1 + \frac{\varphi}{a_y} \cdot h_1 \cdot \pi \tag{15}$$

h_{k} is the height distance between the driver's head and the roll axle, a_y is the lateral acceleration, φ is the roll rate, and h_1 equals the length of 1 m.

CV8:

$$\mathrm{AWP} = \frac{\ddot{\varphi}}{\delta_{\mathrm{H}}} \cdot h_{\mathrm{k}} \cdot \frac{1}{\pi} + \frac{\dot{\varphi}}{\delta_{\mathrm{H}}} \cdot h_1 + \frac{\varphi}{\delta_{\mathrm{H}}} \cdot h_1 \cdot \pi \tag{16}$$

h_{k} is the height distance between the driver's head and the roll axle, δ_{H} is the maximal steering angle, φ is the roll rate, and h_1 equals the length of 1 m.

CV9: The TB value is the product between the peak response time $T_{\dot{\psi}_{\max}}$ and the stationary vehicle sideslip angle β_{stat}.

$$\mathrm{TB} = T_{\dot{\psi}_{\max}} \cdot \beta_{\mathrm{stat}} \tag{17}$$

The TB value is not valid for vehicles fitted with rear-wheel steering.

CV10: $\Delta\delta_{\max,2}$ is the dissymmetry of the steering angle history of the second lane change and $T_{0,(\delta,a_y)}$ the time delay between steering angle and lateral acceleration.

$$\mathrm{KD} = \Delta\delta_{\max,2} + 2.5 \cdot T_{0,(\delta,a_y)} \tag{18}$$

CV11: $\dot{\Psi}$ is the yaw rate and $\dot{\Psi}_{\mathrm{ref}}$ the reference yaw rate.

$$\frac{\dot{\Psi}}{\dot{\Psi}_{\mathrm{ref}}}(t_0 + 1.5\,\mathrm{s}) \tag{19}$$

t_0 is the time of the accelerator pedal kick.

6 SUMMARY

This chapter describes how the handling quality of passenger cars can be measured by subjective and objective evaluations. Beginning with an introduction of the basic principles of the subjective vehicle handling perception, practical advice for the planning, conduct, and analysis of subjective evaluations with groups of selected customers or professional drivers was given.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The compliance with the presented rules can significantly increase the informative value of these evaluations. Objective evaluation methods based on characteristic values promise a more efficient way to assess vehicle handling and facilitate evaluations in virtual simulation environments. Therefore, this chapter has provided an overview on possible methods of objective evaluation. The necessary analysis of subjective or objective data and the identification of the relationship among both by means of correlation and regression analyses were described. The correlation coefficient has proven to be a reasonable measure for the statistical significance of the relationship. Finally, on the basis of an extensive meta-study, an overview on customer-relevant vehicle handling measures was given. In order to ensure customer relevance, only characteristic values that have been validated by real test-drives and regression analyses were considered. The identified measures have been summarized for practical use in the form of Tables 2–5.

REFERENCES

Abe, M. and Kano, Y. (2008) A study on vehicle handling evaluation by model based driver steering behavior. FISITA 2008-03-022.

Barthenheier, T. (2004) Potenzial einer fahrertyp- und fahrsituationsabhängigen Lenkradmomentgestaltung. *Fortschritt-Berichte VDI Reihe*, **12**, 584.

Bortz, J. (2010) *Statistik für Human- und Sozialwissenschaftler*, 7th edn, Springer Verlag, Heidelberg.

Botev, S. (2008) Digitale Gesamtfahrzeugabstimmung für Ride und Handling. *Fortschritt-Berichte VDI Reihe*, **12**, 684.

Breuer, J. (2009) Bewertungsverfahren von Fahrerassistenzsystemen in *Handbuch Fahrassistenzsysteme* (eds H. Winner, S. Hakuli, G. Wolf), Springer Verlag, Heidelberg, pp. 55–68.

Bubb, H. (2003) Fahrversuche mit Probanden - Nutzwert und Risiko. *Darmstädter Kolloquium Mensch & Fahrzeug*, TU Darmstadt.

Cohen, J. and Cohen, P. (1983) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd edn, Erlbaum, Hillsdale, NJ.

Decker, M. (2008) Zur Beurteilung der Querdynamik von Personenkraftwagen. Thesis. TU Munich.

Dettki, F. (2005) Methoden zur objektiven Bewertung des Geradeauslaufs von Personenkraftwagen. Thesis. Universität Stuttgart.

Dibbern, K. (1992) Ermittlung eines Kennwertes für den ISO-Fahrspurwechsel in Versuch und Simulation. *Fortschritt-Berichte VDI Reihe*, **12**, 164.

Dreyer, A. (1990) Untersuchung des Fahrverhaltens von PKW bei kombinierten Lenk- und Bremseingaben. Thesis. RWTH Aachen University.

Eckstein, L. (2011) Vertical- and lateral dynamics of vehicles. Lecture Automotive Engineering II, RWTH Aachen University.

Engels, A. (1993) Geradeauslaufkriterien für Pkw und deren Bewertung. Thesis. TU Braunschweig.

Farrer, D.G. (1993) An objective measurement technique for the quantification of on-centre handling quality. SAE Paper 930827.

Harrer, M. (2007) Steering feel—objective assessment of passenger cars—analysis of steering feel and vehicle handling. Thesis. University of Bath.

Harrer, M., Pfeffer, P.E., and Johnston, N.D. (2006) Steering feel—objective assessment of passenger car analysis of steering feel and vehicle handling. FISITA 2006V165.

Heißing, B. and Brandl, H.J. (2002) *Subjektive Beurteilung des Fahrverhaltens*, Vogel Fachbuch Verlag, Würzburg.

Henze, R. (2004) Beurteilung von Fahrzeugen mit Hilfe eines Fahrermodells. Dissertation. Technische Universität Braunschweig, Schriftenreihe des Instituts für Fahrzeugtechnick TU Braunschweig, 7.

Huneke, M., Pascali, L., Strecker, F., *et al.* (2010) Identifikation von Fahrdynamikmodellen zur Fahrverhaltensbewertung und deren Anwendung in der Erprobung und Simulation bei der Fahrdynamikentwicklung. *15. VDI-Tagung Erprobung und Simulation in der Fahrzeugentwicklung, VDI-Berichte* 2016.

Jürgensohn, T., Willumeit, H.-P., and Irmscher, M. (1999) Fahrermodelle als Hilfsmittel zur Objektivierung von subjektiven Bewertungen der Fahrbarkeit. *Fortschritt-Berichte VDI Reihe*, **22**, 1.

Käppler, W.-D. (1993) Beitrag zur Vorhersage von Einschätzungen des Fahrverhaltens. *Fortschritt-Berichte VDI Reihe*, **12**, 198.

Kawagoe, K., Suma, K., and Watanabe, M. (1997) Evaluation and improvement of vehicle roll behavior. SAE Paper 970093.

Kobetz, C. (2004) Modellbasierte Fahrdynamikanalyse durch ein an Fahrmanövern parameteridentifiziertes querdynamisches Simulationsmodell. Thesis. TU Wien.

Kudritzki, D. (1989) Zum Einfluss querdynamischer Bewegungsgrössen auf die Beurteilung des Fahrverhaltens. *Fortschritt-Berichte VDI Reihe*, **12**, 132.

Kudritzki, D. (2000) Möglichkeiten zur Objektivierung subjektiver Beurteilungen des Fahrzeugverhaltens in *Subjektive Fahreindrücke sichtbar machen: Korrelation zwischen CAE-Berechnung, Versuch und Messung von Versuchsfahrzeugen und –komponenten* (ed. K. Becker), Expert-Verlag, Düsseldorf, pp. 11–26.

Lienert, G. and Raatz, U. (1969) *Testaufbau und Testanalyse*, 3rd edn, Verlag Julius Beltz, Weinheim.

Loth, S. (1997) Fahrdynamische Einflußgrößen beim Geradeauslauf von Pkw. Thesis. TU Braunschweig.

Magnusson, D. (1975) *Testtheorie*, 2nd edn, Deuticke, Wien.

Meljnikov, D. (2003) Entwicklung von Modellen zur Bewertung des Fahrverhaltens von Kraftfahrzeugen. Thesis. Universität Stuttgart.

Meyer-Tuve, H. (2009) Modellbasiertes Analysetool zur Bewertung der Fahrzeugquerdynamik anhand von objektiven Bewegungsgrößen. Thesis. TU Munich.

Mitschke, M. and Niemann, K. (1972) *Die Regeltaetigkeit des Autofahrers bei Kursabweichungen Deutsche Kraftfahrtforschung und Strassenverkehrstechnik*, VDI-Verlag, Düsseldorf, 221.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Mummendey, H.D. (2003) *Die Fragebogen Methode*, Hogrefe-Verlag, Göttingen.

Neukum, A., Krüger, H.-P., and Schuller, J. (2001) *Der Fahrer als Messinstrument für fahrdynamische Eigenschaften?VDI-Berichte Nr. 1613: Der Fahrer im 21. Jahrhundert*, VDI-Verlag, Düsseldorf.

Pfeffer, P.E. and Harrer, M. (2011) Lenkgefühl: Die Kunst der Beschreibung in *Lenkungshandbuch: Lenksysteme, Lenkgefühl, Fahrdynamik von Kraftfahrzeugen* (ed. M. Harrer), Vieweg+Teubner, Heidelberg.

Pfeffer, P.E. and Scholz, H. (2010) Present-day cars—subjective evaluation of steering feel. *Chassis.tech plus 2010*, Munich.

Pietsch, R., Schimmel, C., and Heißing, B. (2009) Objective assessment of handling performance. C*hassis.tech 2009*, Munich.

Pietsch, R. and Heißing, B. (2009) Modellbasierte Beurteilung des Lenkgefühls in *Subjektive Fahreindrücke sichtbar machen IV: Korrelation zwischen objektiver Messung und subjektiver Beurteilung in der Fahrzeugentwicklung* (ed. K. Becker), Expert-Verlag, Essen.

Rasmussen, J. (1983) Skills, rules and knowledge; signals, signs and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, **SMC-13** (3), 257–266.

Redlich, P. (1994) Objektive und subjektive Beurteilung aktiver Vierradlenkstrategien. Thesis. RWTH Aachen University.

Riedel, A. and Arbinger, R. (1997) Subjektive und objektive Beurteilung des Fahrverhaltens von PKW. *FAT-Schriftenreihe*, 139.

Sagan, E. (2003) *Zur Beurteilung von Fahreigenschaften in fahrdynamischen Testverfahren*R*eifen, Fahrwerk, Fahrbahn 2003*.

Scharpe, B., Prokop, G., Golloch, D., and Schick, B. (2012) Design in the Loop—bridging the gap between subjective perception and objective evaluation to achieve desired handling characteristics. C*hassis.tech plus 2010*, Munich.

Schick, B. and Bunz, D. (2002) Fahrzeuggierstabilität beim Kurvenbremsen aus hohen Geschwindigkeiten. *brems.tech 2002*, Munich.

Schimmel, C. and Heißing, B. (2009) Fahrerbasierte Objektivierung subjektiver Fahreindrücke in *Subjektive Fahreindrücke sichtbar machen IV: Korrelation zwischen objektiver Messung und subjektiver Beurteilung in der Fahrzeugentwicklung* (ed. K. Becker), Expert-Verlag, Essen.

Schimmel, C. (2010) Entwicklung eines fahrerbasierten Werkzeugs zur Objektivierung subjektiver Fahreindrücke. Thesis. TU Munich.

Schweers, T.F. and Röth, H. (1995) *Entwicklung eines objektiven Testverfahrens für Pkw mit Antriebsschlupf-Regelung Reifen, Fahrwerk, Fahrbahn 1995*, VDI-Verlag, Düsseldorf.

Seyed-Ghaemi, A. (2005) *Bewertungskriterien zur objektiven Festlegung des Wankverhaltens aus Fahrdynamikmessungen*, Thesis, Landshut.

Wagner, A. (2003) Ein Verfahren zur Vorhersage und Bewertung der Fahrerreaktion bei Seitenwind. Thesis. Universität Stuttgart.

Wallentowitz, H. (1979) Fahrer-Fahrzeug-Seitenwind. Thesis. TU Braunschweig.

Winner, H., Barthenheier, T., Fecher, N., and Luh, S. (2003) Fahrversuche mit Probanden zur Funktionsbewertung von aktuellen und zukünftigen Fahrerassistenzsystemen. *Fortschritt-Berichte VDI Reihe*, **12**, 557.

Wolf, H.J. (2008) Untersuchung des Lenkgefühls unter besonderer Berücksichtigung ergonomischer Erkenntnisse und Methoden. Thesis. TU Munich.

Zomotor, A., Braess, H.-H., and Rönitz, R. (1997) Verfahren und Kriterien zur Bewertung des Fahrverhaltens von Personenkraftwagen—Ein Rückblick auf die letzten 20 Jahre. *ATZ Automobiltechnische Zeitschrift*, (12), 1998 – 03.

Zong, C., Guo, K., and Hsin, G. (2000) Research on closed-loop comprehensive evaluation method of vehicle handling and stability. SAE Paper 2000-01-0694.

Zschocke, A.K. (2009) Ein Beitrag zur objektiven und subjektiven Evaluierung des Lenkkomforts von Kraftfahrzeugen. *IPEK Forschungsberichte Band 34*. Thesis. Universität Karlsruhe.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto021
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The Potential for Handling Improvements by Global Chassis Control

Thomas Raste and Peter E. Rieth
Continental Corporation, Frankfurt, Germany

1 Introduction 1
2 Active Systems 1
3 Vehicle Dynamics Model 5
4 Handling Improvements 6
5 Generic Motion Control Architecture 8
6 Future Trends of Handling Control 12
7 Conclusion 13
References 14

1 INTRODUCTION

From a driver's perspective, controlling of a vehicle means controlling the speed and the path curvature (Figure 1). In exceptional circumstances, for example, in emergency evading situations, also the orientation of the vehicle has to be controlled. In a narrower sense, vehicle handling refers to vehicle dynamics such as cornering and swerving and includes the vehicle stability.

In day-to-day use, speed control by braking or accelerating is largely decoupled from handling, because the steering wheel is turned slowly to steer the car. When driving on race tracks or in emergency evading situations, the steering rate reaches very high values. Skilled drivers are able to manage steering rates of 1000 deg/s or more. In those situations, the capability to follow a path is strongly depending on the speed. The most limiting factor is the tire–road friction, which restricts the lateral acceleration depending on the road conditions. Equation 1 shows how lateral (centripetal) acceleration a_y, vehicle speed v, radius R of the path, yaw rate ω, and sideslip ("spin") angle rate $\dot{\beta}$ are related

$$a_y = \frac{v_2}{R}, \quad R = \frac{v}{\omega + \dot{\beta}} \tag{1}$$

2 ACTIVE SYSTEMS

Active systems, which mean systems with auxiliary power and electronic control, are well suited to improve the handling of the vehicle. Especially, active braking has been established as the most effective active safety system. Electronic stability control (ESC) is now becoming mandatory in many countries all over the world. However, also active systems primarily designed for comfort are able to contribute to a better handling of the vehicle (Figure 2).

2.1 Requirements for active systems

Passive components affecting the handling of the vehicle need targets for the design in every stage of the development process (Figure 3). The feedback of extensive evaluations is used "offline" to tune the passive parts. Active systems lay a foundation for a new perspective in this development process. They offer the opportunity to determine their effectiveness "online" during driving based on the design targets. The active systems, therefore, need to be as generic as possible to minimize additional costs when setting up a new configuration (Andreasson, 2007).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5



Figure 1. The driver is controlling the vehicle's speed, path, and orientation. (Reproduced by permission of Continental.)

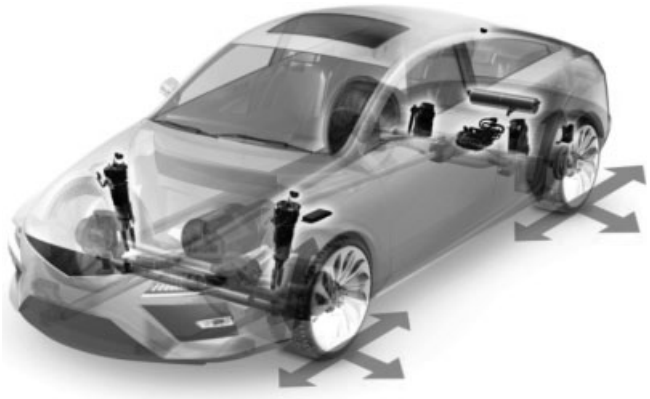


Figure 2. Vehicle with electronic air suspension and electronic adjustable damper. (Reproduced by permission of Continental.)

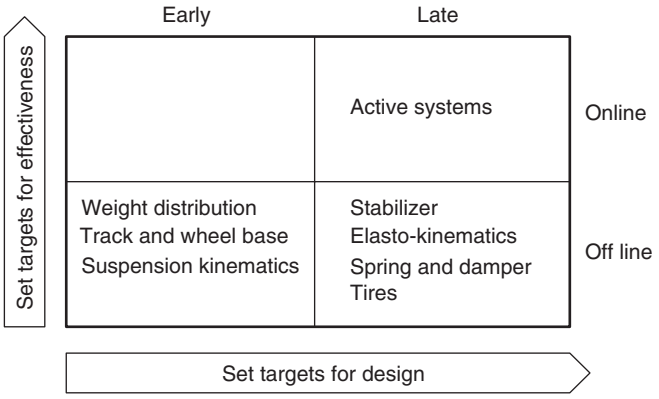


Figure 3. Target setting in the vehicle development process for passive components affecting vehicle handling in comparison to active systems. (Reproduced by permission of Continental.)

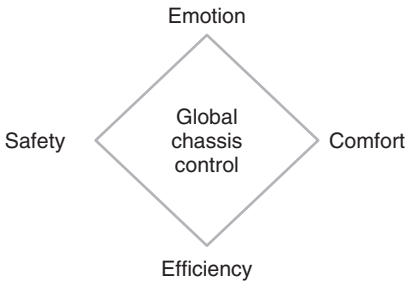


Figure 4. Trade-offs for global chassis control. (Reproduced by permission of Continental.)

The requirements for global chassis control are not unique. Figure 4 illustrates that there is usually a trade-off among emotion, safety, comfort, and efficiency. While safety will never be compromised, the remaining aspects are variable in positioning. There is a trend toward individualization and personalization of vehicle handling functions. The driver can select from different profiles with the push of a button. End consumers can thus buy just one vehicle and yet have the experience of driving different vehicle types. Above and beyond this, for example, for hybrid vehicles to come, a smooth blending of the friction brake and the generator brake for decelerating allows efficient driving with lower CO_2 emissions on a day-to-day basis (Bauer, Raste, and Rieth, 2007).

Active systems are safety-relevant components, which comprise the risk of malfunction. To minimize the risk, the development process has to be according to the safety standard ISO 26262 "Road vehicles—Functional safety." The criticality associated with a function of the system is the result of a hazard analysis and risk assessment and is classified by the Automotive Safety Integrity Level (ASIL). The classification reaches either from QM (quality management, not safety relevant) or from ASIL A (lowest level) to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Safety requirements and ASIL of typical active systems.

Active System	Safety Requirement	Typical ASIL
ESC	Avoids dangerous false brake intervention	ASIL D
AFS	Avoids dangerous false steer angle	ASIL D

Reproduced by permission of Continental.

ASIL D (highest level). The ASIL is then inherited by the software and hardware elements that realize the function and defines the safety requirements that must be fulfilled during concept phase, product development, and production and operation of the system. Typical safety requirements and ASIL classification can be found in Table 1.

2.2 Portfolio of active systems

In Figure 5, a portfolio of currently available active systems and their effectiveness in the regions of normal driving and at the friction limit is shown. The effectiveness of the individual standalone systems can be extended significantly by networking with other active systems or surrounding sensor systems (Raste, Bauer, and Rieth, 2008). Currently, the following topics are being further developed:

- specification of the areas in which the vehicle dynamics should be defined by global chassis control;
- composition of the best active system portfolio for a specific vehicle or a family of vehicles;
- partitioning of control functions on a certain electronics architecture with the need for managing complexity.

The path to a consistent, cross-vendor coordination approach for chassis control systems is still far away. However, there is consensus on the objectives: under normal operating range, the controller ensures maximum comfort and driving pleasure. In this case, the vehicle manufacturer determines all degrees of freedom for the individual setting of the vehicle character. In the friction limit range, all available actuators in the system are included in a coordinated manner to reach one target: to support the driver optimally for accident avoidance.

2.3 Potential of active systems

Figure 6 illustrates the contact patch tire forces and velocities of a single wheel during driving and how the wheel contributes to the total yaw moment of the vehicle. The wheel is steered by the angle δ and the actual direction of travel with velocity v_R defines the sideslip angle α. The resultant horizontal force F_R points toward the opposite direction of the contact patch sliding velocity v_G. The sliding velocity components v_{Gx}, v_{Gy}, each related to the longitudinal velocity v_{Rx} of the wheel center, define the tire

Effect plane	Active system	Normal driving range				Friction limit range		
		Ride comfort (z,θ,φ)	Agility (y,ψ)	Operational comfort	Ride safety (y,ψ)	Stability (x,y,ψ,φ)	Stopping distance	Traction
Horizontal	ESC electronic stability control		+	+	+	o	o	o
	ATV active torque vectoring		o	o	+	+		o
	ARK active rear axle kinematics		o	o	+	+	+	
	AFS active front steering		o	o	+	+	+	
	EPS electric power steering			o	+	+	+	
Vertical	EAS electronic air suspension	o		o		+		
	ARS active roll stabilizer	o	o			+		
	EAD electronic adjustable damper	o	o			+	+	+
	ABC active body control	o	o		+	+	+	+

Figure 5. Portfolio of active systems and their effectiveness. The symbol "o" denotes the main improvements of the active system when acting standalone and "+" shows the potential improvements by networking with other active systems or surrounding sensor systems. (Reproduced by permission of Continental.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

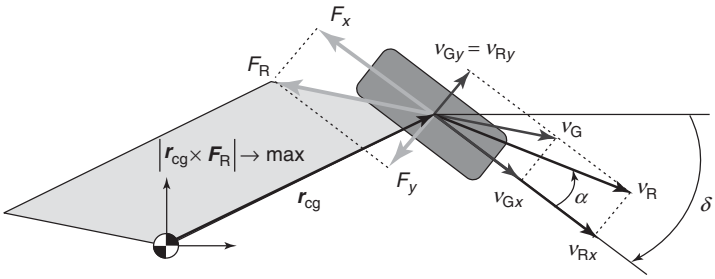


Figure 6. Tire forces, wheel and contact patch velocities, and contribution of the wheel to the yaw moment around the center of gravity of the vehicle. (Reproduced by permission of Continental.)

slip and sideslip angles, which, on the other hand, determine the magnitude of the forces F_x and F_y.

The magnitude of F_R is limited by the friction circle (Kamm's circle). The radius of the friction circle is determined by the product of the tire–road friction coefficient μ and the vertical tire force F_z. The portion of the yaw moment generated by each individual wheel is determined by the scalar product of the total tire force vector F_R and the position vector r_{cg}, denoting the distance between the vehicle center of gravity and the wheel center. The potential of an active system to maximize the scalar product is to increase the resulting area

- by aligning position and force vector orthogonally via steer angle (active rear axle kinematics (ARK), AFS, and electric power steering (EPS));
- by increasing the magnitude of the longitudinal force vector by wheel individual brake or propulsion intervention (ESC and active torque vectoring (ATV));
- by decreasing the magnitude of the lateral force vector by wheel load distribution (EAS, active rear wheel steering (ARS), EAD, and active body control (ABC)).

Figure 7 shows the potential of active brake, steering, and suspension systems selected from Figure 5 to generate additional yaw moments when activated during steady-state driving with constant radius (Schiebahn, Zegelaar, and Hofmann, 2007). It is evident that in the friction limit range, ESC has the highest potential to stabilize an oversteering vehicle. The active steering systems are highly effective to reduce lateral forces within the friction limit range. In case of AFS, this leads to a high turning-out yaw moment, whereas, in the case of ARK, this leads to considerable turning-in yaw moment. Both steering systems show high potential with opposite effectiveness within the normal driving range but only poor authority to increase lateral forces within the friction limit range. The potential of the active suspension system ARS depends as a first approximation on lateral acceleration.

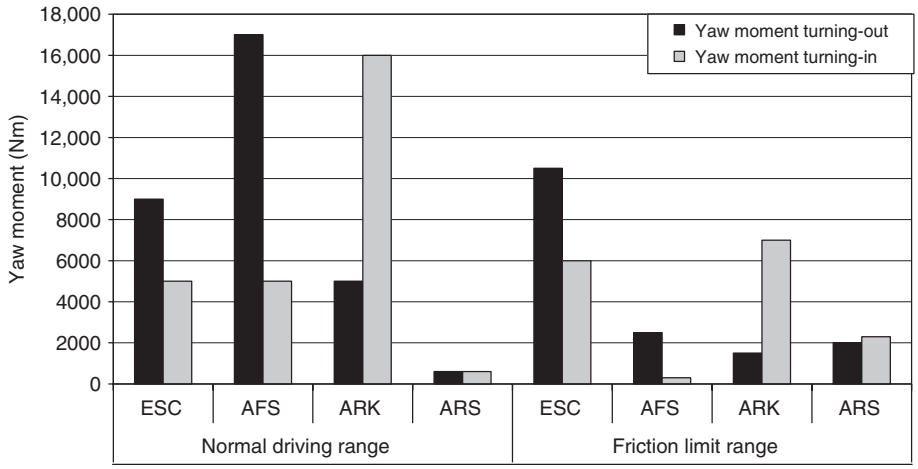


Figure 7. Potential of active brake, steering and suspension systems to generate additional yaw moment to force a vehicle in or out a curve, respectively, when driving in steady state with constant radius. (Reproduced by permission of Continental.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3 VEHICLE DYNAMICS MODEL

In this section, a mathematical model describing the lateral motion of the vehicle is defined. The model is only valid for lateral accelerations below $0.4g$ and is illustrated in Figure 8.

3.1 Equations of motion

To derive the equations of motion, it is assumed that the center of gravity is on ground level and the steering angle δ_F on front axle and δ_R rear axle is small. The equations of motion of the lateral and yaw motions are given by

$$mv(\dot{\beta} + \omega) = F_{yF} + F_{yR} \tag{2}$$

$$J_z\dot{\omega} = l_F F_{yF} - l_R F_{yR} - \frac{b_F}{2}F_{xFL} + \frac{b_F}{2}F_{xFR} - \frac{b_R}{2}F_{xRL} + \frac{b_R}{2}F_{xRR} \tag{3}$$

where F_{yF} and F_{yR} are the combined front and rear lateral tire forces. The model parameters are the vehicle mass m, the moment of inertia around the z-axis J_z, the distances l_F and l_R from the front and rear axles to the center of gravity, and the front and rear track widths b_F and b_R. The lateral tire forces are assumed to be a linear function of the sideslip angles of the wheels with the effective cornering stiffness C_F and C_R

$$F_{yF} \approx -C_F\alpha_F, \text{ where } \alpha_F = -\left(\delta_F - \beta - \frac{l_F}{v}\omega\right) \tag{4}$$

$$F_{yR} \approx -C_R\alpha_R, \text{ where } \alpha_R = -\left(\delta_R - \beta + \frac{l_R}{v}\omega\right) \tag{5}$$

3.2 State-space equation

The state-space equation establishes a relationship between the system's current state and its input, and the future state of the system. The state-space form of the vehicle model is linear and time invariant if the vehicle speed v is considered to be a constant parameter. The system state vector x includes the sideslip angle β and the yaw rate ω. The input vector u contains the two steer angles and the four longitudinal forces illustrated in Figure 8. In addition to these six inputs, two more inputs are introduced, a (virtual) lateral force F_y and a (virtual) yaw moment M_z. These additional inputs are used as virtual control commands for controllability analysis in Section 5.2. The state-space equation of system (A and B) can be derived from Equations 2–5 and is written as

$$\dot{x} = Ax + Bu \tag{6}$$

$$Ax = \begin{bmatrix} -\frac{C_F + C_R}{mv} & -\frac{l_F C_F - l_R C_R}{mv^2} - 1 \\ -\frac{l_F C_F - l_R C_R}{J_z} & -\frac{l_F^2 C_F + l_R^2 C_R}{J_z v} \end{bmatrix} \begin{bmatrix} \beta \\ \omega \end{bmatrix} \tag{7}$$

$$Bu = \begin{bmatrix} \frac{C_F}{mv} & \frac{C_R}{mv} & 0 & 0 & 0 & 0 & \frac{1}{mv} & 0 \\ \frac{l_F C_F}{J_z} & -\frac{l_R C_R}{J_z} & -\frac{b_F}{2J_z} & \frac{b_F}{2J_z} & -\frac{b_R}{2J_z} & \frac{b_R}{2J_z} & 0 & \frac{1}{J_z} \end{bmatrix} \times \begin{bmatrix} \delta_F \\ \delta_R \\ F_{xFL} \\ F_{xFR} \\ F_{xRL} \\ F_{xRR} \\ F_y \\ M_z \end{bmatrix} \tag{8}$$

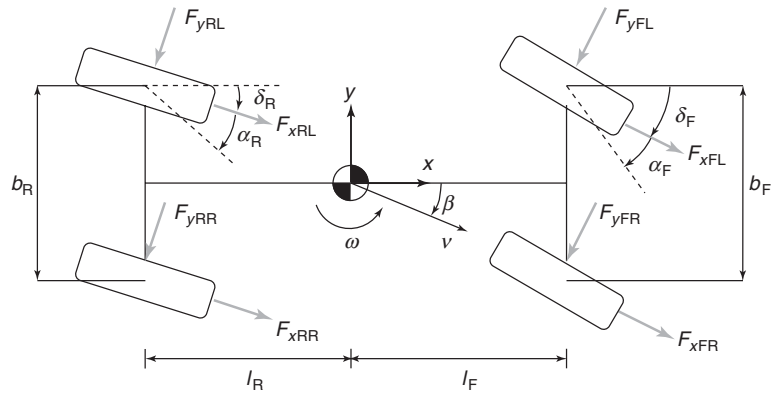


Figure 8. Vehicle model with the tire forces F_x and F_y, the vehicle speed v, the yaw rate ω, and the vehicle sideslip angle β. (Reproduced by permission of Continental.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The Equations 6–8 can be simplified for stationary operation with $\dot{x} = 0$ and the steer angle inputs only ($F_x = 0$, $F_y = 0$, $M_z = 0$). Under the assumption that the system matrix A is regular (a necessary condition is that $v > 0$), the stationary-state variables (β, ω) can be determined from Equation 6 by inverting A and introducing the parameters wheel base l and understeer gradient K_{us}:

$$x = -A^{-1}Bu \tag{9}$$

$$\beta = K_F\delta_F + K_R\delta_R, \text{ where } K_F = \frac{l_R\left(1 - \frac{ml_Fv^2}{C_Rl_Rl}\right)}{l + K_{us}v^2}$$

$$\text{and } K_R = \frac{l_F\left(1 + \frac{ml_Rv^2}{C_Fl_Fl}\right)}{l + K_{us}v^2} \tag{10}$$

$$\omega = K_\omega(\delta_F - \delta_R), \text{ where } K_\omega = \frac{v}{l + K_{us}v^2}$$

$$\text{and } K_{us} = \frac{m}{l}\left(\frac{l_R}{C_F} - \frac{l_F}{C_R}\right) \tag{11}$$

3.3 Vehicle parameter determination

The moment of inertia J_z can be approximated with the ease to measure total vehicle length L, the wheel base l, and the vehicle mass m. The distances l_F and l_R from the center of gravity to the axles are calculated using the weighted rear axle vehicle load m_R (Equation 12).

$$J_z = 0.1269 \cdot m \cdot l \cdot L, \; l_F = \frac{m_R}{m} \cdot l, \; l_R = l - l_F \tag{12}$$

The tire parameters are determined from a stationary circle vehicle driving test with constant radius R and zero steer angle at rear axle ($\delta_R = 0$). From these tests, two gradients with reference to the lateral acceleration a_y have to be derived from measurement plots. The required gradients are the sideslip angle gradient $d\beta/da_y$ and the steering angle gradient $d\delta_F/da_y$ ($=K_{us}$). The steering reduction ratio i_s between the front steer road wheel angle and the driver's steering wheel angle δ_H is assumed to be constant, that is, $\delta_H = \delta_F i_s$. The steering reduction ratio and the cornering stiffness can be calculated consecutively from

$$i_s = \frac{\delta_H.R}{l}, \; C_R = -\frac{ml_F}{\frac{d\beta}{da_y}l}, \text{ and } C_F = \frac{ml_RC_R}{K_{us}.lC_R + ml_F} \tag{13}$$

A set of vehicle parameters for a typical sedan car can be found in Table 2.

Table 2. Vehicle model parameters for a generic sedan car.

Symbol	Value	Unit	Description
m	1770	kg	Vehicle mass
J_z	3140	kg m^2	Vehicle yaw moment of inertia
L	4.841	m	Vehicle length
l_F	1.575	m	Distance from center of gravity to front axle
l_R	1.313	m	Distance from center of gravity to rear axle
b_F	1.562	m	Track width of front axle
b_R	1.590	m	Track width of rear axle
C_F	94672	N rad^{-1}	Effective cornering stiffness of front axle
C_R	160880	N rad^{-1}	Effective cornering stiffness of rear axle
i_s	17.6	—	Steering reduction ratio

Reproduced by permission of Continental.

4 HANDLING IMPROVEMENTS

The operational range for handling improvements is separated into two distinct areas, separated by the yaw rate at the friction limit range as illustrated in Figure 9. The "deep slip" range above the critical friction limit yaw rate is the application range of ESC.

Considering the basic idea that "the steering wheel is a yaw rate demand" (Blundell and Harty, 2004), improvements of handling mean that all the following are true:

- The path curvature can quickly and easily be altered with open-loop commands.
- The path curvature can be adjusted by the driver with only small closed-loop content.
- The vehicle response to steering input is predictable for the driver.

4.1 Handling improvements during normal driving

The alteration of the path curvature can easily be achieved by increasing the yaw gain, such that the driver steering input is small. This strategy is only applicable up to a medium speed. The yaw rate's normal driving range decreases significantly with vehicle speed, because the available tire–road friction is saturated at high speed quickly when the steering wheel angle input is too high (Figure 9). For normal tires on dry roads, this dangerous saturation level is reached at a lateral acceleration of 1g. The acceleration levels of 0.7, 0.3, and 0.1g correspond to wet, snowy, and icy roads. The strategy at high speed, therefore, must be to decrease the steady-state yaw gain at higher speed. From Equation 14, it can be seen that the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

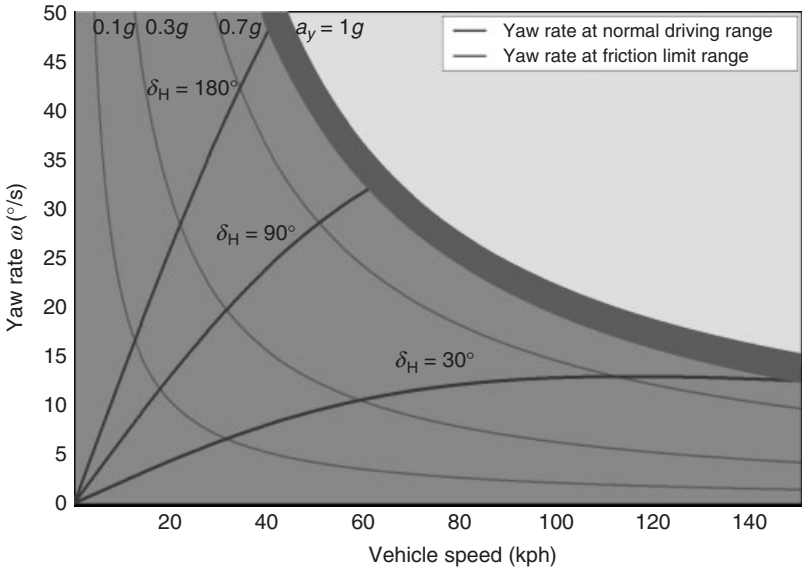


Figure 9. Operational range for handling improvements by global chassis control. (Reproduced by permission of Continental.)

yaw gain can be adjusted either by variation of the steering reduction ratio i_s or using a rear wheel steering. BMW uses both means together in their integral active steering (Herold *et al*., 2008).

$$\frac{\omega}{\delta_H} = K_\omega \left(\frac{1}{i_s} - \frac{\delta_R}{\delta_H} \right) \tag{14}$$

The potential for transient handling improvement is demonstrated with a simulated step-steer maneuver. This maneuver demonstrates that altering a path curvature quickly is not in conflict with a good damping behavior, when the controller is designed carefully. The handling controller used for the demonstration of the control potential is acting on the rear axle steer angle. A feedback of the vehicle's state $x = [\beta \ \omega]^T$ with gain matrix K_x improves the closed-loop dynamics of the vehicle, and a feedforward control part K_H is applied for a vanishing steady-state rear axle steer angle (Figure 10). The control law is given by the following equation, with parameters only valid for the given maneuver speed of 150 kph.

$$\delta_R = K_H \delta_H - K_x x, \text{ where } K_H = -0.0239$$
$$\text{and } K_x = (-0.0764 - 0.0896) \tag{15}$$

The controller increases the response of the vehicle by increasing the yaw moment of approximately 45%. This significant improvement can be seen from the yaw acceleration in Figure 10 and is due to the out-of-phase (= negative) steering of the rear axle at the beginning of the maneuver. This steering input leads to an initially negative lateral force on the rear axle and, therefore, a yaw moment contribution, which enhances the turning-in yaw moment compared to the vehicle with inactive controller. Further advantage of the controller is the reduction of overshoot in yaw rate, sideslip angle, and lateral acceleration. This considerably improved damping behavior provides for an increased safety margin toward the friction limit range. Furthermore, the decreased delay between the steering wheel and the vehicle sideslip angle rate leads to an improved perception of agility by the driver.

4.2 Handling improvements during friction limit driving

At the limit of friction, where safety becomes relevant, the handling controller determines how the vehicle remains stable. All available actuators, that is, those listed in Figure 5, are incorporated and coordinated to reach this goal. The active chassis gives the driver optimal support for avoiding accidents. In the region beyond the limit of friction, the main task of the control system is to prevent the car from heavily skidding such that the car remains on track.

During normal driving, car drivers usually expect a linear yaw response of the vehicle with small phase lag. Most drivers have no experience of loss of linearity caused by saturation of tire forces. If saturation happens at the rear axle, the sideslip angle will increase quickly and, therefore, causes a hazardous driving problem for many drivers. The primary task of the control system should be to keep the vehicle sideslip angle small. An average driver feels uncomfortable when the magnitude of the sideslip

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

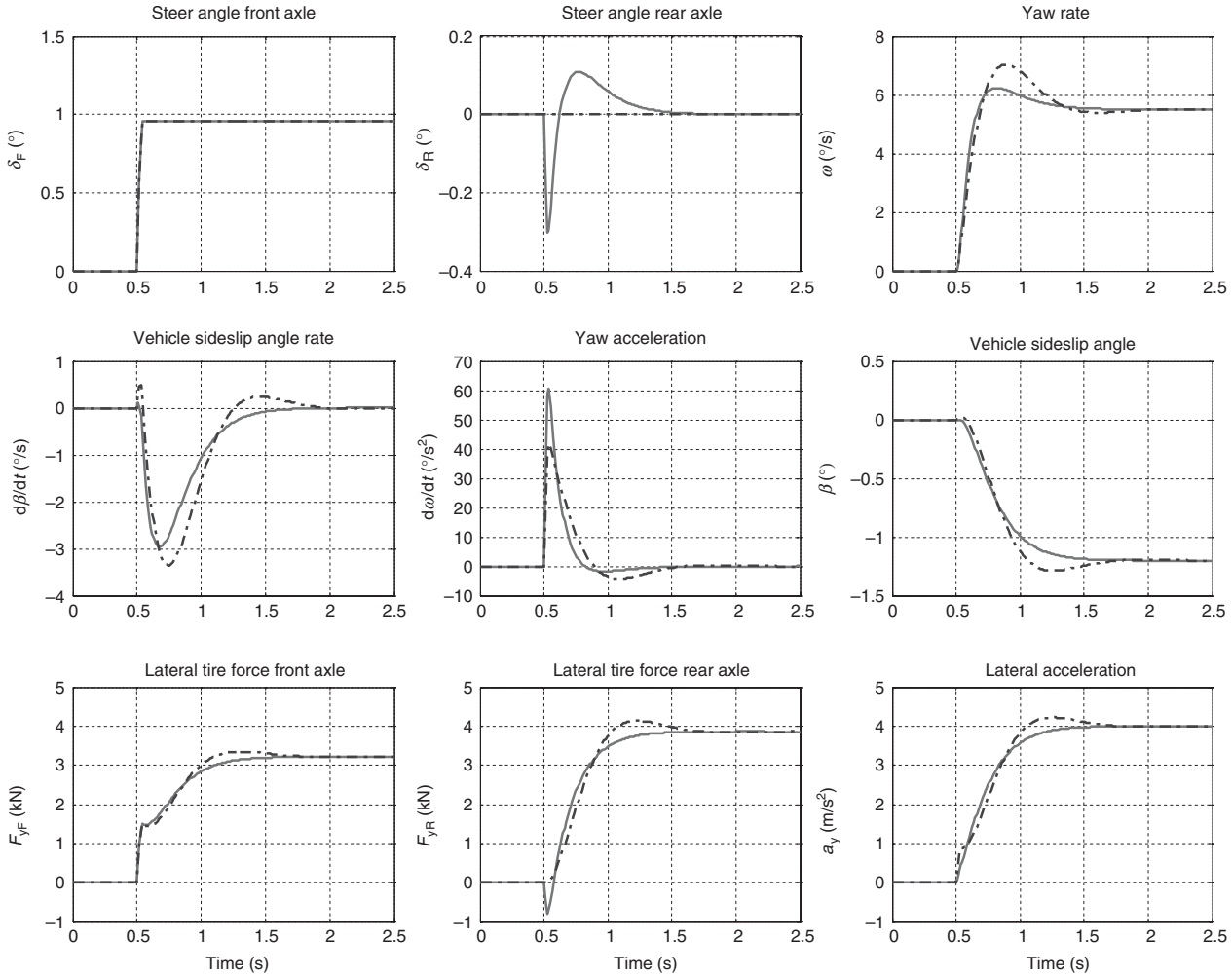


Figure 10. Vehicle states and inputs for a simulated step-steer maneuver with vehicle model Equations 6–8 reaching a steady-state lateral acceleration of $0.4g$ with a maximum steering wheel angle rate of $\delta_H = 500$ deg/s at a vehicle speed of $v = 150$ kph. The light gray solid line and the dashed dark gray line correspond to the controller active and the controller inactive, respectively. (Reproduced by permission of Continental.)

angle exceeds 3°. The state-of-the-art ESC systems limit the sideslip angle indirectly. ESC uses a reference yaw rate limited by the actual acceleration to account for the tire saturation. Additionally, the rate of change of sideslip angle is calculated and also limited. Figure 11 shows the results of a double-lane change vehicle test performed on a low friction surface. With the global chassis control (GCC), the vehicle response is stable and predictable to the driver at that particular speed level. ESC standalone will achieve a similar behavior, but at a much lower speed level. In the data plots of Figure 11, the scaling factor $k = 7$ is used for visualization purposes only. The scaling factor has been introduced to "normalize" the yaw rate for easy comparison of steering angle stimulus and yaw rate response. The unusual delay of the yaw rate response in the ESC standalone case becomes very obvious.

5 GENERIC MOTION CONTROL ARCHITECTURE

5.1 Overall control system

The fundamentals of the integration and coordination of the active systems into a hierarchically structured overall control system are shown in Figure 12. The first subtask in the overall control system is to determine the driver's intention. For that purpose, sensors on brake, steering, and accelerator pedal are interpreted to derive appropriate

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

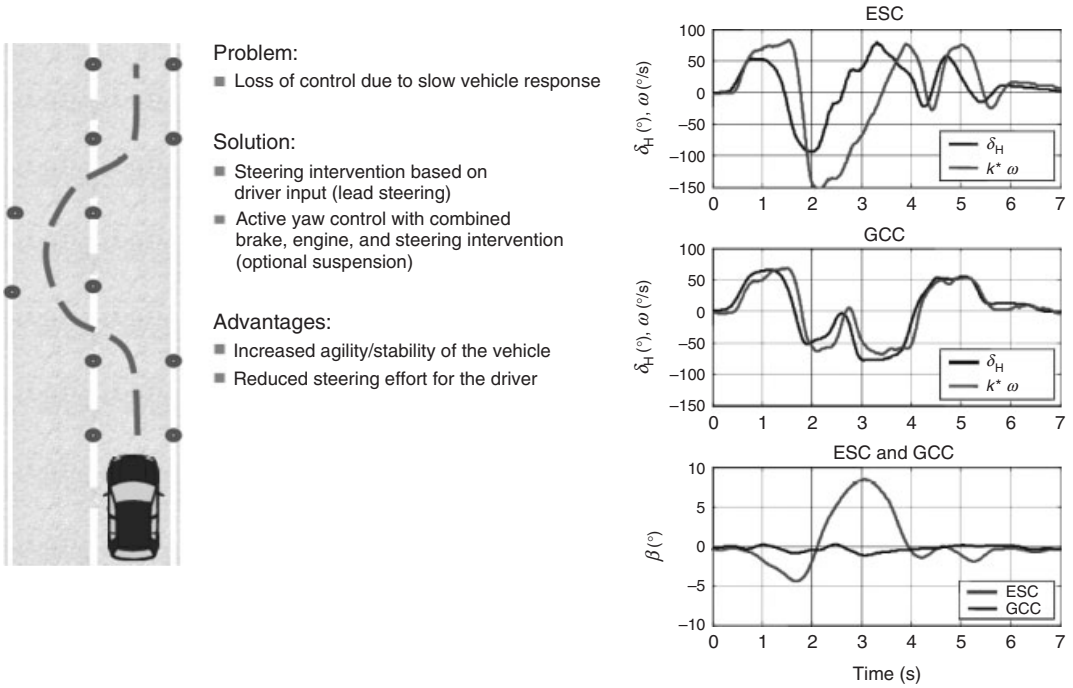


Figure 11. Comparison of vehicle test data from double-lane change (ISO 3888–1) at 70 kph on snow with conventional ESC versus GCC with integrated control of steering at front and rear axles. The light gray solid line and the dark gray line correspond to the GCC controller active and the conventional ESC active, respectively. (Reproduced by permission of Continental.)

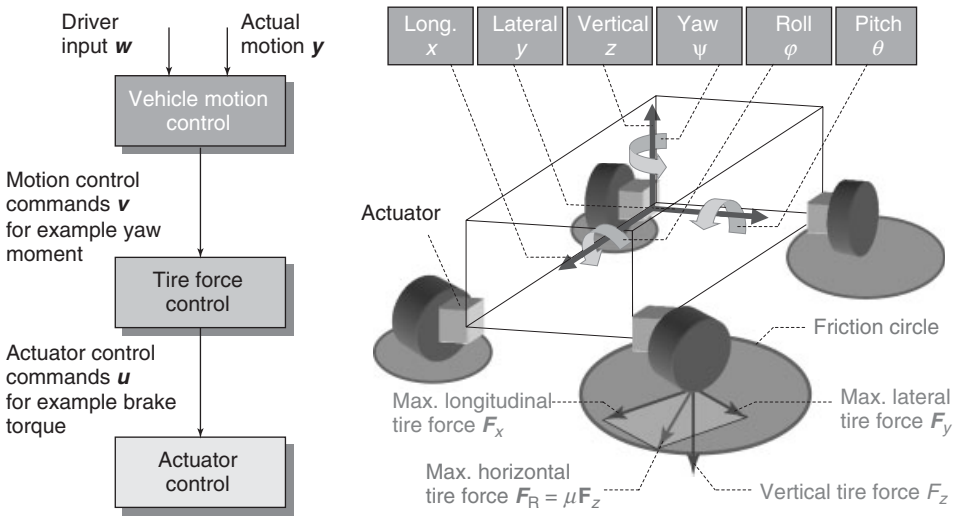


Figure 12. Overall control system. (Reproduced by permission of Continental.)

reference signals. In a second subtask, the driver's intention is compared to the actual vehicle motion, which is measured by inertial and speed sensors. If there are deviations, they are adjusted by calculating target forces and moments to change the actual vehicle's translational and rotational motions according to the driver's intention. The task of the tire force control is to distribute the motion control commands onto the individual wheels in order to change the tire forces in the contact patch area. The tire forces are adjusted by electromechanical or electrohydraulical actuators. A general limitation for the maximum achievable horizontal tire force is the friction circle (Kamm's circle),

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

which depends on the tire–road friction and the load at each wheel.

Tire force control is the task to distribute the motion control commands among the tire forces at each wheel. Mathematically, this is the task to solve an underdetermined, typically constraint system of equations. The ambiguity of the actuator effects is characteristic for vehicles with multiple active systems onboard. For tire force control, a control allocation approach is generally useful, when different combinations of actuator commands can produce the same motion result. When the number of actuators available exceeds the number of degrees of freedom being controlled, the vehicle is called *over-actuated*. A vehicle equipped with ESC falls under this category, because four individual brake actuators control three horizontal degrees of freedom. Control allocation of over-actuated vehicles involves generating an optimal set of actuator control commands while minimizing the control effort and complying with the position and rate constraints of the actuator.

5.2 Selection of actuator configurations

It is a challenging and complex question of how to select an optimal set of actuators. For handling control, a possible answer can be achieved by analyzing the lateral full state (β, ω) controllability property of the vehicle (Raste *et al.*, 2010). The analysis is based on the state-space Equations 6–8. To simplify the analysis, not all longitudinal forces but only the differential forces $\Delta F_{x\mathrm{F}} = F_{x\mathrm{FL}} - F_{x\mathrm{FR}}$ and $\Delta F_{x\mathrm{R}} = F_{x\mathrm{RL}} - F_{x\mathrm{RR}}$ are considered. The analysis requires the controllability Gramian matrix W_c, which can be found as the solution to the Lyapunov matrix Equation 16 (Kailath, 1980).

$$AW_\mathrm{c} + W_\mathrm{c}A^\mathrm{T} = -BB^\mathrm{T} \tag{16}$$

The system state (β, ω) is controllable if the matrix W_c has full rank. If the matrix W_c has at least one eigenvalue equal to 0, then it cannot have full rank and, therefore, the system state is not controllable. Figure 13 shows the normalized smallest eigenvalue $\lambda_{\min}$ of the controllability Gramian matrix W_c as a function of the vehicle speed. The results can be interpreted as follows:

- Controllability of the full state (β, ω) is given in the whole speed range if front and rear steer angle controls are used, which is approximately equivalent to a virtual control input (F_y, M_z) at low speeds.
- Controllability of the full state (β, ω) with longitudinal differential forces only, which is equivalent to a virtual yaw moment control M_z, is not given at low speeds.

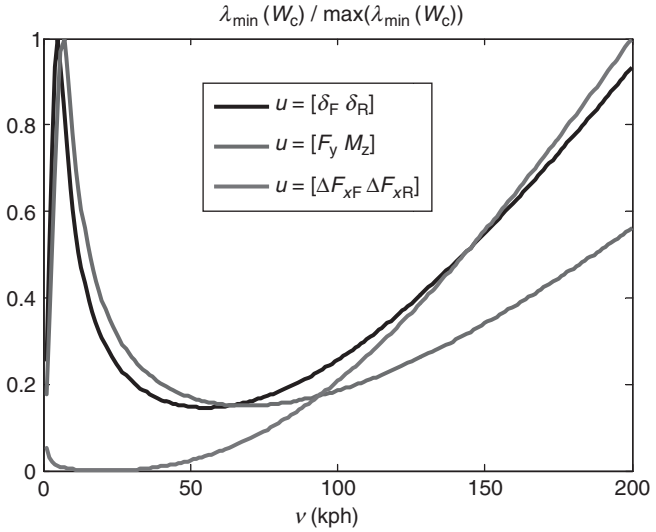


Figure 13. Controllability analysis based on the smallest eigenvalue $\lambda_{\min}$ of the controllability Gramian matrix W_c for different control input vectors. The black solid line corresponds to steer angle control input. The light gray solid line and the dark gray line correspond to virtual control input and longitudinal differential forces control input, respectively. (Reproduced by permission of Continental.)

- Simultaneous path and orientation controls, that is, full state (β, ω) control over the whole speed range, need both yaw moment actuators (e.g., ESC or ATV) and lateral force actuators (e.g., EPS, AFS, or ARK).

5.3 Compatibility of actuator configurations

The results from the controllability analysis are very useful to design the control system as generic as possible. A potential strategy to make the control compatible with various configurations of actuators is illustrated in Figure 14. The control feature A is originally designed for a rear wheel steering. It is assumed that δ_F is determined by the driver steering input. The task of the converter function is to transform the motion control commands into a target force and moment. These virtual motion control commands F_y and M_z might be used as a common basis for arbitration, for example, when requested by different control features. The control allocation procedure distributes the virtual motion control commands onto real commands for a given actuator configuration. The procedure is successful if the actuators generate a control effect as closely as possible to the virtual control demand.

To verify the feasibility of the strategy, a constraint control allocation library for MATLAB /Simulink has been

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

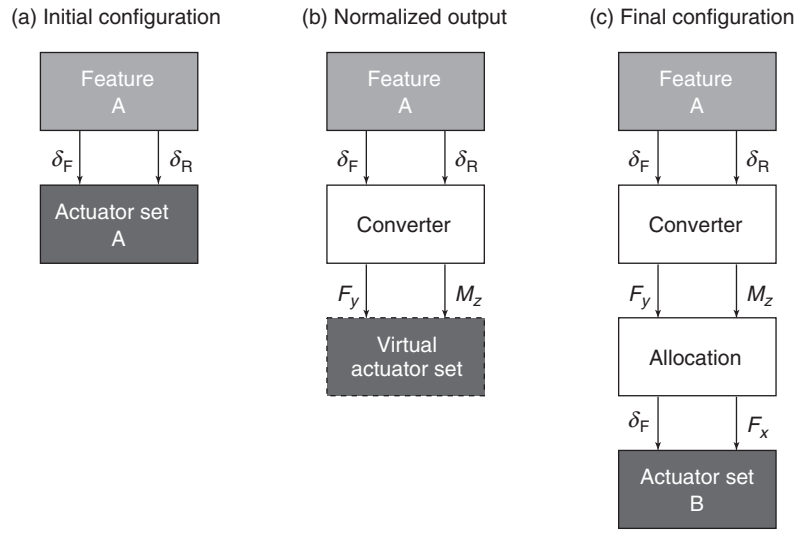


Figure 14. (a–c) Potential strategy to make a control feature compatible with various actuator configurations. (Reproduced by permission of Continental.)

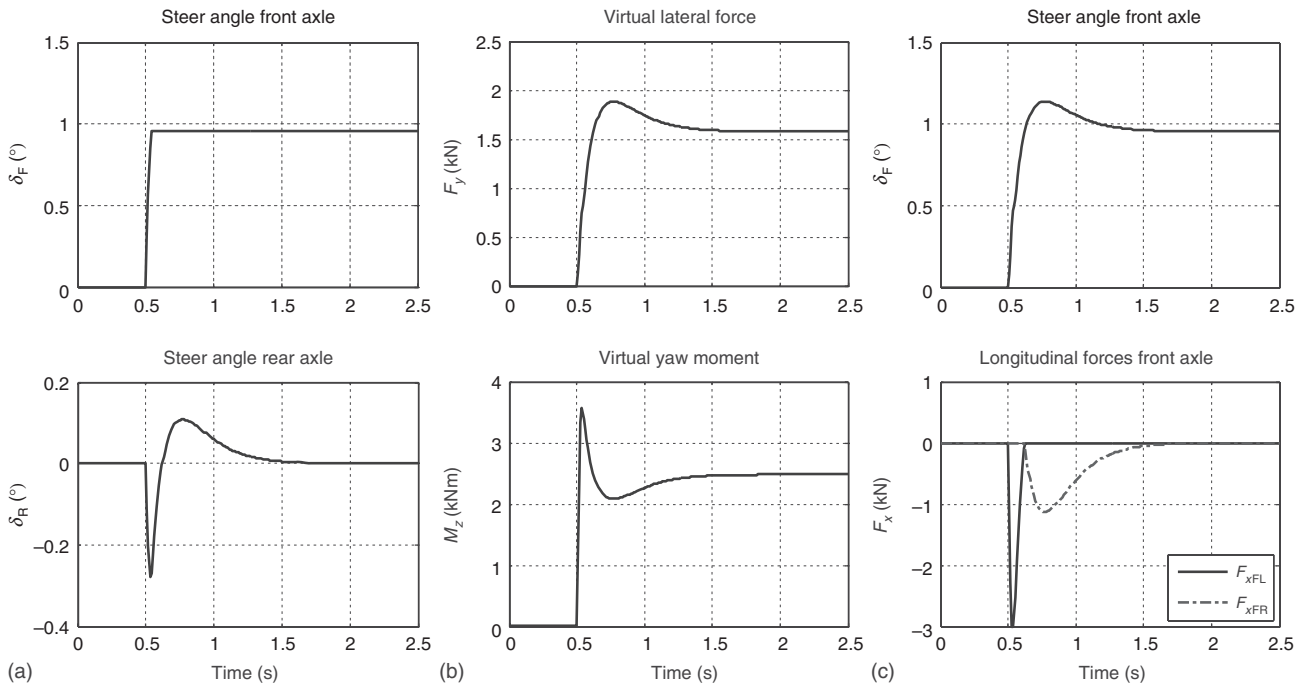


Figure 15. Simulated step-steer maneuver based on vehicle model Equations 6–8 and control feature (Equation 15). The differing input for actuator set A (a), virtual actuator set (b), and actuator set B (c) reproducing identical vehicle state behavior is shown. (Reproduced by permission of Continental.)

applied (Harkegard, 2003). Figure 15 illustrates the results with the same step-steer maneuver, which has been used in the previous section (Figure 10). It is worth to point out that the vehicle state variables β and ω show identical behavior for all three actuator sets. The feature is the transient control with the control law given by Equation 15. The steering signals are converted into the virtual motion control commands shown in Figure 15b. The constraint control allocation is able to consider the position and rate limits of each actuator. In the simulation, a constraint in the positive

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

direction of the force has been used to make the feature available for a brake-based active system. The resulting actuator commands are shown in Figure 15c. The driver input is now slightly modified. Further, simulations have to clarify if the deviation can be expected to be compensated by the driver or if the driver has to be supported by additional active systems, for example, EPS.

6 FUTURE TRENDS OF HANDLING CONTROL

In normal driving and also in hazard situations, the driver is supposed to be the master of control and, therefore, properly the source for reference signals. However, what happens if a driver does not react appropriately? In urban traffic situations, one of the main causes of accidents is driver distraction. The automotive industry has identified the great potential of active systems assisting the driver in such critical situations.

6.1 Emergency brake assist

The emergency brake assist (EBA) shown in Figure 16 intervenes if the driver is inattentive and shows no sign of having recognized the danger of an impending collision. The city version of EBA-City is active at speeds of up to 30 kph. It features an optical sensor that uses infrared beams to monitor the road space in front of the vehicle, up to a distance of about 10 m. Its electronics calculate the distance to the vehicle in front. If there is a risk of collision, EBA initially prepares the brakes, issues a warning to the driver, and then, if time starts to run out, automatically applies the brakes. If the maximum speed differential between the vehicles is no more than 15 kph, a rear-end collision with the vehicle in front can be avoided in most cases. If a collision is unavoidable, automatic emergency braking can significantly reduce the impact velocity and thus the severity of the accident.

6.2 Emergency steer assist

Automotive engineers currently develop systems assisting the driver in hazardous handling situations, when there is no time left for braking (Hartmann, Eckert, and Rieth, 2009). Figure 17 illustrates a typical use case for handling assistance. A driver quickly initiates an evading maneuver to avoid a collision with a suddenly appearing object in front. Although the conventional ESC mitigates the severity of the situation significantly, there is still potential for further improvement. The main benefit comes from surrounding sensors such as radar sensors, which are used to identify an imminent collision. In Figure 17, the object distance information is used to preset an ARK just in time for handling with optimal effectiveness.

The simulation data presented in Figure 18 provides a further insight into the potential of different active systems in combination with predictive handling control. While systems such as adaptive damper or active stabilizer have only limited handling potential, the rear-wheel steering considerably reduces vehicle spinning-out and increases yaw damping when applied early enough. This positive effect on vehicle safety is accompanied by a slightly increased driver-steering effort. The preferred choice of assistance in such a case is to stimulate the haptic channel of the driver via EPS.

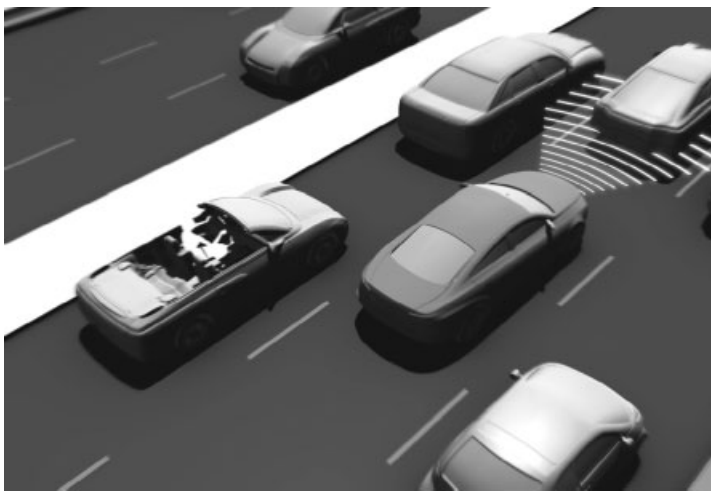


Figure 16. The emergency brake assist for urban areas is already well established. (Reproduced by permission of Continental.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 17. Animation of an emergency evading use case. The vehicle with conventional ESC is skidding off the track and the vehicle with ESC and ARK and predictive handling control remains on track. (Reproduced by permission of Continental.)

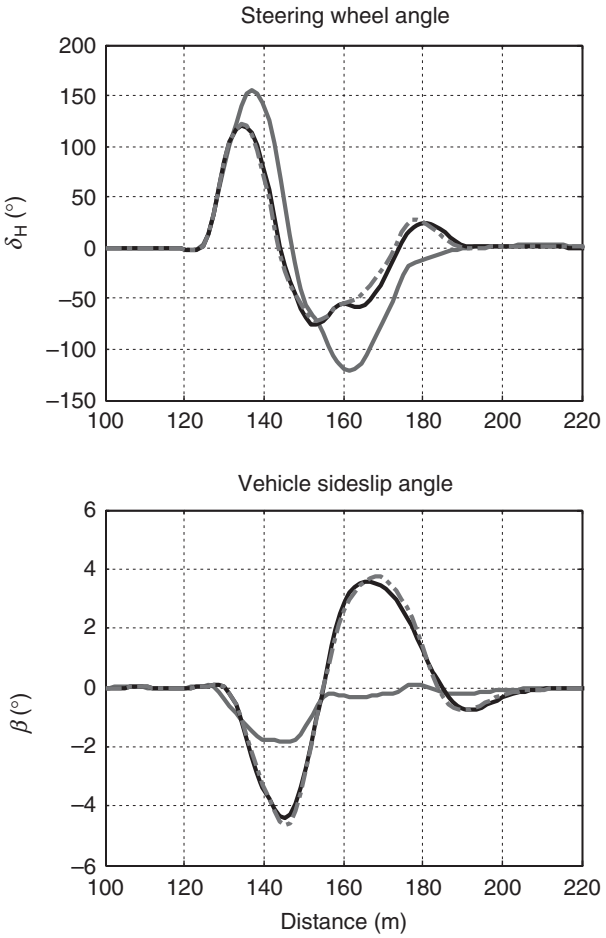


Figure 18. Emergency evading maneuver simulated with a complex nonlinear vehicle and driver model to compare the effectiveness of different active systems when applied with predictive handling control. The light gray solid line corresponds to ESC and ARK, the dark gray dashed line corresponds to ESC and EAD, and the thin black solid line corresponds to ESC and ARS. (Reproduced by permission of Continental.)

6.3 Handling during highly automated driving

Advanced driver assistance systems can offer remedies. On the one hand, they can support the driver in demanding and difficult situations and, on the other hand, develop room for freedom during monotonous driving situations, which are often accompanied by the risk of decreasing attention. Especially, the latter is a potential field of application for highly automated driving.

In the case when the driver is distracted and inattentive, an option of handling assistance could be to carry out the maneuver autonomously. Figure 19 presents a solution based on the procedure shown in Figure 14. The steering maneuver is executed autonomously without the need of steering the front wheels (e.g., when EPS is not available). The control commands are allocated, in the present example, to both friction brakes on the left side and the rear steering. The vehicle behaves in exactly the same way as it would have with steering the front wheels and the controller given by Equation 15 acting at the rear wheels.

Where does the assistance in handling situations go from here? Will drivers understand the meaning of haptic steering feedback in critical situations? What happened if they do not react at all? Autonomous vehicle braking technology has been launched over the last few years with great success. However, a rapid introduction of autonomous vehicle steering is not very likely. The legal situation and the functional safety management raise many questions at the moment. A first step toward a technically feasible solution has been presented by exploiting the opportunities given by the inherent redundancy of multiple active systems.

7 CONCLUSION

Global chassis control delivers significant benefits in normal driving and particularly, in emergency situations. The configuration and the coordinated interaction of the active systems are the key success factors for enhancing the vehicle performance. International standards such as ISO 26262 ensure quality and safety of the overall control system at the highest level. In the near future, the vehicle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

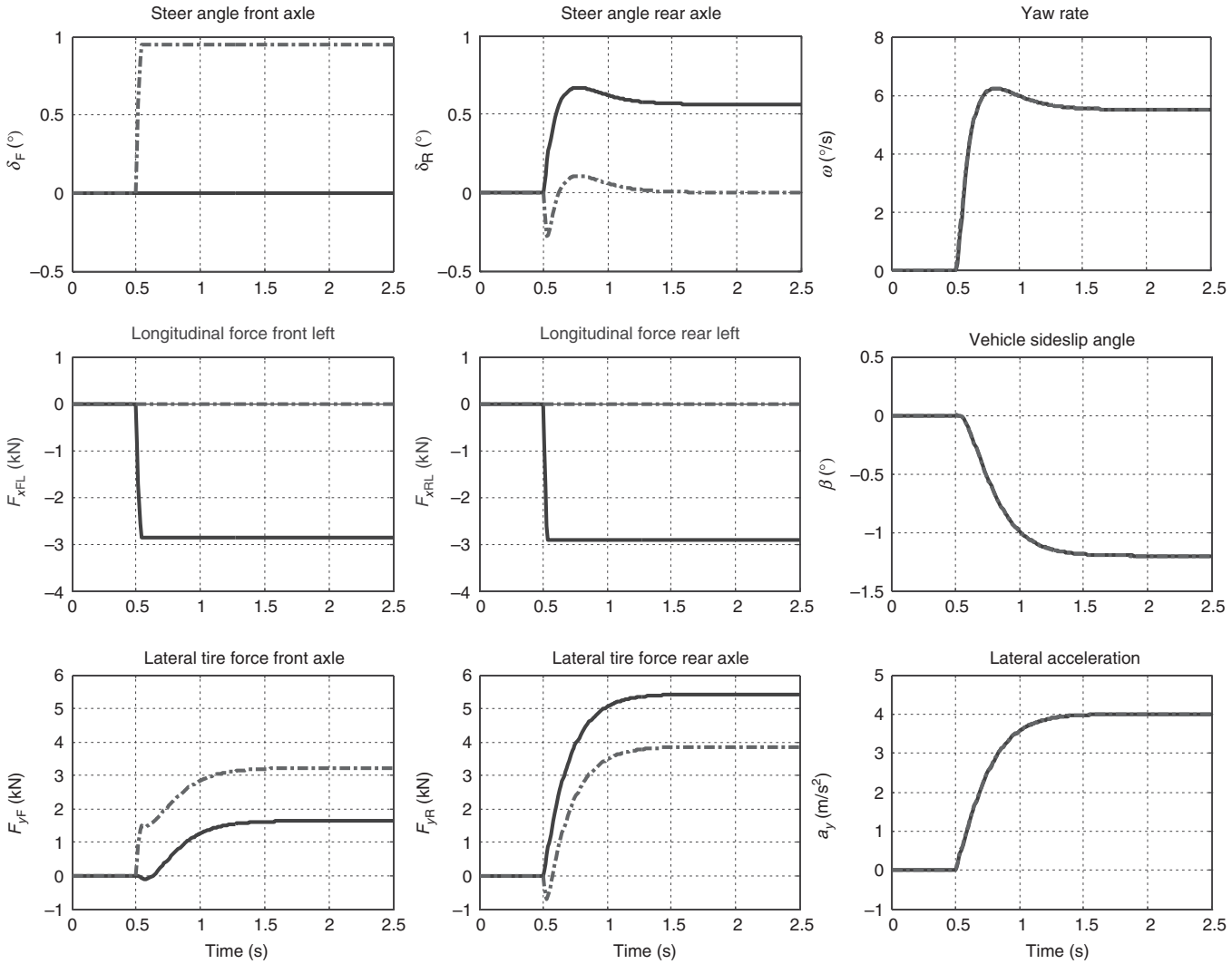


Figure 19. Allocation of control commands in order to substitute a distracted driver. Simulated step-steer maneuver based on vehicle model Equations 6–8 reproducing identical vehicle state behavior for two different input vectors. The light gray solid line corresponds to the maneuver without driver input using longitudinal forces and rear steer angle only. The dark gray dashed line corresponds to the maneuver with driver input on front steer angle and control law given by Equation 15 for the rear steer angle. (Reproduced by permission of Continental.)

is fitted with sensors for monitoring the surroundings, such that predictive motion control interventions become possible.

REFERENCES

Andreasson, J. (2007) On generic road vehicle motion modelling and control. Ph.D. Dissertation. Royal Institute of Technology, Vehicle Dynamics, Stockholm, Sweden.

Bauer, R., Raste, T., and Rieth, P.E. (2007) System integration of hybrid powertrains. *ATZelectronic*, **2** (04), 6–10.

Blundell, M. and Harty, D. (2004) *The Multibody Systems Approach to Vehicle Dynamics*, Elsevier, Burlington.

Harkegard, O. (2003) Backstepping and control allocation with applications to flight control. Ph.D. Dissertation. Department of Electrical Engineering, Linköping University, Linköping, Sweden.

Hartmann, B., Eckert, A., and Rieth, P.E. (2009) Emergency Steer Assist—Assistenzsystem für Ausweichmanöver in Notsituationen. *12. VDI-Tagung Reifen-Fahrwerk-Fahrbahn*, VDI-Berichte No. 2086, pp. 131–148.

Herold, P., Schuster, M., Thalhammer, T., *et al*. (2008) Integral Active Steering. Synthesis of Agility and Sovereignity. *FISITA World Automotive Congress*, Munich.

Kailath, T. (1980) *Linear Systems*, Prentice-Hall, Englewood Cliffs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Raste, T., Bauer, R., and Rieth, P.E. (2008) Global Chassis Control: Challenges and Benefits within the Networked Chassis. *FISITA World Automotive Congress*, Munich.

Raste, T., Kretschmann, M., Eckert, A., *et al*. (2010) Sideslip Angle based Vehicle Control System to Improve Active Safety. *FISITA World Automotive Congress*, Budapest.

Schiebahn, M., Zegelaar, P., and Hofmann, O. (2007) Theoretical Generation of Additional Yaw Torque. *11. VDI-Tagung Reifen-Fahrwerk-Fahrbahn*, VDI-Berichte No. 2014, pp. 101–119.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto022
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Automated Driving

Adrian Zlocki
Institut für Kraftfahrzeuge (ika), Aachen, Germany

1 Introduction 1
2 Requirements for Automated Driving Vehicles 1
3 Advanced Driver Assistance Systems 5
4 Infrastructure-Based Automated Driving Vehicles 9
5 Fully Automated Driving Vehicles 11
6 Summary 14
Related Articles 15
References 15

1 INTRODUCTION

First assistance functions for automobiles were introduced shortly after the automobile itself had been invented, for example, the introduction of the automatic engine starter in 1913. In addition, the idea of automated driving vehicles was already developed in the early twentieth century. One of the first steps toward this idea was the introduction of the first cruise control system in the Chrysler Imperial as the so-called auto pilot in 1958 (Rowsome, 1958). This mechanical system was able to keep a constant speed, set by the driver, and therefore was able to take over the longitudinal vehicle control on a motorway. With the development of microelectronics and the introduction of computer-controlled systems in the last quarter of the twentieth century, the possibilities to realize automated driving vehicles became available. On the basis of these possibilities, initiatives were started such as PROMETHEUS (PROgraMme for a European Traffic of Highest Efficiency and Unprecedented Safety), the American National Automated Highway System Consortium (NAHSC), the Intelligent Vehicle Initiative (IVI), and the Japanese Advanced Cruise-Assist Highway System Research Association (AHSRA) in the end of the 1980s. At present, automated driving basically can be divided into three main categories:

- Advanced driver assistance systems (ADAS): ADAS partly take over the driving task and support the driver in complex and monotonous situations.
- Infrastructure-based automated driving vehicles: These self-driving vehicles operate in a dedicated infrastructure such as dedicated lanes or closed areas automated.
- Fully automated driving vehicles: This type of vehicles are operated driverless.

2 REQUIREMENTS FOR AUTOMATED DRIVING VEHICLES

Automated driving vehicles need to fulfill technological and legal requirements. The necessary technology for automated driving vehicles needs to provide sophisticated information about the surrounding environment and the current vehicle status. This information is processed to determine the situation awareness and provide the correct action or even reaction by the actuators of the vehicle.

In contrast to conventional controlled vehicles, automated driving vehicles need to perceive the surrounding environment, process the information, and react to them fully automated (Figure 1). This reaction is targeted to the normal car operation and the supervision and avoidance of dangerous situations.

While the driver depends on his own recognition for information perception, for automated systems, vehicle and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

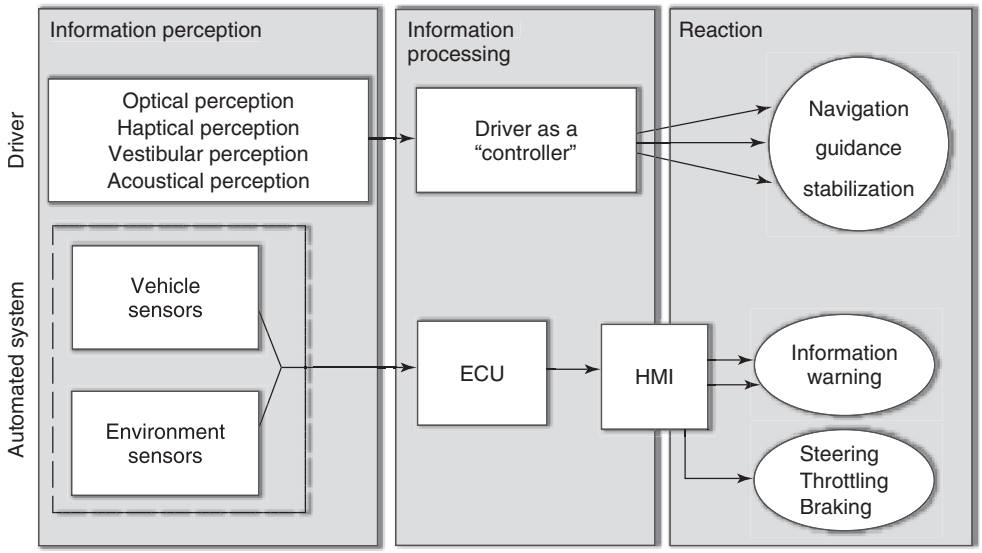


Figure 1. Vehicle control comparison between driver and automated system.

environment sensors are necessary for the perception. These sensors detect the driving condition (speed, acceleration, yaw rate, etc.) and the surrounding traffic situation (other traffic participants, objects in the driving path, current position, etc.). Information processing takes place in an electronic computer unit (ECU). The driver activates the system by means of an appropriate human–machine interface (HMI). Actuators such as the steering system, the accelerator, and the brake actuator or even the automatic gear shifter transfer the reaction of the system into changes of the vehicle's motion.

2.1 Sensor technology

Sensor technology is one of the keys for automated vehicles. The sensors need to be able to detect the vehicle's surroundings and to measure the dynamic status of the other vehicles in relation to the own vehicle. The frontal area of the vehicle is divided into a near field and a far field. The near field is the area up to 50 m around the vehicle. In this area, a wide observation angle is substantial. The far field goes up to 200 m around the vehicle. The range of the sensors and the evaluation of the relative velocities to others are important factors for the far field. Furthermore, position sensors and the communication with the infrastructure and with other traffic participants can provide additional information.

Different principles for environmental detection are available. The most important ones are radar (Radio Detection and Ranging), laser (Light Amplification by Stimulated Emission of Radiation), and image processing by camera picture analysis. Ultrasonic sensors are also necessary for very short ranges.

Radar sensors operate with electromagnetic waves, using frequencies in the centimeter or micrometer distance for object detection. Frequencies for automotive applications of radar sensor are in the 24 GHz and the 76–77 GHz bands.

The frequency-modulated continuous wave-radar (FMCW-radar) represents the most common used radar principle. The measurement of the distance to the relevant target of the FMCW-radar is based on the phase shift between the transmitted and the received signal with a periodic phase. In case the transmitter and the receiver are moved with a relative velocity, a frequency shift occurs between the transmitted and the received frequencies. This shift is due to the Doppler effect. The frequency increases in case the target is approached. The frequency difference depends on the relative velocity, which can be determined based on this effect.

Lidar sensors, which are based on laser technology, use the reflection of transmitted electromagnetic waves with lengths from 0.78 to 1 μm and thus in the infrared range, invisible to the human eye. In automotive technology, two methods are common, the transit time method and the laser-Doppler-shift method, which is based on short laser impulses being emitted by the laser. The Doppler-shift method is used to measure additionally the relative velocity. The frequency shift can be generated by the superposition of the reflected light impulse and a reference impulse of the same laser source. Another type of laser sensor is the laserscanner. Laserscanners use mechanically rotating mirrors. Owing to this, these sensors can scan a wide area

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and therefore they have high opening angles up to 270°. The own vehicle limits the complete 360° view, unless the sensor is mounted on the top of the vehicle.

Image processing is also used as a sensor. The environment is recorded by means of an optical device. Afterward the image data are converted into electrical signals by means of semiconductor image sensors (charge-coupled device or complementary metal oxide semiconductor). The digital image is analyzed by means of object detection algorithms. A very promising application for image processing systems is used by sensor fusion with other sensor systems, as it is used for collision mitigation and collision avoidance systems.

Vehicle-to-vehicle and vehicle-to-infrastructure communications provide additional information. This is done for short distance (e.g., intersection assistance) or even long distance (μ-factor information of the road surface or traffic sign information). Various communication technologies with different frequency bands are available today (e.g., Wireless LAN IEEE 802.11, HiperLAN, CALM etc.). A frequency range from 5.855 to 5.925 GHz is assigned for vehicle-to-vehicle communication in Europe.

Sensors for vehicle status detection, for example, position, friction coefficient of the road, and rain, have become of high importance for modern assistance systems. In particular, digital map data are used for many different ADAS as they contain attributes about the road infrastructure and legal information such as, that is, velocity limitations.

2.2 Actuator technology

In order to convert the signals of the ECU into the desired vehicle reaction, the vehicle needs to be equipped with actuators for steering, braking, and acceleration. Apart from these main groups of actuators, additional actuators (e.g., curve light and pre-crash) are available for ADAS.

At present, steering support is provided by means of hydraulic power steering (HPS). Additional support is provided by the adaptive HPS, the so-called Servotronic. The Servotronic steering system is able to reduce the steering support with rising vehicle velocity and therefore can offer an easy moving steering wheel for parking in the low velocity range as well as a comfortable steering behavior at high velocity.

Compared to the HPS, the electrohydraulic power steering (EHPS) consists of an electronic powered hydraulic pump, which provides the steering force. The control is managed demand controlled.

The electric power steering (EPS), also called *Servoelectric*, provides the steering force to the servo hydraulic power assistance with the help of an electric actuator. The electric motor and the related transmission exemplarily are fixed with the steering column, with the steering gear pinion or with the gear rack. The EPS provides the possibility of a steering boost by any dimension.

Additional steering angle can be applied by the active front steering. The additional angle is applied by means of a planetary gearbox, integrated into the steering system. This superposes the driver's steering intention with any desired angle. The mechanical connection between the steering wheel and the wheels leads to a direct response in lateral dynamic driving conditions.

By-wire steering systems decouple the steering-wheel input from the front wheels. The concept is based on performing steering maneuvers without mechanical transfer of steering torque. There is no mechanical connection between the steering wheel and the front wheels existent. In order to achieve necessary safety requirements, complete system redundancy is necessary. The vehicle has to remain steerable, even in the case of a system error.

The necessary brake interference for ADAS and automated driving vehicles can be realized on the basis of

- Brake booster:
  The brake booster is able to regulate a desired target brake pressure hydraulically and can therefore brake the vehicle without usage of the brake pedal by the driver.
- Electrohydraulic brake (EHB) system:
  The EHB system can divide the brake force up to each single wheel. The advantage of EHB system is the improved operating time because the control unit recognizes the demand for more braking by observing the pedals or their operating rate and acts before a fully hydraulic brake could build up pressure.
- Electromechanical brake (EMB) system:
  The EMB system is a rearrangement of the brake system to a completely electronically operated brake with dry function for the brake disks. This brake system offers subfunctions for automatic driving (e.g., automatic parking or automatic emergency braking).
- Hybrid brake system:
  The hybrid brake system represents a combination of a hydraulic and an EHB at the front axle and a dry electrical brake at the rear axle of the vehicle. An advantage of the hybrid brake system is the simplified package at the rear axle by the omission of hydraulic components.
- Brake-by-wire system:
  In brake-by-wire systems, the desired deceleration by the driver is electronically passed on from the pedal to the brake system. The by-wire-brake system works electrically without a hydraulic medium. This allows a simple electronic interface for driver assistance systems

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

as well as the realization of additional comfort functions and package advantages. Redundancies are absolutely necessary. A high quality battery and energy management is necessary because the brake system does not work on hydraulic basis.

Automated acceleration can be provided by means of electronic acceleration pedals (e-gas). These pedals are equipped with electric motors, pedal sensors, and electronic throttle-position controllers. The pedal sensor transforms the accelerator pedal position into an electronic signal, which is transferred into an appropriate position of the throttle. Modern ADAS pedals provide force feedback functionality, for example, for giving drivers feedback in critical driving situations or urging them to reduce driving velocity.

2.3 Legal aspects

From the today's point of view, automated vehicles are facing different legal issues. In a very first but important step, it must be understood that legal issues will usually strongly depend on the nature of access rights to the roads: as far as automation is to be used on roads with unrestricted access, the legal issues mentioned in the following are strongly relevant. In case, however, that the testing of the automation is done on test grounds with only restricted access by the public (e.g., enclosed working areas or test tracks), the risks for life, health, and property of employees as well as visitors and even trespassers will need consideration. It potentially raises the issue of liability in case of an accident. Apart from this, the limiting issues in case of higher automation levels generally are the national regulatory law on the conduct of the driver, product liability, road traffic liabilities, and vehicle type approval (Gasser, 2010a, b).

The most important starting point to identify those legal issues, which are relevant in case of automation on publicly accessible roads, is to clearly describe the level of automation provided. The ADAS, which are already available today, show quite easy how high the level of legally permissible automation already is today. In terms of regulatory law of conduct of the driver, this automation level is possible, as the driver always remains in the position to take over control immediately and at any point. The driver therefore necessarily observes the surrounding traffic, which belongs to his tasks from the legal situation, found in the terms of regulatory law on driver's conduct. As far as systems actively intervene into the task of driving, it is their overrideability that ensures compatibility with regulatory law (such as the Vienna Convention on international level). This will usually be an issue in case of systems, addressing near accident situations (Gasser, 2010a).

Apart from this, product liability takes an important role in the implementation of the automation for road transport. As far as automation is meant in the way that the vehicle remains under control of the driver and does not reach the level of autonomy, the risks, combined with the estimation of foreseeable use, in terms of product liability remain fundamental. In technical terms, any higher level of automation will require precautional technical measures in terms of reliability, in order to avoid running into product liability. Currently, a nationally focused project group has investigated the legal issues combined with the increasing amount of automation in road traffic, lead by BASt (Federal Highway Research Institute) in Germany. The report is expected in 2011 and it will cover the issues of automation in traffic in general. The report will not be restricted to interventions into the task of driving. These are rather well researched already and discussed with respect to the international law (Vienna Convention on Road Traffic) (Gasser, 2010a).

For near-accident situations, the most important aspects discussed so far are again regulatory law, highlighted in the discussion on the meaning of the Vienna Convention on Road Traffic for intervening systems as well as—most important again—product liability.

Art. 8 (1) of the Vienna Convention from 1968 on Road Traffic postulates that

> "*Every moving vehicle or combination of vehicles shall have a driver*."

Consequently, Art. 8 (5) VC constitutes the driver's obligation to be able to control his vehicle permanently:

> "*Every driver shall at all times be able to control his vehicle or to guide his animals*."

Art. 13 (1) VC substantiates this obligation with regard to speed and distance between vehicles; Art. 13 (1) VC says (in extracts):

> "*Every driver of a vehicle shall in all circumstances have his vehicle under control so as to be able to exercise due and proper care and to be at all times in a position to perform all manoeuvres required of him [. . .]*."

National road traffic regulations such as the German Road Traffic Regulations reflect this basic idea of permanent controllability.

Interventions of automated systems in the vehicle guidance, which do not comply with the driver's will and which cannot be corrected and overridden are therefore considered as incompatible with controllability in terms of the Vienna Convention (Seiniger *et al*., 2011).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

13	Uniform provisions concerning the approval of vehicles of categories M, N, and O with regard to braking
13-H	Uniform provisions concerning the approval of passenger cars with regard to braking
79	Uniform provisions concerning the approval of vehicles with regard to steering equipment

Figure 2. Relevant regulations concerning the type approval of added functionality. (Reproduced from Seiniger *et al*., 2011. © The interactIVe Consortium.)

The second crucial aspect concerning near-accident interventions into driving is product liability. Liability claims arising from damages caused by a defective product may be based on three distinct liability systems: product liability (based on the Product Liability Directive 85/374/EEC), contract (contractual liability), and/or tort (extracontractual liability) in EU Member States (Seiniger *et al*., 2011).

With regard to the liability deriving from the above-mentioned sources of law, a product should comply with the state of the art in science and technology—in order to be able to prove that this state of the art was adhered to during the design, the construction, and the production processes and with that in order to reduce product liability risks, relevant systems of rules such as the RESPONSE 3 Code of Practice (Knapp *et al*., 2009), and technical standards such as the FDIS/ISO 26262 (Sauler and Kriso, 2009) should be observed. From a product liability point of view, it is recommendable to design near-accident interventions in a way allowing the driver to override automated braking and/or steering interventions any time the driver wishes to do so.

On EU-level, Product Safety Law is based on the General Product Safety Directive (GPSD) 2001/95/EC. Owing to its character as a directive, the GPSD had to be transposed into national law by the individual EU Member States. Art. 2 GPSD defines terms such as "product," "safe product," "dangerous product," "recall," and "withdrawal" for the purposes of the GPSD (Seiniger *et al*., 2011).

A very technically oriented legal aspect to be assessed in detail is that of type approval. The mandate to approve vehicles for traffic belongs to the government of each country. However, European countries accept requirements defined by the United Nations Economic Commission for Europe's World Forum for the Harmonization of Vehicle Regulations (UN ECE WP.29) because of the transposition of the EU directives 2007/46/EC, 2002/24/EC, and 2003/37/EC.

There are two different types of vehicle regulations: the 1958 agreement system, which requires vehicles to be certified by an independent technical service (Europe, Japan, rest of the world), and the 1998 agreement (the United States, China, most of the 1958 states), which requires the vehicle manufacturers to certify their vehicles themselves.

The 1958 agreement with its ECE regulations covers most of the world with the exception of the United States and China. It is considered as the most important set of vehicle regulations.

Intervening systems and automated systems act on the vehicle brakes, throttle, and steering systems. The following regulations are of relevance, concerning the type approval of the added functionality (Figure 2).

3 ADVANCED DRIVER ASSISTANCE SYSTEMS

ADAS compensate the known weaknesses of human drivers and support them in their driving tasks. The motivation for the introduction of ADAS and the research of automated and automated vehicles are manifold. The four most important objectives are as follows:

- Increase of driving comfort: Comfort systems relieve the driver from annoying and monotonous tasks in order to ease the drive. Those systems already have a positive effect on traffic safety, as they are for instance keeping the distance to other cars and they avoid quick changes in speed.
- Increase of vehicle safety: ADAS support the driver by increasing the active safety of a vehicle. Active safety can be separated into perception safety (driver support in order to perceive relevant traffic information of the vehicle to other traffic participants), control and interaction safety (driver support to interact with vehicle control elements and perception of vehicle status information), driving safety (support the driver in critical driving situations), and condition safety of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the vehicle (support the driver during the drive, that is, by comfortable climate conditions).

- Improvement of traffic efficiency: ADAS support the improvement of street capacity. Thus, traffic jams can be prevented or are dissolved faster. In addition, the vehicles, which are approaching the traffic jam, can be redirected automatically.
- Reduction of environmental impact: ADAS support the reduction of fuel consumption and noise emissions, for example, assistance systems can give suggestions for gear changes, acceleration maneuvers, or early reduction of vehicle velocity depending on the situation and traffic. In addition, for the optimization for complex drive train structures (hybrid engines), the operating strategy can be implemented.

Support systems or automated systems require the interaction with the environment and an action of the driver, for at least the activation of the system. ADAS, which take over part of the driving task, are part of the closed control loop.

The driver support is divided into one of the three levels of the driving task (Donges, 1982). ADAS are classified, based on their functions and their type of driver support, into the levels navigation, guidance, and stabilization (Figure 3).

On the navigation level, the driver decides on his route inside an existing road network. While driving, the navigation includes the perception of necessary information to maintain the route. If needed, an adjustment of the route due to changed boundary conditions can be pursued. The time demand for the driver on the navigation level is not critical and amounts to over 10 s.

On the guidance level, the driver adapts his driving style to the perceived course of the road and the surrounding traffic. The guidance level comprises subtasks such as lane keeping, following, overtaking, and reaction to road signs. The tasks of this level can be divided into lateral and longitudinal guidance and are realized between 1 and 10 s.

The stabilization level is characterized by changes of the chosen driving strategy in vehicle-related control variables by the driver (steering movement, accelerator, brake, and gear choice). There is a permanent comparison between set and current value of speed and lane position. "Stabilization" for the driver means the avoidance of unsupervised momentum of the vehicle. The stabilization level is highly time critical with a time demand below 1 s. On the basis of this, time criticality automatic intervening systems are used for the stabilization level. The driver does not carry out the necessary action for the vehicle stabilization on his own. Systems on the stabilization level are mostly equipped with a possibility to activate or deactivate the HMI. In some cases, there is no visible HMI during driving (e.g., ABS and ESP).

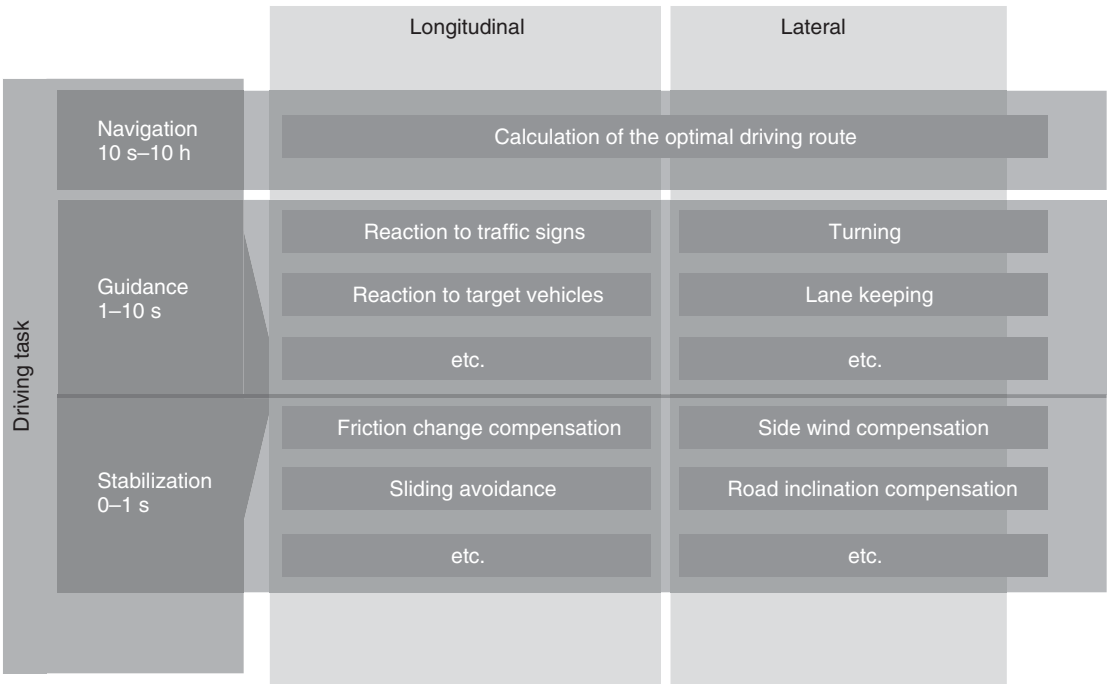


Figure 3. Classification of ADAS in three levels of the driving task.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

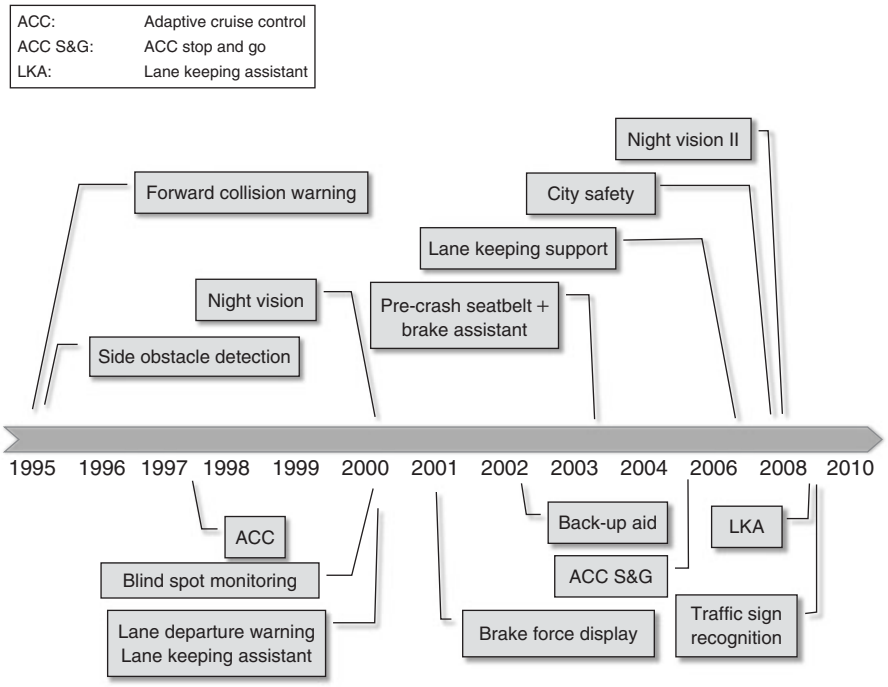


Figure 4. Market introduction of ADAS.

3.1 Overview of ADAS

One of the first ADAS introduced in the market was a collision warning system for heavy goods vehicles in 1995, which was based on a 24 GHz radar sensor for object detection and forward collision warning system. Figure 4 shows the market introduction of different ADAS on guidance and navigation level.

Currently, various different ADAS are available in premium vehicles. They are also being introduced in middle class vehicles. Examples are systems for parking support, driving at nighttime, longitudinal, and lateral vehicle control.

One of the first systems introduced to provide continuous support was adaptive cruise control (ACC). ACC is an enhancement of the cruise control, which was one of the focus research fields of the PROMETHEUS project. The system particularly supports drivers on motorways and country roads by keeping a safe distance and controlling the velocity with respect to vehicles in front. In case the vehicle approaches a slower vehicle, the system intervenes by means of reducing the throttle first and then operates the brake system in order to keep a safe distance to the preceding vehicle. Therefore, the ACC uses its distance sensor in order to detect the range and range rate of the vehicle in front. The desired values regarding the distance are mostly dependent on the velocity and the drivers can adjust the distance by defining a time gap to the vehicle in front. The acceleration and the deceleration behaviors of the vehicle are controlled depending on the driving status by an ECU. The driver is able to turn the function off or override it at any time by pushing the accelerator or the brake pedal. In case no target vehicle is present, the ACC is used as a conventional cruise control.

The market introduction of ACC for different vehicle manufacturers in Europe is given in Figure 5 as an example.

3.2 Research on ADAS

In addition to available ADAS, research on future systems is ongoing in different research areas. The most important ones are as follows:

- Safety (e.g., collision mitigation and avoidance)
- Environmental protection (decarbonization and noise reduction)
- Integrated HMIs and integrated interaction strategies
- New mobility concepts
- Automation

With regard to automated driving vehicles, research on platooning provides the next step to automation. A platoon is described by at least two vehicles, which are electronically connected without any mechanical

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

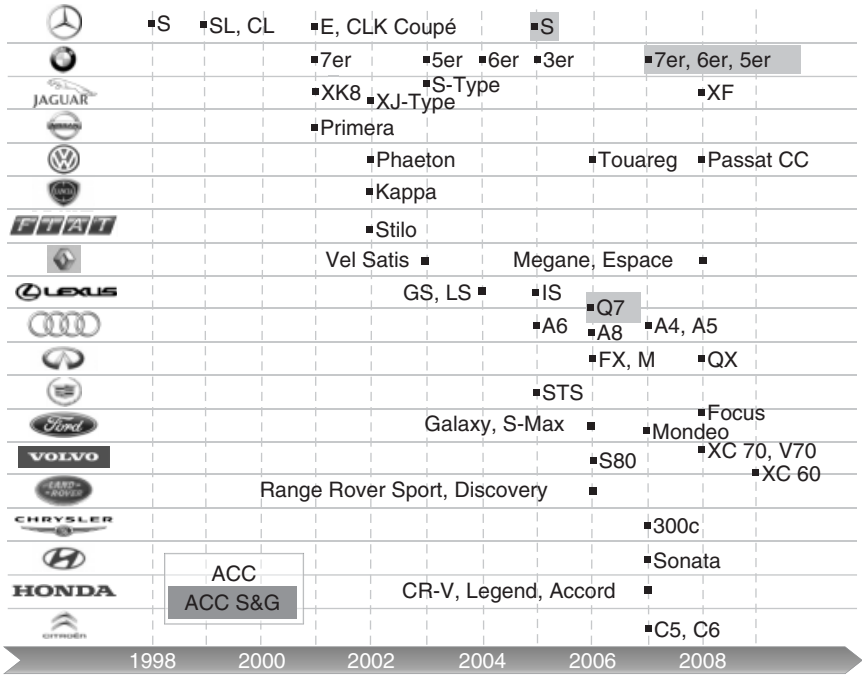


Figure 5. Market introduction of ACC in Europe.

connection between the vehicles. Work in this area was undertaken by VW called *CONVOY Driving* in the PROMETHEUS project (Ioannou 1997), by the PATH (Partners for Advanced Transit and Highways) project of the Institute of Transportation Studies (ITSs) of the University of California in Berkeley together with the California Department of Transportation (Caltrans) in the Unites States in 1996 and the ARTS project (Advanced Road Transportation System), which was funded by the Japanese Ministry of Construction. The ARTS project was merged to the ITS (Transportation System) project in 1995.

The lateral vehicle control of the automated vehicles in these projects is based on magnetic studs, which are integrated into the infrastructure in the form of a dedicated guide way every approximately 3 m. The nails provide the trajectory of the vehicles. Longitudinal vehicle guidance in the platoon is provided by distance and velocity detection in each single vehicle. This information is transmitted via vehicle-to-vehicle communication or vehicle-to-infrastructure communication.

The aim of PATH was to develop long-term strategies in order to cope with the immense traffic in California. Between 7 and 10 of August 1997, a vehicle platoon consisting of eight Buick LeSabres vehicles was demonstrated. The distance between each longitudinal and lateral controlled vehicle was 6.5 m at a velocity of up to 96 km/h. The distance accuracy was measured in the range 10 cm at constant driving and 20 cm at acceleration and deceleration maneuvers (PATH Program, 1997). However, even today, there is no impact on the road traffic in the United States.

In contrast to the PATH approach (only lateral control by means of magnetic nails, longitudinal control provided only to following vehicles based on vehicle sensors), the point-follower approach (infrastructure/vehicle sensors, no target vehicle necessary) demands additional communication between the road infrastructure and the vehicles. This



Figure 6. Automated driving bus using a guide wire in the ground of the driving way. (Reproduced by permission of MAN.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

communication can be provided by LCX (Leakage CoaXial Cable) cables at the surface of the road. The LCX cable provides data transmission and acts as an antenna. Field tests were conducted on a test tracks in Germany and Japan (Gehring, 2000). Figure 6 shows field tests with a MAN bus demonstrator using a guide wire in the ground of the driving way in 1977 (Zeit, 1980).

In Europe, truck platooning was performed for the first time in the European research projects PROMOTE CHAUFFEUR 1 and 2 of the fifth European framework program. The following distance between the three used demonstrator trucks was between 6 and 16 m at a velocity of 80 km/h. On the basis of the reduced air resistance, fuel in the range between 15% and 20% was reduced compared to conventional driving.

The control approach was called *tow-bar* principle, which was realized by means of vehicle-to-vehicle communication in the 5.8 GHz band and an infrared camera system. The distance to the target vehicle as well as the velocity and the acceleration of each vehicle was transmitted to all following vehicles in the platoon. The infrared camera determined the relative lateral position between the vehicles.

Next to the "tow-bar" principle, the so-called CHAUFFEUR assistant was introduced in PROMOTE CHAUFFEUR 2. The assistant was a combination of an ACC system and the lateral control in order to support the driver of the first vehicle in the platoon.

The final demonstration of all functions was done on the IVECO test track in Balocco, Italy, on 7 May 2003 (Bonnet *et al*., 2003).

Similar to the PROMOTE CHAUFFEUR project, the national funded German research initiative KONVOI performed research on truck platooning. Next to technical research questions, covering the steering hardware and the function development for automated longitudinal and lateral control impacts on surrounding traffic, the strategies to build the platoon, acceptance studies, and impacts on truck drivers were investigated in detail. In addition, legal aspects for an introduction of truck platoons were analyzed. The demonstration of a truck platoon with four tractor–semitrailer combinations took place on the German motorway in 2009 (Figure 7) (Deutschle, 2008).

SARTRE is a European project, cofunded by the seventh framework program of the European Commission, that targets to develop strategies and technologies to allow vehicle platoons to operate on normal public highways with significant environmental, safety, and comfort benefits.

In the SARTRE platooning system, a human driver like in KONVOI drives the lead vehicle manually. The other vehicles follow the trajectory of the lead vehicle. Thus, the lead vehicle driver has a huge responsibility. In order to improve the safety in a platoon, different ADAS systems are installed in the lead truck to support the driver.

In order to improve the comfort and support the drivers to find suitable platoons, a back office unit calculates the meeting point for the potential following of the platoon and provides navigation instructions to the drivers (Deutschle, 2008; SATRE project, 2011).

4 INFRASTRUCTURE-BASED AUTOMATED DRIVING VEHICLES

Operation of automated vehicles in dedicated infrastructure provides several advantages compared to fully automated driving vehicles. Dedicated infrastructure can separate automated vehicles from ordinary traffic participants, which solves integration issues. Furthermore, dedicated infrastructure can be equipped with sensors, for example, LCX cables for platooning. Basically two different scenarios are possible for automated vehicles in dedicated infrastructures:

- Guide ways to separate automated vehicles from other traffic participants
- Guide ways to support automated vehicles or ADAS

Different vehicle types are possible to be operated in these scenarios:

- Personal rapit transit (PRT)
- High tech busses
- Vehicle platoons
- Automated vehicles

Infrastructure-based automated driving vehicles provide new mobility concepts especially in the field of public transport, that is, PRT and high tech busses.

PRT is a fully automated transport system using small driverless vehicles (Figure 8). The vehicles navigate automatically along a network of dedicated guide ways. PRT



Figure 7. KONVOI truck platoon demonstration on the German motorway in 2009.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) (b)

Figure 8. (a,b) Examples for PRT systems. (Reproduced from Bly and Lowson (2009). © European Commission.) (Reproduced by permission 2getthere B.V. (http://www.new.2getthere.eu).)

offers a driverless taxi, providing on-demand. By this, a nonstop transport from the origin station to any other destination station selected by the passenger on the whole dedicated guide way network will be possible. Therefore, PRT can be described as a horizontal lift or elevator. All PRT systems under current development are electric powered and thus (locally) emissions free. PRT systems aim to have empty vehicles waiting at stations for arriving passengers, so that there is little or no waiting time. Because the electric vehicles are quiet and emit no exhaust pollutants, they can be routed through buildings.

One system installed at the airport in Heathrow is "ULTra," developed by Advanced Transport Systems of Thornbury, the United Kingdom. The four-seater vehicles have four rubber-tired wheels and the size of a small car. They are powered by electric motors, running from lead-acid batteries. ULTra vehicles are driving on a 2 m wide concrete or metal track. Batteries are recharged in the stations. The maximum driving velocity is 40 km/h, and the vehicles can climb a 10% gradient and negotiate 5 m radii. They can go reverse and steer backward, offering great flexibility in maneuvering (Bly and Lowson, 2009). Another PRT system in operation is located in Masdar, Abu Dhabi (2getthere, 2011).

High tech buses in separated lanes are also already in operation. An example for such a BRT (bus rapid transit) system is the Phileas program in Eindhoven/Veldhoven. Phileas is a new concept for comfortable passenger transport on high frequency dedicated bus lanes (Figure 9). The guide way is fitted with magnetic markers for electronic lane assistance and precision docking. Phileas has hybrid electric propulsion, a large transport capacity, and precision docking, which makes it possible for passengers to quickly enter and exit the vehicle. Thus, the stop times are short and the average speed can be kept as high as possible. In comparison with tram or metro systems, the investment and maintenance costs for the infrastructure are low. Overhead wires and rails are not needed. Phileas combines the advantages of tram and metro systems with the flexibility and low costs of a bus system (Phileas, 2011).



Figure 9. Phileas high tech bus on a dedicated guide way (Phileas, 2011). (Reproduced by permission of VDL Bus & Coach bv.)

In industrial applications, different types of automated and driverless vehicles are already utilized since many years. These are vehicles, which transport goods on dedicated driving routes in factories or work yards.

In 2001, the FOX GmbH in Germany to transport large amounts of euro pallets on a factory yard has modified two commercial vehicles. The vehicles were equipped with additional sensors for automated driving. The driving route of 190 m is marked by transponders in the floor. These

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

transponders are detected by the vehicles with an accuracy of about 5 mm. Therefore, the lateral control of the vehicles on the driving route is possible within a 2 cm range at a velocity range of about 5 km/h. A laser scanner detects obstacles in front of the vehicles. In addition, the vehicles have sensors integrated in the front bumpers. The bumpers are made of soft foam. In case of contact between an obstacle and the bumpers, the sensors activate the braking system and the vehicles are able to halt within a range of 0.4 m after contact. The vehicle position and movement is transferred to a control room by means of communication technology. The vehicles are driving automatically, but a driver can override the overall system at any time (Götting KG, Abt. FOX, 2008).

5 FULLY AUTOMATED DRIVING VEHICLES

Automated driving vehicles are under research since the 1970s. Different research activities demonstrated automated driving according to the latest state-of-the-art technology available at the time of the activity. In recent years, sensor technology for environment detection and electronically controlled actuators became available in production vehicles. Vehicle manufacturers use their knowledge on the necessary technology and provide demonstrator vehicles for automated driving in order to underline the public perception of the company image.

As legal issues still do not allow automated vehicles to be driven on public roads, first systems in low velocity ranges are available at dedicated infrastructures such as company depots, work yards, or for demonstration purpose within inner city limits, separated from conventional traffic.

In order to stimulate recent research on automated vehicles and automated driving by universities and research institutions, different challenges have been initiated by different stakeholders (e.g., military and governments). On the basis of the experience of the research work and the availability of sensor systems for path planning and obstacle detection, automated driving vehicles demonstrators have been set up to fulfill manifold purposes, such as creating development platforms, organizing driver training, or defining new markets. Furthermore, new mobility concepts are being investigated by means of automated driving vehicles.

5.1 Research activities

In the last decade of the past century, different research activities on automated vehicle guidance were conducted with the results of automated driving demonstrations. Within these activities, prototype vehicles were built up and tested. Some tests of the prototypes were carried out as journeys on public roads in real traffic:

- VAMP: test on a journey of approximately 1600 km from Munich (Germany) to Odense (Denmark) in 1995 (Maurer *et al*., 1996);
- NavLab 5: test on a journey of approximately 4587 km from Pittsburgh (USA) to San Diego (USA) in 1995 (Bertozze, Broggi, Fascioli, 2000);
- ARGO: test on a journey of approximately 2000 km in the MilleMiglia tour (Italy) in 1998 (Bertozze, Broggi, Fascioli, 2000);
- Google Cars: test drives of approximately 1600 km without human intervention and more than 230,000 km in total (Markoff, 2010).

In the recent past, challenges were announced by different stakeholders. The American DARPA (Defense Advanced Research Projects Agency) initiated the DARPA Grand Challenges and the DARPA URBAN Challenge in order to demonstrate the state of the art for automated driven vehicles. In Europe, the Grand Cooperative Driving Challenge (GCDC) was initiated to combine different European research activities in cooperative and automated driving.

5.1.1 VaMoRs/VAMP

One of the first automated driving vehicles was called *VaMoRs* (Versuchsfahrzeug für autonome Mobilität und Rechnersehen) of the German Bundeswehr University of Munich. The second vehicle from Daimler was called *VITA*.

Between 1985 and 2004, a transporter of the type Mercedes DB 508 was converted in order to control electronically the lateral and longitudinal dynamics of the vehicle. The core of the research activity was machine vision, which was performed by means of camera systems. The vision system detected objects and lane markings in a distance of up to 120 m in front of the vehicle. By means of this collision avoidance, automated distance control, lateral control, and lane change maneuvers were implemented. Next to machine vision, additional sensors of a digital map and satellite position system were used in order to detect intersections and turnings.

The achieved know-how of the VaMoRs project was used to build up the vehicles VAMP (VaMoRs Passenger Car) and VITS-II. These Mercedes Benz 500 SEL vehicles were equipped with a radar distance sensor additionally to the machine vision system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

After thousands of test kilometers automated driving for testing purpose, a demonstration was conducted from Munich (Germany) to Odense (Denmark) in 1995. The test route had a length of approximately 1600 km. A total of 95% of the test track was driven in automated mode at velocities of up to 180 km/h. In total, approximately 400 lane changes were performed during the test drive (Maurer *et al.*, 1996).

5.1.2 No hands across America

In 2005, the "No Hands Across America" tour was performed by the Robotics Institute of the Carnegie Mellon University, Delco Electronics, and AssistWare Technology. The prototyped vehicle called *NavLab 5* drove approximately 4587 km from Pittsburgh to San Diego (USA).

The vehicle was equipped a windshield mounted camera, a GPS receiver, and a radar sensor for obstacle detection, which were used for lateral control of the vehicle. Longitudinal vehicle control was performed manually by the driver.

NavLab 5 was able to drive 98.2% of the journey (4503 km of a total of 4587 km) with automated lateral control. The system proved to be robust with respect to real road and traffic conditions. The major problems encountered were due to rain, low sun reflection, shadows of overpasses, construction zones, road, and road markings deterioration (Bertozze, Broggi, Fascioli, 2000).

5.1.3 The ARGO project and the VisLab intercontinental autonomous challenge

From 1996–2001, the University of Parma worked on the ARGO project, which had the goal to enable automatic lane following on motorways for a modified Lancia Thema. In 1998, a demonstration of a 2000 km long journey on the motorways of northern Italy with an average speed of 90 km/h took place. A total of 94% of the time the car was in fully automated mode, with a longest automatic stretch of 54 km. The vehicle was equipped with only two black-and-white cameras and used stereoscopic vision algorithms to perceive its environment.

On the basis of the results of the ARGO project, the idea for the intercontinental autonomous challenge was born. In total, 13,000 km was driven from Parma (Italy) to Shanghai (China) with four automated controller electric vehicles (two vehicles traveling and two vehicles as backups). The challenge was performed from 26 July 2010 to 28 October 2010 including two different continents with changing geographical morphology, traffic conditions, weather, infrastructures, and so on.



Figure 10. Vehicles of the VisLab Intercontinental Autonomous Challenge. (Reproduced by permission of Vislab.)

Each test vehicle was equipped with seven cameras, four laserscanners, GPS receivers, and vehicle-to-vehicle communication. The vehicles are shown in Figure 10.

During the journey, all vehicle data and all perception data were logged. The amount of data added up to about 50 TB and provided a data set for a variety of different driving situations, weather conditions, and infrastructure conditions.

In difficult situations, for example, bad weather or heavy and chaotic traffic conditions, driver had to take over control and switch to manual mode because not sufficient information from the perception system had been available (Broggi, 2011).

5.1.4 DARPA grand challenges

The DARPA Grand Challenge took place in the years 2004 and 2005. The task of the Grand Challenges was to cover a track of about 150 and 132 miles through the desert in fully automated mode with no human intervention. The route was announced shortly before the start of the challenges so that the competing teams could not tune their automated vehicles according to the track.

In the first challenge of 2004, no vehicle was able to cover the route. The best team was only able to cover around 5% of the track. In 2005, four vehicles reached the destination within the given time limit of 10 h.

The winner of the Grand Challenge 2005 was a VW Touareg called *Stanley* of the Stanford University. The vehicle was equipped with four laser scanners, 24 GHz radar sensors, mono- and stereo image processing systems, and a GPS system. The vehicle movement was measured by means of an internal measurement unit. The sensor

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

data and the control algorithms were processed by seven Pentium M processors with 1.6 GHz calculation speed each (Thrun, 2006).

5.1.5 DARPA urban challenge

The DARPA Urban Challenge was conducted in 2007 on an old military area close to Victorville in the state of California in the United States. The main differences to the Grand Challenges were the involvement of other traffic participants and the urban scenario with road regulations and urban infrastructure. Especially, the behavior of the other traffic participants, which were vehicles driven by stuntmen, needed to be considered. Pedestrians and bicycles were not part of the Urban Challenge. The route was shortened to 60 miles, which needed to be covered in <6 h.

The winner of the DARPA Urban Challenge was the team Tartan Racing of the Carnegie Mellon University with the vehicle called *Boss* (Urmson *et al*., 2008).

5.1.6 GCDC in Europe

In contrast to the DARPA Grand Challenges and the Urban Challenge, the European GCDC has no military background. It was initiated by TNO, a governmental research organization of the Netherlands and High Tech Automotive Systems (HTASs). The goal was to create a liaison between research groups on the topic of cooperative and automated driving. The first GCDC was held at Helmond in the Netherlands in 2011. It took place on the highway A270, which was closed for public traffic while the challenge was performed.

The focus of the first GCDC was on longitudinal vehicle control for platooning by means of vehicle-to-vehicle communication for cooperation. Different communication and sensor hardware and different types of vehicles needed to be part of the control strategy. In order to provide boundary conditions, the communication protocol, the signals to be communicated, safety measures, the wireless communication, and a mandatory message set needed to be used by all participants. Each participant could choose the vehicle, in-car architecture, the environmental sensors, the control strategy, and the usage of the communicated information freely.

In total, 11 teams competed in the four given scenarios (platoon stability, joining a platoon, joining a platoon at a traffic light, and merge on intersections). The teams were grouped into 2×2 platoons on parallel lanes. The organizers provided the lead vehicles, which were defining the driving velocity of the platoon. Points were provided for different criteria such as string stability and smoothness. The first GCDC was won by team AnnieWay of the Karlsruhe Institute of Technology in Germany (Grand Cooperative Driving Challenge, 2011).

5.1.7 Google cars

The Internet Company Google also works on automated vehicle in terms of using artificial intelligence software for driverless vehicle control. The vehicle combines information from the company's Google Street View service with sensor signals from video cameras inside the vehicle, a LIDAR sensor on top of the vehicle, radar sensors on the front of the vehicle, and a position sensor attached to one of the rear wheels. The system drives at the speed limit, which is stored on its digital map and maintains its distance from other vehicles using the environmental sensors. The driver can override the automated driving function at anytime (Markoff, 2010). In 2010, seven test vehicles have driven 1600 km without human intervention and more than 230,000 km with only occasional human intervention. Next to the technical feasibility demonstration, Google lobbied for two changes of state laws that made the state of Nevada (USA) the first state, where driverless vehicles can be legally operated on public roads. The first bill is an amendment to an electric vehicle bill that provides for the licensing and testing of automated vehicles. The second bill will provide an exemption from the ban on distracted driving to permit occupants to send text messages while sitting behind the wheel (Markoff, 2011).

5.2 Examples for automated vehicle demonstrators and research platforms

In recent years, car manufactures, computer companies, and universities established a variety of different research platforms. On the basis of the experience of the DAPRA challenges, research on automated driving is proceeding in manifold ways.

At the University of Braunschweig, Germany, automated driving in the city's inner ring road of Braunschweig is demonstrated in the so-called *Stadtpilot* project. The goal is to drive fully automated in the traffic flow and to behave according to traffic rules (Saust *et al*., 2011).

The Volkswagen Group Electronics Research Laboratory and Stanford University are working on the Audi TTS Pikes Peak automated vehicle. The vehicle is a modified Audi TTS Coupé Quattro called *Shelly*. The goal is to improve vehicle automation and explore capabilities of current and future driver assistance systems by automated drive up the legendary 12.42 miles Pikes Peak Hill Climb route in Colorado, the United States. The focus of the research

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

work is on path planning and vehicle stability. The route to Pikes Peak consists of paved and graveled roads and can contain weather changes at all time. Therefore, vehicle guidance of the Audi TTS Pikes Peak is working entirely on differential GPS and vehicle state information such as speed and acceleration measured by wheel-speed sensors, an accelerometer, and gyroscopes, but without additional sensors for environmental detection. The control algorithm for path planning and vehicle stability compares the sensor data to a digital map of the route to determine possible deviations and necessary maneuvers. The resulting action for longitudinal and lateral vehicle control is performed by the vehicles production hardware, which already exists in Audi production vehicles (Blackman, 2010).

The VW Golf 53 + 1 is an automated driving robot, which was developed to perform reproducible driving maneuvers in the high dynamic range on test tracks (Kompaß, 2008). Reproducible driving maneuvers allow vehicle development by means of a driver-independent analysis of vehicle dynamics (function or system), driver behavior (impact of the system), and boundary conditions (environment). The driving robot is able to take over the function of the driver with sufficient accuracy and reproducibility. Longitudinal and lateral vehicle dynamics are controlled by the robot and therefore provide automated driving on a predefined course. The driving trajectory is determined by means of a differential GPS. The transmission of the additional correction signal allows a positioning in the centimeter range. Traffic cones mark the driving track, which are detected by a laser scanner in the front of the VW Golf. The vehicle ECU fuses the GPS and the laser scanner data and calculates the driving trajectory on the track. Within the driving path, the ideal racing line is calculated for lateral vehicle dynamic control. These calculations result in desired values for the maximum driving velocity and the vehicle acceleration. The vehicle is equipped with a brake booster, which allows high dynamic braking maneuvers (Kompaß, 2008).

The BMW "Track Trainer" was designed by BMW for race driver training (Kompaß, 2008). The ideal racing line of the racetrack is calculated off-line. The calculation considers position data from GPS as well as vehicle parameters from test runs of professional drivers. An inertial measurement unit compensates the GPS position in order to detect inaccuracies of the positioning. The vehicle ECU controls longitudinal and lateral vehicle dynamics and keeps the vehicle automated on the ideal racing line of the course. The driver experiences the vehicle behavior for some test rounds on the racing course and therefore learns how to stay on the ideal racing line. In a second phase, the driver takes over vehicle control and is supported by the vehicle (haptic and acoustic advises) in case he leaves the ideal racing line. All vehicle data are logged during the test drive. Afterward the driving style can be analyzed in detail and recommendations can be given by the system (Kompaß, 2008).

5.3 Cybercars

Owing to the usage of vehicles in highly populated urban environment, severe problems with respect to pollution, noise, and safety arise. A new mobility concept in terms of Cybercars uses the advantages of automobiles combined with automation technology. The idea of Cybercars was developed in the European funded research project "Cybermove" (Cybermove project, 2004). Cybercars are slow-moving automated driving vehicles for existing road infrastructure. The design of the vehicles allows easy and clean transportation within inner cities. Different persons, depending on the vehicle size, can share the electric vehicles. With existing technology, the velocity is limited to 30 km/h and therefore sufficient for inner urban usage. Some of these vehicles can also allow traditional manual driving in order to run among normal traffic. In these cases, the vehicles are called *dual-mode vehicles* and their automated capabilities allow them to be used in platoons, for example, in order to collect the cars.

Several companies and research organizations have been involved in recent years in the development of Cybercars. The first systems have been put in operation in the Netherlands at the end of 1997 and have been running successfully 24 h a day. Several other systems have been implemented in European research projects on this topic, see Cybermove project (2004), Cybercars project (2006), and CityMobil project (2011). The focus of the research is on the development of tools and systems to enable the driverless vehicles to perform the necessary driving maneuvers in a cooperative manner, that is, in cooperation with each other and also in cooperation with conventional driven vehicles (Vlacic and Parent, 2009).

6 SUMMARY

A first step toward automated driving is the introduction of ADAS into the market. Currently, the development of new assistance functions and the improvement of existing functions are ongoing. Especially, progress in sensor and actuator technology allows a shift from pure comfort function toward active safety and continuous driver support. In the recent past for longitudinal and lateral support in all velocity ranges from parking to city driving up to high speed driving became available.

Research activities in the field of automated driving demonstrate technical solutions for fully automated

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

systems. At present, these are being utilized in niche markets. However, in the evolution of today's ADAS toward automated driving systems, the legal aspects and the assessment of the systems need to be considered. High market penetration of ADAS in the future will be the foundation toward the fully automated driving vehicle.

RELATED ARTICLES

Active front steering for passenger cars
Steer by Wire, Potential and Challenges
Brake Systems, an Overview
Global Chassis Control in Passenger Cars
Vehicle Safety, Functional Safety, OBD Diagnosis
Applications of Radio Wave Technologies to Vehicles
Applications of Image Recognition Technologies to Vehicles
Active Safety, Pre-collision Safety and Other Safety Products (millimeter wave, image recognition, laser)
Human Machine Interface Design in Modern Vehicles
Intelligent Transport Systems: Overview and Structure (History, Applications, and Architectures)
Evolution and Future Trends
Technologies—Data Acquisition: Data fusion
Applications—Intelligent Vehicles: Driver Information Driver Assistance
Applications - Intelligent Vehicles: Autonomous Vehicles
Conclusions, Visions, Upcoming Standards

REFERENCES

Bertozze, M., Broggi, A., and Fascioli, A. (2000) Vision-based Intelligent Vehicles: State of the Art and Perspectives. *Robotics and Autonomous Systems* **32**, 1–16.

Blackman, C. (2010) Stanford's robotic Audi to brave Pikes Peak without a drive. Stanford Report, issue no. 3 (February).

Bly, P. and Lowson, M. (2009) Outline description of the Heathrow Pilot PRT scheme. Deliverable D1.2.2.2, CityMobil project.

Bonnet, C., Schulze, M., and Dieckmann, T. (2003) Promote Chauffeur 2 - Final Report. Deliverable D24, Promote Chauffeur 2 project.

Broggi, A. (2011) The VisLab Intercontinental Autonomous Challenge. *Keynote Speech at CityMobil Conference*, La Rochelle, France, CityMobil project.

CityMobil project (2011) http://www.citymobil-project.eu/ (accessed 18 October 2013).

Cybermove project (2004) http://www.cybermove.org (accessed 18 October 2013).

Cybercars project (2006) http://www.cybercars.org/ (accessed 18 October 2013).

Gasser, T.M. (2010a) 'Die "Projektgruppe Automatisierung": Rechtsfolgen zunehmender Fahrzeugautomatisierung' 11. Braunschweiger Symposium AAET, Proceedings, Intelligente Transport- und Verkehrssysteme und –dienste Niedersachsen e.V., 978-3-937655-23-9, Braunschweig.

Gasser, T.M. (2010b) Legal Aspects of Driver Assistance Systems' 19. Aachener Kolloquium Fahrzeug und Motorentechnik 2010, RWTH Aachen University, Aachen.

Deutschle, S. (2008) Das KONVOI Projekt - Entwicklung und Untersuchung des Einsatzes von Lkw-Konvois, Aachener Kolloquium Fahrzeug und Motorentechnik 2008, Aachen.

Donges, E. (1982) Aspekte der aktiven Sicherheit bei der Führung von Personenkraftwagen *Automobil-Industrie*, **27** (2), 183–190.

Grand Cooperative Driving Challenge (2011) http://www.gcdc.net (accessed 18 October 2013).

Gehring, O. (2000) Automatische Längs- und Querführung einer Lastkraftwagenkolonne. Dissertation. Institut A für Mechanik der Universität Stuttgart, Stuttgart

Götting KG, Abt. FOX (2008) automatisierte Serienfahrzeuge, Infobroschüre der Abteilung FOX, Götting KG, Lehrte.

Ioannou, P. (1997) *Automated Highway Systems*, Plenum Press, New York. ISBN: 0-306-45469-6

Knapp, A., Neumann, M., Brockmann, M., *et al*. (2009) RESPONSE 3: Code of practice for the design and evaluation of ADAS; version 5, http://www.acea.be/images/uploads/files/20090831_Code_of_Practice_ADAS.pdf (accessed 18 October 2013).

Kompaß, K. (2008) *Fahrerassistenzsysteme der Zukunft - auf dem Weg zum autonomen PKW, Forschung für das Auto von Morgen*, Springer Verlag, Berlin.

Markoff, J. (2010) Google cars drive themselves, in traffic *The New York Times*, 9.10.2010,

Markoff, J. (2011) Google lobbies Nevada to allow self-driving cars, The New York Times, 11.05.2011

Maurer, M., Behringer, R., Fürst, S., *et al*. (1996) A Compact Vision System for Road Vehicle Guidance. Proceedings of ICRP 1996, IEEE.

PATH Program (1997) Vehicle Platooning and Automated Highways, PATH Fact Sheets, University of California, Berkeley.

Phileas (2011) http://www.apts-phileas.com (accessed 18 October 2013).

Rowsome, F. (1958) What it's like to drive an auto-pilot car. Popular Science Monthly, USA, April.

SATRE project (2011) www.sartre-project.eu (accessed 18 October 2013).

Sauler, J. and Kriso, S. (2009) ISO 26262 – Die zukünftige Norm zur funktionalen Sicherheit von Straßenfahrzeugen, http://www.elektronikpraxis.vogel.de/themen/elektronikmanagement/projektqualitaetsmanagement/articles/242243/ (accessed 18 October 2013).

Saust, F., Wille, J. M., Lichte, B., and Maurer, M. (2011) Autonomous vehicle guidance on braunschweig's inner ring road within the stadtpilot project. Intelligent Vehicles Symposium (IV), IEEE.

Seiniger, P., Westhoff, D., Fahrenkrog, F., and Zlocki A. (2011) Legal Aspects, Deliverable D7.3, interactIVe project, www.interactive-ip.eu/ (accessed 18 October 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Thrun, S. (2006) Towards self-driving cars, 7. Braunschweiger Symposium AAET, 22./23.2.2006, DLR Braunschweig, Gesamtzentrum für Verkehr Braunschweig e.V.

Urmson, C., Anhalt, J., and Bagnell, D. (2008) Autonomous driving in urban environments: boss and the urban challenge *Journal of Field Robotics*, **25** (8), 425–466. DOI: 10.1002/rob.20255, Wiley Periodicals, Inc

Vlacic, L. and Parent, M. (2009) Cybercars2 - final report V2.0. Cybercars2 project.

Zeit (1980) Immer in der Spur, newspaper article, 09.05.1980

2getthere (2011) http://www.new.2getthere.eu (accessed 18 October 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto023
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Brake Systems, an Overview

Christoph Drexler and Ralf Leiter

TRW Automotive, Koblenz, Germany

1 Introduction	1
2 Types of Brakes	1
3 The Brake Process	14
4 Brake Dynamics	15
5 The Friction Process	20
6 Summary	21
Related Articles	22
Reference	22
Further Reading	22

1 INTRODUCTION

Today, some highly powered vehicles are on the road, often without it being clear what the required capacity of the brake system of these vehicles is. Thus, for example, when braking a modern vehicle of the upper middle class with an engine output of 130 kW (178 HP) to a complete stop from a speed of 220 km/h, the brake force must be more than three times the engine output. In addition to reliability and safety of the brake system, it is also required that this process takes place with absolutely no noise, is easy for the driver to apply and especially easy to control, as well as being comfortable for all passengers.

Even though the vehicle brakes have reached a very high level of development, more than 50% of all technical defects in motor vehicles are caused by the brake system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.

This article is © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

The first vehicles (including the means of transportation of antiquity) were already equipped with devices to stop the vehicle from rolling while stationary. This was the start of parking brakes. Thus, the vehicle brake is surely as old as the wheel itself. As these vehicles increased speed (Roman chariot races), an operating brake was developed that was required to stop the vehicle from rolling when stationary, as well as for decelerations, when driving downhill or when stopping.

2 TYPES OF BRAKES

Generally, the brake is a device to slow down the vehicle:

- until it stops
- to secure the parking vehicle
- to control (reduce) the speed of the vehicle.

During the brake application, kinetic energy is converted to heat. The parking brake prevents rolling of the vehicle. When operating during standstill, this brake works without wear or heating. It can be executed as a friction brake or also as a locking device like in an automatic transmission gear box.

According to the physical work principle, vehicle brakes can be subdivided into three main groups:

- mechanical friction brake
  by solid state friction (wear) of two rigid bodies.
- hydrodynamic brake
  - by inner friction of a medium: water, oil, and fluid friction.
- electrodynamic brake
  - by magnetic effect of the electrical field: generator and eddy brake.

2.1 Friction Brake

In this contribution, only friction brakes are considered, which are working on the principle of mechanical solid friction. They can be used as a parking brake as well as a dynamic brake (stop the car, control the speed). In the case of dynamic braking, it works strictly as a wearing brake. Disk, drum, belt, or cable brakes can be used. Here, only disk and drum brakes will be dealt with.

2.2 Disk brakes

The disk brakes usually consist of a disk (gray iron, cast aluminum, carbon fiber pellets, and ceramic reinforced with carbon fibers), which is mounted to the shaft. The disk brake is predominantly made as a partial disk brake. The friction lining only covers a sector of the disk ring, and not the entire perimeter. The brake calliper encloses the disk such as a gripper. Assisted by one or several hydraulically operated pistons, steel or aluminum segments with applied friction material that are axially movable are pressed against the disk. Very high surface press forces are possible due to the opposing tensioning force (Figure 1).

With increasing wear of the friction linings, the pistons extend further and readjust the wear path. The brake clearance, which must be guaranteed after each operation, remains constant. Additional fluid volume required for this flows from a reservoir. Protective caps on pistons and guide bolts prevent the brake from seizing due to dirt and corrosion. Despite high local temperatures, few fluctuations in adhesion factor occur, compared to drum brakes. This will be shown later.

Disk brakes have good cooling from air flow, as long as the installation space in the vehicle permits air access. With high thermal load, inner ventilated and/or perforated disks are used. The heat expansion is not critical, as it works toward the tension forces. Cracks in the brake disk are considered as production or design faults.



Figure 1. Disk brake. (Reproduced with permission from TRW.)

The disk brake provides a simple and easy accessible device for maintenance and checking for wear. The design of the disk brake is further explained in the example of the brake Colette, a frequently used floating calliper brake. The brake (Figure 2) consists of a housing part, which can be adjusted using two guidance bolts, and a stationary support. The brake linings are hold on the support (4) in this type of brake, and they are held in position with springs (1). These springs mainly prevent noisy rattles and clicks (2 and 3). The lining material that is glued to the back plate is between 10 and 15 mm thick and should be replaced when there is residual thickness of 2 mm. Between the lining back plate and the piston, most disk brakes have a noise damping panel.

The floating calliper brake has proved to be well-suited for small installation spaces in vehicles with negative king pin offset. The housing only contains (one up to three) pistons on the disk side. When operating, the hydraulic pressure works on the pistons and the housing evenly. The piston moves out and presses the neighboring lining against the disk. At the same time, the floating housing, perpendicularly arranged to the disk, is moved in the other direction, so that the second lining is pulled against the other disk side. The holding of the brake must be rigid, which limits the realizable disk diameter, as the rim determines the maximum available space. The movement of the floating calliper is arranged by the guide bolts. The bolt on the disk inlet side positions the housing to the support, whereas the second provides a fixed seat with its rubber coating and prevents rattling in the bore while driving. The relationship of lining size to design size is very high compared to other disk brake designs.

To change the lining (example component, Figure 3), merely release the hex screw. Afterward, the housing can be folded away and the linings can be pulled out away from the disk (upward). In addition, the support also carries the linings in circumferential direction, so that no circumferential forces must be transferred via the guide bolts.

New types of designs have a lining without retainer springs on the upper lining edge. One Nirosta spring sheet on each side position the lining in nonbraking condition and increases its lateral mobility to the disk. The response time of the brake improves, and the residual friction torque of the linings are reduced when the brakes are not in operation.

The movements of the lining parallel to the disk are highly limited. A groove together with the spring sheet on both ends prevents unscrewing while braking. This also minimizes the lining travel to the disk for braking and clicks (as well as squeaking) no longer occur when changing

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

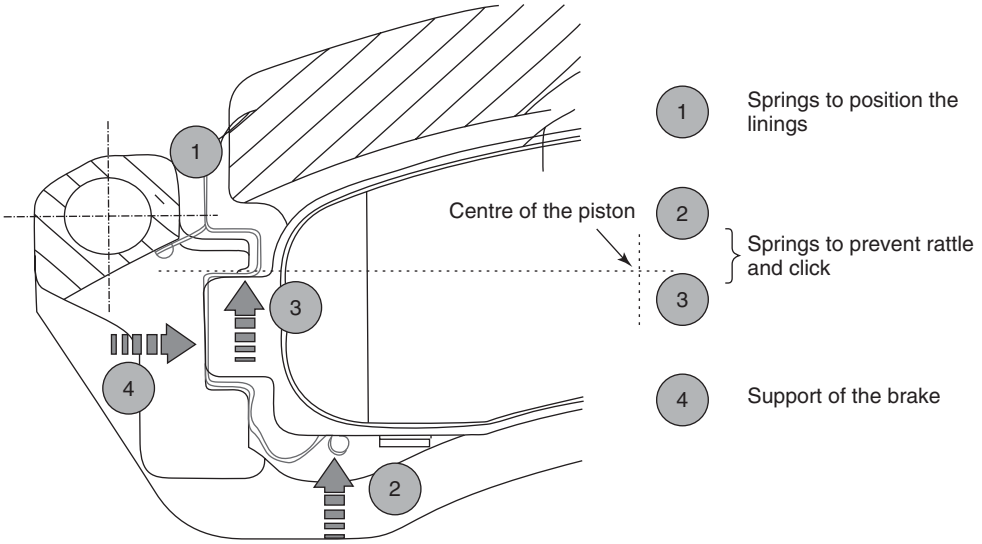


Figure 2. Floating caliper brakes Colette II. (Reproduced with permission from TRW.)

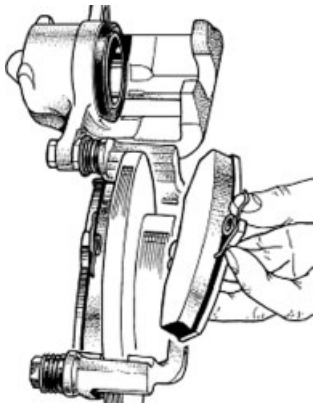


Figure 3. Changing linings Colette I. (Reproduced with permission from TRW.)



Figure 4. Asymmetrical disk brake. (Reproduced with permission from TRW.)

direction of rotation. On uneven roads there is no rattle as well.

To improve lining wear, the latest brakes (Figure 4) no longer have symmetrical designs. The piston is offset opposite the lining back plate, which also identifies the milling groove necessary on the wheel side for processing the housing. The brake plate is only reinforced on the outlet side, which is also where most of the forces are to be transferred. This leads to additional weight reduction.

The brake lining is eccentrically placed on the back plate corresponding to the piston axle. Opposite the back plate, the piston is also radially offset, using a partial sheet that is simultaneously a damping plate, or the piston is displaced opposite the lining's center of gravity. This sheet, in bare or rubberized execution, can be placed or glued to the back plate. This dampens lining vibrations that result during the braking process in the friction path. Minimizing sounds is becoming more and more significant in modern brakes. For this reason, special design measures are even being introduced for damping noise (dynamic vibration absorber). The friction process is the trigger for all brake noises. Each wear optimization that affects the friction process also leads to noise reduction, as the origins of specific noises are prevented.

The frame brake (Figure 5) is a significantly more rigid execution of a floating calliper brake. Design and function are similar to a floating calliper brake. In the RC 5424/13 design with a high performance spectrum, the housing was shifted to the outside and the carrier with linings to the inside. This makes the housing bridge especially

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

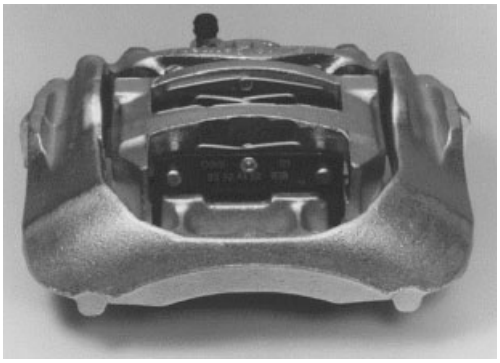


Figure 5. Frame brake. (Reproduced with permission from TRW.)

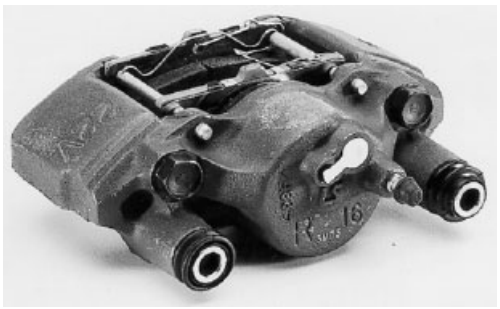


Figure 6. Semi-integrated disk brake. (Reproduced with permission from TRW.)

rigid. The tension forces on the disk are transferred in the plane of force (piston axle), so that there is no radial bending up. However, the benefit of a thinner bridge is offset by width, and the disk diameter can be designed to be larger. The cross member must be very rigid (solid), which increases the weight. The linings are carried by the support in circumferential direction and they are positioned with central springs in the center of the lining. The stable lining pressure is radially and tangentially well supported. Despite the large size, the lining size is somewhat small. This brake is installed on the front axle in larger vehicles like, for example, the VW Transporter with permitted total weight up to 2.8 t.

The FIS16 brake (Figure 6) is a semi-integrated floating frame calliper brake of the middle performance spectrum (semiabutment). Semi-integrated means that only the outer lining is supported in the housing and the housing guide bolts must transfer these circumferential forces. In contrast, the inner lining is directly supported by an integrated steering knuckle carrier. Fully integrated brakes with support of both linings in the housing are also in series.

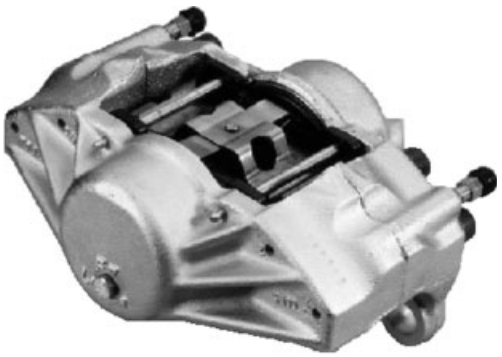


Figure 7. Two-piston-fixed caliper brake. (Reproduced with permission from TRW.)

Today, all brakes are provided with surface protection against corrosion. This can consist of galvanizing followed by yellow chromate conversion coating, of special paint coats or similar.

Another design is the fixed calliper brake (Figure 7). Instead of a sliding housing, a stationary housing is used. There is no support. The tension forces are applied by the opposite pistons. Fixed calliper brakes are built with up to eight pistons per side. All pistons are hydraulically connected with each other. For this, it is necessary to bring brake fluid through the bridge over the disk to the other side of the calliper. If there is longer braking (e.g., in the mountains), or with many successive hard stops, the hot brake disk can heat up the brake fluid in the bridge up to steam formation. As steam is compressible, the brake system loses effect heavily (fluid fading).

The design, which is extremely solid because of the fixed connection between housing halves, allows very large tension forces with minimal housing deformation (high performance spectrum). The response behavior is especially good due to this. The two or three hydraulically connected pistons per lining are individually optimally pressed. The lining wear is thus very even, (no diagonal, tangential, or difference wear). The noise production is also decreased. The clearance when releasing the brake is the same for both linings (inner and outer) and is also quickly deployed. The difference wear is also less when the brakes are not applied. Application noises do not occur.

It is not possible to install fixed calliper brakes for vehicles with negative kin pin offset because of lack of space. The ball joints of the wheel suspension and the brake disk slide too closely together or work on the same spot in the wheel disk.

A completely different version of the disk brake is the wrap-around brake. It has the great advantage of moving the friction process to the largest possible radius, and thus

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to achieve a higher brake torque than outer disk brakes. It is a high performance brake. However, as there are disadvantages from difficulties in manufacturing and also in assembly and maintenance (due to poor accessibility), this brake may not be considered in this context. As a summary, this type of disk brake is very expensive and technically too complex for use in series production.

2.3 Drum brakes

The drum brake (Figures 8–10) is predominantly designed as an inner shoe brake. By spreading apart radially, the brake shoes are pressed against the inner side of the rotating brake drum. The shoes can be housed on the carrier in different ways: we differentiate between brake shoes with single or double pivot, members or friction pads. The drum brake shown here contains friction pads with glued-on linings. Earlier drum brakes exclusively had riveted linings.

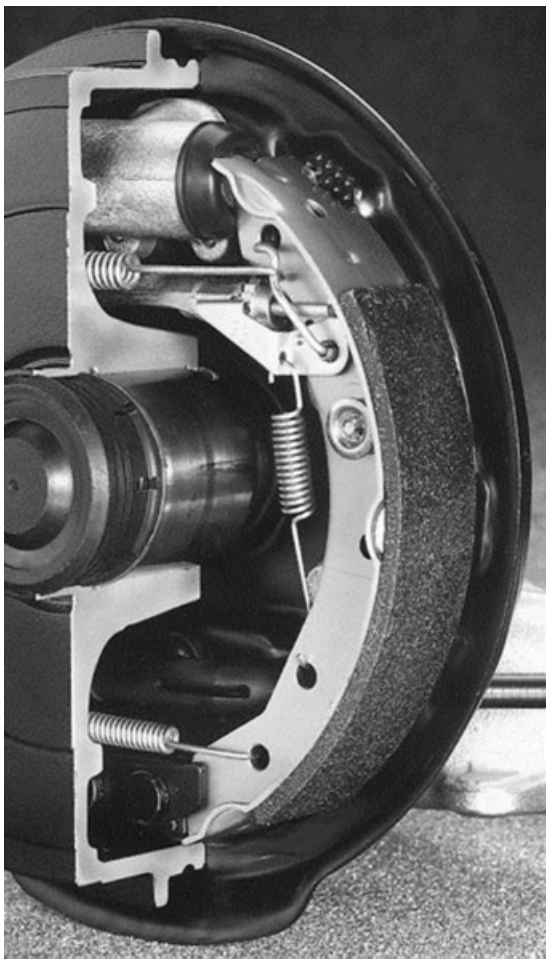


Figure 8. Drum brake. (Reproduced with permission from TRW.)



Figure 9. Simplex drum brake. (Reproduced with permission from TRW.)

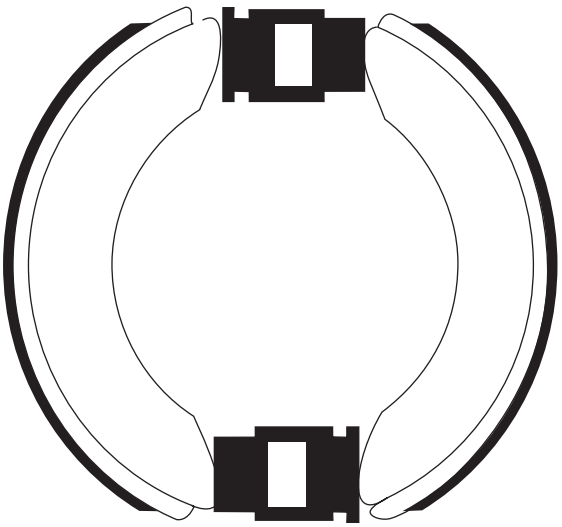


Figure 10. Duo duplex brake. (Reproduced with permission from TRW.)

The linings generally have a very long service life; despite this, a resetting device is required for wear, because of the longer application path. Drum wear is also minimal. Different designs are dependent on the mounting and operation of the brake shoes. The parking brake function is easily realized with an additional mechanical operation. The brake factor is high due to self-reinforcing effect; however, it can vary significantly because of the adhesion factor fluctuations during operation. Altogether, the drum brake is a cost-effective design.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Today, the simplex brake is the most commonly used drum brake. Figure 9 shows the light-weight version, with the anchor plate, the support and the brake wheel cylinder made of aluminum. When the drum turns counterclockwise, the left brake shoe becomes the leading shoe with self-reinforcing effect (accumulating) and the right becomes the trailing shoe with reduction (self-attenuation). By a hand brake lever and a pressure rod, that simultaneously contain the automatic resetting device, the drum brake can be operated mechanically using a cable pull (parking brake). The upper and lower return springs pull the linings away from the drum after releasing the brake and provide clearance again. The retaining spring provides the correct position of the brake shoes along with the lower support.

During operation, high adhesion factor fluctuations can occur, which effects the brake characteristic C*.

The longer application path of the shoes due to lining wear is automatically reset. With the aid of an adjusting lever, a pinion is operated over the application path, which steadily spreads the brake shoes apart. The necessary clearance is always maintained.

If there is a longer application process, the drum expands as a result of the heat. If there is an automatic clearance reset now, the brake would jam after cooling off. To prevent this, a thermal-bimetal (thermoclip) compensates for the longer application path because of temperatures, and the resetting is suspended.

The cooling application of the drum brake, especially those of the brake shoes, are not as good as with a disk brake. The thermal expanding of the drum negatively effects the braking behavior. This results in an extension of the application path of the brake shoes, which is compensated with increased piston travel in the wheel cylinder. This causes a greater volume intake of brake fluid and with it a longer pedal travel. (If the drum is already heated, the driver first steps into emptiness at the start of operation.) The incorrect application of the shoes in the drum decreases the actual lining application length and decreases the brake lining parameter. This results in uneven brake behavior (temperature fading). For this reason, a maximum temperature of of 400°C is permitted at the drum brake during application. Depending on the manufacturer, the disk brakes are between 750 and 1000°C.

Different designs of the drum brake are dependent on the mounting and operation of the brake shoes. Here, the five most important types of drum brakes are more closely described.

The simplex brake (Figure 9) was already described. Both rotation directions have the same total brake force.

Duplex brakes have two single wheel cylinders and two leading shoes; however, they only provide a double reinforced brake force when driving forward. When driving in reverse, the brake force is significantly less (decreasing two times).

The duo duplex (Figure 10) brake is similarly built as the duplex brake. However, it is equipped with two double wheel cylinders and two leading shoes for both rotating directions. This ensures the same strong brake force each time. When changing rotation direction, usually application noises occur.

The power brake consists of a double wheel cylinder and the floating housing of two accumulating sliding shoes when driving forward due to the support force of the primary against the secondary shoe. This leads to a very high self-reinforcement, however only when driving forward. When driving in reverse, it corresponds to a simplex brake. Here, application noises also occur when changing rotation direction.

The duo servo brake is designed similar to the power brake. Dependent on the rotation direction, the two shoes are supported by a support bearing. With this, there are two leading brake shoes for both rotation directions with very high self-reinforcement. Application noises occur when changing rotation direction. This type of brake is frequently used with a mechanical expander lock as a parking brake.

2.4 C* value of brakes

The brake torque is the product of the average friction radius r_{Bm}, the piston surface A_K, of the hydraulic pressure p_{hydr} and the brake lining parameter C*. This parameter describes the relationship between the circumferential force and the tension force, which is dependent on the design and the adhesion factor (Figure 11). As the braking torque should, if possible, remain the same during the brake process, a constant lining-friction parameter (linig adhesion factor) is required.

This requirement is best fulfilled by disk brakes. Here, the C* value is equal to twice the adhesion factor. The brake lining parameter depends on the self-reinforcing factor of the drum brake (geometry). It reacts very sensitively against adhesion factor fluctuations. Figure 11 shows the dependence of the brake lining parameter C* on the lining adhesion factor with different brake designs. As adhesion factor fluctuations are unavoidable, and always occur, but the torque created by the brake should be as even as possible, only the simplex brake is still considered for practical application in modern cars, except for the disk brake. With correct design and moderate self-reinforcement, the simplex brake only differs from the disk brake a little bit. The remaining drum brakes with high self-reinforcements are unsuitable because of the difficult dosing, but were used in the past. The advantage has been that no brake booster

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

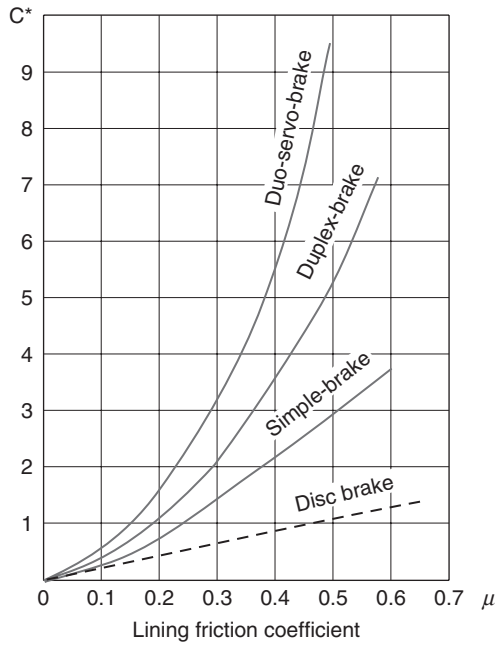


Figure 11. C* values of different brakes. (Reproduced with permission from TRW.)

have been necessary. Today, the brakes with the minimal C* values require additional reinforcement of the driver's foot force through a separate brake booster.

2.5 Leightweight design

In addition to the main task of light-weight design, to reduce the vehicle weight in total and to decrease the fuel consumption, there is especially the requirement to reduce the unsuspended masses on the axles of a vehicle. This includes the brakes and the brake disks as significant parts. Assuming a proved design, the parts containing iron with high density (e.g., steel and cast parts) are replaced with other materials. As a 1 : 1 substitution is usually not possible because of strength, wear, or thermal reasons, the components must be modified in design, corresponding to the requirements of the new material. This often leads to compromises, and the maximum possible weight reductions (with Al 62%) are not reached, as the modified components have more volume.

The following materials are used or under development as light-weight materials:

- Aluminum in various combinations, predominantly as ALMMC (aluminum metal matrix composite), a composite that is reinforced with 20–50 vol% silicon-carbide. This improves the mechanical characteristics, and the melting point increases.
- Ceramics as porcelain from silicate, exclusively used for pistons. Heat-resistant technical ceramics made of silicon-nitride is too expensive by 10-fold.
- Plastic, made of venyl and epoxy resin with good mechanical and thermal characteristics.
- In addition, pistons made of Nirosta sheet metal or anodized aluminum and magnesium are under development.

The modifications due to the new materials are, for example, as follows:

- The brake disk, previously manufactured from gray iron, has a larger thickness in the friction circle so it does not fail thermally. However, in many applications, this material cannot be used, as the thermal stresses overall are too high. Application area is the rear axle.
- On the floating calliper housing, previously made of spherical graphite iron, the bridge strength was increased by ribs when made of aluminum, to keep the expansion during operation to a minimum.
- The wall of the piston made of plastic, ceramics, or aluminum is also thicker than in the previous steel piston, to keep the pressures at 160 bar.
- The weight of the lining rear plate is additionally reduced by bulges. In the former steel plate, there were still reserves here. In addition, the connection with the lining material is improved and the form closure is improved.

The high thermal stress of the brake with the aluminum brake disk is opposite the low melting point and less strength of the material. Therefore, only the use of fiber or particle-reinforced alloying such as ALMMC is possible. The maximum operating temperature of 450°C must be absolutely maintained; above that, the material loses its strength and already reaches its melting point at 550°C. Owing to the lower density, with the same volume, ALMMC only has a very low absolute heat storage ability compared to gray iron, despite the higher specific heat capacity.

The time history of the temperature during a stop, calculated using the finite element method, indicates the distribution of energy within the disk (Figure 12). The temperature in the friction path center increases as the first and quickest; in the disk head, it does not increase until the end or after the stop. The heat released through convection during the stop (radiation or heat conduction into the disk head) can be ignored, independent of material, as during the highest case it only amounts to 5% of the energy conducted in only a few seconds during braking. Therefore, the brake

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

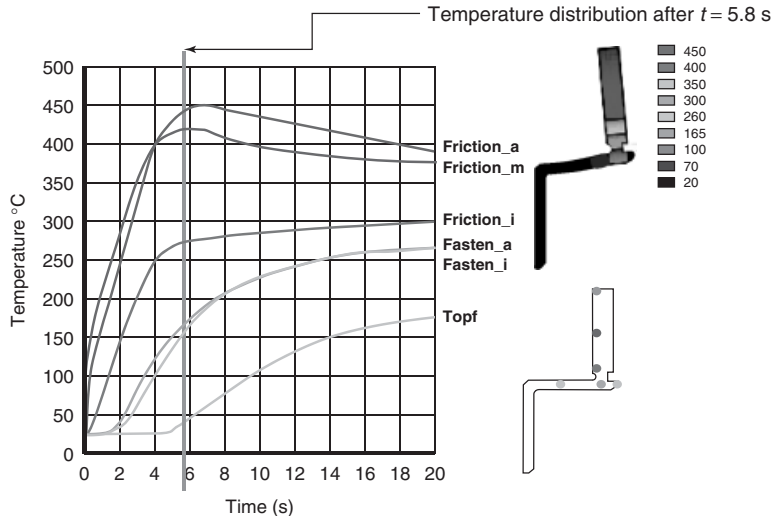


Figure 12. Heating of the disk during a brake process. (Reproduced with permission from TRW.)

disk during a stop is to be considered purely as a heat accumulator, which gets filled.

After braking, in the cooling phase, a material with good heat conduction brings advantages. The temperature profile within the disk circle is better balanced, which causes a quick convective heat release and more convenient heat conduction into the head. The heat conduction into the head must be limited to prevent excessive heating of the wheel bearing; the grease would leak out in fluid form.

Owing to the temperature distribution, different expansions occur, which leads to deformations and tensions of the disk. Figure 12 shows the deformations in a high performance test with successive stops, a vehicle is braked several times in a row with a high deceleration of 70% of gravity and is then accelerated as fast as possible. The first five stops are performed from 90% of the maximum speed to 80 km/h and the last stop to 0 km/h.

Comparing gray iron to an iso-volumetric ALMMC disk, it becomes obvious that the aluminum disk is quickly heat saturated. The temperature in the gray iron disk continues to increase. Owing to the lower storage capacity of the ALMMC, the temperature difference at the same stops is larger. Owing to the higher heat conductivity of the ALMMC, the cooling between stops occurs better than with gray iron.

The temperature of the disk head near the wheel always increases after a stop, an indication that the heat conductivity during a stop can be ignored. The maximum temperature of 180°C clearly shows that the neighboring parts, such as wheel bearing or rim also experience heating, which also occurs more quickly especially with longer braking, and therefore prevails longer than with the gray iron disk.

Test bench tests with ALMMC disks have shown that the thermal loads together with the mechanical loads can be managed with aluminum. The strength of such disks, especially against abrasion, is ensured by silicon carbide particles. A black lining layer forms on the disk surface, so that the lining material rubs against itself. The brake wear is also heavily reduced by this. There is also less wear on the brake linings that are specially developed for aluminum disks.

The tests performed for rear wheel brakes (high performance and fading successive stop brakes as well as the simulation of a slow alpine descent: long-term braking for 45 min) were successful. When exceeding the performance (in this case with double vehicle mass), the disk does not completely fail. On the surface, the melting temperature is reached and deep ridges result. The protective coating is destroyed. An iso-volumetric exchange ALMMC for gray iron will not be possible in most cases. This material can only be used when still additional installation space will be available. Highly loaded brake disks, as are necessary for high powered heavy vehicles on a front axle, cannot be made of ALMMC with the same dimensions.

Brake disks in sandwich construction are under development for use in highly loaded front axle brakes. To decrease the high thermal and wear-inflicted loads of the disk surface, an aluminum brake is protected at the corresponding spots with resistant layers. This principle is already used in brake drums, in which the friction surface is embedded in the inside drum with a gray iron ring. Thus, known friction pairs can be used. Aluminum is then usable without particle reinforcement and is also easier to recycle. The achievable weight reduction, however, is less.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

By half-empirical iterative layer calculation, a thermotechnical pre-design was performed, and the transient temperature profiles for all applications were determined. Such a sandwich disk is only safe for operation, if the maximum temperature in the contact surfaces of both materials is under the maximum allowable temperature of the aluminum, and if between both materials there is no thermal contact resistance. The gray iron layer must not overheat. During cooling, a part of the heat can flow in the groove; however, for convection, the protective layers are insulating. Owing to thermal tensions and aging effects, chipping of the protective layers can occur. For most of the planned application areas, the thermal load is at a critical range. However, the sandwich disk is not yet finished in development. The long-time effects through time, temperature, and corrosion aging have not yet been researched. For this, to evaluate the multiparametric operating load collectives, the correlations, and the specific test procedures, tests must still be developed for.

Increasing engine power and increasing cost pressure have led to a decline in light-weight construction in recent decade. The tire diameters grew as fast as the brake disk diameters. More lining volume for higher operating life required broader partial disk segments and thus broader and heavier brakes. The CO_2 laws and increased efforts to reduce fuel consumption will turn this trend back. Light-weight construction is gaining significance again, and people are willing to pay for it.

2.6 Brake circuit design

To prevent the loss of brake power by leaks in the brake system, the system is divided into two separate circuits for safety reasons. From all the possibilities, only the best-known five brake circuit divisions are presented here:

TT. Simple two-wheel brake system (front, rear axle division or also named: black, white division). Each circuit supplies the brakes on one axle. This is a simple system design, there are no special requirements for the axle design—the typical division for upper middle class and upper class cars. In case of a failure in the rear axle circuit: only small brake performance derease, in case of a failure in the front axle: large decrease in deceleration, as the rear brakes are not very powerful. There will not be a moment around the vertical car axis.

X. Diagonal two-wheel brake system (diagonal distribution) per circuit. One front wheel and one opposing rear wheel are braked by one circuit. This is also a simple design. If there is a failure in one circuit, this results in a moment around the vertical axis of the vehicle. In case of a cicuit failure, the brake power is always cut in half. The moment around the vertical car axis can be independently compensated by designing the axle with a negative king pin offset, otherwise the driver must countersteer. This is the usual division of small vehicles up to middle class.

HT. Two-wheel brake system. The front axle requires a four-piston fixed calliper or a two-piston floating calliper or a piston of the floating calliper is connected to a different brake circuit. If there is a circuit failure, only the half brake force can be applied to the front axle with constant brake pressure and the same piston surface.

LL. Three-wheel brake system. Each circuit effects half of the front axle and one rear wheel. In case of failure, the front axle brakes half and on the rear axle there is always one nonbraked wheel that can take the side forces.

HH. Four-wheel brake system. Here, as with the front axle, four-piston fixed callipers or two-piston floating callipers must be installed on the rear axle. Each circuit works on the half front and the half rear axle brakes. The effect on both individual circuits is identical.

The design of the brake circuits depends on, which cause of failure is significant and whether an antilock braking system (ABS) is to be installed. As the standard-production application of ABS and ESP (electronic stability program), the HH division is considered too complicated, the LL division is inconvenient. The HT division has poor emergency brake characteristics (over braking of the rear axle) if there is a failure at the half front axle. All systems with two circuits on the front axle are only convenient with good thermal design, the brake fluid must be connected over the disk to the other pistons if there are fixed callipers, which makes these easy to overheat by the hot disk (evaporation). This leads to failure of both circuits.

Heavy and power vehicles of the upper class with rear wheel drive and motor in front have at least a TT division. For front wheel drive with front motor, a vehicle is equipped with floating callipers because of the (necessary) negative king pin offset and the diagonal division (X) is used.

2.7 Parking brake

The parking brake must work independent of the hydraulic service brake. By adding a mechanical actuation, the parking brake is integrated into the service brake. The primary task of the parking brake is to hold the parked car on a slope. By law, it is required that this is still possible with a slope of 18% in front or rear position of the fully loaded vehicle (ECE in Australia 30% slope).

Parking brakes are predominantly installed on the rear axle of the vehicle, as these brakes are less heavily stressed due to the brake force distribution (which is why they

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

are also smaller) and an installation is simplified by the wheels that are not steered, in contrast to the front axle. The integration of a parking brake into a drum brake was already explained earlier. Operated by a bowden cable, the brake shoes are pressed against the drume with the aid of a lever. Integration is more difficult in disk brakes because of the smaller free space ratio. The installation of a separate smaller duo servo drum brake in the head of the brake disk (drum in disk) provides one possible solution that is exclusively used as a parking brake. Figure 13a shows an example for such a small servo drum brake. Figure 13b shows the arrangement of the drum brake in the disk brake.

Another solution is the integration of a mechanical device into the housing and the piston of a floating calliper brake. From the outside, this parking brake can only be identified because of the hand brake lever and the loop of the cable support (Figure 14a). In Figure 14b, there is the floating calliper housing with the hand brake lever and the cable support. The piston, which brings up the tension forces to the friction linings and the disk, is operated either hydraulically by the pressure of the brake fluid or mechanically by the hand brake lever using the integrated mechanism.

For hydraulic operation by the service brake, the piston is moved and pressed against the lining. The reaction force moves the housing away from the disk and presses the outer lining against the disk. The hydraulic pressure generates the tensioning forces. For mechanical operation, the main shaft is rotated, which is activated by the hand brake lever. Instead of the hydraulic pressure, a ball rolls up a ramp and spreads the piston shaft away from the housing with the piston. Both linings are pressed against the disk. In contrast to hydraulic operation, now the ball generates the tension forces within the ramp-shaped groove. The mechanical operation is independent of the hydraulic operation. However, both effect the same operation of the brake, exactly as with the drum brake.

In the inner part of the piston, there are resettings necessary for the mechanical operation. It is activated during one of each brake procedures by both operating types

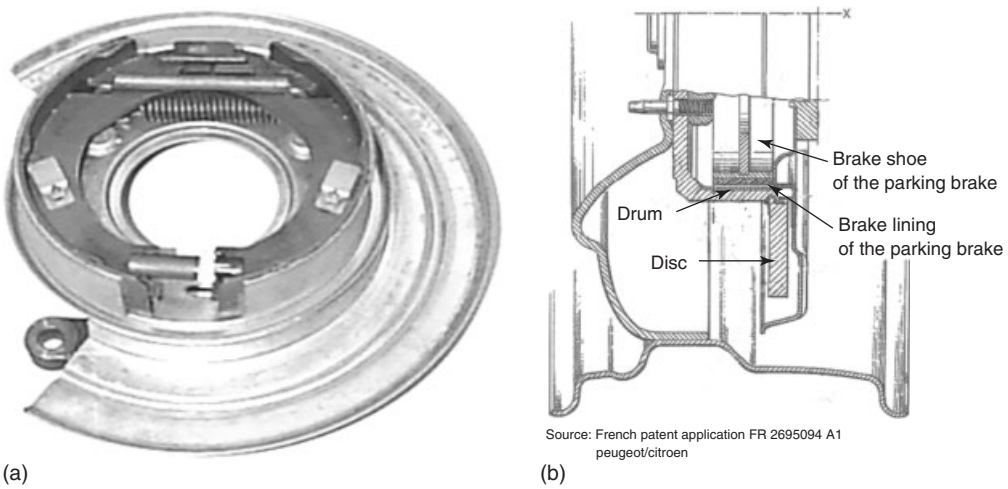


Figure 13. "Drum in disk" parking brake. (a) Drum brake as part of drum in disk. (b) Combination of drum brake and disk brake. (Reproduced with permission from TRW.)

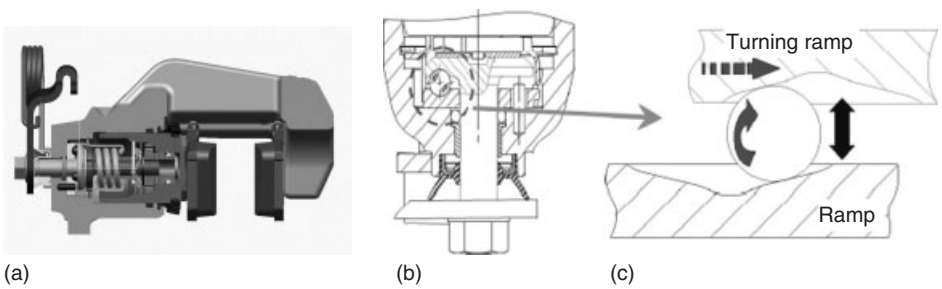


Figure 14. (a–c) "Ball in ramp" parking brake with three balls and the two ramps. (Reproduced with permission from TRW.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

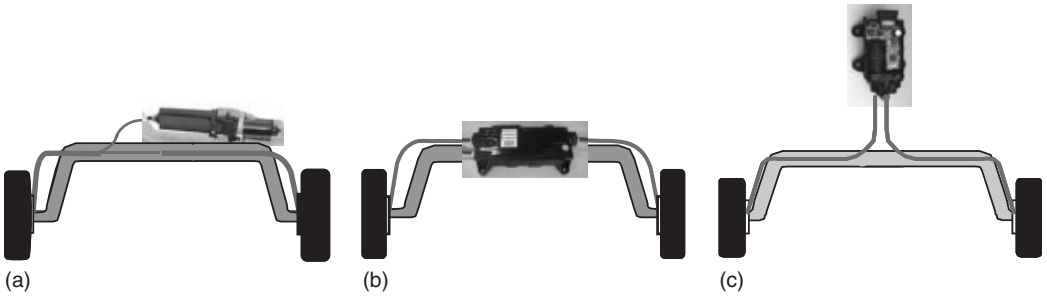


Figure 15. (a–c) Various cable systems for operating parking brakes. (Reproduced with permission from TRW.)

(hydraulic and mechanical) and extends the piston out of the housing mechanically with decreasing lining thickness (due to wear). Owing to the adjustment, there is a constant base clearance at any time between the piston and the mechanical operation when the brake is released. As the mechanically generated base clearance of the piston is larger than the hydraulically created clearance from roll back, the clearance between disk and linings is the same as with a disk brake without parking brake.

The parking brake devices work when operating the cable using hand lever, foot lever, and an electromechanical actuator. The cable puller operation is available for reaction force cable systems (a: also called *conduit systems*), transverse force cable systems (b: also called *cross-pull*), and parallel pull cable systems with force balance bars (c: also called *forward pull*) (Figure 15).

The integrated electric parking brake (EPB) replaces the cable operation using a small electric motor with transmission at the brake and the adjustment device via software. This brake is an additional step to decrease operating forces and to obtain design space in the vehicle interior. The electric operation allows the use of minimal forces, which not only physically weak and disabled people appreciate.

Instead of ball ramp and the automatic adjustment device, a spindle is inserted in the EPB, on which there is a compression nut, which—secured from rotating—is pressed against the piston head or is pulled back from there again. The motion thread is self-locking. A maximum tension force of 23 ± 2 kN can be generated with the device. With this tensioning force, a "slip stick" effect begins in the thread of the spindle. These high tension forces must be considered when developing brake disks (Figures 16 and 17).

The spindle is supported with the flange against the axial bearing on the bottom of the brake housing. The flange has an end stop that prevents the pressure nut from tensioning against the flange. This is important for manual and electric lining change and for the end of the line—calibration (calibration on assembly line).

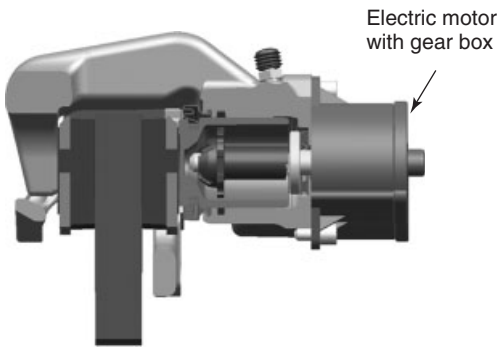
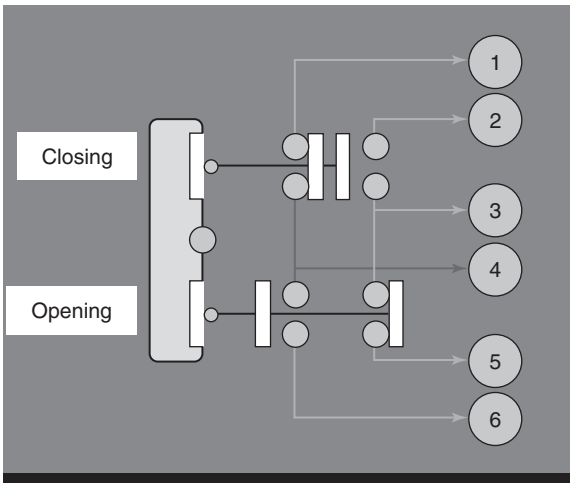


Figure 16. EPB (electric parking brake) in cross section. (Reproduced with permission from TRW.)



	1/4	1/6	2/3	2/5	3/5	4/6
Neutral	1	0	0	0	1	0
Opening	1	1	0	0	0	1
Closing	0	0	1	1	1	0

Figure 17. Wiring of the electric park brake. (Reproduced with permission from TRW.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For emergency release, the actuator can be unscrewed and the brake is opened using a Torx wrench—that is a factory solution. Generally, before an emergency release, it must be considered, how and where the vehicle will next be securely parked even if no power is available.

The variant produced by TRW is operated using dual buttons with neutral setting, mounted in vehicle longitudinal direction. The front position (press) closes, and the rear position (pull) opens. If it is not operated, the switch stays in neutral position. The contacts are redundant, each with a normal open (NO) and a normal closed (NC) contact per switch position. Using the NO contact, a capacitor is placed so that the NO connections can also be constantly monitored with a pulsed test voltage. Thus, a single fault can be securely and immediately detected.

The processor is a Star12 from Motorola with on-chip flash memory that is widely used. The motors are each switched to one FET (field effect transistor). Changing rotation direction is realized using two relays. The real flowing currents are measured with each one shunt, likewise the voltage level and polarity. The control unit (Figure 18) can have up to four end levels for direct control of pilot lights.

For ignition "on," the control unit is in active mode and continuously performs self-tests. After ignition "off" and a shut-off delay, the control falls into a power-saving sleep mode (maximum 200 μA), from which it awakens operating either by a switch or by ignition. A second processor that simultaneously executes the monitoring function controls this and independent of the main processor can lock the end levels. The "end of line" configuration and the diagnostics occur using the CAN (controller area network) bus.

For vehicles with integrated starting assistance, an additional inclination sensor can optionally be provided in the control unit. If this is installed, it can differentiate between the static and dynamic mode by evaluating the vehicle movements during the drive even with failure of the CAN velocity signals.

The software runs in a 20 ms main loop, from which all function modules are continuously addressed.

This is strictly a status machine, whereby the different modes can only be exited using predetermined events (mode control). Subroutines, such as the composition of switch commands with other system signals (demand calculator/clamp force controller) or the motor control (motor controller), are set up as status machines.

The main function is the static closing of the brake with ignition on. The switch must be operated for that. The software checks whether the ignition is switched on and whether the wheel speeds are zero. After 100 ms (5 cycles), the command is accepted and transferred to the motor control. This enables all functional motors in tension direction and monitors the tension flow. This measures the voltages. While driving over the clearance, the motor current is determined and considering the voltage, the motor temperature is estimated. Both voltage and temperature are used for the formation of a correction value for the cut-off current that is then added to the no-load current. A fully loaded vehicle on a 30% incline

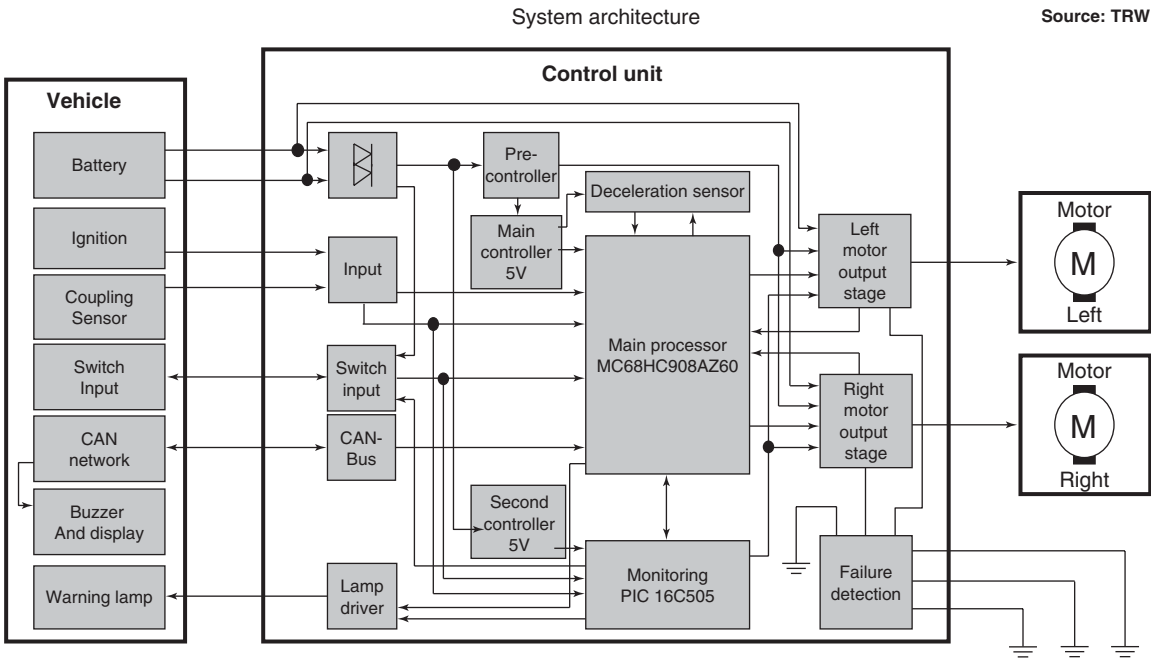


Figure 18. Block diagram of control unit. (Reproduced with permission from TRW.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is assumed to define the cut-off current. Figure 19 shows the time history of a tensioning and a release process.

During the drive, a temperature model runs for the rear brake disks to detect excessively hot disks. By repeated tensioning, the system provides the opportunity after approximately 3 min to realize a one-time (cycle-related) force super elevation beyond the nominal tension force. Rollaway detection can also trigger multiple tensions, which lead to high tension forces, especially, if there is not enough energy available because of undervoltage.

To open the parking brake, the relays are switched in release direction and the FETs are energized. As soon as the current curve has fallen to the no-load current while opening, this point is retained as the new relative zero point, and after clearance adjustment time, the motor is switched off. This time is determined by declamp-voltage and declamp-current (to estimate motor speed).

If the ignition is off, and the switch is operated, only tensioning the parking brake is possible, but not opening (child safety lock) It becomes difficult for drivers that switch off the ignition while driving downhill; wheel speed signals are no longer available from the CAN bus. A CAN follow-up function ended by the EPB ensures that the electric brake does not close until the vehicle is standing. Even if wheel speeds fail, the brake is not closed until the ignition is off and vehicle standstill has been detected.

The total software overview is shown in Figure 20.

If the service mode is activated with ignition on via the diagnostic function of the service mode, both pressure nuts move back to their end stop. Then the linings can be replaced.

The dynamic deceleration is passed to the operating brake system by the EPB using CAN message, which then executes the braking using the ESP system and takes over again with a speed <7 km/h.

Probably, the most convenient function of the EPB is the automatic release when starting. Using the integrated incline sensor, the required engine torque is determined and when reaching this threshold, the brake is opened. For manual transmission, a clutch position sensor is engaged. From the gradients of the clutch position (so the release speed), the grip of the clutch is calculated and the brake is released when there is sufficiently high torque.

The starting assistance is also important for automatic vehicles. Today, less and less creep torques already allow automatic vehicles to roll back on smaller inclines (12%) If the parking brake is not used, semiautomatic vehicles generally roll back when moving the foot from the brake pedal to the accelerator. Using the starting assistance, the rolling-back can reliably be prevented. Undesired starting is prevented by additional signals, for example, starting with the starting assistance is only permitted for a driver whose seatbelt is buckled. Manually releasing EPB is only possible

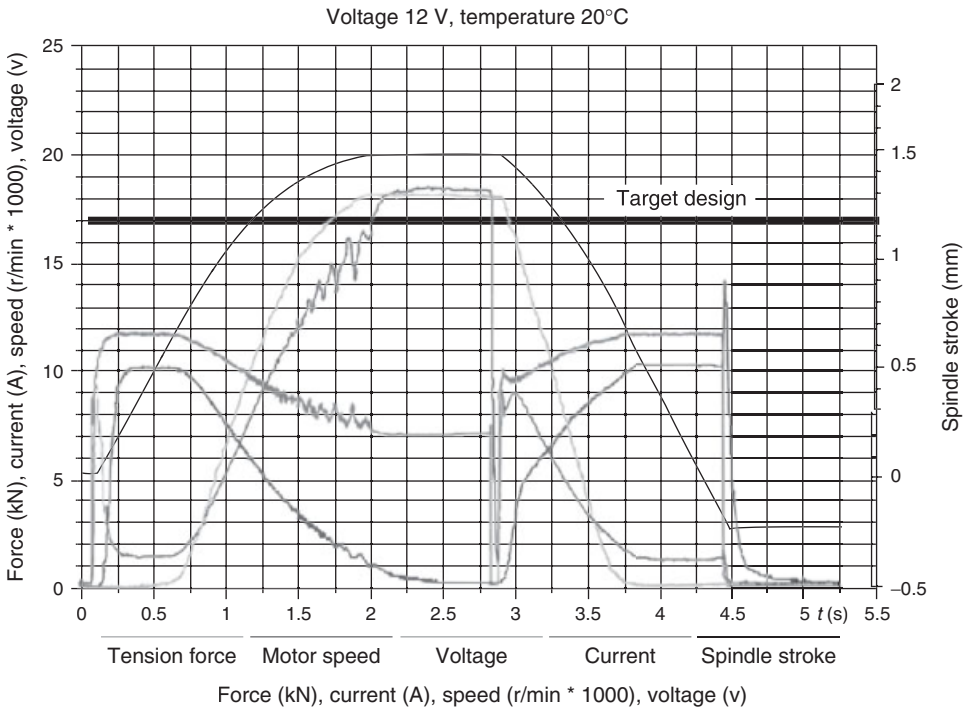


Figure 19. Tensioning and release processes of the EPB. (Reproduced with permission from TRW.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

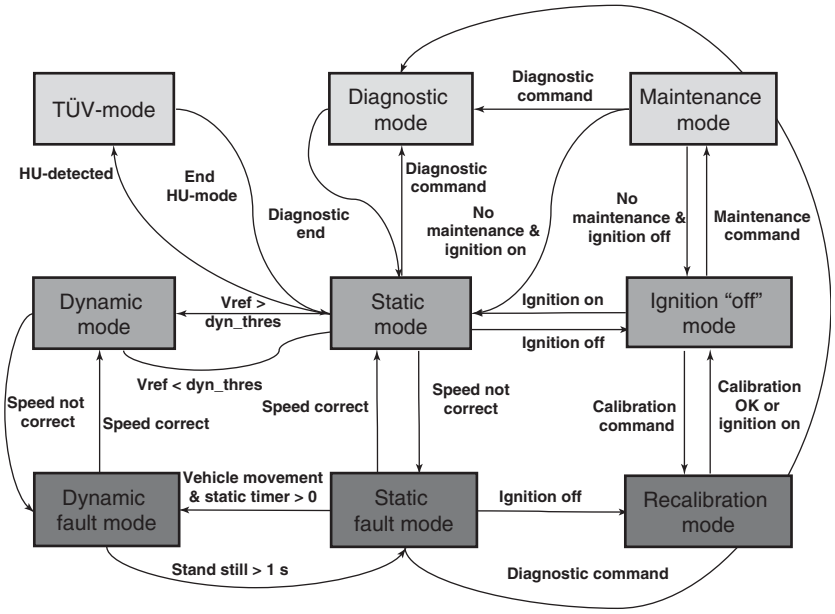


Figure 20. Software condition diagram. (Reproduced with permission from TRW.)

by pressing a button if either the brake or the accelerator is pressed down.

With the typical techniques such as continuous ROM test, RAM test, external watchdog, and second processor, the security of the control unit is ensured up to ASIL D corresponding to ISO 26262. In addition to voltage supply, the redundant operating element, the fault lamp, and the actuators are also monitored for proper function. Using a constant current source (approximately 2 A), the performance output stages are tested for switching on and off. The relays for pole change can only be tested one time when the ignition is switched on, owing to their service life.

All signals are subject to a plausibility test as much as possible before processing. If the ignition is switched on, the CAN must also be there, if the EPB motors are switched on, shortly thereafter the voltages must fit with correct polarity. If the brake opens, the current must decrease; if the brake closes, the current must increase; if the driving speed decreases heavily, there is either an incline or the brake light switch is on. Changes to the drive speed must synchronously go with the signals.

If there are missing signals, if possible, replacement parameters are used. If the vehicle speed is missing via the CAN, the corresponding filtered incline sensor signal is used to detect standstill. If terminal 15 (voltage in the system) is missing, CAN is present and notifies the switched-on ignition, this is used.

The faults are classified as follows:

1. casual faults (e.g., CAN loss): the EPB system continues to work and indicates the faults;
2. interfering faults (e.g., communication with the motor control unit interrupted): the function (here, drive away) is switched off, and the EPB system continues to work and indicates errors;
3. channel fault (e.g., channel to motor defective): the EBP system tries to open the brake with the next release command and switches the affected channel off, the other channel continues to work, and the fault is indicated;
4. system fault (e.g., processor fault): the EPB system switches off, and the fault is indicated.

With this system, the parking brake function is also electrified now. In the coming years, an integration of the electronic steering in the ESP control unit will happen—many functions and signals are "related," the safety requirements are similar.

3 THE BRAKE PROCESS

The stopping distance required during the brake process is composed of

- rolling distance: the vehicle continues to roll without braking; only air and rolling resistance are working—no motor brake torque. The amount of the rolling distance depends on

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

  - time to turn toward view
  - reaction base time
  - time to implementation
  - response time.
- threshold path: braking begins, the brake torque increases
  - threshold time.
- braking distance: the vehicle is decelerated
  - time of braking (complete or partial braking).

All distances together cover the stopping distance.

The introduction and the execution of a braking process in traffic occurs due to visual perceptions. This process is composed of three phases:

- Objective reaction summons: start of visibility of the object
- perception of the object: in general peripheral, on the edge of the field of vision of the driver, only rarely in central field of vision of the driver (time to turn toward view)
- Object fixation: Recognizing the risk that originates from the object, start of muscular reaction

The fact that conscious decisions and appropriate reactions are never made based on peripheral perceptions has fundamental significance. Conspicuous objects in the peripheral field of vision always first trigger a glance. The foveal area generally only has an opening angle of 1°. If there is more than a 0.5° deviation from the driver's field of vision, a view adjustment is made. Object fixation and a conscious reaction are only possible after the *time to turn toward view* (Figure 21).

The *reaction time* begins after turning toward the view. The reaction base time lies between the object fixation and the start of muscular reaction. It is a process with the same behavioral patterns for all individuals. The time span is independent of outer boundary conditions (e.g., brightness and weather conditions) and independent of the situation.

Afterward, with the start of muscular reaction, the *implementation time* begins. The accelerator is released; thus, the first appropriate movement of the foot happens.

The first contact with the brake pedal is the start of the application or *response time*. The brake shoes/lining are applied. The implementation and application times are short compared to the reaction time.

As the brake pressure increases, the *threshold time* begins, where the increase of the hydraulic pressure is significant for the gradients of the deceleration. Normally, the wheels of the vehicle lock before reaching the maximum pressure. In modern passenger cars, the ABS activates beforehand and prevents locking, keeps the wheels in optimal slip range and secure the maneuverability of the vehicle during the brake application.

The following is the *brake time*, which lasts as long until the entire emergency brake process ends with the vehicle at standstill. It is dependent on the output speed, the degree of deceleration, and the achievable friction values at the tires (at 250 km/h and highest possible vehicle deceleration, the process takes approximately 12 s). This time span is solely determined by the capacity of the brake system and the vehicle, assuming that the driver does not reduce the brake pressure.

In contrast, the *reaction time* depends on human movement sequences, physical–technical laws, and the abilities and conditioning of the driver. Additional factors affecting the driver are of special significance: road and weather conditions, traffic volume, distraction, high degree of information, and stress.

The brake process starts as soon as the driver identifies the reason that requires braking. From this time up to the rise of brake pressure, the vehicle practically continues to roll without brakes. At this point, the modern driver assistance systems start—a shortening in reaction time by 0.5 s reduces the probability of an accident up to 80%. The brake effect from air, rolling, or motor resistance can be ignored here, as their effect is very minimal compared to the power of full braking.

In emergency or sudden braking, the driver applies a pedal force of 250 N and moves the brake pedal for approximately 80 mm. The brake master cylinder, connected to the brake pedal, generates a hydraulic-volume shift in the direction of the brake system because of this pedal movement. Owing to the compressibility of the brake linings and the deformation of the brake callipers, the entire brake system absorbs a volume of approximately 8 ml. At the same time, a brake pressure of 80 bar occurs. This leads to a deceleration of approximately 0.9 g (90% of gravity). In panic situations, more than 500 N pedal force and more than 160 bar hydraulic pressure were measured—however, the deceleration did not increase further from this.

The deceleration, with which an average driver brakes, lies in the range 0.01–0.4 g. Only in absolute emergency or exceptional situations are up to 0.8 g achieved. During their entire driving practice, most drivers never brake above 0.5 g. However, the brake system is designed for up to 1.2 g.

4 BRAKE DYNAMICS

The control of the braking process is decisive for the availability of a vehicle in practice, to be able to ensure safety in traffic. A vehicle must be designed in a way that it does not produce unforeseen reactions even during panic

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

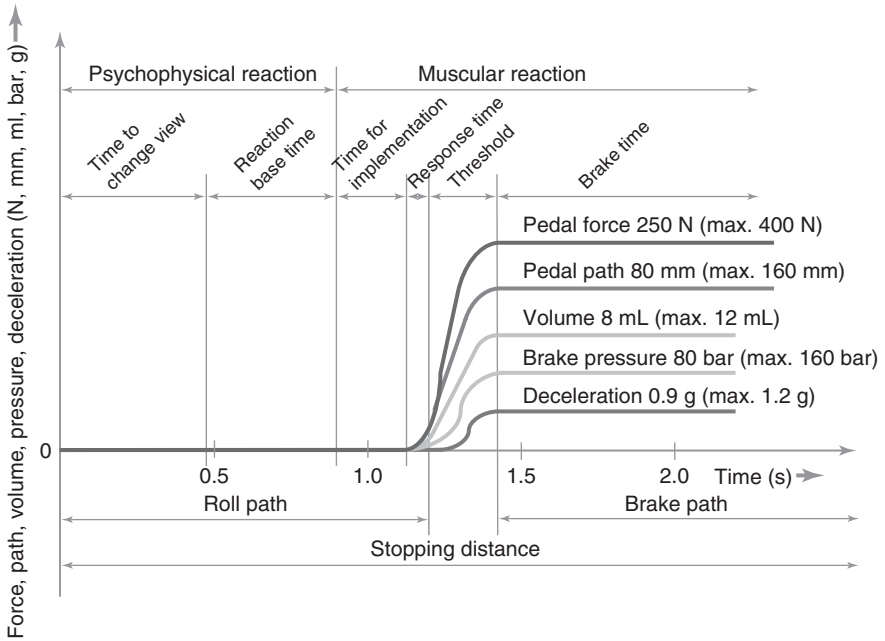


Figure 21. The brake process. (Reproduced with permission from TRW.)

braking and always behaves like the driver expects from his or her experience. Only then the vehicle is safe.

The knowledge of the outer brake dynamics of the vehicle is of special significance for designing a brake system. This is dependent on the brake force distribution that follows from the design parameters of the vehicle. For installation or for the demand-oriented change of the brake force distribution, corresponding control devices are installed in the brake system.

The brake system must be able to brake the vehicle with maximum deceleration. The driver must be able to do this with an application of average forces. The existing limits are the contact area between the wheels and the road, which are only the size of a palm of the hand, and the adhesion coefficient of the tires of $\mu_B = 1.1$ at most. With an assumed friction coefficient of $\mu_B = 1.0$, a maximum vehicle deceleration of $1\,g$ ($=9.81\,m/s^2$) can be reached. In theory, with specific rubber, the μ-factor may even be higher (race cars).

The brake application must be stable, the vehicle must not become unstable or even worse, that the rear axle swerves. A short braking distance naturally occurs with the highest possible deceleration. The processes during braking will not have any reactions on the driver or the behavior of the vehicle (no noises, no rubbing, no steering wheel torsional vibrations, and no pulsing of the pedal). The response of the system improves with only minor volume absorption in the callipers, which also decreases the brake pedal travel (operation, roll back, and widening the brake calipers). Pedal travel going against zero is not appropriate because of ergonomic reasons.

The brake force will be well dosable, so that with light pedal operation, no strong deceleration occurs. A slight decrease in the liner friction coefficient while braking, which forces the driver to step on the pedal again, is preferred. The support by brake boosters should allow every driver to perform locked braking with average use of force in emergencies. On the other hand, the increase must not lead to poor dosing with minimal deceleration. The booster devices are realized with progressive, degressive, or linear characteristics, depending on the manufacturer. In addition, all brake actions will never generate noise (inside and outside of the vehicle) from the liner and the disk wear. Traffic light systems with signals for blind people had to be changed from beeping to clicking because there was confusion with brake noises.

This especially includes the following:

- rubbing: vibrations due to speed fluctuations from differences in thickness of the brake disk;
- squeaking, chirping, and muh: average to high frequency noise or friction fluctuations;
- rattling and clicking: from the moveability of the linings, with pot holes, when applying the linings.

For the quality of a vehicle in practice, the stability is decisive in case of emergency braking. Only with sufficient

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

brake stability, the vehicle is prevented from skidding, and the steerability is ensured with good cornering forces. Regardless of size, type of motor, or design of any vehicle, the rear wheels may not lock before the front wheels lock (legal requirements are setting a limit with minimum 8 m/s^2 deceleration). When the front wheels lock, the vehicle can no longer be steered; however, it stays in its track. Because modern, electronically controled brake balance no longer has these mechanically secure fallbacks, only partial brakes are allowed for (displayed) failure.

The brakes (brake linings) may not lead to strong friction coefficient changes (high temperature fading), as in addition to the loss in brake power, especially a change in the brake balance by fading of front to rear axle occurs, which can endanger the stability of the vehicle. The same can happen because of evaporation of brake fluid (fluid fading).

The tire characteristics must also be adapted to the vehicle. A vehicle that stably brakes with approved tires (first the front axle locks) can become unstable during emergency braking by installing tires with higher maximum adhesion coefficient on the front axle. Then the rear locks first. If the rear axle skids due to premature locking of the wheels, it is impossible to react to this using steering movements. The vehicle immediately goes out of control (Figure 22).

To calculate the vehicle behavior, a rigid two-wheel model is used, and the method with minor disturbance is applied. In doing so, the vehicle is to be considered a rigid body, deformation of the axles or the wheel suspensions are ignored. Owing to a disturbance, the vehicle has the sideslip angle ß between the longitudinal axis of the vehicle and the driving direction. This driving condition generates sideslip angles at the front and the rear wheels (same amount like ß) and then side forces are acting between the road and the tires. The masses and the lateral force affect the vehicle's center of gravity. These counteract the deceleration forces and the side forces on front and rear axles.

The yaw angle or the yaw acceleration can be determined from the balance of forces in x- and y-directions and the moment balance. A positive value means an increase in the sideslip angle. This can become unstable, finally a skidding of the vehicle may happen. A reduction of the sideslip angle points toward stability. The vehicle body will returned to the driving direction. The smaller the moment of inertia J_z, the faster the vehicle turns (e.g., when the engine is in the center of the vehicle).

By relating the brake forces to the vehicle mass, the yaw acceleration in dependence of the deceleration can be determined. The simulation results show immediately under what conditions the vehicle becomes unstable (for instance in Figure 23: 70% g). The additional load of the vehicle is also of great influence, as not only the overall weight, but also the brake balance changes. The behavior of the rigid two-wheel model with ideal tires is a straight line (neutral steering behavior).

For the evaluation of a vehicle, the deviation from the straight line is decisive.

At a deceleration of more than 80% of g, the empty vehicle begins to skid (in the example of Figure 23). The yaw acceleration is positive and increases extremely

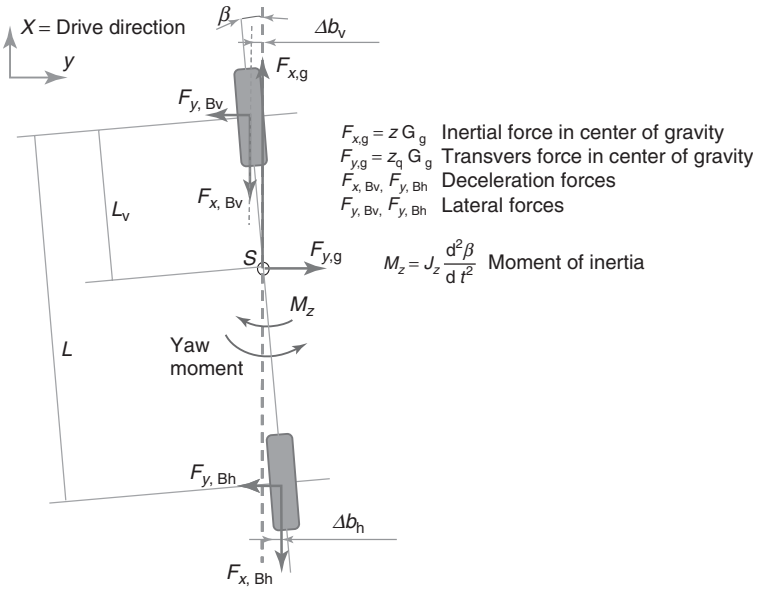


Figure 22. Simple calculation model. (Reproduced with permission from TRW.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

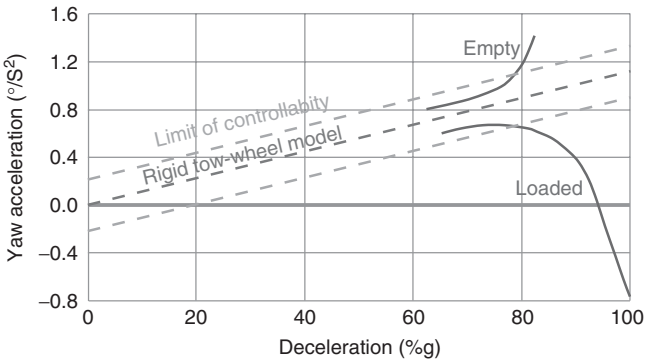


Figure 23. Vehicle stability model.

quickly. A yaw acceleration of $1^\circ/s^2$ corresponds to a yaw angle of approximately $57^\circ/s$ after 1 s and an increase of the angle by 30°. A normal driver cannot handle this. An acceptable value for yaw acceleration is $0.25^\circ/s^2$. Then, the braked vehicle can still be stabilzed by countersteering.

In real vehicles, the elasto-kinematic design of suspensions and stabilizers delivers inherently stable behavior. For the control of the vehicle by a normal driver, the brake behavior in case of circuit failure or with differing low adhesion surfaces (right/left, μ—split) is decisive. At standstill, the weight force in the center of gravity and the wheel loads at the front and the rear axles are acting on the car. The center of gravity has the distance l_1 to the front axle. The wheel base is indicated as I, and the hight of the center of gravity above the road is named h.

During the brake process, the mass forces zG_g (equivalent to vehicle mass multiplied with the deceleration) and the brake forces are added to the front and rear axles. This assumes that neither the position of the vehicle to the road nor its center of gravity change.

The dynamic balance in driving direction leads to the deceleration factor z ($z = a_x/g$ with a_x as longitudinal deceleration) . As a result of the height h of the center of gravity and the inertial force in the center, a dynamic axle load distribution $\pm\Delta G$ results from the rear axle to the front axle. The equations for dynamic axle load arise from the equilibrium of moments around the wheel contact points of the front or rear axles. These equations can be found in Kane (2008).

Assuming the same utilization of traction (that means the same μ) on the front and the rear axles, the equations can be used to generate a diagram that shows the ideal distribution of the brake forces (means: the same friction coefficient at the front and the rear axles during deceleration). This is a diagram (Figure 24) with the brake force related to the total weight, at the front axle as abscissa and at the rear axle as ordinate.

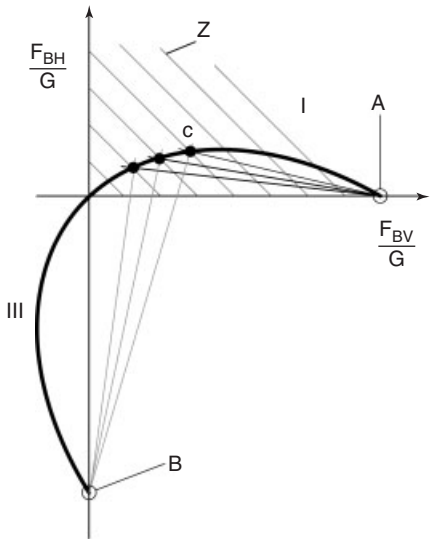


Figure 24. The brake force distribution diagram.

On this parabolic curve, the ideal case of the same adhesion utilization on front and back axles is always realized. It is solely described by the data: center of gravity height, center of gravity position between the axles, total vehicle mass, and wheel base, and it is existing in the first (braking) and third (acceleration) quadrants of the coordinate system.

The deceleration factor z can be drawn in as lines of constant braking at an angle of 45°. On the parabola of the ideal axle force distribution, $\mu = z$ always applies. The lines of same coefficients of adhesion of the tires are also straight lines that go through the "point of ideal operation" (point C).

There are two useful physical explanations for no brake forces at the rear axle:

1. $z = 0$, there is no braking that means no deceleration (coordinate origin).
2. $F_{BH}/G = 0$, the rear axle is completely released (point A).

No brake forces at the front axle can also be explained: The same applies to acceleration, here z is negative

1. $z = 0$, there is no braking that means no deceleration (coordinate origin).
2. $F_{BV}/G = 0$, the front axle is completely released (point B). This means that the car is accelerated (not braked) and therefore the related deceleration z is negative.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The deceleration factor z must be drawn in Figure 24 to be able to plot the green lines (starting in B) and the red lines (starting in A) of constant friction adhesion between tires and road surface. All red straight lines of constant friction adhesion between rear axle and road must cross point A (lifting the rear axle). There is no more friction at the rear axle in this point. Point B is crossed accordingly by straight lines of constant friction at the front axle. During acceleration, the front axle will be lifted under the conditions in Point B.

In point C, the straight lines of friction utilization meet each other. The parabola defines the ideal brake distribution that means for instance the full utilization of friction at the rear axle (may be 0.4) and at the front axle (may also be 0.4). The parabola is valid for $z = \mu$.

Practical statement: The installed brake force distribution in the first quadrant counts as stable, in case the installed distribution line is below the ideal distribution until $z = 1$ (by legal requirements $z = 0.8$). If a friction coefficient of 0.4 is assumed (the same value at front and rear axles), then the brake force P can increase along the line of the installed brake force distribution until the line crosses the green line of the friction utilization of the front axle (P_1 in Figure 25). The front axle locks at this point and the deceleration can be seen (Figure 25).

In case the brake pressure is still increased, the brake force at the front axle is no longer increasing. The braking point is going upward along the green line of the constant friction utilization of the front axle until the red line for the friction utilization of the rear axle (same value like at the front) is crossed. In Figure 25, this is shown as point P_4. Now the rear wheels are also locked. The gradient from P_1 to P_4 is not vertical, as the lines for the constant friction coefficient at the front axle are crossing in point B (comp. Figure 24).

For the evaluation of the brake behavior of a vehicle, the ideal distribution in the first quadrant is used. The installed brake force distribution is considered with respect to the ideal distribution.

As an example for a vehicle, the ideal brake force distribution is shown in Figure 26. This is like the characteristic in the first quadrand of Figure 24. This ideal distribution is also named as outer dynamics of a vehicle. The brake force distribution installed in the car can be regarded as information about the inner dynamics of the vehicle.

The brake force distribution installed provides information about the inner dynamics of the vehicle. The distribution is evaluated from the parameters of the brakes on front and rear axles and it normally represents a straight line. The intersection point of the straight line with the line of the ideal distribution indicates the deceleration (z-value) at which both axles look at the same time. As long as the line for the installed brake distribution is below the ideal one, the front axle will lock first if the road friction will not be sufficient. If the brake operates above the ideal distribution line, the rear axle will lock first and the vehicle will be immediately unstable. The intersection point of the installed distribution with the ideal brake force distribution indicates the so-called critical braking z_{kr} by which the two axles lock simultaneously.

For safety reasons, most vehicles are designed so that the critical braking (locking of all four wheels) is at $z = 0.9–1.0$. As the drivers normally avoid these high deceleration values, the vehicle normally will not become unstable by braking.

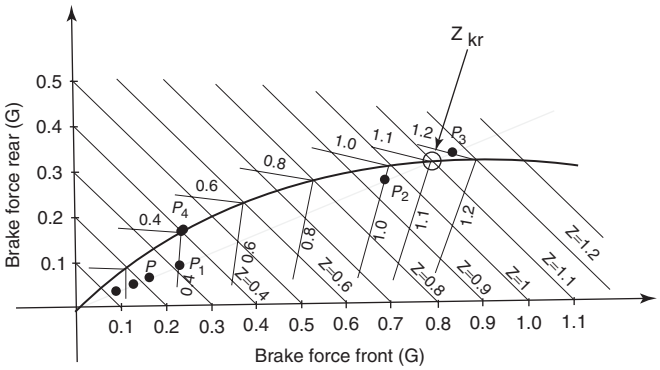


Figure 25. Brake pressure increase and locked wheels at $\mu = 0.4$.

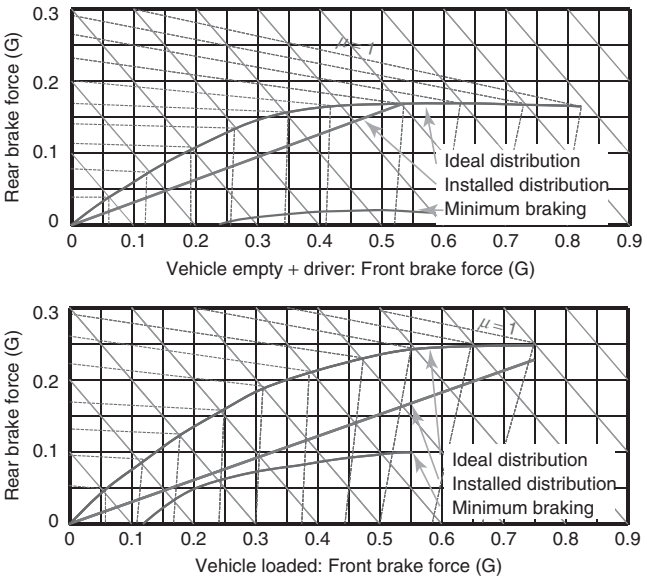


Figure 26. Brake force distribution diagram with vehicle empty + driver and vehicle loaded.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In addition, the minimum brake force distribution given by law is drawn in this diagram. With these minimal values, the requested deceleration will be reached. However, they are far too much low for practical driving.

To prevent the rear axle wheels from locking, which means to keep the installed brake force distribution below the ideal distribution, valves to reduce the brake pressure at the rear axle have already been installed for many years. These components respond to specific brake or vehicle data.

If an increasing pressure is given to the brake system via the master cylinder, the pressure-dependent valve only allows no or only a small increase in pressure to the brakes of the rear axle if a specific brake pressure has been exceeded.

The load-dependent valve works like the pressure-dependent valve. It is mainly used in vehicles with front wheel drive with front motor and low rear axle load, as the brake force distribution can significantly change there because of the vehicle payload. The spring deflection of the body is used to adjust the valve, thus the pressure switching point changes according to axle load (dynamic). In reality, the load dependence is a travel dependence (spring path dependence).

For valves with an acceleration-dependent switching point, the brake force reduction is dependent on the vehicle deceleration. High operating speeds of the brake pedal distort the mode of the valve's action. For diagonal brake circuit distribution, it is necessary to control both circuits in parallel, which is very inconvenient. These valves can also be installed between the wheels of an axle to react to lateral accelerations. For driving in extreme curves, this can prevent locking of the inner wheel.

For a vehicle without brake force control devices or ABS, the limit of brake stability is therefore achieved shortly before the locking limit of the rear wheels at the highest possible deceleration. Figure 26 shows a fixed distribution that enables such a design; it already cuts the ideal distribution with braking of 0.7.

The development is clearly moving from brake pressure control devices and going to real control systems. The antilock system that today predominantly works on all four wheels separately or on both wheels of an rear axle (select low) provides reliable brake stability with almost optimal vehicle deceleration and gives still steerability of a car, even during severe braking.

Together with an electronic brake force distribution, the braking power can ideal be used in every emergency situation, independent of load condition and adhesion coefficient of the tires, as long as the driver only demands the brake system sufficiently, generates corresponding high pressure, and also keeps the pressure while braking.

If the engine power increases with a constant vehicle weight, then generally more powerful brakes are required. If the vehicle weight increases with the same engine power, then the brakes can usually be retained from the point of view of brake force distribution and subsequent braking behavior.

5 THE FRICTION PROCESS

The friction process results from the resistance of the relative movements of two bodies. With relative movements without lubricant, dry friction or solid friction results. The amount of the resistance depends on the size, the texture, and characteristics of the contact surfaces. This includes

- the material combination (strength and deformation behavior, heat conduction, and chemical characteristics);
- the design combination (shape, surface profile, and roughness);
- the operating conditions (kind of stressing: duration, time history, pressure, temperature, and relative speed);
- the environmental medium (chemical affinity between friction partners and environmental medium in connection with the mechanical–thermal energy in friction contact);
- the type of manufacturing process;
- the degree of deformation;
- the formation of reaction layers; and
- the degree of movement (the relative speed and static and dynamic adhesion factor).

Owing to the roughness of technical surfaces, the contact of fixed bodies is always only "discrete." Compared to the nominal, in reality, there is a much smaller contact surface, as the contact only occurs on the roughness peaks. The true contact surface is formed from the sums of these partial surfaces (percentage: contact area). The result is a much higher surface load than the surface pressure formed by the nominal contact surface. This results in elastic and plastic deformations of the roughness and the surface profile, which increase with normal force. In contrast to the nominal specific surface pressure, however, the real surface pressure does not increase proportionally with the normal force. Deformations and wear lead to hardening of the surface layer, a increase of the contact ratio and cause a change in the friction characteristics (embedding).

The friction force is composed of a deformation and a shearing part: the elastic part of deformation causes vibrations of the roughness peaks affected by the friction process. These vibrations are irreversibly converted into

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

inner energy (heat). The softer body, whose roughness peaks have a lower spring stiffness, receives the largest portion of the energy. Whether it also retains and depends on the heat conductivity and the heat capacity of both bodies (heating- up).

The shear force portion of the friction force is due to adhesive force from cold welding and adhesion forces between the oxide layers. The adhesive forces are dependent on the mixture of the oxide layers that in turn are dependent on the mechanical and chemical characteristics of both surfaces, as well friction energy density. The friction energy density can locally lead to very high flash temperatures of several thousand degrees centigrade in the friction path. This stimulates the formation of oxide layers with lower hardness and lower shear resistance that results in a decrease of the adhesion factor (lining fading). This process is a great disadvantage for the vehicle brake, as the brake performance is directly dependent on the adhesion factor. However, a certain continuous renewed oxide layer by wear, also called *friction carbon layer*, is necessary for a constant, stable adhesion factor (Figure 27).

Brake linings consist of a formula of varying materials that are bound with phenolic resin. During manufacturing, the mixture substance is pressed onto the lining back plate either cold or warm. After the installation into the vehicle, the linings do not yet deliver the target adhesion factor. Not enough friction contact surfaces lead to high local friction temperatures with low adhesion factors. After a little while, the linings are ground in and optimal adapted to the disk and to the movements of the brake housing (embedded). In addition, the disk surface did also run in. In general, the adhesion factor also improves after the lining has experienced higher temperatures once during the braking processes and has hardened due to outgassing. However, during the entire lifetime of a lining (approximately 50,000 km), the adhesion factor does not in anyway remain constant.

The adhesion factor depends on the contact pressure, the temperature, and the speed of the vehicle and the disk (with constant vehicle weight). After violent braking, which affects most of the decrease in adhesion factor, the brake linings have the ability to regenerate in the course of the stops following. They are stable against extreme adhesion factor fluctuations (fading) and are not sensitive to wet conditions and salt. The wear of a brake lining should be as low as possible to realize a high lifetime and a narrow design. The friction process, however, leads to oxidation processes that accelerate the mechanical abrasion as well as wear. Often, a specific wear is necessary to ensure comfort characteristics.

The temperature sensivity of the wear from the lining material should be small. Tears or edge breakage due to thermal stress can be ignored in today's series linings. In this context, the coordination of friction partners is important, so the brake disk or the brake drum on the lining. Structural conditions and hardness can shift the wearing process from "lining *eats* the disk" to "disk *eats* the lining." In general, the disk or drum lifetime is twice that of the lining lifetime. The characteristics of the lining materials are as follows:

- high breaking strength against chipping
- high shear strength against loosening from the back plate
- low natural frequency with good damping characteristics
- an average compressibility for good comfort, which, however, keeps the brake from becoming too soft
- low heat conduction to isolate against the brake fluid
- small heat expansions to prevent rolling away while using the parking brake
- no formation of residue on lining or disk/drum
- corrosion resistance.

Today, comfort behavior is more and more important. Thus, a buildup of friction vibrations is critical, if it leads to rattles and steering wheel vibrations. The response and threshold behavior of the brake depends significantly on the friction and compressibility. Compressibility leads to increased pedal travel, and adhesion factor changes effect pedal force changes. A modern lining must not make any noises out of the friction combination and the entire brake system during the whole lifetime; the lining must rathehave damping effects on the vibration system.

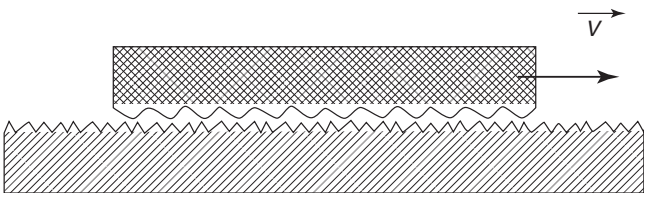


Figure 27. Schematic presentation of the friction process. (Reproduced with permission from TRW.)

6 SUMMARY

Modern vehicles have disk brakes for performance reasons and drum brakes for cost reasons.

Large energy storage devices are made out of gray iron for performance reasons and energy storage devices are made out of alternative materials for weight reasons

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

(aluminum, fiber reinforced ceramics, and carbon fibers). The latter can only store a clearly lower amount of energy with the same design space. Friction brakes ensure safe braking up to standstill, and these brakes can securely hold the vehicle mechanically and without energy conduction.

The secure design of the brake systems is optimized by electronic control systems (ABS, ESP, EVB—electronic brake force distribution, BA—brake assistant). Comfort and safety functions (EPB, ACC—adaptive cruise control—speed and distance contol unit) reduce operation force and support the driver while driving and braking.

The man–machine interfaces are more and more understood (response behavior, pedal feel, and indicator lights).

Everything that the driver experiences in brake technology in the vehicle is a compromise decided by the manufacturer between costs, lifetime, and a performance, which usually goes far beyond the legally required levels.

RELATED ARTICLES

The Development of Alternative Brake Systems
Carbon Fibre Reinforced Siliconcarbide: A New Brake Disk Material
The Cooperation of Regenerative Braking and Friction Braking in Fuel Cell-, Hybrid- and Electric Vehicles
Global Chassis Control in Passenger Cars
Chassis Control Systems—A Look into the Future

REFERENCE

Haken, K.-L. (2008) *Grundlagen der Kraftfahrzeugtechnik*, Hanser Verlag, Germany.

FURTHER READING

Bill, K.-H. and Breuer, B.J. (2012) *Bremsenhandbuch*, Springer Verlag, Germany.

Bill, K.-H. and Breuer, B.J. (2008) *Brake Technology Handbook*, USA, SAE International.

Limpert, R. (2011) *Brake Design and Safety*, 3rd edn, USA, SAE International.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto024
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The Development of Alternative Brake Systems

Bernd D.M. Gombert, Martin Schautt, and Richard P. Roberts

RG Mechatronics GmbH, Seefeld, Germany

1 Introduction 1
2 Potential Benefits 3
3 Brief Review of Alternative Systems 4
4 Challenges 8
5 Example: Electronic Wedge Brake 9
6 Conclusions 10
Glossary 11
References 11

1 INTRODUCTION

Brake system development has always been characterized by the search for and development of new technologies. Although this initially concentrated on the optimization of systems driven by pure muscle power, over the past 150 years, there has been a steady increase in the importance of powered brakes, as a response to both heavier vehicle weights and higher speeds. There has also been repeated interest in self-reinforcement as a means of lowering brake forces, either for the driver or for the actuation system. Critical to the success of such developments has been the provision of good controllability, repeatability, and comfort for the driver. Economic viability and a well-considered safety strategy are of course also essential to any such system. Such hurdles have resulted in many promising systems not being brought to the mass market, but either being restricted to niche applications or disappearing altogether. Some ideas have then reappeared as the technological boundaries have expanded, for example, due to the introduction of powerful but inexpensive microprocessors.

The future of the automobile will be significantly shaped by networked sensors and electronics and increasingly powerful driver assistance systems. These will exert even more influence on the major vehicle controls than is currently the case, intervening in the drive train, brakes, and steering. Starting with ABS (anti-lock braking system) and ESC (electronic stability control), it has been possible to control the vehicle in ways, which are simply not possible for the driver, who has neither the information available (e.g., individual wheel speed sensors), nor the ability to intervene appropriately (e.g., braking of individual wheels for ESC), nor the necessary speed of reaction. This trend will continue in the foreseeable future, with the continued improvement of existing systems, more requirements from more electric vehicles (e.g., blending of regenerative and friction braking), and the introduction of new lower bandwidth systems that exploit the possibilities offered by affordable new sensors.

For existing vehicle systems, such as ABS, TCS (traction control system), and ESC, the braking system must be able to control the force at each wheel individually and autonomously from the driver input, which can be zero. This will continue to be the minimum requirement for future braking systems. A simple electronic interface will increasingly be required between the many vehicle level systems and the high performance brake system. At the same time, the overall use of energy should be reduced to minimize the fuel consumption and production of CO_2. A significant step in this direction can be made using components that only use energy when they are actually required, which typically means a move toward more electric systems. At the same time, the desire to increase vehicle safety will ensure that the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

performance requirements will tend to increase rather than reduce.

As a response to these requirements, the automobile industry has already introduced a variety of systems that can be signaled electrically and include some form of power supply, which is independent of the driver's mechanical energy. The simplest form of such a system is the pump in an ESC system, which permits braking pressure to be produced without intervention from the driver. In order to proportion the braking forces between friction and regenerative braking, a smart booster may be used, allowing a relatively straightforward modification to existing braking systems. At the high performance end of the spectrum is the electrohydraulic brake (EHB), where, in normal operation, the pedal only functions as a simulator, which generates an electronic demand for the system. The hydraulic pressure is supplied from an accumulator and the pressure in the individual brake cylinders is controlled by means of proportional servovalves. These examples illustrate one of the key features of "by-wire" systems that the control and energy paths are separate, in contrast to a standard mechanical brake. The signaling is electronic, but the power supply can be hydraulic (EHB), pneumatic (smart booster), or electric. Electric-powered solutions can be either "wet" or "dry." An example of the former is using an electric-driven pump to provide hydraulic pressure, either

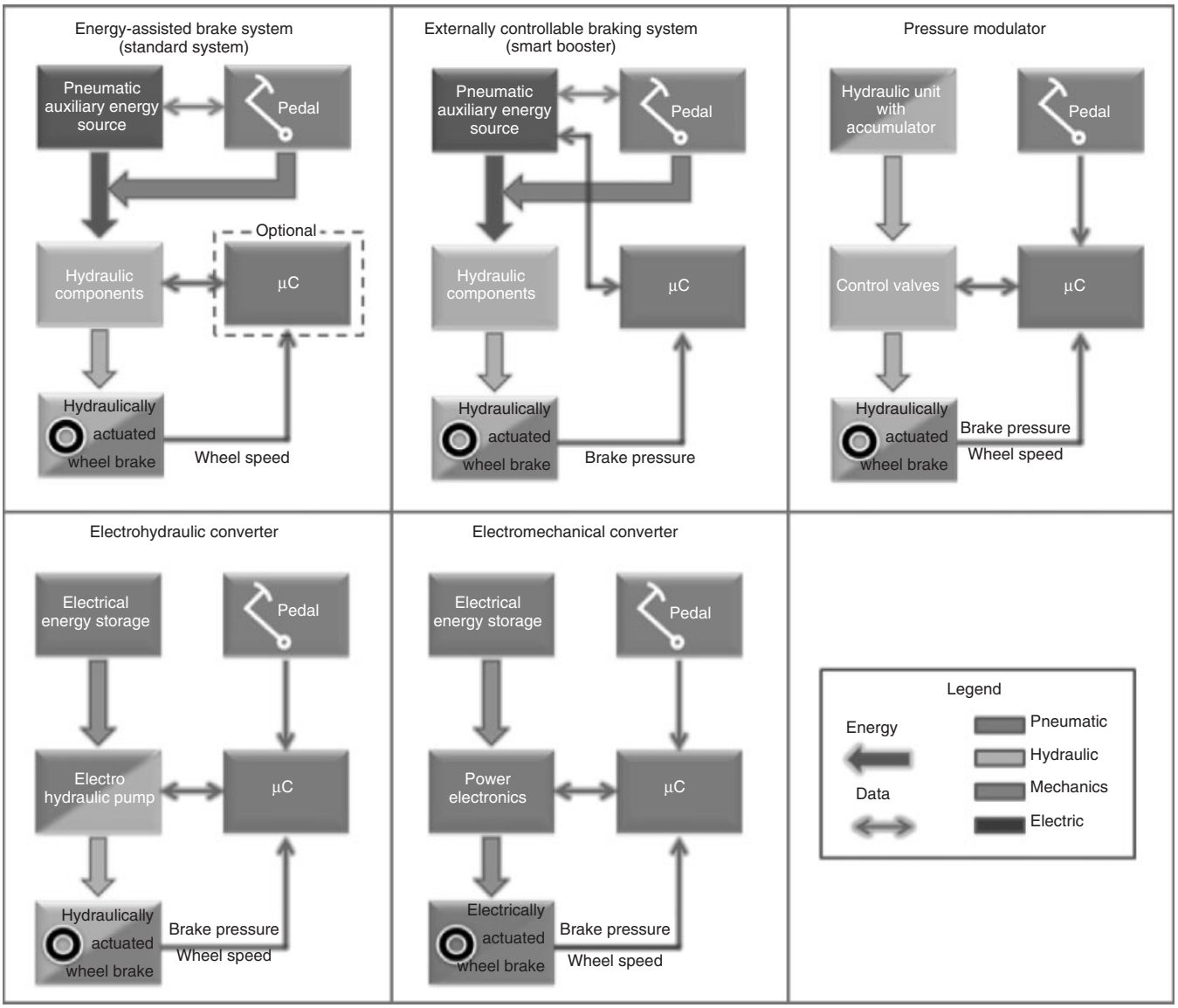


Figure 1. Signaling and power options for brake systems.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

on demand or for storing in an accumulator. In "dry" systems, the power is transmitted by purely mechanical means, obviating the need for hydraulics.

An overview of typical possibilities for powering an automobile brake system is provided in Figure 1. Of these five architectures, three can be considered by-wire systems: the pressure modulator, the electrohydraulic converter, and the electromechanical converter. The first two of these use hydraulic wheel brakes, whereas the last one is the only "dry" system illustrated.

Hydraulic systems have the great advantage of being able to store energy at a high energy density over long periods of time without significant losses. This is both an advantage for the system energy supply (a hydraulic accumulator can store much more energy in a smaller volume than a battery) and for the actuation itself. Thus, a constant wheel force can be held simply by closing a valve, without requiring a constant input of new energy, except for that, needed to overcome leakage. This is clearly true of the EHB solution (pressure modulator). Other solutions (electrohydraulic converter), where the hydraulic system can be thought of as a form of gearbox, do not share this advantage, as they have to actively hold the brake pressure. For booster systems, this requires maintenance of a certain amount of suction from the engine, in order to support the driver.

Practical electromechanical systems do not share this advantage and also have to actively hold the brakes closed. This requires current and hence electric power. For a classic electromechanical brake (EMB), this static requirement can be minimized using higher gear ratios, but this conflicts with the actuator dynamics required for ABS systems. Both the achievable rate of change of brake force and the actuator acceleration can be an issue here. This sets an upper bound to the possible gear ratio. The lower bound is effectively set by volume and weight constraints for the brake actuators. Even motors with modern rare earth magnets have a limited torque density and a direct drive system is simply not feasible. Hence, some form of gearing will always be required.

The designer of a classic EMB is therefore faced with a dilemma: if the gear ratio is too low, the long-term power requirements will be higher, which will have an impact on the long-term fuel consumption of the vehicle (not to mention higher requirements on the power electronics, etc.); should it be too high, then achieving the necessary dynamics for effective ABS control will demand high power peaks. These will have very little impact on the fuel consumption directly but will significantly affect the dimensioning of the vehicle power supply. Confronted with this problem, a general consensus has developed within the automobile industry that conventional EMBs are a possible solution for the rear axle but that a 42 V power supply is required, if they are to be used for the front brakes. Consequently, conventional EMBs are still under consideration within the industry. However, alternative approaches have been sought in order to meet all requirements with 14 V power supplies.

Within this chapter, we consider various approaches to the problem of dry brake-by-wire (BBW) actuators. In nearly all cases, an attempt has been made to reduce the required static actuator force. This means that a lower gear ratio is required between motor and brake pad and that the dynamic requirements can consequently be met more easily. The additional benefit is that many components can be reduced in size, providing actuator cost and weight savings.

2 POTENTIAL BENEFITS

The first question that arises is why such developments could be of interest. After all, the performance of existing brake systems is already quite impressive. Some points have been touched on in the introduction but we will expand on them here. It needs to be stated in advance that there is no single overriding reason why such systems have to be introduced. As a result, it will be a judgment on the part of OEMs and suppliers as to when/whether a switch is made. However, when this point is reached, none of the major players can afford to be left behind.

Firstly, it is clear that the coming years will see increasing electrification of vehicles. A wide variety of systems is possible, with mild hybrid systems at one end of the spectrum and electric or fuel cell vehicles at the other. The former will become common in the mass market, whereas the latter will probably be restricted to small production runs in the near future. For all vehicles, there will be a trend to lower fuel usage and reducing CO_2 output, as a response to rising fuel prices and as part of a political solution to achieving climate goals. Within the European Union, this has resulted in the regulatory requirement to reduce fleet average CO_2 production to 95 g/km by 2020. The significant savings offered by hybrids and electric vehicles in city traffic, where a significant proportion of drivers live and commute, will mean that more electric vehicles have an important role to play in meeting such goals. Such developments are also being pushed politically.

As a result of this increasing electrification, more electric energy (and power) will be available in vehicles. Minimizing its use is important both for reducing CO_2 output for more conventional vehicles and for ensuring

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

maximum range for minimum battery size in electric vehicles. Efficient use of the available energy and power requires the following:

- *Minimizing* quiescent power losses (power on demand)
- *Maximizing* the efficiency of the individual components
- *Controlling* power consumption via
  - Intelligent power/battery management
  - Intelligent consumers.

The consumers themselves have to be prioritized so that essential systems (e.g., steering and braking) are always provided with sufficient power, whereas comfort systems, especially those with a large power draw (e.g., seat heating), can be either controlled or disabled, should the need arise. This requires significant information flow around the power management system and demands that the critical systems can signal their needs before the power situation becomes critical.

Thus, we can see that

- the primary signaling interfaces are *already* electronic;
- the primary power interface is *becoming* electric;
- energy-efficient solutions are required, both in operation and when "off";
- more central intelligence will be required for power management;
- more local intelligence will be required to enable the power management to function properly;
- more communication will be required between "better integrated" vehicle systems, to permit optimum use of the available power and energy.

All of these factors point to more automotive by-wire solutions in the future. The main advantages of electrically powered systems are that, if correctly implemented, they only need to draw power when it is actually needed (e.g., electric power-assisted steering) and packaging, installation, and maintenance can be simplified. They are also clearly compatible with situations such as a hybrid vehicle driving in electric mode.

The issues discussed earlier are valid for any vehicle system. However, what would be the advantages of a dry BBW system? The first obvious issue is removing the hydraulic fluid itself. This saves time and hence costs during the production process and in servicing, particularly as there is no longer a need to bleed the brake system. It is no longer necessary to replace and dispose of the fluid several times during the vehicle lifetime. The brake booster and vacuum pump can be removed from the engine compartment, saving considerable volume. The lack of hydraulic architecture permits more flexibility in the system design and contributes to a more modular system, which is easier to assemble.

Perhaps the most important gains can be obtained in the pedal region. The pedal is reduced to a feel simulator system with sensors. This removes any sort of direct connection with the brakes themselves. This has significant benefits for the passive safety and for reduced noise transmission into the passenger compartment. Pedal force and travel is no longer constrained by the requirements of the hydraulic brakes themselves, so the travel can be reduced without any significant effect on the system. The feel can then be made exactly the same for a range of vehicles and can be potentially be tuned via software to individual requirements (e.g., 5% female vs 95% male). Mounting the pedal unit should be simpler and the location of the supporting structure is less constrained by other components. Finally, there will be no pulsing of the pedal during ABS interventions—some see this as an advantage and some as a disadvantage.

As regards performance, electric systems are by nature more linear than hydraulic systems and less sensitive to temperature (viscosity of hydraulic fluid). Individual wheel braking is of course implemented as standard. Experience suggests that it is possible to make an all-electric braking system perform as well as a very "highly tuned" hydraulic system and significantly better than most conventional systems. Traction control interventions in particular can be made smoother than existing systems. Tuning can be performed by means of software, which makes the process quicker and more flexible. Thus, the same hardware can be "branded" differently by adapting the software. The additional electronics provide new possibilities for system diagnosis and fit well with the concepts of intelligent power management.

3 BRIEF REVIEW OF ALTERNATIVE SYSTEMS

Two separate issues are addressed in this section:

- The various proposed brake systems
- Proposed vehicle configurations.

Different developers have taken different routes to realizing the goal of a dry BBW system, all probably with slightly different motivations. As yet, none of these systems has been applied to a mass production vehicle, although some have come close. Our review here is based on published information, which means that some developments are probably excluded, either for lack of such information or because it has not come to the attention of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the authors. Many suppliers and OEMs choose to be very discrete about their developments.

3.1 Brake actuators

We will split the brake actuators into two groups: those, which do not use self-reinforcement and those, which do. While self-reinforcement offers significant benefits, the brake response can easily be influenced by varying friction—coefficient between pad and disk, requiring higher gain controllers to ensure repeatable performance. There are also other risks, which we will address later. As a result, much development has focused on more conventional actuators, which were seen as lower risk solutions.

3.1.1 Actuators without self-reinforcement

3.1.1.1 Conventional electromechanical brake. Conventional EMBs consist of a motor, gearbox, and a ball- or roller screw, which replaces the hydraulic piston. They have been investigated by a large number of organizations, and one sample of collaboration between industry and academia is given in the article by Schwarz *et al*. (1998). As discussed earlier, a consensus has developed that a 42 V supply is required to power such actuators for the front axle of a passenger vehicle. Development of such systems has stalled; therefore, the prospect of a vehicle using pure EMB has receded.

Nevertheless, there is still interest in these actuators. The main reason is that it is relatively straightforward to include a latch mechanism in the gear train and so to produce an integrated parking brake. The actuator can be sized for the maximum parking brake force and used on the rear axle of the vehicle with a conventional 14 V supply. As braking forces when driving are generally relatively low on the rear axle because of stability considerations, the actuator dynamics are adequate for existing ABS/TCS applications.

3.1.1.2 Maximum torque brake (Delphi). The maximum torque brake (MTB) was an attempt to solve the problem of using an EMB on the front axle of the vehicle. The idea is relatively simple—instead of one brake disk, the MTB uses two. This doubles the number of friction interfaces and hence approximately halves the actuator force required compared to the EMB. Assuming the same motor, sized on long-term torque requirements, the gear ratio can therefore be halved. As nearly all the inertia is in the motor itself, this change results in approximately double the rate of change of brake force and four times the acceleration of the conventional solution. This provided sufficient dynamics to allow the MTB to be installed on the front axle with a conventional 14 V power supply (Smith and Hudson, 2003).

The most difficult engineering challenge here is in ensuring that all the sliding components move correctly relative to one another under all conditions and throughout the vehicle lifetime.

While development effort on the MTB seems to have ceased, this approach is clearly one way of overcoming the limitations of the conventional EMB. The concept can also still be powered by hydraulics if required, perhaps offering some economies of scale. In addition, it is certainly possible to make such concepts work: aircraft brakes have long relied on stacks of rotors and stators to provide the necessary energy absorption. These systems are also now moving in the direction of electric actuation.

3.1.2 Actuators with self-reinforcement

A second path to reducing the actuator forces and improving the dynamics is to use self-reinforcement. This can ideally result in the condition that no actuator force is required to produce any given braking force. While this was seen as a hard boundary for hydraulic actuators (which were only designed to push), for electromechanical actuators, this is actually the point of neutral stability. When the self-reinforcement goes beyond this point, the actuator is unstable and requires a controller to stabilize it. This means that, in the event of a failure, some independent means is required to ensure that the brake fails open.

A second issue arises with self-reinforcement, namely what happens when it disappears, typically due to the vehicle stopping. This is particularly critical on a slope. For example, if a car is braked to a halt going up a hill, then, once it has stopped, it will try to roll backward. Thus, the force, which should normally be assisting the brake, acts against it and tries to open it. Some strategy is required to deal with this problem smoothly and safely.

3.1.2.1 Electronic wedge brake. Considerable development work was conducted on the EWB by Siemens VDO in the period between 2004 and 2007. It was designed to use self-reinforcement in both the stable and unstable regimes, so as to obtain the maximum benefit from the technology (Hartmann *et al*., 2002). Adjustment for wear, a parking brake function and a fail open function were implemented in an additional mechanism to the main brake actuator. It was possible to demonstrate vehicles containing all the main modern brake control functions (ABS, TCS, and ESC), which could operate using a 14 V supply.

At this point, it is worth considering why this can be the case. It has been noted by other authors that the energy

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

required to expand the caliper is not that high, with the implication that self-reinforcement is redundant. We will examine the benefits by means of calculations based on simple assumptions.

Firstly, for a conventional EMB, we will assume a linear caliper stiffness of 20 kN/mm and a maximum clamping force of 35 kN. The energy required to achieve the maximum clamping force is therefore

$$E = 0.5 \times F_{\text{Max}} \times \Delta x_{\text{Cal}} = 0.5 \times 35{,}000 \times \frac{35{,}000}{20 \times 10^6} = 30.6\,\text{J} \quad (1)$$

Assuming the maximum force is for a fading case and that this needs to be achieved in 0.2 s, then the power required per brake to reach this force is an average of approximately 150 W over this 0.2 s, ignoring the mechanical losses.

For a wedge brake with actuation in the plane of the disk, the actuator requires lower forces to achieve the same clamping force but has to move further. Hence, if α is the wedge angle and μ is the coefficient of friction between pad and disk, then

$$E = 0.5 \times F_{\text{Max}} (\tan\alpha - \mu) \times \frac{\Delta x_{\text{Cal}}}{\tan\alpha} = \left(1 - \frac{\mu}{\tan\alpha}\right) \times 30.6\,\text{J} \quad (2)$$

Thus, excluding the mechanical losses, the wedge brake will always require less energy than the conventional solution as long as $\mu < 2\tan\alpha$. For high friction coefficients, the energy works the opposite way round to a conventional brake: the EWB requires energy to open the brake rather than to close it.

To consider the dynamic case, we have to make some more assumptions. We will assume that the wedge brake is constructed with $\tan\alpha = 0.35$ and that it is designed to operate normally for $0.15 \le \mu \le 0.55$. This means that the actuator force is reduced to 20% of that required by an EMB. Assuming we use the same motor, dimensioned so that it can hold the maximum steady torque indefinitely for both applications, then the EWB requires a gearing factor 5 less than the EMB for the same application. If the EWB must rotate at 250 r/s to meet the requirements for rate of change of force, then the EMB must rotate at

$$\omega_{\text{EMB}} = 5\tan\alpha \times 250 = 437.5\ \text{r/s} \quad (3)$$

We will assume that the motor has an inertia of 25 kg mm^2 and that this dominates the inertia of both systems. Hence, the kinetic energy required to achieve the same dynamics for the two brakes is

$$\text{KE}_{\text{EMB}} = 0.5 \times 25 \times 10^{-6} \times 437.5^2 = 2.39\,\text{J} \quad (4)$$

$$\text{KE}_{\text{EWB}} = 0.5 \times 25 \times 10^{-6} \times 250^2 = 0.78\,\text{J} \quad (5)$$

If we now consider that the maximum speed must be reached in the order of 10 ms to produce good ABS control, then the average mechanical power required for the EMB is 239 W whereas that for the EWB is 78 W. The peak power will of course be considerably higher than this.

Thus, the main advantage of the wedge brake is in providing superior dynamics for the same power. This has been demonstrated in dynamometer and vehicle tests. The penalties are as follows:

- More complicated mechanics (forward/backward braking, adjustment, and emergency release functions)
- More complicated control.

3.1.2.2 Cross-wedge brake (Mando). Mando has also been developing a wedge brake to allow braking with a 14 V supply (Kim, Kim, and Kim, 2009). The so-called cross-wedge mechanism has been used, with the aim of producing even pad wear and avoiding the use of rollers, which were felt to be vulnerable to contamination. High force amplification was not a goal, and a subjective view of the wedge design based on the publications seen suggests that the brake should normally operate in the stable regime ($\tan\alpha \approx 0.75$). This would obviate the need for an emergency release mechanism (assuming the friction coefficient can be guaranteed). However, the use of a worm gear to change the direction of the drive through 90° raises the question of whether the system can be back-driven open. An electric parking brake (EPB) function has been integrated into the brake (Kim *et al*., 2010), which is thought to be implemented by disabling the motor and allowing the worm gear to hold the force. If so, this would cause problems with the system safety requirements, which should demand that a faulty brake is opened if the vehicle is moving.

Results have been presented, which show that the brake can be used for ABS and ESC control, although the authors admit that the performance is not yet equivalent to a conventional brake system (Kim *et al*., 2010).

3.1.2.3 VE brake. The Vienna Engineering group has proposed a solution, which relies on a nonlinear lever mechanism to provide a limited degree of self-reinforcement (Putz, 2010). Because of the mechanism, both the overall mechanical advantage and the effective wedge angle vary with position. It is designed to produce a higher force ratio at higher forces but should not become unstable. Thus, no

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

additional release mechanism is required, as long as this can be guaranteed. A parking brake function can be integrated in the brake, although its precise implementation has not been specified in the published literature. At the moment, only dynamometer results have been published, but it is clear that vehicle tests will follow (Vienna Engineering, 2011).

In the view of the authors, the main challenge here is that the bearings in the lever mechanism must be able to support the forces acting on the active pad and rotate freely for the lifetime of the brake.

3.2 Vehicle concepts

A second aspect of the development of advanced brake systems is the configuration to be applied in the vehicle. As mentioned earlier, none of the systems has yet reached series production, so all of the solutions mentioned here must be regarded as provisional.

3.2.1 Rear EMBs only

In this configuration, the front axle of the vehicle retains an existing hydraulic brake system, whereas the rear axle uses EMBs, ideally with integrated parking brakes. As a means of introducing the BBW systems to the market, this approach has several advantages:

- The design requirements for the hydraulic system are simplified.
- The rear axle suits the strong points of the EMB and permits a 14 V power supply.
  - No additional parking brake required.
  - Simple integration of hill-holder type functions within the brake system.
  - Fail-silent and adjustment functions without any additional mechanisms.
- The safety concept can build on that for conventional systems.
- Service experience with such a hybrid concept will help generate reliability data for later systems.

Continental has, for some years, been developing an electrohydraulic combi (EHC) brake based on this approach (Neunzig and Linhoff, 2009).

3.2.2 Front and rear EMBs

At the start of interest in dry BBW systems, this was naturally the favored configuration. As discussed earlier though, it was determined that such an approach is only practical when a 42 V supply is used. Consequently, at the time of writing, such configurations do not appear to be being seriously investigated.

3.2.3 Rear EWBs only

Mando is investigating a vehicle with their cross-wedge EWBs on the rear axle (Kim *et al*., 2010). It is not known whether this is intended to be representative of a production configuration or it is only intended for prototype testing. In the opinion of the authors, it must be the latter, because this configuration gains the minimum advantage from self-reinforcement while being most susceptible to its disadvantages (see, e.g., the following section).

3.2.4 Front and rear EWBs

This configuration was successfully investigated by Siemens VDO. The performance of the system was extremely good and more than lived up to expectations.

The main challenges, which arose, were mainly the result of changes in direction of, or loss of, self-reinforcement. As discussed earlier, on coming to a halt forward on an up slope, the direction of the self-reinforcement changes. If nothing is done, the EWBs will draw a high current and may not be able to hold the vehicle stationary. The solution adopted was to split the braking directions of the EWBs just before the vehicle stopped: the front axle remained in the forward direction whereas the rear axle was set to brake backward. Should the direction of the self-reinforcement change, then there is always one set of brakes, which will hold the vehicle. The overall braking force was held approximately constant while this occurred. On stopping, a brake current limit was introduced to prevent the system using too much power. Because of the high degree of self-reinforcement, the brakes, which were acting in the "correct" direction, could always hold the vehicle using this reduced current.

This works well in most circumstances, but it is more tricky in the case of rear-wheel drive cars, particularly automatics. Here, the driven axle can continue pushing against the closed brakes, increasing the current required to hold them closed. Should traction control be required at low speed, then there is also the danger that the rear brakes will continually be changing direction to allow for the (potentially) different directions of self-reinforcement. Switching round the directions of the braking on the front and rear axles is not attractive because the system is then always trying to brake with the "wrong" axle at low speed.

An additional motor was required for the EPB function. To avoid producing a conventional EMB for this case, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

EPB force was reduced and distributed between all four wheels. This is not ideal for the critical case of holding the vehicle on a slope with a loaded unbraked trailer attached.

Such issues will be relevant to all the proposed systems that use self-reinforcement. Clearly, the lower the level of the self-reinforcement is, the less critical the problem becomes.

3.2.5 Front EWB and rear EMB

MOBIS has prepared a prototype vehicle in which the axles each have a different dry by wire brake (Cheon *et al*., 2010). It appears to be a well-thought-out solution, because it uses the actuators where they show most potential:

- The EWBs on the front axle require only a 14 V supply, as with the Siemens vehicles.
- The EMBs on the rear axle implement the parking brake function and are not sensitive to changes in direction.

Using this type of concept, it would be possible to increase the braking force on the rear axle once the vehicle has stopped, so that there is no danger of the vehicle moving.

While for production systems, the two different types of actuators might be regarded as a cost disadvantage, in terms of the system safety, the possibility of common mode failures is significantly reduced.

4 CHALLENGES

It is clear from what has already been written that there are benefits to be had from introducing new BBW systems and that there are solutions, which are technically feasible. However, while there has been done considerable research on this topic during the past decade and a half, there is no such system on the mass market yet. This alone demonstrates that there must be significant hurdles to overcome. So, what are they? We will list some of the major issues.

4.1 Somebody has to be first

This category contains a range of issues, for which we could think of no better title. Because no such system exists, there is no single set of requirements, which can be applied to it, so there is no well-defined set of rules to work by. Within ECE-R13H, it can be seen that there are passages, which have been added to accommodate the EHB. For any developers of future EHBs, it is therefore possible to follow these guidelines and it should be possible to obtain certification for a large number of markets. While the first one to market a dry BBW system may have the advantage of being able to help write these regulations, this requires considerable extra effort and liaison with a number of safety authorities. There is also the possibility that competitors may not only offer constructive criticism in such cases!

Any such system will now be subject to ISO/DIS 26262. This also represents a challenge, because the standard is still new and not all the associated documentation is yet available (mid-2011). In some cases, it may still be necessary to fall back on the IEC 61508, because not all the processes are completely clear, especially for such a complex system as this. This will again require additional effort and specialized personnel.

The fact that there is no internationally agreed path to certification is a major obstacle. Until such a process is agreed, there will remain a residual risk of having to repeat tests and certification work, even if the development is almost complete.

Assuming these issues are overcome, there is then the issue that the OEMs may require two independent suppliers for such systems. This is obviously not likely to be the case, and this means that at least one supplier will have to have an OEM firmly "on-board" before deciding to go ahead with production.

4.2 Costs

Costs will always be a challenge, especially now that even relatively high technology brake systems are more-or-less sold "by the kilogram." A supplier has to be willing to invest a lot of money up front to produce a system and must be very sure that there will be sufficient return on it. Quantifying exactly the scale of investment is also not easy, because of the issues mentioned earlier.

In uncertain economic times, this requires some courage, not least because existing brake systems will not stand still, neither in capability nor in price. Moreover, it will be more difficult to leverage low labor costs for a new, high technology system than for well-understood, conventional systems. Finally, competitors will be certain to make life more difficult when the new product is ready to be launched.

Thus, as always in the automobile industry, cost control will be a key factor in a successful development process.

4.3 Reliability/availability

A system containing many electronic components will have to be very well designed and tested to achieve reliability

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

rates comparable with a "dumb" hydraulic system. There is simply more that can go wrong. In addition, some of the electronic components are likely to be located in harsh environments (wheel well, with shock and vibration, and large temperature changes), which will not make this any easier.

4.4 Conclusion

In general, it is clear why there is a marked reluctance to be first in the field, despite the potential of dry BBW systems. It is perhaps not surprising that an "outsider," such as Siemens VDO, has to date been the most willing to push their development: they have the most to gain and the least to loose. Existing suppliers are often satisfied with the status quo, which is more predictable and which (still) produces reasonable margins, especially for "higher level" brake functions. In Germany, in particular, there is also the memory of the introduction of the EHB, which has made suppliers and OEMs alike more risk averse. At the same time, nobody wants to be left too far behind should the market take off.

Both these factors perhaps explain Continental's development of the EHC brake, which represents a stepping stone in this process. There is an understandable desire to minimize risk and also to ensure that the hardware is really ready for production before introducing it.

5 EXAMPLE: ELECTRONIC WEDGE BRAKE

Here, we discuss some of the experiences gathered by the authors during the development of the EWB. We split the section into subsections devoted to a particular topic.

5.1 Mechanical design

A good mechanical design is clearly the basis for any successful product. In the case of the wedge brake, there were several difficult problems to be solved in a relatively confined space.

The first obvious issue is that the brake must be actuated in two directions. This demands motion perpendicular to the direction of travel of the roller screw. The solution chosen was to attach the roller screw to the wedge with reinforced leaf springs (to prevent buckling). The forward actuation was designed to be in the plane of the wedge angle, so that normally no perpendicular motion was required. In the reverse direction, the motion was proportionally quite large, but the travel was reduced because smaller braking forces were needed. A view of this design is provided in Baier-Welt and Schmitt (2007).

The next major issue was the force sensor. Here, a new development was undertaken to produce a sensor integrated in the caliper using cheap Hall sensors to measure the relative deflection of a loaded and unloaded part. The noise level from this sensor was noticeably lower than that of the load cells used in the early prototypes, but it was more sensitive to drift. Some detail about this development is provided in Baier-Welt (2007).

The most complicated single development was that of a mechanism to implement the following three functions:

- Emergency release (fail silent)
- Pad wear adjustment
- Parking brake.

This was built around a second motor, a solenoid, and a cam. During normal operation, a roller sat on the cam slope, supporting the brake clamping force. The cam was actively locked by means of the solenoid. In the event of a loss of power, a spring opened the solenoid and the cam released the brake force. The pad wear adjustment relied on interaction between the solenoid and the secondary motor and could not be conducted under load. Finally, the parking brake used a stable depression in the cam surface to hold load without power. Illustrations of these functions are provided in Baier-Welt (2007).

5.2 Actuator control

Control was obviously a major issue in the development of the EWB, simply because the wedge can be unstable. Any controller has to be able to address this fact. The taken approach was to design a controller, which could stabilize the brake, even if the friction coefficient between pads and disks was 1. This also required an estimate of this friction coefficient, so that the motor could not be overpowered and pulled in.

Two different approaches were taken, which were to some extent dictated by the mechanical design. The early Beta prototypes were designed with two motors so that drive-train backlash could be actively removed (Roberts *et al*., 2003, 2004). They were also stiff in the axial direction. The latter property allowed the controller to be based on a conventional cascaded design, with an inner current loop, followed by a rate loop, and an outer force control. The motors were controlled either to remove backlash if the friction was close to the ideal value or to cooperate if more axial force was required. Brief details are provided in Ho *et al*. (2006) and Lang *et al*. (2006).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The preproduction prototypes were all based around a single motor. On the one hand, this simplified some of the control, but on the other hand, the mechanism required for forward and backward brakings was much more flexible in the axial direction than in the earlier hardware. For this reason, it was necessary to use a state "feedback" controller. The system was stabilized for a μ of 1.0 and a conventional PID controller was then used to achieve the desired performance. This process is discussed in Fox *et al*. (2007).

5.3 Vehicle control

Vehicle level systems were designed and optimized with considerable use of simulation tools. This, together with the well thought out structure, enabled a comprehensive system to be produced in a relatively short time. It also simplified the process of adapting the control to different vehicles.

A hierarchical structure was built up, with slip controllers at the lowest level to manage ABS and traction control interventions, and a vehicle dynamics controller at the level above this. This implemented the brake force distribution, ESC, and some wedge-specific functions (e.g., direction control). This structure helps produce relatively smooth interventions, which were not only of benefit to passenger noise and comfort but also suit electromechanical actuators in general, where the motor inertia is an important influence on performance. More details are provided in Semsey and Roberts (2006).

5.4 Functional safety

A description of the approach to functional safety is provided in Schaffner *et al*. (2006). The brake system requirements were taken as far as possible from ECE-R13H, whereas IEC 61508 was used to cover the electrical and electronic components. At this stage, ISO 26262 was still a working draft and it was mainly used for categorizing risks into ASILs (automotive safety integrity levels). Three major categories of failures were found to lead to ASIL-D (most severe):

- Unintended braking
- Insufficient braking
- Braking (or lack of it) leading to vehicle stability problems.

After consideration of a number of potential architectures, a solution was proposed based on.

- three pedal sensors (fail-operational)
- two EPB button sensors (fail-silent)
- a central electronic control unit with three microcontrollers (fail-operational)
- two energy management systems, each with an associated backup battery, attached to the main vehicle power supply. Each power channel feeds one brake diagonal, the central electronic control unit, and the pedal unit.

In the case of a single fault, the objective was to achieve either fail-operational or fail-degraded behavior (reduced performance). For a second fault, it had to be possible to bring the vehicle to a controlled stop. For a failure where the next failure could lead to a critical loss of performance, an Auto-Stop function was proposed. This would have led to additional hazards of its own.

5.5 Electronics

The design of the electronics followed the functional safety requirements. A fail-operational central control unit was designed and produced. The wheel unit electronics featured an independent "safing control" processor and were developed in two phases: firstly an external version to check the circuitry and then an integrated version. Considerable thermal modeling and testing was done to support the latter development. It was found that, in order to meet the EMC targets, integration within the actuator was a "must." More information can be found in Zelger (2007).

6 CONCLUSIONS

In this contribution, we have tried to lay out some of the background to the development of dry BBW systems. The technical motivation for such systems belongs to an increasing development over the past few years: more and more manufacturers look seriously at hybrid vehicles and electromobility. Although the level of research being conducted is not currently as high it perhaps has been, there are still a number of organizations working on the problem. There are a variety of different solutions under investigation, both in terms of actuator technology and vehicle architecture. There has even been some new actuator concepts been developed in recent years, relying on limited self-reinforcement.

It is nevertheless easy to see why such systems have not yet been brought to market. There is nothing simple about introducing a revolutionary new product into an established market for a safety-critical application. However, the authors are still in the firm opinion that it is only a question of when this breakthrough occurs, not if. The "fit" between this technology and the trends in automobile development

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is simply too good. Once this occurs, the very obstacles, which have discouraged many from taking this step, will act to secure a competitive advantage for the successful organizations.

GLOSSARY

ABS	Anti-lock braking system
BBW	Brake-by-wire
EHB	Electrohydraulic brake
EHC	Electrohydraulic combi brake
EMB	Electromechanical brake
EPB	Electric parking brake
ESC	Electronic stability control
EWB	Electronic wedge brake
OEM	Original equipment manufacturer
MTB	Maximum torque brake
TCS	Traction control system

REFERENCES

Baier-Welt, C. and Schmitt, B. Electronic Wedge Brake. *IQPC 3rd Annual Innovative Braking Conference*, Frankfurt am Main, Germany, 26th April 2007.

Baier-Welt, C. *High Performance Electrical Wedge Brake Actuator*. Vehicle Dynamics Expo, 5th May 2007.

Cheon, J.S., Jeon, J.H., Kim, J.S., *et al*. (2010) Brake by wire system configuration and testing using front EWB (electric wedge brake) and rear EMB (electro-mechanical brake) actuators. SAE Paper 2010-01-1708.

Fox, J., Roberts, R., Baier-Welt, C., *et al*. (2007) Modeling and control of a single motor electronic wedge brake. SAE Paper 2007-01-0886.

Hartmann, H., Schautt, M., Pascucci, A., and Gombert, B. (2002) eBrake—the mechatronic wedge brake. SAE Paper 2002-01-2582.

Ho, L.M., Roberts, R., Hartmann, H., and Gombert, B. (2006) The electronic wedge brake—EWB. SAE Paper 2006-01-3196.

Kim, J.G., Kim, M.J., and Kim, J.K. (2009) Developing of electronic wedge brake with cross wedge. SAE Paper 2009-01-0856.

Kim, J.G., Kim, M.J., Chun, J.H., and Huh, K. (2010) ABS/ESC/EPB control of electronic wedge brake. SAE Paper 2010-01-0074.

Lang, H., Roberts, R., Jung, A., *et al*. (2006) The road to 12V brake-by-wire technology. *VDI-Berichte Nr.*, **1931**, 55–71.

Neunzig, D. and Linhoff, P. Electro Hydraulic Combi Braking System (EHC). *IQPC 5th Annual Innovative Braking Conference*, 2009.

Putz, M. (2010) VE mechatronic brake—investigations of a simple electro-mechanical brake. SAE Paper 2010-01-1682.

Roberts, R., Gombert, B., Hartmann, H., *et al*. (2004) Testing the mechatronic wedge brake. SAE Paper 2004-01-2766.

Roberts, R., Schautt, M., Hartmann, H., and Gombert, B. (2003) Modelling and alidation of the mechatronic wedge brake. SAE Paper 2003-01-3331.

Schaffner, J., Doerricht, M., Hartmann, H., and Gombert, B. *Approach to a Functional Safety Concept for the Electronic Wedge Brake*. BremsTech, 2006.

Schwarz, R., Isermann, R., Böhm, J., *et al*. (1998) Modeling and control of an electromechanical disk rake. SAE Paper 980600.

Semsey, A. and Roberts, R. (2006) Simulation in the development of the electronic wedge brake. SAE Paper 2006-01-0298.

Smith, A.C. and Hudson, S.M. (2003) A new, high torque brake design using sliding discs. SAE Paper 2003-01-3309.

Vienna Engineering. *TechnischeUniversität Braunschweig bestellt brake-by-wire System*. www.vienna-engineering.com, 2011.

Zelger, C. (2007) Electronic Wedge Brake—Actuator Integrated Electronics. *IQPC 3rd Annual Innovative Braking Conference*, Frankfurt am Main, Germany, 26th April 2007.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto025
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Carbon-Fiber-Reinforced Silicon Carbide: a New Brake Disk Material

Andreas Kienzle and Hubert Jäger

SGL Carbon GmbH, Meitingen, Germany

1 Introduction 1
2 Carbon Fiber and C/C Brakes 2
3 C/SiC Brakes 2
4 Production Ways to CF/SiC Materials 3
5 Elements of a Ceramic Brake Disk 4
6 Production of C/SiC Brake Disks 5
7 Material Behavior 8
8 C/SiC as Brake Disk Material 10
9 Outlook 11
10 Summary 11
Related Articles 11
References 12

1 INTRODUCTION

Brake disks in cars are safety relevant parts, which require high reliability and a high lifetime. The task of a brake disk is to transform the kinetic energy during braking by friction into heat, which is absorbed by the brake disk and the brake pads. The absorbed heat must be dissipated in the surrounding area by thermal conduction, heat radiation, and convective flow of heat. A brake system of this kind is accordingly limited by the friction characteristics of the brake materials and its ability to store and remove the heat. In general, brake materials must have very good thermomechanical properties, high and constant friction characteristics, and good resistance against abrasion. With the development of the first cars in the late *nineteenth* century, the kinetic energies are still very low but never the less, the cars must be stopped safely. Therefore, there was an increasing demand in equipment, which allows their safe stopping. The first cars used the running surface of the wheels directly for braking. Brake linings out of wood or leather were pressed mechanically by a lever on this running wheel surface. This brake was substituted by the brake drum made out of metal. Brake disks for braking need around another 50 years before they were used in cars. Nevertheless, the first patent regarding to brake disks was granted in 1902 to Lanchester (1902): This patent describes the first time the idea to use a brake disk, which is mounted on the wheel hub. The brake linings were pressed mechanically on the disk surface during braking. The first brake disks out of gray cast iron were used for race applications (Jaguar winning the 24-h race of Le Mans in 1953). Alfa Romeo introduced the first brake disks in road vehicles. The BMW 502 3.2l was the first German car equipped with disk brakes on front wheels, which can be ordered as an option followed by Mercedes 1961 in the 300 SL.

In addition, the development in the material optimization of gray cast iron and in the production process technologies for the production of ventilated disks allowed reducing cost and improved the quality constancy of the material. Nevertheless, the main disadvantages of the gray cast iron disks, that is, like high weight due to the high material density of 7.2 gcm^{-3}, corrosion sensitivity by oxygen and water, limited high temperature stability, and high wear rates, could not be solved. In addition, the speeds, which are attained nowadays by such vehicles, are constantly increasing. Since 1950, the top speed of upper class

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

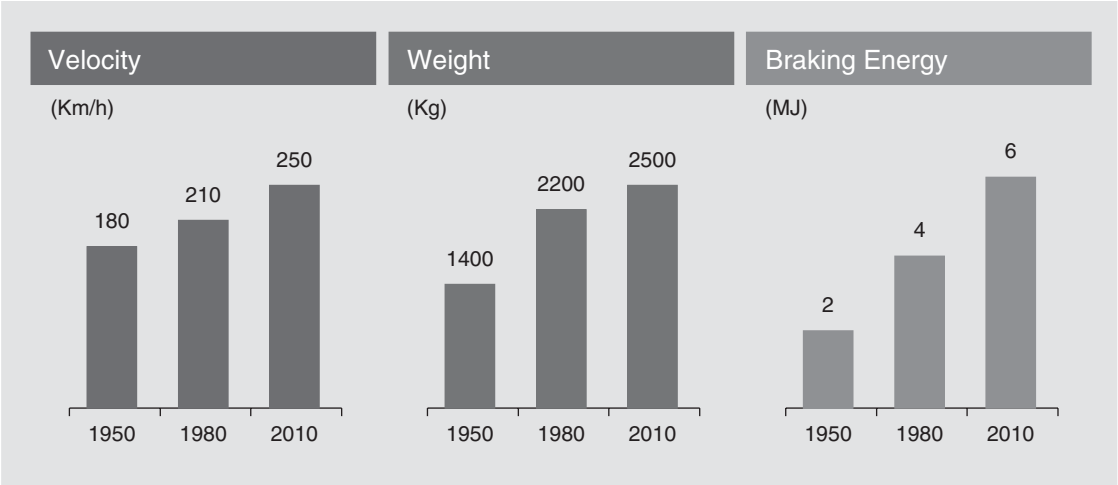


Figure 1. Increase of kinetic energy of upper class limousine shown here since 1950 up to the year 2010 leads to a demand of the automotive industry to look for new brake disk materials. (Reproduced by permission of SGL Carbon GmbH.)

limousine increased from around 180 to 250 km/h in the year 2010. The 250 km/h was a limit decided by the car manufacturer (OEM) for tire reasons. In the same time, also the car weight increases from around 1400 to 2500 kg. This leads to higher kinetic energy from 2.38 MJ in the year 1950 to 6.0 MJ in the year 2010 (Figure 1). This high energy results in a lower lifetime of the metal disks because of the increasing thermal loads during braking. This heat can also negatively result in a reduction of the friction coefficient during braking, called *fading*, with the result of longer stopping distances.

This leads to an increasing demand of the automotive industry to look for new brake disk materials and systems, which are stable under these rising braking conditions.

2 CARBON FIBER AND C/C BRAKES

With the development of the carbon fibers in the early 1960s, a new type of composite material, the carbon-fiber-reinforced carbon (CFC) material, which is made out of these fibers, leads to a new high temperature stable material class. These materials show very low weight because of their low density and found fast interest for high energy braking systems such as race cars or aircraft brakes and railway applications. The used carbon fibers are produced out of polyacrylnitrile polymer (PAN) or pitch-based fibers by carbonization at temperatures up to 1800°C under inert atmosphere. They show very high mechanical strength and high stiffness and additionally they show a very high temperature stability >2000°C in inert atmosphere without losing their mechanical performance. The disadvantages of these fibers are their behavior against oxygen. Owing to the oxidation reaction of carbon with oxygen under the formation of CO and CO_2, they are only stable up to 500–600°C under air. Nevertheless, the C/C brake disks and CFC pad materials developed for race car applications show excellent high friction coefficients that are stable even under extreme loads. The disadvantages of these disks are the low friction coefficient on cold disks surface and also in the wet state of the disk. In addition, the high wear rates especially in the cold disks combined with the high production costs limit this disk material mainly for race or aircraft applications. These disadvantages avoid a broad entrance of this CFC material class in the area of the normal road vehicles.

3 C/SIC BRAKES

The problems of the C/C brake disks could be solved by the development of a new type of brake material: the fiber-reinforced SiC-based ceramic brake disks (CF/SiC). Here, the carbon fibers are embedded in a ceramic matrix made out of SiC. Normally, ceramic materials suffer from their brittle behavior under mechanically and thermally induced stresses, which limit their applications. Different ways to improve fracture toughness of ceramics were developed in the past decades. The toughening of SiC ceramic by C-fibers or ceramic fibers such as SiC-fibers is the most effective way to reduce such brittle behavior (Bader, 1993; Evans, Zok and Davis, 1991) and opens the wide field for the use of this material. First trials to use these types of CF/SiC materials started in the past decade of the last century, based on public-funded projects in Germany for the development of ceramic brake disks for high speed trains.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

First patents for CF/SiC brakes were granted for Krenkel and Kochendörfer (1994). They used woven fabrics for the C/C performs production. After the infiltration of this preforms with liquid silicon, they get the C/SiC material, which could be used as brake disk. The commercialization of CF/SiC brake disks started with the development of the short fiber-reinforced CF/SiC materials. This type of material was first developed at SGL and closely parallel at Daimler Chrysler and described by Gruber and Heine (1997) (SGL) and Haug *et al*. (1997) (Daimler Chrysler) in the patents. This material shows good friction coefficients in cold and wet disk states. Owing to the high hardness and low porosity of the material, it shows additional low wear rates in cold and hot disks. Their low weight (more than 50% reduction of unsprung mass compared to cast iron), high hardness and high stability of resulting friction coefficients, high corrosion resistance, and long lifetime are the main advantages for using CF/SiC ceramics as brake disk material or clutch material for automotive applications and their increasing demands. At the Frankfurt Motor Show in 1999, the carbon-ceramic brake disk was shown the first time to the public. In 2001, Porsche AG was the first car producer who installed the carbon-ceramic brake disk as series equipment into the 911 GT2. Since that time, also other premium cars use the advantages of CF/SiC brake disks. At present, for nearly all high end sport cars and luxury limousines for instance from Ferrari, Porsche, Audi, Bentley, Bugatti, and Lamborghini, CF/SiC brake disks are available (Figure 2). Ferrari equips today all new cars by series with the CF/SiC brake systems. In the year 2011, in total around 80,000 CF/SiC brake disks were produced.



Figure 2. CF/SiC Brake disk. (Reproduced by permission of SGL Carbon GmbH.)

4 PRODUCTION WAYS TO CF/SIC MATERIALS

There are different ways available for the production of carbon-fiber-reinforced silicon carbide, which are summarized in Figure 3. These technologies show different ways to bring in the ceramic matrix in a porous carbon fiber preform under the formation of C/SiC. The technologies used are chemical vapor infiltration (CVI), liquid silicon infiltration (LSI), and the polymer infiltration and pyrolysis (PIP), which bring the ceramic SiC matrix in a carbon fiber preform. These technologies can be used alone or in combination and are explained in the following sections.

4.1 Chemical vapor infiltration

Starting from a porous carbon fiber preform, the ceramic matrix can be impregnated by a gas-phase deposition reactions chemical vapor infiltration (CVI) at temperatures $>1000^{\circ}$C. To control the interface of the fiber to the ceramic matrix, a special interface layer must be deposited on the fiber surface. In most cases, this is a thin carbon layer (several micrometer thickness) deposited by CVI. Subsequently, the types of gases are changed and the ceramic matrix is deposited. Very slow infiltration rates are necessary for this reaction to keep the infiltrating channels open. After several days of infiltration, the surface must be brushed to open closed porosity on the surface. These together with the slow infiltration rates lead to long production cycle times of several weeks and to the high

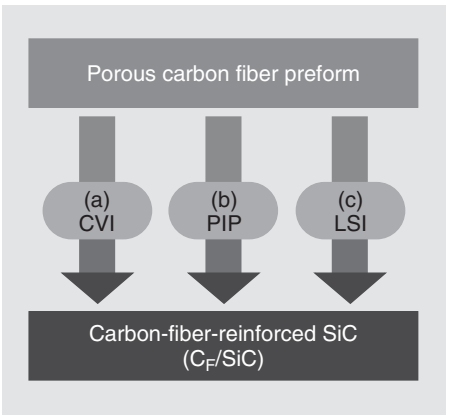


Figure 3. Production ways for carbon-fiber-reinforced silicon carbide materials starting from a porous carbon fiber preform infiltrating the ceramic matrix by (a) chemical vapor infiltration (CVI), (b) polymer infiltration and pyrolysis (PIP), and (c) liquid silicon infiltration. (Reproduced by permission of SGL Carbon GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

costs of this technology. The resulting parts produced by this technology show excellent mechanical behavior and a matrix porosity up to 15 vol%.

4.2 Polymer infiltration and pyrolysis

The impregnation of the porous carbon fiber preform with inorganic silicon-based polymers such as polysilans or polycarbosilanes used in the polymer infiltration and pyrolysis (PIP) technology is another way to build up the SiC ceramic matrix around the coated carbon fiber preform. The ceramic composition after the pyrolysis depends on the used inorganic polymer. These polymers are infiltrated under vacuum followed by a pressure cycle in the preform. After the pyrolysis of such impregnated preforms at temperatures up to 1000°C under inert atmosphere, the polymer decomposes and is transformed into a porous amorphous ceramic SiC material. To get a mechanical stable material, these impregnation and pyrolysis cycles must be done several times. Each cycle increases the density and closes the residual open porosity and raising the mechanical stability of the resulting CF/SiC material. Normally up to 5–6 cycles are necessary to get enough density, stability, and porosity lower than 10%. Production times for the PIP route are therefore 1–2 weeks. High process and polymer costs and the polymer availability limit currently this production way. To reduce the infiltration cycles, additional fillers in the preform can be introduced. Use of inorganic polymers such as polysiloxanes or polysilazanes lead to other ceramic matrix materials such as ceramics in the ternary systems Si-C-O and Si-C-N, which are currently not used for brake disk applications. This can be an option for the future.

4.3 Liquid silicon infiltration

The fastest way to produce CF/SiC brake disks is the reaction way via the LSI process. Here, the liquid silicon is infiltrated at temperatures higher than the melting point of pure silicon (>1420°C), in the porous C/C preform. To increase the infiltration speed, the infiltration is done under vacuum. The liquid silicon is infiltrated by capillary forces in several minutes into the porosity of the preform and reacts with the carbon matrix in a strong exothermic reaction under the formation of β-silicon carbide (Figure 4).

The residual porosity of the material is filled with the unreacted excess silicon. This LSI production way allows a fast production of C/SiC parts. The resulting material shows a low porosity of <1%. A small dimensional change only during the transfer from the CFC part to the ceramic part is a great advantage of this advanced technology.

For the CF/SiC brake disk, the LSI process is currently the mainly used process for production.

5 ELEMENTS OF A CERAMIC BRAKE DISK

The ceramic C/SiC brake disk of today is normally built up out of the following elements (Figure 5):

- The CF/SiC carrier body is the main element of the brake disk. It has the task to transfer the forces and to store and transport the heat, produced during braking. Therefore, the material of the carrier body must have a high mechanical stability and ductility and a good

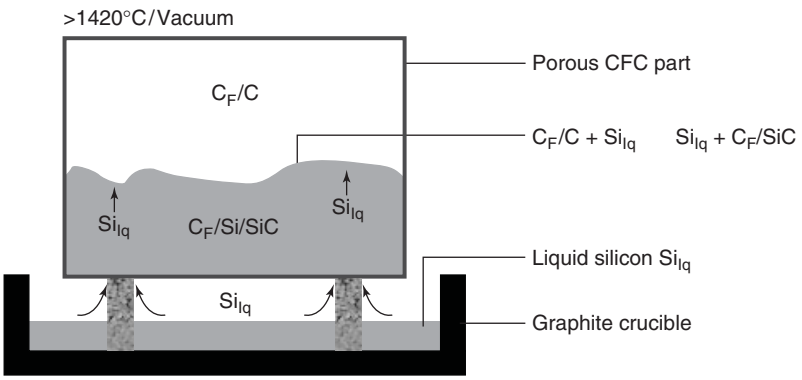


Figure 4. Schematic drawing of the liquid silicon infiltration process of CFC (w = porous carbon wick). (Reproduced by permission of SGL Carbon GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

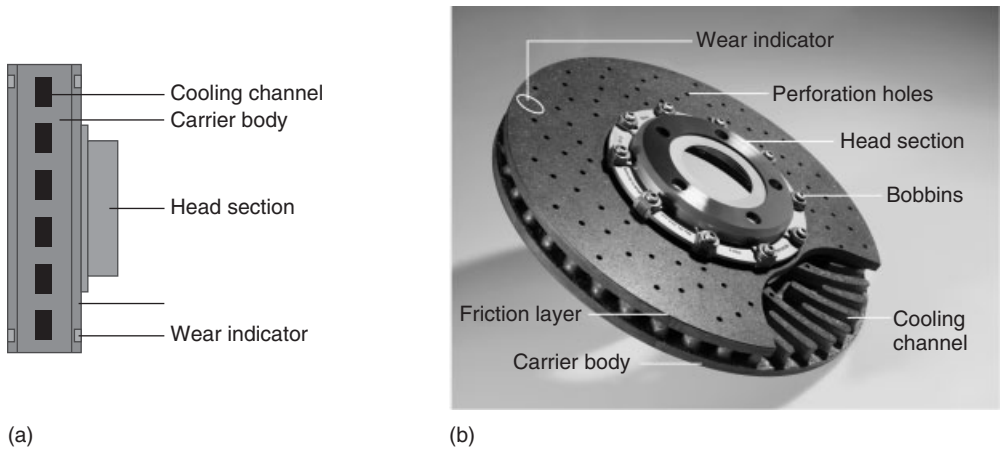


Figure 5. (a,b) Assembly up of a typical C/SiC brake disk. (Reproduced by permission of SGL Carbon GmbH.)

heat conductivity to transport the heat from the friction layer to the cooling channels. C/SiC material of the carrier body has typically higher carbon fiber content than the separate produced friction layer. Moreover, longer carbon fiber bundles are used than in the other disk parts because of the higher strength needs.

- The cooling channels in the carrier body have the task to allow a fast cooling of the brake disk. Therefore, the heat has to be transported as fast as possible out of the disk by the air passing through the cooling channels. The design of the cooling channel can be optimized to the demands coming from the car such as top speed and weight of the car.
- The friction layer is the contact area of the brake disk to the brake linings. Friction coefficient, wear, and lifetime of the disk are depending on the interaction of the friction partners. The friction layer material has a higher hardness than the carrier body material and typically a low carbon fiber content. If carbon fiber bundles are used, they are smaller and shorter than in the carrier body. As the C/SiC material has already low wear rates, brake disks can also be applied without a special friction layer.
- The head section today is typical out of metal. It is mounted on the brake disk by special bobbins, which allow compensating the different thermal expansion of the metal head section and the ceramic disk during braking. The head section is the contact area of the disk to the hub. The materials used for the head section are stainless steel or aluminum.

The dimensions of the produced C/SiC brake disks today start with outer dimensions of 350 mm up to above 420 mm and a thickness of above 32 mm, depending on the car weight and maximum speed.

6 PRODUCTION OF C/SIC BRAKE DISKS

6.1 Protection of carbon fiber bundles against reaction with liquid silicon

To reduce the direct reaction of the carbon fiber with the liquid silicon during the infiltration step, the carbon fibers must be protected. Therefore, the production of short fiber-reinforced CF/SiC brake disks starts with the preparation of the short fiber bundle with a special protection against liquid silicon. During the past decades, different technologies were developed to protect the fiber bundles against the reaction of liquid silicon (Gruber and Heine, 1997; Haug *et al*., 1997; Krätschmer *et al*., 2004). One promising way to prevent the carbon fiber bundles siliconization is the filling of the porosity in the bundles with various types of carbon. One example is the infiltration of the fiber bundles with phenolic resin or pitch followed by a pyrolysis step (Figure 6).

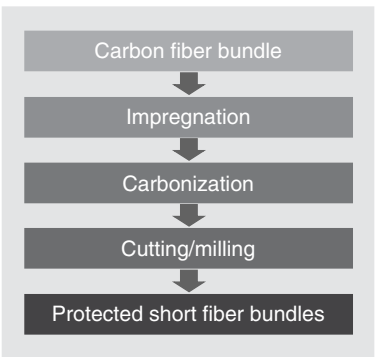


Figure 6. Production of carbon fiber bundles protected against liquid silicon. (Reproduced by permission of SGL Carbon GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

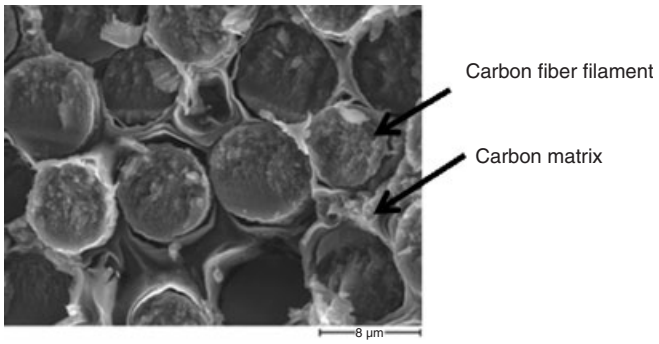


Figure 7. Cutting surface of a densified carbon fiber bundle by impregnation with pitch followed by a carbonization step. (Reproduced by permission of SGL Carbon GmbH.)

The carbon fiber bundles are used in the form of fiber rovings, woven fabrics, or pressed short fiber plates. The bundles are filled with phenolic resin or pitch and then carbonized up to temperature of 1200°C. Additional impregnation and carbonization cycles lead to a decrease in porosity in the bundles and an increase in fiber bundle stability and integrity. Figure 7 shows the SEM image of the cutting surface of an impregnated and compacted C-fiber bundle. The space between the single C-fiber filaments of the fiber bundle is filled with a dense carbon layer of pitch-based carbon. These dense carbon structures in the bundle give a good protection of the carbon fiber filaments against the reaction with liquid silicon during the siliconization process. The schematic production way to a ventilated C/SiC disk with friction layer is shown in Figure 8.

6.2 Mixing and molding step

For the carrier body of the disks, these stabilized fibers are mixed with phenolic resins as binder and further carbon-containing fillers such as graphite powder or coke powders.

After the mixing step is finished, the press masses are ready for the filling step. In the case of a production of solid disks without cooling channels, the press mass is filled directly in the mold and densified under pressure to the necessary green density. The plates therefore are heated up to 200°C where the phenolic resin starts to cross-link. After the cross-linking of the resin, the final CFRP disk is removed out of the mold and ready for further processing.

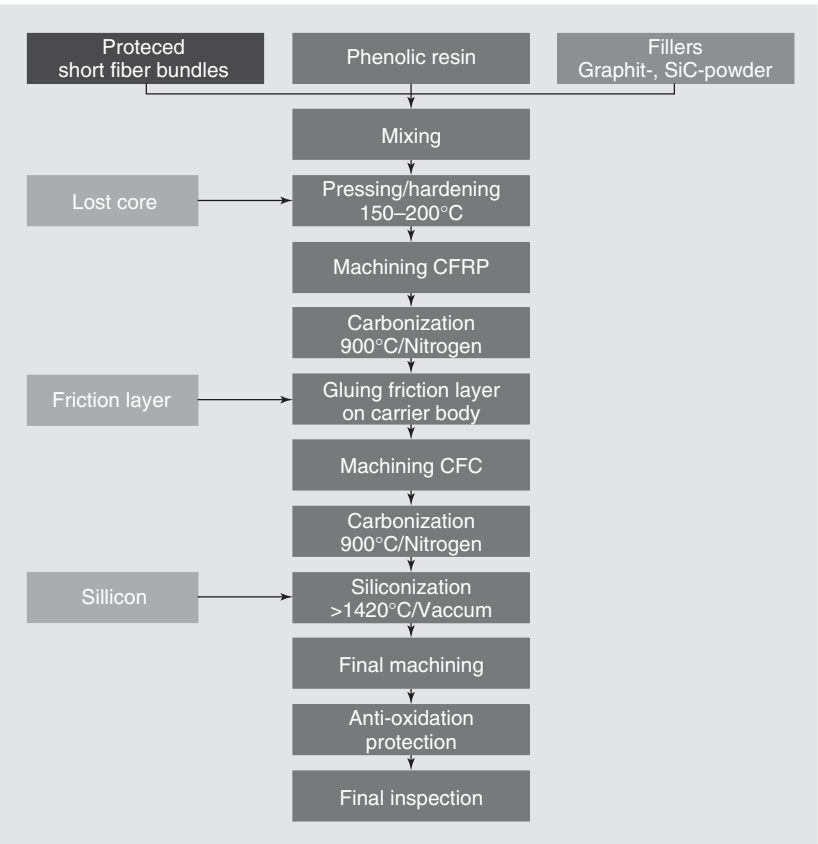


Figure 8. Production process of a ventilated C/SiC brake disk with friction layer. (Reproduced by permission of SGL Carbon GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For the preparation of ventilated disks, more effort is necessary. Here, different technologies were developed. For cooling channels with an easy and radial geometry, the solid disks could be machined after the molding step. The disadvantages of these machining of the cooling channels are the high material loss and the high machining effort. To avoid these machining steps and also to allow more complex cooling channel geometries, three different production technologies are used today:

(a) The disk could be produced out of two parts with ribs on each part (Martin, 1999). The cooling channels of the ventilated disk are formed after gluing the two disks together in the contact areas of the ribs.
(b) For basic geometrical radial cooling channel, geometries drawable cores out of metal are used (Pacchiana and Goller, 2001). After the pressing step, the cores are removed mechanically out of the CFRP disk.
(c) For complex formed cooling channels, the lost core technology is used (Bauer *et al*., 2002). Therefore, a core is produced with the geometry of the designed cooling channels. This core material is decomposed during the heat treatment of the disks completely. A disk with near-net-shaped cooling channels remains for further processing.

6.3 Carbonization

The pressed CFRP disks are removed out of the molds and are machined with conventional machining tools in the outer and inner diameter and thickness. After this premachining step, the disks are carbonized at temperature up to 1200°C under inert gas atmosphere. During the heating up to 1200°C, the organic binder material is thermally decomposed and transformed in porous glassy carbon such as carbon material under evolution of organic species with lower molecular weight. The weight loss of the used phenolic binder systems is typically in the range 40–60 wt%. The main weight loss of the binder is thereby in the temperature area between 400 and 600°C. These temperatures must be carefully crossed to avoid damages of the ceramic fiber-reinforced ceramic (CFRC) disks.

6.4 Friction layer

For disks with friction layer, the friction layer is glued on the porous CFRC carrier body. The friction layer itself is produced in a separate way and a special recipe is used. Therefore, a mixture of carbon fiber bundles, fillers, and phenolic resin as binder is mixed to form a press mass analogous to the carrier body press mass. In difference to the carrier body typically lower, carbon fiber bundle content is used. In addition, the used fiber bundles are smaller in their dimensions than in the carrier body material (Gruber, Heine and Kienzle, 2001). The friction layer is molded equal to the carrier body, on a press at temperatures up to 200°C. The produced CFRP parts were carbonized in the next step under inert atmosphere and then glued on the porous carrier body by a separate pressing and heating step. After the fixation of the friction layer, the final green machining step starts (perforation holes, etc.) and the disks get their finished design. After an additional carbonization step to transform the glue to carbon, the disks are then ready for the siliconization step.

6.5 Siliconization

The siliconization process of the CFC disks is performed in graphite crucible on wicks out of porous carbon material. The crucibles are filled with the calculated amount of solid silicon and are heated up to temperatures of 1700°C. The heating is done under a vacuum atmosphere in special siliconization furnaces. At a temperature of 1420°C, the silicon starts to melt and the molten silicon is absorbed by the wicks and transported by capillary forces into the porous CFRC disk. The infiltration into the body starts immediately and, in a highly exothermic reaction, the CFC disk is infiltrated completely and the matrix carbon and also some fiber filaments on the surface of the fiber bundles react with liquid silicon under the formation of SiC. The formed layer around the fiber bundles protects the fibers in the inner part of the bundle against further reaction with the liquid silicon. After the infiltration is completed, the furnace is cooled down to room temperature. Owing to the lower density of liquid silicon compared to solid density, some of the excess silicon is pressed out of the disks and forms some bigger Si drops, which must be removed after cooling. The density of the resulting C/SiC disk after siliconization is in the range 2.2–2.4 gcm^{-3} and depends on the used recipe and the CFC density of the disks.

In the next production step, the C/SiC disks are machined in inner and outer diameters and on the friction layer using diamond tools. The holes for the fastener elements are drilled in the C/SiC state to guarantee the necessary high precision. To protect the carbon fiber bundles in the disk against oxidation at temperatures >500°C during braking, the disks are further impregnated with an anti-oxidation solution. The C/SiC disks are then assembled with the metal head section using connecting elements made out of stainless steel and the disk is finally machined to the defined tolerances of parallelism between the head section/hub area and the friction area. The assembly is balanced by a groove, machined in the coverage of the brake disk. Every finished

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

disk is quality controlled via sound test to detect internal defects. The dimensions of the disks are measured and additionally the thickness of the friction layer is controlled by measurement. The disks are then ready for shipping.

7 MATERIAL BEHAVIOR

7.1 Microstructure

An overview of the microstructure of a typical short fiber CF/SiC material from the carrier body of a brake disk is shown in the micrograph of Figure 9. The fibers are oriented perpendicular to the press axis *z*. In the *xy*-plane, the fibers are distributed with a random orientation or depending on the filling step. The fiber bundles are embedded in a SiC matrix (gray color). The porosity of the porous CFRC structure is filled with silicon (white areas).

Owing to the fiber protection of the bundles, the reaction with the silicon takes place only on the surface of the fiber bundle under the formation of a dense silicon carbide layer. This dense silicon carbide layer protects the fiber against further reaction during the infiltration. The different thermal expansion coefficients of the carbon fiber bundles, the silicon, and silicon carbide matrix lead to the formation of microcracks formed during the cooling of the material from 1420°C to room temperature. The microcracks are stopped in the neighboring fiber bundles. In addition, there are some areas of unreacted amorphous carbon matrix in the microstructure, which are also embedded in a formed SiC layer.

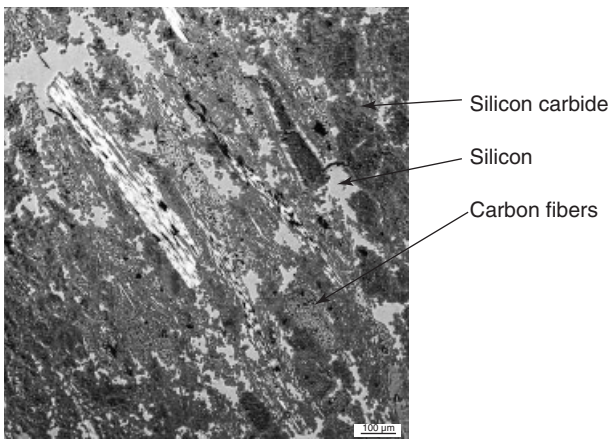


Figure 9. Micrograph of carbon-fiber-reinforced SiC. Sample of carrier body from a brake disk material. (Reproduced by permission of SGL Carbon GmbH.)

7.2 Material properties

In Table 1, the material data of typical short fiber CF/SiC materials in comparison to gray cast iron are summarized. In addition, the typical values of 2D reinforced CF/SiC material are mentioned. This material is produced by the siliconization of 2D CFC material out of woven C-fiber phenolic prepregs by lamination and carbonization. Short fiber-reinforced material shows a lower mechanical stability (strength, modulus, and elongation at maximum stress) than the long fiber-reinforced CF/SiC material. The higher density of the short fiber-reinforced material and the higher content of silicon and silicon carbide lead to increases in heat conductivity through the material. The density of the CF/SiC materials with values of 2.3 gcm^{-3} is significantly lower than the metal with densities of >7.2 gcm^{-3}. This lower density gives a significant weight advantage. If, on the front and rear axles, the CF/SiC brake disk is used instead of metal disks, weight reductions of up to 20 kg by car can be achieved for disks with 400 mm diameter. The differences in physical properties of the two types of CF/SiC materials are summarized in Table 1 together with gray cast iron GG-20.

The specific heat capacity of CF/SiC material is around 40% higher than cast iron. The combination of the high heat capacity, low Young's modulus, and high heat conductivity results in a high thermal shock resistance of the CF/SiC material, compared to the GG 20.

Table 1. Summary of the material data of short fiber- and long fiber-reinforced CF/SiC.

	I C/SiC *Short Fiber*	II C/SiC *Long Fiber*	GG-20
Density (g/cm^3)	2.3	1.9	7.2
Four point bending (MPa)	70–80	200	220
Strength modulus (GPa)	30	70	110
Elongation at break (%)	0.3	0.45	0.3–0.8
Heat conductivity *z*-axis (W/(mK))	35	10	54
Thermoshock resistance (W/m)	$>27,000$	$>27,000$	<5400
Maximum operating temperature (°C)	1400	1400	700
Specific heat capacity/ weight (kJ/kgK)	0.8	0.8	0.5
Thermal expansion ($*10^{-6}$ 1/K)	2.5	1.5	9.0–12
Phase content			
Si (wt%)	13	10	
C (wt%)	32	48	
SiC (wt%)	55	42	

(Reproduced by permission of SGL Carbon GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

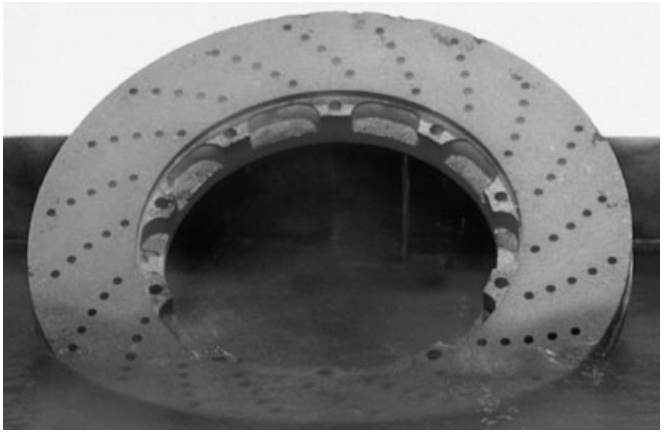


Figure 10. Thermal shock test of C/SiC brake disk. (Reproduced by permission of SGL Carbon GmbH.)

Figure 10 shows a disk, which is heated up to 1000°C and then dipped into water with one-half. While the side in the water is cooled down below 100°C, the non-dipped part is still red hot. Silicon carbide ceramic without fiber reinforcement would not survive such extreme thermal shock conditions. Another big advantage of the CF/SiC materials is the increase in mechanical stability with increasing temperatures (Figure 11). During the heating up to temperatures of 1200°C, the mechanical strength of CF/SiC increases up to 80% higher values, compared to respective room temperature values. In addition, the stiffness of the material is also increasing. In comparison to CF/SiC, the gray cast iron loses at temperature of 700°C around 40% of the mechanical room temperature strength and also stiffness.

Under dynamic load tests, CF/SiC materials show a reduction in stiffness in the range 20% but without reduction in the mechanical strength of the material (Thielicke, 2005). Owing to the heterogeneous material structure of the CF/SiC and the formed microcracks resulting from the different thermal expansions of the materials during the cooling step, critical stresses are reduced. In addition, newly formed cracks from an applied load are bridged or bypassed by carbon fiber bundles and the cracks are stopped. These effects also result in a low notch sensitivity of the CF/SiC materials. The high hardness compared with the good mechanical behavior and the increasing mechanical behavior with increasing temperature makes the short fiber material an ideal candidate for the use as brake disk material. The CF/SiC ceramic material combines the beneficial behavior of a hard ceramic material, which is responsible for the low wear, good tribological behavior, and long lifetime of the brake disk with the advantages of fiber-reinforced material that shows high fracture toughness and damage tolerance. The resulting quasi-ductile properties of the ceramic composite material ensure the resistance to high thermal and mechanical stability. Limitations are coming from the oxidation of the carbon fibers at temperature over 500°C under air. This oxidation can result in a decrease of the mechanical strength of the CF/SiC and may have a negative impact in the lifetime of the

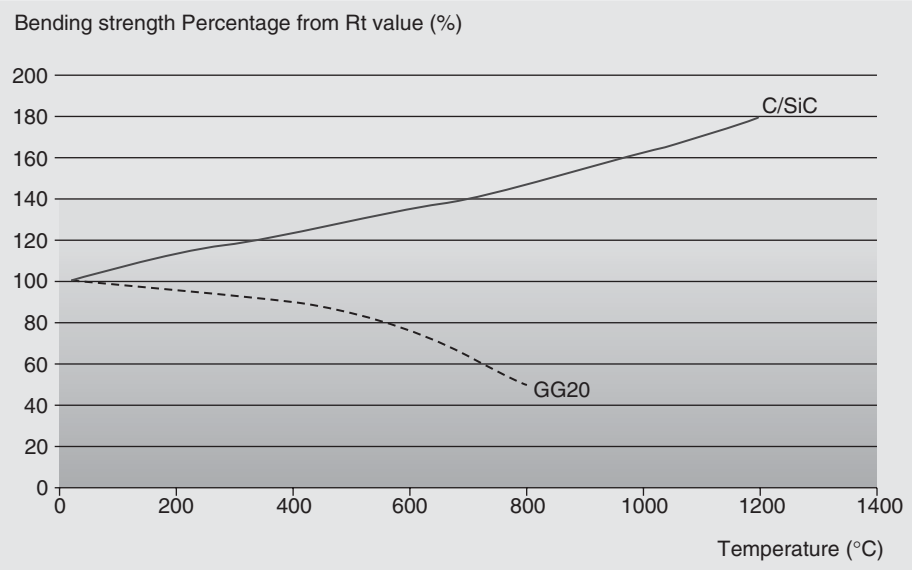


Figure 11. Bending strength behavior of GG 20 and C/SiC with increasing temperature up to 1200°C as percentage of the room temperature value. (Reproduced by permission of SGL Carbon GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

disks. Anti-oxidation agents such as aluminum hydrogen phosphates are preventing long time any negative oxidation effect during application.

8 C/SIC AS BRAKE DISK MATERIAL

8.1 Dimensioning and design of the disk

For the design of a CF/SiC disk for a new car, identical tests must be performed as for gray cast iron disk. Main control parameters are the maximum speed and the maximum load and their distribution on front and rear axles and the time to stop the car from maximum speed. The brake disk layout is largely determined by the response during full braking at v_{max} when the tensile stresses hit their peak because of the superimposition of the maximum braking and centrifugal forces.

The resistance of the brake disk mainly depends on their design of the chamber interface and interior disk rim as well as on the shape of the webs and cooling ducts. In addition, the cooling conditions of the disks are important factors. In usage, the brake disks are mainly strained by centrifugal loads and braking forces. Therefore, they must withstand lateral forces as well and further consideration must be given to thermal loads and stresses occurring at the interface between the brake disk and the brake disk chamber. Calculations for assembled carbon-ceramic brake disks include the design of the head section connection. The necessary dimensions of the disks are further calculated to ensure all possible mechanical load situations during the brake use over the lifetime of the car. This prevents that neither the disk itself nor any other component in its direct neighborhood is exposed to excessive thermal loads. To remove the heat as fast as possible out of the disks, the cooling channel geometry must be optimized using numerical methods for each car model. First prototypes are produced from the optimized disks and cooling channel dimensions and tested on a test bench also under extreme load conditions (Figure 12). The mechanical stability, braking performance, and heat dissipating behavior of the tested brake disks are the main results from these tests.

These results are used to optimize the simulated and calculated behavior of the disks and consequently modifications in the construction are transferred. The loads over the lifetime of the car are simulated on the disks and their mechanical stability is tested. These tests are done with disks directly after the production and also after surviving the simulated lifetime, for example, by bursting tests. Finally, right after perform dynamometer tests and analyzing the results, the tests on the car are done. They cover not only high speed runs on a test circuit but also mountain passes, descents, and road tests in normal traffic under winter and summer conditions. During these test runs, the test driver evaluates the brake behavior, the braking performance, and the braking comfort.

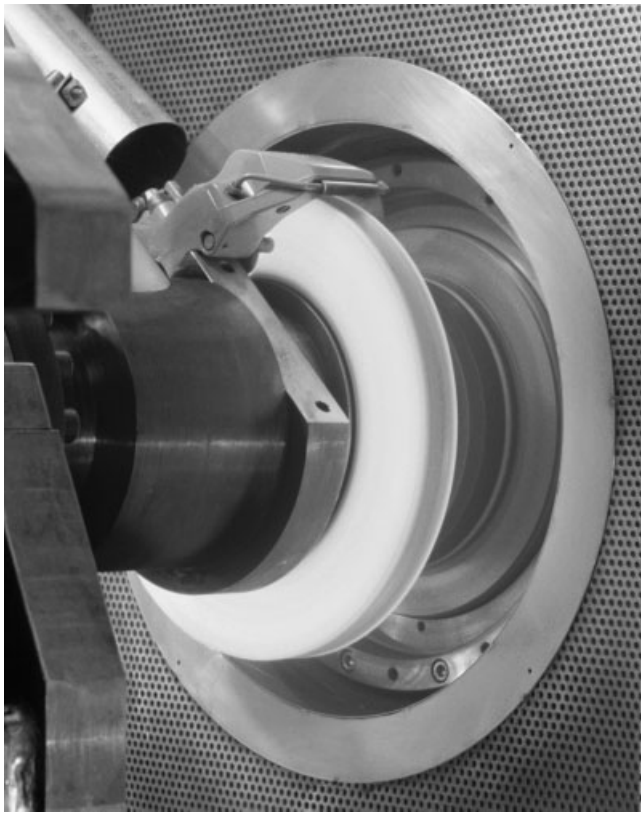


Figure 12. Overload friction test of a C/SiC disk on a test bench. (Reproduced by permission of SGL Carbon GmbH.)

8.2 Brake caliper and brake pad

For the CF/SiC brake disks today mainly fixed caliper with up to four opposing piston pairs are used. The pistons clamp the CF/SiC disk during braking between the pad materials. To avoid overheating of the hydraulic fluid of the caliper, the pistons are equipped with a layer with low heat conductivity between the brake pad and the piston. At present, mainly organic bounded pads are used as brake pad. The material is matched to the friction behavior of the ceramic surface of the disks. Metal bounded pads and ceramic bounded material pads were tested for the C/SiC brake disks and show good friction performance. Owing to comfort problems, such as noise evaluation during braking, various improved pad materials are still under development and testing.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

8.3 Advantages C/SiC brakes

In comparison to the standard metal brake disk, this ceramic brake disk has several advantages:

Low Density. The low density of around 2.3 gcm^{-3} leads to a significant reduction in weight of the brake disks compared to gray cast iron (−30% to 70%).

High Abrasion and Corrosion Resistance. Gray cast iron disks show a reduction of the disk thickness during lifetime by wear and corrosion, which can be measured easily. In comparison to this, the CF/SiC brake disk shows only low wear rates because of the high hardness of the SiC ceramic and consequently only a low reduction of disk thickness during lifetime of the car is measurable. The significantly lowered wear of disk and pad material offers as additional advantage the reduction of break dust. This is a very positive contribution for the environment.

Noise during Braking. High thermal loads during cooling after high energy breakings at temperatures over 700°C form cracks and deformations of the metal disks. These deformations result in noise during braking and lead to disk changes. Owing to the low thermal expansion and the high temperature stability of the C/SiC disk, there is no significant change in dimensions or geometries of the disks measurable during braking, which could reduce the performance of the disks.

Friction Behavior and Fading Stability. With increasing temperatures, the CF/SiC disks show additionally a slight increase of the friction coefficient during braking in comparison to the gray cast iron disks. This friction behavior results in an improved fading stability of the CF/SiC disks compared to a metal disk. The fast build up of the friction coefficient and the higher possible values of the friction coefficients of >0.45 compared to gray cast iron (from $\mu = 0.3$ up to 0.45) lead to an excellent positive pedal feeling for the driver and results in significantly shorter stopping distances especially from high road speeds.

8.4 Other automotive friction applications of CF/SiC materials

Owing to the high strength and ductility of the long fiber CF/SiC material and the high hardness of this material, it can be applied as material for clutch disk systems. This type of disk was for the first time used in the high end sports car Porsche Carrera GT. The main advantages for CF/SiC in clutch applications are the low wear rate under extreme load, the low weight, and the high and stable friction coefficient. The disks can be designed in smaller dimensions and allow lowering down the drive train and the center of gravity of the vehicle with the better performance of the driving characteristics. Currently, this ceramic clutch material is mainly used in high end race cars.

9 OUTLOOK

The current CF/SiC disks are mainly used in high end sports and luxury road vehicles because of high costs of the disks. To enter the market of normal cars, the price of the disks must come down. This will happen with the development of new brake systems allowing easier disk geometries and resulting in a less complex production. Of ongoing importance is the increase of the automation of the production processes. Therefore, new production technologies must be developed, implemented, and qualified. With their high lightweight potential, the C/SiC brake disk in combination with their discussed braking performance, the CF/SiC material is the right brake disk material for cars in future.

10 SUMMARY

Since 1950, the top speed of upper class limousine increases from around 180 to 250 km/h in the year 2000. In the same time, the car weight also increases from around 1400 to 2500 kg. This leads to higher kinetic energy, which results in a lower lifetime of the metal disks and fading problems, with the result of longer stopping distances.

In the past decade, the carbon-fiber-reinforced silicon carbide brake disk material is developed. The material is produced by liquid siliconization of porous CFC perform with liquid silicon under vacuum. The resulting material shows a low weight because of the low densities of around 2.3 gcm^{-3} and high hardness. At the Frankfurt Motor Show in 1999, the carbon-ceramic brake disk was shown the first time to the public. At present, for nearly all high end sports and luxury cars, CF/SiC brake disks are available. In this contribution, the production processes of the CF/SiC disks are described and an overview of the CF/SiC material and brake disk behavior is given.

RELATED ARTICLES

Brake Systems, an Overview
The Development of Alternative Brake Systems

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

REFERENCES

Bader, M.G. (1993) Reinforcing fibers: the strength behind the composites *Materials World*, **1**, 22–26.

Bauer, M., Gruber, U. Heine, M. Huener, R. Kienzle, A. Rahn, A., and Zimmermann-Chopin, R. (2002) Process of Manufacturing Fiber Reinforced Ceramic Hollow Bodies. EP 1300376 B1.

Evans, A.G., Zok, F.W., and Davis, J. (1991) The role of interfaces in fiber reinforced Brittle Matrix composites *Composites Science and Technology*, **2**, 3–24.

Gruber, U., Heine, M. and Kienzle, A. (2001) Friction or Slip Body Comprising Composite Materials, Reinforced with Fibre Bundles and Containing a Ceramic Matrix. EP 1216213B1.

Gruber, U., Heine, M. (1997) Mit Graphitkurzfasern verstärkter Siliciumcarbidkörper. DE1971010105.

Haug, T., Kienzle, A., Schwarz, C., Stöver, H., Weißkopf, K., Dietrich, G., and Gadow, R., (1997) Verfahren zur Herstellung einer faserverstärkten Verbundkeramik. DE 19711829 C1.

Krätschmer, I., Kienzle, A., Wuestner, D., Domagalski, P., Haeusler, A. (2004) Polymer Bound Fiber Tow. EP 1645671B1

Krenkel, W., Kochendörfer, R. (1994) Verfahren zur Herstellung einer Reibeinheit mittels Infiltration eines porösen Kohlenstoffkörpers mit flüssigem Silizium. DE 4438455 C1.

Lanchester, F.W. (1902) Improvements in the Brake Mechanism of Power propelled Road Vehicles. GP 26407.

Martin, R. (1999) Bremsscheibe aus Faserverbund Werkstoff. DE 19925003 A1.

Pacchiana, G. P. and Goller R.S. (2001) Mold and Procedure for Manufacturing of a Braking Band with Ventilation Ducts in Composite Material. EP 1412654B1.

Thielicke, B. (2005) Mechanische Charakterisierung von Faserkeramiken für Bremsscheiben *Konstruktion*, **9**, 20–21.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto026
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The Cooperation of Regenerative Braking and Friction Braking in Fuel Cell, Hybrid, and Electric Vehicles

Zhuoping Yu and Lu Xiong

Tongji University, Shanghai, China

1 Introduction 1
2 History of the Regenerative Braking System 2
3 Cooperative Working of Regenerative and Friction Braking Systems 3
4 Three Parts of the Cooperative Regenerative Braking System 7
5 Related Research Fields in this Regards 12
6 Cases about the Cooperative Regenerative Braking System 13
7 Summary 14
Related Articles 15
References 15

1 INTRODUCTION

It is well known that when the brakes are applied in a car to slow it down, some energy is wasted. The kinetic energy converts into heat and becomes useless. However, the energy crisis is becoming more and more serious at the moment and carbon emission is reaching a peak. Therefore, energy must be used as efficient as possible.

The beginning of the twenty-first century could very well mark the final period in which internal combustion engines are commonly used in cars. Now, automakers are trying to apply new energy technologies, such as pure electric, hybrid, plug-in hybrid, hydrogen fuel cell, and other alternative fuel energies such as biofuel. At the same time, automotive engineers have beaten their brain out to wring the maximum efficiency out of aerodynamic streamlining of the bodies and use of lightweight materials. Among them, regenerative braking is one of the most important technologies.

When driving a car, you have to hit the brakes occasionally to stop the car or adjust the speed. For a conventional car, about 80% of its energy converts into heat through friction. However, with a system that can recapture much of the car's kinetic energy and convert it into some other restorable and reusable energy, regenerative braking is able to capture as much as half of that wasted energy and put it back to work. Therefore, the fuel consumption can be reduced by 10–25%.

Figure 1 shows the energy conversion by braking system. The conventional braking systems convert the kinetic energy into friction heat and acoustic energy, for example, braking squeal, which cannot be reused. To convert the kinetic energy into some storable and useable energy, a lot of various schemes of regenerative braking have appeared. For instance, the kinetic energy of the vehicle can be converted into electrical energy by the electric motor and stored in the battery or supercapacitor or mechanically converted to the kinetic energy of a flywheel (e.g., KERS in some F1 racing car), etc.

The regenerative brake with electric motors is the well-developed one than other variants of regenerative braking systems. It is used in electric vehicles, primarily pure

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

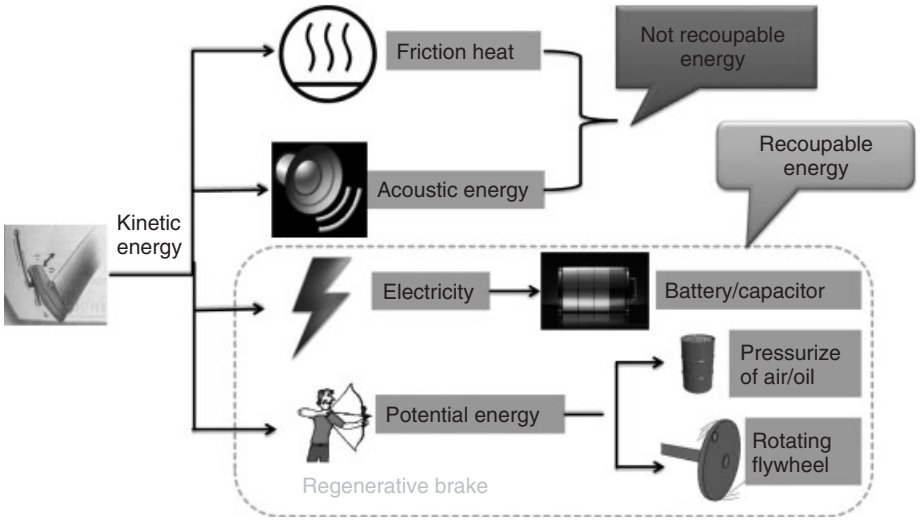


Figure 1. Energy conversion by braking systems.

electric vehicles and hybrid electric vehicles, whose battery can be used for longer periods of time without the need to be recharged by external charger or get gas service. The traction motor or the electric machine would work as a generator when using the regenerative braking, it could produce a brake torque, meanwhile producing electrical energy and stored it into energy storage components, namely, battery or supercapacitor.

2 HISTORY OF THE REGENERATIVE BRAKING SYSTEM

The origin idea of regenerative braking dates from the nineteenth century. In the most common form, there is an electric motor working as an electric generator. In electric railways, the generated electricity is put back to the supply system, whereas in battery electric and hybrid electric vehicles, the energy is stored in a battery or bank of capacitors for later use.

2.1 The Krieger electric landaulet

Louis Antoine Krieger (1868–1951)'s horse-drawn cabs were incipient examples of regenerative braking system. There is a drive motor with a second set of parallel windings in each front wheel. The motors are free to revolve a short distance either way around the large spur wheel and drive the wheels through spur gearing. Under the driver's seat is the main battery and under the passenger seat is an additional one. While descending a hill, there is certain provision for charging the battery (Figure 2).



Figure 2. Krieger electric landaulet. (Reproduced from Harris & Ewing, 1906. Library of Congress, Prints & Photographs Division, photograph by Harris & Ewing.)

2.2 "Regenerative control" in tramway

In order to reduce electricity consumption, tramway operators brought the Raworth system of "regenerative control" into use in the early 1900s in British cities such as Devonport (1903), Rawtenstall, Birmingham, Crystal Palace-Croydon (1906), and many others. While slowing down the car or keeping its speed in the downhill, the motors worked as generators, braking the car and capturing the kinetic energy at the same time. There were also wheel brakes and track slipper brakes in the tramcars in case of failure

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 3. Train carrying iron ore transported between Kiruna and Narvik. (Reproduced from Gubler, 2009. © David Gubler.)

of electric braking systems. The tramcar motors were shunt wound in several situations, and the systems on the Crystal Palace line utilized series-parallel controllers. However, an embargo was promulgated on this form of traction in 1911, because of a serious accident at Rawtenstall. Fortunately, the regenerative braking system was reintroduced 20 years later.

2.3 Application of regenerative braking in railways

Regenerative braking has been widely used on railways for many decades. The advantages of the regenerative system are the reduction of energy consumption, the reduction of wearing of brake shoes and wheel tires, and the consequent lowering of maintenance costs. The disadvantage is that the motors are larger and more costly. There are many examples of remunerative regenerative braking abroad in mountain railways and on lines where the gradients are heavy. From Riksgränsen on the national border to the Port of Narvik, the trains use only 20% of the regenerated energy. This regenerated energy is sufficient to power the empty trains back up to the national border. Any excess energy from the railway is pumped into the power grid to supply families and businesses in the region, and the railway is a net generator of electricity (Figure 3).

3 COOPERATIVE WORKING OF REGENERATIVE AND FRICTION BRAKING SYSTEMS

In a traditional braking system, brake pads produce friction with the brake rotors to slow or stop the vehicle. Additional friction is produced between the slowed wheels and the surface of the road. This friction is what turns the car's kinetic energy into heat. With regenerative brakes, on the other hand, the system that drives the vehicle does the majority of the braking. When the driver steps on the brake pedal of an electric or hybrid vehicle, these types of brakes put the vehicle's electric motor into reverse mode, causing it to run with reverse torque, thus slowing the car's wheels. While running with reverse torque, the motor also acts as an electric generator, producing electricity that is then fed into the vehicle's batteries. These types of brakes work better at certain speeds. In fact, they are most effective in stop-and-go driving situations. However, in a cooperative regenerative braking system, there still exist two sub-braking systems, that is, traditional friction braking system (usually hydraulic braking system) and regenerative braking system based on electric motors. Generally, the reasons are as follows:

1. The regenerative braking torque is not large enough to cover the required braking torque; furthermore, when the motor speed rises, regenerative braking torque will decrease because of the flux-weakening control.
2. The regenerative braking cannot be used for many reasons such as high state of charge (SOC) or high temperature of the battery to increase the battery life.
3. Regenerative brake cannot brake to stop (because of its work principal).

The relationship between regenerative and friction brake system is shown in Figure 4.

At the beginning of braking, the motor (light gray line) could meet the driver's rapidly rising demand for the total braking torque (black line), when the motor reaches its full potential, the hydraulic braking (dark gray line) begins to intervene. Continue with the braking, vehicle speed decreases, and the motor torque raises a little (light gray line). The motor torque declines rapidly (light gray line) when the vehicle slows down, and the braking torque is fully provided by the hydraulic brake (dark gray line) when close to stopping.

3.1 Three ways to cooperate the two sub-braking systems

There are three ways to realize the cooperative working of friction and regenerative braking systems (Zheng, 2010; Zhang and Ning, 2009).

The first way is to add the regenerative braking system directly onto the original hydraulic braking system without changing the arrangement of the original hydraulic braking system (Figure 5). The advantage of this way is simplicity.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

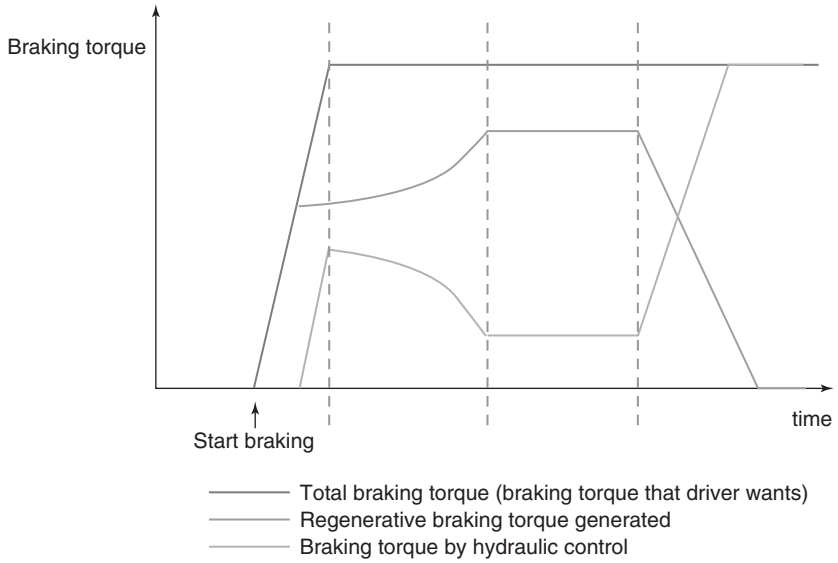


Figure 4. The relationship between regenerative and friction brake system.

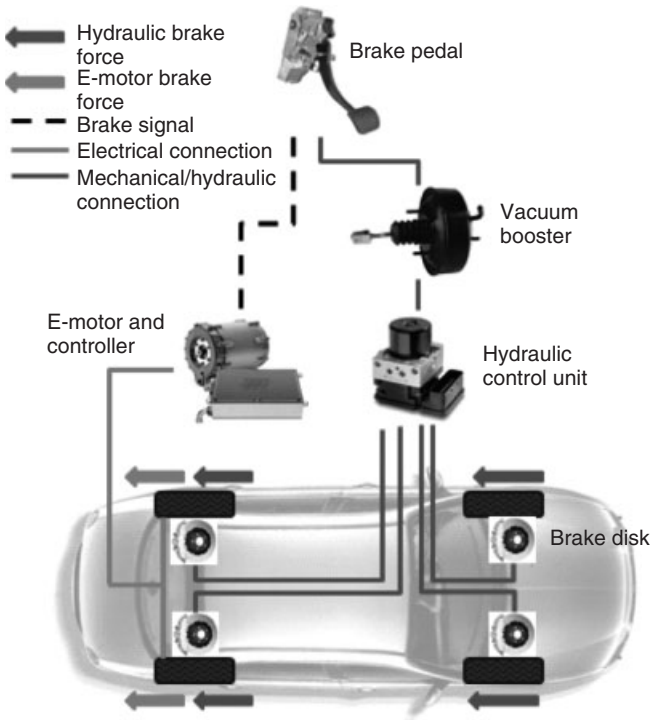


Figure 5. Add the regenerative braking system directly.

It is not necessary to rebuild a braking system. In this way, the structure is simple and what need to be changed can be minimized. However, the disadvantage is the low energy recovery efficiency because of the current braking force distribution design (Zhang and Ning, 2009).

The second way is brake-by-wire technology, where many of the functions of brakes that have traditionally been performed mechanically will be performed electronically (Zhang and Ning, 2009) (Figure 6). Here, electronic braking pedal and integrated sensors are used. When drivers step down the brake pedal, the sensor records the force, displacement, velocity, and other necessary signals to identify drivers' brake intention. The controller would give command to the pedal displacement simulator in order to supply the designed friction brake force. In this way, the friction brake force can be controlled actively to maximize the energy recovery efficiency. At present, different automakers have come up with different circuit designs to handle the complexities of regenerative braking. However, in all cases, the most important part of the braking circuitry is the braking controller.

The third way is redesigning the current hydraulic braking system by a large margin (Zhang and Ning, 2009). By adding and controlling the valves on the hydraulic braking pipelines, the hydraulic braking pressure can be controlled in time to generate the designed friction brake force. In terms of theory, this design of cooperative regenerative braking system can achieve the highest energy recovery efficiency.

3.2 The structure of the cooperative regenerative braking system

Regenerative braking system is used in the vehicles driven fully or partly by electric motors. Electric motor supplies the regenerative brake force. One of the most interesting

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

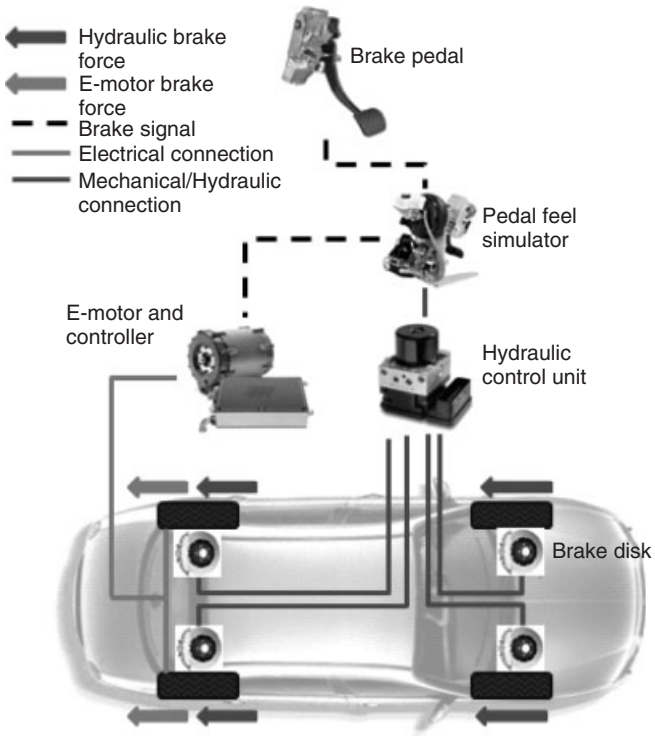


Figure 6. Brake-by-wire technology.

properties of an electric motor is, when it is running in one direction, it converts electrical energy into mechanical energy that can be used to perform work (such as turning the wheels of a car), but when the motor is running in reverse mode, a properly designed motor becomes an electric generator, converting mechanical energy into electrical energy. This electrical energy can then be fed into a charging system for the car's batteries, shown in Figure 7.

A cooperative regenerative braking system includes a frictional braking device for applying a friction brake torque to the wheels, and a regenerative braking device for applying a regenerative brake torque to drive wheels of the vehicle. Here, frictional braking device means a system which operates on the principle of friction mechanism to generate brake force regardless of using disk brake structure or drum brake structure, hydraulic or pneumatic actuating device (usually it is hydraulic braking system). The regenerative braking device means the system which operates to transform the vehicle kinetic energy to electricity energy, usually it is electric motor. A cooperative regenerative braking system is usually used in the following three kinds of new energy vehicles, such as pure electric, hybrid electric, and fuel cell vehicles.

3.2.1 Regenerative braking in battery electric vehicle

Battery electric vehicles use the motor as a generator when using regenerative braking: it is operated as a generator during braking and its output is supplied to electrical loads and it provides the braking effect. This energy can be saved in a storage battery or a supercapacitor and can be used to propel the motor (Figure 8).

3.2.2 Regenerative braking in hybrid electric vehicle

A hybrid vehicle is a vehicle that uses two or more distinct power sources to propel the vehicle. The term most commonly refers to hybrid electric vehicles (HEVs), which combine an internal combustion engine and one or more electric motors. Hybrid system is divided into series hybrid, parallel hybrid, and series–parallel hybrid.

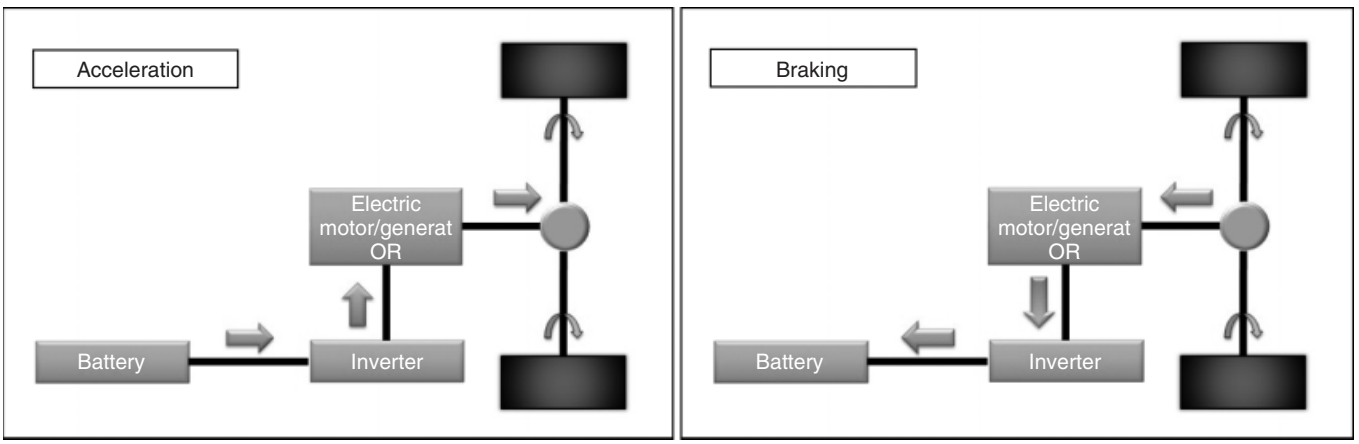


Figure 7. The working property of electric motor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

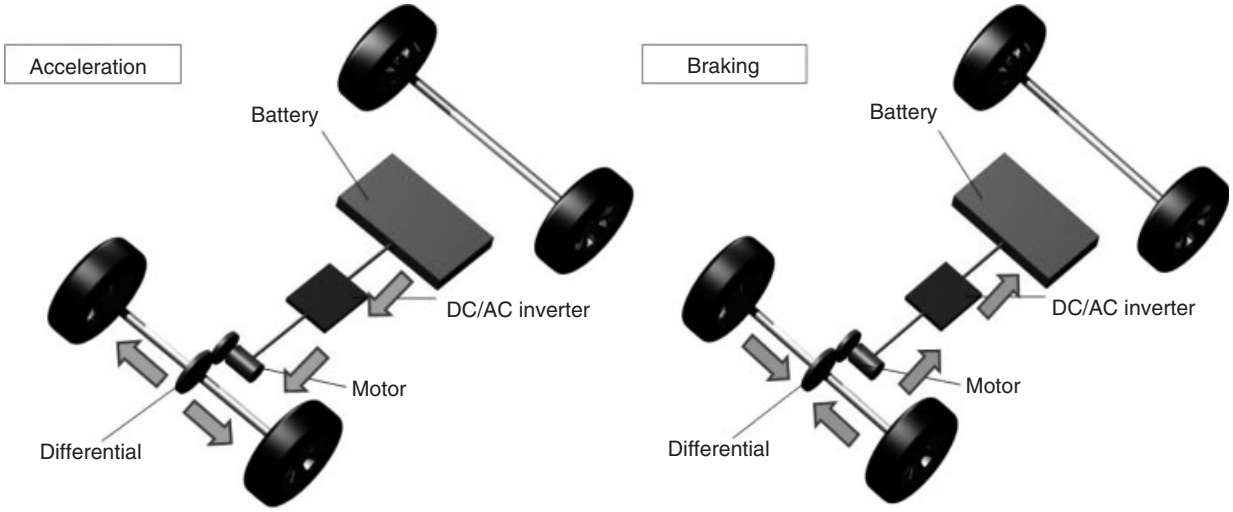


Figure 8. Cooperative regenerative braking system of the pure electric vehicle.

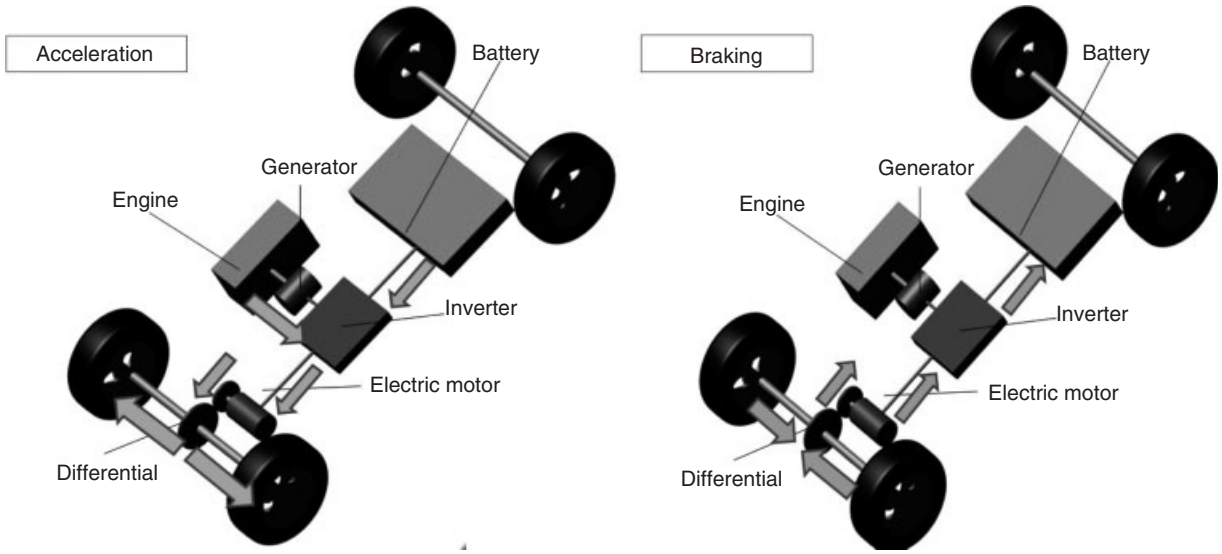


Figure 9. Cooperative regenerative braking system of the series-hybrid vehicles.

Series-hybrid vehicles are driven by the electric motor with no mechanical connection to the engine. Instead, there is an engine tuned for running a generator when the battery pack energy supply is not sufficient for demands (Figure 9).

In a parallel hybrid, the single electric motor and the internal combustion engine are installed so that they can power the vehicle either individually or simultaneously. In contrast to the power split configuration, typically, only one electric motor is installed. Most commonly, the internal combustion engine, the electric motor, and the gearbox are coupled by automatically controlled clutches. For electric driving, the clutch to the internal combustion engine is open, whereas the clutch to the gearbox is engaged. While in combustion mode, the engine and motor run at the same speed (Figure 10).

In a series-parallel hybrid electric drive train, there are two motors: an electric motor and an internal combustion engine. The power from these two motors can be shared to drive the wheels via a power splitter, which is a simple planetary gear. The ratio can be from 0% to 100% for the combustion engine, 0% to 100% for the electric motor, or an anything in between, such as 40% for the electric motor

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

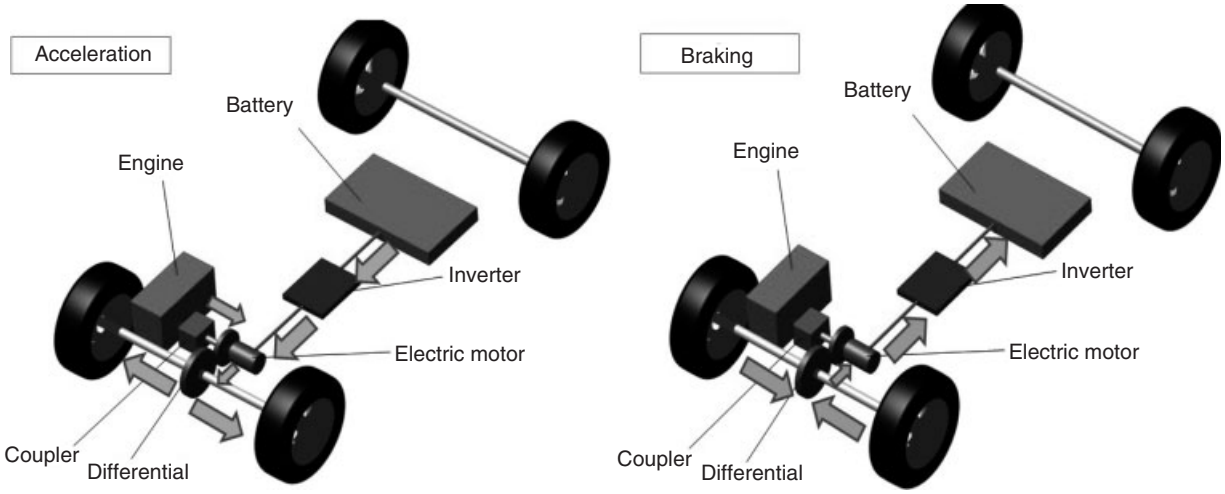


Figure 10. Cooperative regenerative braking system of the parallel-hybrid vehicles.

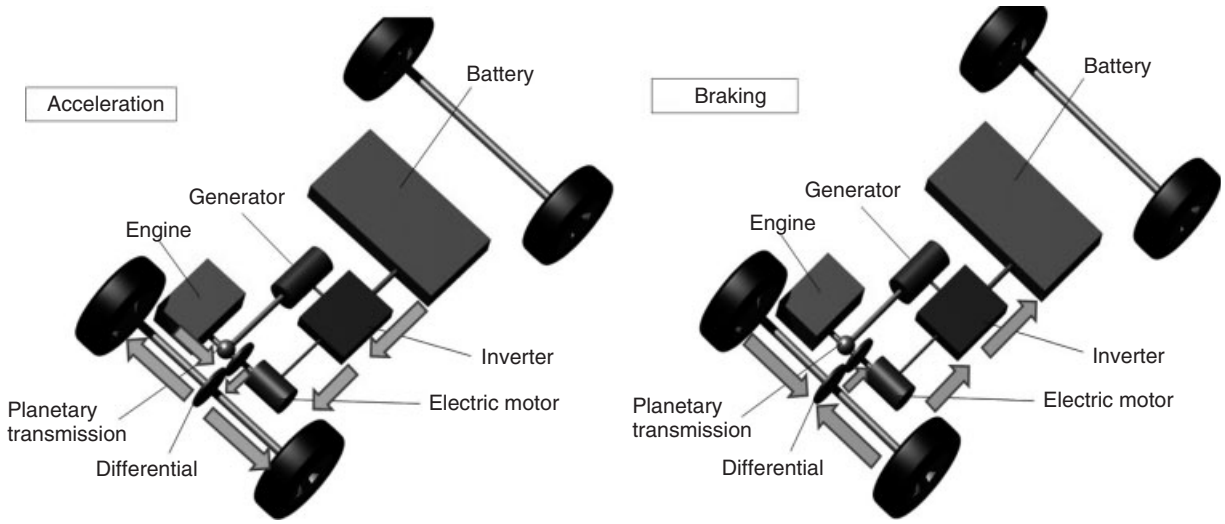


Figure 11. Cooperative regenerative braking system of the series–parallel hybrid vehicles.

and 60% for the combustion engine. The electric motor can act as a generator charging the batteries (Figure 11).

3.2.3 Regenerative braking in fuel cell vehicle

A fuel cell vehicle is a type of hydrogen vehicle that uses a fuel cell to produce electricity to power its electric motor. Fuel cells in vehicles create electricity to power an electric motor using hydrogen and oxygen from the air. The electricity produced by the fuel cell is delivered to the electric drive system in the vehicle, which converts electric power into the mechanical energy and drives the wheels of the car (Figure 12).

4 THREE PARTS OF THE COOPERATIVE REGENERATIVE BRAKING SYSTEM

From the point of view of control system theory, the cooperative regenerative braking system is made of three parts: signal part, control part, and actuator part (Zhang and Ning, 2009), shown in Figure 13.

1. *Signal Part.* The signals include the brake pedal position, vehicle speed, energy storage (e.g., battery) status, hydraulic pressure in the pipeline, pneumatic pressure in braking valves, etc., shown in Figure 14.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

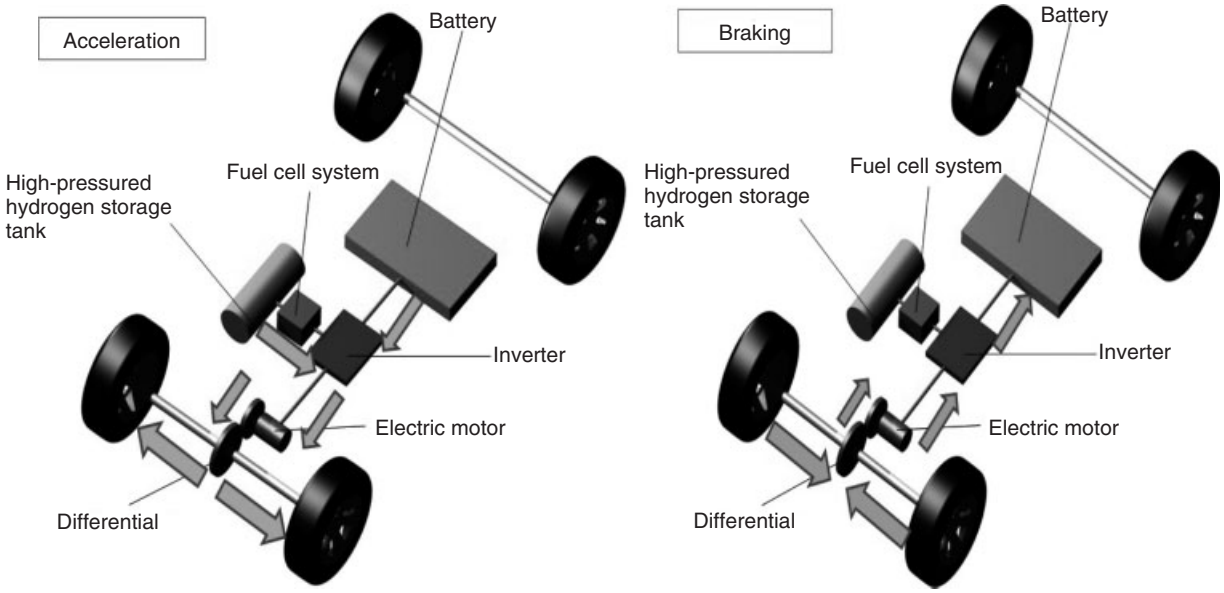


Figure 12. Cooperative regenerative braking system of the fuel cell vehicles.



Figure 13. Three parts based on control system.

The most important task in this part is to identify the driver's right intention in time.

2. *Control Part*. Electronic devices that control the regenerative braking function remotely by transmission parts and electronic control unit (ECU).
3. *Actuator Part*. Considering that there are two sub-braking systems, hydraulic or pneumatic braking system parts belongs to the friction braking system. Electric motor belongs to the regenerative braking system.

4.1 Signal part

The vehicle can be regarded as a closed control circuit consisting of driver, vehicle, and traffic condition. The driver gives the command based on the current traffic condition and personal intension. By steering, shifting gears or stepping down gas pedal or brake pedal, the driver can speed up or slow down the vehicle to adapt to the current outside environment and meet personal demands. Braking intention can be shown as how the driver controls the brake pedal.

4.1.1 The categories of braking intention

The braking intention can be divided into three categories: mild braking, moderate braking, and emergency braking (Zhang *et al*., 2009), shown in Figure 15.

Mild braking means the driver brakes, slows down, or stops the vehicle actively. It is common to see this kind of braking intention on the cross road when the driver is turning around the corner. In this case, the pedal angular velocity is small and the hydraulic pressure in the pipeline increases slowly (Yu *et al*., 2005).

Moderate braking means driver brakes passively. In this situation, the pedal angular velocity is big and the hydraulic pressure changes sharply. It is common to see this kind of braking intention on the highway when the front vehicle suddenly brakes or in case the vehicle needs to be stopped within a short brake distance. Sometimes, this kind of braking would be withdrawn very soon.

Emergency braking means the driver has stopped the car rapidly till the car completely stopped.

4.1.2 The ways to identify braking intention

In the cooperative regenerative braking system, by collecting signals such as brake pedal step force, pedal displacement and velocity, or hydraulic pressure change

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

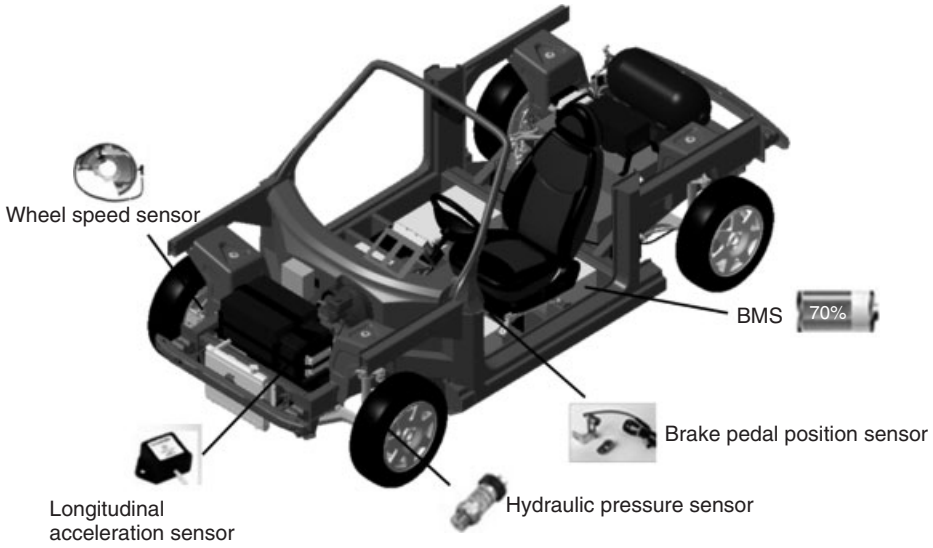


Figure 14. Signal parts.

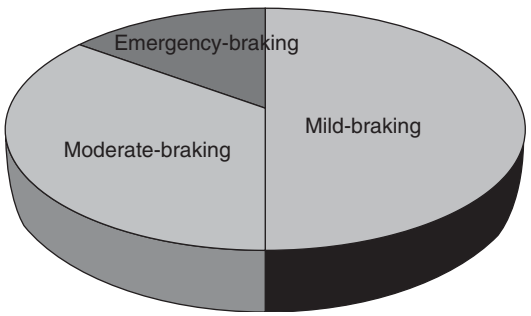


Figure 15. The categories of braking intention.

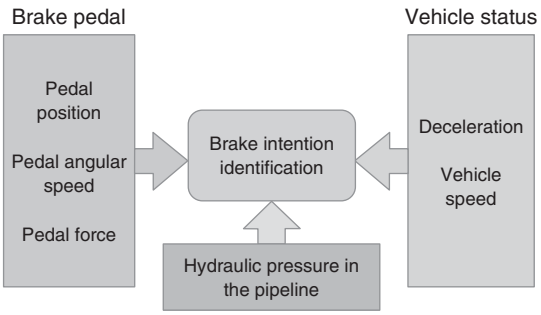


Figure 16. The signals that are used to identify the driver's braking intention.

rate in pipeline, shown in Figure 16, the driver's braking intention can be identified (Zhang *et al.*, 2009). At the same time, the rotation velocity of each wheel should also be monitored, to assure the vehicle is under safe condition.

4.2 Control part

4.2.1 Requirements for the braking system

Because there are two sub-braking systems in the cooperative regenerative braking system, its complexity is higher than the traditional braking system. The requirements for the cooperative regenerative braking system are as follows (Xiao, 2009):

1. sufficient brake force according to the driver's intention supplied;
2. the recovery efficiency of braking energy maximized;
3. the road adhesion coefficient best exploited to achieve the highest braking efficiency;
4. the driver's requirement about braking comfort met.

4.2.2 The role of the controller

The regenerative braking controller monitors the rotational speed of the wheels and the difference in that speed from one wheel to another. In vehicles that use these kinds of brakes, the brake controller not only monitors the speed of the wheels but also can calculate how much torque (rotational force) is available to generate electricity to be fed back into the batteries. During the braking operation, the brake controller directs the electricity produced by the motor into the batteries or capacitors. It not only makes sure that an optimal amount of power is received by the batteries but also ensures that the inflow of electricity is not more than the batteries can handle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The most important function of the brake controller, however, may be deciding whether the motor is currently capable of handling the force necessary for stopping the car (Lampton, n.d). If it is not, the brake controller turns the job over to the friction brakes, averting possible catastrophe. In vehicles that use these types of brakes, as much as any other piece of electronics on board a hybrid or electric car, the brake controller makes the entire regenerative braking process possible.

4.2.3 Influence of the position of the electric motor

As mentioned in Section 3, there are three types of cooperative regenerative braking system according to where the regenerative brake force acts. So obviously the position of electric motor has effect on the braking efficiency and vehicle stability. When the regenerative brake force acts on the front wheels, it can achieve better vehicle stability and braking efficiency because it can make better use of the road adhesion coefficient (Yu *et al*., 2008). When the regenerative brake force acts on the rear wheels, vehicle's stability is worse and it is not good for energy recovery. Therefore, Section 4.2.4 control strategy is based on the principle that electric motors supply regenerative brake force on the front wheels at first. When brake force is not enough, regenerative brake force would act on the rear wheels. In the case of emergency, electric motor and hydraulic braking system would work together to supply brake force.

4.2.4 Control strategy

For the vehicles, on which electric motors are just added to the original friction braking system, the control strategy is to control the electric motor to generate the designed brake torque that equals to the insufficient value between the realistic braking force distribution curve and ideal curve.

For the vehicles, on which the cooperative regenerative braking system is realized by "brake-by-wire" technology, the control strategy is more complicated. With the clutch connecting, the backward drag force from combustion motor should be considered. The following control strategy is under the circumstance without considering the backward drag force from combustion motor. In the case of mild or moderate braking, recovering kinetic energy is regarded as the priority goal. Therefore, in the case of mild braking, all the brake force would be supplied by electric motor. In the case of moderate braking, electric motor would supply brake force at first. The hydraulic braking system serves as a "backup" in case that the brake force is not enough. In the case of emergency braking, braking efficiency would be regarded as priority goal. In such circumstance, electric motor and hydraulic braking system would work together to supply the maximum brake torque within the shortest time (Zhang and Ning, 2009). The interconnectedness and mutual influence of electric motor braking and hydraulic braking should be considered. In this case, it needs help from ABS to keep the wheels unlocked. Considering system control simplification and motor protection, the motor's regenerative braking function of composite brake systems is usually turned off when ABS works (Figures 17 and 18).

4.3 Actuator part

Usually, for the friction braking system, hydraulic braking system is the actuator part whereas for the regenerative braking system, it is electric motor.

4.3.1 Electric motor

According to the PWM control theory, by switching the arm of the inverter conduction power tube, the armature current can be changed in opposite direction under the same magnet poles to drag the electric motor reversely. In this way, the power battery can be charged while part of the electricity is consumed by motor internal resistance in forms of heat. The energy stored in the battery can be used in idling, start, acceleration condition to achieve the purpose of energy saving and pollution reducing.

4.3.1.1 The limitations of electric motor regenerative braking.

1. Regenerative braking relies on the electric motors. Therefore, this kind of braking system cannot apply on the wheels that are not driven by electric motors (Zhang and Ning, 2009);
2. Because brake torque supplied by electric motors cannot exceed the maximum electric motor torque at current rotational speed, if pure regenerative brake torque cannot meet the current braking requirement, friction braking system should supply the rest brake torque (Zhang and Ning, 2009);
3. The energy regenerative efficiency relies on the battery SOC. The upper and lower limits of SOC depend on the internal resistance of battery. When SOC is too high, the power battery should not be charged in order to keep its normal operational life cycle. The recovery power of braking energy should not exceed the charge power of the battery (Zhang and Ning, 2009).
4. In the case of regenerative braking, when the rotational speed of electric motor is lower than the rated speed, the electric motor works with the rated brake

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

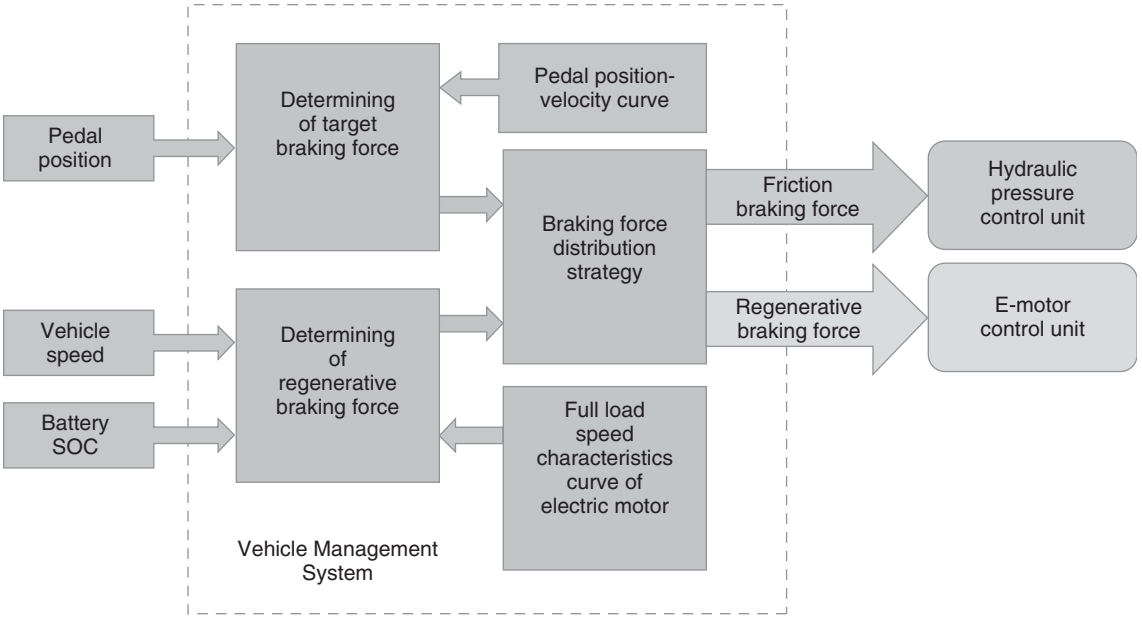


Figure 17. Control strategy.

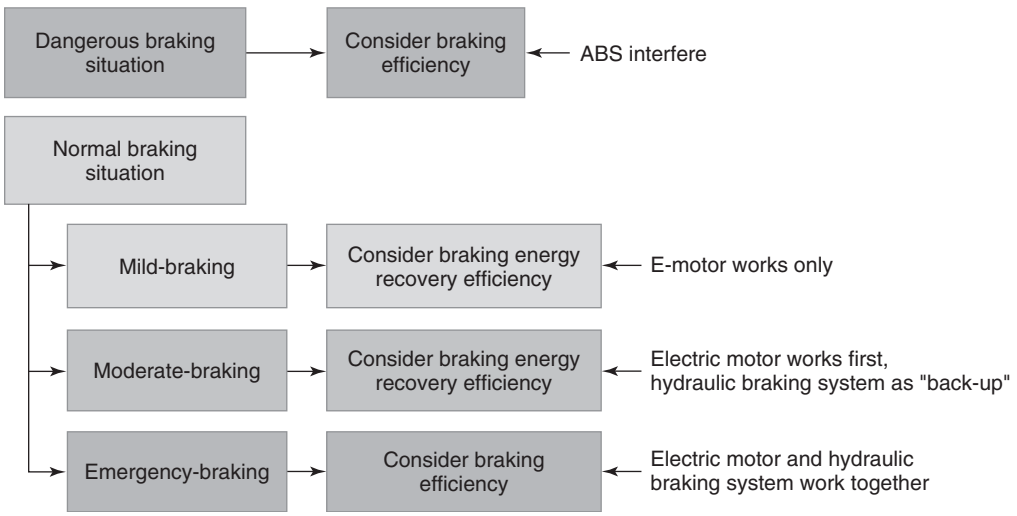


Figure 18. The control strategy under different braking intentions. (Modified from Zhang and Ning, 2009.)

torque. When the rotational speed of the electric motor is higher than the rated speed, the electric motor works under the rated power. When the rotational speed of the electric motor is too low, the regenerative braking fails to work because of the too low electromotive force. In this case, the regenerative braking force decreases to zero immediately (Zhang and Ning, 2009).

5. By emergency braking, the required braking torque may be many times higher than which the traction motor can supply. In this case, the conventional braking system is indispensable for a compensation of the braking torque to achieve the desired deceleration.

4.3.1.2 Different kinds of the electric motors. Most commonly used electric machines in electric vehicles are permanent magnet synchronous motors (PMSM), AC induction motors (IM), and switched reluctance motor (SRM). IM is simple-structured, rugged, low cost, and relatively mature in speed vector control technology. However, compared to the permanent magnet motor, it has a lower efficiency and power density.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

PMSM uses permanent magnet to replace its excitation system, and it has a higher power density, a higher efficiency, and a wider speed range. Vector control is often used when the drive system works in a low speed range, whereas the flux-weakening control is used in a high speed range.

SRM system features a compact and solid motor, suitable for high speed operation, with a simple and low cost drive circuit, performance reliable in a wide speed range with relatively high efficiency, and it can easily realize four-quadrant control. The disadvantage is the torque ripple and noise. In addition, by contrast to PMSM, the power density and efficiency is lower.

4.3.2 Hydraulic braking system

The traditional hydraulic braking system can be divided into two categories, with variable β and with fixed β (Zhang and Ning, 2009). Here, β is the braking force distribution coefficient. With variable β, the realistic brake force distribution curve is much closer to the ideal distribution curve. The technique of variable β is called *electric braking distribution* (*EBD*, shown in Figure 19). At present, with the application of ABS, wheel lock cases can be avoided. However, the biggest disadvantage is that ABS cannot realize regenerative braking. All braking energy is wasted.

In the cooperative regenerative braking system, if the hydraulic braking force can be controlled, the realistic brake force distribution curve can be much nearer to the ideal curve by adjusting the hydraulic brake force. Yet, currently in many cases, the cooperative regenerative braking system is realized by adding electric motors directly onto the original hydraulic braking system, the frictional brake force cannot be adjusted. In this case, the electric motor regenerative brake force can be adjusted in order to get near to the ideal distribution curve.

5 RELATED RESEARCH FIELDS IN THIS REGARDS

The cooperative regenerative braking system brings us more energy efficient vehicles, but at the same time, it brings more challenges for automotive engineers to ensure the vehicle safety, control stability, and riding comfort.

5.1 Brake feel

Brake feel (Zhang and Ning, 2009) can help the driver make the judgment, which includes the brake pedal displacement, the resistance force of brake pedal, and the vehicle braking performance which the driver can feel (Figure 20).

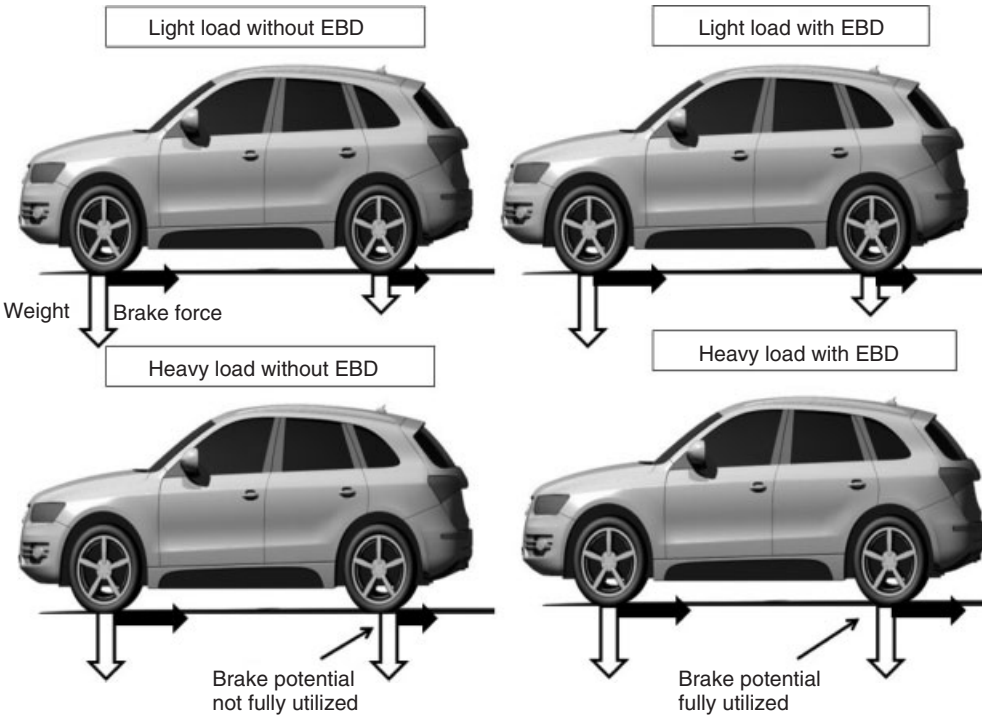


Figure 19. The influence of using EBD.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

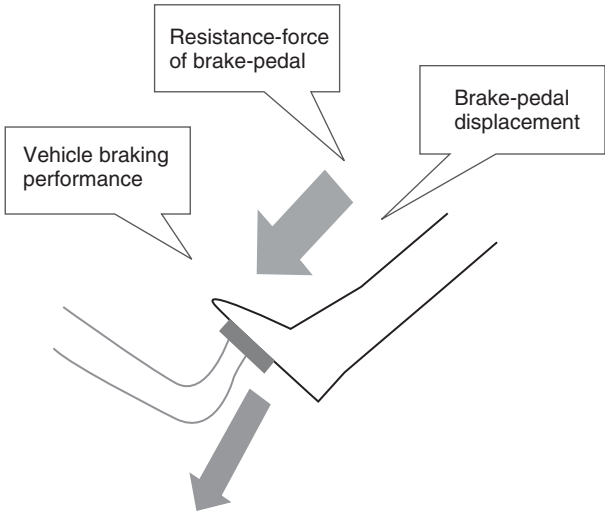


Figure 20. Brake feel.

Currently, some cooperative regenerative braking systems are realized by adding the electric motor systems directly onto the traditional friction braking system. Compared with hydraulic braking system, the electric motors respond faster. However, it brings a new problem, different brake feel. Therefore, if it needs to make sure of the same brake feel, automotive engineers have to postpone the active time point of the electric motor. Another problem is how much brake force should electric motor supply. The uncertainty increased with the change of the brake deceleration. To solve those problems, it is common to install an electric motor under the brake pedal to simulate the feedback of brake pedal force and displacement for the driver.

In addition, in the case of regenerative braking, brake fluid cannot flow to the wheel brake cylinder because of valve resistance. It results in a smaller brake pedal displacement. As the regenerative braking changed to friction braking because electric motor cannot supply torque when the rotational speed is too low, the valve switched from closed to open, it brings a sudden change in brake pedal displacement. At this point of time, definitely, it brings a sudden and significant bake feel change.

5.2 The influence of the cooperative regenerative braking system on ABS

To maximize the braking efficiency, hopefully the wheels would be in a lock/unlock state. ABS is a system that adjusts the wheel slip rate to make sure that the wheels in the right state. Because electric motor has advantages in fast response and easy control, usually the electric motors would act at first to supply the regenerative brake force (Zhang and Ning, 2009). If it is not enough, the traditional friction brake system would work as a "backup."

The motor's regenerative braking function is usually turned off when ABS works in order to simplify motor control and protect motor. For four-wheel-driven vehicles, in-wheel electric motors can work together with hydraulic system to realize the function of ABS to adjust each wheel's slip rate by controlling motor current and the wheel's brake torque in order to keep the wheel's ideal status.

For central electric-motor-driven vehicles, electric motor can only take the job of EBD. The function of ABS can only be realized by the hydraulic brake system. In the vehicles on which the electric motors are just directly added onto the traditional friction braking system, controlling the pressure of the brake fluid can adjust the wheel's slip rate.

5.3 The energy recovery sufficiency

Electric motor plays the role of energy transform in the cooperative regenerative braking system. There should be some parts to store the retrieved energy such as battery. Basically, there are three energy storage devices: battery, supercapacitor, and flywheel. At present, battery and supercapacitor would be the prior choices. Either one or some different kinds of devices can be used together in vehicles (Figure 21).

Energy recovery efficiency depends on the battery property, electric motor working property, the charging speed, etc. In the cooperative regenerative braking system, over charging and fast charging happen quite often. It makes the working environment of electric motor and battery very bad. In this case, the key point to get higher energy recovery efficiency lies in the electric motor control strategy as well as the battery energy storing technology.

6 CASES ABOUT THE COOPERATIVE REGENERATIVE BRAKING SYSTEM

Although the regenerative braking technology was first used in trolley cars, it has subsequently found its way into such unlikely places as electric bicycles and even Formula One racing cars. At present, these kinds of brakes are primarily found in hybrid vehicles, such as the Toyota Prius, and in fully electric cars, such as the Tesla Roadster. Followings are some cases showing how the cooperative regenerative braking system works in the hand of different automakers.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

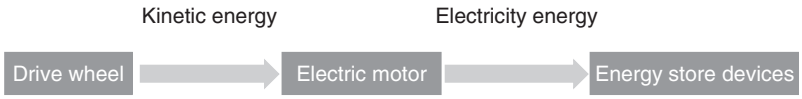


Figure 21. The theory of regenerative braking system.

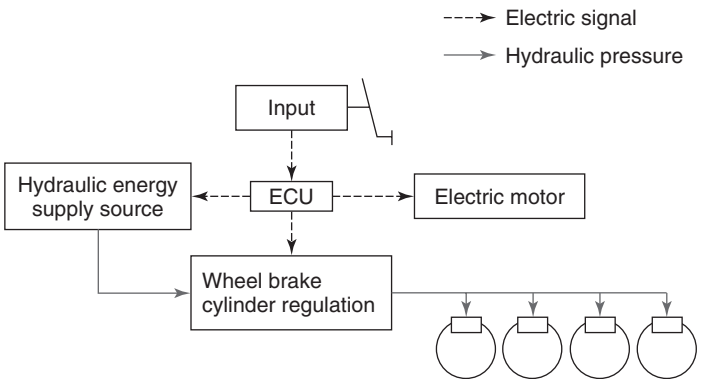


Figure 22. The regenerative braking system of Prius.

6.1 Toyota Prius

It is well known that Toyota Prius is the most popular hybrid electric vehicle in the world with the highest degree of industrialization and largest market share. The cooperative regenerative braking system in Prius is working based on the linear electromagnetic valve developed by Toyota, shown in Figure 22 (Lampton, n.d).

The design is based on the concept "brake by wire" (Zhang and Ning, 2009). Because the input from brake pedal controlled directly by the driver is separated from the brake fluid pressure, the pressure in the wheel brake cylinder can be controlled completely without the influence from the movement of brake pedal. When the driver steps down the brake pedal, the pedal displacement simulator transmits the information to the ECU instead that the pedal pushed the piston in brake main cylinder directly. The ECU decides how much brake force should the hydraulic braking system supply. With this feedback, the piston in the brake main cylinder would be pushed forward to generate the designed pressure.

6.2 Continental

Engineers in Continental developed a regenerative braking system for electric and hybrid vehicles (http://www.conti-online.com/generator/www/de/de/continental/engineering_services/themes/brakes_chassis/download/one_pager_regenerative_braking_pdf_uv.pdf). The design is also based on the concept "brake by wire." The components in this brake system include ESC with eGap (generator), electric vacuum pump, simulated brake actuation, and electromechanical brakes (Figure 23).

Cooperation between brake and power train system include

1. monitoring of driver brake request;
2. decouple driver demand from friction braking;
3. recuperation interface between brake and power train.

Different solutions for regenerative braking is as follows:

1. brake actuation with pedal force simulator (full by-wire braking);
2. Use of electrical brake caliper (by-wire braking at one axle);
3. eGap (the generator) realized by electronic brake system (conventional hydraulic brake system).

7 SUMMARY

Although there is still a long way to go for the cooperative regenerative braking system to achieve the same degree of the success and popularity as ABS, one thing seems certain: the application of regenerative braking is an irreversible trend. As hybrids with electric motors and regenerative brakes can travel considerably farther on a gallon of gas, some has achieved more than 50 miles per gallon at

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

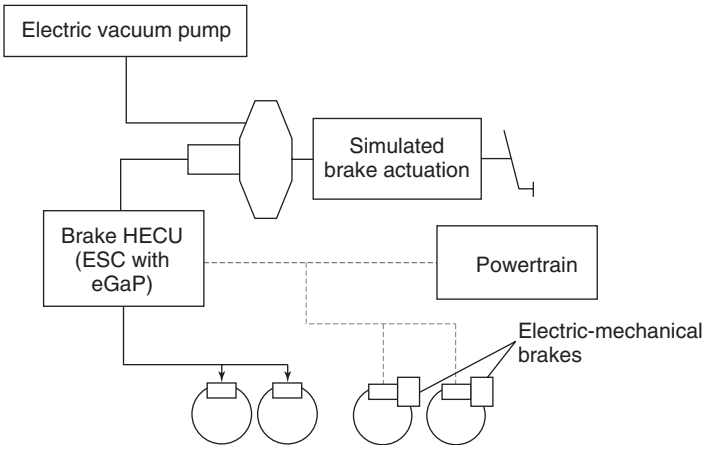


Figure 23. The components in the Continental regenerative braking system.

this point. That is something that most drivers can really appreciate.

RELATED ARTICLES

Overview of Electric, Hybrid and Fuel Cell Vehicles
Power and Energy Requirements for Electric and Hybrid Vehicles
Regenerative Braking Systems
Energy Management Systems of EVs
Energy Management Systems of HEVs
The Development of Alternative Brake Systems

REFERENCES

Gubler, D. (2009) IORE beim Torneträs, http://en.wikipedia.org/wiki/File:IORE_beim_Tornetr%C3%A4sk.jpg (accessed 9 October 2013).

Harris & Ewing, Inc. (1906) Senator Wetmore in Automobile, http://it.wikipedia.org/wiki/File:SenatorWetmoreInAutomobile_retouched.jpg (accessed 9 October 2013).

Lampton, C. (n.d) How Regenerative Braking Works in *HowStuffWorks*, http://auto.howstuffworks.com/auto-parts/brakes/brake-types/regenerative-braking.html (accessed 9 October 2013).

Xiao, K. (2009) Algorithms research on braking force distribution of cooperative braking system *Beijing Vehicle*, **32** (2), 42–46.

Yu, Z., Xiong, L., and Zhang, L. (2005) Matching research on electrohydraulic cooperative brake *Automotive Engineering*, **27** (4), 455–462.

Yu, Z., Zhang, Y., Xu, L., *et al*. (2008) Brake force coordination distribution methods simulation in cooperative braking system *Train Technology*, **5**, 1–5.

Zhang, Z. and Ning, G. (2009) Vehicle mechanical-hydraulic cooperative braking system *Shanghai Vehicle*, **11**, 42–46.

Zhang, Y., Yu, Z., Xu, L., *et al*. (2009) Cooperative braking system brake force distribution strategy research based on the brake intentions *Automotive Engineering*, **31** (3), 244–249.

Zheng, H. (2010) An research on anti-lock braking control method of 4WD electrical vehicle with in-wheel motors based on variable structure control. Tongji University Graduation Thesis for Master Degree, 3.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto028
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Torque Vectoring for Drivetrain Systems

Claus Granzow and Matthias Arzner
ZF Friedrichshafen AG, Friedrichshafen, Germany

1 Torque Vectoring 1
2 Torque Distribution in the Driveline 1
3 Influencing Vehicle Handling with Torque Vectoring 3
4 Mechanical Concepts of Modern Torque-Vectoring Axle Drives 5
5 ZF Vector Drive© Concept 5
6 The Mitsubishi Concept 5
7 The Magna Concept 7
8 Torque Splitter and Limited Slip Differentials 7
9 Actuating Systems 7
10 Further Details of a Torque-Vectoring System Using the Example of ZF-Vector Drive© 9
11 Discussion and Conclusion 12
Related Articles 12

1 TORQUE VECTORING

The term *torque vectoring* describes the active generation of a vehicle yaw moment by a directed distribution of input torques over the left and right sides of the vehicle. Torque vectoring can be used to influence the degree of yaw of the vehicle and in this way to actively control the driving dynamics (active yaw control).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 TORQUE DISTRIBUTION IN THE DRIVELINE

To avoid stress on the driveline and tire wear during cornering, a transmission is needed that provides free speed compensation. These compensating or differential transmissions create a fixed torque distribution over the two outputs, while making different rotational speeds possible.

Relating to one axle, that is, in the transverse vehicle direction, the differential distributes the torque 50%/50% on the two wheel sides. In the longitudinal vehicle direction of a multi-axle-driven vehicle, differentials are also used for the necessary speed compensation. Not only equal distribution, but also asymmetric torque distributions are possible here (e.g., 33%/67%, 38%/62%). They are selected depending on the traction potential of the axles and the desired handling behavior.

Although differentials are unavoidable for passenger cars, a permanently fixed torque distribution has disadvantages in many driving situations. Widely different road-tire friction coefficients of the wheels on one or different axles can cause individual wheels to build up high slip that limits the torque transfer capability. The wheel with the lowest friction coefficient determines the total transferable torque of the driveline. This means that free speed compensation impairs the traction behavior on different road surfaces. For instance, if one wheel is on ice, it builds up a high level of tire slip, whereas all the other wheels on asphalt cannot transfer any more torque than the wheel rotating on the ice. This gives the vehicle a poor start-up performance.

This can be remedied by limited slip differentials (LSDs) that can prevent the compensating effect in critical driving situations. Principally, a distinction can be made between three types of LSDs:

- Shiftable black–white differential locks that can lock the speed compensation by 100% (automatically or manually).
- Self-locking differentials either in torque-sensing or in speed-sensing design. The first type locks the speed compensation as a function of the torque to be transmitted, whereas the locking rate of the second type represents a function of the differential speed.
- LSD that can be activated externally. These systems are usually electronically controlled and feature a hydraulic or electromechanical actuator. Usually, the locking rate is continuously variable between 0% and 100% and is determined by a strategy depending on the driving situation.

2.1 Definition of locking value

The locking value S is a characteristic value that defines the degree to which the compensating effect of the differential is prevented. The definition is as follows:

$$S = \frac{\text{Locking torque } T_B}{\text{Input torque } T} = \frac{T_{right} - T_{left}}{T_{right} + T_{left}} \quad (1)$$

The possible value range is between 0% and 100%. An ideal open differential has a locking value of $S = 0\%$, whereas a fully blocked differential has a locking value of $S = 100\%$. Normal open differentials have locking values of approximately $S = 5–10\%$. Torque-sensing limited slip differentials have a constant locking value.

Limited slip differentials can be used mainly to increase traction by compensating any slippage that occurs. The drive torque is largely transmitted to the vehicle side or vehicle axle that still has potential for transmission. The originally fixed torque distribution between the differential outputs (e.g., 50%/50%) is superimposed by the differential locking torque. As a basic principle, a limited slip differential only enables torque transmission from the side that is rotating faster to the side that is rotating more slowly. When a vehicle drives round a curve with an even friction coefficient, actuating the limited slip differential tends to lead to understeering. The limited slip differential can be used to stabilize the driving dynamics.

Torque-vectoring systems expand the function of controllable limited slip differentials. The directed asymmetric torque distribution over one axle can be used to apply a yaw moment around the vehicle vertical axis. Beside the differential locking effect, a torque-vectoring system must also be capable of transmitting torques to the shaft that is rotating faster and therefore increasing the difference in speed of rotation between the left and right vehicle sides. During cornering, it is possible to affect an inward turning; this means a more agile vehicle handling.

2.2 Definition of vectoring torque

The difference between the two-wheel torques (in the direction of travel) is called the *vectoring torque of an axle*:

$$T_{TV} = T_{right} - T_{left} \quad (2)$$

Positive vectoring torques lead to a counterclockwise yaw moment of the vehicle around the vertical axis. The maximum vectoring torque depends on the vehicle moment of inertia around the vehicle vertical axis as well as the required operating range of the torque-vectoring system. Typically, affecting traction at low speeds requires higher values of vectoring torques than affecting driving dynamics.

2.3 Hang-on clutches and torque splitters

As an alternative to differentials with superimposition units (limited slip differentials or torque-vectoring units), there is also the option of achieving speed compensation using a slipping multidisk clutch. Here, there is no differential, because the required torque is completely transmitted via a multidisk clutch operated during slipping. A basic distinction is made between clutches for torque distribution in longitudinal and transverse configurations.

Clutches that distribute the torque between different axles replace an interaxle differential and are called *hang-on clutches*. A permanent all-wheel-drive system with hang-on clutch requires a specially designed clutch as well as a suitable actuating system and operating strategy. The torque flow to the hang-on or secondary axle can be freely selected within the range of the speed conditions according to aspects of traction and driving dynamics. The strategy must also achieve the speed compensation in case of cornering.

The differential relating to one axle as speed compensation in the transverse direction can be functionally replaced by a twin clutch system. Here, the two wheel drive torques are set by the respective clutches and used individually for influencing the traction and driving dynamics. Twin clutch systems are also known as *torque splitters* because they are only able to individually distribute the input torque to the two drives. In contrast, torque-vectoring systems with the above-mentioned superimposition units can generate wheel differential torques and therefore yaw moments independently of loads.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

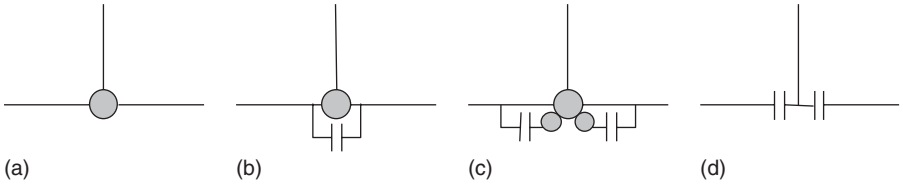


Figure 1. Diagram of the basic elements of transverse torque splitting. (a) Open differential with fixed torque distribution (50%:50%), full speed compensation. (b) Differential with lock and basic distribution (50%:50%), active compensation of differential speeds, torque transfer from faster to slower sides, mainly used to increase traction. (c) Differential with superimposition unit and basic distribution (50%:50%), additional torque can be superimposed, torque transfer to faster side also possible, active yaw rate control possible. (d) Twin clutch system without basic distribution, individual distribution of the input torque possible, no generation of differential torques in coast situation, also covers active longitudinal distribution (hang-on system).

2.4 Speed error

Another essential design criterion apart from the maximum vectoring torque is the speed error. The speed error is the maximum speed difference of the two output shafts in percentage until which torque transmission to the faster rotating output shaft is possible. At the vehicle level, this is the minimum curve radius at which agility-increasing vectoring activities are still possible. While stabilizing or traction-increasing interventions are possible under any speed conditions, agility-increasing interventions are not possible below a minimum curve radius.

Very large speed errors collectively cause increased friction power on the clutches. That is why, it is important when determining the speed error to weigh up the smallest curve radius that is still relevant for torque-vectoring interventions and what collective friction power can be collectively covered by the clutches.

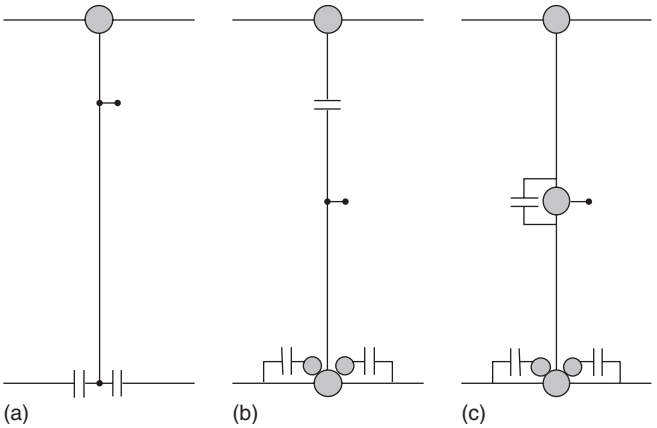


Figure 2. Diagram of driveline examples with active transverse torque distribution. (a) Twin clutch on rear axle, front axle as primary axle (e.g., Honda SH-AWC). (b) Rear axle torque vectoring and hang-on front axle (e.g., BMW X6). (c) Rear axle torque vectoring and Torsen center differential (e.g., Audi Sportdifferential).

2.5 Torque distribution strategies/designs

Figure 1 shows the basic elements of transverse torque distribution. Together with the various options for longitudinal torque distribution, this can be used to arrange various drivelines. Figure 2 shows some driveline configurations available on the market.

3 INFLUENCING VEHICLE HANDLING WITH TORQUE VECTORING

The free distribution of torques between the four vehicle wheels can be used to actively influence the vehicle handling. The way this works is comparable to the steering principle of a tracked vehicle. When the driver of a tracked vehicle wishes to change the travel direction, different track speeds are applied to the inner and outer tracks. The different speeds cause the vehicle to turn around its vertical axis.

Torque-vectoring systems in a passenger car driveline influence the direction of travel in a similar way. Depending on the system characteristics, this can occur depending on or independently of the current input torque of the vehicle. Simple torque splitters can only influence the ride direction when there is input torque. To utilize the effect, the driver must consciously select trailing throttle or acceleration mode.

Torque-vectoring systems with superimposition function can generate the required differential torque without limitation by the input torque. They offer a much wider scope for influencing the driving dynamics. As they work independently of the input torque, they can also be seen as components purely for influencing vehicle dynamics, comparable with steering systems.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.1 Yaw moment by torque vectoring

The vectoring torque, also called *wheel difference torque*, occurs due to different longitudinal forces on the two wheels of a vehicle axle. In the simplest case, these are two longitudinal wheel forces with identical magnitude and preceding plus or minus signs. However, in most cases, the differential torque is superimposed by symmetrical drive or braking forces.

According to the track width of the vehicle, the wheel difference force leads to a free torque around the vehicle's vertical axis. This torque influences the yawing motion of the vehicle and can be used to change the direction of travel.

The complex friction conditions on the vehicle tires can influence the effect of torque vectoring especially during cornering. In respect to the traction limit given by the circle of forces, the lateral wheel force can be weakened by superimposed longitudinal wheel force. Superimposed vectoring torque reduces the amount of lateral wheel force in the same way as additional drive or braking torque. In any case, the additional longitudinal force causes the amount of lateral wheel force and generates an additional influence on the vehicles yaw behavior.

Depending on the place of installation of the torque-vectoring device at front or rear axle and the sign of the vectoring torque, this effect can reinforce or weaken the effect on the vehicle yaw moment.

3.2 Inward- and outward-turning differential torque

The yawing motion of the vehicle around its vertical axis can have a stabilizing effect or can increase the dynamics. In principle, a difference is made between an inward-turning and an outward-turning torque.

Inward-turning torque supports vehicle cornering by reinforcing the yawing motion around the vertical axis. The inward-turning torque reduces the curve radius. The outward-turning torque slows the vehicle yawing motion and increases the curve radius. This is why the outward-turning torque is used to stabilize the vehicle. Figure 3 describes the relation between direction of yaw torque and vehicle behavior during cornering.

3.3 Influencing self-steering using torque vectoring

For reasons of safety and controllability, modern vehicles built to today's standards are designed to have a general understeering behavior at the driving limits. This means that with increasing lateral acceleration, the steering angle

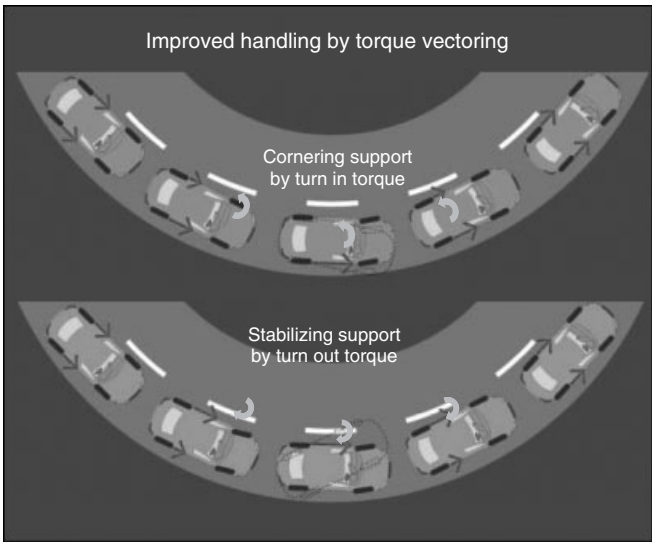


Figure 3. Influencing the driving dynamics with the torque-vectoring function. Agility-increasing and stabilizing yaw moment.

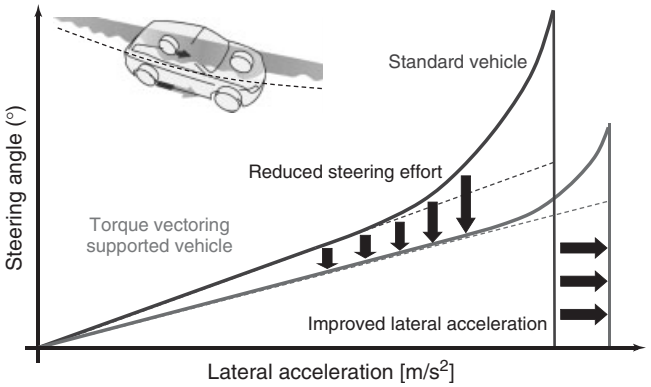


Figure 4. Altering the self-steering response by superimposing the torque-vectoring function.

needed to further reduce the curve radius increases disproportionately. The maximum possible cornering forces of the front tires limit the minimum curve radius. The vehicle stability limit is indicated to the driver by a stable vehicle with slipping front tires. The relation in terms of driving dynamics between the steering angle and lateral acceleration is also called the *vehicle self-steering effect* (Figure 4).

This understeering tendency of the vehicle is intensified when a braking or input torque is additionally applied to the steered axle. Especially in the case of front-wheel or all-wheel-drive vehicles, this is called *load understeering*.

In combination with a suitable, dynamic steering system, a torque-vectoring system offers the option of freely influencing the self-steering effect. In most cases, it is used

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to reduce or compensate for the understeering tendency described. If, due to overloading or high cornering speed, the front axle reaches the limits of its cornering force, the yawing motion of the vehicle can be increased by the application of a free yaw moment around the vertical axis. The vehicle continues to follow the driver's steering command and behaves in a sporty and neutral way.

In particular, drivers often find the responsive reaction of the vehicle to their steering commands and the very late or even complete absence of understeering a very positive characteristic.

In general, the torque-vectoring unit is actuated by means of a driving dynamics control. In a similar way to a brake-based stability system (ESC), this monitors the driving condition of the vehicle by comparing a measured vehicle yawing rate with an ideal yawing rate calculated from the steering angle and the vehicle speed. The great advantage of torque-vectoring-based regulation of driving dynamics is the system's continuous mode of operation. It corrects the direction of travel gently and smoothly without any speed-reducing braking interventions.

3.4 Driving dynamics limits of torque-vectoring systems

As with any driven wheel, the torque transmission capability is limited by the friction coefficient and the vertical tire force. The vertical tire force is particularly important because it can change dramatically during cornering.

Constant cornering produces a positive wheel load distribution on the two wheels on the outside of the curve. The force potential gained here can be used to supply the outer wheels with additional input torque. In this way, the wheel load distribution toward the outside of the curve supports the generation of an inward-turning yaw moment.

In the opposite case, the load is reduced on the two inner wheels of the vehicle during high lateral acceleration. The accompanying reduction in wheel force potential limits the force buildup of a stabilizing yaw moment. Therefore, a torque-vectoring system is particularly suited to creating an inward-turning yaw moment.

4 MECHANICAL CONCEPTS OF MODERN TORQUE-VECTORING AXLE DRIVES

Generally, a torque-vectoring system is understood as an axle drive that can generate a positive or negative wheel differential torque irrespective of the drive situation.

The basic power distribution is always provided by a differential compensating transmission. The differential decouples the two wheel speeds and distributes the input torque between both wheels, ideally 50% per wheel. The function can be performed by a conventional bevel gear differential. For reasons of space saving and effectiveness, a spur gear differential is used in some applications.

The torque-vectoring function itself is achieved with the help of so-called *superimposition units*. They make it possible to apply alternative power flows inside the transmission and enable controlled power distribution between the two wheels.

In general, torque shifting is required bidirectionally between both wheels of an axle (inward- and outward-turning redistribution), so two separate superimposition units are installed in a transmission. Here, each unit takes care of one redistribution direction. Depending on the transmission concept, the units can be installed symmetrically on the left and right of the basic axle transmission (see ZF and Magna concept as described in Sections 5 and 7), or arranged nested on one side of the axle drive (Mitsubishi).

A single superimposition unit always consists of a friction clutch and a transmission stage. While the friction clutch provides the actuation moment necessary for torque distribution, the transmission stage ensures the speed error required for initiating and intensifying the torque-vectoring effect.

5 ZF VECTOR DRIVE© CONCEPT

The ZF Vector Drive© axle drive consists of a conventional bevel gear differential and two symmetrically arranged superimposition units. The superimposition unit consists of a planetary gearset, which itself consists of two sun gears, one planet carrier, and three planetary gears. The inner sun gear is connected to the differential cage, whereas the outer sun gear is connected to the relevant side shaft. A multidisk brake acting between the planet carrier and the transmission housing acts as a torque modulation element (Figures 5 and 6).

With the help of a roughly 11% speed error in the planetary gearset, the application of a braking torque achieves a power flow from the axle drive via the planetary gearset to the individual wheel.

The ZF Vector Drive© transmission was used for the first time in volume production in 2008 in the BMW X6.

6 THE MITSUBISHI CONCEPT

The classic, asymmetric torque-vectoring transmission design is called the *Mitsubishi concept*. In the current version, it features a planetary-designed axle drive. The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

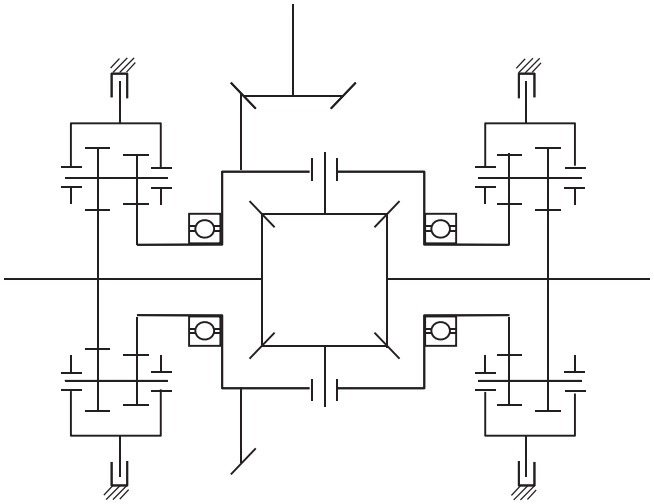


Figure 5. Schematic diagram of the ZF Vector Drive© system.



Figure 6. Sectional drawing of the ZF Vector Drive© system.

advantages of this design are the low self-locking value, the narrow construction shape, and the direct action on the opposite output shaft.

Added to the axle drive is a nested superimposition unit installed on the right side. The input and output of the superimposition unit are each connected to the two side shafts of the axle drive. The superimposition unit itself consists of two multidisk clutches and two transmission stages allocated to the individual clutches (Figures 7 and 8).

The difference is that the transmission stage on the inner clutch generates increased slippage compared to the wheel differential speed, and the transmission stage on the outer clutch provides a reduced clutch slippage compared to the wheel differential speed. Depending on which clutch is

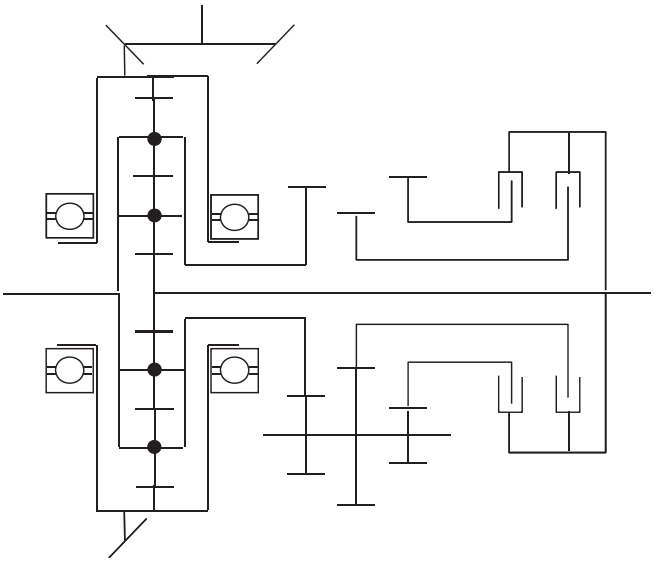


Figure 7. Schematic diagram of the Mitsubishi torque-vectoring system.

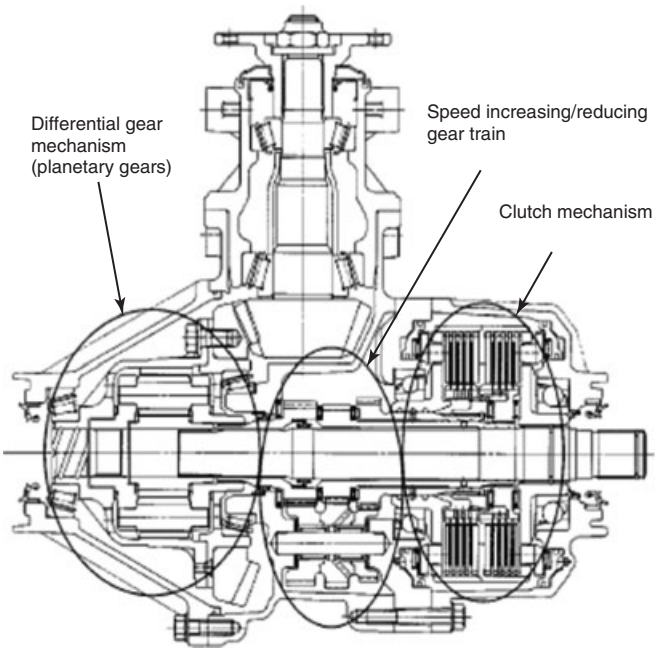


Figure 8. Sectional drawing of the Mitsubishi torque-vectoring system. (Photo: Mitsubishi.)

actuated, the different differential speeds of the two clutches can create a specific power flow directly between the two rear wheels.

The first time a torque-vectoring transmission was used in volume production was by Mitsubishi in 1996 in its Lancer Evolution 4 model. That makes Mitsubishi along

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

with Honda one of the pioneers in the application of this technology. The planetary basic differential design presented here has been in use in the Lancer model since the Lancer Evolution 8 generation.

7 THE MAGNA CONCEPT

The Magna concept also has a symmetrical design, consisting of a bevel-gear-based drive unit and two individual superimposition units attached outboard. The superimposition units each feature two individual, single-stage transmissions connected by a multiplate clutch. Each of these transmission stages consists of a sun gear and a ring gear, which has a radial displacement in respect to the sun gear. The sun gear of the inner transmission stage is connected to the differential cage, and the sun gear of the outer transmission stage is connected to the output shaft. To generate the required speed error within the superimposition unit, the ratios of the two stages differ slightly by about 10%.

Friction clutches are used to control the torque distribution. The two wet-running clutches are each installed between the two ring gears of each superimposition unit and also run with a radial offset to the drive units' output axis (Figures 9 and 10).

When the clutch is actuated, the positive speed error creates a power flow from the axle drive to the wheel via the first transmission stage, the clutch package, and the second transmission stage.

The Magna torque-vectoring transmission was introduced for the first time in 2009 in the Audi S4 under the name "Sportdifferential", and then made available as an optional extra in further Audi models.

8 TORQUE SPLITTER AND LIMITED SLIP DIFFERENTIALS

The two types torque splitter and limited slip differential are special forms of torque-vectoring axle drives. They also enable splitting of the input torque between the wheels, although with the above-described limitations (Figure 11).

The torque splitter does not use a conventional differential-based axle drive. The power or torque splitting is achieved directly by two control elements that control individual wheels. The control elements usually consist of friction clutches that are connected either directly or with the help of planetary gearsets. In both cases, a sophisticated ratio selection must ensure that both clutches have a differential speed in order to generate a positive power flow to the individual wheel.

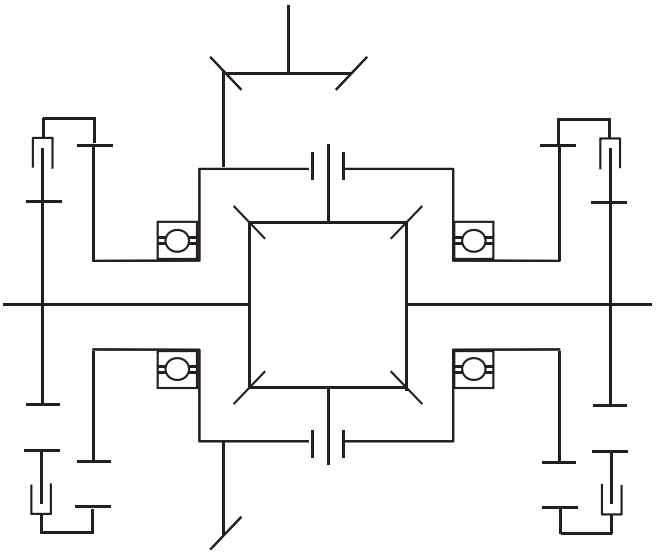


Figure 9. Schematic diagram of the Magna torque-vectoring system.

Usually, an electronically controlled limited slip differential consists of a bevel gear differential and an additional multidisk clutch. The clutch is installed between the differential cage and one of the side shafts. It enables a direct flow of power between the differential cage and the side shaft.

9 ACTUATING SYSTEMS

A further distinction between different torque-vectoring transmissions is the selected concept of actuating system. The whole range of mechatronic designs can be used to actuate the multidisk clutches installed. Apart from classic hydraulic actuation, electromechanical and electromagnetic actuations are also used for products in volume production.

Each of these technologies comes with advantages and disadvantages. Electrohydraulic actuation features a high power density and enables central energy generation for both clutches. Compared to separate actuators for each clutch hydraulic actuation can have favorable effects in terms of installation space and weight, but have to be weighed up against disadvantages in efficiency and dynamics at low temperatures.

Electromechanical actuation is attractive due to its integrated position sensing with very precise controllability and high actuating dynamics. Another advantage is the temperature resistance especially under cold conditions. Viscosity

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

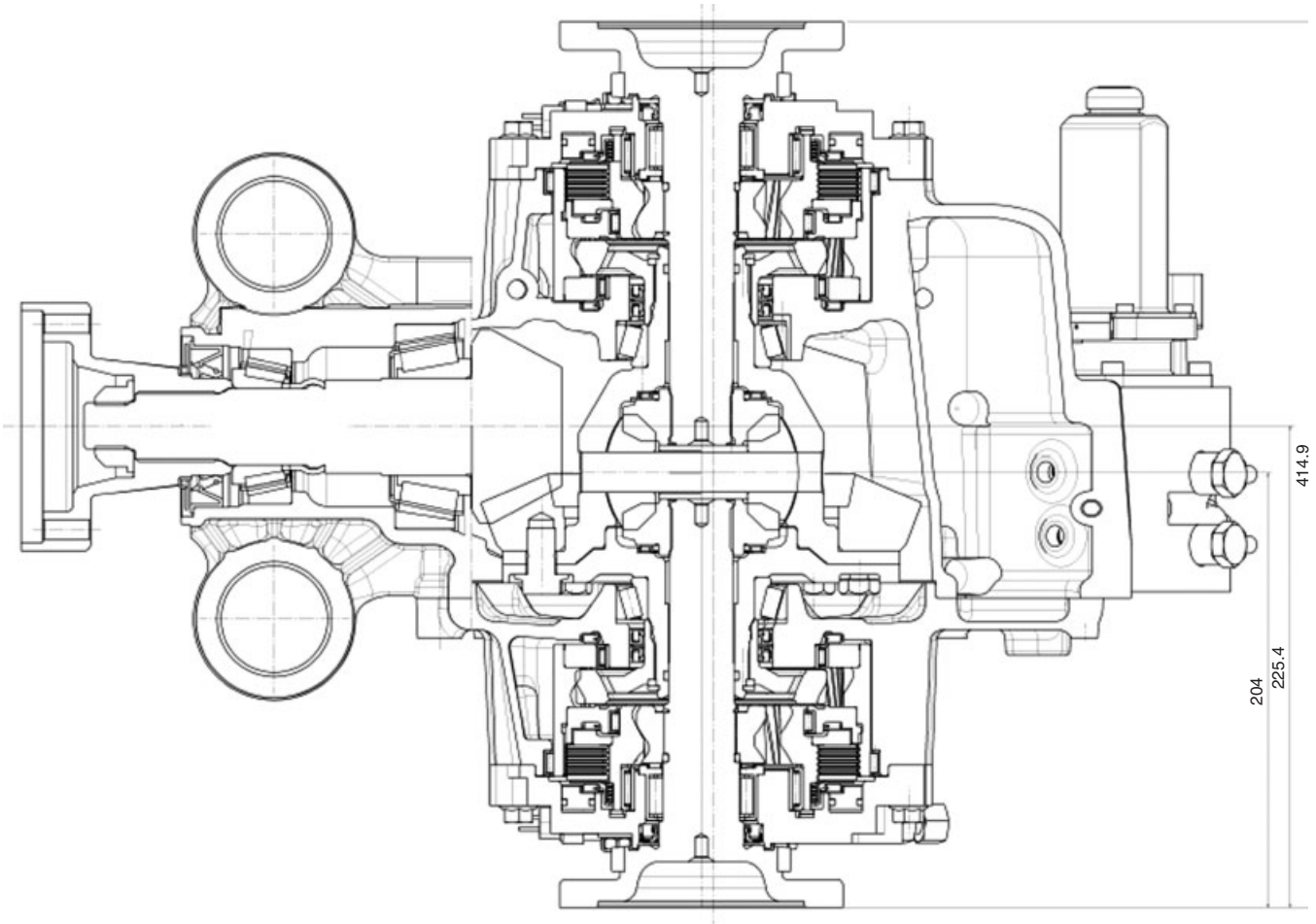


Figure 10. Sectional drawing of the Magna torque-vectoring system. (Reproduced from Sackl, W. and Sankar, M. (2006) 'Simulation and Definition of an Active Yaw Control Device', presented at *7th All Wheel Drive Congress*, Graz, Austria.)

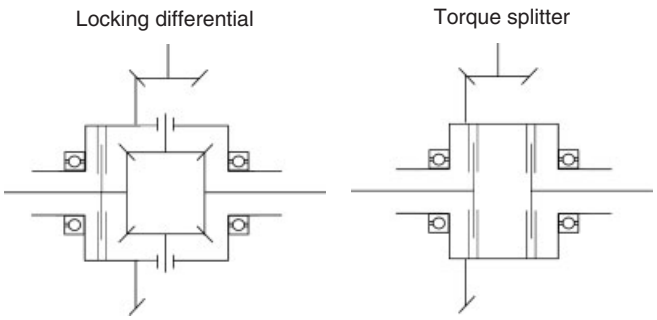


Figure 11. Schematic diagram of a limited slip differential and of a torque splitter.

effects only have minor influence on the actuation function. Compromises must be made in the package and weight because the actuator mechanics cannot be installed freely on the drive unit.

Both electrohydraulic and electromechanical actuation use electric motors to generate the control power. When high dynamic handling is required, the motors provide for short-term higher energy takeoff from the vehicle main power system.

The advantage of electromagnetic systems is their better package and high dynamics. The concentric arrangement of the actuator coils around the side output shafts allows for compact, space-saving installation. The disadvantages of direct magnet actuation are higher weight and lower energy density. To generate high clutch torques, large and heavy coils must be supplied with high currents. These disadvantages can be avoided by using electromagnets to actuate only a pre-control clutch. The pre-control clutch turns a ball-ramp mechanism, using the rotary motion from the driveline. Clutches based on this principle are a common and very cost-effective solution for many all-wheel applications.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

10 FURTHER DETAILS OF A TORQUE-VECTORING SYSTEM USING THE EXAMPLE OF ZF-VECTOR DRIVE©

10.1 Torque Bias Effect

The direct influence of the self-locking coefficient on the control accuracy of a modern torque-vectoring drive unit is called *torque bias effect*. This effect occurs when the drive unit is included in the power flow for generating the differential torque (see ZF and Magna concept as described in Sections 5 and 7).

Depending on the type of load (trailing throttle, acceleration, or coast) and the required vectoring torque, a residual torque occurs within the drive unit. This residual torque relates to the differential itself and is not transferred by the superimposition unit.

Owing to the locking value S of an open bevel gear differential, the drive unit generates an outward-turning differential torque. The sum of the outward-turning differential torque and the vectoring torque by the superimposition unit determines the overall torque split between the two output shafts.

Depending on the value of the self-locking coefficient, the differential torque in the (very important) case of inward turning is weakened, and in the (less important) case of outward turning, it is strengthened. A self-locking coefficient that is as low as possible and remains constant over the entire lifetime is key to the constant and efficient functioning of a torque-vectoring system.

Figure 12 shows the torque flow of the ZF Vector Drive© system during accelerated cornering in connection with an inward-turning differential torque.

10.2 Actuation

Active control of self-steering properties and of the driving-dynamics-related vehicle response through highly dynamic transversal distribution of wheel torques requires an actuator that can perform very quick and precise actuations. Electromechanical actuating systems are compact and very efficient, with low power consumption. The torque of the E-motor is boosted by a spur gearset before it is transformed into an axial force via a ball ramp. The axial force is used to pressurize the multiplate disc and thus for the transfer of the specifically desired torque (Figure 13). The geometry of the ball ramp can be designed in such a manner that ensures the fastest-possible free travel with optimal resolution and controllability of the operating range. The adjustment time from 0% to 90% of the maximum possible differential torque of 1800 Nm is just 80–100 ms. Figure 14 gives an example of control accuracy measurement with a step-type vectoring- torque request.

10.2.1 Electric motor

In addition to the requirements in terms of dynamics and precision, the E-motor must fulfill the requirements relating to safety and the environmental impact. The safety concept for a de-energized (failure) condition requires that the torque-vectoring moment is immediately reduced and that the transmission behaves like an ordinary rear drive unit

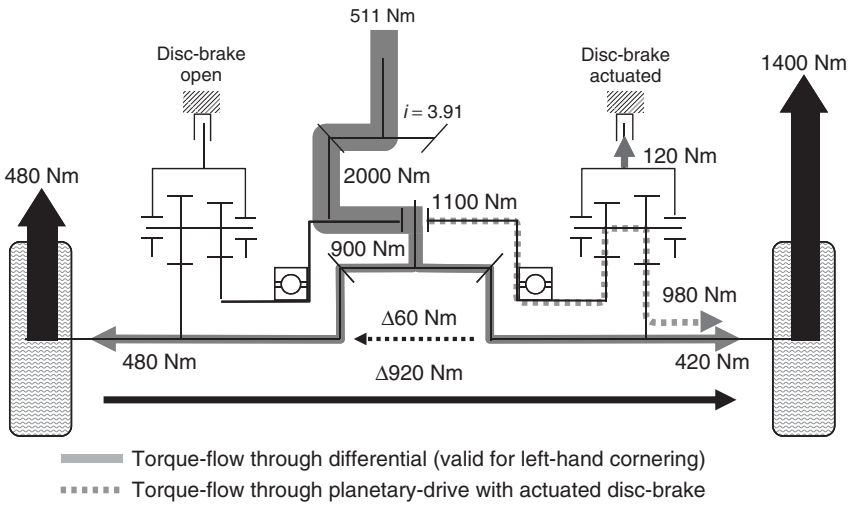


Figure 12. Torque flow and torque bias effect in the ZF Vector Drive © system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

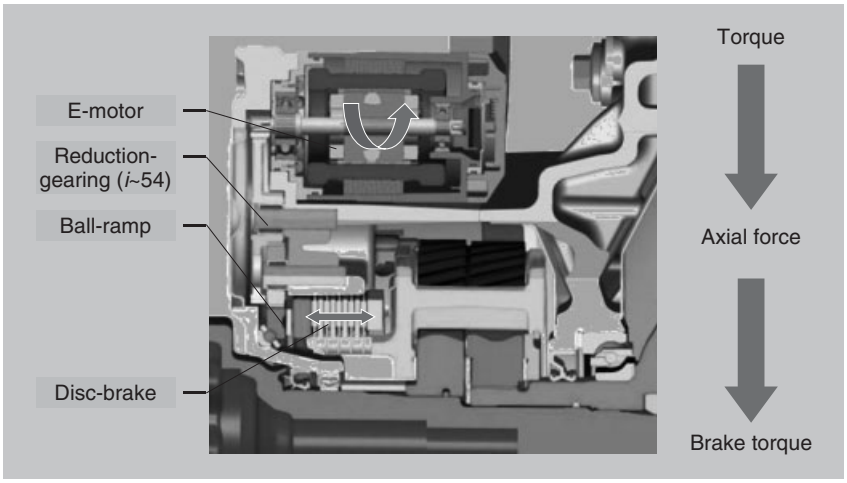


Figure 13. Actuation concept of the ZF Vector Drive © system.

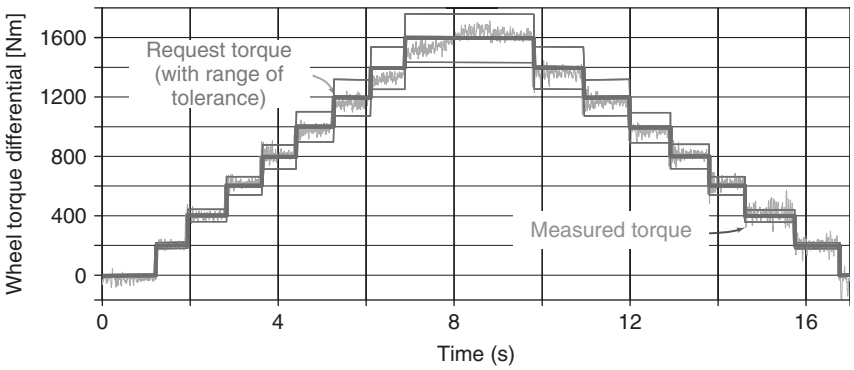


Figure 14. Control accuracy sequence of the ZF Vector Drive © system.

with open differential (fail-safe). As regards the E-motor, this means that the automatic opening of the multiplate disc must not be hindered by a locking moment. Use of the E-motor in the area of the differential implies extraordinary strains in terms of leak tightness, vibration, and temperature.

All these requirements are optimally fulfilled by asynchronous E-motors. These E-motors do not need a brush, nor do they need permanent magnets. They are thus free from any locking moment and ideally support de-energized opening in the event of a failure. Moreover, they are wear-free and provide a high level of ruggedness. Operation is possible with the E-motor open toward the drive unit, which permits free oil flow between motor and space of drive unit. The oil of the drive unit thus also lubricates the bearings of the rotor shaft, and separate venting of the E-motor is not necessary any more.

10.3 Electronics and safety concept

A dual-controller unit with integrated power electronics controls the drive. The vectoring torque calculated by the driving dynamics function is processed further in respect to present operating conditions. This means the controlled electric motors apply the multidisk brakes to generate an accurate vectoring torque.

Angle-of-rotation sensors sense the motor position and permit precise and prompt actuation of the multidisk brake. The software includes compensation functions for temperature, speed, and aging effects.

Owing to the intervention in the driving dynamics and stability of the vehicle, the safety concept was classified according to SIL 3, which corresponds with an error response equivalent to that of an active steering angle intervention. The safety concept implemented in the control

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

unit constantly monitors all units and ensures an immediate reaction to any sudden faults, so that the vehicle response is never compromised.

A dual-controller control unit implements the safety concept on the basis of mutual monitoring through redundancy functions. If the results from the two controllers differ, the vectoring torque will immediately be reduced to establish a safe condition of the system. This ensures that an error will not result in a safety-critical driving condition, which might endanger the driver and which he might not be able to handle. To meet the high system requirements, a precise tuning of the safety functions especially in respect to the units' mechanical characteristics is necessary.

10.4 Performance

Modern driving dynamics functions are based on vehicle models, and use sensor information such as steering angle, lateral acceleration, yaw rate, and vehicle speed to calculate the necessary vectoring torque highly accurately. In general, the overall control consists of two main control elements. The pre-control element sets the optimal and reproducible wheel differential torque for any driving situation. At the same time, the yaw rate control element corrects any deviations between the target and actual yaw rates, ensuring the required driving dynamics are maintained. This ensures a high degree of agility, controllability, and safety even at the stability limits of the vehicle. High setting accuracy and dynamics of the torque-vectoring system is essential to provide these benefits (Figures 14 and 15).

10.5 Efficiency factor and drag torques

As part of current efforts to reduce energy consumption and carbon dioxide emissions, efficiency, including reducing losses that occur within the axle drive, plays an increasingly important role. Every torque-vectoring system creates additional losses due to its components, gearing, and bearings. However, the levels of loss are different depending on the torque-vectoring design and the driving situation. The goal is to keep the extra consumption as low as possible in the driving cycle that is relevant to the driver.

Figure 16 shows the proportion of drag power caused by the two TV units of the ZF Vector Drive© system examined. The driving situation this is based on is straight-ahead driving at constant speed without actuation, such as occurs on freeways. Compared to the total losses from axle drives, the losses here are relatively low. The reason for this is that in this driving situation, the planetary gearset rotates as a block so that there is no gearing roll-off. Furthermore, a careful choice of oil and lining material means the fast-running multidisk clutches have a very low drag torque.

10.6 System networking with other driving dynamics control systems

Modern vehicles can feature a large number of other control systems apart from torque-vectoring systems, which make it possible to actively influence the driving dynamics. Active steering systems offer an additional superimposition steering angle that can be used to influence driving dynamics. Active rear axle steering systems can be used for active sideslip-angle control. Active roll stabilization impacts the self-steering response. Furthermore, braking interventions on selected wheels can also be used in addition to ESC to influence driving dynamics. What all these systems have in common is that they can electronically control the movement of the vehicle around the vertical

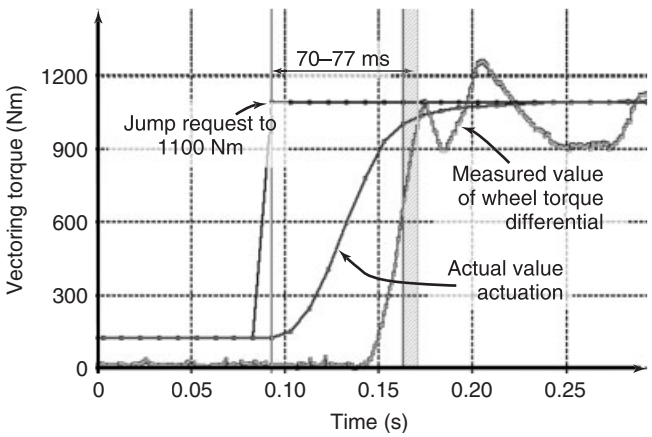


Figure 15. Dynamics characteristic of the ZF Vector Drive © system.

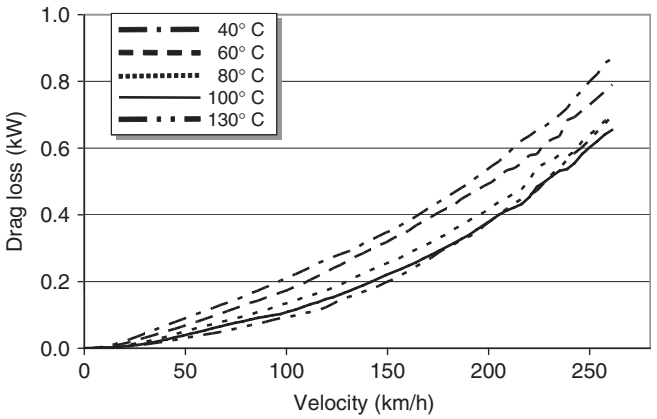


Figure 16. Drag power contribution of the two torque-vectoring units in the ZF Vector Drive ©. The figure shows the measured drag power at constant speed and straight-ahead driving without actuation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

axis. Super ordinate strategies link up the individual driving dynamics systems and help optimally coordinate interventions. This not only increases functionality by using interventions that supplement or support each other but also ensures maximum availability and increased driving safety.

11 DISCUSSION AND CONCLUSION

Torque vectoring for drivetrain systems represents a powerful vehicle technology to improve vehicle performance. The systems allow the intelligent distribution of propulsion torque to influence traction and vehicle dynamics. Different concepts, approaches, and designs have been developed and are nowadays available on the customer market. Several car manufactures use the technology of torque-vectoring to further improve the precision and agility of their modern and sporty vehicles.

The microcontroller-based control strategy of modern torque-vectoring systems allows a defined tuning of the vehicle behavior. The possibility of asymmetric torque distribution generates a yaw moment on the vertical axis and provides a strategy-based input for the intended vehicle behavior. Depending on the application and the intention of the manufacturer, different priorities in performance gain can be chosen during the development of the vehicle.

Beside the dynamic improvement, the optimization of the traction behavior still plays a crucial role, as this is one of the most important motivations for all-wheel-drive vehicles. Torque vectoring can also support the demand for increased traction and off-road performance by the lateral distribution of propulsion torque.

The technical effort to implement a torque-vectoring system at a modern vehicle unfortunately increases the weight and cost of the car. Additional gearsets, clutches, and actuators incorporate losses, which have negative effects on the efficiency and emissions. The losses are quite different, depending on the realized concept but nevertheless, they can restrain the technology from a wide spreading on the present car market.

RELATED ARTICLES

The potential for handling improvements by global chassis control
Global Chassis Control in Passenger Cars
Tyre Modelling
Clutch Wet
Basic Open Differentials
Passive and Active Limited Slip Differentials
Torque Transfer with AWD Systems
Clutch Actuation
Axle Systems

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto029
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Global Chassis Control in Passenger Cars

Hideo Inoue[1], Takashi Yonekawa[2], Masayuki Soga[1], Kenji Nishikawa[1], Eiichi Ono[3], and Makoto Yamakado[4]

[1]*Toyota Motor Corporation, Toyota, Japan*
[2]*Toyota Motor Corporation, Susono, Japan*
[3]*Toyota Central R&D Labs. Inc., Nagakute, Japan*
[4]*Hitachi, Ltd., Ibaraki, Japan*

1 Introduction 1
2 Hierarchy of Integrated Vehicle Systems 2
3 Evolution of Chassis Control Devices 4
4 Evolution of Control Using VDIM 6
5 Methods of Designing Integrated Controls 12
6 Future Trends of Integrated Systems 14
7 Conclusion 17
Related Articles 18
References 18

1 INTRODUCTION

Full-scale global chassis control systems are being developed by various companies for use in passenger cars. This development started in Europe in 1983 with the Daimler-Benz 4Matic four-wheel drive (4WD) system, which used online calculation of vehicle states. Calculation results from a bicycle model and yaw rate measurement results from the wheel speeds were compared to determine whether to engage the front axle to the drive train using electronically controlled clutches. This system used longitudinal slip and lateral vehicle dynamics for control for the first time (German patent DE 35 05 455 CS, 1985). Other early examples include the integrated active suspension, active rear steer (ARS) system, traction control system (TCS), and antilock brake system (ABS) in the Toyota Soarer in 1991 (Tanaka *et al*., 1992) and the introduction of an in-vehicle local area network (LAN) with the 4WDi-Four and ARS system in the Toyota Crown Majesta in 1992. In addition, in 1992, BMW offered dynamic stability control (DSC), the first electronic stability control (ESC) in its 850i sports car. This system controlled the throttle and the ignition timing after comparing the vehicle behavior to a bicycle model. Mercedes and Toyota as ESC released additional brake intervention controls in 1995. These were followed in 1997 by the release of the Toyota Hybrid System (THS) in the Prius, which featured the first example of technology that combined braking performance with the recovery of regenerative energy. Subsequently, various companies are now developing global control systems that integrate powertrain, steering, and active braking systems. These include the integrated chassis management (ICM) system developed by BMW, the vehicle dynamics management (VDM) system developed by Bosch, the vehicle dynamics integrated management (VDIM) system developed by Toyota, and so on. Such systems are beginning to be considered as fundamental technology for enhancing safety and environmental performance as well as driving enjoyment. In addition, advances in vehicle environment recognition technology have led to the development of driver assistance systems such as precrash safety (PCS) and the like. This chapter describes an outline of integrated vehicle systems, focusing on global chassis control. It also

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

presents an overview of the integration of driver assistance systems and global vehicle dynamics controls to enhance safety technology and discusses the value of the regenerative energy recovery and integrated vehicle dynamics control functions of hybrid electric vehicle (HEV) systems in enhancing environmental technology.

2 HIERARCHY OF INTEGRATED VEHICLE SYSTEMS

The process of driving a vehicle consists of three related factors: the driver, the vehicle, and the traffic environment. The component technologies that make up integrated vehicle control must be structured simply and rationally while considering the requirements to enhance the overall value of the vehicle in terms of safety, driving enjoyment, and environmental performance. Figure 1 shows the hierarchical structure of an integrated vehicle system. The system consists of the following five parts.

1. VDM
2. driver assistance management
3. energy management
4. human–machine interface (HMI) management
5. occupant protection management.

Currently, the most important of these parts are VDM and driver assistance management. It is likely that further advances will occur in the field of sensing, such as cooperative communication among vehicle environment recognition systems using autonomous sensors such as radar and camera, navigation systems, and roadside infrastructure. Such advances are particularly likely for integrated safety systems and VDM is a key technology supporting this progress.

2.1 Integrated safety

This section briefly discusses the concept of integrated safety. Figure 2 shows the trends in active and passive safety technology. Advances in the field of passive safety technology include optimized body structures, enhanced restraint systems such as airbags, and measures for various different types of accident formats (i.e., compatibility in crashes with different vehicle types, pedestrian safety, and the like). However, active safety has grown in importance in recent years. The aim of active safety systems such as ESC, VDIM, and PCS is to help reduce the number of accidents that occur. Systems that coordinate with roadside infrastructure using communication technology may also be developed in the future. On the basis of these trends, it is likely that active and passive safety systems will be introduced together as integrated packages rather than as separate technologies. The aim of integrated safety management is to create a simpler overall system that helps to seamlessly operate each function together (Figure 3). For

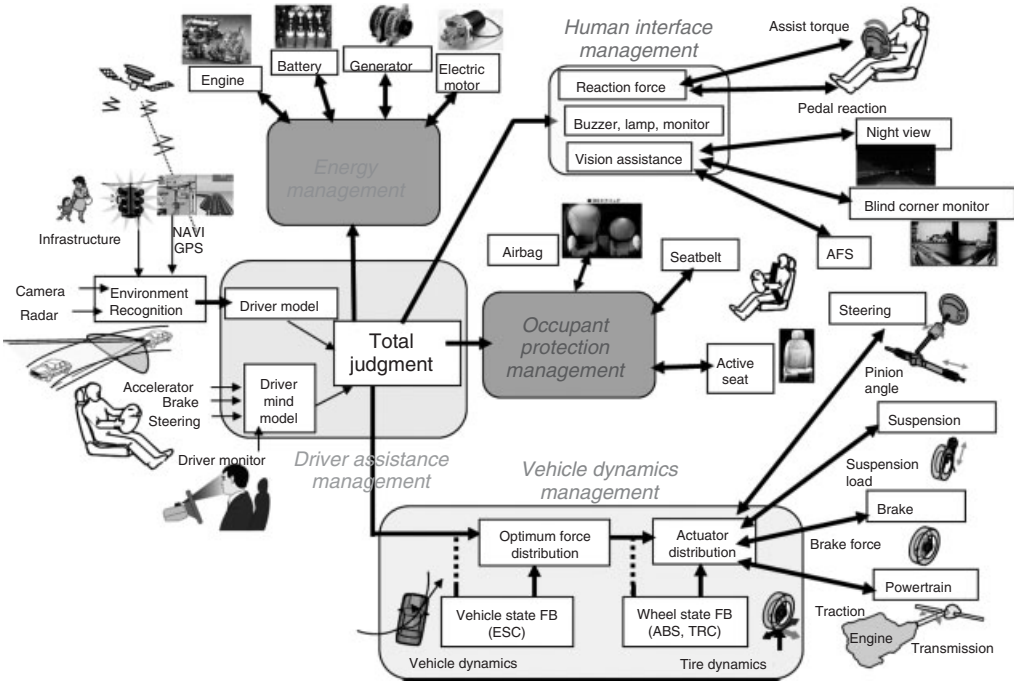


Figure 1. Hierarchical structure of integrated vehicle control.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

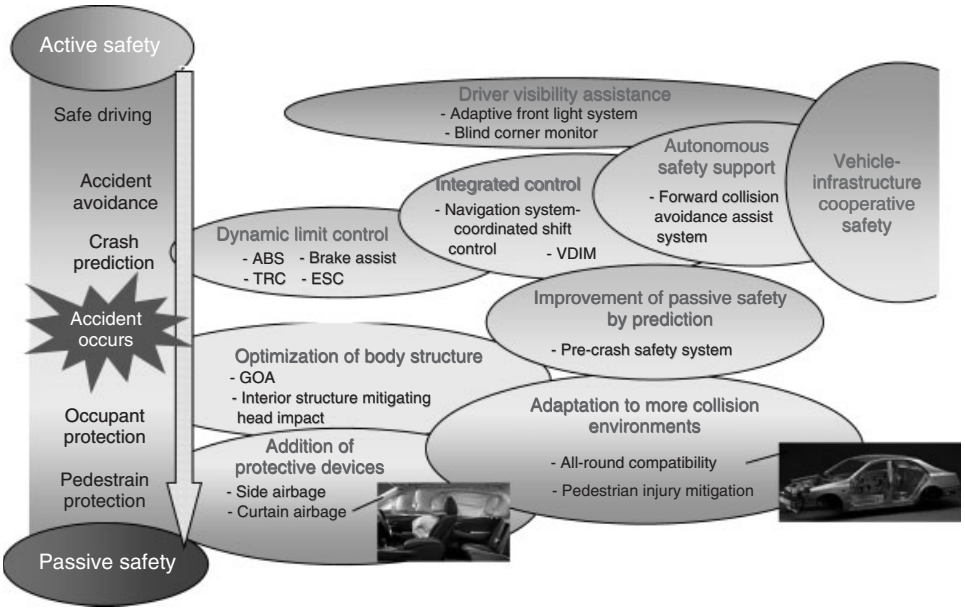


Figure 2. Trends in passive and active safety technology.

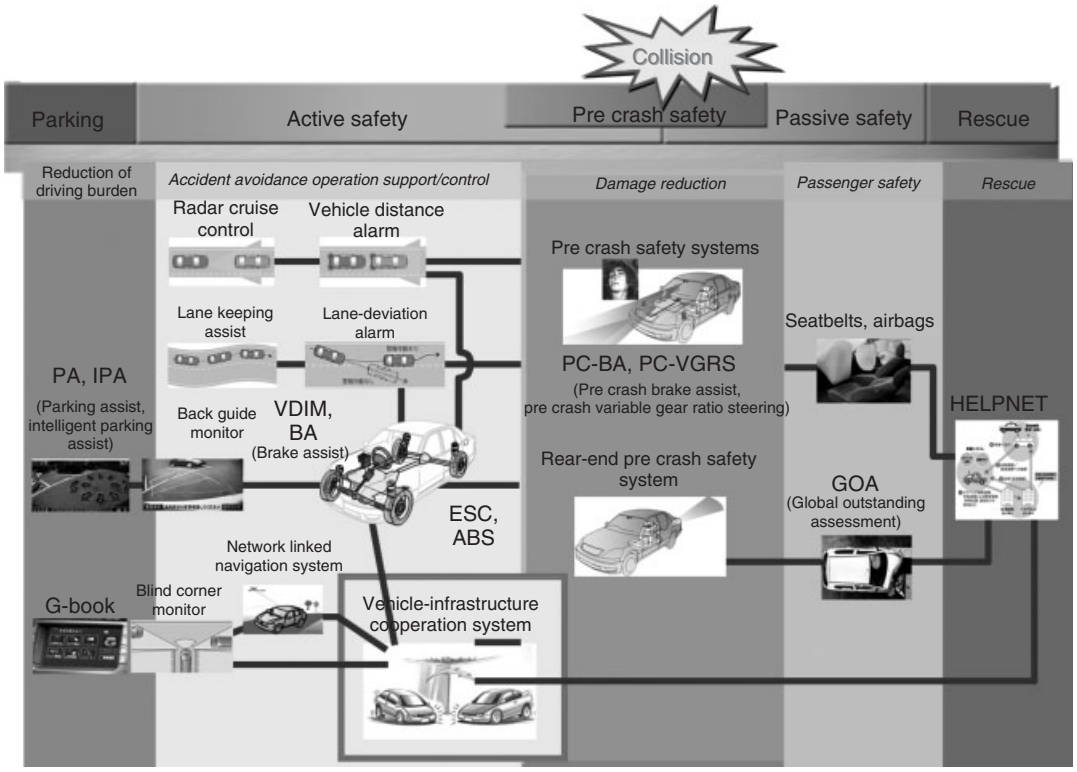


Figure 3. Integrated safety management.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

this reason, VDM will play a fundamental role acting as the muscles and nerve system of the vehicle.

3 EVOLUTION OF CHASSIS CONTROL DEVICES

The development of more advanced chassis control devices that act as the muscles of the car to control the tires is an essential part of enhancing vehicle dynamic performance. This requires expanding the degree of freedom of control over the forces generated at each tire contact point from the longitudinal direction to the lateral and vertical directions. This can only be achieved by smooth and highly responsive control in all directions. The following sections describe recent development trends in chassis control devices and examples of their application.

3.1 Brake control

The development of brake control systems has accelerated because of the adoption of ABS and electronic brake distribution (EBD) systems as standard equipment. In combination with advances in ESC technology, the major effect of integrated brake control systems in helping to reduce accidents has been confirmed in the real world. As a result, the United States and other countries have begun to mandate their usage and such systems are likely to become standard equipment in the future.

The popularization of HEVs in recent years has also contributed to the development of electronically controlled brake (ECB) systems (Figure 4) with the aims of achieving linear hydraulic brake control and improving response (Nakamura, 2002). This is a brake-by-wire system in which the hydraulic brake pressure at each wheel is isolated from the brake pedal and braking control is performed by a high pressure source through linear solenoids. This system facilitates braking force coordination during regenerative braking in an HEV. As its smooth and highly responsive controllability can also be utilized as part of ESC systems, ECB systems are also spreading to other vehicle types in addition to HEVs.

3.2 Drive system control

This field has seen progress in active controls such as systems that transfer torque to the front and rear wheels in a 4WD vehicle, as well as limited slip mechanisms in center, front, and rear differentials. In recent years, systems for distributing driving force to the left and right wheels using a speed-increasing mechanisms installed inside the differential have been developed. The characteristics of these systems are being actively used to improve cornering performance.

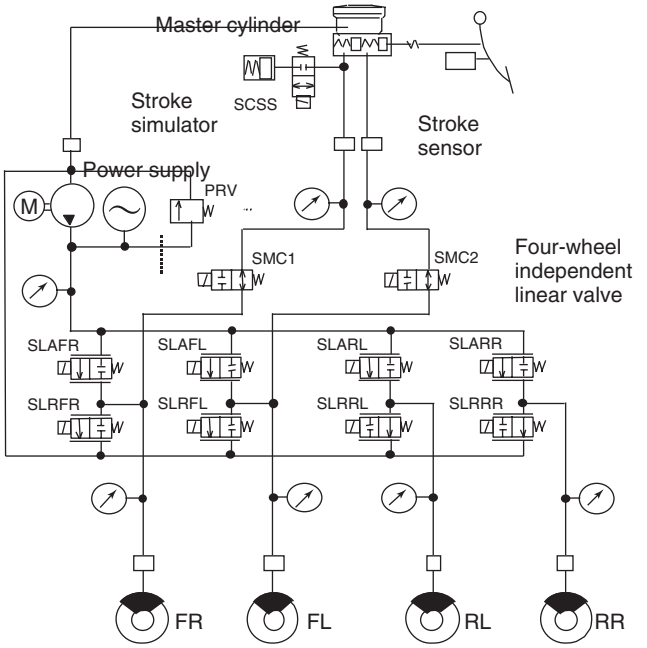


Figure 4. ECB structure.

In contrast, the electrification of driving forces has started in HEVs and electronically controlled 4WD vehicles to enable smother and more responsive driving force control. The practical application of motors capable of driving each wheel independently in the future may also help to further enhance dynamic performance.

3.3 Steering control

Electric power steering (EPS) systems are being rapidly adopted to improve fuel efficiency, as well as for use in HEVs and the like. In addition to simply replacing hydraulic power steering, the role of EPS is spreading to new functions that make active use of its capability to freely vary the assistance force.

For example, it has already been introduced in some vehicles to perform the following functions in coordination with ESC. EPS can assist the driver's steering effort to facilitate countersteering when the rear wheels slip or to prevent understeering when the front wheels slip. It can also be used as an actuator in functions that help the driver keep the vehicle in its lane by assisting the driver's steering effort. Alternatively, EPS can also function as an actuator in automatic driving systems such as Intelligent Parking Assist.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

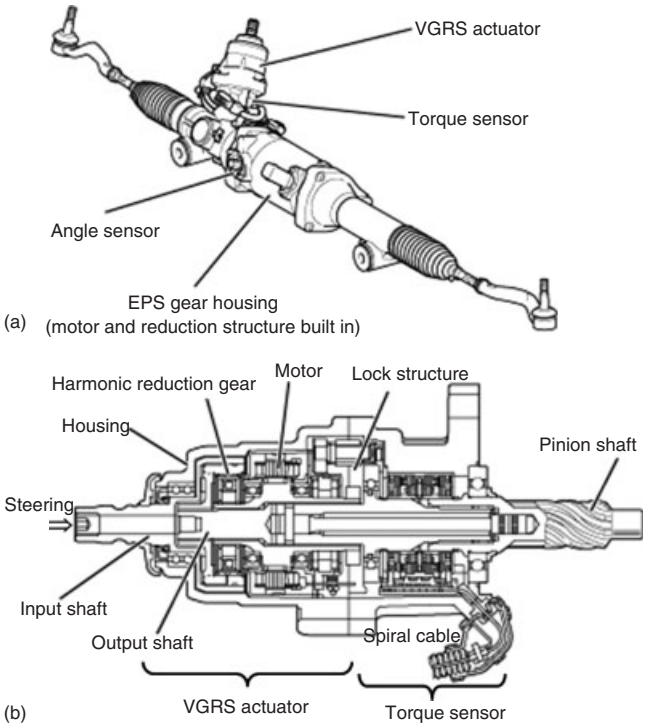


Figure 5. (a,b) VGRS actuator.

In addition, variable gear ratio steering (VGRS) systems have been developed that vary the steering response characteristics of the vehicle yaw angular velocity (Figure 5). Conventionally, the steering characteristics of the yaw angular velocity were determined by the specifications of the suspension. VGRS is capable of varying these characteristics to a constant optimum level in accordance with the driving environment. Furthermore, ESC devices are starting to be used, which utilize the capability of VGRS to control the turn angle of the front wheels independently of the steering wheel angle.

ARS has also been adopted on some vehicles, and its enhancement of dynamic performance has been verified. In recent years, vehicles have been released that combine ARS with front-wheel active steering systems to further improve dynamic performance (Katayama, 2007; Kojo, 2002; Ono, 2007).

3.4 Suspension control

The suspension of a vehicle consists of springs, shock absorbers, and link mechanisms. However, there is a long history of control technology applied to suspensions to improve both ride comfort and vehicle stability. Examples include the semiactive suspensions introduced in the 1980s for controlling shock absorber damping force. Improvements have continued since then. Recent years have seen the development of electronically controlled active stabilizer suspension systems that actively reduce vehicle roll (Figure 6). Progress is also being made toward the development of electronically controlled fully active suspensions. As a result, the use of actuators in suspension control is spreading, and these are expected to help improve vehicle performance through the active variation of vehicle attitude and vertical load.

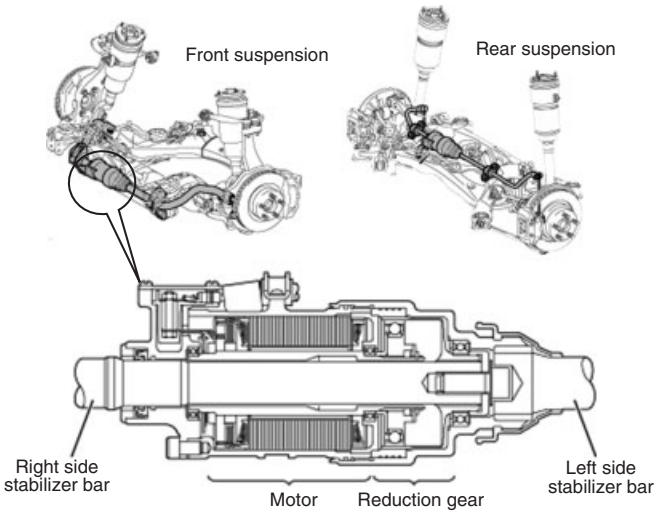


Figure 6. Active stabilizer actuator.

3.5 Evolution of vehicle environment recognition technology

One current trend is vehicle environment recognition technology that supports the cognitive processes of the driver. The development of environment recognition sensors such as radar and cameras is advancing rapidly.

Forward recognition technology using radar is already used by cruise control systems. PCS systems have been developed that help the driver to avoid accidents through coordinated control with the brakes (Tsuchida, 2007). A more accurate system that combines information from radar with camera images (Figure 7) and a system that coordinates information with a camera that detects the orientation of the driver's face have also been developed.

Lane Keeping Assist, Intelligent Parking Assist, and other driving assistance systems that use image-recognition technology have already been commercialized. However, the number of systems in this field is likely to grow further as researchers study ways of utilizing information provided by roadside infrastructure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

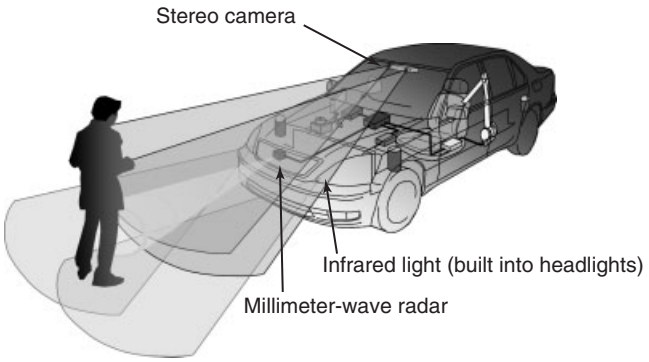


Figure 7. Forward recognition system using radar and stereo camera.

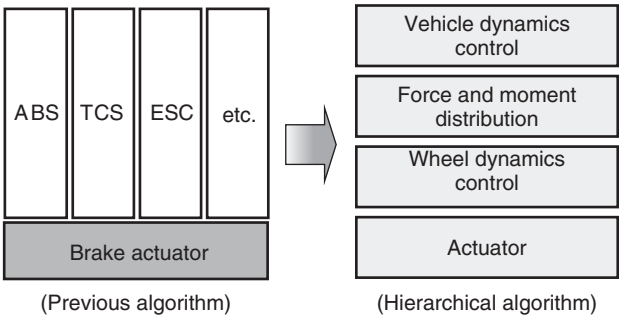


Figure 9. Hierarchical control algorithm.

4 EVOLUTION OF CONTROL USING VDIM

4.1 Concept of VDIM

Since the 1980s, various attempts have been made to enhance vehicle dynamic performance using active chassis control. In 1986, the Daimler-Benz 4Matic system was the first to use lateral dynamics for control purposes. Direct yaw moment control systems with active braking such as ESC enable good performance in the critical limit region (Shibahata *et al*., 1993; Koibuchi *et al*., 1996; Van Zanten, 1996). The aim of the next generation of vehicle dynamic control systems is to provide seamless vehicle maneuverability and stability at all times through the integrated control of driving forces to all four wheels. Figure 8 shows the concept of Toyota's VDIM system. This illustration is called the *ball in a bowl* concept (Hattori *et al*., 2002). The ball corresponds to the state of the vehicle, which is maintained within a bowl constructed by the control. The inside of the bowl is the stable region, and the outside is the unstable region. In conventional systems, the walls of the bowl are constructed from independent functions such as ABS, ESC, and TCS, which form sheer boundaries before an emergency occurs. As a result, although these functions are capable of stabilizing vehicle motion, the motion may be discontinuous in some cases. In contrast, VDIM realizes smoother behavior because the conventional control systems are restructured to form a continuous and smooth wall.

As vehicle control systems are becoming more diversified, the algorithm is required to perform cooperative control of many systems, such as the drive train, braking, and steering, easily. Accordingly, the compatibility of the algorithm with various system configurations is important. The hierarchical control system structure shown in Figure 9 has been adopted for VDIM (Hattori, 2002; Fukatani, 2005).

The first layer (vehicle dynamics control) calculates the target forces and moments of the vehicle to achieve the desirable vehicle motion corresponding to the driver's pedal input and steering wheel angle. There are several examples of research for the first layer. In the critical limit region, the determined target resultant force and moment also satisfy robust stability conditions to avoid vehicle spin (Ono *et al*., 1998). However, in the moderate region, Yamakado *et al*. (2010) have proposed a target longitudinal acceleration/deceleration model determined by predicted lateral jerk to improve driving enjoyment. The target resultant force and moment of the vehicle motion are distributed to the target tire forces of each wheel based on the friction circle of each wheel in the second layer (force and moment

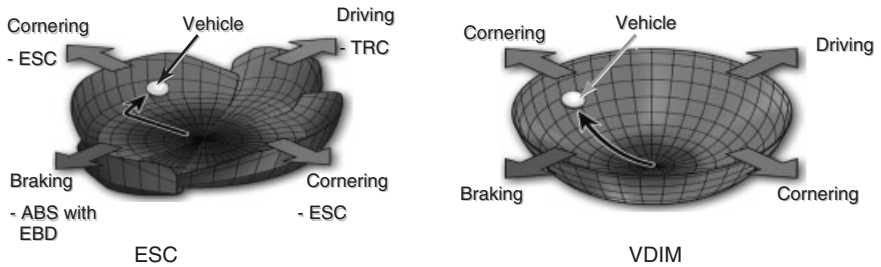


Figure 8. Evolution of control using VDIM.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

distribution). The third layer (wheel dynamics control) controls each wheel motion to achieve the target tire force. There are redundant degrees of freedom in the second layer. The vehicle dynamics performance in the critical limit region depends on the force and moment distribution algorithm, which uses these redundant degrees of freedom.

The motion of a vehicle in the three degrees of freedom (longitudinal, lateral, and yaw) is controlled by the steering and traction or braking forces from the four tires. If each tire can be individually steered and operated for traction or braking, the control system has redundancy. Vehicles move using the friction between the tires and the ground. The frictional forces at the tires have limits dependent on the conditions of the road surface. These limits are called the *friction circle*, and a tire cannot exert force on the road surface in excess of the friction circle. To extend the limits of the performance of vehicle dynamics, it is necessary to ensure that the forces exerted by all tires work efficiently in cooperation with each other. The problem of integrated control of vehicle motion then becomes how to best use the redundant degrees of freedom. As the friction circle has nonlinear constraints because of the limitations on the frictional forces at each wheel, the distribution of the longitudinal and lateral forces and the yaw moment (vehicle forces and moments) to each tire force becomes a nonlinear problem.

Ono *et al*. (2009) proposed a vehicle dynamics integrated control algorithm using an online nonlinear optimization method for four-wheel-distributed steering and four-wheel-distributed traction and braking systems. The proposed distribution algorithm calculates the magnitude and direction of the tire forces to satisfy constraints corresponding to the target resultant forces and moments of vehicle motion and also to minimize the maximum μ rate (=tire force/friction circle) of each tire. This research demonstrated the convexity of this problem and guaranteed the global optimality of the convergent solution of the recursive algorithm. This implies that the theoretical limitation performance of vehicle dynamics integrated control can be reached.

Comparing it with general quadratic programming can show the efficiency of the proposed algorithm, which calculates the theoretical limitations of vehicle forces and moments. The following minimization problem of the sum of squares of the μ rate may be considered as a benchmark. This is an extension of the problem described by Mokhimar and Abe (2003). In this simulation, the generated vehicle longitudinal forces are compared for straight-line braking on a split road with different coefficients of friction μ (μ =1.0, 0.2).

Figure 10 shows the tire forces of a vehicle controlled by the proposed method and a vehicle controlled by quadratic programming. Both of the controls achieve the reference braking force within a moderate area when the reference braking force is 7000 N. However, unlike the vehicle controlled by quadratic programming, the vehicle controlled by the proposed method can also achieve the reference braking force in the critical region when the reference braking force is 10,000 N.

4.2 Configuration of VDIM

Figure 11 shows the overall configuration of the VDIM system. Using in-vehicle sensors to detect the yaw rate and steering angle, VDIM collects together various items of data for optimally controlling the brakes and front steering to stabilize the vehicle attitude. It achieves a steer-by-wire function using an active front steering (AFS) system with two actuators (VGRS and EPS) to control the steering angle of the front wheels and the reaction torque of steering. It also acts as a wide-ranging safety control function in coordination with the ECB system.

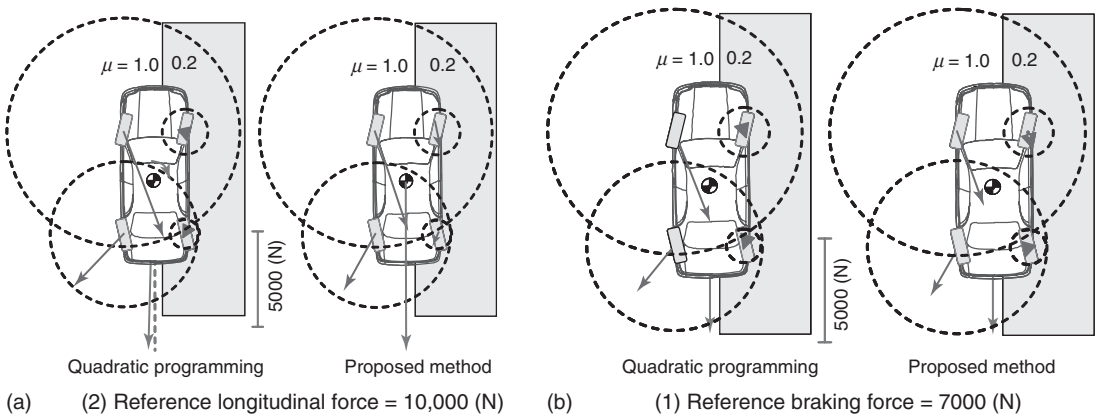


Figure 10. (a,b) Straight-line braking on split μ road.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

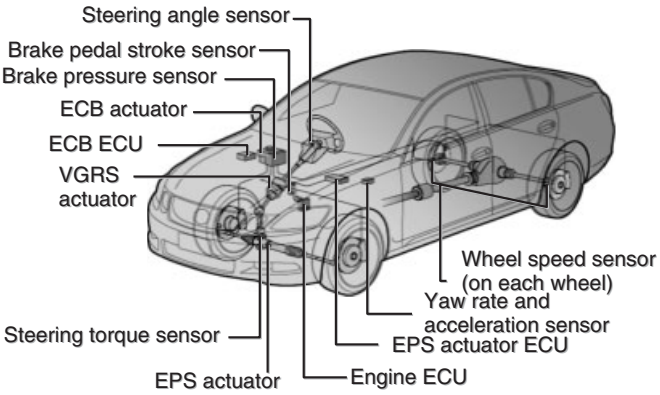


Figure 11. Configuration of VDIM system.

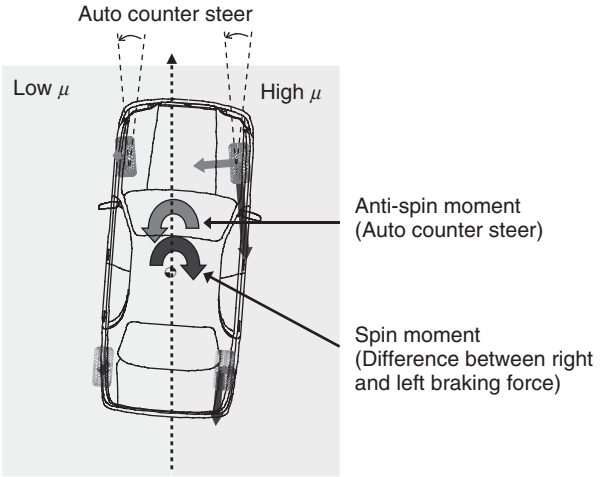


Figure 13. Corrective steering control during braking on split μ road.

Figure 12 shows the layout of functions for controlling vehicle dynamics and behavior using VDIM. Creating this layout diagram clarifies the role of each function individually and in combination with other functions. VDIM enables true integration of vehicle control by emphasizing the development of each function and its actions.

4.3 Performance of VDIM

Figure 13 shows the outline of control when driving on a road with different coefficients of friction (μ) under the left and right wheels. As shown in the figure, when the driver brakes, spin moment is generated in accordance with the difference between the left and right braking forces, resulting in deflection of the vehicle. On this type of road surface, a vehicle dynamics control system that uses just longitudinal forces cannot achieve a high degree of stability and braking simultaneously with driving performance. In contrast, this can be achieved by an AFS system that is capable of controlling the vehicle in the lateral direction.

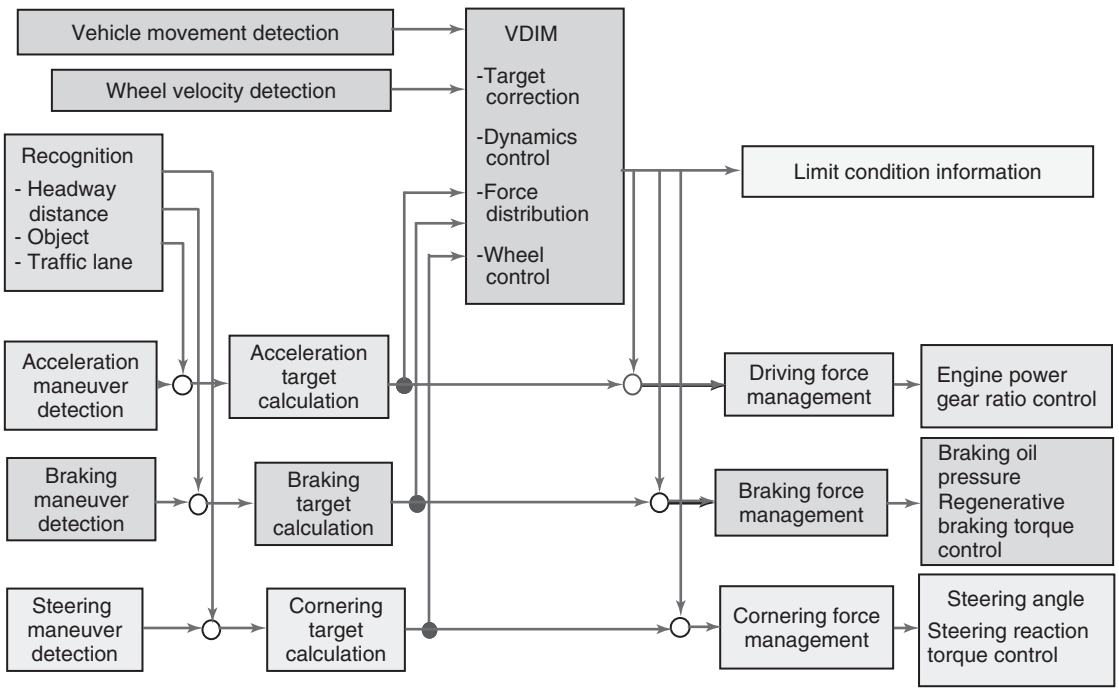


Figure 12. Layout of VDIM software functions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Figure 14 shows the test results of straight-line braking on a split μ road. It shows the normalized steering angle and peak yaw rate after braking (state without control = 1). With the active steering control, the peak yaw rate was reduced by approximately 50% compared to the rate without the control. In addition, the AFS control requires less corrective steering effort from the driver and generates a lower peak yaw rate on the vehicle than conventional ABS.

AFS has a large effect on side slipping of the front wheels. In particular, for a vehicle braking on a road with different coefficients of friction under the left and right wheels, AFS is capable of maintaining both vehicle stability and driving force. It also helps to suppress vehicle spin behavior when the driver performs an evasive maneuver (Figure 15). AFS is more effective at suppressing spin as it can generate moment by controlling the slip angle of the front wheels in addition to the conventional control that generates moment to reduce spin using braking force.

Figure 16 shows the results of a test comparing the effects of VDIM and ESC. In the test, the vehicle was steered mechanically through a slalom using constant steering operations while accelerating on an artificial low friction surface ($\mu = \sim 0.3$) simulating a snowy road. At the limit regions, the slip angle with VDIM was half that of ESC (+TCD), which shows that VDIM is capable of greatly improving the vehicle stability.

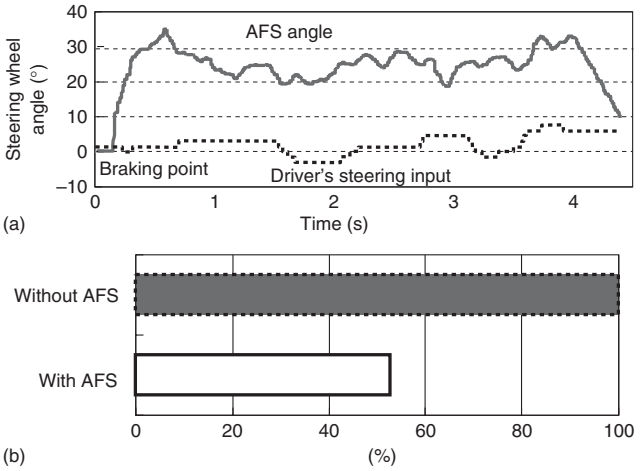


Figure 14. Effect of AFS on braking on split μ road. (a) Time domain data example and (b) normalized peak yaw rate.

4.4 Further evolution in control using VDIM

4.4.1 Enhancement of collision-avoidance performance using environment recognition technology

Toyota developed a collision-avoidance support system in 2006 that aimed to help reduce accidents by assisting evasive maneuvers by the driver. This system uses a forward monitoring function consisting of a millimeter wave radar and stereo camera to judge the risk of a collision with an object and assists the driver to avoid the collision by varying the steering gear ratio and braking. This system consists of a block that detects objects and judges the collision risk, a block that determines the evasive maneuver by the driver, and a block that controls vehicle behavior (Figure 17). The system operates as follows (Figure 18).

1. If a high collision risk is judged, the system warns the driver to take evasive action.
2. The system reduces the VGRS steering gear ratio to assist the driver's evasive steering maneuver.
3. If a high collision risk is judged, the system begins automatic braking. When the driver performs an evasive steering maneuver, deceleration during the maneuver is achieved by reducing braking force with a gradual gradient. In this event, VDIM controls the steering and brakes appropriately in accordance with the vehicle state.

Figure 19 shows the results of a double lane change test with and without system operation. The system reduces the driver's steering wheel angle and steering velocity to enhance the object evasion capability of the vehicle.

4.4.2 G-vectoring

VDIM is basically a feedback control that aims to restore the ball to the bottom of the bowl in the ball in a bowl concept. In contrast, G-vectoring control (GVC) is a feed-forward control mechanism that enables the ball to start rolling toward the edge of the bowl in coordination with driver's maneuvers. In 2008, Yamakado and Abe identified an original trade-off strategy between longitudinal traction and cornering force using jerk information to observe an expert driver's voluntary braking and turning actions (Yamakado and Abe, 2008). This strategy was used to develop GVC, which is basically a mechanism for achieving automatic longitudinal acceleration control in accordance with vehicle lateral jerk caused by the driver's steering maneuvers (Yamakado *et al*., 2010). With GVC, the direction of the resultant acceleration (G) changes seamlessly (i.e., by vectoring) in the same way as it does when an expert driver is behind the steering wheel. In this way, ideal vehicle motion can be achieved. The following

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

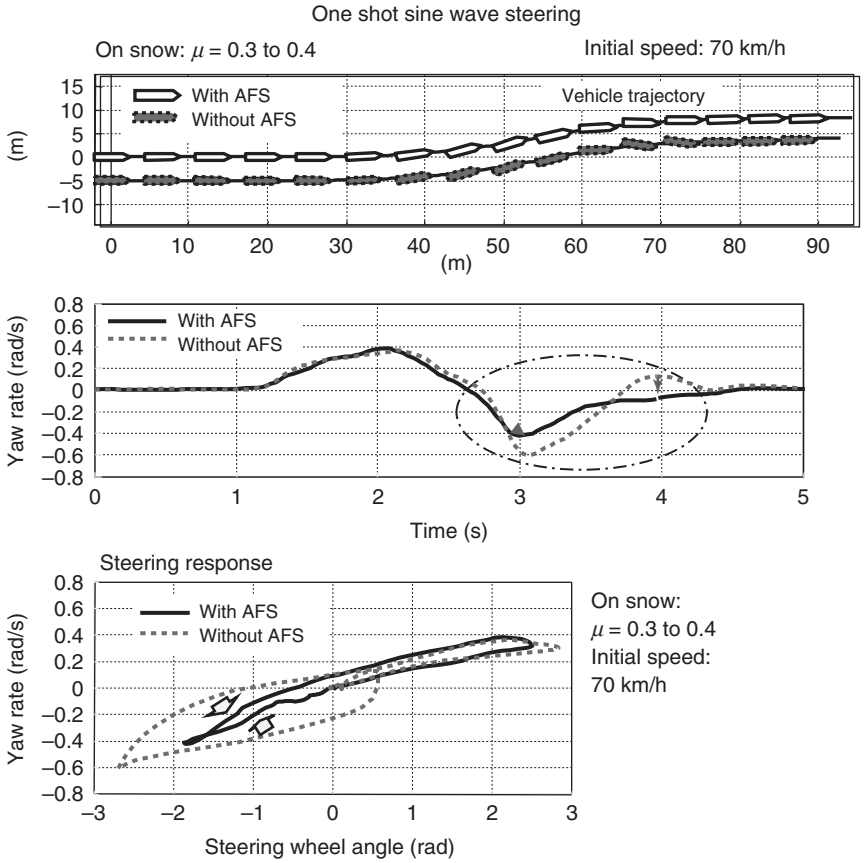


Figure 15. Comparison of vehicle behavior when changing lanes.

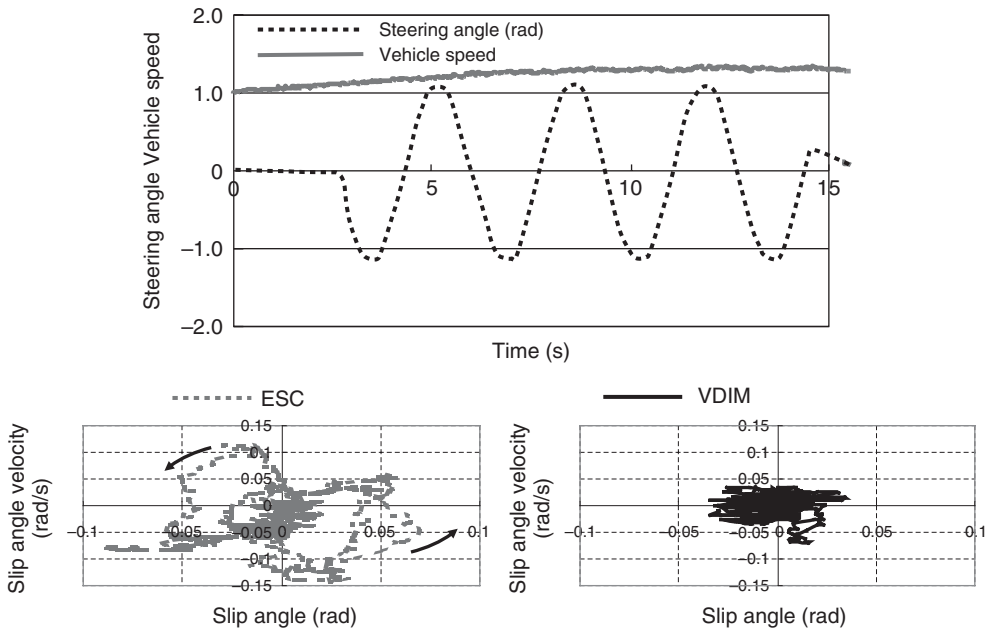


Figure 16. Vehicle stability with ESC and VDIM.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5
www.it-ebooks.info

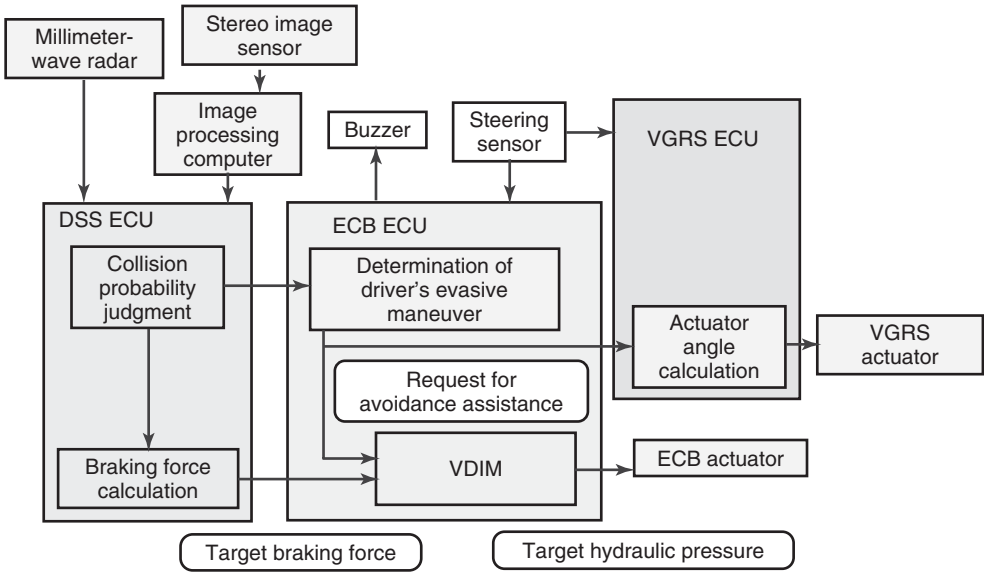


Figure 17. Functional configuration of object avoidance assist system.

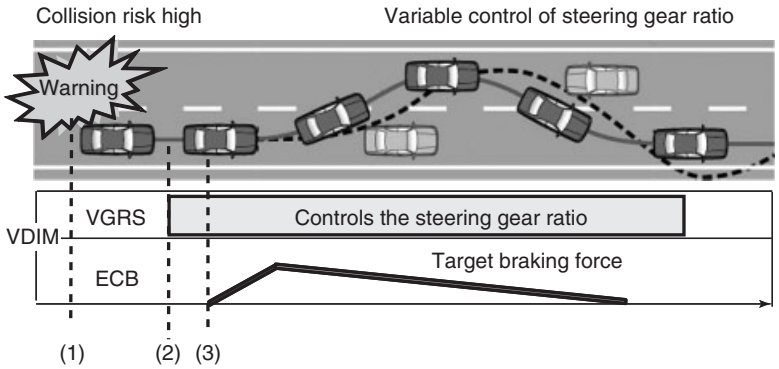


Figure 18. Control procedure of object avoidance assist.

equation was proposed as the fundamental equation for GVC.

$$G_{xt} = -sgn(G_y \cdot \dot{G}_y)\frac{C_{xy}}{1+Ts}|\dot{G}_y| \quad (1)$$

where G_{xt} is the longitudinal acceleration command, C_{xy} the gain, and G_y the lateral jerk.

Figure 20 illustrates the GVC concept. When the vehicle starts turning a corner, it starts braking simultaneously as lateral jerk increases (vehicle positions 1–3). After that, the braking stops during steady-state cornering (vehicle positions 4 and 5) because the lateral jerk becomes zero. The vehicle begins to accelerate when it begins to return to straight-ahead driving (vehicle positions 6 and 7). If a bowl were fixed to the vehicle, a ball in the bowl would move smoothly along the level curve as shown at the top of the figure (ball positions 3–5) because of the change in inertial force caused by the acceleration of the vehicle. Considering the shift in wheel vertical load between the front and rear wheels (which is caused by the acceleration and deceleration of the vehicle), the handling of the vehicle when entering a corner and its stability when leaving a corner will be improved.

Figures 21–23 show the measurement results obtained for cases with and without GVC. These results show that applying GVC makes it quite possible to emulate expert driving.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

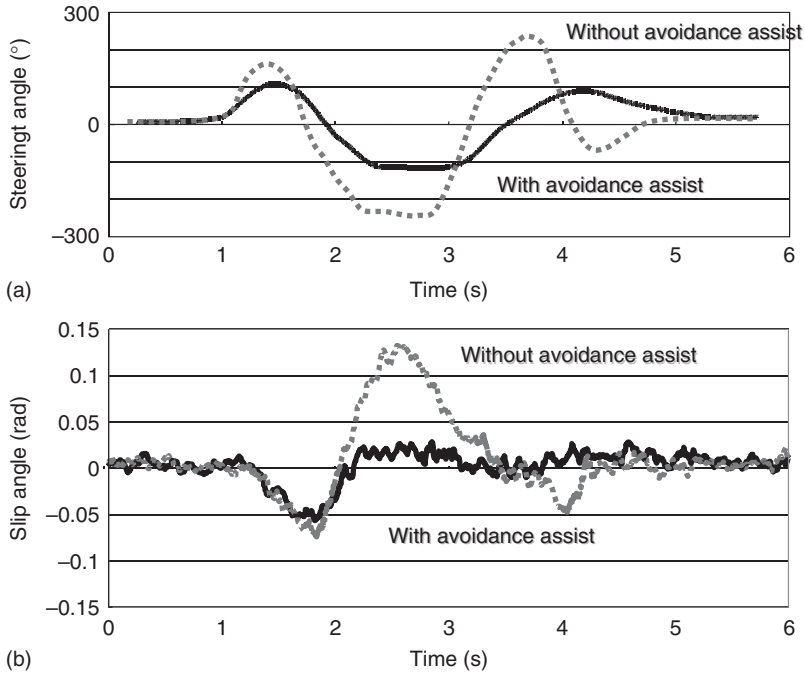


Figure 19. (a,b) Steering angle comparison when avoiding object.

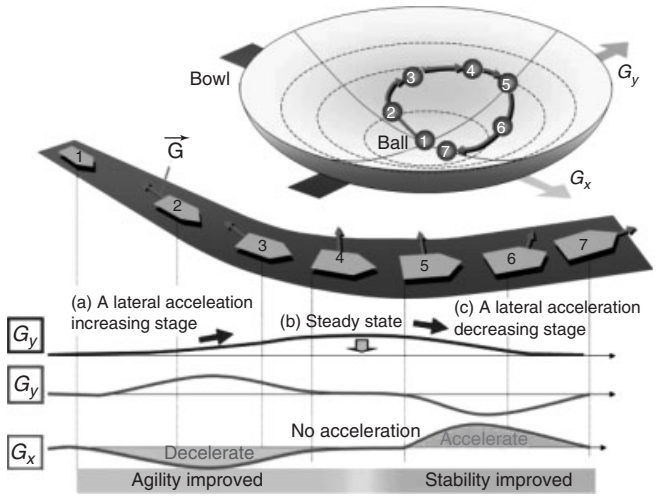


Figure 20. G-vectoring control concept.

5 METHODS OF DESIGNING INTEGRATED CONTROLS

The previous sections have described how integrated controls have a large potential to enhance vehicle dynamic performance. However, the integration of controls increases the complexity of the systems, substantially increases the scale of development and the work hours required, and makes it more difficult to secure reliability. Consequently, the original goal of improving performance also becomes more difficult to achieve. For this reason, the key concepts for designing integrated control are to create a hierarchy and to mask and abstract information.

5.1 Creating a hierarchy of function and application levels

When constructing a control, it is important to consider the failsafe conditions and other operations in the event of an abnormality, in addition to approaching the control from the standpoint of the normal targeted performance and operation. This is particularly important for integrated controls. The control has to be constructed on a layered hierarchical basis, categorizing the functions to be integrated and the functions that operate independently and autonomously. In creating the hierarchy, it is convenient to envision the relationship between the actions of a person's hands and feet and the reflexes of the brain and spinal column. Actions that require fast reactions and actions with a fixed pattern are achieved by spinal column reflex and the situation is reported to the brain. The brain directs overall actions by sending commands to the hands and feet. VDIM was created based on this concept (Figure 24). The roles are determined in sequence from the control devices equivalent to the lower order muscles, and the configuration is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

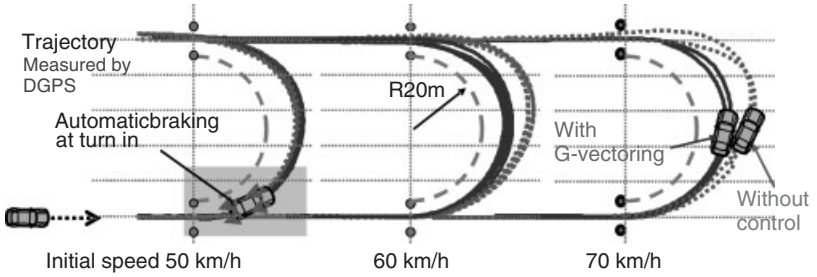


Figure 21. Trajectory evaluation.

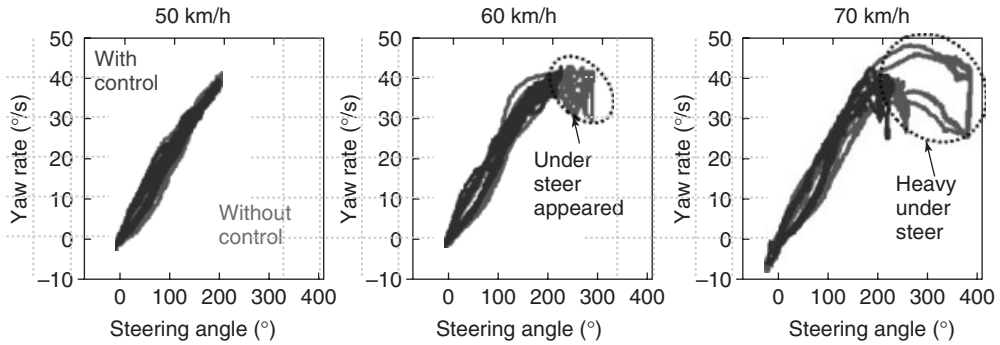


Figure 22. Steering angle versus yaw rate.

designed as much as possible to allow independent action and the selection of functional operation.

5.2 Masking and abstracting information

When constructing various functions in a hierarchical structure, another key point is the masking and abstracting of information when collecting information and transmitting commands. From this standpoint, it is simple to envision the relationship between a team manager and the team members in a corporate organization. Normally, the team members perform work based on instructions from the manager. However, the manager does not have a detailed grasp of each specific aspect of the work of the team members and the manager does not give specific detailed instructions about that work. Therefore, if the manager is ill or in another abnormal situation, the team members can operate autonomously to a certain level. In addition, the manager above the team manager is capable of running the organization without knowing the last detail of the work of the team members. Therefore, this manager can produce results that would not be possible individually. A simple and highly reliable system can be constructed by making use of this type of information collection and command transmission system.

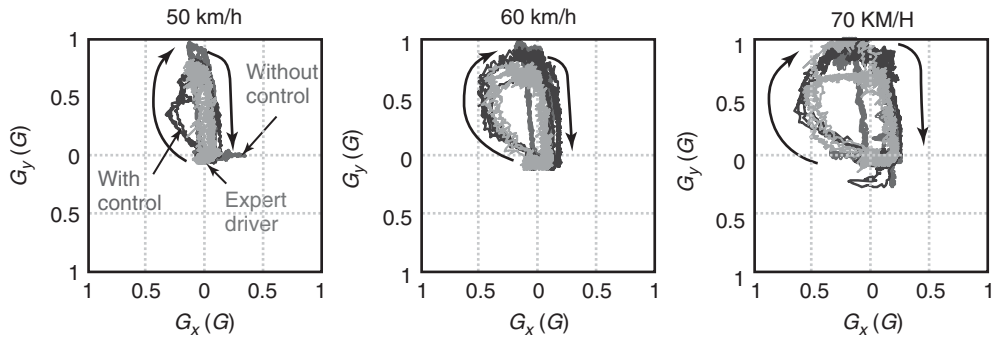


Figure 23. "g–g" diagram.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

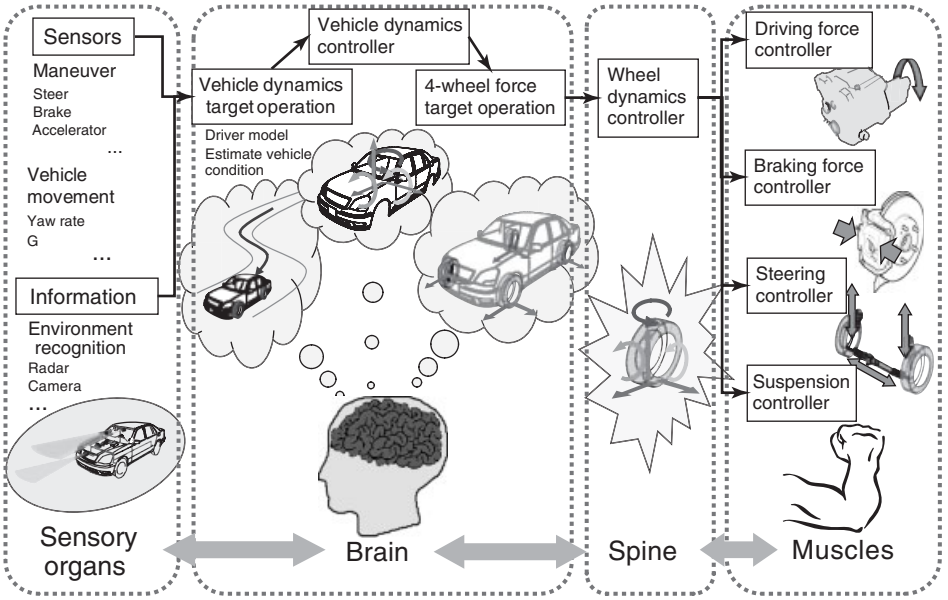


Figure 24. Hierarchy of VDIM and outline of components.

If the upper level system relies on detailed information from the lower level system, it becomes difficult to change the system configuration or add new functions. For example, in the case of brake control, the upper level system can simply control the total braking force even when in combination with a motor or another device. This is carried out by masking internal operations based on an abstraction and normalization concept. This concept positions braking force and braking *G* above hydraulic pressure and hydraulic pressure above the solenoid current of the actuator. Furthermore, if a theoretically different brake actuator is added into the system, there is only very little impact on the upper level system.

5.3 Creating packaging level hierarchies

The adoption of software platforms and operating systems (OS) is advancing to absorb differences in inputs and outputs, as well as in the communication between hardware and software, and to create freedom in application package locations for integration into ECUs. Conventional software structures are constructed differently based on the approach of each automaker and software supplier. However, standardization efforts are under way to achieve integration. Future development will also have to consider these trends.

5.4 AUTOSAR activities

AUTOSAR (automotive open system architecture) is an enabling technology for integrating systems in a vehicle. As mentioned in Section 5.3, AUTOSAR defines the basic software architecture, which consists of a hardware abstraction layer (HAL), system/communication services, and a runtime environment (RTE). RTE is an embodiment of a virtual functional bus (VFB) that enables the integration of application software and the physical allocation of applications into each ECU (Figure 25).

The system/communication services provide standard functions such as communication and OS. HAL absorbs the differences between microcontrollers, sensors, and the like. Figure 26 shows the basic AUTOSAR architecture.

AUTOSAR is being advanced by a development partnership, in which many global automotive companies are participating in AUTOSAR activities to develop AUTOSAR specifications as a worldwide de-facto standard (Figure 27).

6 FUTURE TRENDS OF INTEGRATED SYSTEMS

6.1 Expansion of active safety and driver assistance

The previous sections have highlighted the significant contribution of integrated control technology in active safety systems. The systems are likely to grow even more important in the future. Normally, the driving process consists of cognition, judgment, and action phases. An error in even one of these driving processes may result in an

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

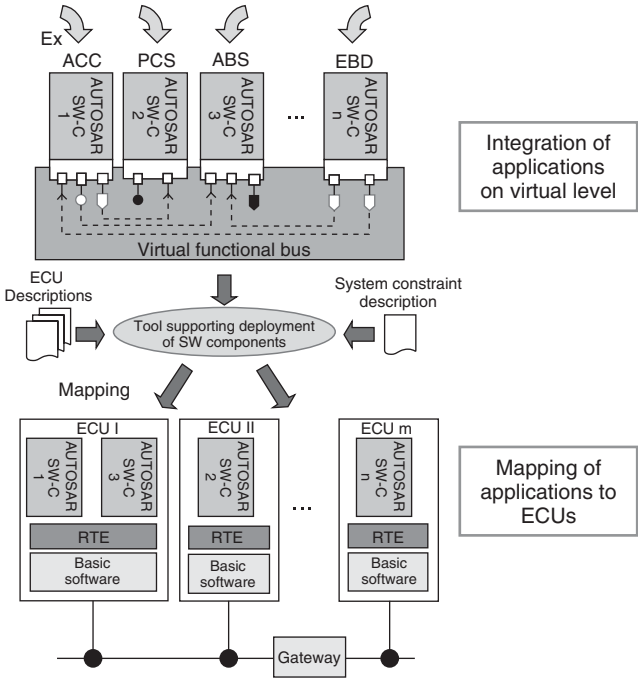


Figure 25. Integration of applications.

accident. Consequently, active safety technology is being developed to support each phase (Figure 28).

Figure 29 shows the technological areas of the various control systems that have already been commercialized. The direction of brake technology has already changed from brake assist (BA) type systems to PCS and other automatic braking systems. However, there are still many undeveloped areas on the horizontal axis, which is likely to be the direction of development in the future.

Although the physical limits are determined by the performance of the tires, brakes, and the like, technologies such as ESC have been developed that support driver operations at these limits. These technologies are now also being integrated with steering controls. In the future, it is likely that development will continue toward the commercialization of vehicle dynamic control that can stretch the possibilities at the physical limits. This development will be based on research such as the verification of theoretical limits when four-wheel independent steering and four-wheel independent braking and traction systems are combined with force control technology through suspension control.

In addition, technology to assist the cognition and judgment phases is likely to become more sophisticated as recognition technology advances. Coordination between dynamics control to enable automatic evasive maneuvering in the lateral direction will probably progress. In preparation for these developments, vehicle dynamics control must have the capability to freely control vehicle behavior. Integrated control technology will play a major role in accomplishing this aim.

6.2 Future trends

Figure 31 shows a matrix depicting the concept for total vehicle system integration. In addition to the integration of energy within the vehicle, further integrated controls are being considered that incorporate the vehicle, driver, and the traffic environment in the same way as active safety. Possible methods of helping to improve the environment

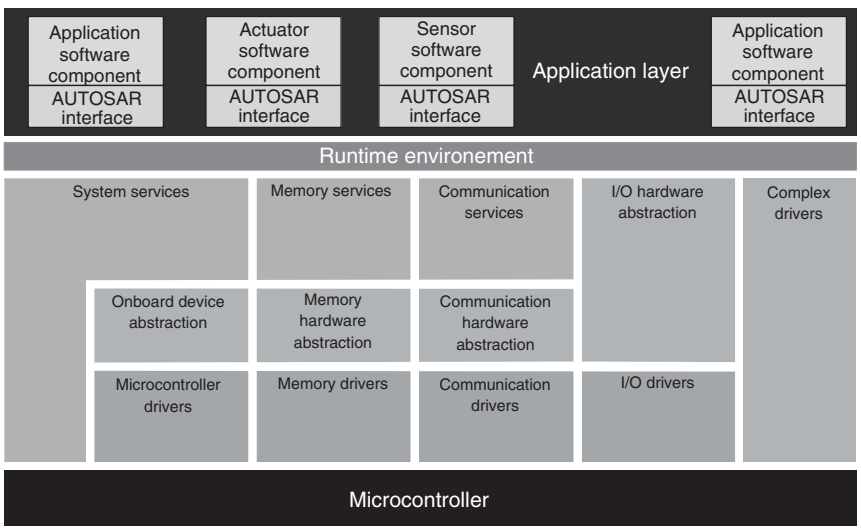


Figure 26. Basic AUTOSAR architecture.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

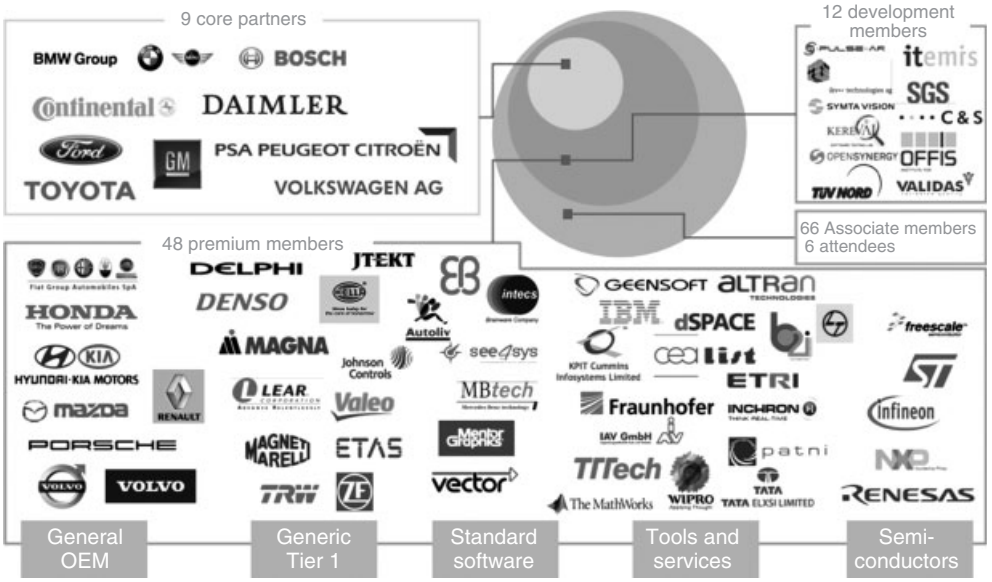


Figure 27. AUTOSAR partnership and members (2010).

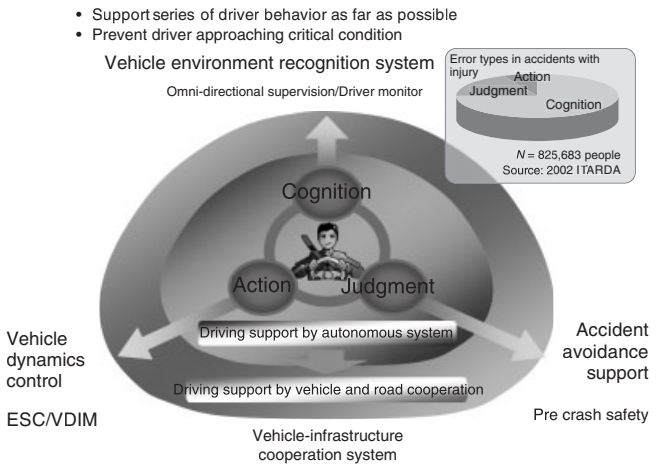


Figure 28. Future trends in active safety technology.

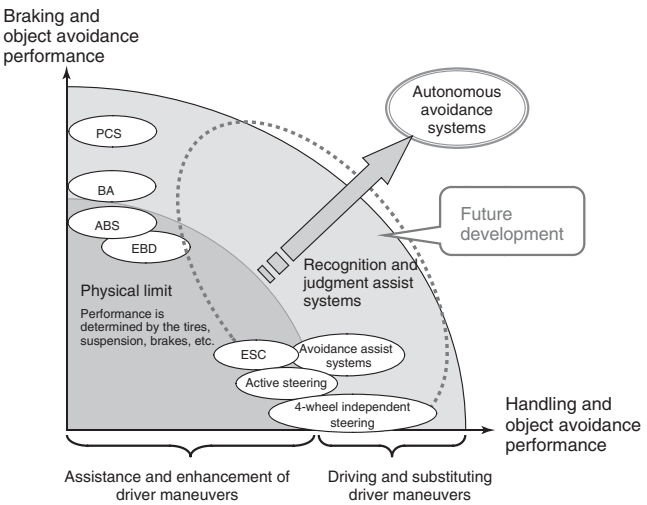


Figure 29. Current control systems and future development areas.

include the shift lever indictor and CO_2 reduction control using ACC. Integrated functions based on traffic signal controls and ITS (information technology services) that also factor in the traffic environment may be developed that help to alleviate congestion.

From the standpoint of active safety, the evolution from integrated vehicle dynamics controls to integrated active safety controls will focus on the development of driver monitoring functions in PCS systems. In the future, these technologies will develop into systems that provide assistance to individual drivers as appropriate in coordination with the driver and integrated safety systems that are coordinated with the traffic environment and infrastructure. The matrix in Figure 30 is also related to the achievement of sustainable mobility in terms of vehicle safety, the environment, and comfort. In addition to the autonomous and infrastructure-coordinated driving environment detection technologies shown in Figure 31, an area of growing importance will be individual applications designed in accordance with navigation system and traffic control ITS information and the state and personal characteristics of the driver. Therefore, an integrated HMI that incorporates individual information, instructions, alerts, and warnings will be the key for creating an integrated control that helps to enhance safety, the environment, and comfort.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

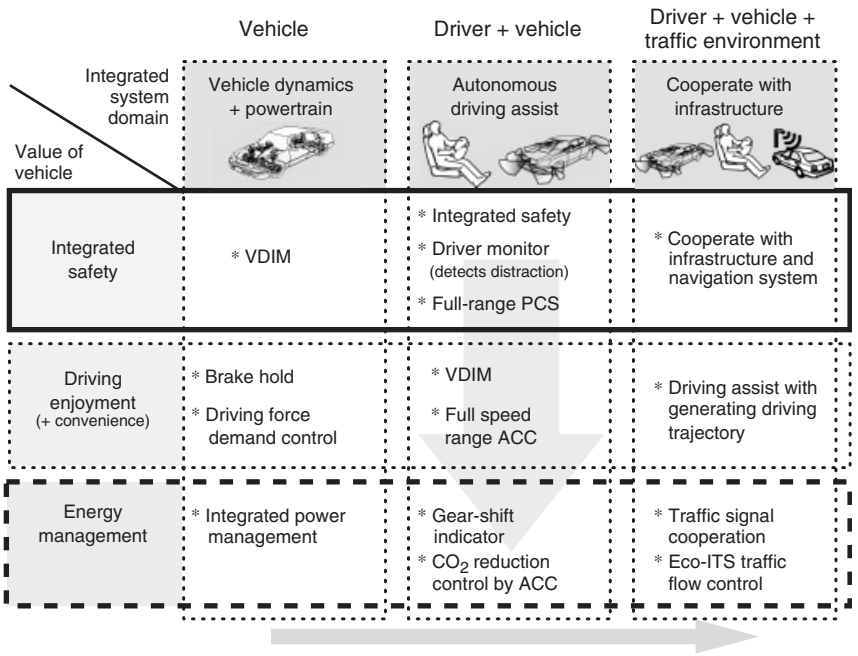


Figure 30. Total integrated vehicle system concept.

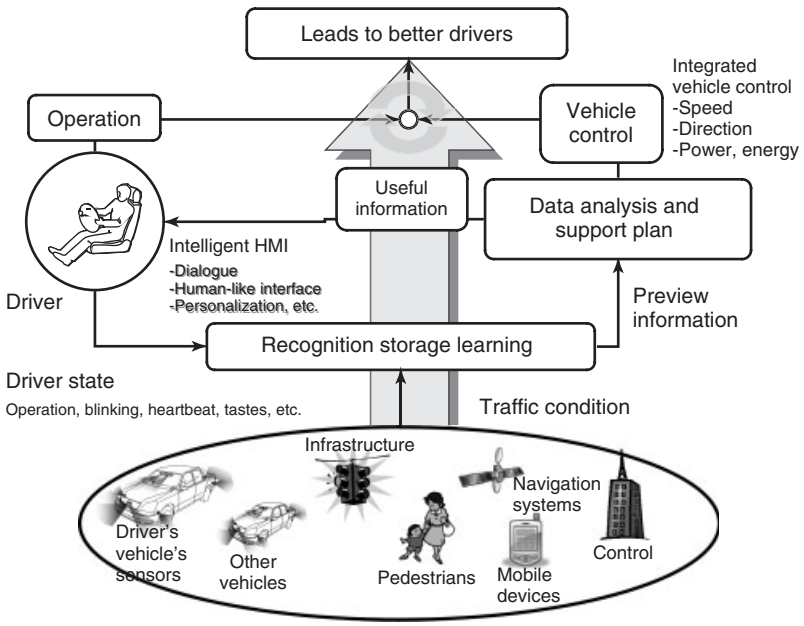


Figure 31. Integrated driver assistance concept.

7 CONCLUSION

Dynamic control technology for controls related to the suspension, steering, braking, driving forces, and the like is advancing relentlessly. At the same time, cognition and judgment functions equivalent to the eyes and brain are also being rapidly developed. However, the number of people hurt or killed in traffic accidents remains at a high level.

Therefore, this technology has to be consolidated and applied properly to fulfill the responsibility of automakers to develop vehicles that do not cause accidents. Control systems are required that are highly reliable, flexible, and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

have the potential for widespread use. This chapter has described the trends and configurations of these systems. Comfort and driving enjoyment are essential parts of a vehicle and these must not be sacrificed. For this reason, the development of these systems will continue while enhancing basic vehicle performance.

RELATED ARTICLES

New Electrical Power Steering Systems
Steer by Wire, Potential and Challenges
Automated Driving
Torque Vectoring by Drive Train Systems
In-Vehicle Network
Interfaces between Sensors and ECUs
Various Kinds of Actuators and Signal Interfaces
ECU Chassis (Steering)
Chassis ECU (Vehicle dynamics, ABS)
Body ECU (airbag)
Chassis ECU (ACC and sensor)
Active Safety, Pre-collision Safety and Other Safety Products (millimeter wave, image recognition, laser)
Human Machine Interface Design in Modern Vehicles
Technologies—Communication: Broadcast
In-vehicle sensors
Applications—Intelligent Vehicles: Driver Information Driver Assistance
Applications—Intelligent Vehicles: Autonomous Vehicles
Applications—Intelligent Roads and Cooperative Systems: Urban Traffic Management

REFERENCES

Fukatani, K. (2005) Vehicle dynamics integrated control system: overall vehicle modeling and control, Modelling and control of the car understood thoroughly; Workshop teaching materials, published by Nihon Machinery Society, 43–50.

German patent DE 35 05 455 CS (1985) Vorrichtung zur automatischen Zu- oder Abschaltung von Antriebselementen eines Kraftfahrzeuges.

Hattori, Y. (2002) Force and Moment Control with Nonlinear Optimum Distribution for Vehicle Dynamics. *AVEC*, 02024.

Hattori, Y., Koibuchi, K., and Yokoyama, T. (2002) Force and Moment Control with Nonlinear Optimum Distribution for Vehicle Dynamics. *JSAE Transactions*, **35** (3), 215-221.7.

Katayama, K. (2007) Development of four-wheel active steer system. *Nissan Technical Review*, **60**, 5–9.

Koibuchi, K., Yamamoto, M., and Fukada, Y. (1996) Vehicle stability control in limit cornering by active brake. SAE Paper 960487.

Kojo, T. (2002) Development of Front Steering Control System. *AVEC*, 149.

Mokhimar, O. and Abe, M. (2003) Effects of an optimum cooperative chassis control from the view points of tire workload. *Proceedings of Society of Automotive Engineers of Japan Annual Congress*, **33**(03), 15–20.

Nakamura, E. (2002) Development of electronically controlled brake system for hybrid vehicle. SAE World Congress 2002-01-0300.

Ono, E. (2007) Improvement in Safety and Comfort by Active Four Wheel Steering. *Proceedings of the Society of Automotive Engineers of Japan*, 11-07.

Ono, E., Hosoe, S., Tuan, H., and *et al*. (1998) Bifurcation in vehicle dynamics and robust front wheel steering control. *IEEE Transactions on Control Systems Technology*, **6**(3), 412–420.

Ono, E., Hattori, Y., Aizawa, H., and *et al*. (2009) Clarification and achievement of theoretical limitation in vehicle dynamics integrated management. *Journal of Environment and Engineering*, **4**(1), 89–100.

Shibahata, Y., Shimada, K., Tomari, T., and *et al*. (1993) Improvement of vehicle maneuverability by direct yaw moment control. *Vehicle System Dynamics*, **22**, 465–481.

Tanaka, H., Inoue, H., and Iwata, H. (1992) Development of a vehicle integrated control system. 24th FISITA Paper Award.

Tsuchida, J. (2007) The advanced sensor fusion algorithm for pre-crash safety system. SAE International World Congress 2007-01-0402.

Van Zanten, A.T. (1996) Control Aspects of Bosch-VDC. *AVEC 96*, pp. 573-608.

Yamakado, M. and Abe, M. (2008) An experimentally confirmed driver longitudinal acceleration control model combined with vehicle lateral motion. *Vehicle System Dynamics*, **46**, 129–149.

Yamakado, M., Takahashi, J., Saito, S., *et al*. (2010) Improvement in vehicle agility and stability by G-vectoring control. *Vehicle System Dynamics*, **48**, 231–254.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto030
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Chassis Control Systems

Gerhard Klapper and Ralf Leiter

TRW Automotive, Koblenz, Germany

1 Introduction 1
2 Steering 1
3 Suspension and Damping 7
4 Chassis Control by Using Vehicle Dynamics Theory 12
5 Central Computer Compared to Distributed Intelligence 15
6 Communication of Distributed Intelligences 16
7 Summary 20
References 20

1 INTRODUCTION

The appreciation of vehicles and the differentiation between the manufacturers and models is mainly gained from the increased use of more and more complicated control systems, including their electronic components. This trend will rapidly accelerate in the near future, as indicated by the top models from various manufacturers already on the market. In some companies, there is still the "world car" idea; but in the meantime, nearly everybody has also learned that different regions require different coordination of the chassis. Moreover, the greatest dream is still to achieve this using software coding.

In general, the target of the chassis development can be described as monitoring the vertical wheel forces in a way that the necessary longitudinal and lateral forces can be transferred to the road with simultaneous maximum comfort and safety for passengers and minimal fuel consumption. The classic chassis presents many compromises in design and performance; the use of electronics permits design shifts to improve the performance even further.

Legal requirements and environmental protection can be achieved by bringing microprocessors into the vehicle. The exhaust laws of today can only be met by regulating ignition timing and injection quantity in dependence of the catalyst sensor. This was followed by safety systems such as antilock braking system (ABS) and the airbag, which brought additional microprocessors into the vehicle. Over time, components became so reliable that now complex control strategies can be implemented with simple control systems.

This chapter defines the classic components as the chassis system:

steering,
suspension and damping, and
chassis and brake control.

These components cover the controls for longitudinal and lateral guidance of the vehicle, while simultaneously controlling the vertical movement.

2 STEERING

The steering of a vehicle presents the most sensitive components in the control loop of driver-vehicle. The driver has decisive influence over the behavior of the machine to be operated. A vehicle is moving dependent on the feel and the individual capability of the vehicle's driver. Next to the general vehicle design, steering is the most important feedback to the impressions and control information reported to the driver.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.

This article is © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

The driver receives feedback about the street, the environment, and the vehicle behavior from the respective steering design that he can immediately incorporate into his behavior to operate the vehicle safely. This is a variety of visual (e.g., deviations from the wanted position on the lane), acoustic (e.g., warning strip noises from the tires), and dynamic information (e.g., the torque on the steering wheel, the yaw angle velocity, or also the lateral acceleration) that the driver must simultaneously receive and process. The driver can derive and implement the necessary control interventions with regard to steering the vehicle, releasing the acceleration pedal or even pushing the brake pedal. Since the drivers themselves function as dynamic controllers, it is important that they do not make the entire system unstable. Their basic dynamic behavior is a second order system, which can become unstable.

Today's steering systems do not only ensure the sensors and feedback of important information for driving a vehicle, but depending on execution, they also provide much additional assistance for the driver. This applies to the typical simple servo-support (auxiliary power units to relieve the steering forces the driver must use) until the steering movement support that is mainly dependent on the speed (additional steering angle), in some situations even from the yaw rate sensor.

The latest step forward in steering has been realized by electrical power steering systems. Even those systems can be developed further. The different types of electrical power steering are making this obvious.

2.1 Electrically powered steering

Especially for smaller vehicles with less axle loads, electric drives for power steering are becoming reality. With regard to positioning and space required, the electric drive is superior to the hydraulic drive. This is also true with respect to fuel consumption.

Electric drives reduce the range of their components to:

- cable harness,
- drive motor,
- clutch,
- transmission, transmission housing,
- steering wheel torque sensor,
- control unit or partial integration in present control unit, and
- mechanically functioning steering gear.

In addition, advantages of the electric assistant steering systems are:

- reduction in the number of parts (no pump, pump drive, oil container, and hydraulic lines),
- less weight,
- less space required,
- independent of vehicle motor, operation as possible with combustion engine shut off,
- less loss in driving power,
- better adaptation to the power requirements, reducing gasoline consumption,
- steering force regulation as needed: speed or load dependence is possible, and
- no need for inconvenient hydraulic maintenance.

There are also disadvantages that limit the complete introduction of the electric power steering in the automotive industry. They are evidenced by:

- limited performance, dependent on available energy sources (12-V electrical system);
- considerable protection needed from system failure due to use of electronic components;
- increased development expenses due to software and hardware adaptation work; and
- ever-increasing manufacturing costs (compared to hydraulic power steering systems).

Potential installation positions for the electric steering assist are more numerous than with hydraulic systems. Figure 1 gives an impression of the different possibilities. Electric motors can also be arranged:

- directly under the steering wheel on the steering column;
- like all hydraulic systems on the steering gear and on the rack-and-pinion; or
- for example, using a belt on the recirculation ball gear.
- in parallel to thes steering rack.

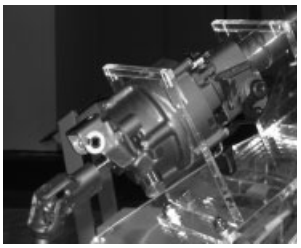
Electric power steering systems can also be preassembled modularly, as they are independent of hydraulic lines, which enable verifiability during production and can significantly reduce the production costs again.

The various manufacturers of such systems on the market use a variety of different drive and sensor concepts. Unlike with hydraulic systems, the steering angle and also torque applied are recorded by sensors and converted to electric signals. The control unit evaluates these signals and converts them into corresponding commands for the drive motor. This now works on the mechanical steering gear with the torque associated with the command.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



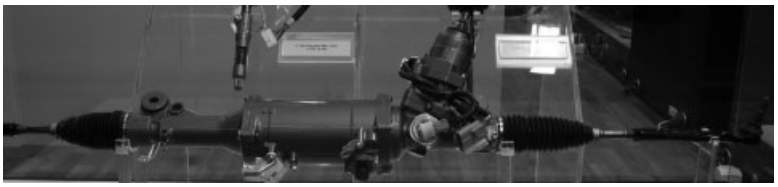
(a) Electro-hydraulic steering



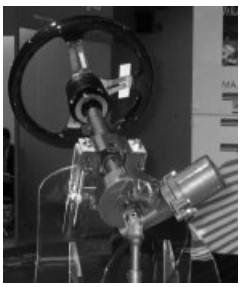
(b) Electric actuator in line with the upper part of the steering column



(c) Electric motor at the pinion



(d) Electric actuator in line with the rack



(e) Electric motor rectangular to the upper part of the steering column

Figure 1. (a–e) Potential installation positions for the electrically powered power steering. (Reproduced with permission from Henning Wallentowitz. © H. Wallentowitz.)

2.1.1 Column drive (drive arranged on the steering column)

In this application (Figure 2), the electric motor is directly fastened to the steering column. The unit can be pre-manufactured separately from the vehicle assembly and delivered as a module to the assembly line.

The additionally required electronic control unit can either be integrated into the module or be positioned separately from the servo unit in the vehicle. Depending on the vehicle concept, the control unit can also be integrated in other control units. Synergies with this procedure are evident in the reduction of electronic components. Reductions in costs are the result.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 2. Steering tube drive. (Reproduced by permission of TRW.)

The steering tube or the steering spindle between the steering wheel and the steering gear are powered.

To record the necessary input sizes, a torque sensor, or in individual cases, a rotary sensor and possibly according to requirements also speed sensors are integrated into the system.

2.1.2 Rack-and-pinion drive (with concentrically arranged motor)

The application in this case requires a rack-and-pinion steering with side pickup. Accordingly, other potential rack-and-pinion variants are conceivable, however, with more effort for the realization. Placing a coaxially working motor requires corresponding installation space in the middle region of the rack-and-pinion mechanics. Figure 3 shows one possibility. Often, the engine/gearbox unit requires this space.

2.1.3 Belt drive

In order to save space and still be able to directly grasp the rack, this variant (Figure 4) is available for electrical application of the assistance. The overall design of the unit provides significantly more clearance. In individual cases, the correct variant of those possible must be determined by an installation study, due to difficult installation dimensions in the front of the vehicle. The corresponding suggestion is therefore to evaluate cost/effort.

As is typical for the introduction of electronic components in safety-relevant vehicle systems, the demand for quality controls and safety tests increases considerably.

This is supposed to prevent critical driving conditions and drive situations, caused by malfunctions in the electronic components.

Such malfunctions can occur with:

- Too high current load at the motor
- Motor blockade due to seizing components such as bearing or rotor body

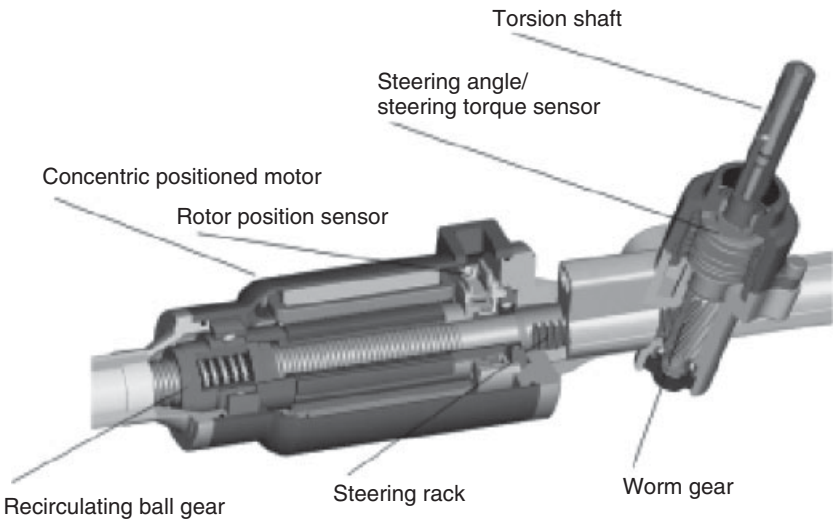


Figure 3. Steering gear drive sectional view. (Reproduced by permission of TRW.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

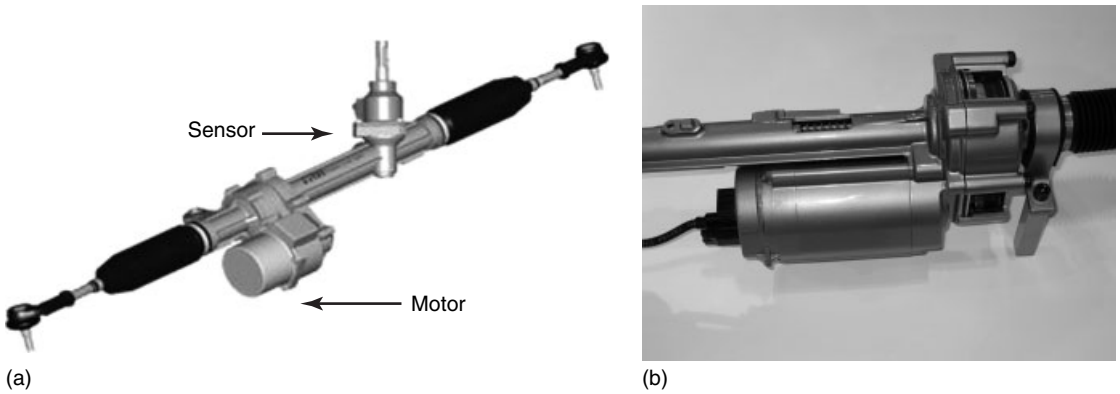


Figure 4. (a, b) Belt drive. (a: Reproduced by permission of TRW, b: Reproduced by permission of H. Wallentowitz.)

- Manipulation of electronics or the electrical system
- Sensors that give incorrect signals

To avoid all these fault possibilities, vehicle manufacturers exclusively use high quality materials and components. In addition, in production, up to 100% checks can be used (production of safety components).

In addition, redundant systems are required and introduced, which can balance the failure of parallel working systems on one hand, and on the other hand find use for control purposes of main systems and switch off the system if there are deviations in the signals.

Furthermore:

- The current values are technically limited;
- Voltages and sensors are permanently checked; and
- Torques are monitored.
- As further safety equipment, plug connectors that cannot be manipulated and so-called *electric couplers* are used.

Self-diagnostic programs constantly monitor all-important functions. This occurs by comparing the actual announced values with the specified target values or target value windows.

Fault detection shuts down the auxiliary systems and simultaneously notifies about faults using the system lights in the dashboard. The driver also recognizes the fault function by the increased steering forces. These behave exactly like with a strictly mechanical steering system without power steering.

Since the 1980/1990s, there have been cars with rear wheel steering in the market (mechanically powered (Abe, Ogura, and Sato, 1988), or different hydraulically powered systems (Donges, 1989; Kuroki and Irie, 1991). Now, there are systems coming to the market with electrically powered rear wheel steering systems. They may be considered in the next paragraph.

2.2 Active rear steering

Four-wheel steering is a known method to improve the maneuverability at low speeds (opposite impact to reduce the curve radius) and to enable better vehicle stabilization at high speeds. The simplest algorithms adjust the steering angle of the rear axle linear to the steering wheel angle. Advanced algorithms support the development of the yaw moment, dependent on speed, first with opposite turning, but then control back to the same direction, releasing the yaw angle and increasing driving stability (Abe, Ogura, and Sato, 1988).

The active control of slip angles at the rear wheels allows a targeted increase of the cornering forces. This can influence the steering of a vehicle (e.g., development of sideslip angle in the vehicle's center of gravity). Side wind influences can also be automatically compensated, without the driver needing to intervene.

Four-wheel steering can make the vehicle significantly more agile, but it can also change lanes without (in reality with reduced) a yaw movement occurring. That opens new possibilities to the algorithms of automatic accident prevention. The first electronically operated system has been introduced by BMW in the beginning of the 1990s (Donges, 1989; Wallentowitz, Donges, and Wimberger, 1994).

2.3 Superimposed active steering

Another modern steering system, which is in the market since 2003 (Köhn *et al*., 2003), is the superimposed steering system. AUDI followed with an own solution in 2007. This

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 5. Electric superimposed steering. (Reproduced by permission of EBM-Papst.)

is the first step to combine steering commands generated by the driver with steering movements coming from an electronic.

The superimposed steering makes it possible to realize a steering intervention on the front axle independent of the driver, without separating the mechanical coupling between steering wheel and front axle. An additional gear, an electric motor, sensors, and a control unit, expands the normal steering system. Figure 5 shows the additional gearbox in such a system. This system now enables the continuous change of the steering gear ratio—dependent on the driving situation detected. The effective steering angle at the wheels can be larger or smaller than the driver adjusted with the steering wheel. This just depends on whether the auxiliary system is steering in the same direction as the driver, or whether it is steering in the opposite direction. Thus, in city traffic, there is less steering required by the driver; at higher speeds, the vehicle can be better stabilized, thanks to an automatic and limited "counter-steering" of the system.

If the electric motor is not actuated, there is a direct mechanical connection between the steering wheel and the front wheels, as with conventional steering. Thus, such "intelligent" steering fulfills the existing legal requirements.

2.4 Steer by wire

In the last level of the current development, the driver is mechanically and hydraulically completely separated from the power units of the steering systems (like with the control of engines and automated transmission gear boxes). Steering angles are accepted as input parameters by the driver using the steering wheel or in the future, the so-called *joystick*, but are converted into electronic control commands for the actuators, moving the wheels.

Currently, law still requires permanent connection either mechanically or hydraulically (then limitation) vehicle speed to be able to react as needed if the electronics fail. This is called *fully redundant safety measures*. However, the entire vehicle industry is working on other "clever" systems to avoid this expense in the future and to be able to change the laws accordingly.

A complete separation of the input interfaces of the power aggregates would lead to:

- Enormous savings in materials,
- Reduction of parts, and
- Crash behavior improvement

The environment can also profit from this, as no poisonous hydraulic fluids would have to be disposed of. Rubber seals are not used in the large circumference; expensive manufactured tubes and hose lines are no longer needed; as well as a part of the mechanics, using a lot of energy during manufacturing, will not be necessary any longer.

The steering reaction can from this time forward be coordinated to all potential input parameters of the variety of sensors already present in today's vehicles (steering angle and steering torque sensors, yaw rates information, or camera images). Even safety-promoting interventions of the electronics into the handling (already realized in partial areas) are conceivable.

The potential configurations with redundant mechanical backup systems could look as shown in Figure 6.

In the basic idea, the "steer by wire" will not have any mechanical connection between the driver and the front wheel. Obviously, it is extremely important that in this case several redundant electronic systems must efficiently, and above all, safely prevent the steering from failing.

Figure 7 gives an impression of such an idea. Nissan is just now suggesting a similar system, but in addition to the electronic backup, there will be still a mechanical backup with a clutch, which is normally open (Kubota).

Ultimately, the vehicle manufacturers and their development partners have not yet reached series introduction of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 6. Fully electronic steering with mechanical backup. (Reproduced by permission of TRW.)

these systems. The cheaper and always mechanically redundant hydraulic power steering used in series today is still the main percentage of power steering systems used on the market. Electric/electronic steering systems suffer the enormous price pressure under which the vehicles must be built today. An additional hurdle is coming from the numerous electronic components, which are installed mainly for comfort reasons. This stresses to the limit of the currently used 12-V electrical systems of the vehicles. Fundamental changes of the electrical infrastructure of today's vehicles are currently being controversially discussed from this viewpoint (Amsel, 2003; Wallentowitz and Amsel, 2003).

3 SUSPENSION AND DAMPING

Advanced suspension systems are realized by air springs and by hydraulically actuated conventional steel springs. This chapter describes some of the systems and the possibilities for advancements.

3.1 Air spring

The trend for more comfort in the passenger compartment and less vibration in the vehicle's suspension, independent of load, for continuous control of the height level as well as level adaptations to entry and loading assists, air suspensions in passenger cars, after a first application in the 1960s, became common equipment in passenger cars of the upper class in the 1990s. The suspensions of the 1960s are not comparable to the modern systems.

In semitrucks, where, thanks to the pneumatic brake system, the air energy is already available, air suspension for street transportation vehicles has become routine. As the air compressor and the air dryer are consuming a lot of energy, even here changes are wanted.

Today, the compressors still mostly work in open systems. The air is taken from the environment, compressed and either pressed directly into the air spring or is kept available in a reservoir. Excess air (when lowering the vehicle structure) is returned to the environment. The vehicle structure can only be raised relatively slowly and with limited frequency (load of battery and compressor). Closed systems found a solution where the air is pumped

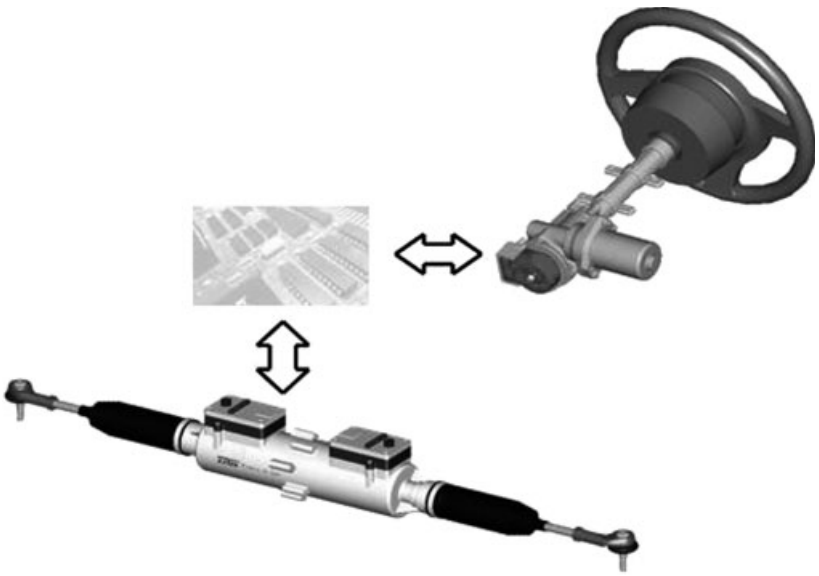


Figure 7. Fully electronic steering with electronic backup. (Reproduced by permission of TRW.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 8. Air spring with variable effective area. (Reproduced by permission of Meritor.)

back and forth between the spring and reservoir at high pressure. This clearly permits faster-level changes with less needed energy. This technology now also allows lowering the chassis while driving (lower air resistance) or raising the vehicles body (more clearance with poor street surface or driving out in the country).

An interesting variant is changing the effective area of the air springs with an additional interior air bellow. This allows the spring stiffness to increase threefold in a few milliseconds or to decrease without needing much air volume. Figure 8 shows this spring in soft condition (a) and in stiff position (b). There is only compressed air necessary to blow up the small air spring, which changes the rolling piston diameter (Lloyd, 2010).

This "active air spring" enables various operating modes, adapted to the vehicle driving conditions. The main advantage can be created during cornering. The wheel on the outer side of the curve not compress as much with a stiff air spring; the car body remains more horizontal. During acceleration or braking, the anti-dive and anti-pitch can be organized by these air springs. Unfortunately, these types of air springs are not commonly used in the market. For keeping the car horizontal, today active stabilizers are used. Their functionality is described in the next paragraph.

3.2 Active stabilizer

Conventional mechanical chassis are always a compromise between comfort and driving safety. When driving in curves, the vehicles roll around the rolling axle. The roll angle depends on the lateral acceleration. The softer the stiffness of the body springs and the larger the distance between center of gravity and rolling axle, the more roll angle can be built up. This soft roll stiffness can even cause roll over (elk test) if there are quick directional changes and simultaneous minimal roll damping. To decrease the tilt angle, the stabilizer bar rigidity must be increased.

The stabilizer is an additional connection between chassis and structure. The roll moment when driving in curves leads to the wheel load differences becoming larger with increased lateral acceleration (but the sum of wheel load of an axle remains the same). The wheel on the outside driving around a curve compresses and the wheel at the inner side rebounds. Its coupling rod rotates the stabilizer, and the forces at the bearing point create a torque, which counteracts the roll movement of the car's body.

The rigidity of the stabilizer has considerable influence on the roll stiffness of the vehicle. With the coordination of front and rear stabilizers, the vehicle behavior is changed between understeering and oversteering. Additional influential parameters for the driving behavior are the stiffness of the springs between chassis and body (as already mentioned above) and the position of the roll centers on the front and the rear axle. The specific roll center height results from the design of the used axle. The roll centers may be above the road, on the road, or even below. They are (mostly) virtual points.

As the stabilizer represents an additional spring, which is also active for one-sided deflection, named copying, active stabilizers have been invented. They reduce copying and they can in addition reduce the roll angle of the vehicle during cornering. Active stabilizers are realized with electric or hydraulic actuators in two or more stages. Yaw rate, lateral acceleration, steering angle, and speed are

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

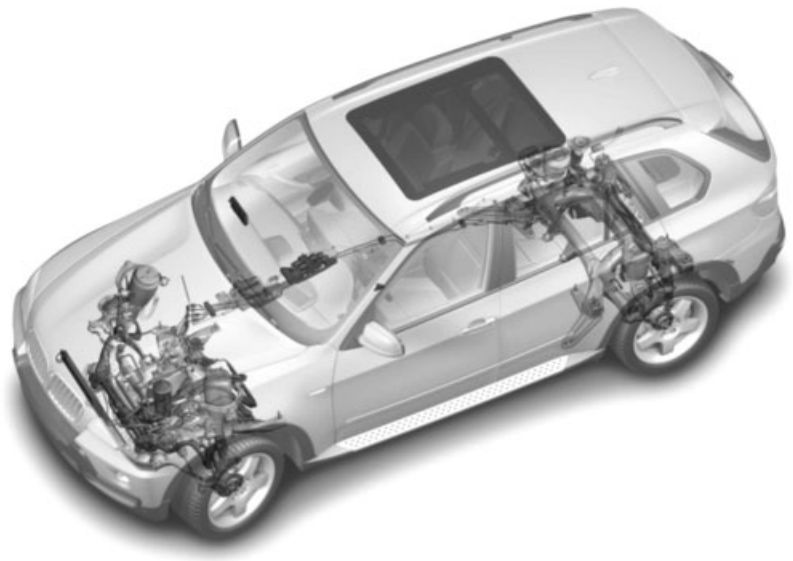


Figure 9. BMW with active chassis control (Dynamic Drive) (Jurr *et al*., 2001; Konik *et al*., 2000). (Reproduced by permission of BMW AG.)

considered as incoming parameters. Two pressure sensors, six valves, and one electronic controller provide the "right" driving behavior, which is adjusted using hydraulic pressure from a motor-powered tandem pump. Figure 9 gives an impression about the installation of the system in a car.

Hydraulic motors can even actively intervene in the chassis regulation with a very quick electric tandem pump unit. However, the pump power must be approximately as great as that of the electric hydraulic power steering. This leads to short-term electrical system load >100 A.

After switching on, the control units run a self-test and perform a functional self-test on the connected valves and sensors. Body control does not occur with standing vehicle. At approximately 20 km/h, the systems are active. Cornering is detected by the lateral acceleration signal, steering wheel angle, and speed. With up to 180-bar hydraulic pressure, counter-torque of 600 Nm (front axle) and until 800 Nm (rear axle) is generated on the stabilizer. At higher speeds, understeering drive behavior is preferred; thus, a higher roll torque amount is required on the front axle. To increase agility, a higher roll torque is temporarily set on the rear axle.

Simpler active stabilizers, especially with all-terrain vehicles, can only decouple the two stabilizer ends on driver command and permit an increased linking of the axles, which significantly improves the traction out in the country. Logically, this occurs at speeds below 50 km/h. When driving straight-ahead, the comfort is improved by decoupling (minimizing the copying tendency), but for normal cornering, these stabilizers work like rigid coupled ones.

3.3 Hydraulic chassis

The hydro-pneumatic suspension (a gas suspension with interconnected oil column) was already used in 1953 in passenger cars from the Citroen Company. Every wheel is equipped with a suspension/damping element, which contains nitrogen in the upper half of the hydro-pneumatic accumulator. The gas (constant mass of the gas) compresses and releases during the wheel movements and takes care of the reaction forces with respect to the wheel loads. A rubber membrane separates the gas from the hydraulic oil in the lower half of the strut.

During compression, the piston presses the oil through the pressure valve, arranged in the cylinder. During rebound, the gas presses the oil down through the harder adjusted rebound valve. In principle, the strut is an upside-down one-tube damper. The design level position in the vehicle is set using a lever-operated valve by adding or releasing oil, independent of load.

In the mid 1990s, Citroen added an active hydraulic stabilization system to its hydraulic suspension. This system was able to improve the driving comfort, but all in all, hydro-pneumatic suspension was replaced by air spring suspension. Only special solutions are still on the market, such as the system mentioned in the following paragraph.

3.4 ABC chassis

The active body control of Daimler AG is an electric-hydraulic chassis to meet the goal to always keep the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 10. ABC chassis components (Wiesinger). (Reproduced by permission of Daimler.)

vehicle structure at the same height (Wiesinger). In addition to steering angle, wheel speed, yaw moment, and longitudinal and lateral acceleration, the height of the vehicle body is sensed against the wheels. The control unit "observes" the body behavior, and it can almost completely compensate the vertical body movement, the roll, and pitch with assistance of hydraulic cylinders on the support of the coil springs. The base points of the steel springs are hydraulically controlled with frequencies until 5 Hz—the shock absorbers are responsible for the wheel damping.

Electric-magnetic stop and control valves permit the individual control of each strut. With this system, individual wheel loads can be specifically increased or decreased, depending on what the drive situation requires. It is also easily possible to use wheel load control to optimize the brake and side forces in the tire contact area. Acceleration sensors on the vehicle body support the control strategy. Roll torque can be distributed to front and rear axle. A car with ABC suspension does not have a stabilizer bar.

Figure 10 shows the chassis components of such a suspension.

Right before opening the driver's door, the vehicle is lifted to entry height. When driving in curves, the outer springs are pressed together. To reduce roll, the spring base point is lifted until the compression stroke of the spring has been compensated. While braking, this base point tracking occurs on the front springs that are compressed. On the rear axle, the steel spring releases and the base point shifts downward. During acceleration, the base points are shifted just opposite. This compensates pitch and bounce movements.

Level sensors record bumps and correct for these. In a speed range between 65 and 140 km/h, the vehicle body is continuously lowered to 15 mm. Another 40-mm clearance is available with the push of a button. At 80 km/h, the crosswind stabilization is available, which can, however, be canceled again by quick and strong steering movements.

The advantages of such a system are: the car body remains horizontal, even during cornering, accelerating, or braking; mechanical stabilizers are no longer required; a radial piston pump provides the system with up to 210-bar oil pressure; and individual pressure sensors monitor the control movements. One disadvantage is the additional energy consumption to operate the system.

Pre-scan system is just now in series, which allows the ABC system to work with foresight by laser scanning the lane in front of the vehicle, but such systems are still under development.

3.5 Adaptive damping

In 1987, BMW was the first manufacturer in Europe to introduce adaptive damping. The previous systems could only select between several fixed damping characteristics (sport, normal, and comfort), but the BMW EDC has already worked with continuous adjustment. Today's systems also do this. Control programs that use the various areas of the shock absorber parameters are used in every case. In addition to automatic control, the driver can tune the system toward sportier or more comfortable operation as desired using a switch. However, in any way, the system reacts continuously. The damper characteristics are adapted to the driving situation as much as possible.

The former situation with the fixed settings of the damper can be explained by the so-called *conflict diagram* (Figure 11). Every combination of body spring stiffness and damper characteristic delivers driving comfort and driving safety results above the limiting curve. There is no simultaneous optimization of driving comfort and driving safety possible. To overcome this limiting curve was only possible with adaptive dampers.

Three acceleration sensors (two on the front axle, one on the rear axle), the steering wheel angle sensor, and the wheel speeds are the input parameters for the system, which then can adapt the chassis to the driving situation using the four adjustable dampers.

Usually, forces are calculated from the signals. Different simultaneous active control loops try to determine the optimum damping for the respective driving situation. There are vertical, lateral, and longitudinal control algorithms installed. The algorithm with the hardest setting determines the damping characteristics to be set on the shock absorbers. If possible, the soft setting is used until another setting is required.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

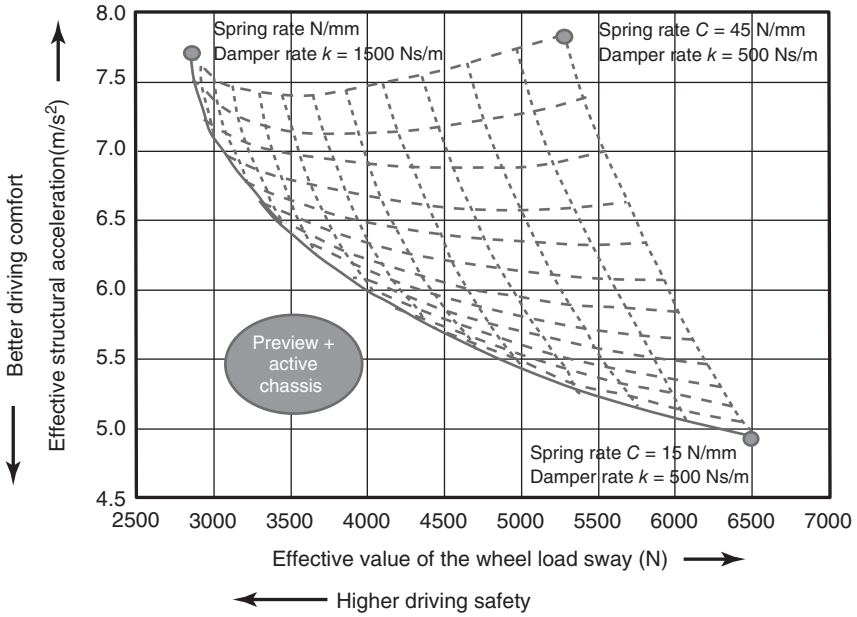


Figure 11. Conflict diagram passive wheel suspension.

The vertical dynamics controller (also called *hub controller*) reacts to the movements of the vehicle in vertical direction. Input parameters are wheel and body accelerations and their time integrals, that means wheel and body speeds. The wheel vibrations are between 10 and 16 Hz. When driving on uneven roads, irregularities occur in the wheel speeds, which can be used to estimate the quality of the street surface. The vertical body speed (the body vibrates approximately with 1 Hz) is determined with consideration of the driving speed, the frequency range of the street surface, the load, and the size of the excitation. Every axle is first individually controlled; a downstream parallel controller then provides for parallel movements.

The longitudinal controller reacts to accelerations and brake processes. The wheel signals (two per cable directly from electronic stability program (ESP), two additional ones via controller area network (CAN), and the vehicle speed) are the input parameters to calculate the static and the dynamic portion of the longitudinal acceleration.

On the basis of this, most of the dampers for both axles are adjusted the same to better support the vehicle.

The lateral dynamic controller detects yaw moments very early on using driving speed and steering angle. The dampers, which are adjusted harder (separate regulation of front and rear axle), already permit better vehicle support during cornering.

The vehicle roll with one-sided compression and rebound of the wheels while driving straight is called *copying*. It is reinforced by the axle stabilizer, which transfers the compression of one wheel (as a result of unevenness, the vehicle structure is lifted on this side) to the other wheel on the axle and also allows this wheel to compress. This increases the roll angle of the body. This is uncomfortable. The vehicle body reacts to this one-sided ground unevenness when in the comfort setting. The right and left body accelerations that are measured above the front axle detect this copying. All dampers are set to be harder and work against this unwanted body movement.

The diminishing mechanical damping characteristics are compensated by a tolerance adaptation. Thus, even defect dampers can be identified. If there are systems faults, an average damping is set that would correspond with a passive chassis, or if there is a total failure, the dampers automatically go to the hard settings.

For comfort reasons, the goal is to drive as long as possible with "soft" dampers and only switch to "hard" for safety reasons. This takes energy from the body and the wheel movements. In Figure 12, a cross section of a damper piston with an adaptive control mechanism can be seen.

Whether the damper rate is set using magnetically influenced rheological fluid or by a proportional valve is only a secondary issue as seen from a technical view.

The recuperation of the energy, now converted into heat in the dampers and radiated into the ambient air, is currently discussed as source for electricity generation. Whether it will be used in series is yet to be seen.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

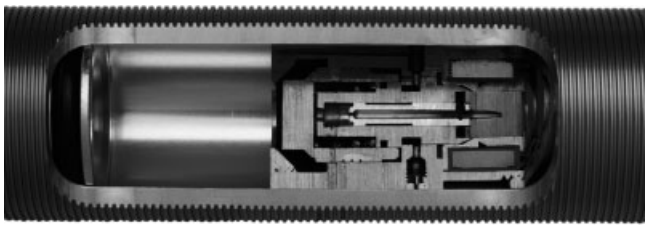


Figure 12. Adaptive damper, cross section of the piston. (Reproduced by permission of Meritor.)

3.6 Tire pressure monitoring

Air is a structural component of today's tires. It provides both lateral rigidity (cornering stiffness) and vertical stiffness (spring affect) of the contact surfaces between tire and road. If there is air loss, the tires lose a substantial part of this rigidity and they deform more than the design allows.

If there is not enough air in the tires, they deform more while rolling. This causes friction between the reinforcement materials (textures) in the tire and the rubber. Heat is generated in the tires. This heat leads to degeneration of the tire materials until pyrolysis—finally leading to total failure of the tire.

The sidewall reinforcements of modern tires are so large that visual monitoring cannot detect a pressure loss of 20%. Even daily pressure monitoring has its limits with strong temperature fluctuations between day and night. The pressure difference can easily amount to 0.4 bar (from 21 to 0°C). While the majority of drivers (82%) claim to check tire pressure at least every 3 months, tests have shown that only 13% actually do. The Ford-Firestone case made this clear. Even today, many accidents caused by tire damage can be led back to tire pressure that is too low. This was found out by checking the other tires of a car after an accident. In addition to lack of maintenance, more than 30% of the pressure gauges at the gas stations indicate the pressure too high. After the cost for fuel, tires are in position two on the maintenance cost of a car.

The longitudinal and the lateral dynamic behavior of the car depends on an intact interface between the tires and the road. If this is disturbed, these vehicle dynamics cannot reach their full capacity. Therefore, the demand is made more and more to combine the tire pressure monitoring (already legally required in the United States of America and Europe) with the brake regulating system, to adapt the controller of longitudinal and lateral dynamics to the pressure loss. This adaptation already starts with the simple determination of the real vehicle speed, which like wheel speed measurement is dependent on the wheel circumference. An additional potential function would be the reduction in the maximum speed due to pressure loss.

3.7 Tire pressure adjustment

With the availability of compressors fed by the electrical system (for instance, for air suspension) and the minimal probability of perception of indicator lamps, it is logical that an active tire pressure control will be useful in chassis control. These technologies are already available for military vehicles and for fire engines today.

The road condition identification of the ESP supply the input parameters for air pressure selection in the stages mud, sand, snow, country road, expressway, and escape. Obviously, the driver can also preselect the pressure. By entering a tire model and the automatic wheel pressure detection (e.g., using the air suspension pressure), the correct pressure is automatically selected. The controlling occurs per axle until permanent refilling for acute loss of pressure. This requires a permanent monitoring. The normal cycle of the pressure measurement every 15 min is reduced to 15 s in case of pressure loss detection. The maximum compressor power is used to stabilize the tires and thus the vehicle. If the system cannot balance the pressure, the driver is warned after 2 min.

4 CHASSIS CONTROL BY USING VEHICLE DYNAMICS THEORY

A wide field of chassis control was opened, when the microcomputers became more powerful at the beginning of the 1980s. Until that time, the ABS (antilock devices for brakes) electronics used specific hardware, which limited the flexibility of the control tasks.

With the expanded electronics, it was possible to use the bicycle model and to evaluate the car behavior as a result of speed, vehicle mass, wheelbase, center of gravity position, and above all the cornering stiffness of the front and the rear axle. Figure 13 shows the yaw amplification factor as function of vehicle speed (Wallentowitz, 2007).

Understeering vehicles have a declining characteristic and are defined with the so-called *characteristic velocity*. With the equation in Figure 13, it is possible to evaluate the vehicle characteristic by using the data. In case the yaw rate, the vehicle speed, and the steering wheel angle are measured, the real relationship between these data can be compared to the theoretical ones (evaluated by online simulation of the equation). In case the real relationship is below the characteristic line, the car is more understeering, in case this point is above, the car tends toward oversteering.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

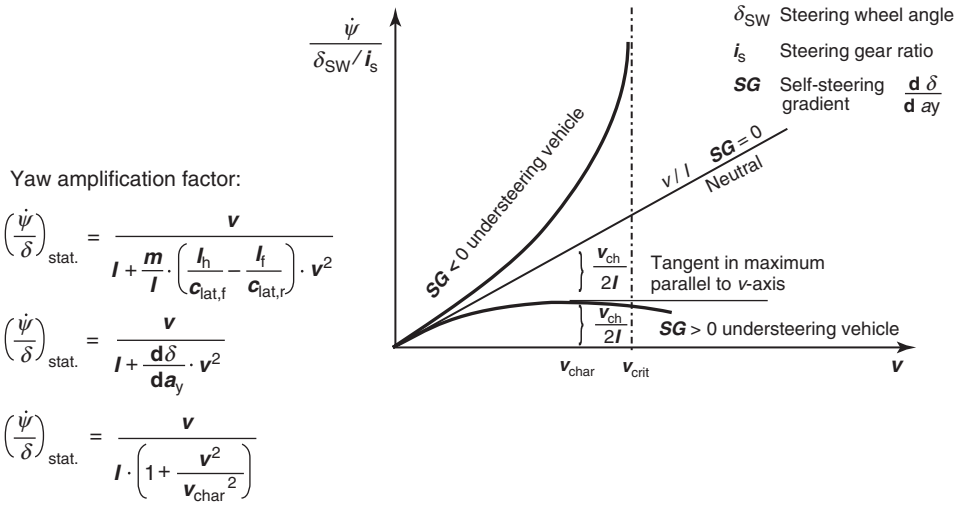


Figure 13. Yaw amplification factor as a function of vehicle speed. (Reproduced by permission of Henning Wallentowitz (2007).)

These theoretical background was used first in the Daimler-Benz All-Wheel-Drive 4Matic—starting in 1986 and in BMW dynamic stability control (DSC) system (dynamic stability program) in 1992 (Debes *et al*., 1997). The next step was the introduction of ESP in the market, which has been done by BOSCH in 1996 (van Zanten, 2000). Other companies followed. There is also the name TRC used.

The comparison of these data derived measures for switching on the all-wheel-drive or controlling the engine. The other systems used at that time only monitored the longitudinal dynamics. Thus, the ASR (traction control system while starting to drive) was realized over a motor interface, to quickly decrease the excessive motor torque at the source. In addition, braking was used to control the spinning wheels, to reach traction with the wheel that is not spinning. The drag torque control was possible due to the active throttle in modern cars.

4.1 Electronic stability program

Since 1996, the ESP is in the market. In addition to the DSC function (Debes *et al*., 1997), it uses brake applications to steer the car in the direction the driver intents to steer to. Under oversteering identification (comp. Figure 13), one front wheel is applied, under understeer, one of the rear wheels is braking, to keep the right driving direction. The values of yaw rate and sideslip angle in the center of gravity are the decisive information. To identify the actual sideslip angle, the measurement of vehicle speed and yaw rate is necessary. The sideslip angle itself is computed during driving.

In addition to the basic functions, electronic brake variator (EBV) (electronic brake force distribution), brake assist system (BAS) (electronic brake assistant), ABS (antilock system), dynamic traction control (DTC) (traction control while driving), ASC (automatic stability control), and DSC,

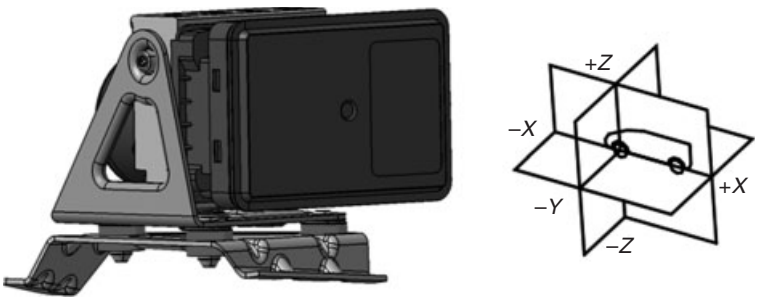


Figure 14. ESP with integrated sensors for yaw rate and longitudinal and lateral acceleration. (Reproduced by permission of TRW.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

modern ESP systems also have CBC functions (curve brake control—with consideration of the varying wheel loads in the curves and active braking of inner curve wheels), electronically controlled deceleration (ECD) (electronic braking due to other systems such as wheel sensors or electrical parking brake), HDC (driving down mountains), AAS (automatic trailer stabilization), and FLR (the control of driving performance—reduction for brakes that are already overloaded).

Even the brake lining-wear sensors are now evaluated by the ESP, in passenger cars still with one or two steps, in semitrucks already continuously. The originally separate electric parking brake is integrated into the ESP—the control unit must now execute the complete dynamic management (including the brake and warning light control) and the standstill management (including auto-hold) for the vehicle (Figure 14).

Internal temperature models of the brake disks are required for the fading brake support (FBS) (brake fading support) as well as for HTR (tension algorithms of the parking brake at high temperatures).

4.1.1 Networking of ESP and electrical steering

One function of ESP is steering of the vehicle by using the brakes. The driver enters the steering wheel and defines the direction. Single brakes are applied to follow this direction. Now, an automatically generated steering angle of the front wheels can be as effective as the brake intervention, but much more comfortable. And the combination of both actions can be even more effective.

If the vehicle oversteers, the combined system with an earlier steering intervention can possibly avoid the brake intervention. If the vehicle understeers, the steering wheel torque drops and the ESP begins with engine and brake interventions at the rear axle. An active feedback by steering already before interventions from ESP can help the driver avoid the situation.

If the combined system detects braking with different friction coefficients on the right and left lane side, it can automatically take on the necessary steering angle correction, without waiting for the driver. Driving and braking of a vehicle with a flat tire requires more skill to brake than the normal driver has. At the same time, a steering wheel turn is required. If the tire pressure monitor signals which tire is affected, the steering can compensate for the difference torque. By this additional steering input, the stopping procedure remains balanced. This allows quick stopping with high driving stability. This performance will be further improved with an additional networking with the lane assistant (usually realized optical via video cameras). This can realize an automated stabilization of the vehicle in the lane—when networked with the radar system even in the foresighted safe free lane. Fully automatic rescue support will thus be possible in the future.

4.1.2 Networking of ESP and transmission

The introduction of gears, for which not only the distribution of the drive torque between front and rear axle is possible but also the distribution between right and left drive wheel (torque vectoring), opens completely new possibilities for vehicle stabilization. In addition to steering and braking interventions, one-sided drive forces can now also influence the yaw moment of the vehicle. The necessary information signals (torque distribution between front and rear, friction utilization of the different wheels and the wheel speeds) are available from the ESP electronics. Thus, torque vectoring can be a subroutine of ESP.

4.2 EV (electric vehicle) and HEV (hybrid-electric vehicle) and electric brakes

The next level of complexity is achieved with hybrid and electric vehicles (EVs) and electric brakes. At commercial semitrucks, it is typical that in case of drive stability control, all disturbing systems (such as retarders, automatic transmissions, and exhaust brakes in engines) are switched inactive to have only strict brake control of the friction brakes. Now, from the development of the (not yet introduced) electric brakes, it is common knowledge that differences in torque caused by tolerances in spindles, gears, and electric motors are already noticeable with light braking. The car is slightly pulled sideward. Different cable lengths and two different energy reservoirs (X-split of the batteries) reinforce these symptoms as well as different temperature left/right.

If the manufacturing tolerances could be caught up with a final calibration, then there are also influences during driving such as temperatures (one-sided sun, batteries, cable, and motors) and uneven chassis and tire wear. The necessary corrections of offset and tolerances are done by long-term corrections (per ignition run and per brake procedure).

If two drive motors are installed on one axle, the same correction principle can and must also be used for drive torque and generator torque. The necessary prepared signals (wheel speed, yaw rate, lateral acceleration, longitudinal acceleration, steering wheel angle, and steering torque)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

are already available in the ESP as control unit for the longitudinal and lateral coordination, as well as for the necessary compensation algorithms. It is to be expected that the complete compensation of tolerances in drive and brake systems must be performed in the ESP.

4.3 Wheel hub electric motors

Ferdinand Porsche already used wheel hub motors more than 100 years ago in his Lohner vehicles. However, afterward, the drive concepts were characterized by a centralization of the torque generation. The idea of wheel hub motors did not return until EV and hybrid-electric vehicle (HEV). The unsuspended masses are significantly larger by adding the electric motors to the wheels. This is the same problem as with the increasing combustion engine power: larger wheels were the consequence. The 20-inch wheels mainly serve to accommodate larger and heavier brakes. This decreases the axle natural frequency and the vehicle body structure is more excited, which leads to decrease in comfort. At the same time, the wheel load fluctuations increase—which finally leads to semi-active or active chassis due to the better controllability.

The use of decentralized motors now allows a vehicle stabilization even using the drive. In doing so, targeted one-sided drive torques are created to move the vehicle in the direction the steering wheel angle sensor is indicating.

All technology steps listed above can each only improve individual compromises of the conventional chassis design. Their combination, however, cannot be used functionally independently of each other. Using the steering wheel, drivers specify the driving direction and with the accelerator pedal the vehicle speed is determined via the drive train. The necessary forces in the vehicle center of gravity (the mass is estimated first or derived from the air suspension pressures or wheel deflection suspension) deliver the measured lateral and longitudinal accelerations. From there, the necessary forces are calculated at the four-wheel contact points (all three force vectors). These can now (depending on vehicle equipment) be influenced by various systems. The used one is determined by the energy requirement of each individual function. The system with the least energy requirements is selected (potentially, the vehicle can be steered more simply with the wheel individual motors than with a steering intervention). In addition, those systems with the best efficiency are advantageous. A comfort calculation is superimposed to keep roll, yawing, and vertical movements within limits. As superior hierarchy element, the safety functions control the system behavior. This determines whether the maneuver can be driven this way, or whether a different focal point must be set, such as relations to other traffic participants (avoiding accidents).

5 CENTRAL COMPUTER COMPARED TO DISTRIBUTED INTELLIGENCE

First, electronics for single systems have been installed in the vehicles. The next step was to combine functions. A motor control unit can simply be expanded to the function of a speed control system by reading in additional sensors. The typical "butter soft" shifting of the automatic transmission today already requires a more complex interaction of the control algorithms of the engine and the automatic transmission.

In the start of the 1990s, some engineers tried to put a large central computer into the vehicle. That was also supposed to take on further control tasks, such as transmission control or the ABS.

The main arguments for suggesting this were that distributed intelligences waste a large part of the performance for communication. In addition, it was intended to reduce costs. A large processor is cheaper than two smaller ones. In addition, redundancies are necessary inevitably the central unit only needs one (e.g., hardware: housing, power supply, EEPROM and sensor inputs; and in the software: operating system and diagnostics). While in distributed systems. Theoretically, synergies take effect with the summary of several functionalities.

It was also already tried to implement the idea of the central control unit by summarizing comfort control units from several suppliers. Inconsistencies regarding the resource administration, the product responsibility, the protection of each supplier's knowledge, the cost and coordinate of change management during series production, and the allocation of errors in the warrantee processing brought the specially founded company to a quick end.

The reality went a different direction. In any case, increased functionality means more sensors and actuators, thus, more cables and a higher complexity. If the *one*, central control unit is in a central site, all cables must be routed there. The electric mirror adjustment might be an example: the button is in the left door and the motors are in the left and in the right door. With one central computer, this positioning resulted in many meters of cable to the central control unit and back again. Cables are weight, space, fault sources, and not least costs.

An additional argument against the centralization is the complexity. Every control task demands that the associated sensors are scanned in a fixed pattern and/or processed at very specific times. The sensor requirements of an ABS or an engine controller are significantly different. If sporadic tasks are added, such as controlling an exterior mirror, this places considerable demands on the processor; the corresponding processor performance must be available for the potential concurrency. The controllability of so

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

many layers of requirements is much more complex in one program. The standard PC operating system is certainly an example for such complex programs. The complexity is not controlled so well, that everything is always available at critical moments.

Like with any software, in the chassis systems, there remains a residual error probability of approximately 2500 ppm at the start of series production (aviation is at approximately 250 ppm). This means in a million rows of C-Code, there are 2500 errors that are equally distributed to light, average and critical errors. Half of the errors are caused by incorrect specification, thus, surprising nonfunctionality or failures that were not planned. The probability of errors is approximately 10 times higher than the probability of hardware failures.

To be able to control the entire complexity, to not have to route all cables to one site, to switch off systems in case of error, to compensate for malfunctions of individual systems, every task requires its own control unit. The knowledge of the supplier and its protection lies multiplied in the software. The allocation of legal and financial responsibility is easily possible with single systems.

Finally, it should be noted that the current specifications of the automobile manufacturer already now exceed the capabilities of the microprocessors coming on the market in 2 years. This is true with regard to the memory capacity as well as the loop-time. All demanded algorithms can no longer be processed by the time slot specified by the driving dynamics. The electronic engineers like to compare software to the ideal gas: the software immediately fills all available storage and consumes every available calculation time, without generating noticeable advantages to the end customer.

This makes all safety consideration unanswerable when central computer systems will be in the vehicle. In the future, there will not be any "computer center" in the car.

6 COMMUNICATION OF DISTRIBUTED INTELLIGENCES

To avoid the disadvantage of multiple sensors or actuators with distributed intelligences, data must be exchanged. Let us take the example of cruise control: The engine control unit has obtained the speed signal from the speedometer. The ABS constantly determines the wheel speeds. What could be easier than for the cruise control to use the ABS speed signal? Logically, the speedometer signal should also be generated from the ABS speeds. Doing this generates the advantages for the automotive manufacturer that the transmission pickup can be omitted and the costs could be reduced.

To use data multiple times, a means for transport is required. There should be a communication system that is inexpensive and secure. The vehicle manufacturers and the large suppliers had already recognized the necessity of such a system at an early stage and began doing research in the 1970s.

6.1 CAN bus

In the early 1980s, Bosch began testing the bus systems on hand to see whether they were suitable for vehicles. After none of the existing systems satisfied the requirements of Bosch, they began specifying their own system in 1983. Daimler soon joined and Intel was gained as a hardware supplier. In early 1986, CAN was born.

CAN is a serial multi-master system. Every device connected to the bus determines itself when and how many messages are placed on the bus. The Ethernet network in the office networking is also one such serial multi-master system. However, only one message can be transferred to the bus at a time using the serial system. With the Ethernet, everything is regulated by "trial and error". If two messages collide with each other, both senders must try again using a specified schematic. With a higher busload—many participants try to transmit something at the same time—it can happen that no messages can be transmitted.

CAN is avoiding this with a nondestructive prioritized arbitration. In addition, a series of fault-detection mechanisms are integrated that ensure that transmission faults can be detected securely and generally can be avoided because of repetition of the message.

To simplify the handling, large parts of this standard were realized in hardware. Today, the processors used in the automotive industry are available with up to three integrated CAN interfaces. Every bus participant must be able to transmit its messages at the right time, complete and error-free.

At the start of networking, this was still simple: if the automatic transmission and the engine control unit communicate in 20-ms rhythm, there is never any danger that a message would be lost.

Later, the ABS was added to enable starting assistance together with the engine control unit and to be able to realize ESP functions with additional control units. However, these safety functions are only a small part of the control units in the modern vehicle. Mirror adjustment, window lifter, and electric seat adjuster are raising the need for communication exceedingly. The amount of information is limited by the bus width and the cycle speed of the bus. With a serial bus, the calculation is simple. For 500-kBit/s

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

bus frequency (the standard frequency for high speed CAN) and 20-ms cycle time for the critical signals, the following results:

$$500\ \text{kBit/s} \times 0.02\ \text{s} = 10,000\ \text{Bit}$$

This information rate could be transferred. Those are approximately 76 messages with each 130 bit message length (8 byte user data + overhead)

6.1.1 Structuring

The messages that must be transferred to the bus can be subdivided into important, less important, urgent, and less urgent messages. The messages of safety relevant systems are naturally extremely important, most also urgent. However, non-safety relevant controls can also require a high data transfer speed. The critical systems have an uninterrupted need to communicate—there are closed control loops that expect and send data in fixed time intervals—whose information environment is deterministic. Lost messages are not acceptable.

The vehicle manufacturers first routed two busses in the vehicle, so that critical messages are separated from the less critical messages:

- A high speed bus for the critical messages that are characterized by all messages are routed in cycles on the bus and their busload never exceeds 60% (e.g., drive train bus).
- A low speed bus (comfort bus) that must handle the rest of the communication needs.

To administer the comfort bus, generally, a network management is used to ensure that every participant receives its messages in acceptable times on the bus. Here, the bandwidth of the busses in critical conditions (all simultaneous messages are to be released) is not enough.

The bandwidth is a cost factor. By executing the comfort bus as a low speed bus, the system costs are reduced. Today, in Europe, the CAN is the most widely used bus system in vehicles. However, there were also alternatives, as, for example, VAN in France or in the USA Standard J1850 for Class B Network.

The CAN has a constant cost share with its safety, realized in hardware. If the requirements for cost and safety are lower, the local interconnect network (LIN) bus will be in favor, which halves the costs for a bus node.

6.1.2 Wireless

For networking the vehicle with the environment, for example, for cell phone, personal digital assistant (PDA) connection, and so on, bluetooth (e.g.) is already being used. This type of data transfer is unsuitable for safety systems, because of how it is influenced by electric and magnetic fields.

6.1.3 New requirements

In applications that are currently still covered by CAN, there are new requirements coming up. Driving stability systems that connect active suspension systems with electric steering and with ESP, engine control and adaptive cruise control, to be able to perform active corrective maneuvers, require closed control circuits that cannot be managed by an event-controlled CAN.

The main problem is that the time allocation of the messages cannot be guaranteed due to an accumulation of events or faults. (Highly prioritized messages receive priority and defective messages are repeated.) The variety of participating control units bursts the bandwidth that is available. High transmission rates from up to 10 Mbit/s are required.

The CAN is no longer able to cover the safety requirements set by the X-by-Wire functions (brake-by-wire, steer-by-wire). In some systems, there will be no mechanical fallbacks any more. In the currently used EHB systems (electric-hydraulic brake, the brand name at Daimler is SBC), some brake pressure can still be generated in emergencies, using the pedal. However, it is obvious that the development has reached a limit, which now requires new solutions to expand the technology.

6.2 Flexray

The bus system that is step by step being implemented in the automobile is called *Flexray*. It is a communication system as it has a fixed communication protocol and fixed hardware. It permits a deterministic, collision-free data transfer with guaranteed latency and jitter period with a scalable synchronous and asynchronous data transfer. The latency is the time between an event and the (delayed) reaction following. Jitter is the first derivative of the delay; the back and forth and the cycle frequencies during transfer of digital signals are designated by it (Figure 15).

Minimal latency and jitter times save signal filter times at the recipient. The support of redundant transfer channels, as well as the error tolerance and time-controlled actions implemented in the hardware, permit a quick fault detection and short reaction times.

Instead of arbitration such as with CAN, the Flexray has a fault tolerant synchronous time basis as global time basis of all systems in the vehicle. Owing to the support of Star Technology, partial switch-offs can be realized.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

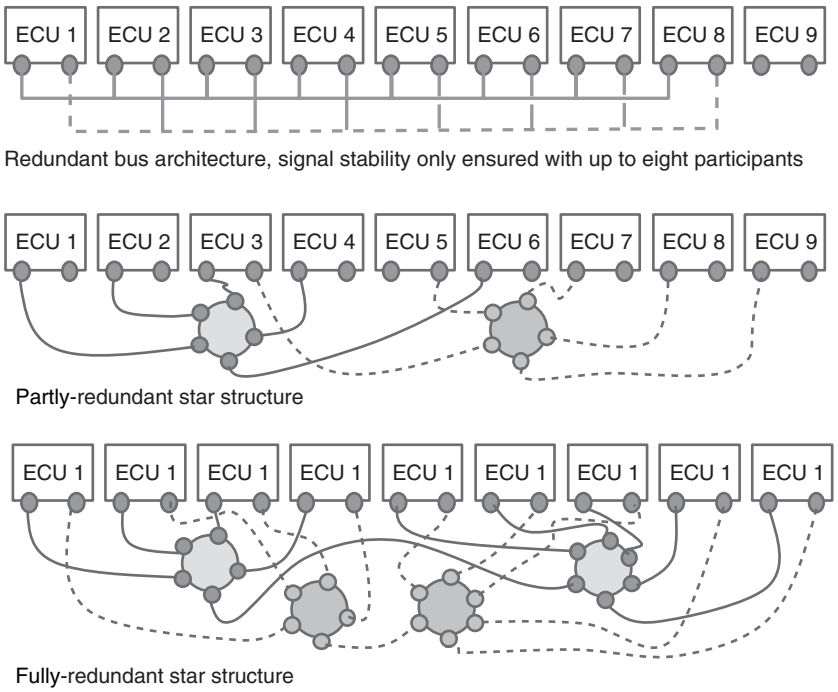


Figure 15. Flexray networking structures. (Reproduced by permission of R. Leiter.)

The Flexray can offer approximately 20 times the bandwidth of the CAN (until 10 Mbit/s) and it permits short cycle times (starting with approximately 1 ms). The consequences are doubling the costs for the bus nodes and the necessity that these bus systems must already be designed during design of the E/E architecture of the vehicle, including all signals and transfer time points. The automobile manufacturer is the only one who can do this. He must rebuild the necessary expertise (including the detailed knowledge of the individual system). Subsequent changes (e.g., adding new bus participants) are extraordinarily costly and time-consuming.

6.3 Gateway

The accumulation of control units forces structuring. As described above, as no bus fulfills all requirements, the bus systems are fit into the vehicle corresponding to the tasks. This means that a vehicle today can easily have five different busses installed; thus, the vehicle has an IT—infrastructure that is not inferior to that of the average production operation.

To make information that is required by many control units available to all systems, for example, the setting of the ignition key or the vehicle speed, that, for example, switches away the television receiver during driving, the information must be transferred to the various bus systems. Gateways take on this task.

6.4 Networking

Without the network, many of the currently available functions, such as vehicle stability control or adaptive cruise control, would not be possible. The new developments build on progressing networking in the vehicle. The comfort for the driver is increased; active safety and driver support systems can be realized with it.

6.5 Hierarchies in the control systems

The limited options of individual control units require cooperation that must be regulated. This will lead to a grouping in function areas—exactly like the supplier and automotive manufacturer are set up in their departments.

Thus, in the lower level of the driver's domain, there are the longitudinal and lateral dynamics, the vertical dynamics, and the drive and energy recovery. In the comfort domain, we have the air-conditioning/heating and the operating forces; and in the communication domain, the geophysical position, the site-related information, the information from other vehicles and local hosts, the driver's intention detection, driver feedback, and information for or about the driver (e.g., recording health status) (Figure 16).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

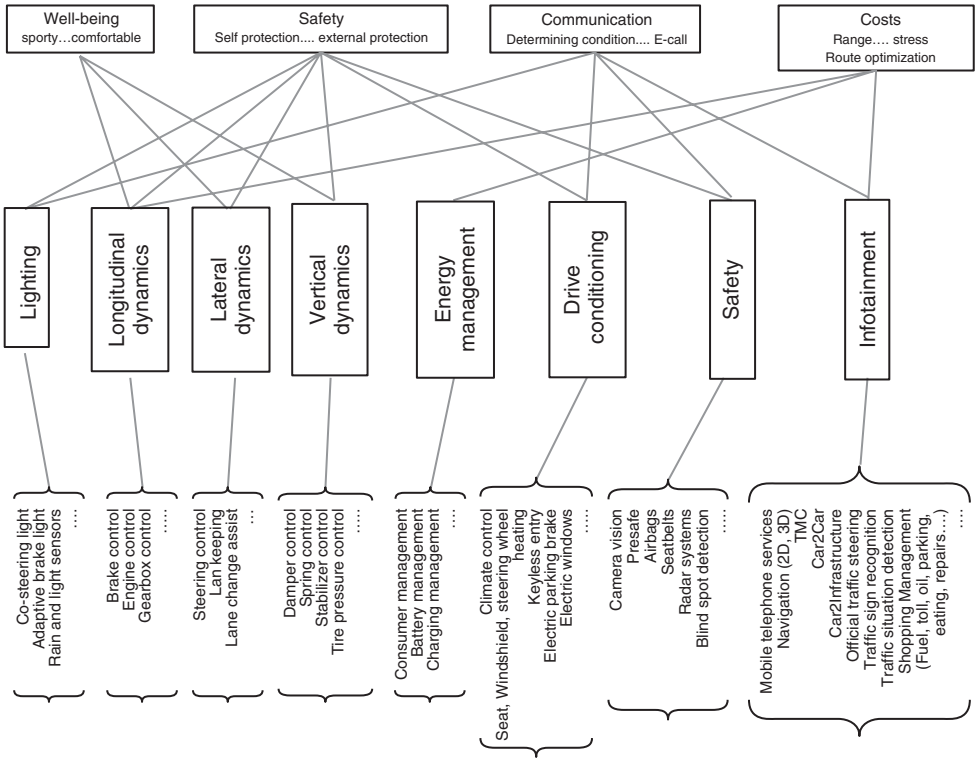


Figure 16. Domain structure. (Reproduced by permission of R. Leiter.)

Each domain has its own controller in which the tasks are joined. In addition, the vehicle manufacturer can place the characterizing master computer that sets the standard specialties specific to the vehicle, and in the case of fault, provides alternative strategies.

6.6 Shortened development times, tight resources

The new technology cannot be introduced without secondary effects. The classic development systems were tested in the vehicles long before the start of series production. Then, high quality and safety standards of the automotive manufacturer were brought into series.

Currently, the manufacturers' desire to innovate is keen and fast. Not just *one* new system is integrated in a new vehicle.

The shortened times to market introduction now mercilessly show their negative sides. All these new systems are specified early on, but naturally due to overall short resources, are only finished at the time when the final integration tests are supposed to be performed. Problems that no one thought of do not appear until then.

An example: the electric parking brake contains a logic that does not operate the brake in dynamic condition when the accelerator pedal is pressed while driving (preventing the driving against the brake). On the CAN, the electric parking brake algorithm only shows the respective brake condition once it has been reached. The cruise control should be abandoned by operating the parking brake (dynamic delay). As the cruise control pays attention to a brake-information, that was not visible at the start of the action, it tries to counteract this force at the start of braking and increases the engine power. The parking brake sees the gas pedal information that the cruise control uses to increase the motor power and it does not execute the command to decelerate. Consequence: the networking function does not work.

The reasons for this problem have been small misunderstandings of the specifications. Since such things do not show up until a time when the actual finished system should run endurance tests, the test times cannot be adhered to, with known and unknown consequences.

6.7 Integration and simulation

The way out from the dilemma is simulation and integration tests on simulation basis. To do this, HIL (hardware in the loop) and SIL (software in the loop) systems are combined. Available control units are connected to vehicle simulation (HIL). In these vehicle simulations, the high

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

level algorithms of the new systems are incorporated as software simulations, so that the overall vehicle can be tested long before the first prototype.

These techniques are available, yet not developed so widely that they could also detect the last problems. Unfortunately, the work never reflects the current development status of the communication participants, but rather the specification that always left interpretation gaps. The human language is not able to sufficiently define the IT: there are communication difficulties between the vehicle dynamics specialists and the "software workers" that are to implement the ideas of the vehicle dynamics specialists.

6.8 Black box and the operating system

With the high degree of networking, it is no longer possible that the automotive manufacturers purchase black boxes from their system suppliers and also shift the responsibility on them.

To be able to successfully integrate all systems, the vehicle manufacturer must have intimate knowledge of the subsystems and provide sufficient engineering resources. The vehicle manufacturer becomes the system-integrator, even if several functions are calculated in one control until. He can only accept the responsibilities connected with this, if he knows the processes in the control unit sufficiently well. The "game rules" must be followed. Accepting the prescriptions of operating systems such as AUTOSAR can do this. Industry started with OSEK, to define software standards. Currently, AUTOSAR 4.0 is used. If the suppliers used to have their own operating systems (and the responsibility for them), today they have to follow the strategy of the automotive manufacturers. The search for errors is very difficult in such complex systems, where due to networking, an overlapping of influence fields happens. Thus, there is also a possibility for faults.

6.9 Risk factor driver

The variety of system functions is supposed to support the driver and increase both his comfort and his safety. However, every automotive manufacturer has its own operating philosophy, also to keep its brand identity.

Most drivers do not take the time to study the operating instructions, and they expect self-explanatory system functions. The same system can function differently with a different manufacturer; even within a fleet of a manufacturer, there could be function differences due to different generations of systems (e.g., autonomous cruise control systems of generation 1 functioned between 30 and 180 km/h, the same technology in generation 2 now works until standstill).

There are many opportunities for faults by the driver, when a foreign vehicle must be controlled.

Once the driver is accustomed to the design of a safety function, he must be able to count on it. If, in a critical driving situation, partial systems fail, that can be critical. The system may overextend most drivers.

7 SUMMARY

The complexity of chassis control in cars is exploding.

Networking is one of the basic technologies of the innovation explosion.

The diagnostic systems and the diagnostics are not keeping up.

REFERENCES

Abe, M., Ogura, M., Sato, T. (1988) Mechanism for Steering Front and Rear Wheels of Four-Wheel Vehicle. Patent: US 4792007 A.

Amsel, C. (2003) Schutzkonzepte und Designregeln für den Kurzschluss 42 V/14 V im dualen Bordnetz. Dissertation RWTH Aachen, Schriftenreihe Automobiltechnik.

Debes, M., Herb, E., Müller, R., *et al*. (1997) Dynamische Stabilitäts Control *DSC* der Baureihe 7 von *BMW*. *Automobiltechnische Zeitschrift 99*, **3**, 4.

Donges, E. (1989) Funktion und Sicherheitskonzept der Aktiven Hinterachskinematik von BMW. Tagung "Allradlenkung bei Personenwagen" Haus der Technik, Essen.

Jurr, R., Behnsen, S., Bruns, H., *et al*. (2001) *Das aktive Wankstabilisierungssystem Dynamic Drive*, Springer Verlag, Wiesbaden.

Köhn, P., Wachinger, M., Fleck, R., *et al*. (2003) *Aufbau und Funktion der Aktivlenkung von BMW*, Tagung Fahrwerktechnik des Haus der Technik, München, Juni.

Konik, D., Bartz, R., Bärnthold, F. *et al*. (2000) Dynamic Drive – Das neue aktive Wankstabilisierungssystem der BMW Group; 9. Aachener Kolloquium Fahrzeug und Motorentechnik.

Kubota, Y. (2013) Nissan to install electronic "steer-by-wire" in Infiniti cars, http://www.reuters.com/article/2012/10/17/us-nissan-technology-idUSBRE89G03A20121017 (accessed 4 September 2013)

Kuroki, J. and Irie, N. (1991) HICAS: Nissan Vierradlenkungsechnologie, Fortschritte der Fahrzeugtechnik, Band Nr. 7 in *Allradlenksysteme bei Personenwagen* (ed. H. Wallentowitz) (Hrsg), Vieweg Verlag, Wiesbaden.

Lloyd, J.M. (2010) Cross-linked variable piston air suspension. US Patent: US 2010/0230912 A1; September 6.

Wallentowitz, H. (2007) Lecture Vehicle Dynamics 2. "Vertical and lateral dynamics of vehicles" Aachen.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Wallentowitz, H. and Amsel, C. (eds) (2003) *42 V – Power Nets*, Springer.

Wallentowitz, H., Donges, E., and Wimberger, J. (1994) Die Aktive Hinterachskinematik (AHK) des BMW 850 Ci, 850 Csi. *Automobiltechnische Zeitschrift*, Springer Verlag, Wiesbaden, **96**(11), 674–689.

Wiesinger, J. : Active Body Control (ABC), http://www.kfztech.de/kfztechnik/fahrwerk/federung/abc_aktive_body_control.htm; (accessed 14 April 2013)

van Zanten, A. (2000) Bosch ESP systems: 5 years of experience. SAE Technical Paper 2000-01-1633.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto032
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

System Simulation in DSH*plus*, What Applications are Possible?

Christian von Grabe[1], **Olivier Reinertz**[1], **René von Dombrowski**[2], **and Hubertus Murrenhoff**[1]

[1]*RWTH Aachen University, Aachen, Germany*
[2]*FLUIDON Gesellschaft für Fluidtechnik mbH, Aachen, Germany*

1 Introduction 1
2 System Simulation 1
3 Simulation Examples 4
4 Summary 17
Related Articles 17
References 17

1 INTRODUCTION

The rising demands on technical systems in automotive and general applications in mechanical engineering result in products with an elevated degree of complexity and a wide range of integrated functionalities. To fulfill the requirements regarding system performance and reliability, modern technical systems must combine subsystems of different technical disciplines such as pneumatics, hydraulics, electronics, and mechanics as well as informatics and control techniques.

These very different technical subsystems must interact precisely in order to realize the required functionality of the entire system. In contrast to developments in the past, the engineers involved in the development process of such systems can no longer rely on estimated formulas or simple static design techniques. The use of simulation techniques is therefore indispensable in today's development processes. Furthermore, the used system simulation tools are required to provide functionalities for the coupling between different technical disciplines to coordinate the interactions between the implemented subsystems.

To meet the requirements, several software tools such as DSH*plus*, AMEsim, and SimulationX are available to aid the engineer during the whole development process.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 SYSTEM SIMULATION

2.1 Simulation procedure

In the following, the general modeling procedure and the basic calculation methods of one-dimensional simulation tools will be illustrated, using the example of DSH*plus*. System simulation tools provide a detailed description of the real system, which is to be investigated, using sophisticated models of the component and transfer properties in the form of mathematical equations.

To ease the creation and compilation of those equations, most simulation suites provide a graphical user interface, which allows the synthesis of the system from single functional component units. These are available in libraries that contain hydraulic and pneumatic as well as electrical, control engineering, and mechanical elements. The functional units usually constitute a technical component such as a valve or a cylinder, but can as well be used to synthesize complex technical systems. Behind every functional unit,

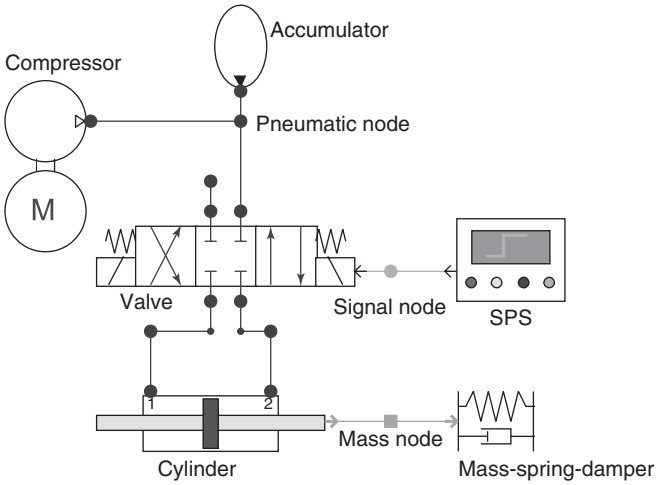


Figure 1. Simple pneumatic system.

such as a high pressure injection pump or a piston pump, for example, lies a mathematical description of the component. The components are thereby available with different levels of detail according to the needs of the simulation focus. To remain on the example of pumps, a pump, for example, can be modeled by just providing a constant volume flow to the system or in very detailed way considering all geometric circumstances and physical effects. In the second case, the volume flow of the pump will show all volume flow pulsations resulting from the number of pistons, for example. This is not necessarily needful for the simulation of entire systems but it might be of great importance for the investigation of noise problems.

If the provided libraries are not adequate to model the desired component functionalities, most simulation environment offer the possibility to create user-specific function units in a common computer language. In case of DSH*plus*, the programming language *C++* is utilized to create components from scratch. This enables the user to implement even the most complex physical models, as long as they are mathematically conveyable and the computation time is economically acceptable. The downside to the use of *C++* is the requirement for the user to be familiar with at least basic programming techniques.

In the further modeling process, different functional units can be connected by the user via nodes in the graphical user interface depending on their physical interaction with other functional units. A simple simulation model of a pneumatic cylinder drive consisting of different functional units and component connections is shown in Figure 1.

On the basis of the connections of the functional units, the mathematical equations of the physical system are automatically generated by the simulation tool. The dynamic behavior of the technical system is represented in a differential equation of the *n*th order or *n* differential equations of first order. The notation in *n*-differential equations of first order is far more suitable for the numerical integration because numerical solution methods utilize the same form of equation.

With a given set of parameters for the system components and starting values for the state variables, the system of equations can be treated as an initial value problem of common differential equations (FLUIDON, 2007):

$$\begin{aligned} \underline{\dot{y}} &= \underline{f}(t, \underline{y}) \\ \underline{y}(t_0) &= \underline{y_0} \end{aligned} \tag{1}$$

with

$$\underline{y} = \begin{pmatrix} y_t\,(t) \\ y_2(t) \\ \dots \\ y_n(t) \end{pmatrix} \tag{2}$$

and

$$\underline{f} = \begin{pmatrix} f_t\,(t, y_1, y_2, \dots, y_n) \\ f_2(t, y_1, y_2, \dots, y_n) \\ \dots \\ f_n(t, y_1, y_2, \dots, y_n) \end{pmatrix} \tag{3}$$

Solving the set of differential equations and integrating the state variables over time provides values to the state variables at a later time step.

2.2 Calculation of pneumatic and hydraulic values

The node-oriented model structure and a concentrated calculation of the state variables within the connecting nodes are the basic principles of all system simulation tools, which therefore are also called *lumped parameters simulation tools*. Figure 2 visualizes the calculation core of such lumped parameter approaches for the example of a pneumatic system. Within the nodes, the pressure derivatives are calculated from the balances of the entering and leaving mass flows. These pressure derivatives are then integrated to pressures, which are transferred to the components. Mass flows, however, are calculated in the components from the difference of the pressures applied to their connections. To define the characteristics of the pressure buildup, for example, in a pneumatic connection node, the starting values for the volume, temperature, and pressure must be specified by the user. In case of volume changing components, as for example cylinders, the volume and its derivative are transferred to the nodes to allow their consideration in the pressure buildup equation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

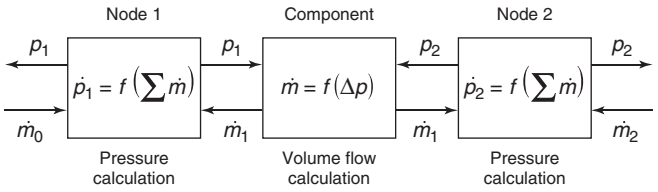


Figure 2. Calculation method of simulations with lumped parameters.

The correlation of volume flow and mass flow is described by the density ρ, which can be calculated by ideal gas equation (Equation 4) using the temperature T, the pressure p, and the gas constant R.

$$\rho = \frac{p}{T \cdot R} \tag{4}$$

The change in pressure of a pneumatic system is caused by mass flows, heat transfer, and volume change. The pressure change is described by the first law of thermodynamics for open systems and results in Equation 5:

$$\dot{p} = \dot{p}_{\dot{m}} + \dot{p}_{\text{th}} + \dot{p}_{\dot{V}} \tag{5}$$

The energy exchange resulting from the mass flows in between the components and the nodes is considered in Equation 6 by the variable $\dot{p}_{\dot{m}}$:

$$\dot{p}_{\dot{m}} = \frac{\kappa}{V} \cdot R \cdot \left(\sum \left(\dot{m}_{\text{in}} \cdot T_{\text{in}}\right) + \sum \left(\dot{m}_{\text{out}} \cdot T_{\text{out}}\right)\right) \tag{6}$$

Furthermore, the heat flow from and to the environment is represented by $\dot{p}_{\text{th}}$ in Equation 7. κ is the polytropic exponent. For air, this can be an adiabatic process and then κ would be near 1.4. In an isothermic process, the κ is 1.0. The heat transmission coefficient α describes the ratio of heat flux per unit area to the temperature difference from component wall with the contact area A to the air inside the balanced system

$$\dot{p}_{\text{th}} = \frac{\kappa - 1}{V} \cdot [\alpha \cdot (T_W - T) \cdot A] \tag{7}$$

The performed work due to the volume change at the balanced system is considered by $\dot{p}_{\dot{V}}$ in Equation 8:

$$\dot{p}_{\dot{V}} = -\frac{\kappa}{V} \cdot p \cdot \dot{V} \tag{8}$$

The temperature is acquired by means of the ideal gas equation as shown in Equation 9:

$$T = \frac{p \cdot V}{m \cdot R} \tag{9}$$

The time derivative of the ideal gas equation delivers the change in temperature over time:

$$\dot{T} = \frac{\dot{p}}{p} \cdot T + \frac{\dot{V}}{V} \cdot T - \dot{m} \cdot \frac{RT^2}{pV} \tag{10}$$

To minimize the computation time of the simulation, the pressure change $\dot{p}$ is calculated first. In a second step, the preceding results are used to determine the temperature change $\dot{T}$.

2.3 Numerical integration

The system of differential equations representing the modeled system describes the derivative of all state variables in all operating points. To solve the system of differential equations of the entire simulation model, DSH*plus* provides different numeric integrators, which allow an explicit as well as an implicit solving on a time-discrete basis.

The explicit solver only uses values, which are already calculated and therefore available from former time steps. The implicit solver in contrast uses the solution of the current time step as well.

The integral interval $[t_0, t_n]$ is divided into time increments.

$$t_0 < t_1 < t_2 < \cdots < t_n$$

with the local time increment of $h_i = t_{i+1} - t_i$ for $i = 0 \ldots (n-1)$. An approximation for $y(t_i)$ is obtained by Equation 11 with $i = 0 \ldots (n-1)$.

$$\underline{y}(t_{i+1}) = \underline{y}(t_i) + \int_{t_i}^{t_{i+1}} f(t, y(t))\mathrm{d}t \tag{11}$$

Furthermore, modern system simulation tools offer solvers, which work with an automatic step size control instead of a constant step size integration. The advantage of this multistep methods is a small time increment in high dynamic operation modes and a bigger time step, when no or only small changes in the model behavior are present. Multistep methods therefore reduce the computation time. As a consequence of the numerical integration, a discretization error has to be encountered for. To reduce this error, DHS*plus* offers an automatic step size control in order to achieve optimal results for highly nonlinear very stiff systems. The step size control adapts the step size in such manner that a specific error tolerance is met.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.4 Typical applications

Typical applications of system simulation tools such as DSH*plus* within the automotive engineering are all fluid-containing systems, such as functional systems, safety systems, or comfort systems. As an example, comfort systems such as air springs or semiactive dampers are virtually analyzed and optimized with the help of system simulation tools. Further, common simulation applications are power steering or braking units as well as safety systems such as ABS (antilock brake system), ESP (electronic stability program), or active roll systems with their related control strategies.

Beneath simulation applications on the entire system level, it is also common practice to look at the single component as a system. Applications on the component level are fuel injection pumps, hydraulic or pneumatic lines, fuel injectors, hydro bearings, or torque converters, for example.

3 SIMULATION EXAMPLES

To illustrate the wide range of possible applications for DSH*plus*, a rather unusual example of a modeling approach for pneumatic seat components is described. The example shows the whole process including data acquisition for component parameterization, the buildup of the simulation model in DSH*plus*, and the comparison of the simulation results with measurements. The example not only reveals the advantages of simulation tools in the design process but also displays the flexibility of modern system simulation tools.

3.1 Example 1: pneumatic seat adjustment

Pneumatically actuated comfort systems for the dynamic adjustment of the shape of automotive seats are characterized by their compactness and consist of miniature pneumatic components. These components are a compressor, a number of valves—depending on the application—inflatable plastic cushions, as well as the necessary connectors and hoses. Figure 3 depicts the functions and the pneumatic components built into an automotive seat.

Commonly, the components mentioned earlier will be delivered to the seat manufacturer by different suppliers. Therefore, a fast, exact, and energy-efficient method is necessary to compare components of different producers or to characterize or assess single components regarding their usability for the automotive seat. The pneumatic elements are low priced components according to their usage in the automotive sector, which is well known as a *mass market*. Their functionality and fabrication result in a large variation of their pneumatic characteristics. This variation of the components' characteristics leads to a strong variation of the system behavior. Therefore, an integral approach for the characterization and simulation to determine the system behavior is of great importance. DSH*plus* is utilized to evaluate the applicability of the components regarding the specified behavior of the complete system. As pneumatic cushions are quite rare system components in typical pneumatic applications, there is no library containing similar components available in DSH*plus*. Furthermore, because of the early development stage of the components, there are no specific datasheets available to parameterize the components. Hence, the parameterization of such a model by estimates would result in an inaccurate description of the system behavior. To ensure accuracy of the model,

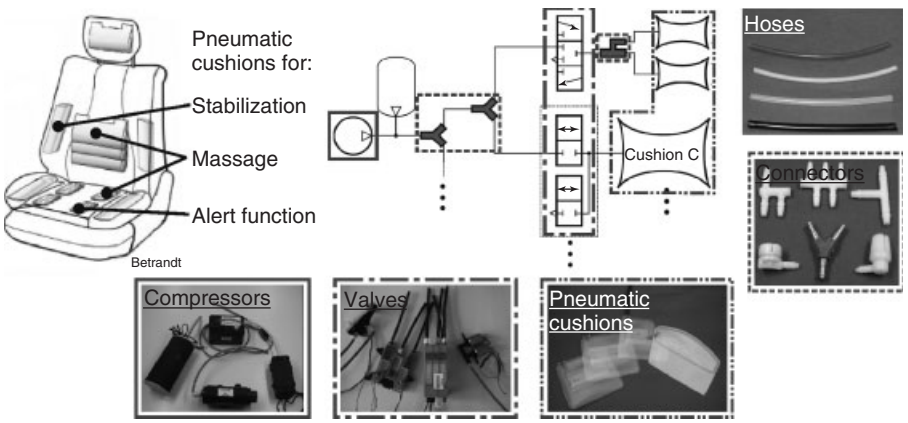


Figure 3. Pneumatically actuated comfort system of an automotive seat with its functions and its builtin components.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

components are calibrated and the results are used to reproduce the exact component behavior in the system simulation in DSH*plus*.

3.1.1 Valves

Figures 4 and 5 depict the mass flow characteristics derived from measurements and the mentioned parameters of an exemplary valve for the flow path $1 \rightarrow 2$ and $2 \rightarrow 3$, respectively.

The result of one gauging procedure is the mass flow characteristic over pressure p_1. Regarding the flow through the valve, p_1 represents the static pressure in front of the valve (input pressure). A set of characteristics results from the variation of the pressure p_2 downstream of the valve. According to the interesting pressure range from p_{Amb} to $1.5\,\mathrm{bar_{Abs}}$, p_2 was varied in 0.1 bar steps. The set is completed by changing the flow direction from "normal" flow to "reverse" flow.

In a second step, the outlet metering edge was measured (Figure 5). In addition, the switching delay t_{SD} was determined (Siebertz *et al.*, 2010). This is an important parameter, which characterizes the valve's opening dynamics. A second parameter, the pressure loss coefficient $K_{\mathrm{V},\Delta p}$, describes the valve's leakage behavior. Both parameters are implemented in the simulation model described in the following sections.

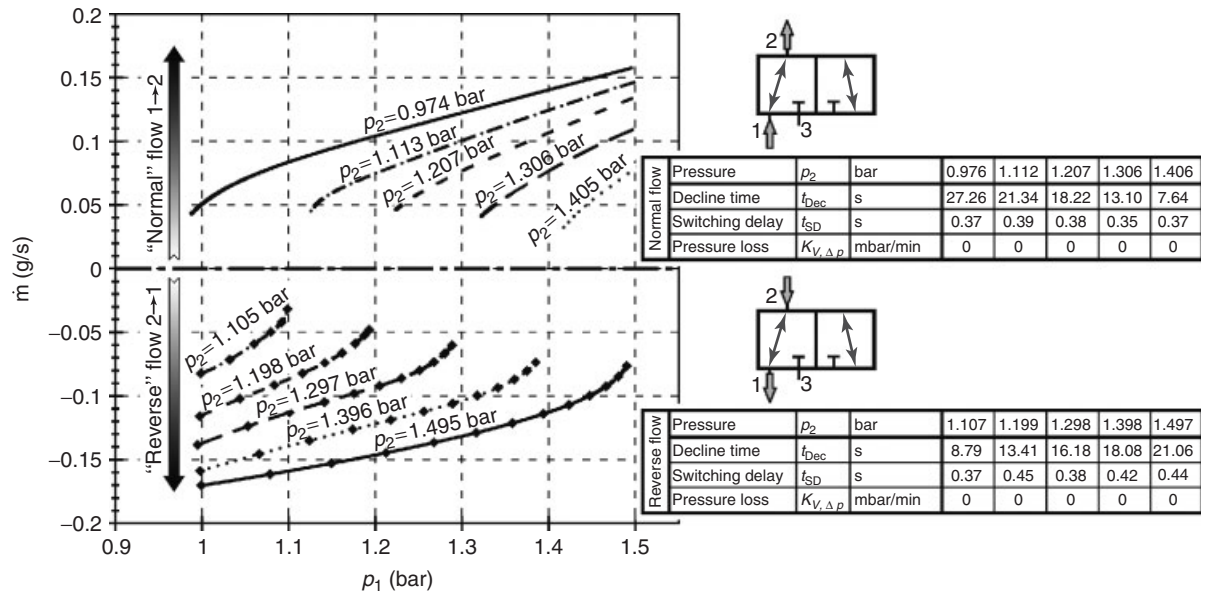


Figure 4. Characteristics and parameters of the flow path $1 \rightarrow 2$ of an exemplary valve.

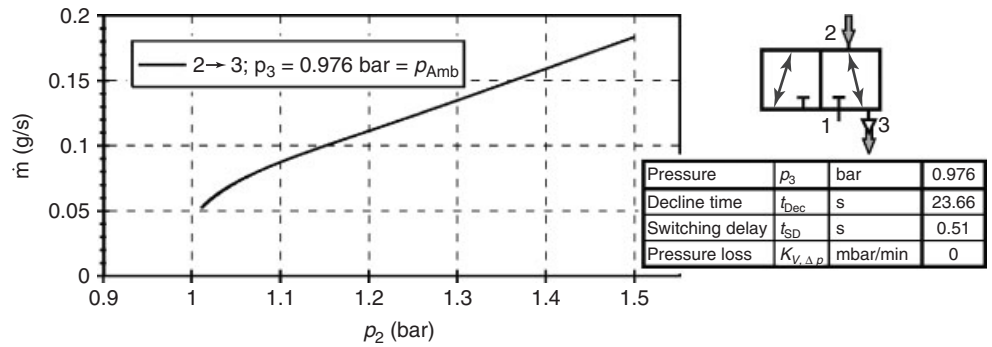


Figure 5. Characteristics and parameters of the flow path $2 \rightarrow 3$ of an exemplary valve.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.1.2 Compressor

The result of the compressor gauging is its conveying characteristics and a few other parameters, as depicted in Figure 6. The characteristics over the relative pressure difference of two exemplary compressors are shown. According to the compressor's performance, the pressurization of the measurement volume is only possible up to the compressor's maximum pressure $\Delta p_{\mathrm{Max}} = p_{\mathrm{Max}} - p_{\mathrm{Amb}}$. The leakage rate $K_{\mathrm{C},\Delta p}$ is determined in order to specify the compressor's quality. Driving the compressor's motor with 12 V, the power consumption of the compressor is quantified by the averaged current value I_{Avg}. While the compressor's characteristic is used for simulation, the other parameters can be used as quality attributes of the compressor, too.

3.1.3 Cushions

The volume as a function of the internal pressure characterizes the seat cushions' pneumatic behavior. It is determined by scaling the cushion while filling it with water, as water can be deemed to be incompressible in the investigated pressure range and its specific weight is well known. To allow the measurement of pressure-dependent characteristics, the water reservoir, connected to the cushion, is pressurized by an entering airflow, whereas the internal pressure of the seat cushion is determined by a pressure sensor at its inlet. Figure 7 shows the behavior of three different cushions, measured without applying a load. It becomes obvious that—after a filling period with minimal pressure rise—a linear relationship of pressure and volume exists. The results show steady-state volumes, whereas the strain of the cushions' plastics leads to a time dependence, comparable to a first-order lag element. The knowledge of the system behavior is necessary to enable a correct implementation of the cushions in DSH*plus*.

3.1.4 Measured data processing

The behavior of the measured compressors and valves has to be transferred into characteristic fields for their use in the one-dimensional simulation. These characteristic fields have—in the case of valves—two input parameters (the inlet and outlet pressures p_1 and p_2) and one output (the mass flow $\dot{m}$). The simulation software DSH*plus* requires tables, providing an output value for each possible combination of input sampling points (FLUIDON, 2007). This

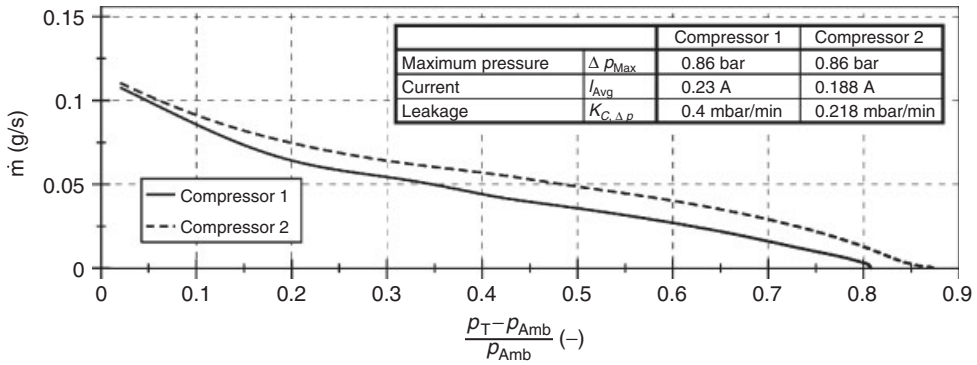


Figure 6. Characteristics and parameters of two exemplary compressors.

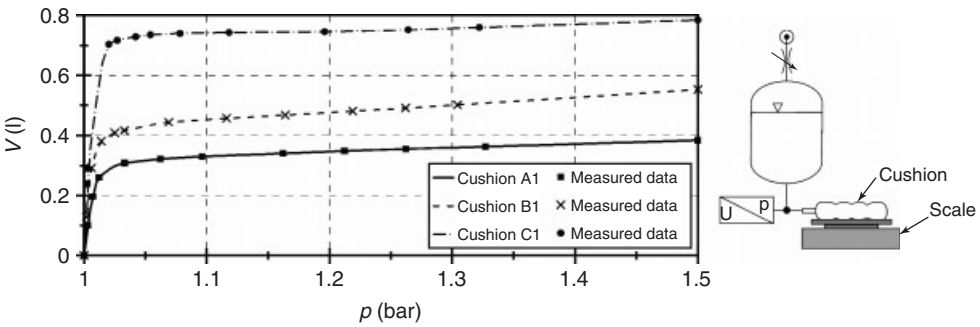


Figure 7. Cushion volume as a function of its internal pressure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

means, in case of the above-mentioned gauging method, that the sampling points of the inlet pressure have to be identical for each measurement, whereas the outlet pressure is varied. The measured inlet pressure is influenced by the mass flow across the measured object. Thus, at different points, the measurements will vary. Common sampling points have to be generated by a spline interpolation algorithm, interpolating from the measured points at defined pressure levels. Here, the "radar-spline" method can be applied to generate additional curves between the measured ones and thus can be used to avoid the above-stated problems, occurring during linear and common spline interpolations. This method considers zero mass flow in case of pressure equality, represented by an additional diagonal in the characteristic diagram. Its endpoints serve as center points for the radar-splines, defined by one center point and crossing points with the present curves. Finally, additional curves between the measured ones can be achieved by interpolation of sampling points generated by radar-splines. In this way, a characteristic diagram with a more continuous behavior (Figure 8) is generated. The level of discontinuity and the difference from the above-described physical law decrease with a higher number of data points in the characteristic map. As the number of data points to be generated with the radar-spline method is unlimited, the model behavior can be optimized by a compromise between model quality and data volume. The data import into DSH*plus* is realized either by a provided characteristic map creator for two-dimensional characteristic maps (one input parameter results in one output parameter) or by an ASCII-text file. The wide compatibility of almost any data acquisition software to ASCII-text files allows the import of measured data from almost any source to DSH*plus*. Furthermore, this method allows the import of three-dimensional characteristic maps by simply appending a two-dimensional characteristic map for every coordinate in the third dimension to the text file.

The deflating valve can be described by a one-dimensional characteristic (Figure 5), as the ambient pressure is considered to be constant ($p_3 = p_{\mathrm{Amb}} =$ constant). The measured compressor mass flow as a function of its outlet pressure can also be imported as a one-dimensional characteristic without further calculations (Figure 6). The inflatable seat cushions are described by their pressure-dependent volume. The dynamical strain behavior is considered by summation of a proportional and a first-order lag element.

3.1.5 Simulation model

The validation of the component models in DSH*plus* can be achieved by simulating and measuring the time variant pressure in a system, consisting of elements measured beforehand. The system includes a compressor, a triple valve (three single 3/2-way valves in one housing), three cushions of two different types, and several hoses. In the DSH*plus* model, each component is represented by a corresponding component model, parameterized by measured characteristics, whereas the volume of the hoses is contained in the depicted volume nodes. Compressor and valves of the simulation model are driven by the measured control signals of the test bench to achieve identical conditions for simulation and measurement. Figure 9 shows the test bench and the corresponding simulation model.

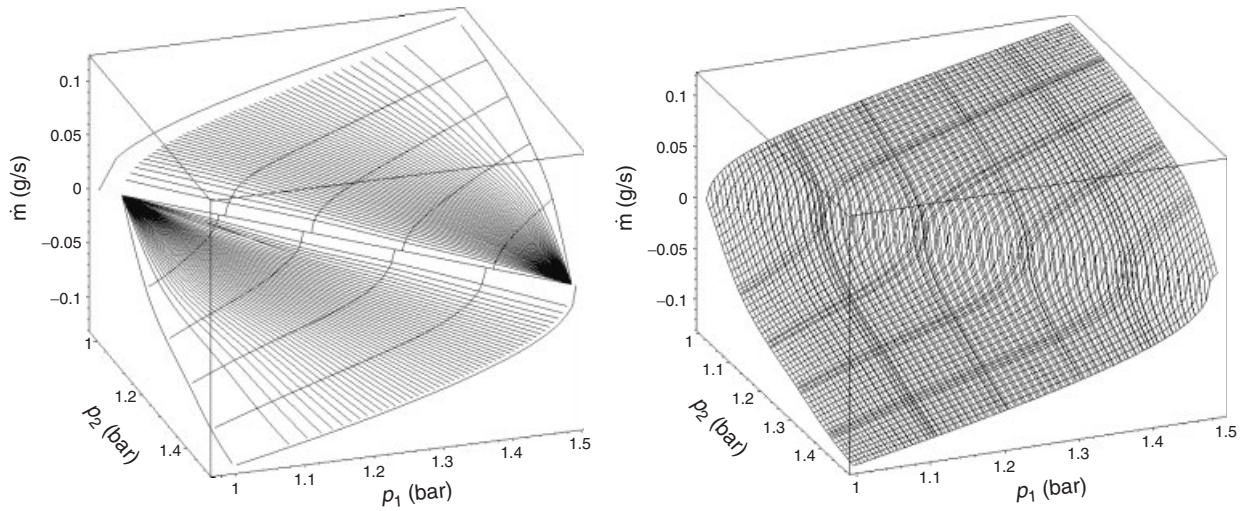


Figure 8. Characteristics map of the valve using the radar-splines method.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

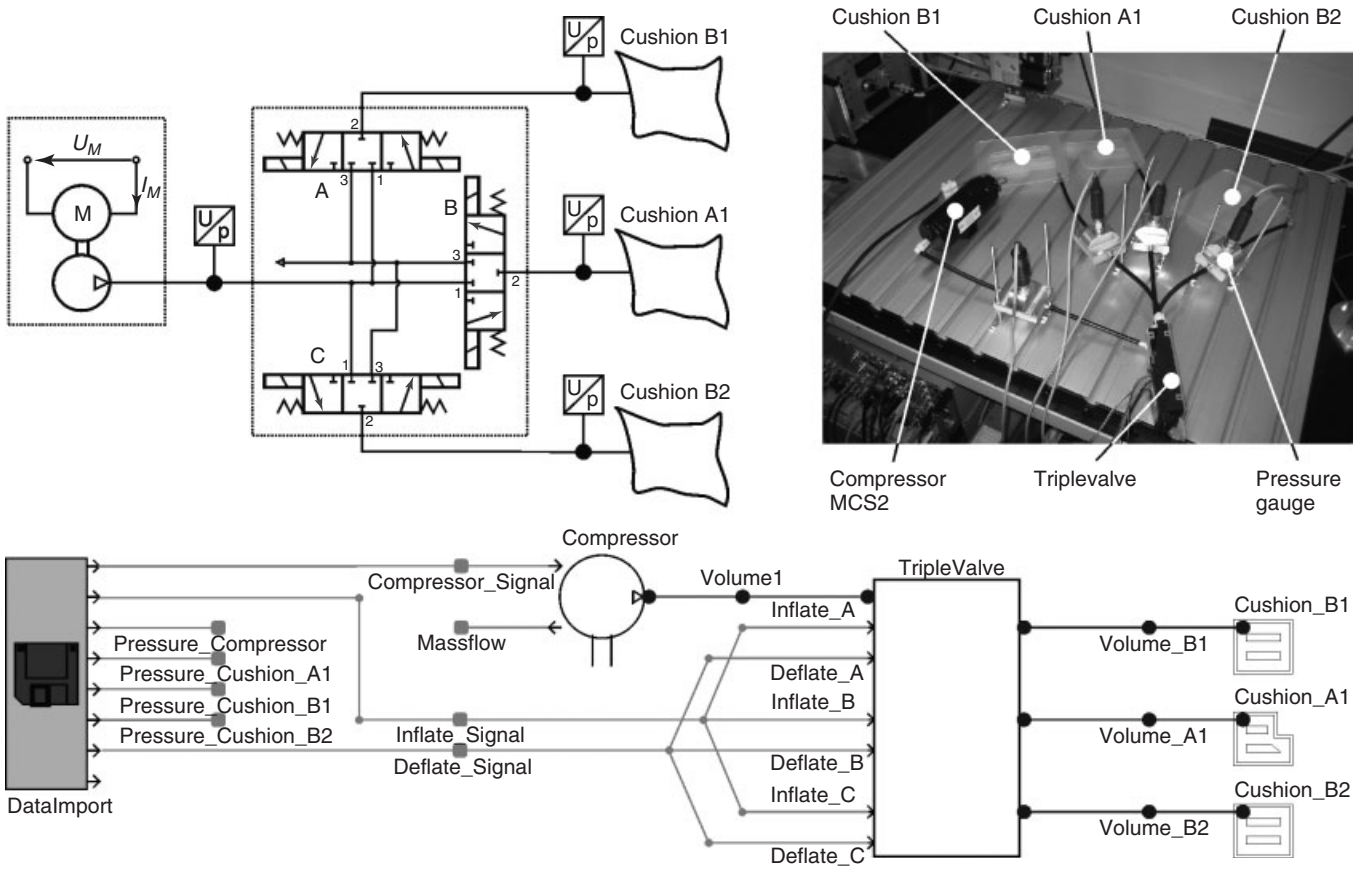


Figure 9. Test bench and corresponding simulation model.

The very good correlation between simulation results and measurements shown in Figure 10 emphasizes the high quality of the gauging method and the modeling. In combination with the fast gauging method for all system components, the simulation model enables the user to optimize the contour adjustment of automotive seats in a very short time. This is a big improvement compared to the state of the art method consisting of trial and error bench testing of hardware components.

The one-dimensional simulation enables the user to conduct parameter studies in a very fast and comfortable way, because no additional pneumatic components are required.

3.1.6 Summary of example 1

As an example of a complex pneumatic system, a pneumatically actuated comfort system for the contour adjustment of automotive seats was presented. A measurement method for the characterization of the system's elements can be used for quality inspection tasks, fault detection, and quantification, as well as the extraction of element-specific maps and parameters. These maps and parameters can be used for parameterization of a simulation model. Pneumatic valves, compressors, and cushions were investigated in order to determine their theoretical model behavior. The measurement method is not limited to pneumatic miniature elements and can be used for common elements of industrial pneumatics, likewise. The good correlation between measured pressure characteristics and its simulated values during the validation phase emphasizes the high quality of the simulation model and the need for highly accurate data to parameterize the components. Therefore, the simulation model can be used for a preestimation of the system's behavior or an investigation of the parameters' sensitivity.

3.2 Example 2: simulation of tubes and hoses

In automotive engineering, tubes and hoses are used most of the time to connect different pneumatic and hydraulic subsystems and components, such as the fuel system, the air conditioning, brake systems, exhaust systems, and power steering unit. These components typically consist of highly

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

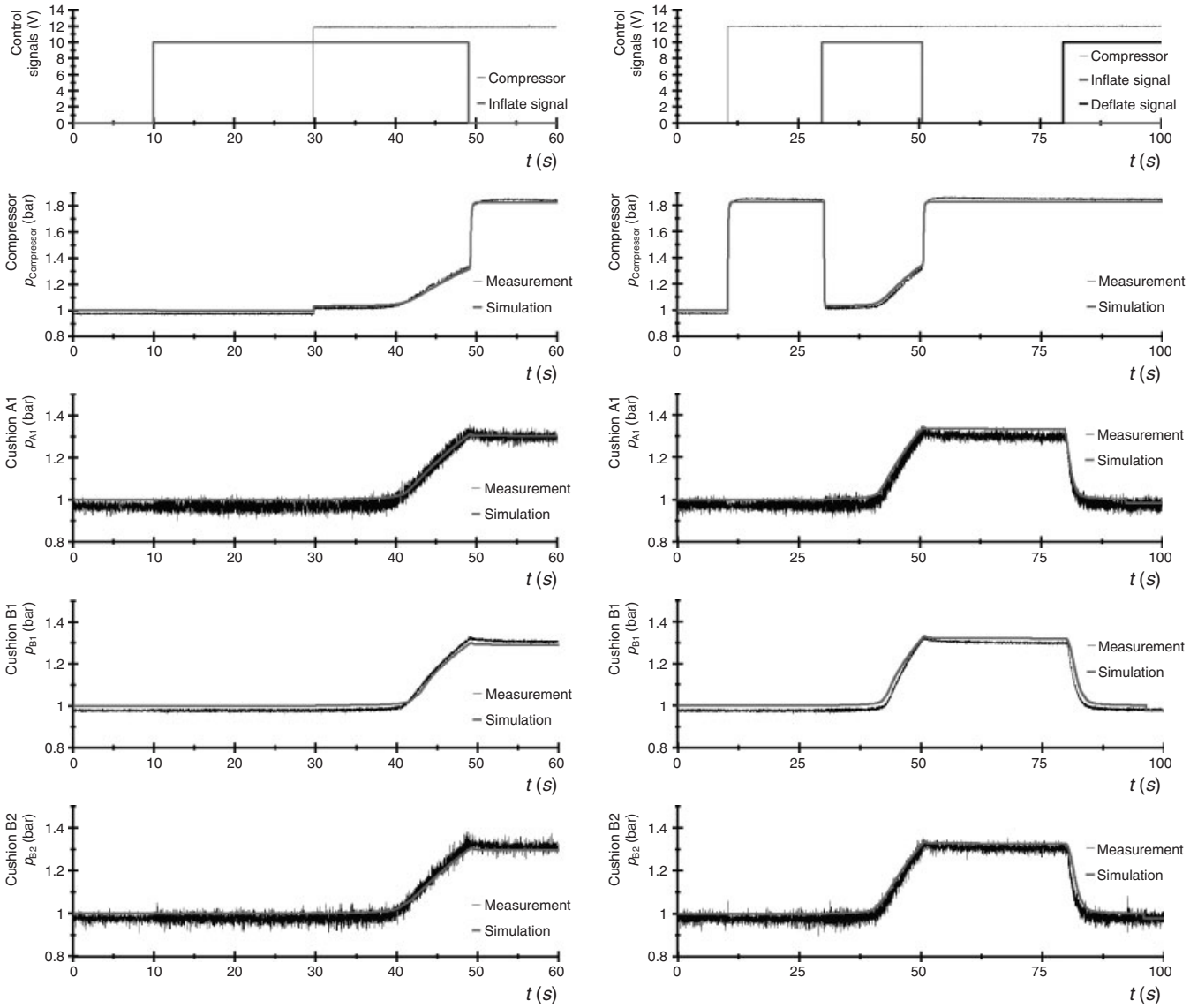


Figure 10. Validation of the simulation model.

integrated components, which must be are arranged on a very limited installation space. These limited installation conditions require flexible pathways and good mechanical features. Tube and hose connections fulfill those requirements and represent very flexible connection elements.

Figure 11 shows a typical expandable automotive line of average complexity, used in power steering units.

Owing to excitation from different sources, such as pumps, valves, and vibrations caused by the engine or other ancillary units, unwanted pulsations in the tubes or hoses can arise. These pulsations can lead to mechanical vibrations, which could cause unwanted noise emission in the form of structure-borne sound. The vibrations can cause a disturbance in the passenger cabin, which deteriorates the comfort of the vehicle. Furthermore, pressure peaks caused by resonance can result in high component strains especially in high pressure connections. These effects are summarized within the term *NVH* (*noise, vibration, and harshness*), which is also used to describe the study and modification of noise and vibration characteristics of vehicles.

Well known as a source of noise within the interior of a vehicle is the hydraulically assisted power steering system. The hydraulic pump of the steering system initiates a pressure ripple into the hydraulic circuit in which the pressure ripples propagate throughout the system as fluid-borne noise. To minimize the noise emitted by the power steering system, different noise reduction techniques are used. The most convenient way to reduce the noise emission

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 11. Typical expandable automotive line with average complexity (Johanning, Baum, and Wurmann, 2009). (Reproduced by permission of Heiko Baum.)

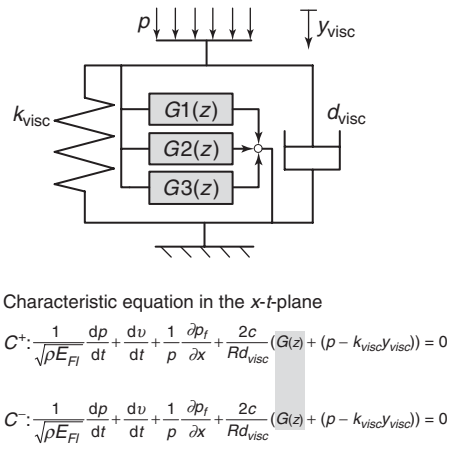


Figure 12. Implementation of parametric resonance into the characteristic equation (Baum and Johanning, 2009). (Reproduced by permission of Heiko Baum.)

is the installation of a tuning cable with a certain length to induce a destructive interference and therefore attenuate the amplitude of the pressure ripple (Visteon, 2005).

In addition to the destructive interference, structural damping can be used for noise reduction. The length of elastic hose elements in the hose assembly is increased, so that the energy from the pressure ripples is absorbed by the expansion of the hose wall. Furthermore, the hose reduces the wave speed in the fluid (Visteon, 2005).

If excessive noise emission and vibrations occur in the final product, the reinstatement work usually takes a lot of effort and time. To compensate or suppress the propagation of these pulsations throughout the connections and components, the mechanical features of the connections must be considered in an early stage of the development process.

In the past, a combination of different methodologies using analytical (Beater, 1999) and experimental approaches were used. However, all of these methodologies show significant disadvantages. The experimental methodology using a trial-and-error approach on a physical model is very time consuming and therefore costly. The analytical methodology uses a set of equations in a closed form and is therefore often restricted to simplified geometry and material properties. Furthermore, the increasing complexity of, for example, the power steering system and the typically nonlinear system behavior complicate the analytical description. The use of a finite element approach allows an accurate description of the isolated tubes and hoses but struggles with the simulation of the whole system because of the complexity of component interactions, which must be regarded.

A methodology without the above-mentioned disadvantages facilitates a time-based modeling approach provided by modern system simulation tools, such as DSH*plus*. In contrast to finite element method, the time-based approach allows the accurate and fast evaluation of the NVH characteristics of the system, including system vibration, airborne noise, and fluid-flow characteristics because the parameters at the boundaries are calculated by physically modeled system components such as the pump and valve of a power steering unit. This even enables the simulation of time-dependent operating conditions such as driving maneuvers.

This makes the system simulation the first choice to perceive and circumvent problems regarding NVH. Furthermore, changes can be applied and tested in an early development stage, which leads to a significant cost saving and reduction of time to market. In addition, system simulation tools can be used to troubleshoot systems in which NVH problems occur. The system simulation allows to clearly identify the source of vibration in a depicted and parameterized system. Furthermore, the system simulation enables the user to develop and test adequate countermeasures in a short period of time.

To obtain accurate results, all relevant physical aspects of the connection element must be implemented and properly parameterized. In the modeling process, this poses the greatest challenge. For example, steel tubing allows a simple phenomenological approach to describe the viscoelastic properties of the tube walls by a simple spring damper model (Müller, 2002), whereas the damping of hoses is essentially influenced by the temperature- and pressure-dependent viscoelastic properties of the material

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

used for the hoses. Furthermore, measurements showed that typical hoses used in power steering units allow volumetric expansion rates of 15–35 ccm/m at 23°C, which is substantially higher than steel tubes.

The viscoelastic properties have a significant influence on the damping and are adjustable by the hose material and the winding characteristics. In a simple spring damper model of the hose wall, the interaction of the characteristic frequency of the connection element and the damping of the connection wall is neglected. The representation of these specific characteristics requires a more complex model. Therefore, a parametric resonance is implemented into the model Figure 12.

The parameterization of the phenomenological models is done by measurements on a test bench, which allows the characterization of a specific material property of a hose.

On the basis of the characterization of different hoses, the characterization for frequently used fittings and hose elements, as shown in Figure 13, is available in the DSH*plus* component library.

By combining these basic elements, almost any desired elastic hose configuration can be modeled. Furthermore, the combination of basic elements allows the integration of resonators such as volume, Helmholtz, and cavity resonator, as well as throttles or orifices. Furthermore, a combination with a tuner cable is configurable and allows the optimization of a whole hose assembly.

3.2.1 Measured data processing

The foundation of realistic simulation results is the correct parameterization of the system components in the system

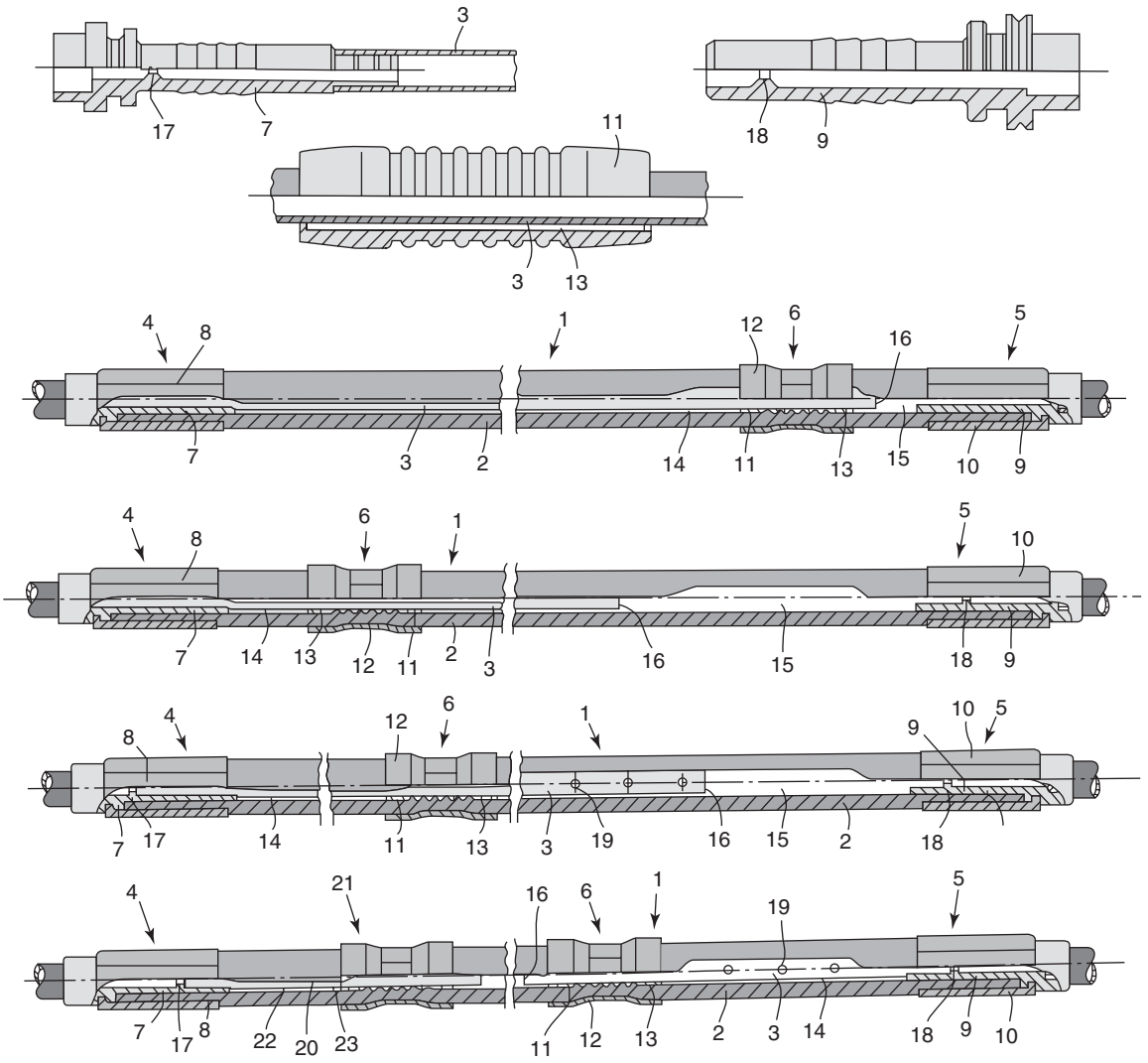


Figure 13. Examples for fittings and flexible hose elements. (Eaton, 1992. Reproduced by permission of Eaton Aeroquip GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

simulation. To correctly parameterize connections in the simulation, their dynamic properties are required. A well-established approach to determine the dynamic properties from experimental data is the quadrupole analysis, which allows the determination of the wave propagation in tubes and hoses in a wide frequency spectrum. On the basis of the measured data, the transmission behavior of the connection can reliably be reproduced in the simulation tool.

The quadrupole analysis is based on a process, originated in the field of electrical engineering, and is used to characterize an electrical network with two pairs of terminals connected together internally by the electrical network, which is to be investigated. This allows the mathematical description of any linear circuit detached from the physical buildup, provided that the circuit does not contain an independent source and satisfies the port condition. The port condition requires that the same current must leave and enter each pair of the terminals.

By the transferability of electrical effects to fluid power effects, the current corresponds to the volume flow and the potential corresponds to the pressure.

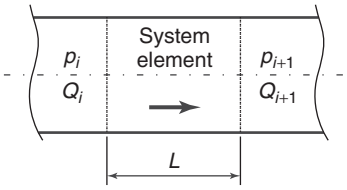


Figure 14. Line segment.

Therefore, a linear fluid technical system element can be described similarly to an electrical system element by four parameters. The transfer matrix consists of the elements T_{ij}, which correlate the pressure p_i and volume flow Q_i of the input side with the pressure p_{i+1} and volume flow Q_{i+1} of the output side of a line segment (Figure 14).

$$\begin{pmatrix} p_i \\ Q_i \end{pmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{pmatrix} p_{i+1} \\ Q_{i+1} \end{pmatrix} \tag{12}$$

The vector of state variables in the frequency domain of $\hat{p}_1$ and $\hat{Q}_1$ on the input side results from the multiplication

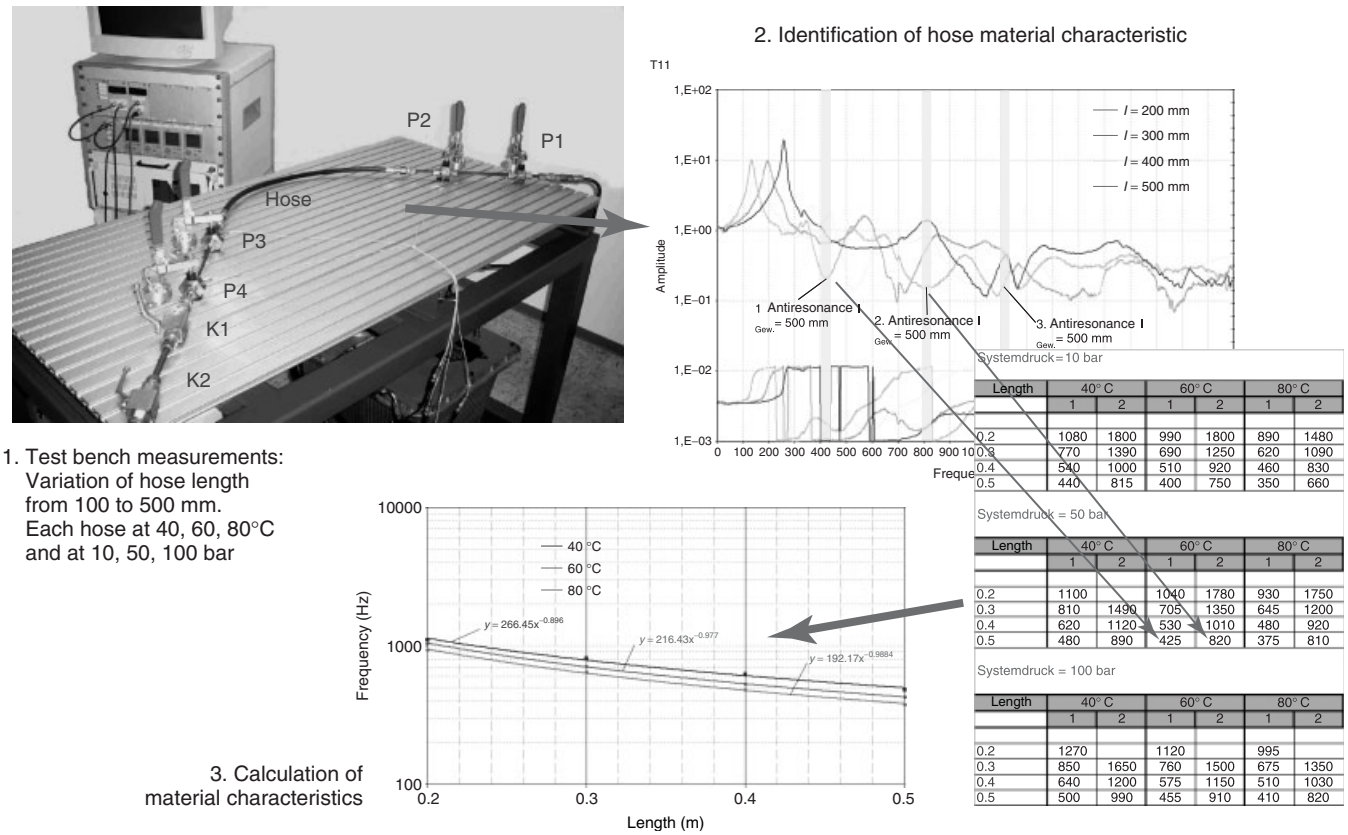


Figure 15. Test bench for identification of hose material characteristics (Baum and Johanning, 2009). (Reproduced by permission of Heiko Baum.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of the transfer matrix T and the state variables $\hat{p}_2$ and $\hat{Q}_2$ on the output side.

To determine the transfer matrix T, the pressure pulsations can be measured directly with high resolution using pressure sensors. An indirect measurement method is necessary to determine the volume flow because volume flow rate sensors either have an influence on the pressure propagation or lack the required dynamic properties. Therefore, the volume flow rate is usually determined in a reference tube, whose characteristics are well known, so that the flow rate can be determined indirectly from the measured pressure signal.

3.2.2 Test bench

The transfer matrix can be determined from the Fourier transformation of the measured pressure signals of any given tube or hose. Figure 15 shows the test bench with a mounted hose segment.

The measured tube characteristics allow the system simulation to consider the connection properties. Hence, any combination of these elements can be simulated. The model validation of a line assembly is conducted on the real test bench, which is also available as fully parameterized component in DSH*plus*, depicted in Figure 16. Therefore, the validation process of newly implemented components is narrowed down to comparing the simulated line characteristics to the measured line characteristics.

The following example shows the optimization process of pulsation propagation in DSH*plus*.

At the beginning of every development, process stands the customer specifications regarding the dynamic system behavior of the assembly of expandable hoses and tubes. The goal of the simulation process is to find an adequate assembly of hoses and tubes, which fulfill the targeted pressure losses, frequency response, or amplitude damping.

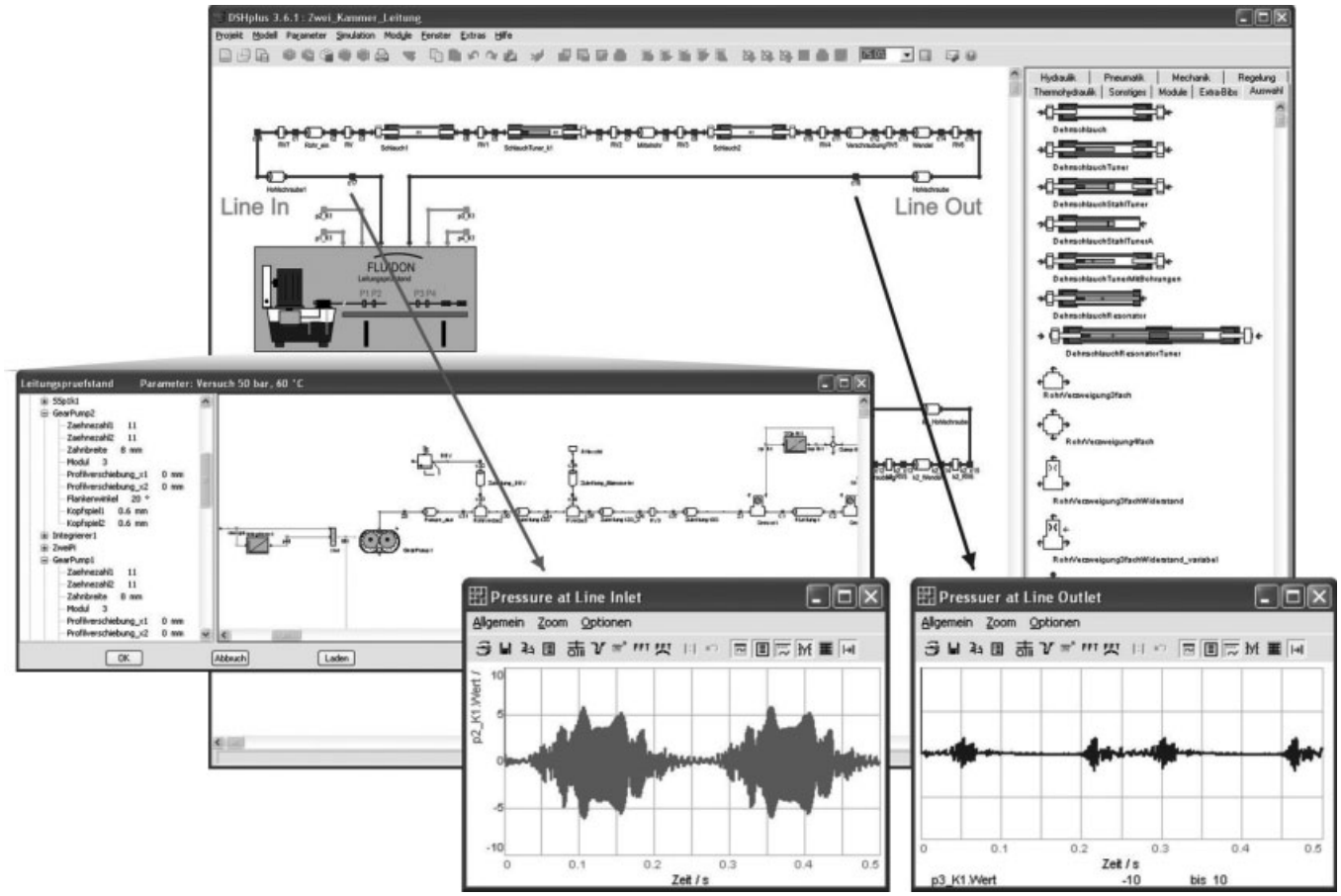


Figure 16. Flexible hose element library and simulation model of test bench with installed line (Baum and Johanning, 2009). (Reproduced by permission of Heiko Baum.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For the final design, only little tuning on the test bench or the vehicle is required. This allows a significant time saving compared to a typical trial and error procedure.

Figure 17 depicts the procedure of the optimization process. Beginning with the CAD data of the packaging space and connection alignment, a first connection design is created, outgoing from developments in former projects.

On the basis of the geometrical data of the CAD model, the assembly is recreated in DSH*plus* from the component library and parameterized by the geometric dimensions of the design. Only the expandable hose sections are parameterized by a data set that describes the elastic properties of the hose material. If the data set of the specific material is not available in the material database, the material properties must be measured once before the design process and are then available for future projects from the material database.

Integrating the design into the virtual hose test bench, which is equilibrated with the real test bench, offers a simple way to validate the simulation model.

Furthermore, the flow conditions of the desired application on the ends of the hose assembly must be set.

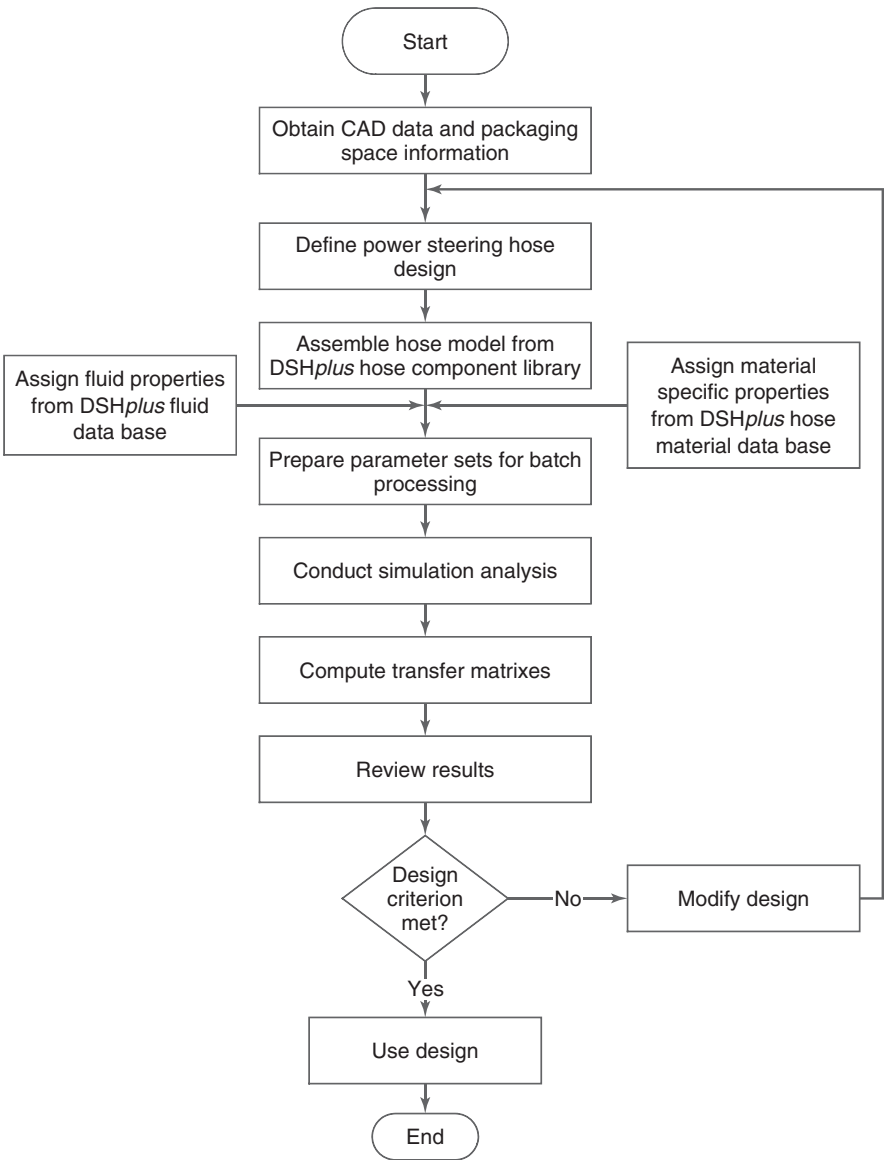


Figure 17. Development methodology to optimize automotive hose assemblies (Baum and Johanning, 2009). (Reproduced by permission of Heiko Baum.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

This can be achieved either by modeling the necessary system components such as valves or pumps or by a characteristics diagram of the connected components. If the parameterization of the hose assembly is finalized, an automated design variation under consideration of the prescribed packaging space and design requirements can be initiated. The quadrupole analysis of DSH*plus* allows the graphical representation of the findings.

Furthermore, DSH*plus* allows a parameter-driven design optimization. The user specifies quality criteria over a certain frequency spectrum, which is then used to drive the parameter variation. Furthermore, the influence of different geometric design parameters can be investigated by a design of experiment (DOE) analysis.

If the developed geometry satisfies the design requirements, a prototype can be build up; otherwise, the user should return to the beginning of the design process and start again with a new starting design.

3.2.3 Application of the tube and hose model

Using the example of a two-chamber automotive line with PTFE tuner and a single-chamber automotive line with steel-flex-tuner, the suitability of DSH*plus* as simulation tool is presented. The two connections are existing stock parts and have been modeled in DSH*plus* to use in a design parameter study. The design parameter study allows the automated variation of geometric lengths and diameters of the tuners, hoses, and tube segments as well as the position of the tuner elements.

To validate the simulation results, the flow characteristics of the connections have been measured on the test bench.

Figure 18 shows the transmission behavior of the two-chamber automotive line. The transmission behavior of the initial configuration A shows a distinct rise in the amplitude ratio in the frequency range 500–700 Hz. The optimization process of configuration B results in an improved damping

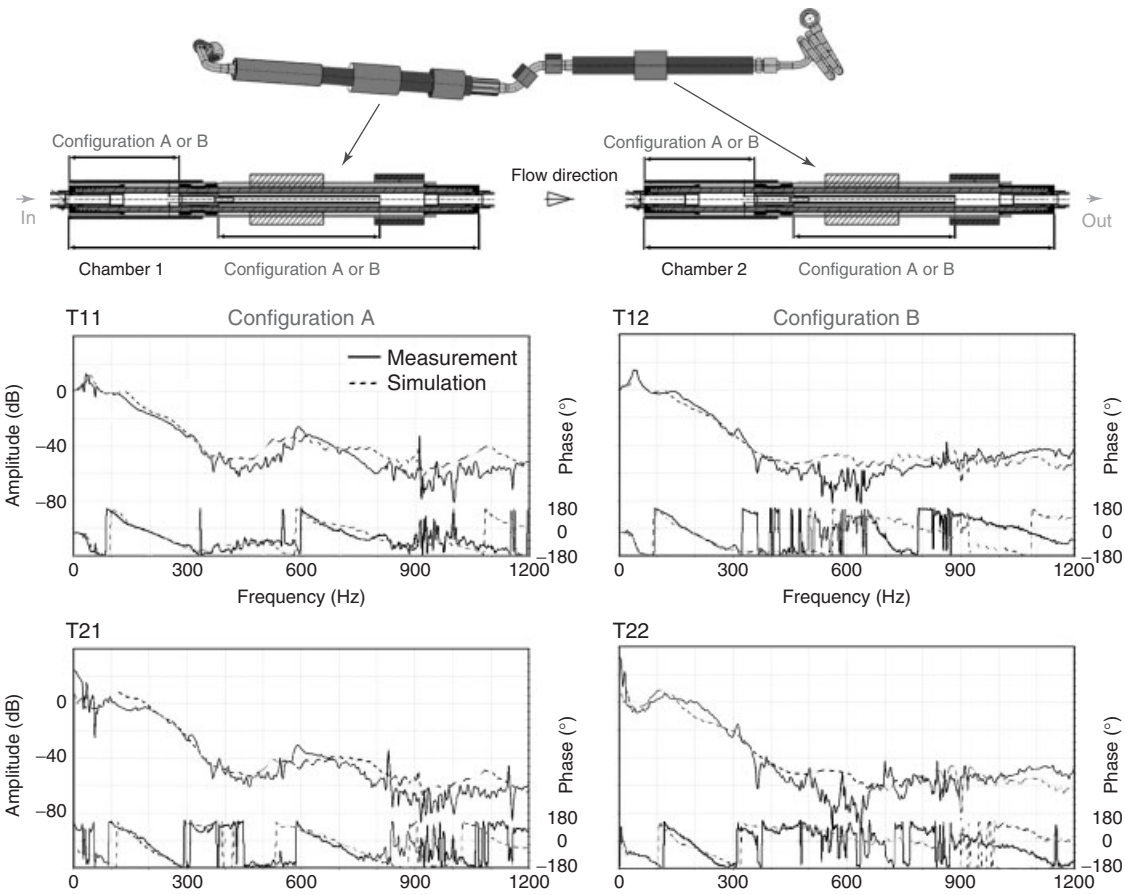


Figure 18. Measuring and simulation of two-chamber automotive line (Baum and Johanning, 2009). (Reproduced by permission of Heiko Baum.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

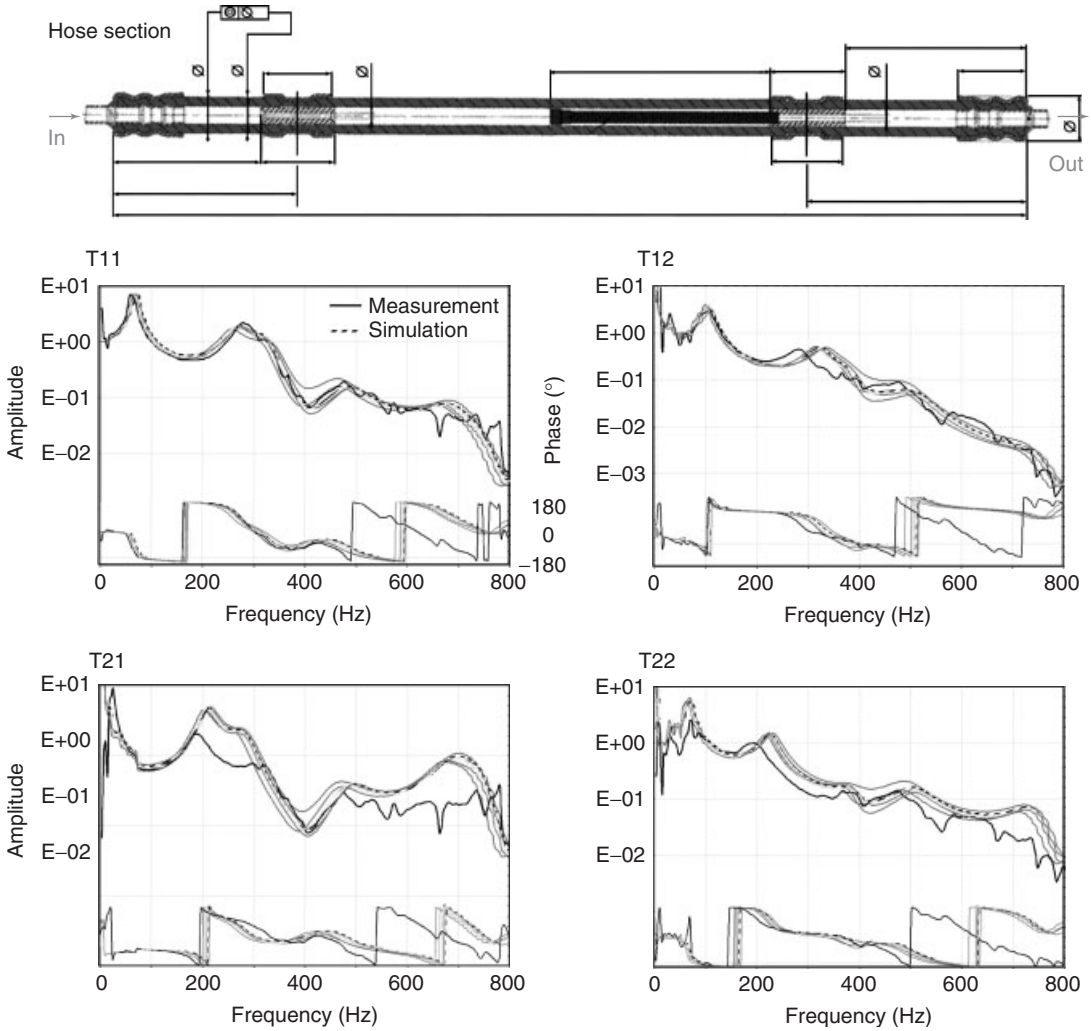


Figure 19. Robustness analysis of a one-chamber line with steel tuner (Baum and Johanning, 2009).

in the frequency range, owing to a variation of the throttle position and length of the tuner cable.

Figure 19 depicts the measurement and simulation of a flexible one-chamber-line with steel tuner.

In the first simulation of the one-chamber line, the exact geometric dimension of the existing line was modeled. On the basis of the performed simulation, a parameter variation of the relevant geometric dimensions, as depicted in Figure 19, is conducted. The limits for the parameter variation of the geometric dimensions were deduced from the tolerance specification of the line. This analysis shows the variation in the dynamic properties of the line within the accepted tolerance span of manufacturing. The simulated variations envelope the curve of the simulation using the exact geometric dimensions. The parameter variation enables the identification of sensitive geometric dimensions. On the basis of the simulation results in an early development stage, appropriate countermeasurements can be considered.

3.2.4 Summary of example 2

The use of system simulation tools accompanying the development process of expandable automotive lines in the early stages allows a significant cost and time saving because protracted trial-and-error studies can be circumvented.

DSH*plus*, among others, has all necessary software modules to aid the designer during the whole design process. The component library for automotive lines allows the user to start with a virtual prototype for a variety of different application fields not only limited to the automotive sector. On the basis of the virtual prototype, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

design and optimization process regarding the specified design requirements can be performed. In order to do this, DSH*plus* offers a comfortable way to perform optimization and robustness – as well as sensitivity studies for the line design. Furthermore, DSH*plus* allows the evaluation of different parameters utilizing methods of DOE. This accelerates the overall design process and significantly reduces the number of time-consuming field tests.

4 SUMMARY

Modern system simulation tools are an indispensible utility in the modern design process, applicable to a wide variety of fields in engineering. System simulation tools such as DSH*plus* provide the user with a certain level of planning security in an early design stage. Furthermore, modern simulation tools provide a possibility to test different design concepts with relatively little effort in time and money. Later in the design process, system simulation tools offer valuable optimization tools to optimize an initial design under consideration of the specified design requirements. The resulting design is optimally attuned to the specifications in a short time span compared to conventional methods of trial and error. In addition, the provided tools allow an elaborate comparison and assessment of different designs.

Furthermore, system simulation can be used to illustrate and evaluate the system behavior of interconnected components. Early system evaluation significantly reduces failure during first startup. Furthermore, the depiction of whole systems allows optimal choice of components before realization. Hence, a cost optimal solution can be found for the manufacturer and for the customer.

RELATED ARTICLES

Petrol Fuel Injection Systems
Diesel Fuel Injection Systems
Hardware-in-the-Loop Simulation

REFERENCES

Baum, H. and Johanning, H.-P. (2009) *Simulation of flexible hose lines for power steering and active chassis application*. Aachener Kolloquium Fahrzeug- und Motorentechnik 2009.

Beater, B. (1999) *Entwurf hydraulischer Maschinen: Modellbildung, Stabilitätsanalyse und Simulation hydrostatischer Antriebe und Steuerungen (VDI-Buch)*, vol. 1999, Springer-Verlag, Berlin Heidelberg.

Eaton Aeroquip GmbH (1992) Expandable Hose that Reduces the Hammering Produced in Hydraulic Systems by Pumps. United States Patent 5094271, Publication Date: 1992-03-10.

FLUIDON GmbH (2007) User Manual DSH*plus* 3.7, FLUIDON GmbH, Aachen, Germany.

Johanning, H.-P., Baum, H., and Wurmann, G. (2009) Simulation von Dehnschlauchleitungen für Lenkung und Fahrwerk. ATZ-Automobiltechnische Zeitschrift, 2009-06.

Müller, B. (2002) Einsatz der Simulation zur Pulsations- und Geräuschminderung hydraulischer Anlagen. Dissertation RWTH Aachen.

Siebertz, K., Reinertz, O., Fritz, S., and Murrenhoff, H. (2010) Rapid Gauging Method and Generic Modelling Approach for Pneumatic Seat Components. *7th International Fluid Power Conference*, 22–24 March, Aachen, 2010.

Visteon Global Technologies, Inc. (2005) Method of Power Steering Hose Assembly Design and Analysis. United States Patent 6917907, Publication Date: 2005-07-12.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto034
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Comparison of the Modeling Techniques for Chassis Applications. An Advice for the User

Frank Heßeler, Alexander Katriniok, Matthias Reiter, Jan Maschuw, and Dirk Abel

RWTH Aachen University, Aachen, Germany

1 Introduction 1
2 Rapid Control Prototyping 2
3 Overview of Modeling Techniques for Different Chassis Applications 4
4 Usage of Different Modeling Techniques in Common Software tools 9
Related Articles 14
References 14

1 INTRODUCTION

Progress in simulation technology leads to a wider application of simulation software for the reduction of testing costs in many engineering disciplines. This trend is also true for vehicle design at all development levels. The level of detail of the models is sufficient for simulation of entire vehicles as well as for detailed simulations of separate mechanical or electrical components. Each engineering discipline applies its own software tools, developed specifically for the purposes of application. An updated and complete overview of single programs is thus not easy to provide and lies beyond the scope of this chapter.

In order to meet the increasing requirements to safety and comfort in a vehicle, pure mechanical systems are complemented or replaced by mechatronic systems. This tendency can be also observed in chassis systems. Typical examples of such systems are antilock braking system (ABS) and electronic stability control (ESC), which are currently part of a standard equipment of each modern vehicle. Moreover, there is an increase in the development and marketing of comfort features, such as adaptive damping or electric power steering. Such systems can be strongly connected with safety critical systems, such as ESC. To exemplify this statement, electric power steering can be named, which receives steering directions from the ESC and thus also implements a safety critical task.

In addition to the pure mechanical construction of the chassis components, the development of mechatronic systems receives a steadily growing attention of the automobile manufacturers, as mechatronic systems can help to meet both safety and comfort requirements. However, interconnection of mechanical and electrical components and integration of the control software make the development even more complex. In this area too, there are a number of modeling technologies for simulating a vehicle. Modern functional developments are based on the methods of Rapid Control Prototyping (RCP). With these methods, it is possible to test the developed control functions during early stages of the development cycle. In addition to established signal-oriented modeling methods, new physically based, object-oriented modeling languages, such as Modelica, are applied.

The purpose of this chapter is to provide an overview of various modeling approaches in the area of vehicle simulation for the application in the functional design and to show how these models can be used in the development

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

process of RCP. In addition to the explanation of the different modeling approaches, two different ways to model these differential equations on a digital computer will be introduced: (1) the established signal-oriented methods, as they are implemented in Simulink and (2) an object-oriented method using the modeling language Modelica.

The chapter is organized as follows: in Section 2, the principles of RCP as a method of modern functional development are elucidated, and the resulting necessity of having various model types is demonstrated. Section 3 gives an overview of popular modeling approaches for the dynamical description of the longitudinal, lateral, and vertical dynamics of a vehicle. An attempt is made to provide the reader with the instructions on various applications to facilitate the choice of an appropriate model. This chapter provides only a general overview and description of the methods and refers the interested reader to the related literature for details. Section 4 presents a comparison between signal-oriented and object-oriented methods for differential equations. On the basis of these methods, the widely used software Simulink by the Mathworks and Dymola are described. This chapter concludes with a summary.

2 RAPID CONTROL PROTOTYPING

Examination of the developments in the automobile industry during the past decade shows that the number of mechatronic systems in a vehicle grows steadily and will continue to grow in the future. The same is also true for chassis control. This trend is particularly encouraged by the increase in safety and comfort that is possible with the use of mechatronic systems. ABS and ESC are the most typical examples of such systems. To a great extent, these functions are realized by software that is integrated as an embedded system into a vehicle. A further trend is a stronger interconnection of separate functions in the vehicle, which makes the functional development of new systems even more difficult and time-consuming.

To guarantee goal-oriented and efficient development of the control software in these complex systems, the concept of RCP has been established for controller development. It combines the advantages of the classic method of V-modeling with the possibility of early testing at separate development stages in a standardized development environment. This testing helps to eliminate the disadvantages of the classic modeling in which certain design errors are only detected rather late in the development cycle. The basic prerequisite here is the use of a continuous tool chain to allow the designer to concentrate on the core competencies. Figure 1 depicts the V-model of a controller design for the graphical visualization of the single steps and their succession in the development.

According to the V-model, the design comprises the following steps: requirement analysis, specification, rough and detailed design, simulation, component test, and system test. In the classic development cycle, these steps are performed one after another, and the tests are performed at the end. In the RCP concept, the tests can be performed much earlier in the development cycle using the corresponding software tools. The development steps need not occur in the above-mentioned succession and can be performed repeatedly, rather quickly, with the assistance

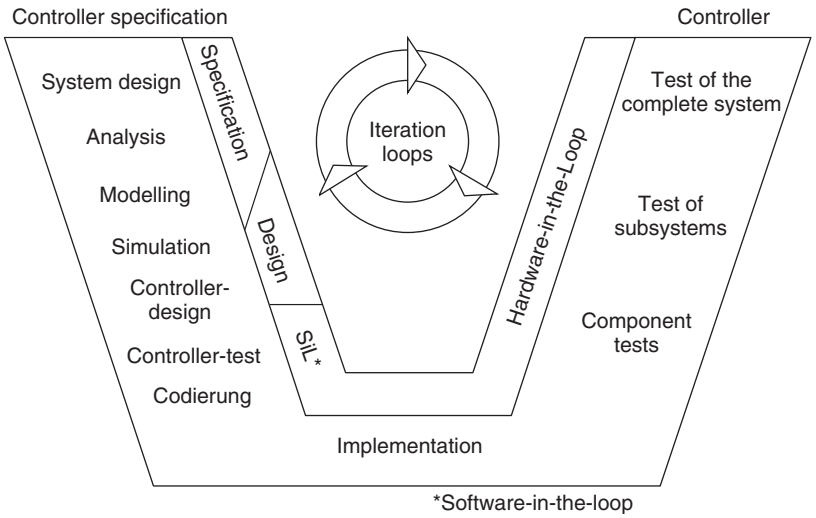


Figure 1. V-model. (Reproduced from Rapid Control Prototyping, (Abel, 2005). With kind permission of Springer Science+Business Media.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

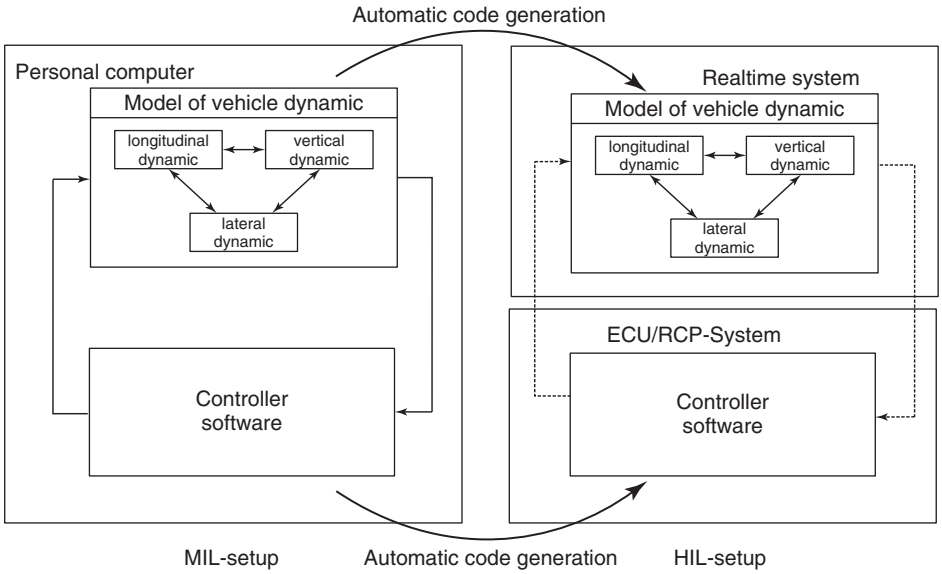


Figure 2. MIL-test.

of the software tools, in order to even more quickly detect and eliminate design and implementation errors.

Essential features of the methods of RCP development are Model-in-the-loop (MIL) and Hardware-in-the-loop (HIL) tests. They help perform tests during various stages of the development cycle, for example, for the purpose of also implementing horizontal iteration loops. Thus, it allows an early creation of controller software in the process of development by using automatic code generation, as well as testing with the available real components or with models of yet unavailable real components.

In Figure 2, two scenarios are compared: the MIL and the HIL tests. The MIL test can be performed during early stages of the development cycle, as the development environment can be, for example, Matlab/Simulink by The Mathworks, in which the controller development also often takes place. The primary goal of a MIL test is the development of the controller's concept and validation of control quality. Therefore, it is important that the MIL models describe the real vehicle behavior as good as possible.

In contrast, in a HIL test, both components (system model and controller model) must be executed in real time. Using automatic code generation, it is possible to transfer the models to a real-time simulation platform. This step is one of the central elements of RCP because the complex manual software design can be automated. It allows making quick changes and tests directly in a real-time system by the press of a button. The primary goal of HIL tests is performing system tests in which the designed controller structure runs on the target platform and is tested there in combination with all other components.

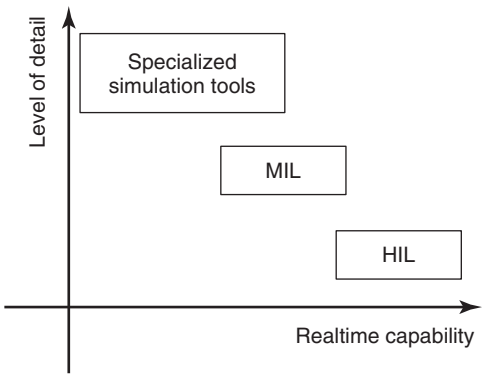


Figure 3. Comparison between different types of models.

Simulation models of the controlled system play a crucial role in all these testing strategies. On the basis of their application, the MIL and HIL models differ in their level of detail and in real-time capability.

Figure 3 shows the relation between real-time capability and level of detail for various model types. Special simulation tools such as ADAMS by MSC Software or Simpack by SIMPACK AG, that can be used to design the mechanical components of a chassis, possess the highest level of detail and also require the highest amount of computation time. Generally, these models are not suitable for simulation of an entire vehicle because of their computation time, and thus are seldom used for controller development.

Concerning MIL models, they can simulate an entire vehicle with all its essential components at a high level of detail. Examples of these models are CarMaker by IPG or CarSim by Mechanical Simulation Corporation. Depending

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

on the selected level of detail, these models are capable of providing real-time performance on very powerful computers. Because the simulation of an entire vehicle they provide is of a high quality, these models are used for MIL tests, which are primarily intended for validation of the control functions. The MIL test need not be run in a real-time environment and can be performed directly on a personal computer. As the controller development is implemented mostly in Matlab/Simulink, the MIL test is also often performed as a co-simulation in Simulink. Most software producers offer corresponding interfaces for Simulink.

Another model type is a HIL model. The main function of the HIL models is real-time performance that, however, reduces the level of detail and thus the accuracy of the models. The model accuracy must be sufficient for the interaction of all designed features and the corresponding diagnosis functions. The control quality does not have the primary importance here, as it has already been tested during the MIL tests. The real benchmark is always to be found in a real vehicle where the quality of the target results of the described process can be judged on the basis of real tests.

Thus, both detailed and real-time capable system models are equally important for modern controller development. In addition to the application of the models for validation of the controller software, these models can also be used in the development of the control concept. A model-based controller design usually requires a linear control model that can be applied either for control design or even as a component of the control concept itself. Thus, the application of simulation models gains further importance, simultaneously making it more difficult to choose a suitable model for a certain purpose. Section 3 deals with various models with different levels of detail for a number of applications. Each model is illustrated with an example of its application to facilitate the reader's choice of a right model.

3 OVERVIEW OF MODELING TECHNIQUES FOR DIFFERENT CHASSIS APPLICATIONS

3.1 Aim of modeling

Developing a chassis application requires a mathematical model that describes the vehicle's dynamic behavior in an adequate manner with respect to the considered application. In particular, the relevant system inputs (e.g., the steering angle) and outputs (e.g., the yaw rate) have to be identified. Subsequently, the dynamic system behavior has to be described in terms of differential equations. The level of detail, that is, the number of differential equations and the number of employed inputs and outputs, inherently depends on the chassis application. From the perspective of control engineering, a process model is required for the purposes of *analysis*, *controller synthesis*, and *validation*, as subsequently explained. As far as process *analysis* is concerned, a plant model is used to analyze the open-loop, that is, the uncontrolled, process behavior. In the linear case, the process' eigenvalues are crucial for the plant's stability, its dynamic time constants, and the damping ratio. For the purpose of *controller synthesis*, a simplified plant model is commonly employed to design the controller in such a way that the closed-loop system fulfills the desired requirements such as stability, offset-free tracking of reference values, the ability to suppress external disturbances, or further performance requirements. Before the controller is applied to the real plant, it is *validated* in numerical simulations. Simplified (linear) plant models are used for controller synthesis, whereas plant models with much higher complexity are employed for validation purposes.

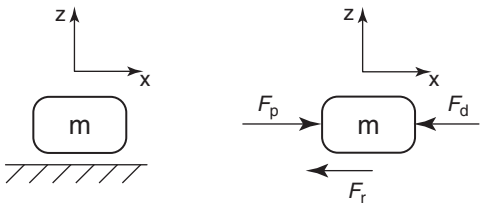


Figure 4. Free body diagram.

In order to illustrate the above-mentioned issues, the process of modeling is subsequently illustrated for several chassis applications having different requirements as far as model complexity is concerned. In this context, applications of longitudinal, lateral, and vertical vehicle dynamics are taken into account.

3.2 Longitudinal vehicle dynamics

In this section, model approximations for applications in which primarily the longitudinal dynamics are of interest will be presented. In particular, linear model approximations, that are often used for the application of adaptive cruise control (ACC), and nonlinear effects, that are needed, for example, to model problems addressed by ABS, will be discussed. In the following sections, the chassis will be modeled as point mass that is subject to the forces applied from the surrounding environment. A free body diagram is given in Figure 4.

F_p, F_d, and F_r represent the propulsion force, the drag force, and the rolling resistance. According to (Mitschke and Wallentowitz, 2004), the main influencing forces and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

consequently the longitudinal vehicle dynamics on flat terrain can be described by:

$$m\dot{v}_x = \underbrace{\frac{T_e\left(n_e, \alpha_{th}\right) \bullet \eta \bullet i_g}{R}}_{F_p} - \underbrace{c_w A \frac{\rho}{2} v_x^2}_{F_d} - F_r \qquad (1)$$

Thereby T_e, n_e, and α_{th} denote the engine torque, the engine rotational speed, and the throttle input, respectively. The gear ratio, efficiency, and static tire radius are given by i_g, η, and R while the drag coefficient, the reference area, and the air density are given by c_d, A, and ρ. The above equation holds for engine operation; for brake operation, the term F_p changes sign and has to be replaced by an actuator dynamics that is controlled by a brake pedal position α_{br}.

3.2.1 Linear approximations

To differentiate between absolute values and differences with respect to the operating point, we introduce the perturbation $\tilde{v}_x = (v_x - v_{x,0})$, and $\tilde{\alpha}_{th} = (\alpha_{th} - \alpha_{th,0})$. If we neglect the velocity dependence of the rolling resistance and assume that the engine rotational speed can be related to the velocity as $n_e = v_x \bullet i_g/(2\pi r)$ with an effective tire radius r, the Taylor-series expansion of Equation 1 results in the following approximation of the velocity dynamics:

$$m\dot{\tilde{v}}_x + \left(c_d A \rho \bullet v_{x,0} - \frac{\eta \bullet i_g^2}{R \bullet 2\pi r} \bullet \left.\frac{\partial T_e}{\partial n_e}\right|_0\right)\tilde{v}_x = \frac{\eta i_g}{R} \bullet \left.\frac{\partial T_e}{\partial \alpha_{th}}\right|_0 \tilde{\alpha}_{th} \qquad (2)$$

Equation 2 exhibits a first-order lag behavior that can be summarized by

$$\tau \bullet \dot{\tilde{v}}_x + \tilde{v}_x = K_v \bullet \tilde{\alpha}_{th} \qquad (3)$$

where the time constant τ and the static gain K_v depend on the above-mentioned vehicle parameters and the operating point. Likewise, the dynamics during braking can be approximated by a linear model. Therefore, the partial derivative of the braking force with respect to the brake pedal position α_{br} has to be included instead of the derivatives of the engine torque.

The above model assumes that the engine torque or a braking pressure (or force) are used as system input that can directly be controlled. However, this is often not the case, and further dynamics have to be included. As a result, another model that is frequently used in literature describes the longitudinal acceleration denoted by a_x. Very often, lower level controls are applied to linearize the resulting dynamics such that this can be approximated again by a first-order lag element as

$$\tau \bullet \dot{a}_x + a_x = a_{x,ref} \qquad (4)$$

where τ denotes the resulting time constant. Concerning the approximation of drivetrain dynamics and underlying linearizing controls further reference is given to (Lu and Hedrick, 2004; Rajamani, 2006) and (Ha, Tugcu, and Boustany, 1989) respectively.

3.2.2 Nonlinear effects

The linear models introduced by Equations 2 and 4 stem from a Taylor-series expansion that is only valid for small deviations from the operating point. Typical driving conditions do normally involve a wide interval of operating points (e.g., different throttle or braking commands at different velocities) and hence cannot be expressed by only one of the models mentioned above. Often, this problem can be dealt with by switching between several different linear models that best describe the current operating point. It should be noted that this is only possible as long as operating conditions change slowly; otherwise, it might be better to use true nonlinear models to describe the drivetrain dynamics.

Besides different operating conditions, another source of nonlinearity arises from the force transmitted between tire and road. For the above-mentioned models, we were basically assuming that any torque applied to the wheels (through engine or brake commands) is transmitted to the road and results in a net force accelerating the mass m. But, the angular speed of the wheels ω and the longitudinal vehicle speed v_x are coupled by a nonlinear friction-slip characteristic that limits the transmittable force between tire and road. If the rolling resistance is neglected, the transmittable force can be expressed by

$$F_x = \mu \bullet F_z \qquad (5)$$

where μ and the tire slip λ are related by a (generally) nonlinear static mapping (friction-slip characteristic). The longitudinal tire slip is defined by

$$\lambda = \frac{\omega \bullet r - v_x}{v_x} \text{ (braking)}, \quad \lambda = \frac{\omega \bullet r - v_x}{\omega \bullet r} \text{ (acceleration)} \qquad (6)$$

For low slip values, a (quasi-linear) ascent of friction (or force) goes along with an increasing tire slip. If the slip exceeds a critical value, friction and hence the transmittable force do not further increase but decrease. Effective force-slip characteristics of a Pacejka tire model are shown in Figure 6a for the longitudinal direction discussed here and in Figure 6b for the lateral direction. A more detailed view of (nonlinear) tire models is given in Section 3.3.2 focusing on the lateral direction. Owing to the decreasing force (beyond the critical slip value), the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

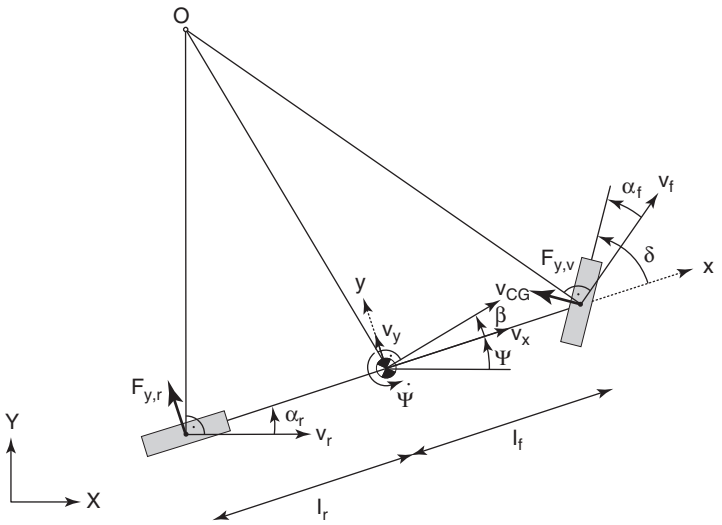


Figure 5. Free body diagram of single track vehicle model.

slip further increases until the wheel spins (during acceleration) or skids (during braking). Using ABS, this problem can be reduced and the maximum force can be achieved during the braking maneuver. As a result, the transmittable force and hence the longitudinal acceleration are limited by a maximum value. For critical maneuvers with high acceleration values, this effect has to be taken into account. As an example, a simplified extension of the linear model in Equation 4 would be an additional saturation of the effective acceleration to account for the related effects.

3.3 Lateral vehicle dynamics

3.3.1 Single-track model

Applications that aim at guiding the vehicle in the lateral direction require a plant model that is basically different from those that have been illustrated in the previous section. When considering advanced driver assistance systems (ADAS) such as lane keeping systems or even autonomous driving systems that are only capable of changing the vehicle's steering angle, a single-track model (see Figure 5) according to (Mitschke and Wallentowitz, 2004) is commonly employed to describe horizontal vehicle dynamics.

The single-track model combines the two wheels of an axle to a single one, while the center of gravity (CG) is placed on the road surface, that is, $h_{\mathrm{CG}} = 0\mathrm{m}$. In Equations 7–9 all rolling resistances as well as aerodynamic drag are neglected. Furthermore, it is assumed that no longitudinal forces are applied at the wheel. In particular, Equations 7 and 8 describe the translational degrees of freedom in the longitudinal and lateral directions with respect to the vehicle reference frame. Furthermore, yaw dynamics are modeled by Equation 9.

$$m\dot{v}_{\mathrm{x}} = m\dot{\psi}v_{\mathrm{y}} - F_{\mathrm{y,f}}\sin(\delta) \tag{7}$$

$$m\dot{v}_{\mathrm{y}} = -m\dot{\psi}v_{\mathrm{x}} + F_{\mathrm{y,f}}\cos(\delta) + F_{\mathrm{y,r}} \tag{8}$$

$$J_{\mathrm{z}}\ddot{\psi} = F_{\mathrm{y,f}}\cos(\delta)l_{\mathrm{f}} - F_{\mathrm{y,r}}l_{\mathrm{r}} \tag{9}$$

In this context, v_{x} and v_{y} denote the longitudinal and lateral velocities at CG with respect to the vehicle reference frame, $\dot{\psi}$ the yaw rate, δ the wheel steering angle at the front axle, m the vehicle mass, J_{z} the mass moment of inertia with respect to the vertical axis, and l_{f} and l_{r} the front and the rear wheel bases. When considering a lane keeping or path following system, additional differential equations are required to model the relative motion between the vehicle and the desired path, see (Keßler et al., 2007). As this section primarily focuses on the issue of modeling lateral vehicle dynamics, these equations are omitted for reasons of clarity. In order to determine the tire force $F_{\mathrm{y,f}}$ at the front and $F_{\mathrm{y,r}}$ at the rear axle, a tire model is required. In the following section, a modeling approach for these tire forces is discussed depending on the magnitude of the considered lateral accelerations.

3.3.2 Tire models

If small lateral accelerations are considered, that is, $|a_{\mathrm{y}}| \leq 0.4g$ on a dry surface, a linear tire model is sufficient to model the resulting tire forces, see (Mitschke and Wallentowitz, 2004). In detail, the side force $F_{y,i}$ at each wheel

$$F_{\mathrm{y},i} = c_{\alpha,i} \bullet \alpha_{\mathrm{i}}, i \in \{f, r\} \tag{10}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is assumed to depend linearly on the tire sideslip angles α_i where $c_{\alpha,i}$ describes the nominal tire cornering stiffness for pure cornering. In this context, the vertical tire load is assumed to be constant. The tire sideslip angles define the difference angle between the velocity vector at the wheel and the longitudinal wheel axis, that is,

$$\alpha_\mathrm{f} = \delta - \arctan\left(\frac{v_\mathrm{y} + l_\mathrm{f}\dot{\psi}}{v_\mathrm{x}}\right) \tag{11}$$

$$\alpha_r = -\arctan\left(\frac{v_\mathrm{y} - l_r\dot{\psi}}{v_\mathrm{x}}\right) \tag{12}$$

When considering larger lateral accelerations, nonlinear tire behavior has to be taken into account, see (Katriniok and Abel, 2011). In literature, the Pacejka Magic Formula tire model (Pacejka and Bakker, 1992) is frequently employed to model the steady-state lateral tire force at pure cornering, i.e. only lateral forces are transmitted at the wheel, as a nonlinear function of the tire sideslip angle

$$F_{\mathrm{y},i} = D_\mathrm{i}\sin\left[C_\mathrm{i}\arctan\left\{B_\mathrm{i}\alpha_\mathrm{i} - E_\mathrm{i}\left(B_\mathrm{i}\alpha_\mathrm{i} - \arctan\left(B_\mathrm{i}\alpha_\mathrm{i}\right)\right)\right\}\right], \quad i \in \{f, r\} \tag{13}$$

In general, this model can also be used to determine the longitudinal tire force as well as the aligning torque. An example of the Pacejka tire model for modeling the side force is shown in Figure 6b.

It can be seen that there are basically three regions in the tire model: one "stable" region of static friction (i.e., region II) and two "unstable" regions of sliding friction (i.e., regions I and III). In particular, the maximum absolute tire force that is physically feasible is limited to $F_{\mathrm{y,i,PMF,max}}$ for a tire sideslip angle of $\alpha_{\mathrm{i,PMF,max}}$ respectively $\alpha_{i,PMF,\min}$. If the absolute sideslip angle increases even more, the tire begins to slide over the road surface such that the resulting absolute tire force is decreased. When comparing both tire models, it can be noticed that the cornering stiffness

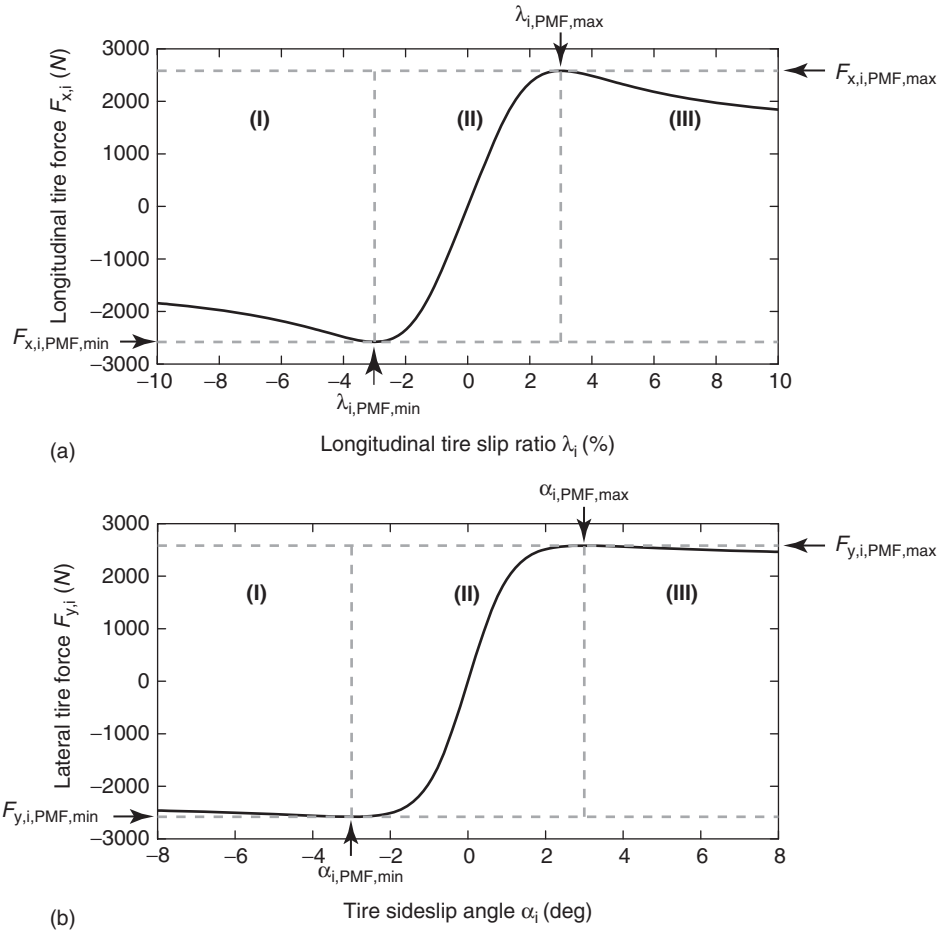


Figure 6. Pacejka Magic Formula tire model.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

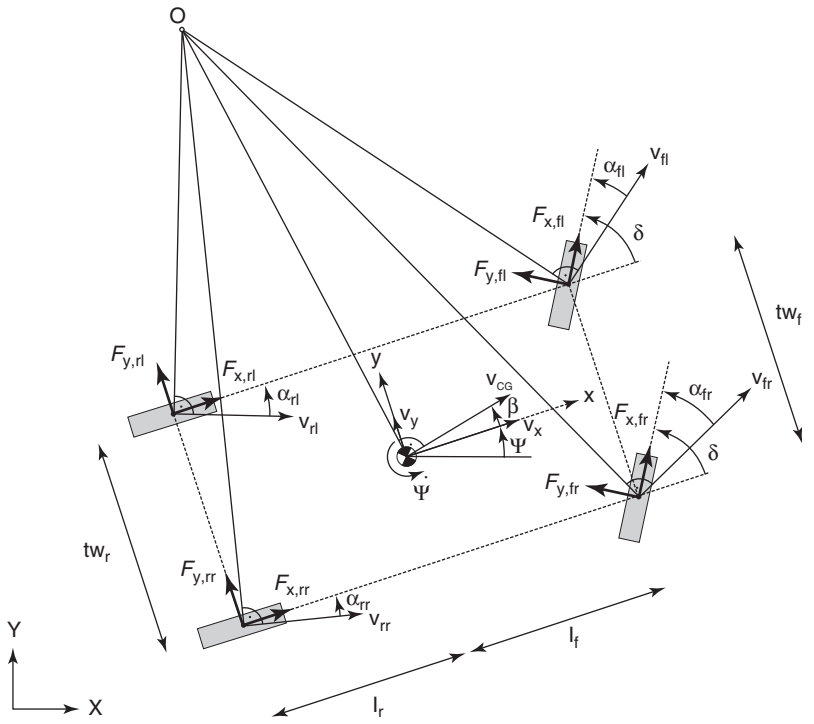


Figure 7. Free body diagram of the two-track vehicle model.

$c_{\alpha,i}$ of the linear tire model corresponds to $dF_{y,i}/d\alpha_i$ for $\alpha_i = 0rad$ with respect to the Pacejka Magic Formula tire model. The same principle, which is illustrated in Figure 6a, holds for the transmission of longitudinal tire forces that basically depend on the longitudinal tire slip λ_i. If the considered application requires that combined slip, that is, transmission of longitudinal and lateral forces at the same time has to be taken into account, there are several approaches to include this issue into the Pacejka tire model, see (Johansson and Gäfvert, 2004). Furthermore, load transfer can also be incorporated into the tire model. Especially, if the chassis application operates at the vehicle handling limits, the application has to be aware of the maximum feasible tire forces that inherently change with the vertical tire load.

3.3.3 Two-track model

So far, this section has mainly focused on chassis applications that are only able to change the vehicle's steering angle. If the simultaneous transmission of longitudinal and lateral tire forces at each wheel has to be considered, for example, for an ESC or torque vectoring, the single-track model is not sufficient to describe horizontal vehicle dynamics in the required level of detail. Therefore, the two-track model that is illustrated in Figure 7 is commonly employed.

In contrast to the single-track model, all the four wheels are considered such that Equations 14–16 contain the forces $F_{x,i}$ (longitudinal) and $F_{y,i}$ (lateral) at each wheel $i \in \{fl, fr, rl, rr\}$. These forces can again be determined by employing one of the tire models described in the previous section. Furthermore, tw_f and tw_r denote, respectively, the front and the rear track widths.

$$m(\dot{v}_x - v_y\dot{\psi}) = \left(F_{x,fl} + F_{x,fr}\right)\cos(\delta) - \left(F_{y,fl} + F_{y,fr}\right)\sin(\delta) + F_{x,fl} + F_{x,rr} \quad (14)$$

$$m(\dot{v}_y + v_x\dot{\psi}) = (F_{y,fl} + F_{y,fr})\cos(\delta) + (F_{x,rl} + F_{x,fr})\sin(\delta) + F_{y,fl} + F_{y,rr} \quad (15)$$

$$J_z\ddot{\psi} = l_f[(F_{y,fl} + F_{y,fr})\cos(\delta) + (F_{x,fl} + F_{x,fr})\sin(\delta)] - l_r(F_{y,rl} + F_{y,rr}) + \frac{tw_f}{2}[(F_{x,fr} - F_{x,fl})\cos(\delta) + (F_{y,fl} - F_{y,fr})\sin(\delta)] + \frac{tw_r}{2}\left(F_{x,rr} - F_{x,rl}\right) \quad (16)$$

3.4 Vertical vehicle dynamics

Finally, this section illustrates the issue of determining a sufficient dynamic plant model considering vertical vehicle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

dynamics. In particular, this section focuses on the application of an active suspension, that is, the vertical movement of the vehicle body (the sprung mass) that results from an uneven pavement should be suppressed. For this purpose, the vehicle and its suspension can be reduced to a quarter car that consists of an unsprung mass m_u—the tire—connected to the road through a spring (i.e., the stiffness of the tire) and a sprung mass m_s—the vehicle body—connected to the unsprung mass through a spring c_s and a damper d_s (i.e., the suspension), see Figure 8. In order to model an active suspension, a force actuator is added between the sprung and the unsprung mass that is able to apply a force F_a to suppress external disturbances of the road.

In this context, z_s denotes the vertical position of the sprung mass, z_u the position of the unsprung mass and z_r the road elevation. According to (Mitschke and Wallentowitz, 2004), the dynamic behavior of the plant can be described with respect to the equilibrium point by the following differential equations

$$m_s\ddot{z}_s = c_s \bullet \left(z_u - z_s\right) + d_s \bullet \left(\dot{z}_u - \dot{z}_s\right) + F_a \qquad (17)$$

$$m_u\ddot{z}_u = c_s \bullet \left(z_s - z_u\right) + d_s \bullet \left(\dot{z}_s - \dot{z}_u\right) - F_a + c_u \bullet \left(z_r - z_u\right) \qquad (18)$$

In Equations 17 and 18, the springs and dampers are modeled as linear elements. If nonlinear effects have to be taken into account, the spring stiffnesses c_u and c_s and the damping ratio d_s have to be modeled as nonlinear functions of the vertical deflection and the vertical velocity. If the consideration of linear elements is sufficient, it inherently depends on the considered chassis application and has to be determined during the development process. If, for example, a controller for suppressing external disturbances of the road by applying the force F_a has been developed using a linear model and tested with satisfying performance on the real plant, nonlinear effects need not be taken into account for controller synthesis. Nevertheless, it might be useful to consider nonlinear effects in the validation model that is used to validate the controller. Finally, it has to be stated that the presented suspension model is well suited for a general analysis of its dynamic behavior, that is, its damping or eigenfrequency. Furthermore, vertical vehicle dynamics can be combined with longitudinal and lateral vehicle dynamics when the presented plant model is combined with single- and two-track models, see (Mitschke and Wallentowitz, 2004).

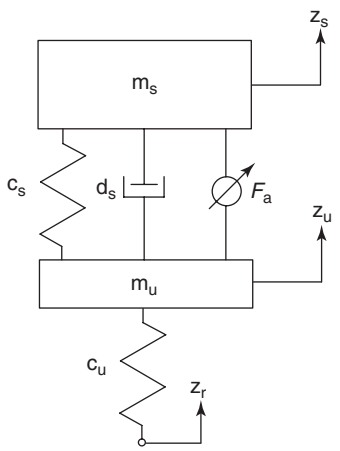


Figure 8. Free body diagram of a quarter car model.

4 USAGE OF DIFFERENT MODELING TECHNIQUES IN COMMON SOFTWARE TOOLS

4.1 Signal-oriented versus object-oriented modeling

Two main modeling concepts are common in many modern simulation tools for simulating dynamic systems: block-oriented (or signal-oriented) and object-oriented modeling. The main differences are found in the way the equations governing the behavior of the system are described and implemented.

A very easy-to-understand method is the block- or signal-oriented approach. For signal-oriented models, cause–effect relationships are modeled directly. A signal-oriented model consists of blocks that are interconnected by signals. The direction of the signals defines causality. Blocks have fixed inputs and outputs, signals only "flow" from block outputs to block inputs; therefore, there will be no feedback from a downstream block to an upstream block if no explicit feedback path is established. In order to be able to create a signal-oriented model, knowledge of the basic governing equations is needed. If an equation contains derivatives, it should be solved for the highest derivative before implemented in order to generate a well-defined block-oriented model. Thus, the lower derivatives can be established by integration of the higher derivatives, preserving causality. For example, if a moving mass is to be considered, its equations of motion should be solved for the acceleration. The velocity is then calculated by integration of the acceleration and the position is calculated by the integration of the velocity.

In Figure 9, an oscillator consisting of two point-shaped masses interconnected by a spring shall serve as an example.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

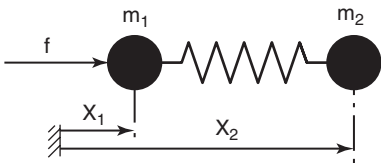


Figure 9. Example: Simple two-mass oscillator.

In order to generate a block model, first the equations of motion of the individual masses are derived:

$$m_1 \bullet \ddot{x}_1 = f_{\text{Spring}} + f$$
$$m_2 \bullet \ddot{x}_2 = -f_{\text{Spring}} \tag{19}$$

Using the equation for the force created by the spring $f = C \bullet (x_2 - x_1)$ and solving for the highest derivates of each state, two differential equations are found:

$$\ddot{x}_1 = \frac{1}{m_1} \bullet \left[f + C \bullet \left(x_2 - x_1 \right) \right]$$
$$\ddot{x}_2 = \frac{1}{m_2} \bullet \left[-C \bullet \left(x_2 - x_1 \right) \right] \tag{20}$$

On this basis, the equations can be implemented using standard blocks, for example, using Simulink. The resulting block diagram can be seen in Figure 10.

As one can see, the structure of block models does not necessarily correspond to the equivalent physical system (Fritzson, 2003); therefore, modeling of complex physical systems can become quite difficult. One main reason for this is the above-mentioned absence of implicit feedback due to the use of unidirectional signals. As of Newton's third law, according to which every action is accompanied by an equal and opposite reaction, almost all real physical systems incorporate a feedback loop that has to be accounted for, which in block-oriented modeling language leads to additional signal paths. Hence, although there is only one physical connection between the two bodies, there are two signal paths - one into and one out of the blocks representing the body. Still, a signal-oriented approach can be a good choice for a physical model. Especially for relatively simple systems, signal-oriented models are easy to understand. Most importantly, however, solving the underlying equations is very straightforward and can be done using a relatively simple ordinary differential equation (ODE) solver. This can especially be useful if a real-time capable model is needed, that is, if the computational effort for simulation needs to be predictable. This is the case here because ODE solvers do not necessarily need to use iterations. Apart from the use as a modeling formulation for physical systems, block models are widely used in control engineering. They are very useful for the description of digital control algorithms, which essentially are calculation instructions and therefore have built-in causality. Especially, as solving the equations is straightforward and well defined, it is possible—under certain restrictions—to automatically generate code from a block diagram that can be implemented on a microcontroller. Therefore, the block-oriented description approach can also be seen as an intermediate step between control engineering theory and controller programming.

Another modeling approach is object-oriented modeling. Here, the very universal object-oriented modeling language "Modelica" shall serve for illustration. A thorough introduction to object-oriented modeling with Modelica can be found in (Tummescheit, 2002). In Modelica,

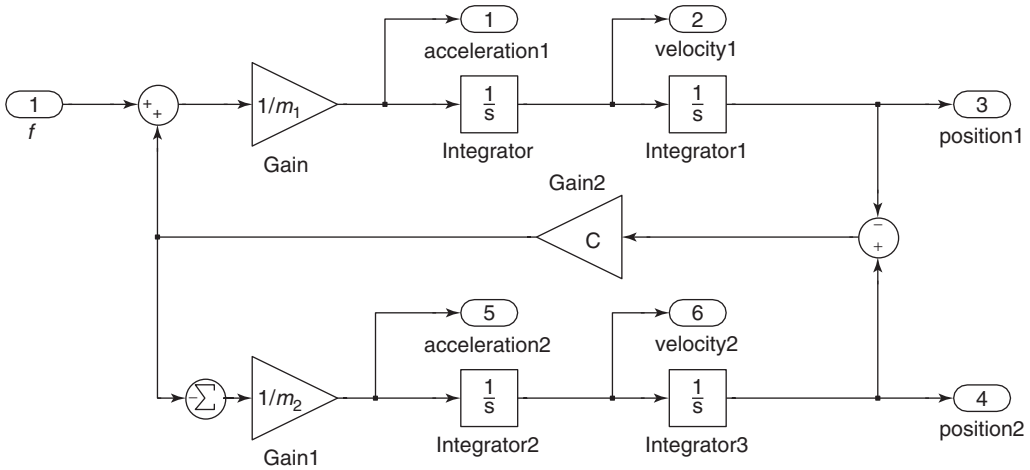


Figure 10. Implementation of two-mass oscillator in Simulink.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 11. Implementation of two-mass oscillator in Dymola.

physical systems are described using so-called objects that can be interconnected. For example, the two-mass oscillator mentioned earlier can be described as a connection between two objects of type "mass" and one object of type "spring." Figure 11 displays the resulting model as implemented in Dymola, a commercial simulation environment for simulation of Modelica models. All objects can have states, parameters, and equations that contribute to the overall model. They also have connectors that define interfaces to other objects. For example, in this case, each mass object has a parameter containing its mass as well as state variables containing its speed and location. The spring is described by an equation that relates its deflection to the force it exerts on its connected objects.

Unlike signals, which have a fixed direction, interconnectors establish additional relationships between connected elements, but do not yet determine the direction of causality. The connector equations account for both cause and effect, so it is not necessary to explicitly incorporate the resulting opposite reaction. Also, conservation of energy is automatically accounted for, greatly simplifying the task of creating models that are physically correct.

An important aspect of object-oriented programming is the concept of classes and inheritance. A class is a description of a certain type of object. For example, the two masses in the example are both of the class "point mass," both having the same structure. So the actual objects are instances of that class which only has to be defined once, leading to high reusability. In addition, class properties can also be inherited by other classes. For example, going back to the case of the quarter-car model, a "tire" class could be created. A tire would contain an unsprung mass of the type mass as well as a spring. Therefore, the class "tire" automatically inherits the states of the tire (velocity, position) and the equation for the spring, but for the end user can be displayed as a tire. It is also possible to incorporate interchangeable components. For example, it would be possible to give the user a choice of springs to simulate the behavior of different types of tires.

Especially when the system to be modeled becomes more complex, high level languages such as Modelica can be advantageous. However, it should be noted that a Modelica model describes physical systems but not the simulation process. Therefore, additional software is required to simulate the model using the equations specified in Modelica. Modelica models can incorporate differential, algebraic, and discrete equations (Tummescheit, 2002). Their solution can require high computational effort and iterative solution methods. Thus, the models may not always be suitable for real-time computation.

4.2 Modeling example in Matlab/Simulink, and Dymola

4.2.1 Implementation in Simulink

In this section, the implementation of a plant model to be used for numerical simulations will be illustrated using Matlab/Simulink. For this purpose, the application of an active suspension, introduced in Section 3.4, is considered again. According to Section 3.4, the dynamic behavior of the plant is described by the following differential equations

$$m_s\ddot{z}_s = c_s \bullet (z_u - z_s) + d_s \bullet (\dot{z}_u - \dot{z}_s) + F_a \tag{21}$$

$$m_u\ddot{z}_u = c_s \bullet (z_s - z_u) + d_s \bullet (\dot{z}_s - \dot{z}_u) - F_a + c_u \bullet (z_r - z_u) \tag{22}$$

One possibility to implement the plant model (Equations 21 and 22) in Matlab/Simulink is to use integrator, gain, and sum blocks that are part of the Simulink library. Figure 12 shows a possible realization in Matlab/Simulink.

Equation 21, describing the vertical movement of the sprung mass, is modeled in the upper part of the Simulink model; the coupling of the two masses through the spring and the damper in the middle and the vertical movement of the unsprung mass, defined by (22), is modeled at the bottom. Considering the sprung mass, the acceleration $\ddot{z}_s$ is computed by dividing the sum of all forces acting on m_s by the sprung mass m_s. Subsequently, the velocity $\dot{z}_s$ and the position z_s of the sprung mass are determined using two integrators. As the spring and damper forces are proportional to the relative distance and velocity of the two masses and the road, these forces are modeled using the sum as well as gain blocks. The system inputs, that is, the road's elevation z_r as well as the actuator force F_a, and the considered system outputs, that is, the vertical velocity and position of both masses, can be recognized as sources and sinks.

In order to obtain a clearly arranged implementation, Simulink allows for creating subsystems. Thus, the plant can be reduced to its inputs and outputs as illustrated in Figure 13.

Another possibility to implement the quartercar in Matlab/Simulink in a clearly arranged way is to use the *state space*-block, which belongs to the Simulink block

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

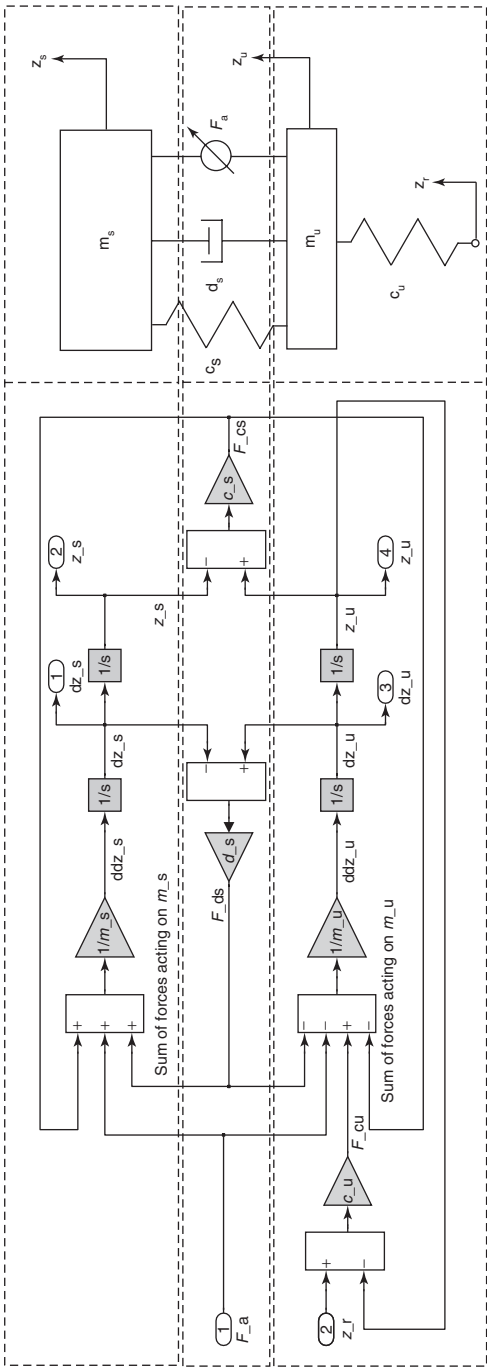


Figure 12. Simulink implementation of the quartercar example using single blocks.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

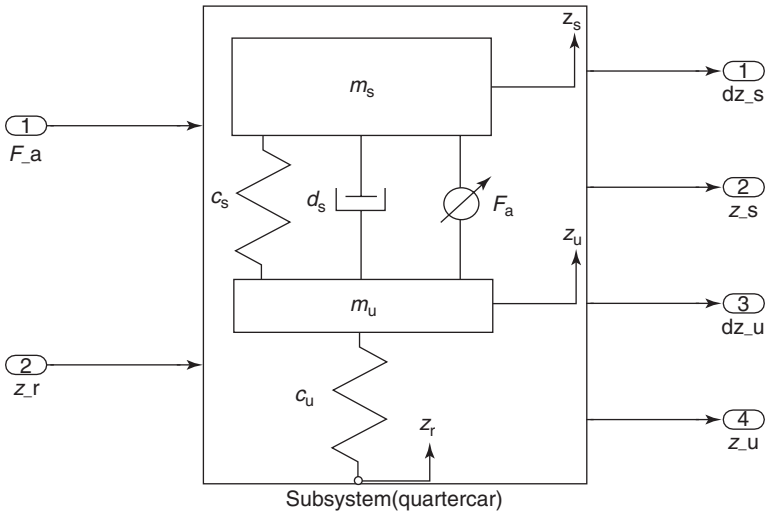


Figure 13. Simulink implementation of the quartercar example using subsystems.

library and supports linear time-invariant systems having the structure

$$\dot{\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{u} \tag{23}$$

$$\boldsymbol{y} = \mathbf{C}\boldsymbol{x} + \mathbf{D}\boldsymbol{u} \tag{24}$$

Using $\mathbf{x} = [\dot{z}_\mathrm{s}\ z_\mathrm{s}\ \dot{z}_\mathrm{u}\ z_\mathrm{u}]^\mathrm{T}$ as state vector, the actuator force $u = F_\mathrm{a}$ as control input, the road elevation $z = z_\mathrm{r}$ as system disturbance, and $\mathbf{y} = [\dot{z}_\mathrm{s}\ z_\mathrm{s}\ \dot{z}_\mathrm{u}\ z_\mathrm{u}]^\mathrm{T}$ as system output, Equations 21 and 22 can be rewritten as

$$\begin{bmatrix} \ddot{z}_\mathrm{s} \\ \dot{z}_\mathrm{s} \\ \ddot{z}_\mathrm{u} \\ \dot{z}_\mathrm{u} \end{bmatrix} = \begin{bmatrix} -\frac{d_\mathrm{s}}{m_\mathrm{s}} & -\frac{c_\mathrm{s}}{m_\mathrm{s}} & \frac{d_\mathrm{s}}{m_\mathrm{s}} & \frac{c_\mathrm{s}}{m_\mathrm{s}} \\ 1 & 0 & 0 & 0 \\ \frac{d_\mathrm{s}}{m_\mathrm{u}} & \frac{c_\mathrm{s}}{m_\mathrm{u}} & -\frac{d_\mathrm{s}}{m_\mathrm{u}} & -\frac{c_\mathrm{s}+c_\mathrm{u}}{m_\mathrm{u}} \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \dot{z}_\mathrm{s} \\ z_\mathrm{s} \\ \dot{z}_\mathrm{u} \\ z_\mathrm{u} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} F_\mathrm{a} + \begin{bmatrix} 0 \\ 0 \\ \frac{c_\mathrm{u}}{m_\mathrm{u}} \\ 0 \end{bmatrix} z_\mathrm{r} \tag{25}$$

$$\mathbf{y} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{z}_\mathrm{s} \\ z_\mathrm{s} \\ \dot{z}_\mathrm{u} \\ z_\mathrm{u} \end{bmatrix} \tag{26}$$

As described earlier, only linear dynamic systems having the structure as in (Equation 23) can be modeled using the state space block. Thus, z_r has to be defined as an input such that Equation 25 can be written as

$$\begin{bmatrix} \ddot{z}_\mathrm{s} \\ \dot{z}_\mathrm{s} \\ \ddot{z}_\mathrm{u} \\ \dot{z}_\mathrm{u} \end{bmatrix} = \begin{bmatrix} -\frac{d_\mathrm{s}}{m_\mathrm{s}} & -\frac{c_\mathrm{s}}{m_\mathrm{s}} & \frac{d_\mathrm{s}}{m_\mathrm{s}} & \frac{c_\mathrm{s}}{m_\mathrm{s}} \\ 1 & 0 & 0 & 0 \\ \frac{d_\mathrm{s}}{m_\mathrm{u}} & \frac{c_\mathrm{s}}{m_\mathrm{u}} & -\frac{d_\mathrm{s}}{m_\mathrm{u}} & -\frac{c_\mathrm{s}+c_\mathrm{u}}{m_\mathrm{u}} \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \dot{z}_\mathrm{s} \\ z_\mathrm{s} \\ \dot{z}_\mathrm{u} \\ z_\mathrm{u} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ -1 & \frac{c_\mathrm{u}}{m_\mathrm{u}} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} F_\mathrm{a} \\ z_\mathrm{r} \end{bmatrix} \tag{27}$$

With this modification, Equations 25 and 26 can be implemented in Simulink as illustrated in Figure 14.

The overview that has been given earlier is just a compendium of the possible modeling techniques that can be employed in Simulink. The particular implementation depends on the complexity of the considered plant. If, for example, a model of an entire car is implemented, it is reasonable to implement, for example, the vehicle

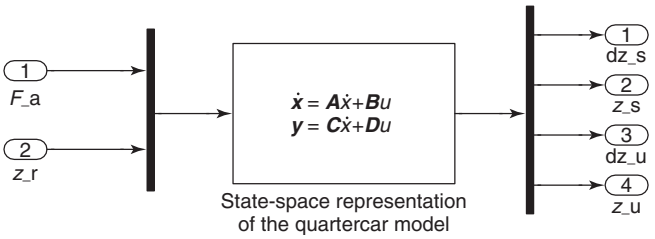


Figure 14. Simulink implementation of the quartercar using a state space representation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

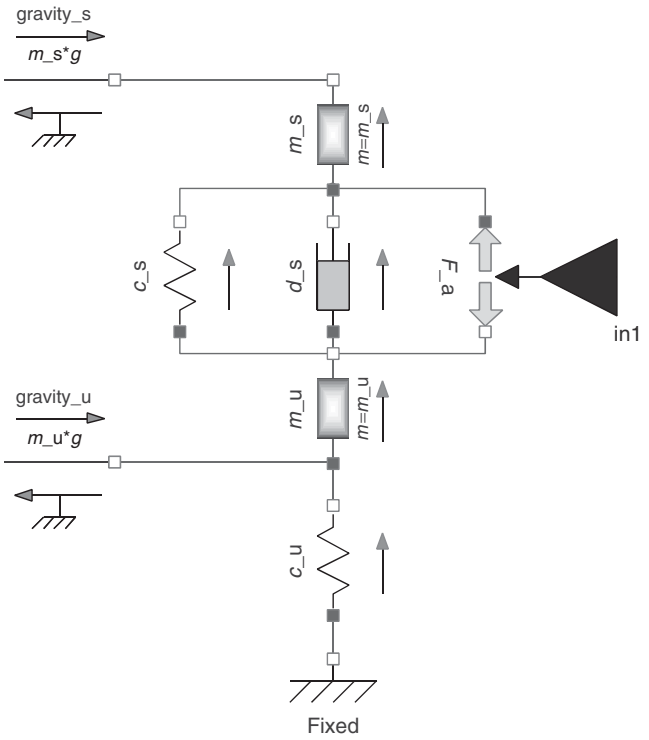


Figure 15. Dymola implementation of the quartercar example.

body, the suspension, and the power train as interconnected subsystems.

As mentioned in Section 3.1, plant models are required and employed for different purposes. For controller design, the model that has been described earlier has an appropriate complexity to be used for a first analysis of the dynamic plant behavior as well as controller design. For the purpose of controller validation, plant models of higher complexity that consider nonlinear effects are commonly employed. In this regard, it is reasonable to run co-simulations of Simulink and third-party software tools. As far as vehicle dynamic control applications are concerned, IPG CarMaker and CarSim^{TM}are two powerful tools that provide several vehicle models as well as validated tire datasets to be used for validation purposes. In this context, both software packages allow for MIL as well as HIL testing procedures.

4.2.2 Implementation in Dymola

The chapter concludes with an implementation of the quartercar model in Dymola (Figure 15). Using blocks that are part of the free Modelica Standard Library Modelica Association (1998–2008), the model can be assembled using the predefined components for the masses, springs, and the damper as well as the interface elements "force" and "force2." The interface element "force"—used here to account for gravity—applies an additional force to its connected element, whereas "force2" creates a generic force between two elements and is used here to introduce the actuator force acting between sprung and unsprung mass. As one can see, the structure of the model is very close to the structure of the physical system it describes, making it more accessible. This is mainly due to the above-mentioned use of connectors as opposed to unidirectional signals. Still, it is also possible to incorporate model parts that have fixed causality. For example, a control algorithm calculating the desired actuator force can be implemented using the signal input "in1."

RELATED ARTICLES

Control Systems and Strategies for Automated Manual and Double Clutch Transmissions
Chassis modeling and optimization by the advanced method ABE
System Simulation in Dshplus, What Applications are Possible?
Torque Vectoring by Drive Train Systems
Global Chassis Control in Passenger Cars
Chassis Control Systems—A Look into the Future
Hardware-in-the-Loop Simulation
Vehicle Safety, Functional Safety, OBD Diagnosis
Applications—Intelligent Vehicles: Autonomous Vehicles

REFERENCES

Abel, D. (2005) *Rapid Control Prototyping*, Springer.

Fritzson, P. (2003) *Principles of Object-Oriented Modeling and Simulation with Modelica 2.1*, Wiley-IEEE Press.

Ha, J., Tugcu, A.K. and Boustany, N.M. (1989) Feedback linearizing control of vehicle longitudinal acceleration. *IEEE Transactions on Automatic Control*, **34**, 689–698.

Johansson, B. and Gäfvert, M. (2004) Untripped SUV rollover detection and prevention. IEEE Conference on Decision and Control, pp. 5461–5466.

Katriniok, A. and Abel, D. (2011) LTV-MPC approach for lateral vehicle guidance by front steering at the limits of vehicle dynamics. In IEEE Conference on Decision and Control and European Control Conference, pp. 6828–6833.

Keßler, G.C., Maschuw, J.P., Zambou, N. and Bollig, A. (2007) Concept for the generation of reference variables and model-based predictive control for the lateral guidance of heavy-duty vehicle platoons. *Automatisierungstechnik*, **55** (6), 298–305.

Lu, X.Y. and Hedrick, J.K. (2004) Practical string stability for longitudinal control of automated vehicles. *Vehicle Systems Dynamics Supplement*, **41**, 577–586.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Mitschke, M. and Wallentowitz, H. (2004) *Dynamik der Kraftfahrzeuge*, Springer, Heidelberg.

Modelica Association (1998–2008) *Modelica Standard Library*, Modelica Association.

Pacejka, H. and Bakker, E. (1992) The magic formular tire model. *Vehicle System Dynamics: International Journal of Vehicle Mechanics and Mobility*, **21** (1), 1–18.

Rajamani, R. (2006) *Vehicle Dynamics and Control*, Mechanical Engineering Series Edition, Springer, New York.

Tummescheit, H. (2002) *Design and Implementation of ObjectOriented Model Libraries using Modelica*. Department of Automatic Control, Lund Institute of Technology.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto035
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Chassis Modeling and Optimization by the Advanced Method ABE

Ingo Albers

ZF Friedrichshafen AG, Dielingen, Germany

1 Introduction 1
2 Theoretical Basics on Independent Wheel Suspensions 2
3 Review 9
References 9

1 INTRODUCTION

Wheel suspensions have the important task to serve as a connection between the road surface and the chassis, guiding the wheel while driving. The suspension has to fulfill many tasks: cushioning, damping, steering, force transmission, powering, and braking.

The way in which a wheel is guided significantly determines the kinematics of the system and hence which type of movement the wheel is capable of doing. One differentiates between inflexible (or rigid) and elastic kinematics. With the inflexible kinematics, it is given that all connecting components such as suspension arms and joints are ideally stiff and the joints only show one or more pure rotation degrees of freedom toward the desired directions. In reality, the elements of wheel suspensions are seldom considered to be ideally stiff. Especially, elastomer bushings have a considerable effect on the kinematics of the wheel suspensions. In addition, structural elasticity of the involved elements is to be considered to precisely analyze the kinematics. The kinematics of a wheel under the influence of a real element rigidity and the so-called soft elements such as rubber bushings, overload springs, or bump stops is usually called *elasto-kinematics*. The elasto-kinematics can be differentiated drastically from the idealized stiff-kinematics, still it is highly suggested that a clear analysis and synthesis of the stiff-kinematics is performed before looking at the real elasto-kinematics.

To make the kinematics of wheel suspensions tangible and especially comparable, numerous values on axle kinematics have been developed in the past, which have established themselves in the daily life of chassis engineers. This applies to both single values, valid for the so-called design position, and complete characteristic data. These trends are identified and evaluated for compression and rebound and steering or action of forces – or torque conditions (braking, powering, cornering, and vertical loads).

Nowadays, within the automotive industry, usually a first model in 3D-CAD and a model in the multibody system (MBS) are defined when designing the layout of an axle. For a fast prelayout without MBS knowledge, for university purposes, for Formula Student suspensions, or in all other cases in which an MBS model is not feasible (e.g., due to high fees), a simple program can be self-created to calculate the stiff-kinematics. To do so, the mathematical approach by Matschinsky (1992) is used, which is complementary to a program family called *ABE* (axle calculation in Excel) (Albers, 2003, 2009).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto037
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 THEORETICAL BASICS ON INDEPENDENT WHEEL SUSPENSIONS

2.1 Necessary degrees of freedom of a wheel suspension

Fundamentally, a stiff body has six degrees of freedom in space and can perform a translational movement in *X*-, *Y*-, or *Z*-direction or a rotational movement around every axis of a Cartesian *X*-*Y*-*Z* coordinate system.

As road roughness has to be compensated and high accelerations of the vehicle body have to be avoided, an independent suspension basically has to have a vertical movement option. The degree of freedom is, thus, $F = 1$. This degree of freedom does not necessarily have to be a parallel movement, but can generally also be a combined stroke, cross, and tilt movement. With such a combination of translational and rotational moving parameters, all parameters are in direct relation to each other, the so-called kinematic constrained motion. A body suspension spring is responsible for saving and releasing the energy. Simultaneously, a damper is dissipating energy and reducing vibrations. The suspension system adjusts to the statically defined weight balance of the vehicle. The spring element represents a highly elastic suspension arm; the degree of freedom $F = 1$ is, thus, appropriate only if a force is applied (Albers, 2003).

If a body is stiffly hung to six suspension arms in a room, which illustrates a wheel with a wheel bearing solidly attached to the vehicle. To maintain the degree of freedom of $F = 1$, one of the six suspension arms has to be removed and one gets the basic type of an independent suspension, the so-called common space 5 rod wheel guidance. The calculation program described here is, thus, only applicable to independent suspension systems, which can be illustrated or be derived from the "common 5 rod suspension." This, for example, also includes the double wishbone axle as well as the common McPherson axle at the front (Albers, 2003).

The guidance of the suspension and definition of the degrees of freedom takes place with the components shown in the following.

2.2 Components of wheel suspensions

A wheel suspension can be considered as a kinematic chain. The wheel hub is the connecting rod. Suspension arms serve as intermediate components of the chain. Joints are the smallest elements of such a chain. The total degree of freedom results from the degrees of freedom of the bodies to be guided, all joints as well as the confining degrees of freedom of the suspension arms. In the following, the different types of joints are shown in Figure 1. How many degrees of freedom is required by the respective joint constrain will be explained. A small "f" identifies the joint degrees of freedom (Albers, 2003).

2.3 Characteristic values of axle kinematics

In the automotive literature, there are numerous different vehicle suspension characteristic values. Matschinsky (1992) describes them in great detail in his PhD thesis, thus only a small overview is given in Tables 1 and 2.

2.4 Velocity status of the wheel carrier

Looking at the respective definitions of the kinematic values shown (Matschinsky, 1992; Albers, 2003), it becomes evident that they can be calculated through either geometrical positions or translational as well as rotational velocity conditions of either the tire contact point or the wheel carrier. The complete velocity status consists of the translational velocity vector $\vec{v}_M$ and the angular velocity vector $\vec{\omega}_K$. The translational velocity is only valid in the reference

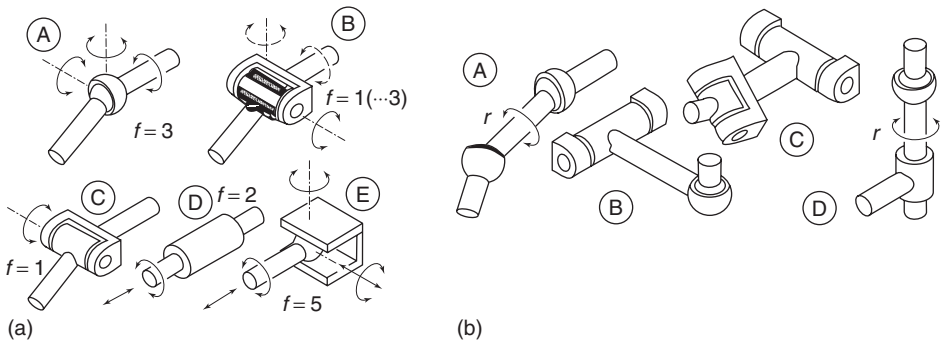


Figure 1. Overview of (a) joint types and (b) steering types. (Reproduced with permission from Matschinsky, 1992. © W. Matschinsky.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto037
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Overview of usual characteristics during compression and rebound

Characteristics During Compression and Rebound	
Basic characteristics	Roll center
Spring travel (mm)	Roll center height (vertical movement) (mm)
Wheel travel (mm)	Roll center position (roll movement) (mm)
Spring ratio (spring travel/wheel travel) (−)	Steering axle/castor axle
Track width (mm)	Castor angle (deg)
Wheel position characteristics	Castor trail (mm)
Toe in (deg)	Castor offset (mm)
Camber angle (deg)	Steering axle/king pin angle
Antidive/antisquat	King pin inclination angle (deg)
Real brake supporting angle (deg)	King pin offset (mm)
Optimal brake supporting angle (deg)	Disturbance force lever (DFL)
Brake pitch compensation (%)	DFL braking (mm)
Real acceleration support angle (deg)	DFL traction (mm)
Optimal acceleration support angle (deg)	Other characteristics
Drive pitch compensation (%)	Suspension oblique angle (deg)
Acceleration balance (%)	Brake support angle (deg)

Reproduced with permission from Albers, 2009. © Ingo Albers.

Table 2. Overview of usual characteristics during steering

Characteristics During Steering	
Basic characteristics	Steering axle/castor axle
Tie rod travel (mm)	Castor angle (deg)
Wheel steering angle (deg)	Castor trail (mm)
Track width (mm)	Castor offset (mm)
Wheel position characteristics	Castor angle (deg)
Camber angle (deg)	Steering axle/king pin angle
Wheel travel (mm)	Castor offset (mm)
Steering characteristics	Steering axle/king pin angle
Wheel steering angle inside (deg)	King pin inclination angle (deg)
Wheel steering angle outside (deg)	King pin offset (mm)
Averaged wheel steering angle (deg)	Disturbance force lever
Toe difference angle (deg)	Disturbance force lever braking (mm)
Ackermann angle (outside) (deg)	Disturbance force lever traction (mm)
Ackermann percentage (%)	

Reproduced with permission from Albers, 2009. © Ingo Albers.

point, which, reasonably, is the center of the wheel (M). Within a solid body, the angular velocity vector is identical, thus it is defined for the wheel carrier body (K) (Albers, 2009).

The wheel carrier including the tire contact point, which is rigidly coupled with the wheel carrier over the wheel, is abstracted by a wheel carrier level, which, at the beginning, is defined by the tire contact point (A), the wheel center (M), as well as a once defined auxiliary point (H) (Figure 2). The auxiliary point is conveniently defined as being part of the wheel axle; in addition, the lines MA and MH are perpendicular to each other. The distance between H and M is simply the wheel radius, hence points A, M, and H construct an orthogonal, equally sided triangle in space (Albers, 2009).

This plane is being incrementally moved in spring and steering activity. In doing so, the kinematic velocity is being analyzed. Through summation by integration of the incremental speeds, the position is also always known.

As an example for a position-orientated calculation, the toe and camber calculation is shown. The track width is a value, which is trivially determined from the positions of the tire contact points. For the identification of the geometrical constraints of points A, M, and H, it is advisable to show a wheel in a random position and then examine it in two layers. It is very important to determine the respective distances correctly in their respective depiction in the layer. Shown in Figure 2 is a wheel with a positive camber and a positive wheel steer angle. A positive wheel steer angle at the left wheel means negative toe-in (Albers, 2009).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto037
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

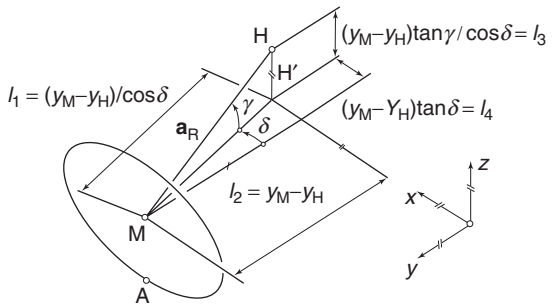


Figure 2. Wheel and wheel carrier in optional, steering situation. (Reproduced from Matschinsky, 1998. With kind permission from Springer Science+Business Media.)

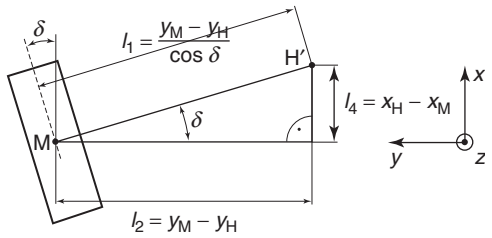


Figure 3. View from above in *X*-*Y*-level.

Viewed from above, the picture is shown in the *X*-*Y*-level, cut at the wheel center M (Figure 3).

The lines named in Figure 2 can easily be found here again and can be depicted geometrically through known values. "H′" is the vertical projection of the point H in the shown level. The respective sections are trigonometrically further explained in the figures.

In the following, the view from the rear upon the wheel axle a_R is illustrated as a cut through the wheel center M and the help point H (Figure 4). Hence, the cut level shown is the *Y*-*Z*-level, turned by the steering angle δ.

With the help of these geometrical correlations, camber- and toe-in with respect to the front wheel steering angle can be determined directly. For instance, when the coordinates of the points M and H of the wheel axle a_R are known, one can calculate the steering angle δ as well as the camber angle γ. From Figures 3 and 4, one gets the following correlation for the steering angle:

$$\delta = \arcsin \frac{x_H - x_M}{\sqrt{(x_H - x_M)^2 + (y_M - y_H)^2}} \qquad (1)$$

The toe-in angle for a left wheel is, according to the preceding definition (DIN 70000), obtainable by switching the algebraic sign. The camber angle can be obtained

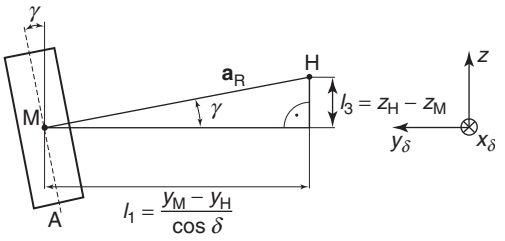


Figure 4. View from behind on the *Y*-*Z*-level of the left wheel.

likewise through the following equation:

$$\gamma = \arctan \frac{z_H - z_M}{\sqrt{(x_H - x_M)^2 + (y_M - y_H)^2}} \qquad (2)$$

In this case, the current coordinates of the wheel center point as well as of the help point H on the wheel axle are needed (Albers, 2009).

2.5 The ABE-core: determining the velocity situation

Now, it is possible to establish the core of the calculation program. Considering a solid body, such as a wheel carrier of a wheel suspension in the case given, its velocity situation is explicitly described by specifying the velocity $\vec{\mathbf{v}}_M$ in the wheel center point and $\vec{\boldsymbol{\omega}}_K$ of the wheel carrier; hence, it consists of six components:

$$\vec{\mathbf{v}} = (\vec{\mathbf{v}}_M, \vec{\boldsymbol{\omega}}_K)^T = (\mathbf{v}_{Mx}, \mathbf{v}_{My}, \mathbf{v}_{Mz}, \boldsymbol{\omega}_{Kx}, \boldsymbol{\omega}_{Ky}, \boldsymbol{\omega}_{Kz})^T \qquad (3)$$

Thus, for a clear calculation, six independent equations are needed as well. The translational velocity of any random point P_i in a body can be calculated directly if the distance between P_i to its point of reference (in this case the wheel center point M) is known as the *respective distance vector* $\mathbf{r}_i$. The current velocity situation $\mathbf{v}_i$ of a random point P_i is calculated through

$$\vec{\mathbf{v}}_i = \vec{\mathbf{v}}_M + \vec{\boldsymbol{\omega}}_K \times \vec{\mathbf{r}}_i \qquad (4)$$

Figure 5 shows the necessary components.

With the calculation algorithm given, it is assumed that the connection rods of the suspension are stiff. Thus, the wheel-carrier- and the vehicle-based velocity can be correlated to each other. The velocity in the direction of the connection rod has to be identical on both sides. This is illustrated with a simple scheme (Figure 6; Albers, 2009).

One can clearly see that there are different velocities $\vec{\mathbf{v}}_i$ and $\vec{\mathbf{v}}_i'$ effective on the wheel-carrier-based point *i* and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto037
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

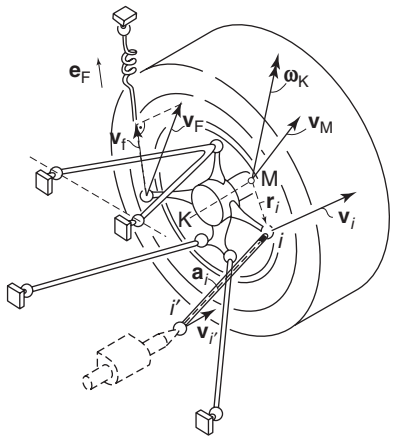


Figure 5. Common velocity situation of a wheel carrier. (Reproduced with permission from Matschinsky, 1992. © W. Matschinsky.)

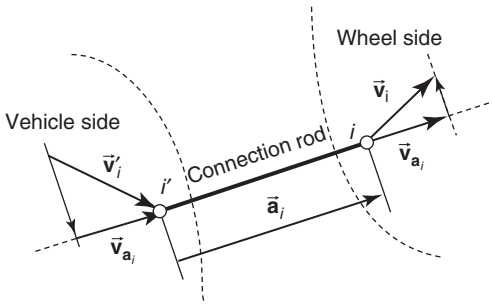


Figure 6. Identical velocity along a connection rod vector. (Reproduced with permission from Albers, 2009. © Ingo Albers.)

on the vehicle-based point i' of a stiff connecting rod, the velocity in the direction of the rod ($\vec{\mathbf{v}}_{\mathbf{a}_i}$), however, has to be identical. The scalar velocity in the direction of the rod results in the common case from the scalar product from velocity vector and steering rod vector, which points from the vehicle to the wheel carrier:

$$\vec{\mathbf{v}}_i \cdot \vec{\mathbf{a}}_i = \vec{\mathbf{v}}'_i \cdot \vec{\mathbf{a}}_i \tag{5}$$

This equation is called the *connection rod requirement*. The connection rod vector $\vec{\mathbf{a}}_i$ is, just like the location vector $\vec{\mathbf{r}}_i$, known from the original coordinates of the wheel suspension. With Equations 4 and 5, the fundamental equation for this discussion is as follows:

$$(\vec{\mathbf{v}}_M + \vec{\boldsymbol{\omega}}_K \times \vec{\mathbf{r}}_i) \cdot \vec{\mathbf{a}}_i = \vec{\mathbf{v}}'_i \cdot \vec{\mathbf{a}}_i \tag{6}$$

which puts the velocity situation of the wheel carrier in correlation to the wheel-carrier- and vehicle-based joints' velocities at the ends of the examined connection rod. This fundamental equation is called *common velocity equation* (CVE) and is an elementary core of the calculation program. The CVE is being used in many different forms and adjusted according to the specific needs (Albers, 2009).

The CVE, formulated for three different basic suspension variants, is as follows:

- Common five-rod suspension with spring element on the wheel carrier (5LR)
- Common five-rod suspension with spring element at a connection rod (5LL)
- McPherson strut (McP)

As an example, the CVE is developed and drafted for a five-rod suspension with connection of the spring element to the connection rod (5LL). For both the other basic types (5LR and McP), refer to the study by Albers (2009).

The CVE originates in its basic form (Matschinsky, 1992) but is further developed to adjust to other applications. Hence, the CVE is formulated in such a way that the spring element incrementally varies its length or the tie rod is incrementally shifted from the vehicle side. The two following moving patterns can be observed:

- *Spring action*: Change in length of the spring elements.
- *Steering action*: Moving of the vehicle-based tie rod joint (equals the connection to a rack and pinion gear box).

In between these two moving actions, on spring movements, the difference is not only that a different spring rod is affected than that on the steering (tie rod) but especially that the length of the spring rod is changed through the stepwise speed increment, whereas upon steering, the length of the tie rod is constant. Here, the steering speed increment is externally impacted, that is, from the vehicle-based side (from a translational operating steering gear box) to the wheel suspension system (Albers, 2009).

For both these common types of movements of the wheel suspension, which can perform independently from each other, the CVE will be developed in such a way that it can be used in a calculation algorithm.

2.5.1 *Derivation of the ABE-core of a wheel suspension system*

With a common five-rod suspension, as shown in Figure 5, the CVE can be freely set up for a spring movement with the indices 1 to 5 for the first five connection rods. The entire derivation originates from the study by Albers (2009).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto037
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The necessary sixth equation generally evolves from contemplating the spring element. This spring element is considered to be stiff in each incremental move and completes the linear equation system with the sixth equation. This spring equation has to be compiled independently from the type of movement.

In case of a spring movement, the spring element is moved out of its design position with the spring speed $\mathbf{v}_\mathrm{f}$. Principally, this spring speed increment $\mathbf{v}_\mathrm{f}$ is responsible for a change in position of the wheel suspension, as it induces the velocity $\vec{\mathbf{v}}$ of the wheel carrier. This again results in a velocity $\mathbf{v}_i$ on all wheel-carrier-based joints, which can be calculated with Equation 4. The wheel suspension system moves, that is, it experiences a change in position. The tie rod does not move in this case, it is said to be a "blocked steering."

A similar consideration is valid for the steering process at which the strut compression velocity $\mathbf{v}_\mathrm{f}$ is put to zero. This activity is called a *blocked suspension*. The tie rod is moved at $\mathbf{v}_\mathrm{L}$ and it induces a movement of the wheel carrier and, thus, the entire wheel suspension.

Generally speaking, at first, a scalar strut compression velocity $\mathbf{v}_\mathrm{f}$ is defined, which points in the direction of the unit vector of the spring and is positive during compression from the design position. Put into the CVE according to Equation 4, the result is

$$\mathbf{v}_\mathrm{f} = (\vec{\mathbf{v}}_\mathrm{M} + \vec{\boldsymbol{\omega}}_\mathrm{K} \times \vec{\mathbf{r}}_\mathrm{F}) \cdot \vec{\mathbf{e}}_\mathrm{F} \tag{7}$$

with $\vec{\mathbf{r}}_\mathrm{F}$ (distance vector from the wheel center to the spring function line) and $\vec{\mathbf{e}}_\mathrm{F}$ (unit vector of the spring element). The connection rod vector of the spring is named with $\vec{\mathbf{a}}_6$ and points, as defined earlier, from the vehicle structure to the wheel carrier and to the pivot joint of one of the connecting rods, and thus contrarily to $\mathbf{v}_\mathrm{f}$.

The case that the spring element is mounted directly to a wheel carrier can be expressed more easily. Here, a more complicated case with the connection of the spring to the connection rod is to be shown. It can easily be imagined that in such a case the connection point between the spring element and the connection rod (suspended rod) does not instantaneously follow the velocity situation of the wheel carrier but receives a coupled movement via the suspended rod. It is, thus, important for the development of the spring equation to include the information on where the spring element is mounted. For the derivation of the ABE-core, the following steering indices are defined: tie rod (1), spring element (6), suspended rod (5), as well as the remaining three connection rods (2, 3, and 4). Figure 7 shows the velocity plan of the suspended connection.

It is possible to apply the intercept theorem here to determine the relation of the values of the connection rod

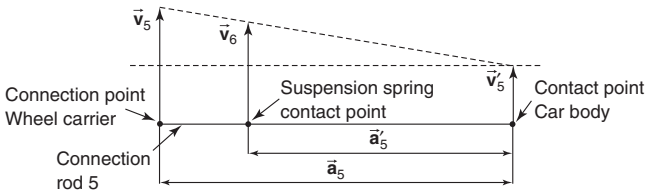


Figure 7. Application of the intercept theorem with a spring, acting to the connection rod. (Reproduced with permission from Albers, 2003. © Ingo Albers.)

vector $\vec{\mathbf{a}}_5$ as well as $\vec{\mathbf{a}}'_5$. The vector $\vec{\mathbf{a}}'_5$ is the connecting vector from the vehicle-based joint to the point of action of the spring at point 6. It is $\mathbf{a}_5 = |\vec{\mathbf{a}}_5|$ as well as $\mathbf{a}'_5 = |\vec{\mathbf{a}}'_5|$. The distance relation is viewed according to amount and converted to

$$\mathbf{v}_6 = \mathbf{v}'_5 + \frac{\mathbf{a}'_5}{\mathbf{a}_5}.(\mathbf{v}_5 - \mathbf{v}'_5) \tag{8}$$

With an introduced leverage $\mu_\mathrm{HBV} = \frac{\mathbf{a}'_5}{\mathbf{a}_5}$, it comes out to

$$\mathbf{v}_6 = \mathbf{v}'_5 + \mu_\mathrm{HBV}.(\mathbf{v}_5 - \mathbf{v}'_5) \tag{9}$$

As the spring compression speed $\mathbf{v}_\mathrm{f}$ is divergent to the connection rod vector, it is

$$\vec{\mathbf{v}}_6 \cdot \vec{\mathbf{a}}_6 = -\mathbf{v}_\mathrm{f} \cdot |\vec{\mathbf{a}}_6| \tag{10}$$

Fundamentally, for a spring element it is, according to the connection rod requirement equation 5

$$\vec{\mathbf{v}}_6 \cdot \vec{\mathbf{a}}_6 = \vec{\mathbf{v}}'_6 \cdot \vec{\mathbf{a}}_6 \tag{11}$$

In combination with Equations 9 and 10, this results in

$$[\vec{\mathbf{v}}'_5 + \mu_\mathrm{HBV} \cdot (\vec{\mathbf{v}}_5 - \vec{\mathbf{v}}'_5)] \cdot \vec{\mathbf{a}}_6 = -\mathbf{v}_\mathrm{f} \cdot |\vec{\mathbf{a}}_6| \tag{12}$$

These are as usual the common requirements of the superposition of translational and rotational speeds $\vec{\mathbf{v}}_5 = \vec{\mathbf{v}}_\mathrm{M} + \vec{\boldsymbol{\omega}}_\mathrm{K} \times \vec{\mathbf{r}}_5$. With this information, the spring equation can be formulated step by step for the ABE-core from Equation 12:

$$\begin{aligned} &\Rightarrow [\vec{\mathbf{v}}'_5 + \mu_\mathrm{HBV} \cdot (\vec{\mathbf{v}}_\mathrm{M} + \vec{\boldsymbol{\omega}}_\mathrm{K} \times \vec{r}_5 - \vec{\mathbf{v}}'_5)] \cdot \vec{\mathbf{a}}_6 \\ &= -\mathbf{v}_\mathrm{f} \cdot |\vec{\mathbf{a}}_6| \\ &\Longleftrightarrow (\vec{\mathbf{v}}_\mathrm{M} + \vec{\boldsymbol{\omega}}_\mathrm{K} \times \vec{\mathbf{r}}_5) \cdot \vec{\mathbf{a}}_6 = -\frac{1}{\mu_\mathrm{HBV}} \cdot [\mathbf{v}_\mathrm{f} \cdot |\vec{\mathbf{a}}_6| \\ &\qquad + \vec{\mathbf{v}}'_5 \cdot \vec{\mathbf{a}}_6 \cdot (1 - \mu_\mathrm{HBV}) \end{aligned} \tag{13}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto037
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

This equation is, component based, completed to

$$\begin{aligned}
&[\mathbf{v}_{\mathrm{M}x} + \boldsymbol{\omega}_{\mathrm{K}y} \cdot (z_5 - z_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{K}z} \cdot (y_5 - y_{\mathrm{M}})] \cdot (x_6 - x_6') \\
&+ [\mathbf{v}_{\mathrm{M}y} + \boldsymbol{\omega}_{\mathrm{k}z} \cdot (x_5 - x_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{K}x} \cdot (z_5 - z_{\mathrm{M}})].(y_6 - y_6') \\
&+ [\mathbf{v}_{\mathrm{M}z} + \boldsymbol{\omega}_{\mathrm{K}x} \cdot (y_5 - y_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{k}y} \cdot (x_5 - x_{\mathrm{M}})] \cdot (z_6 - z_6') \\
&= -\frac{1}{\mu_{\mathrm{HBV}}} \cdot \mathbf{v}_{\mathrm{f}} \cdot \sqrt{(x_6 - x_6')^2 + (y_6 - y_6')^2 + (z_6 - z_6')^2} \\
&= -\frac{1 - \mu_{\mathrm{HBV}}}{\mu_{\mathrm{HBV}}} \cdot [\mathbf{v}_{5x}' \cdot (x_6 - x_6') + \mathbf{v}_{5y}' \cdot (y_6 - y_6') \\
&\quad + \mathbf{v}_{5z}' \cdot (z_6 - z_6')] \qquad (14)
\end{aligned}$$

The CVE from Equation 6 represents, after completely writing the cross product with the indices $i = 1 \ldots 5$, the first five equations of the ABE-core

$$\begin{aligned}
&[\mathbf{v}_{\mathrm{M}x} + \boldsymbol{\omega}_{\mathrm{K}y} \cdot (z_i - z_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{k}z} \cdot (y_i - y_{\mathrm{M}})] \cdot (x_i - x_i') \\
&+ [\mathbf{v}_{\mathrm{M}y} + \boldsymbol{\omega}_{\mathrm{K}z} \cdot (x_i - x_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{K}x} \cdot (z_i - z_{\mathrm{M}}) \cdot (y_i - y_i')] \\
&+ [\mathbf{v}_{\mathrm{M}z} + \boldsymbol{\omega}_{\mathrm{K}x} \cdot (y_i - y_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{K}y} \cdot (x_i - x_{\mathrm{M}})] \cdot (z_i - z_i') \\
&= \mathbf{v}_i' = \mathbf{v}_{ix}' \cdot (x_i - x_i') + \mathbf{v}_{iy}' \cdot (y_i - y_i') + \mathbf{v}_{iz}' \cdot (z_i - z_i') \qquad (15)
\end{aligned}$$

with $\mathbf{v}_i'$ as external, vehicle-based joint moving speed (Albers, 2009).

This equation is, however, only valid when there are no internal connection rod extensions within the connection rods with the indices 1 to 5. To include those in the common matrix as well, the CVE is used once more:

$$\vec{\mathbf{v}}_i \cdot \vec{\mathbf{a}}_i = (\vec{\mathbf{v}}_{\mathrm{M}} + \vec{\boldsymbol{\omega}}_{\mathrm{K}} \times \vec{\mathbf{r}}_i) \cdot \vec{\mathbf{a}}_i = \vec{\mathbf{v}}_i' \cdot \vec{\mathbf{a}}_i \qquad (16)$$

The vehicle-based speed scalar product $\vec{\mathbf{v}}_i' * \vec{\mathbf{a}}$ is replaced by the connection rod change speed $\mathbf{v}_{\mathrm{LL}}$ multiplied with the value of the connection rod vector $\vec{\mathbf{a}}_i$. As the connection rod change speed $\mathbf{v}_{\mathrm{LL}}$ is contrary to the connection rod vector, it is

$$\vec{\mathbf{v}}_i \cdot \vec{\mathbf{a}}_i = -\mathbf{v}_{\mathrm{LL}} \cdot |\vec{\mathbf{a}}_i| \qquad (17)$$

From Equation 16, thus, it follows that

$$(\vec{\mathbf{v}}_{\mathrm{M}} + \vec{\boldsymbol{\omega}}_{\mathrm{K}} \times \vec{\mathbf{r}}_i) \cdot \vec{\mathbf{a}}_i = -\mathbf{v}_{\mathrm{LL}} \cdot |\vec{\mathbf{a}}_i| \qquad (18)$$

Solving the vector product, like in Equation 15, calculating the value of the connection rod vector $\vec{\mathbf{a}}_i'$ and performing the scalar multiplication, one gets the equations for the connection rods 1 to 5 in component form:

$$\begin{aligned}
&[\mathbf{v}_{\mathrm{M}x} + \boldsymbol{\omega}_{\mathrm{K}y} \cdot (z_i - z_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{k}z} \cdot (y_i - y_{\mathrm{M}})] \cdot (x_i - x_i') \\
&+ [\mathbf{v}_{\mathrm{M}y} + \boldsymbol{\omega}_{\mathrm{k}z} \cdot (x_i - x_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{K}x} \cdot (z_i - z_{\mathrm{M}})] \cdot (y_i - y_i') \\
&+ [\mathbf{v}_{\mathrm{M}z} + \boldsymbol{\omega}_{\mathrm{K}x} \cdot (y_i - y_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{K}y} \cdot (x_i - x_{\mathrm{M}})] \cdot (z_i - z_i') \\
&= \mathbf{v}_{\mathrm{LL}i} \cdot \sqrt{(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2} \qquad (19)
\end{aligned}$$

The left-hand sides of Equations 15 and 19 are identical, whereas they differ significantly on the right-hand sides. Equation 15 takes an external joint movement into consideration (point movement, short PM), whereas Equation 19 provides for an internal connection rod length change (LC). As both movements can only be performed separately, the common equation for the first five connection rod indices can be shown by superposition:

$$\begin{aligned}
&[\mathbf{v}_{\mathrm{M}x} + \boldsymbol{\omega}_{\mathrm{K}y} \cdot (z_i - z_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{K}z} \cdot (y_i - y_{\mathrm{M}})].(x_i - x_i') \\
&+ [\mathbf{v}_{\mathrm{M}y} + \boldsymbol{\omega}_{\mathrm{k}z} \cdot (x_i - x_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{K}x} \cdot (z_i - z_{\mathrm{M}})] \cdot (y_i - y_i') \\
&+ [\mathbf{v}_{\mathrm{M}z} + \boldsymbol{\omega}_{\mathrm{K}x} \cdot (y_i - y_{\mathrm{M}}) - \boldsymbol{\omega}_{\mathrm{K}y} \cdot (x_i - x_{\mathrm{M}})] \cdot (z_i - z_i') \\
&= [\mathbf{v}_i'] + [\mathbf{v}_{\mathrm{LL}i} \cdot \sqrt{(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2}\,] \qquad (20)
\end{aligned}$$

In this and in the spring equation 14, it is to be seen that next to the searched velocity components $\vec{\mathbf{x}} = (\mathbf{v}_{\mathrm{M}x}, \mathbf{v}_{\mathrm{M}y}, \mathbf{v}_{\mathrm{M}z}, \boldsymbol{\omega}_{\mathrm{K}x}, \boldsymbol{\omega}_{\mathrm{K}y}, \boldsymbol{\omega}_{\mathrm{K}z})^{\mathrm{T}}$, the velocities $\mathbf{v}_i'$ of the vehicle-based pivot joints as well as the spring compression velocity $\mathbf{v}_{\mathrm{f}}$ and the connection rod length change velocity $\mathbf{v}_{\mathrm{LL}i}$ are also unknown.

Equations 15 and 14 are separated into a matrix to create a linear equation system. It is shown to be

$$\mathbf{A} \cdot \vec{\mathbf{x}} = \vec{\mathbf{b}} \qquad (21)$$

with the square matrix $\mathbf{A} = \mathbf{A}_{(m \times n)}$ as well as the vector $\mathbf{b} = \mathbf{b}_{(m \times 1)}$ with $m, n = 1 \ldots 6$.

$$\begin{pmatrix} \mathbf{a}_{11} & \ldots & \mathbf{a}_{1n} \\ \ldots & \ldots & \ldots \\ \mathbf{a}_{m1} & \ldots & \mathbf{a}_{mn} \end{pmatrix} \cdot (\mathbf{v}_{\mathrm{M}x}, \mathbf{v}_{\mathrm{M}y}, \mathbf{v}_{\mathrm{M}z}, \boldsymbol{\omega}_{\mathrm{k}x}, \boldsymbol{\omega}_{\mathrm{k}z})^{\mathrm{T}} = \begin{pmatrix} \mathbf{b}_1 \\ \ldots \\ \mathbf{b}_m \end{pmatrix} \qquad (22)$$

To do so, the solution vector components from Equations 15 and 14 have to be isolated. For lines 1 to 6

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto037
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of matrix **A**, the following six row entries are to be valid:

$$\begin{aligned}
\mathbf{a}_{i1} &= (x_i - x_i') \\
\mathbf{a}_{i2} &= (y_i - y_i') \\
\mathbf{a}_{i3} &= (z_i - z_i') \\
\mathbf{a}_{i4} &= (y_i - y_\mathrm{M}) \cdot (z_i - z_i') - (z_i - z_\mathrm{M}) \cdot (y_i - y_i') \\
\mathbf{a}_{i5} &= (z_i - z_\mathrm{M}) \cdot (x_i - x_i') - (x_i - x_\mathrm{M}) \cdot (z_i - z_i') \\
\mathbf{a}_{i6} &= (x_i - x_\mathrm{M}) \cdot (y_i - y_i') - (y_i - y_\mathrm{M}) \cdot (x_i - x_i')
\end{aligned} \quad (23)$$

The first five entries of vector **b** are the result of Equation 15 for the indices 1 … 5:

$$\mathbf{b}_i = [\mathbf{v}_i'] + [\mathbf{v}_{\mathrm{LL}i} \cdot \sqrt{(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2}] \quad (24)$$

The sixth vector entry originates from Equation 14 and is to be enlarged by the term $\mathbf{v}_\mathrm{LL}$ through superposition

$$\begin{aligned}
\mathbf{b}_6 = &-\frac{1}{\mu_\mathrm{HBV}} \cdot \mathbf{v}_\mathrm{f} \cdot \sqrt{(x_6 - x_6')^2 + (y_6 - y_6')^2 + (z_6 - z_6')^2} \\
&-\frac{1 - \mu_\mathrm{HBV}}{\mu_\mathrm{HBV}} \cdot [\mathbf{v}_{5\mathrm{x}}' \cdot (x_6 - x_6') \\
&+\mathbf{v}_{5\mathrm{y}}' \cdot (y_6 - y_6') + \mathbf{v}_{5\mathrm{z}}' \cdot (z_6 - z_6')] \\
&+ \mathbf{v}_{\mathrm{LL}6} \cdot \sqrt{(x_6 - x_6')^2 + (y_6 - y_6')^2 + (z_6 - z_6')^2}
\end{aligned} \quad (25)$$

To generate a solvable linear equation system with six variables of this equation system, further boundary conditions, depending on the type of movement, have to be set.

Generally speaking, the body can be seen as stationary with every suspension or steering action. This is a true analog to many real suspension test benches.

For the conventional spring movement, all vehicle-based velocities $\mathbf{v}_i'$ are set to zero. In addition, there are no connection rod length changes except at the spring element; thus, all $\mathbf{v}_{\mathrm{LL}i}$ are set to zero. The spring compression velocity is a quotient, originating from the total spring travel and the number of calculation steps and is, thus, exactly defined. The matrix **A** is a pure geometrical matrix of the reviewed wheel suspension and stays unchanged. Vector **b** simplifies noticeably to

$$\begin{aligned}
&\vec{\mathbf{b}} = (0, 0, 0, 0, 0, 0, \mathbf{b}_6)^\mathrm{T} \text{with} \\
&\mathbf{b}_6 = -\frac{1}{\mu_\mathrm{HBV}} \cdot \mathbf{v}_\mathrm{f} \cdot \sqrt{(x_6 - x_6')^2 + (y_6 - y_6')\,(z_6 - z_6')^2} \\
&\quad - \frac{1 - \mu_\mathrm{HBV}}{\mu_\mathrm{HBV}} \cdot [\mathbf{v}_{5x}' \cdot (x_6 - x_6') + \mathbf{v}_{5y}' \cdot (y_6 - y_6') \\
&\quad + \mathbf{v}_{5z}' \cdot (z_6 - z_6')]
\end{aligned} \quad (26)$$

Matrix **A** and vector **b** are, thus, complete and definitely covered.

For a conventional steering movement, all connection rod length changes $\mathbf{v}_{\mathrm{LL}i}$ are set to zero. The spring compression velocity $\mathbf{v}_\mathrm{f}$ is set to zero as well. The tie rod bar of the wheel suspension is moved in *y*-direction, this is where the steering gear connects. The steering velocity $\mathbf{v}_1'$ (index 1 for the tie rod) is the quotient from the total traverse path related to the number of calculation steps. All other $\mathbf{v}_\mathrm{i}'$ are set to zero. This reduces vector **b** to an indicator of the steering velocity.

$$\vec{\mathbf{b}} = (\mathbf{v}_1', 0, 0, 0, 0, 0)^\mathrm{T} \quad (27)$$

Matrix **A** and vector **b** are, thus, complete and definitely covered for this case as well.

Next to the conventional analysis of the spring movements or steering actions, this calculation core can also be used for a tolerance or sensitivity analysis. It can be investigated how sensitive the system reacts to tolerances in the body-based pivot joints, for example, through production tolerances in body fabrication or differently long suspension arms. To do so, the spring compression velocity $\mathbf{v}_1$ as well as the steering velocity $\mathbf{v}_1'$ is initially set to zero.

For the analysis of different connection rod lengths, the connection rod length $\mathbf{v}_{\mathrm{LL}i}$ changes with the respective index as the set value for the calculation. For the tolerance analysis of the body pivot joints, the respective entry for the vehicle-based speeds $\mathbf{v}_\mathrm{i}'$ is chosen and defined. The tolerance fault can then be iteratively calculated through these new reference inputs.

2.6 Solution and embedding the ABE-core

With all elements of matrix **A** and vector **b** complete after Section 2.5, the solution vector is calculated in the ABE-core. In the programming used, the calculation via the inverted matrix turned out to be the fastest alternative. To do so, the initial system $\mathbf{A}^*\vec{\mathbf{x}} = \vec{\mathbf{b}}$ is multiplied with the inverted matrix on the right-hand side:

$$\begin{aligned}
&\mathbf{A}^{-1} \cdot \mathbf{A} \cdot \vec{\mathbf{x}} = \mathbf{A}^{-1} \cdot \vec{\mathbf{b}} \\
&\Longleftrightarrow \vec{\mathbf{x}} = \mathbf{A}^{-1} \cdot \vec{\mathbf{b}}
\end{aligned} \quad (28)$$

It is possible to calculate the inverted matrix $\mathbf{A}^{-1}$ only when matrix **A** is a regular and not a singular matrix. A squared matrix with the dimension *n* is considered regular only when the rank of the matrix is equal to the dimension *n* [BRO05]. Only then the matrix has *n* linear and independent equations. If the determinant det(**A**) unequal to zero, then

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto037
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

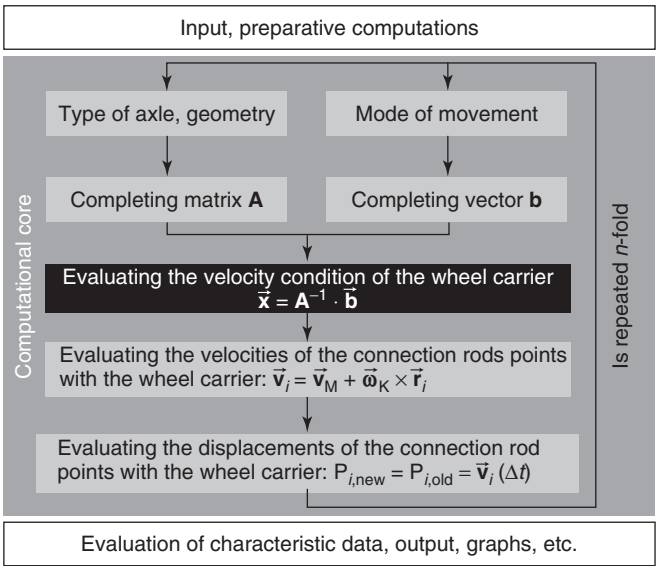


Figure 8. Embedding of the ABE-core.

the matrix is considered to be regular. That is why the determinant is checked before inverting the matrix.

The central ABE-core is centrally embedded to the total flow. It is very flexibly designed and programmed in its basic form as well as constantly called upon by different program modules. It is especially essential to program this calculation block efficiently in view of calculation speed (Figure 8).

This core can, for example, be programmed in Microsoft Excel using a VBA programming language and designed with a respective interface for input and output of data.

3 REVIEW

Chassis modeling and optimization is very often done by commercially available programs, which do not facilitate understanding of the theoretical relationship between input and output. This chapter describes a methodology that avoids these disadvantages and starts with the theoretical background of chassis design to formulate equations and explain the movements of suspensions under vertical and rotational inputs.

The equations, that both explain and describe the wheel movement, are finally programmed in Excel and the kinematic characteristics can be found. Changing the coordinates of the connection points within the suspension linkage, the new chassis characteristics can be easily found rapidly. This design tool, therefore, enables the optimization of chassis systems, even before any hardware has been built. In addition, the engineers can identify the relationship between their design changes and the new chassis behavior.

The technical background of all these procedures are described by Matschinsky (2000) and the computational transfer into a simple and fast simulation program is the merit of Albers (2009).

REFERENCES

Albers, T. (2003) Erstellung eines Berechnungstools zur starrkinematischen Analyse von Einzel-radaufhängungen. Diploma thesis. RWTH Aachen, Aachen, Germany.

Albers, I (2009) Auslegungs- und Optimierungswerkzeuge für die effziente Fahwerkentwicklung. PhD thesis. RWTH Aachen, Aachen, Germany.

Matschinsky, W. (2000) *Road Vehicle Suspensions*, Professional Engineering Pub, Wiley-Blackwell, UK.

Matschinsky, W. (1998) *Radführungen der Strassenfahrzeuge*, Springer Verlag, Berlin, Germany.

Matschinsky, W. (1992) Bestimmung Mechanischer Kenngrössen von Radaufhängungen. PhD thesis. Universität Hannover, Hannover, Germany.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto037
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Gaskets and Sealing

Lucian M. Silvian
Purdue University, West Lafayette, IN, USA

1 History 1
2 Surface Texture Analysis 1
3 General Sealing Requirements 2
4 Gasket Design and Material Selection 6
5 Test Methods 12
6 New Methods of Evaluating and Analyzing Designs 15
Further Reading 18

1 HISTORY

The history of automotive sealing is closely related to the development of the engine. As the combustion ratio and combustion pressure increased, sealing the engine, especially the combustion chamber, became a challenge.

New materials were necessary to help in sealing. In the 1940s and 1950s, for example, a few weeks after purchasing, the cars would be returned to the dealers and the cylinder head-to-block assembly had to be retorqued to ensure a good seal of the combustion chamber. In the 1960s, new materials were developed for the cylinder head gasket, which did not require retorqueing.

Asbestos was one of the materials used for many engine sealing applications. In combination with other materials, it had good sealing characteristics over a wide range of temperatures. However, health issues prompted responsible sealing companies to remove asbestos from their gasket products.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

In addition, beginning in the mid-1970s, the energy crisis, emission reduction, and worldwide competition required a different philosophy in engine design. Increasingly, the large, slow speed high torque engines used in American cars would be replaced by smaller, more fuel efficient, high speed engines that work under higher temperature, higher combustion pressure, and increased service life.

New lighter designs were needed, as engine designers removed material from cylinder heads and engine blocks. To further reduce weight, cast iron was often replaced by aluminum. First, there were the bimetal engines, usually a cast iron block and an aluminum cylinder head, and then more and more total aluminum engines were developed. Engine sealing became a more difficult challenge. New seal designs were needed that could withstand the harsher conditions and longer life requirements. Many engines migrated from composite gasket materials to multilayer steel gaskets.

Other applications such as valve covers, oil pans, and transmission pan gaskets are designed today using molded rubber and some plastic–rubber composites that also simplify engine assembly.

2 SURFACE TEXTURE ANALYSIS

To properly design a functional gasket, one must include the surface finish of the mating components. The irregularities in the surfaces keep the surfaces away from each other creating passages for fluids to flow. Gaskets are designed not only for new engines but also for the aftermarket. New engines are manufactured to a surface finish selected during the design. Engines that are repaired or rebuilt will invariably have rougher surfaces than those of new engines.

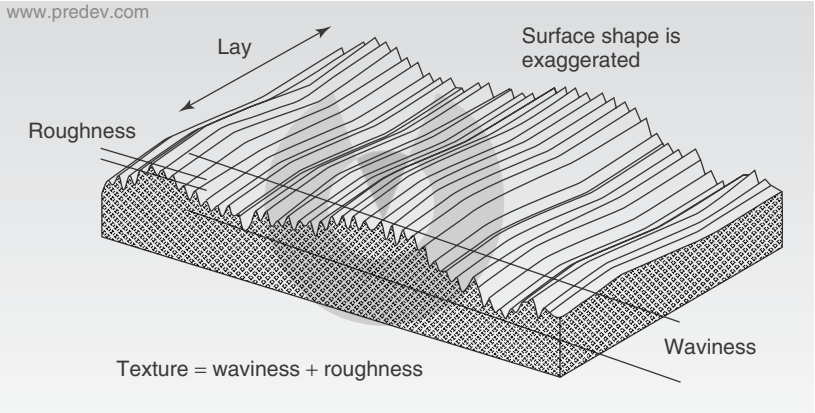


Figure 1. Surface texture characteristics. (Courtesy of Precision Devices, Inc.)

For example, during the 1980s, a typical cylinder head-and-block surface roughness was between 1.5 and 3 μm, whereas a typical surface roughness suggested for remachining was 3–4 μm. This difference in surface finish required different gaskets.

Surface texture is the combination of deviations from the nominal surface. Texture includes *roughness*, *waviness*, *lay*, *and form*. Each term is defined as follows.

Roughness, as seen in Figure 1, includes the finest (shortest wavelength) irregularities of a surface.

Roughness generally results from a particular production process, tool, or material condition. Figure 2 shows the roughness average R_a resulting from various manufacturing processes. One could guess that the lower R_a the higher the expense to produce it. Therefore, it is important for the designer to choose a surface roughness as high as required and as low as necessary.

Waviness includes the more *widely spaced (longer wavelength) deviations* of a surface from its nominal shape. Waviness is the result of machine or work deflections, vibrations, chatter, or various causes of strain in the material.

Texture includes roughness and waviness.

Form is the general shape of the surface, neglecting variations due to roughness and waviness (usually it is referred to as *form error*).

Lay refers to the predominant direction of the surface texture (Figure 3).

A profile is a two-dimensional picture of a three-dimensional surface similar to the representation of a sectioning place cutting the surface. Profiles are ordinarily taken perpendicular to the lay (Figure 4).

The traverse length (A + B + C) of a profile measurement, as indicated in Figure 5, is the total distance traveled by the profiling instrument's pickup during data collection. The evaluation length (B) is the entire length of a profile over which data has been collected. The evaluation length will ordinarily be shorter than the traverse length because of end effects in the travel (A) and (C): motors accelerating and decelerating, electrical filters settling down, etc.

To be of engineering value, surface traces are magnified moderately in the horizontal direction and significantly in the vertical direction in order to be presented on a computer screen or a paper. This difference, shown in Figure 6, leads to a very sharply undulating trace that easily deceives the uninitiated as to the actual shape of the surface.

3 GENERAL SEALING REQUIREMENTS

3.1 Creep

Creep is the tendency of materials to slowly deform under long-term stress that is below material's yield.

As indicated in the accompanying diagram, the creep of a material in time can be divided into four stages (Figure 7).

1. Zero stage is the creep that takes place at assembly. If measured precisely at the end of the torque operation, the torque is already dropping.
2. First stage, or primary creep, starts at a rapid rate and slows with time. This usually takes place and ends after a few hours to a few weeks. It is responsible for the majority of creep loss.
3. Second stage (secondary) creep has a relatively uniform rate. It is the stage that takes place during the life of the assembly. The rate is very slow and is not expected

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

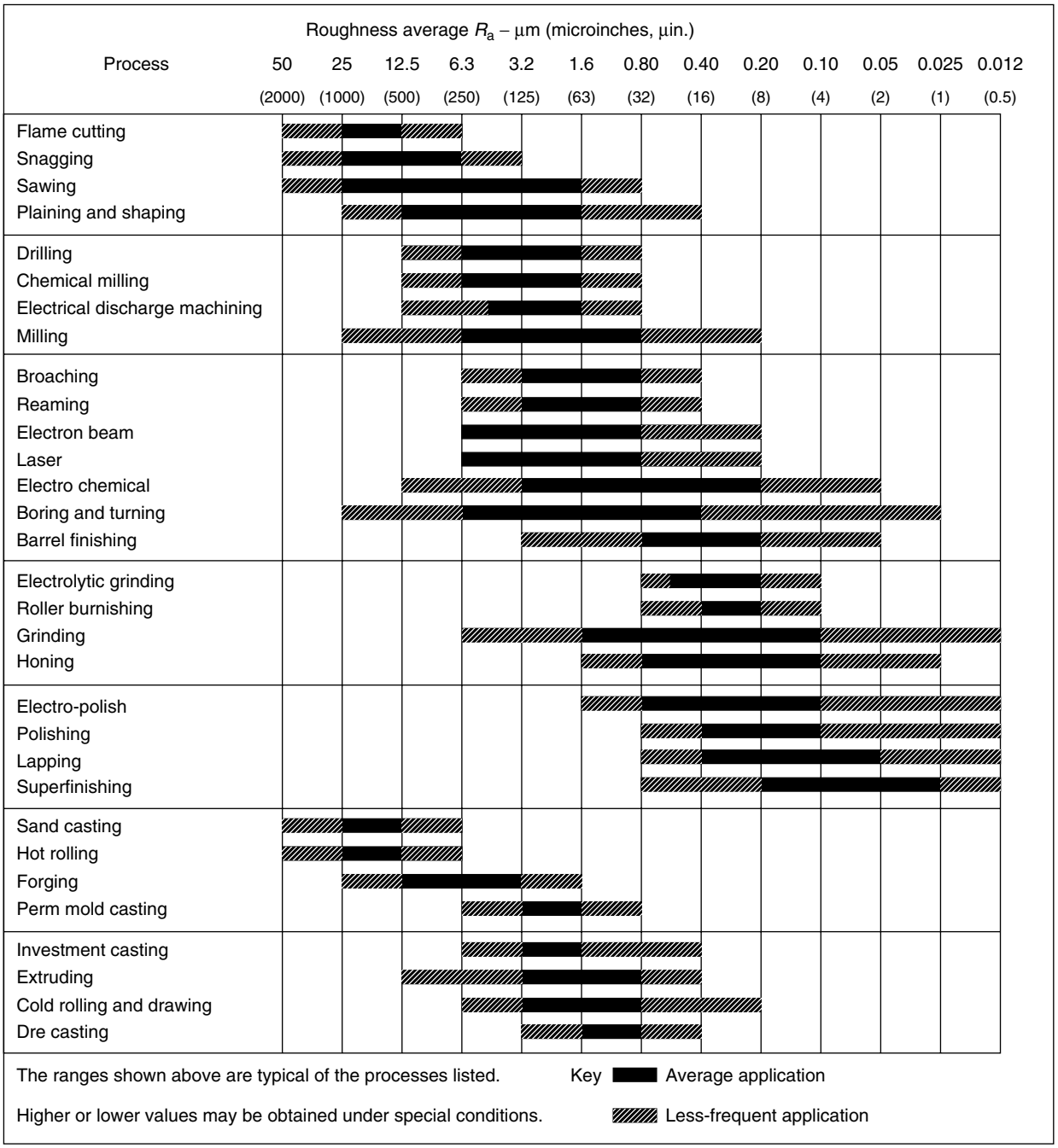


Figure 2. Surface roughness R_a produced by different manufacturing processes. (Reproduced from Black and Kohser (2008), *Degarmo's Materials & Processes in Manufacturing,* 10th ed. © John Wiley & Sons.)

to stop. Careful tests up to 1000 h have demonstrated that creep continues over the life of a gasketed joint.

4. Third stage (tertiary) creep has an accelerating creep rate and terminates by failure of material at time for rupture.

The factors that influence creep the most are application temperature and load. In the case of temperature and load variation such as seen in an internal combustion engine, the effect is even greater. As creep takes place, the joint relaxes, resulting in torque (clamping load) loss over time.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

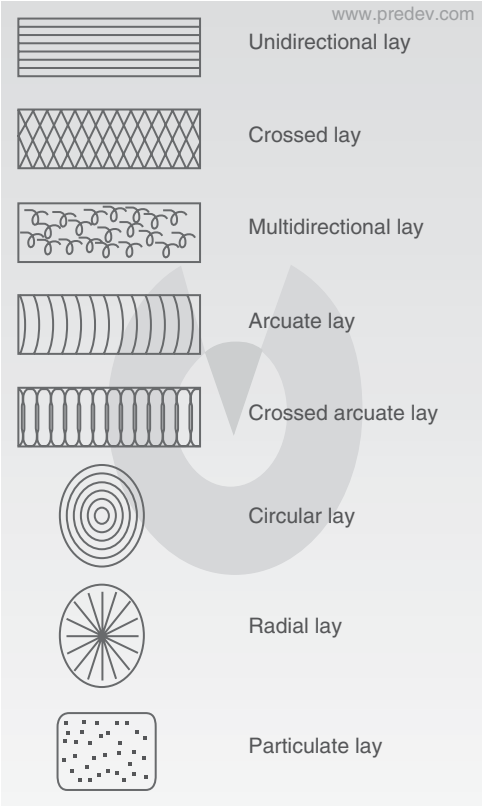


Figure 3. Typical lay. (Courtesy of Precision Devices, Inc.)

3.2 Bolted joint behavior

In an application, a bolted joint is composed of the flanges, bolts, and gaskets. As the clamp load, also called *preload*, is applied, by applying torque to the bolt, the bolt is stretched and the gasket is compressed.

The stress–strength diagrams of the bolt under tension and the gasket under compression could be conveniently arranged as shown in the diagram later.

Assume that the diagram represents a portion of a cylinder block, gasket, cylinder head, and bolt. As the bolt is tightened, it stretches in the linear relationship shown in Figure 9. The bolt behaves like a spring as shown in Figure 8.

The gasket, on the other hand, will compress following a curve as shown in Figure 9 (seal compression). Its behavior is determined by the gasket material and could be imagined as the combined performance of a spring and a dashpot or damper. There are several models to describe it; the most common are the Maxwell and Kelvin models (Figure 10).

For the time being, the deflection in the flanges has been neglected. This will add more complexity to the analysis of the system. At present, computer simulation through finite element analysis (FEA) can be used to analyze the complex behavior of a joint with very good results.

During the power stroke of the engine, the pressure pulse in the combustion chamber further stretches the bolt as shown in Figure 9, reducing the gasket load. If the gasket is unloaded, too much leakage will occur. A good gasket designer will consider this and design a gasket with minimum load to seal much below this point.

In addition, the temperature and cyclic load change the behavior of the gasket and after some time of operation, the stress–strength–shape of the gasket changes to follow the red curve in the Figure 9. Most gasket materials are made in part with elastomers. Elastomers are considered time-dependent materials. It is this behavior that further reduces the load on the gasket toward the minimum load to seal. The designer has to consider this as well when designing gaskets.

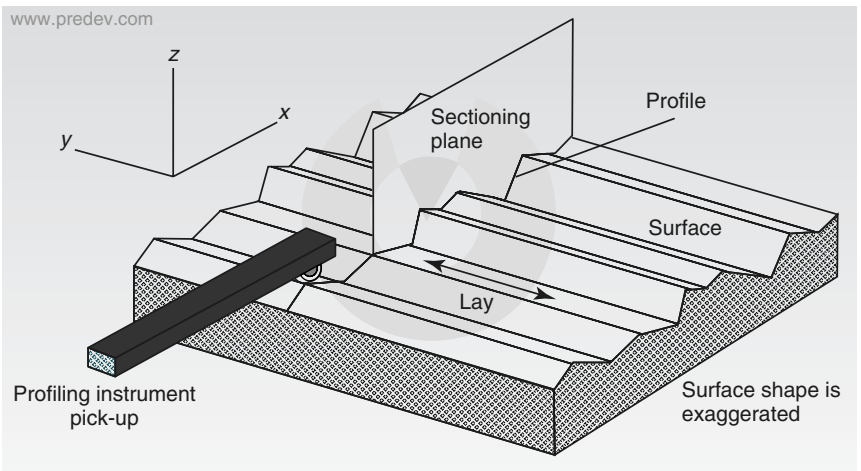


Figure 4. Profile representation. (Courtesy of Precision Devices, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

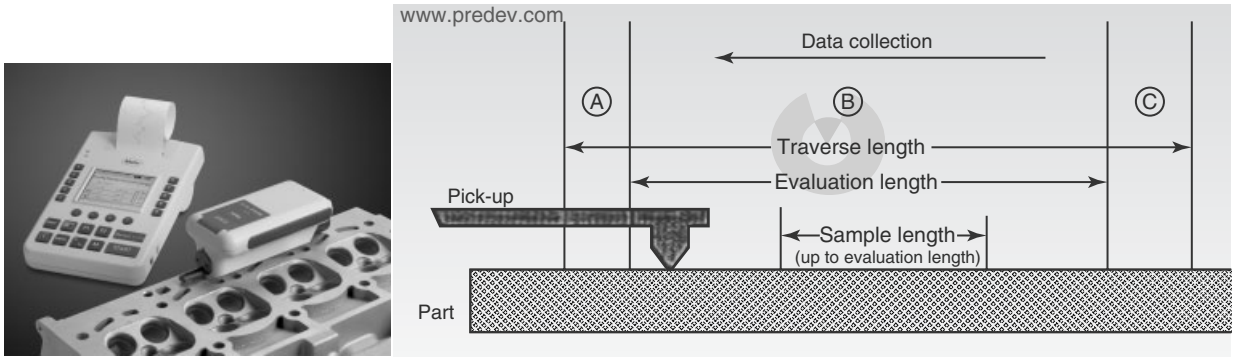


Figure 5. Profile measurement. (Courtesy of Mahr Federal (left) and Precision Devices, Inc. (right).)

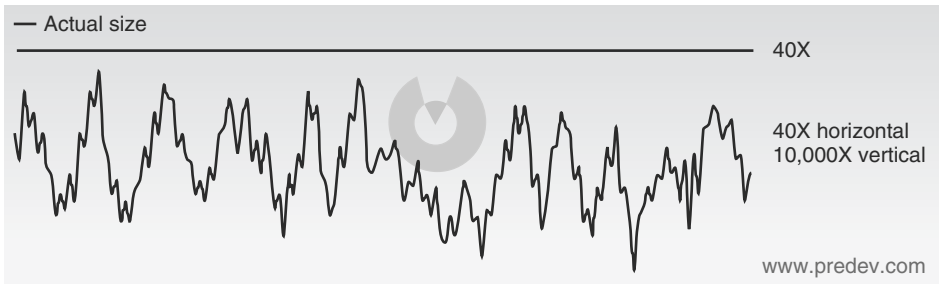


Figure 6. Surface profile trace. Observe the difference between horizontal and vertical scales. (Courtesy of Precision Devices, Inc.)

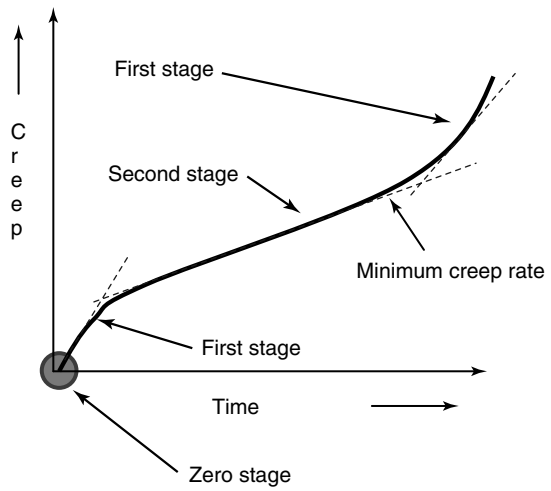


Figure 7. Creep behavior of materials.

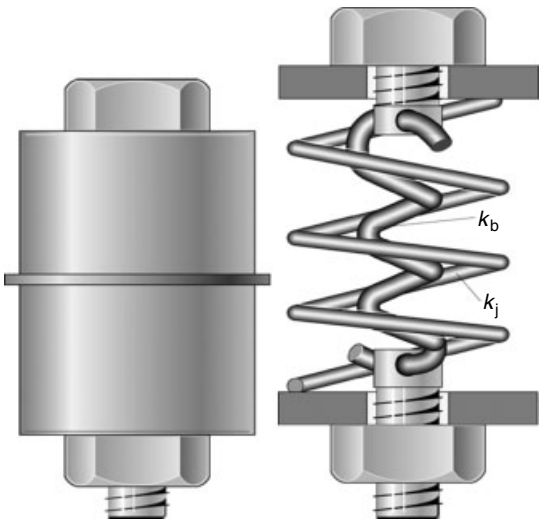


Figure 8. A bolted joint diagram.

3.3 Recovery

The next consideration is what takes place when a seal is loaded and unloaded. Most of the materials used to make gaskets will behave as shown in Figure 11. This is called *hysteresis*. If the load is cyclical, the area under the curve will eventually be reduced. In an internal combustion engine, during the power stroke, as the bolt is stretched, the gasket material is unloaded, and material recovery is essential in maintaining load in the joint, thus providing a permanent seal.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

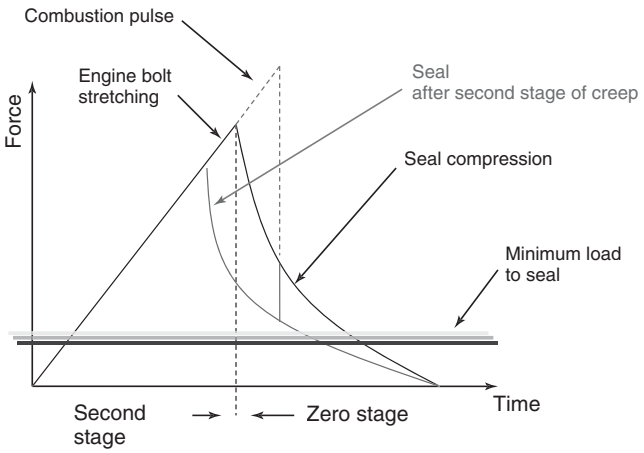


Figure 9. Stress–strain bolted joint diagram.

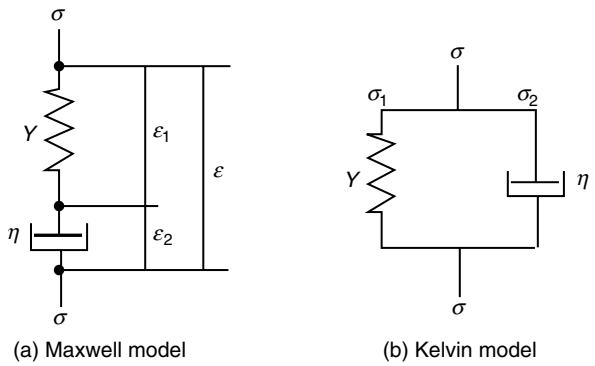


Figure 10. Models of viscoelastic behavior of typical gasket materials.

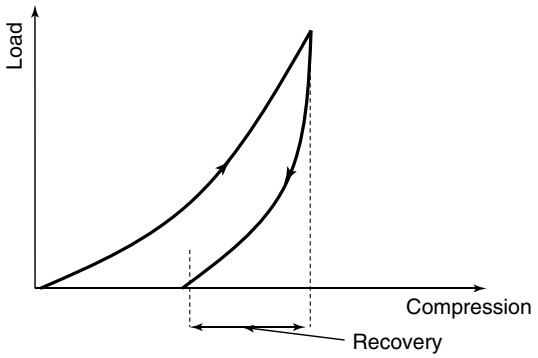


Figure 11. Stress–strain diagram showing material recovery on removing the load.

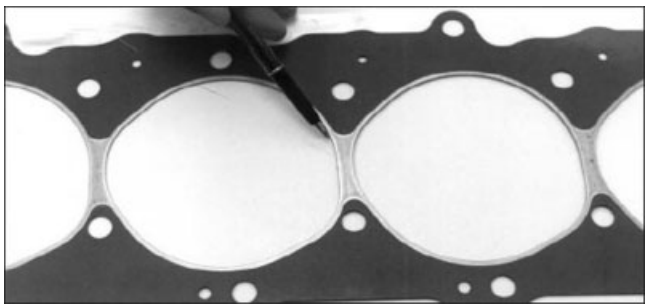


Figure 12. Typical cylinder head gasket used in a gasoline engine.

4 GASKET DESIGN AND MATERIAL SELECTION

4.1 Cylinder head gasket

The cylinder head gasket is the most complex gasket in an engine application. It is required to seal combustion gasses at medium to high temperature and high pressure. Peak combustion pressure starts at about 50 bar in gasoline engines and may be well over 200 bar in high compression diesel engines (Figure 12).

The same gasket seals coolant transfers between the cylinder block and the cylinder head. In many cases, it throttles the flow of coolant to balance coolant flow from one cylinder to another, thus creating a pressure differential up to 6 bar on the gasket, which may apply high forces on the exposed area of the gasket.

Finally, the cylinder head gasket must seal the oil transfer passages between the cylinder block and head.

The cylinder head and block are not designed with uniform thickness throughout, so the deflection of both after assembly creates nonuniform loading on the gasket.

The design pattern of the exhaust and intake valves poses an added challenge as the exhaust valves of adjacent cylinders are often placed next to each other. This creates a hot spot in the thin gasket bridge between two cylinders and a much cooler bridge between two intake valves at the adjacent bridge. This also must be considered during the design.

The head bolt pattern is often defined based on the other engine considerations, with little input from the gasket designer. As the cylinder bore increases beyond the span between head bolts, there may be inadequate and unequal clamping load around the combustion chamber perimeter.

The surface finish of the cylinder head and block again must be considered when selecting a cylinder head gasket design. The considerations identified here result in different requirements in different areas of the gasket.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.1.1 The body

The body of the gasket could be made as a layered composite: a metal substrate called *core* that provides structure and stiffness and soft facing materials on both sides of the core that provide the necessary seal to the face irregularities of the cylinder head and block. The facing material is usually porous. In the application, load is applied to compress the materials and reduce the porosity. If the clamp load offered by the bolts is not enough, enhancements to the gasket body are required.

There are generally two methods of making the composite:

Chemical bonding (Figure 13): the facing material is chemically bonded to the core. It is usually done using laminator equipment. The core is cleaned by a chemical process to insure that its surface is free of any rust, grease, or other foreign materials. The bond that could be water or solvent based is applied to the core, and the facing material is laminated on both sides of the core. Pressure and temperature to cure the bond are required. The advantage of this method is that the body offers uniform loading limited only by the facing material density.

The disadvantages are that if cleaning and curing are not done exactly as prescribed, delamination of the facing material could occur in service.

Perforated core or mechanical bond (Figure 14): the metal substrate is pierced on both sides with a piercing machine. The facing material is applied on both sides of the perforated core and the composite is laminated together to a finished body thickness; as the perforations bend around the material they provide a mechanical bond (Figure 15).

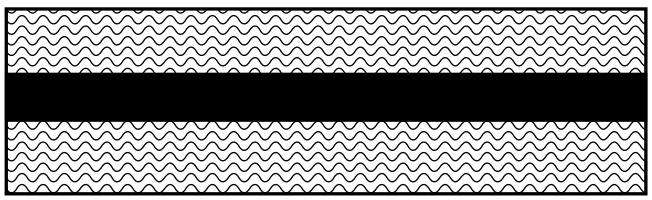


Figure 13. Layered composite made by chemical bonding. (Courtesy of Federal Mogul.)



Figure 14. Perforated core facing material. (Courtesy of Federal Mogul.)

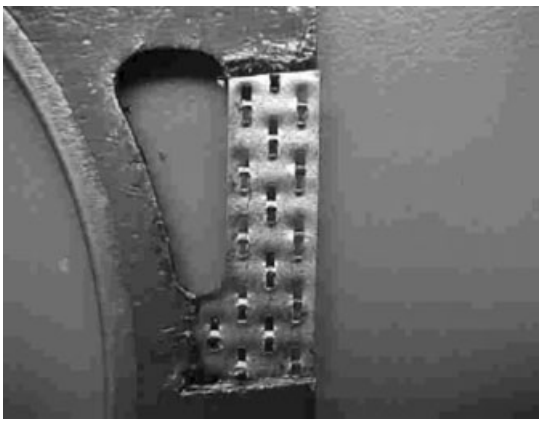


Figure 15. Exposed core. (Compliments of LFP Technologies.)

The density of perforations (number of perforations per square inch) and geometry determine the final stiffness of the gasket body. The advantage of this type, which is a mechanical bond, is that it will not require a very thorough cleaning of the metal surface as required for chemically bonded composite fabrication. The disadvantages are some nonuniformity of loading on the body, and in the narrow areas of the gasket, there may not be enough perforations to keep the facing material attached to the core and the facing material could move laterally.

Facing materials used in both cases are called "*paper*" materials. They are made with special clays and fibers such as Kevlar and elastomers. They are called *papers* because they are made using the same process as the paper gaskets.

Graphite has been used since the 1980s as a facing material. It has the advantage that it could be densified during the lamination process to the desired density for improved sealing. Once the right density is achieved, the gasket requires very little load to properly seal. The disadvantage of graphite is that while it has good mechanical properties in the surface directions (X and Y), it has poor mechanical properties in the vertical direction (Z). Graphite is used in both chemically bonded and mechanically clinched forms.

Most of these facing materials are time-dependent materials. The properties change with time and deteriorate over the life of an engine.

As engines are designed today for longer and longer life cycles, most of these facing materials are being replaced by multilayer gasket materials. These gaskets have the advantage that they can be used with much lower deterioration, meeting increased engine life to overhaul expectations.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

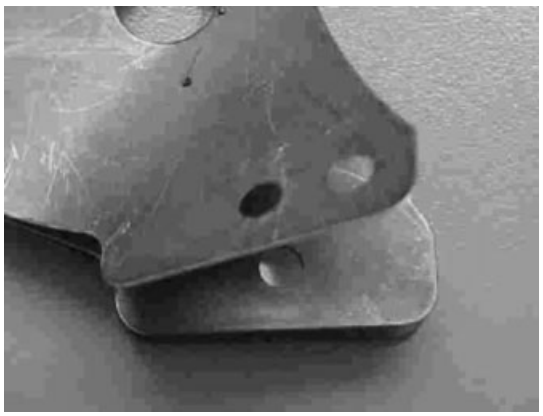


Figure 16. Multilayer metal gasket. (Compliments of LFP Technologies.)

Multilayer gaskets are made using two or three layers of metal as shown in Figure 16. They increase the life of the engine but require better engine surface preparation resulting in a required R_a under 1.016 μm. These gaskets are less conformable and need more uniform loading across the engine head and block and more uniform deck thickness.

When the total load developed by the bolts is not enough to compress the gasket body to perform a good seal, enhancements are designed into the gasket body to help seal.

Screen printing, generally using elastomers such as silicone, increases the density and seal under the screened bead, reducing the load required to seal. It can be used to vary load distribution on the gasket body and selectively build up gasket thickness (Figures 17 and 18).

For multilayer gaskets, beads are designed into the body to mechanically upset the material as shown in Figures 19 and 20. This improves the sealing at problem locations (high temperature gas openings and high pressures). It is also used to reduce bending of joint flanges.

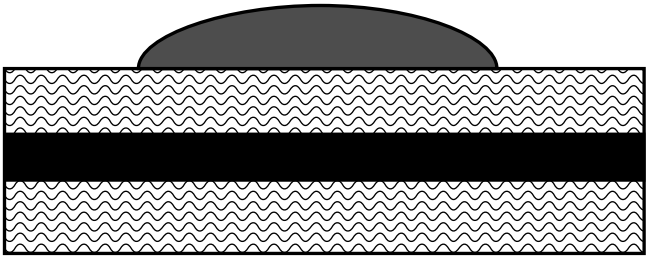


Figure 18. Screen printing. (Complements of Federal Mogul.)

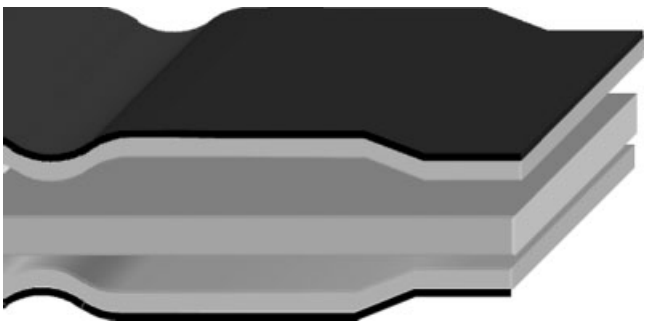


Figure 19. Multilayer metal construction. (Complements of Federal Mogul.)

Coatings are applied to the gasket body to help surface sealing. There are primarily two types of coatings: microsealing and antifriction. Microseal coatings are used as the name suggests to seal surfaces reducing the need for higher clamping loads. They are in general silicone coatings, but some other elastomers are used depending



Figure 17. Screen printing. (Compliments of LFP Technologies.)



Figure 20. Multilayer gasket with beads. (Compliments of LFP Technologies.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

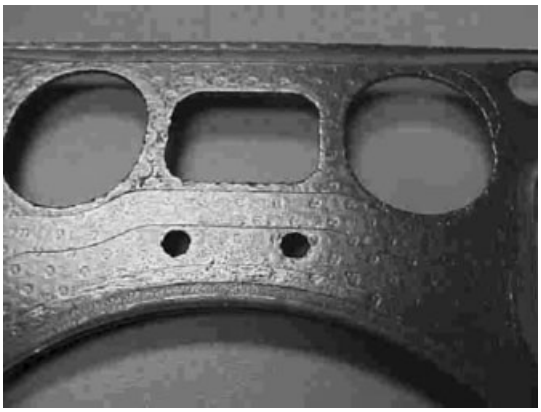


Figure 21. Antifriction coating on a graphite cylinder head gasket. (Compliments of LFP Technologies.)

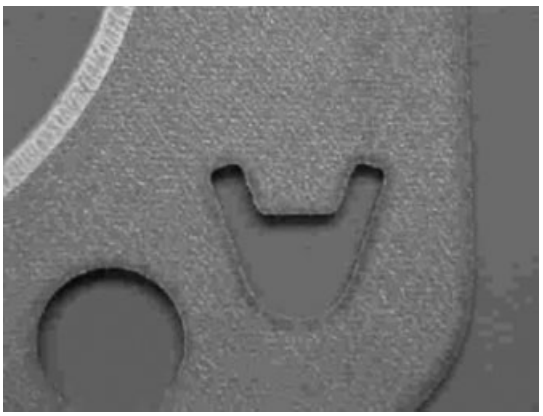


Figure 22. Armored gasket. (Compliments of LFP Technologies.)

on the requirements. An example is a coating needed to seal during engine initial operation such as during the final engine inspection. Once the gasket is exposed to heat for 5–10 min, the gasket will take over and seal for the life of the engine.

Antifriction coatings are used in bimetal or all aluminum engines. The aluminum expands when heated at twice the rate of steel. The metal substrate of the gasket is steel; the aluminum surface irregularities would embed into the soft facing material and would tend to move the facing as the engine heats up thus tearing the facing material apart. In general, these coatings contain molybdenum disulfide and graphite (Figure 21).

4.1.2 Combustion seal

The gasket body seals the coolant and lubricant but the higher pressure and temperature of the combustion chamber requires further measures.

4.1.2.1 General combustion seal. For typical gasoline engines, there is an eyelet made of metal, usually steel. This eyelet or armor has two functions, the first of which is to increase loading. This is accomplished by the additional thickness of the eyelet material. During initial gasket installation, this area will be loaded first and the soft material under the armor will be compressed much more than the rest of the body, thus creating a much higher unit stress. The thickness and the width of the armor together with a precoining operation are essential to the load distribution on the combustion seal and the load balance between the body and the combustion (Figure 22).

The second reason for the combustion eyelet is to protect the facing material from the combustion heat.

For high performance gasoline engines and diesel engines, the addition of the combustion armor alone does not produce adequate clamping loads because of the increased combustion pressure. A round wire ring is then attached to the gasket along with the combustion armor, as shown in Figure 23. In this case, the wire ring dimensional tolerances are critical. The load on the gasket body and rings must be carefully balanced. Computer programs are used to determine the adequate load distribution between the combustion and the gasket body.

In the case of aluminum flanges, further care must be taken to avoid material yielding. Depending on the application, the wire rings are made of steel, stainless steel, or copper.

4.1.2.2 Wire rings. The wire rings are used with a composite body as presented earlier. When multilayer steel designs are used, the combustion seal is developed by

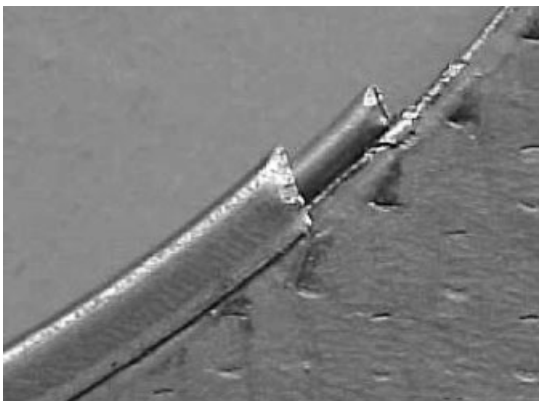


Figure 23. Wire ring under the armor. (Compliments of LFP Technologies.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 24. Embossed metal for added stress in the combustion area. (Courtesy of Federal Mogul.)

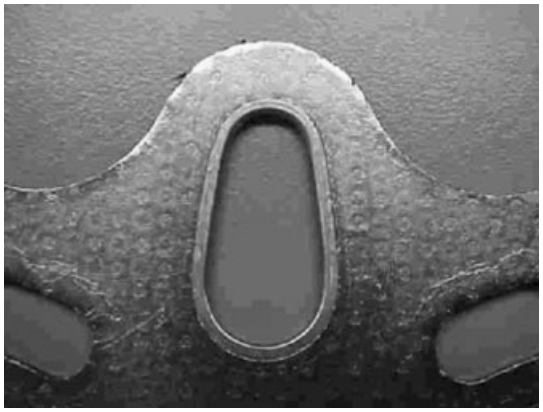


Figure 25. Eyelet used to seal an oil passage in a cylinder head gasket. (Compliments of LFP Technologies.)

mechanically upsetting the material of one of the layers and using a second layer to fold over to create an eyelet (Figure 24).

4.1.3 Oil sealing

Oil sealing on the cylinder head gasket is achieved in many ways. One way as shown in Figure 25 is using a metal eyelet over the gasket body. The eyelet thickness is important in balancing the load on the oil seal and the other openings close to it.

4.2 Intake manifold

The intake manifold gasket seals fuel–air mixture, coolant, and sometimes exhaust gas recirculation (EGR). Many designs are made using similar composite bodies as used for cylinder head gaskets. As the stresses in these gaskets are less than that of the cylinder head gasket, facing materials used may be less expensive. For example, instead of using Kevlar fiber that is very expensive materials with other, less expansive fibers could be used.

For "V" engines, the installation of the intake manifold creates a particular challenge. If the bolt distribution and installation is not developed correctly, it may lift the cylinder head from the cylinder head–gasket–block assembly producing coolant leaks.

4.3 Exhaust gasket

Exhaust gaskets are subjected to high temperatures and flange deflection. The body consists of fiber facing materials or graphite applied on both sides or sandwiched between perforated metals as shown in Figure 26. There are two basic designs: independent gaskets for each port and a single exhaust gasket construction called a *header*.

There are also some single and multilayer exhaust gaskets incorporating embossed designs as shown in Figures 27 and 28. Because of high temperatures, the metal gaskets are made of stainless steel full hard or super hard.

Metal exhaust gaskets work very well and maintain seal if designed properly.



Figure 27. Single layer embossed metal.

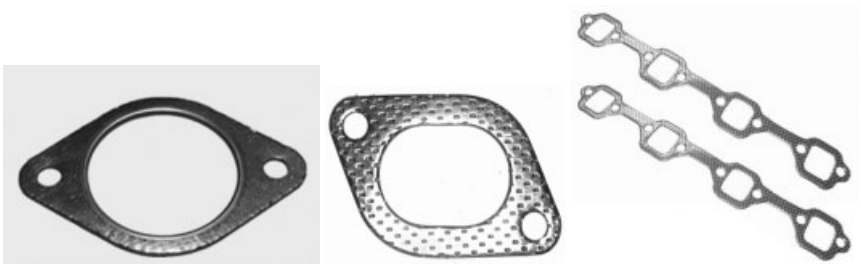


Figure 26. Exhaust gaskets.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

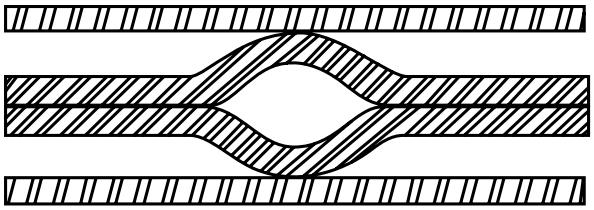


Figure 28. Multilayer embossed metal. (Courtesy of Federal Mogul.)

4.4 Valve cover, oil pan, and transmission pan gaskets

Each of these gaskets is designed to seal oil at low pressures. In the past, the valve covers and oil pans were made of sheet metal and the gaskets were made of cork or cork rubber. Short bolts and vibration were producing very large torque loss resulting in leakage. At present, the covers are more typically die castings that provide added rigidity, and the gaskets are molded elastomers as shown in Figure 29.

In some modern designs, valve covers are made of plastic and the gaskets are molded in place. In this case, torque limiters (load limiters) are required to prevent cracking around the bolt holes and to distribute the load evenly. Torque limiters are inserted during assembly or are molded in place during gasket fabrication (Figure 30).

4.5 O-Rings

An O-Ring is a seal that is in the shape of a torus (doughnut) (Figure 31). It is made of an elastomer, polytetrafluoroethylene (PTFE), or another thermoplastic. It could be solid or hollow. It is used to seal fluids. In engine applications, it is usually used to seal fuel.

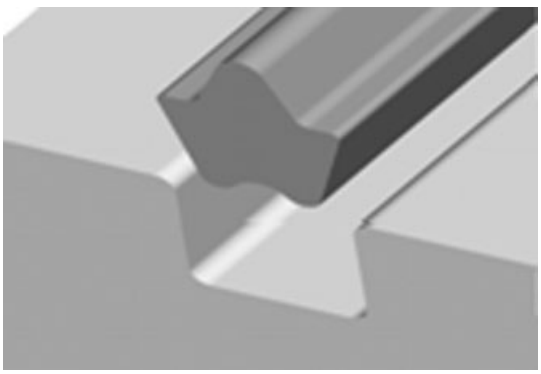


Figure 29. Elastomer molded gasket installation on a cast valve cover. (Compliments of Federal Mogul.)



Figure 30. Torque/load limiters inserted in a molded rubber gasket. (Compliments of Federal Mogul.)

When discussing O-Rings, one must consider the assembly, the O-Ring, and the gland. The gland is cut into one of the flanges and its dimensions are designed for proper function (Figure 32).

The O-Ring is subjected to two stresses: the clamping force, which is produced by the surrounding flanges, and the internal pressure, which is created by the fluid that is supposed to be sealed.

The elastomer is viscoelastic; it will compress little and it will deflect a lot, filling the groove and closing

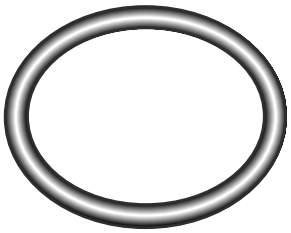


Figure 31. O-Ring.

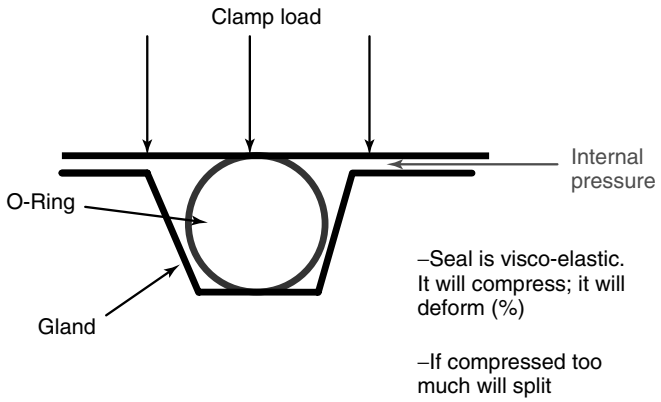


Figure 32. O-Ring and gland assembly.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

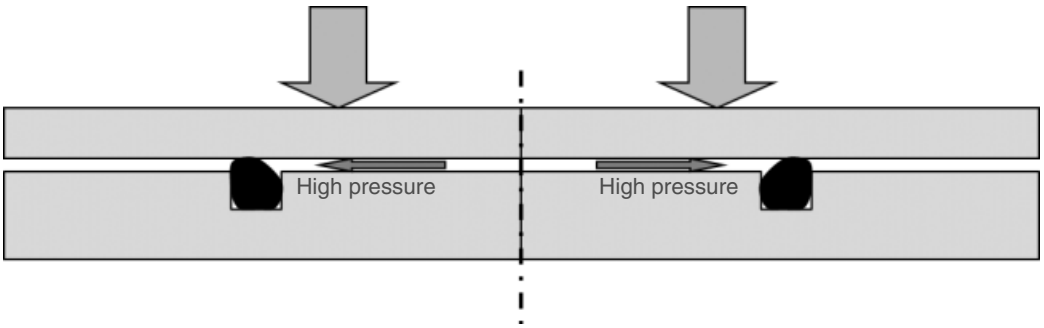


Figure 33. O-Ring deforms under high internal pressure.

any leak passage. If the internal pressure is high enough, the O-Ring will deform into the narrow gap between the flanges.

As the pressure increases, a small portion of the O-Ring may enter into the narrow gap as shown in Figure 33. This is actually the pressure limit of the O-Ring. If the pressure is high, in general above 100 bar, backup rings may be needed.

In static seal applications, the flanges are not subjected to movement relative to each other except by the fluid pressure, temperature changes, and vibration. The O-Rings are usually subjected to constant or pulsating internal pressure in one direction only. An example of a static seal application is sealing fuel lines.

Very important when determining the right O-Ring is to determine the compatibility between the O-Ring material and the fluid it seals. This is important especially today when there is a lot of research on synthetic fuels and oils for automotive and aerospace industries.

5 TEST METHODS

There are many tests used in industry for gasket materials and for design and development of gaskets: some of these tests are standardized under the American Society for Testing Materials (ASTM). Others are developed for a very specific purpose by the gasket manufacturer and usually are approved by the customer before being incorporated into a test program.

5.1 Gasket material tests and their importance

Table 1 shows some of the tests that are suggested and their reasons.

5.1.1 Sealability

Sealability is the most important property and the main reason we use gaskets. Many tests were developed by material and gasket companies for their own purpose, and

Table 1. Standard ASTM and typical tests perform on materials during the design and development of gaskets.

Property To Be Tested	Test Method	Importance and Significance
Sealability	ASTM F37	Resistance to fluid passage
Heat resistance	Exposure testing at elevated temperatures	Resistance to thermal degradation
Fluids compliance	ASTM F104 and F146	Compatibility to fluids
Antistick characteristics	Fixture testing ASTM F607 or as agreed with the customer	Ability to release from the flanges during disassembly
Stress versus compression, spring rates	Load frame testing	Determining loading/compression for sealing the different fluids
Compressibility and recovery	ASTM F36 (sheet) and ASTM F805 (composites)	Determine the stress–strain characteristics for proper load distribution
Creep relaxation and compression set	ASTM F38 and D395	Related to clamp load retention and subsequent loss of sealing
Crush and extrusion properties	Compression test using load frame machines	Verifying the resistance at high loading on gasket narrow areas at room and elevated temperatures

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

some of these tests were adopted by ASTM. The universal test procedure is ASTM F37. Test method A is restricted to liquid leakage measurement. Method B may be used for liquid as well as gas leakage measurement.

Both methods use test specimens, most of the time doughnut shaped, sandwiched, and compressed to a specific "clamp" load, between the surfaces of two smooth steel plates. Fluid pressure is applied to the center of the doughnut. Leakage rate is observed at the outside of the doughnut. More sophisticated test fixtures include pressure transducers to measure pressure decay as a measure of leakage rate.

During the development of a gasketed joint surface, profile of the steel plates may be changed to resemble the surface of the flanges used in the real application. This step is important as it may reduce problems during production start up.

5.1.2 Heat resistance

Some engine applications require high temperatures or high temperature gradients for prolonged periods of time. Thermal degradation of materials could result in leakage that, in some cases, could have catastrophic effects on engine operation. Sealing companies have developed proprietary test methods to evaluate the ability of gasket materials to perform under those conditions.

5.1.3 Antistick capabilities

Antistick capabilities are necessary for gasket removal during engine repairs. If the gasket material sticks to the flanges, scraping off the materials may cause damage to the mating flanges. More importantly, the combination of surface finish with aluminum flanges during heating and cooling cycles could produce a scraping effect on the gasket, resulting in gasket failure.

Various test methods have been developed to determine antistick properties of different materials under various environments. ASTM 607 provides a means of determining the degree to which the gasket material, under compressive load, adheres to metal surfaces. Other methods exist that could determine the friction coefficient under different clamp loads.

5.1.4 Stress versus strain

Stress versus strain, usually called *stress versus compression*, is important as adequate stress is required at various locations. Stress and compression of materials are performed in a load frame. One example of this test equipment is the Instron (Figure 34). It is mechanically



Figure 34. Instron with mechanical jaws set for tensile test. (Photo courtesy of Instron®.)

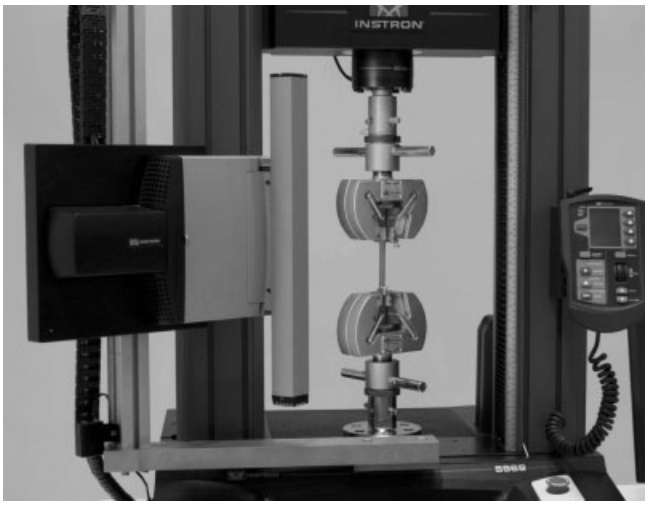


Figure 35. Environmental Chamber. (Photo courtesy of Instron®.)

actuated allowing for the compression or stretch of materials. Adding an environmental chamber, Figure 35 would make the instrument more versatile as it would give the capability of testing materials at different temperatures.

To improve the accuracy of the test, a strain gage is added to the system. As load is increased, strain is not influenced by machine deflection, but strain is measured directly as the material is deformed. This is shown in Figure 36.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

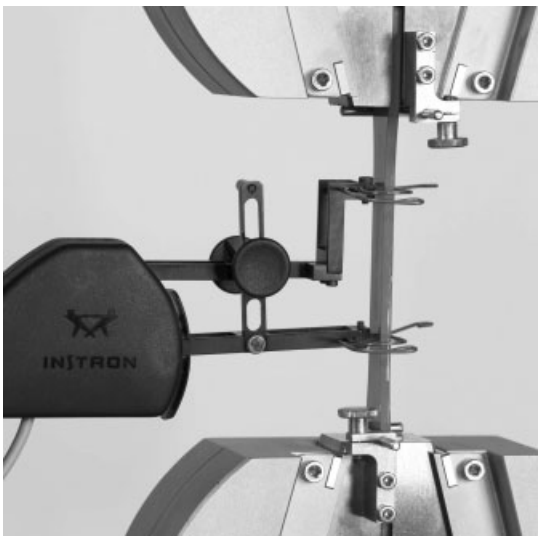


Figure 36. Strain Gage for improved accuracy of measuring the strain. (Photo courtesy of Instron®.)

5.1.5 *Compressibility and recovery*

The recovery is important because the gasket material has to follow flanges movement and deformation. The ASTM A36 test determines the short-time compressibility and recovery of gasket materials at room temperature. Its limitation comes from the fact that most materials' properties change in time. Long-term tests have to be added during the development to ensure gasket performance over the life of the application.

5.1.6 *Creep relaxation and compression set*

Creep relaxation and its relation to torque loss have been discussed earlier. ASTM F38 "Standard Test Method for Creep Relaxation of Gasket Material" is the standard test.

The compression set is determined by measuring a gasket thickness, clamping it in a fixture, subjecting the assembly to a specified temperature profile, disassembling the fixture, and measuring the recovered thickness of the material.

Compression thickness is calculated using the relationship:

$$\frac{T_o - T_f}{T_o} \times 100 = \text{Percent compression set} \qquad (1)$$

where T_o is the original thickness and T_f is the final thickness.

The compression set results in torque loss, thus resulting in stretch loss of the clamping bolts

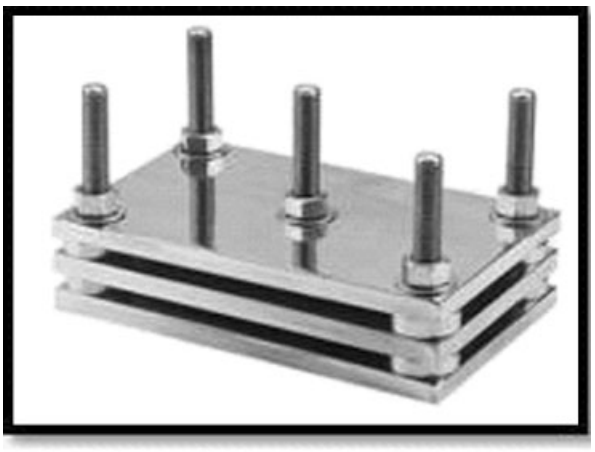


Figure 37. Split resistance fixture.

5.1.7 *Crush and extrusion resistance*

In some cases, certain locations on the gaskets may be subject to very high stresses that may create extrusion. An example may be the narrow area between two adjoining cylinders. Testing for such a phenomenon requires compression test machines and may require specially designed fixtures. One method for testing materials is shown in Figure 37. The test is accomplished using small-diameter steel wire to apply localized stress on materials. In addition, high temperature and sealed-medium conditions are simulated by presoaking the material and then heat soaking the material.

5.2 Gasket testing

Once the material's compatibility to the application and environmental conditions are verified, the designer determines gasket construction, dimensions, and shape. Prototype samples are made and gasket testing begins.

There are two types of testing that should be done:

- Bench testing
- Application or engine testing.

5.2.1 *Bench testing*

Bench testing uses components such as mating flanges along with the gasket, assembled and subjected to different tests to simulate application conditions. Bench testing should be performed ahead of engine tests to weed out gasket designs. Some of these tests are presented in the following sections.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5.2.1.1 Steam tests. Steam tests are used to perform short-duration thermal cycles. Saturated steam and cold water are alternatively circulated through the cooling passages.

5.2.1.2 Vibration tests. Vibration tests are used to simulate actual unit vibration. Such tests are very useful for oil pan and transmission pan gaskets. The vibration table, or shaker, is equipped to load the assembly with specific programs of vibration frequency and amplitude. The unit can be installed in an environmental chamber where the temperature can simultaneously be cycled.

5.2.1.3 Oven test. Hot oven and environmental chambers are used to evaluate sealing performance of gaskets. Some environmental chambers are large enough to accommodate an entire engine and have cycling temperature capabilities from -40°C to 315°C. Environmental chambers are also used to study the effect of flange expansion and antistick capabilities of some coatings.

5.2.2 Application tests

5.2.2.1 Dynamometer test. Automotive and gasket companies have developed several dynamometer tests specifically to evaluate gaskets and seals. They are component driven; several tests were developed specifically for the cylinder head gasket. As an example is the "200-cycle deep thermal cycle" test, developed to evaluate designs under very severe temperature conditions. The coolant is cooled to -28°C and then the engine starts and runs until the temperature reaches 104°C. The engine then stops and coolant is cooled to -28°C. The cycle is designed for a duration of 30 min, and the test runs for 100 h. The most severe part of the test is when, during the heat up the thermostat opens and coolant at -28°C hits the hot gasket. At the end of the test, the combustion chamber is pressurized to 76 bar. It has to be maintained pressure for 10 min. The coolant system is pressurized to 8 bar. The pressure must be maintained for 10 min.

5.2.2.2 Vehicle testing. Vehicle testing is typically done by the automobile manufacturer. Some automotive companies select, statistically, vehicles as they come off the assembly line and form groups of vehicles (fleets). One fleet will be send to a warmer climate and one fleet to a colder climate. Those vehicles are usually leased for up to 160,000 km. As they are retrieved at the end of the lease, the vehicles are returned to the manufacturer and disassembled, and the parts are evaluated. Problem parts are returned to the suppliers and analyzed, and prog rams to improve and resolve issues are put in place.

6 NEW METHODS OF EVALUATING AND ANALYZING DESIGNS

6.1 Failure mode and effect analysis

Potential failure mode and effect analysis (FMEA) is a systemized group of activities intended to:

1. Recognize and evaluate potential failure modes and causes associated with the design and manufacturing of a product.
2. Identify the actions that could eliminate or reduce the chance of potential failure occurring.
3. Document the process. Determines ahead of time what a design must do.

Engineers used the FMEA since the post-World War II era of the 1950s. This era is known for rapid technological advances, but product reliability did not keep up with the advancing technology. At that time, it was primarily used in the aerospace industry (Figure 38).

The main elements are

- failure mode
- possible cause
- effect of failure
- design verification.

They are applied in the order shown in Figure 39.



Figure 38. Environmental chamber.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

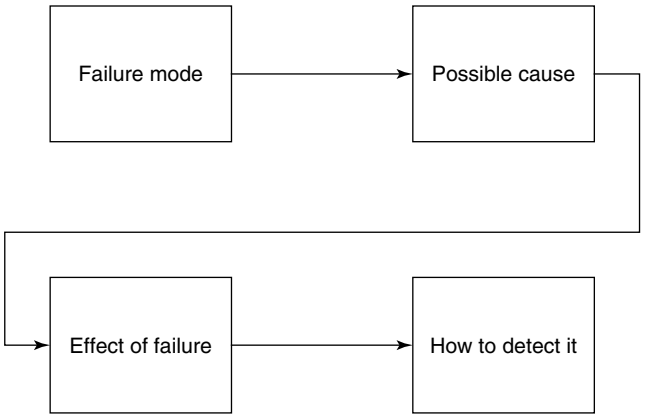


Figure 39. FMEA methodology.

In evaluating the elements, one needs to determine the risk of a failure. The risk is assessed as follows:

- severity
- occurrence
- probability of detection
- RPN—risk priority number.

For each potential failure, a risk number is associated with severity, occurrence, and probability of detection; by multiplying those numbers, an RPN is established. The team will work on the potential failure with the higher numbers.

For each potential failure, a test must be established that would verify the design. At the end of the process, the team tallies the entire test program. This becomes the Design Verification Process and Report (DVP&R), which is the test program to verify the design.

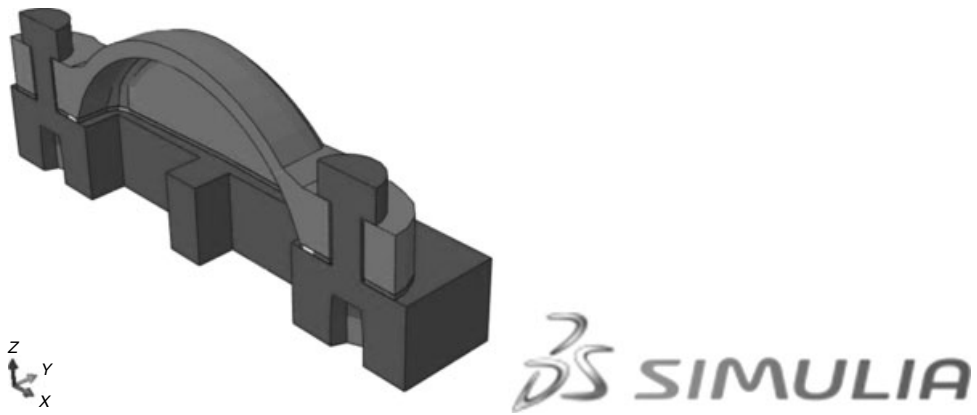


Figure 40. Modeling coolant assemblage with a gasket sandwiched. (Courtesy of SIMULIA.)

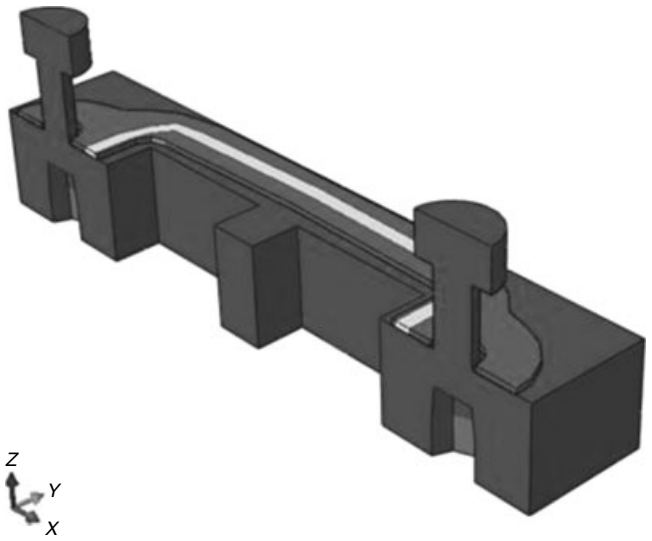


Figure 41. Gasket with the silicone bead represented by gray color. (Courtesy of SIMULIA.)

It is a systematic approach very useful, if it is done right, for younger, less experienced engineers, and designers. It is also useful when new approaches, materials, or designs are being developed, and there is little or no previous experience. FMEA is used for manufacturing processes as well.

6.2 Finite element analysis

During the design process, many alternative designs and options are identified. Testing all those designs is expensive and time-consuming. As competition pushes us to develop products in record time, new methods of analyzing designs and reducing the number of design iterations are needed.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

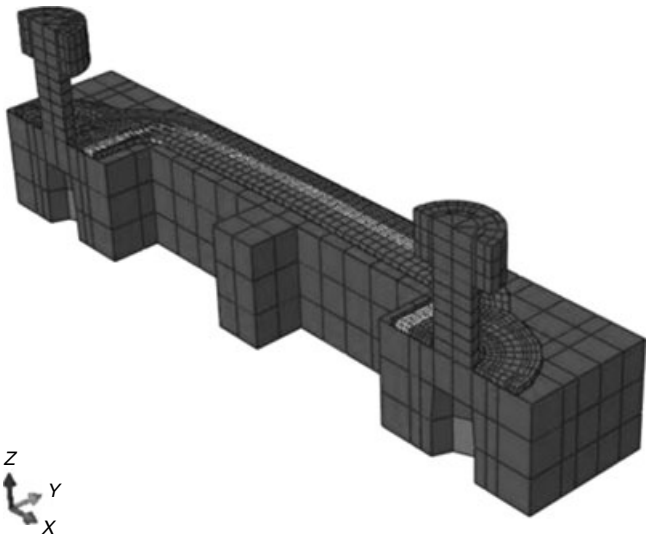


Figure 42. Gasket modeled with the gasket elements. (Courtesy of SIMULIA.)

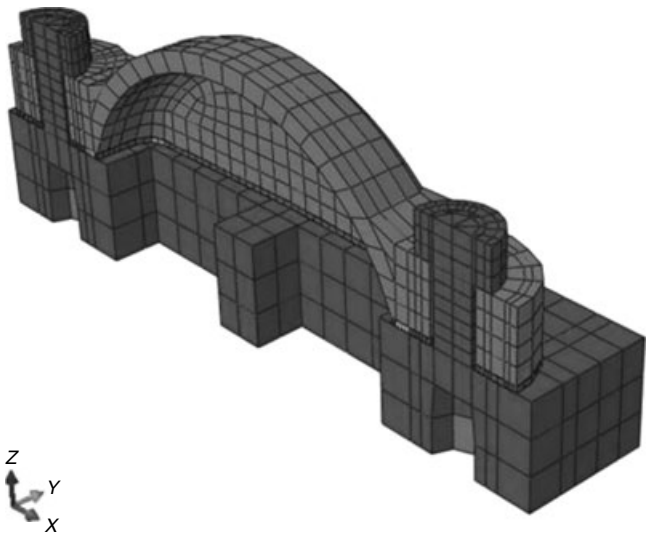


Figure 43. Finite element mesh of the model and the gasket elements. (Courtesy of SIMULIA.)

FEA can help reduce the number of designs by simulating the gasket design and understanding interactions with the mating parts. Designs and materials can be changed in the model so design ideas can be screened without necessitating making real parts and testing them.

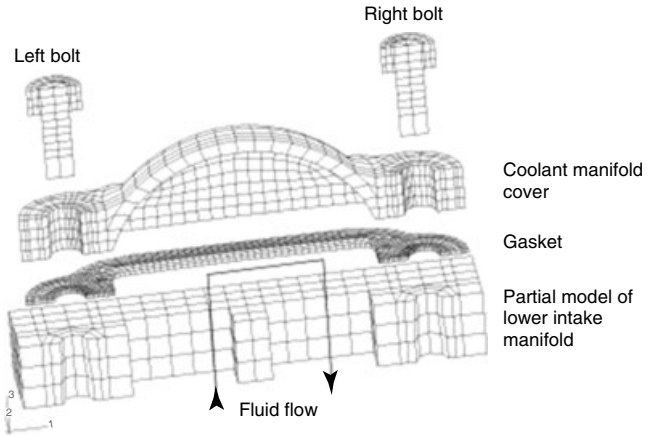


Figure 45. Coolant manifold assemblage. (Courtesy of SIMULIA.)

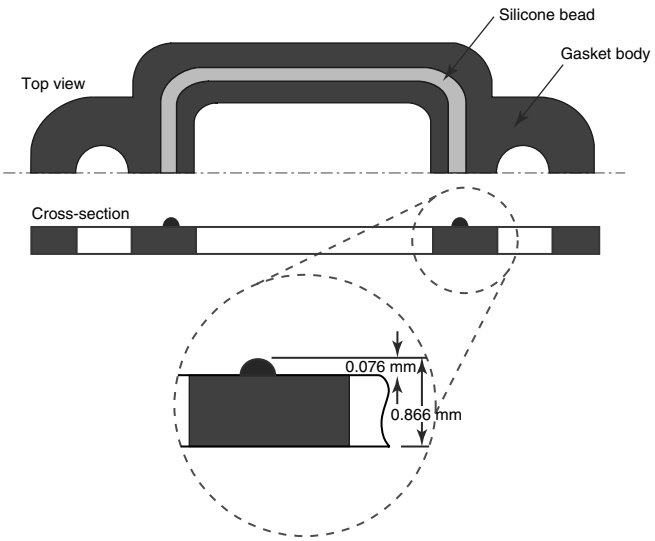


Figure 46. Schematic representation of a silicone bead indicated. (Courtesy of SIMULIA.)

The following is an example of an FEA application to coolant gasket modeling.

The gasket has to seal coolant in an assembly shown in Figure 40. Because the distance between the bolts is large, a flat gasket may not be able to seal in the center region. Therefore, the designer proposed a silicone bead to be added to the gasket as shown in Figure 41. The objective

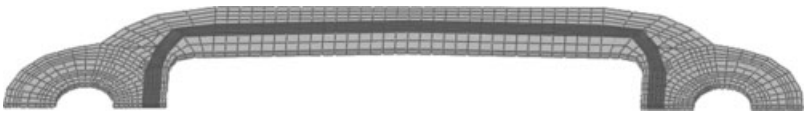


Figure 44. Mesh of gasket with silicone bead highlighted. (Courtesy of SIMULIA.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 47. Contact pressure after the initial fastening. (Courtesy of SIMULIA.)

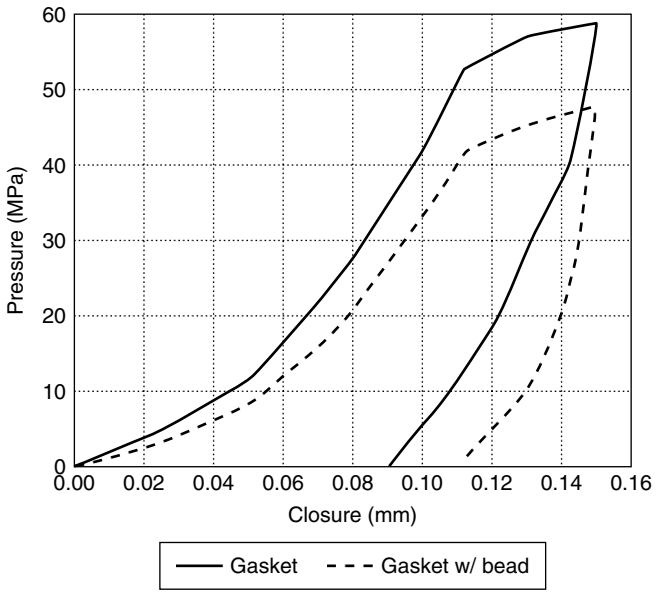


Figure 48. Pressure versus closure behavior for the gasket and gasket with silicone bead. (Courtesy of SIMULIA.)

was to determine the dimensions for the silicone bead and the thickness of the lower flange.

After the model was created, the next step was to create the mesh of the components (Figure 42) and the mesh of the assembly (Figures 43 and 44). Figure 45 shows the passage of coolant through the assembly and the area that needs to be sealed. As the design iterations progress, the bead height is developed and the height is established as shown in Figure 46.

The bead height is tested by simulating the assembly during initial fastening (Figure 47). Loading is represented by color, with higher loading represented by red and lower loading by blue.

Figure 48 shows the loading required to seal with and without the silicone bead. It is clear that adding the silicone bead reduces the required load and therefore reduces bending thus improving sealing. The bead had a recovery capability built into the structure and was expected to perform better than the gasket only design.

The FEA model should be tested under bench test to prove its validity. Companies developing FEA software are maturing in the industry. Special attention must be taken when choosing the software to run FEA because, as discussed earlier, besides the materials' stress–strain curves that could be generated easily today, one must consider the time-dependent nature of these materials. This is not easily simulated by all of the software programs that are in the market today.

FURTHER READING

Bickford, J.H. (1981) *An Introduction to the Design and Behavior of Bolted Joints*, Marcel Dekker.

Brown, M. (1995) *Seals and Sealing Handbook*, 4th edn, Elsevier.

Bouquet, F.L. (1988) *Introduction to Seals and Gaskets Engineering*, Systems Co.

Christensen, R.M. (2004) *Mechanics of Composite Materials*, 2nd edn, Wiley.

Czernik, E.D. (1996) *Gaskets Design, Selection, and Testing*, McGraw Hill.

Goldman, Y.A. (1994) *Prediction of Deformation Properties of Composite Materials*, American Chemical Society.

Miszczak, F.L. and Silvian, L.M. (1990) *Fluid Sealability of Gaskets Materials: New Test Fixture, Instrumentation, and Test Results*, Society of Automotive Engineers: SAE technical Paper 90115.

Mummery, L. (1992) *Surface Texture Analysis: The Handbook*, Hommelwerke GmbH, Germany.

Parker O-Ring Handbook, ORD5700AUSA. ASTM Publication Code Number (PCN: 03-603093-20, Director, Editorial Services: Roberta A. Storer and others.

Parmley, R.O. (1996) *Standard Handbook of Fastening and Joining*, 3rd edn, McGraw Hill.

Shah, V. (1998) *Handbook of Plastic Testing Technology*, Wiley.

Storer, R.A. (1993) *ASTM Standards on Gaskets*, 6th edn. ASTM Publication Code Number (PCN: 03-603093-20, Director, Editorial Services: Roberta A. Storer and others.

Ward, I.M. (1983) *The Mechanical Properties of Polymers*, 2nd edn, Wiley.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto040
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

General Requirement of Traction Motor Drives

Ming Cheng[1] and C.C. Chan[2]

[1]*Southeast University, Nanjing,China*
[2]*The University of Hong Kong, Pokfulam, Hong Kong*

1 Introduction 1
2 Classification 3
3 Design Consideration of Traction Motor 8
4 Control Consideration of Traction Motor Drive 12
5 Conclusion 16
Related Articles 17
References 17
Further Reading 18

1 INTRODUCTION

The traction motor drive is the heart of electric vehicles (EVs). Its role is to convert electric energy to mechanical energy or vice versa, thus to interface energy source (such as batteries) with vehicle wheels. In motor mode, the electrical energy from the battery is converted to mechanical energy such that the vehicle overcomes aerodynamic drag, rolling resistance drag, and inertia resistance. In generator mode, it converts mechanical energy to electrical energy such that the kinetic energy released during vehicle deceleration is converted to electrical energy to charge the battery. Hence, the electric motor drive is the core technology for electric, hybrid, and fuel cell vehicles.

It should be emphasized that traction motors are different from traditional industrial motors due to the fact that traction motors usually require frequent start/stop, high rate of acceleration/deceleration, high torque low speed hill climbing, low torque high speed cruising, and very wide-speed range of operation, whereas industrial motors are generally optimized at the rated conditions. Thus, traction motors are so unique that they are deserved to form an individual class. Hence, the general requirements of traction motor are significantly different from those of industrial motors. Their major differences in load requirement, performance specification, and operating environment are as follows (Chan and Chau, 2001; Chau, Chan, and Liu, 2008; Zhu and Howe, 2007):

- Traction motors need to offer four to five times the rated torque for temporary acceleration and hill climbing, whereas industrial motors generally offer twice the rated torque for overload operation.
- Traction motors need to achieve four to five times the base speed for highway cruising, whereas industrial motors generally achieve up to twice the base speed for constant-power operation, where the base speed is the speed at which the motor delivers the rated torque with the rated voltage.
- Traction motors should be designed according to the vehicle driving profiles and drivers' habits, whereas industrial motors are usually based on a typical working mode.
- Traction motors demand both high power density and good efficiency map (high efficiency over wide speed and torque ranges) for the reduction of total vehicle weight and the extension of driving range, whereas industrial motors generally need a compromise among power density, efficiency, and cost with the efficiency optimized at a rated operating point.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

- Traction motors desire high controllability, high steady-state accuracy, and good dynamic performance for multiple-motor coordination, whereas only special-purpose industrial motors desire such performance.
- Traction motors need to be installed in mobile vehicles with harsh operating conditions such as high temperature, bad weather, and frequent vibration, whereas industrial motors are generally located in fixed places.

Thus, the general requirements of the traction motor drives can be summarized as follows:

1. high torque density and power density;
2. very wide speed range, including constant-torque and constant-power regions;
3. high efficiency over wide torque and speed ranges;
4. high torque for low speed starting and climbing and high power for high speed cruising;
5. fast torque response;
6. multiquadrant operation ability, including forward motoring, forward braking, backward motoring, and backward braking;
7. high reliability and robustness for vehicular environment;
8. low acoustic noise;
9. reasonable cost.

From the functional point of view, a traction motor drive can be divided into two parts—electrical and mechanical. The electrical part consists of the subsystems of motor, power converter, and electronic controller, whereas the mechanical part includes the subsystems of mechanical transmission (optional) and vehicle wheels. The boundary between the electrical and mechanical parts is the air-gap of the motor, where electromechanical energy conversion takes place. The power converter supplies the motors with proper voltage and current and regulates the power flow between the energy source and the electric motor for motoring and regeneration. The electronic controller commands the power converter by providing control signals to it, and then controls the operation of the electric motor to produce proper torque and speed, according to the command from the driver. The electronic controller can be further divided into three functional units—sensor, interface circuitry, and processor. The sensor is used to translate the measurable quantities, such as current, voltage, temperature, speed, torque, and flux, into electronic signals through the interface circuitry. These signals are conditioned to the appropriate level so as to be fed into the processor. The processor output signals are usually amplified via the interface circuitry to drive power semiconductor devices of the power converter. The functional block diagram of a traction motor drive is shown in Figure 1.

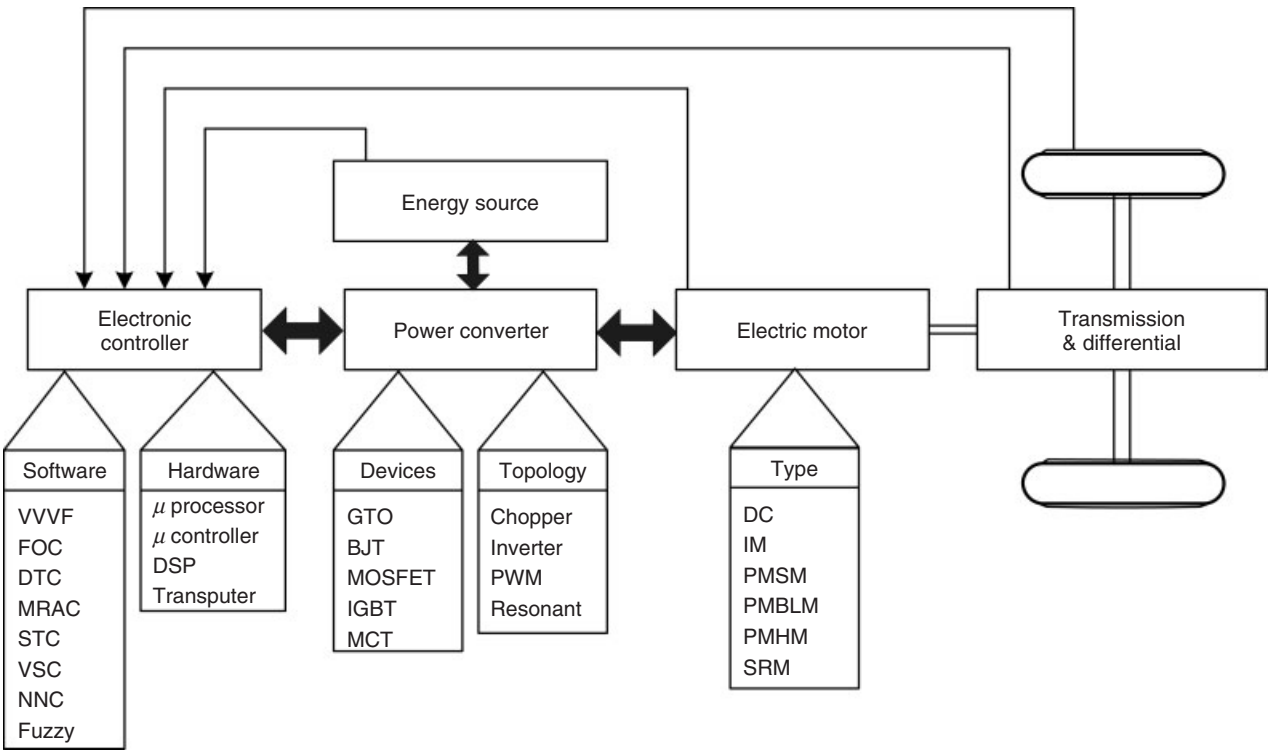


Figure 1. Functional block diagram of a traction motor drive.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2 CLASSIFICATION

As illustrated in Figure 2, those traction motors applicable to EVs can be classified into two main groups, namely the brushed motors and the brushless motors. The former simply denote that they generally consist of the commutator and brushes, mainly traditional DC (direct current) motors, whereas the latter have no brushes.

2.1 DC motor

Traditionally, DC brushed motors have been loosely named as DC motors. There are typically four types of wound-field DC motors, depending on the mutual interconnection between the field and armature windings, namely separately excited, shunt excited, series excited, and compound excited. By replacing the field winding of DC motors with permanent magnet (PM), PM DC motors are generated, which permit a considerable reduction in stator diameter due to the efficient use of radial space and an increase in motor efficiency due to the elimination of the copper loss in field windings. Owing to the low permeability of PMs, armature reaction is usually reduced and commutation is improved. The control principle of DC motor is simple because of the orthogonal disposition of field and armature magnetomotive forces (mmfs).

However, the principle problem of DC motors, due to their commutators and brushes, makes them less reliable and unsuitable for maintenance-free operation and high speed. In addition, winding-excited DC motors have low specific power density. Nevertheless, because of their ability to achieve high torque at low speeds and because they are easy to control, DC motors have ever been prominent in the electric propulsion system. Actually, various types of DC motors, including series, shunt, separately excited, and PM excited, have ever been adopted by EVs.

Recently, technological developments have pushed brushless motors to a new era, which offer the advantages of higher efficiency, higher power density, lower operating cost, increased reliability, and being maintenance-free over DC brushed motors. Thus, brushless motors have now become more attractive in traction motor drive for EVs.

2.2 Induction motor

The induction motor (IM) is a widely accepted brushless motor for EV application because of its robust structure, low cost, high reliability, high efficiency, and free from maintenance as compared with the DC motor drive. There are two types of IMs, namely, wound-rotor and squirrel-cage motors. Because of the high cost, need for maintenance, and lack of sturdiness, wound-rotor IMs are less attractive than their squirrel-cage counterparts, especially for electric propulsion in EVs. Hence, the most common types of rotors for IM are the squirrel cage in which aluminum bars are cast into slots in the outer periphery of the rotor. The aluminum bars are short-circuited together at both ends of the rotor by cast aluminum end rings, which also can be shaped as fans. Figure 3 shows the cross section of an IM.

An inverter is used to control the motor so that the desired torque can be delivered for a given driving condition at a certain speed. Advanced control methodologies, such as field-oriented control (FOC) or vector control and direct torque control (DTC), are popular in IM control for traction applications.

The main advantages of IM include:

1. Robust structure and relatively low cost;
2. Light weight, small volume, and high efficiency.

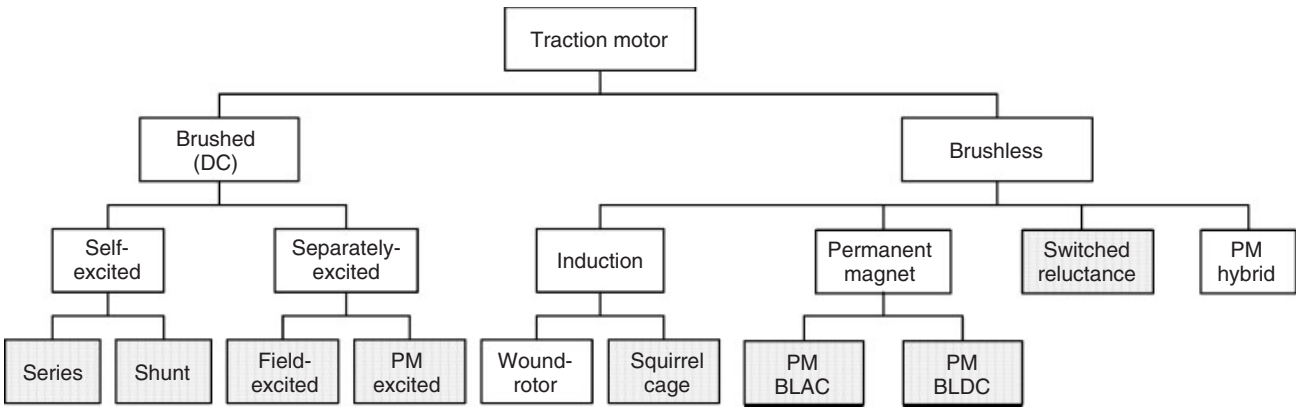


Figure 2. Classification of traction motors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

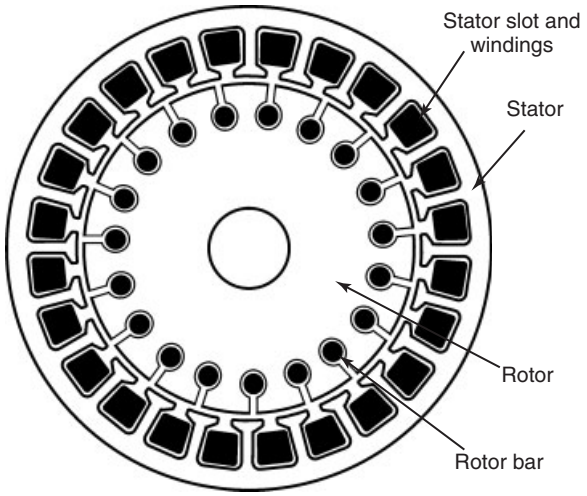


Figure 3. Induction motor with squirrel cage.

The disadvantages include:

1. The limited constant-power range (only 2–3 times the base speed);
2. Relatively difficult control schemes due to the variable equivalent parameters.

2.3 Permanent magnet brushless motor

Permanent magnet brushless motors (PMBLMs) include sinusoidal and trapezoidal back electromotive force (EMF) machines. From control point of view, they are divided into brushless direct current (BLDC) and brushless alternate current (BLAC) motors. Generally, a trapezoidal back EMF waveform in BLDC or a sinusoidal back EMF waveform in BLAC is needed so as to achieve high torque density and low torque pulsation. The PM BLAC motor with sinusoidal back EMF is also called *PM synchronous motor*. As they are essentially synchronous motors, the PM BLAC motor can run from a sinusoidal or pulse width modulation (PWM) supply without electronic commutation.

When PMs are mounted on the rotor surface, they behave as non-salient synchronous motors because the permeability of PMs is similar to that of air. By burying those PMs inside the magnetic circuit of the rotor, the saliency causes an additional reluctance torque, which leads to facilitating a wider speed range at constant-power operation. Figure 4 illustrates the typical topologies of the PM brushless motors. Similar to IMs, those PM synchronous motors usually employ FOC or DTC for high performance applications. Because of their inherent high power density and high efficiency, the PM motors are the choices for traction motor drives in EV applications.

The PM BLDC motors are fed by rectangular alternate current (AC) and hence are also called *rectangular-fed PM brushless motors*. The most obvious advantage of these motors is the removal of brushes, leading to elimination of many problems associated with brushes. The PM BLDC motor has surface-mounted magnets on the rotor and a concentrated fractional stator winding, which results in a low copper loss. Different from PM synchronous motors, these PM BLDC motors generally operate with shaft position sensors. Recently, sensorless control technologies have been developed.

The main advantages of PM brushless motors are:

1. Light weight, small volume, and high power density as the magnetic field is excited by high energy PMs.
2. High efficiency and high reliability.

The main disadvantages include:

1. Comparatively narrow range of constant-power operation due to the difficulty in weakening the air-gap flux. By using some new schemes, the speed range can reach three times the base speed. However, the PM may suffer from demagnetization and possible fault.
2. Relatively high cost due to PM materials, especially in high power application.

It should be emphasized that all the PM machines mentioned above have the magnets located in the rotor and are referred to as *rotor-PM machines*, which are predominated in EV applications due to their outstanding advantages. However, the magnets usually need to be protected from the centrifugal force by employing a retaining sleeve made of either stainless steel or non-metallic fiber. The rotor temperature rise may be a problem due to poor thermal dissipation, which may cause irreversible demagnetization of magnets and ultimately limit the power density of the machine. Recently, in contrast, a new type of PM machines having magnets in stator, referred to as *stator-PM machines*, have reemerged and developed, which can overcome the problems suffered by rotor-PM counterparts (Cheng *et al*., 2011). Conceptually, the stator-PM machines employ the polarized reluctance principle, in which the torque and EMFs are resultant from the flux-switching action of rotor saliencies on a unipolar flux produced by PMs in the stator. As there are no PMs or windings in rotor, these stator-PM machines are mechanically simple and robust, hence suitable for high speed operation. Compared with conventional rotor-PM brushless machine topologies, generally, it is easier to limit the temperature rise of the magnets as heat is dissipated more effectively from the stator. According to the location of the PMs in stator, they can be classified as the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

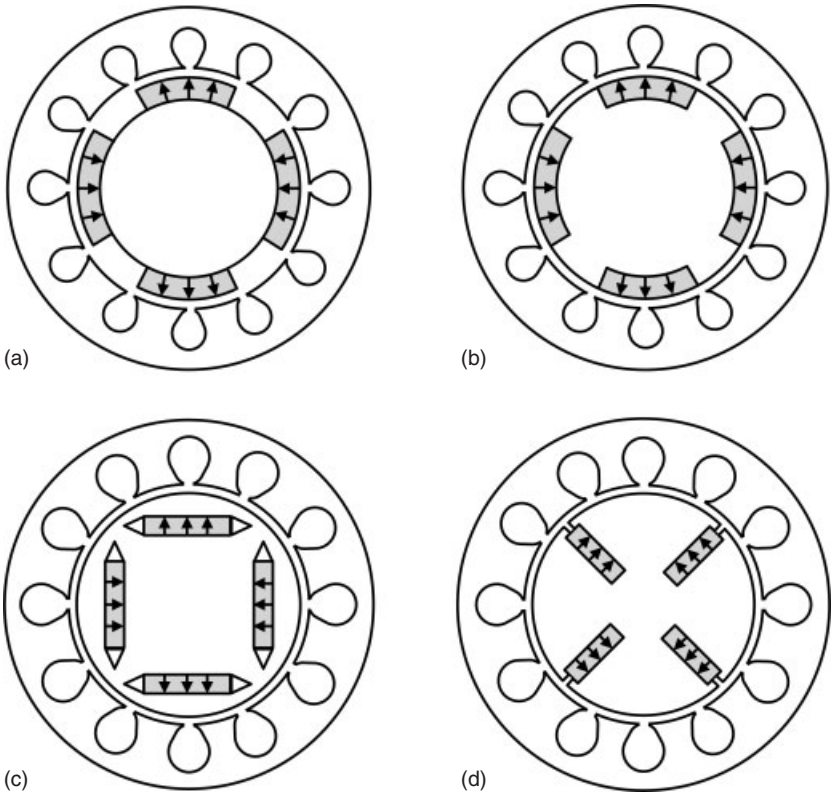


Figure 4. Typical topologies of PM brushless motors. (a) Surface mounted; (b) surface inset; (c) interior radial; (d) interior circumferential.

doubly salient permanent magnet (DSPM) machine (Liao, Liang, and Lipo, 1995; Cheng, Chau, and Chan, 2001), flux-reversal permanent magnet (FRPM) machine (Deodhar *et al*., 1996), and flux-switching permanent magnet (FSPM) machine (Hoang, Ben-Ahmed, and Lucidarme, 1997). They have been recognized to have considerable potential for EV applications.

1. *Doubly Salient PM Machine*. In this DSPM machine, the PMs are located in stator back-iron. Figure 5 shows a 12/8-pole DSPM machine topology (with 12 stator poles and 8 rotor poles). The variation of the flux linkage with each coil as the rotor rotates is unipolar, while the back EMF waveform tends to be trapezoidal. Thus, this topology is more suitable for

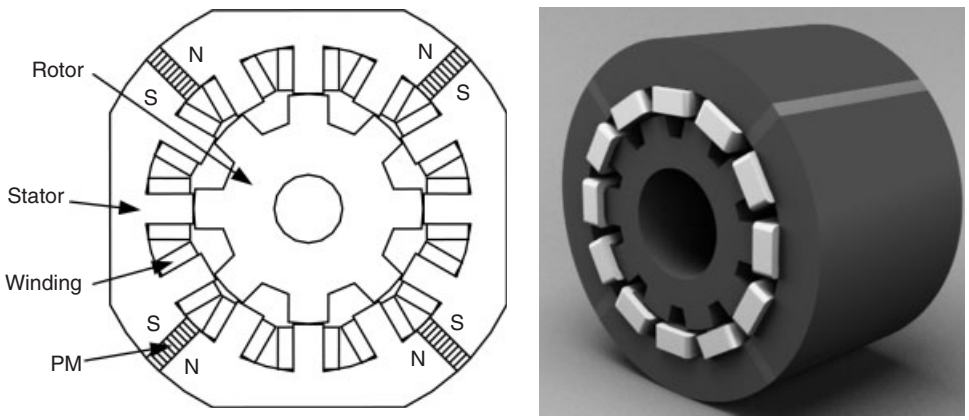


Figure 5. 12/8-pole DSPM machine.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

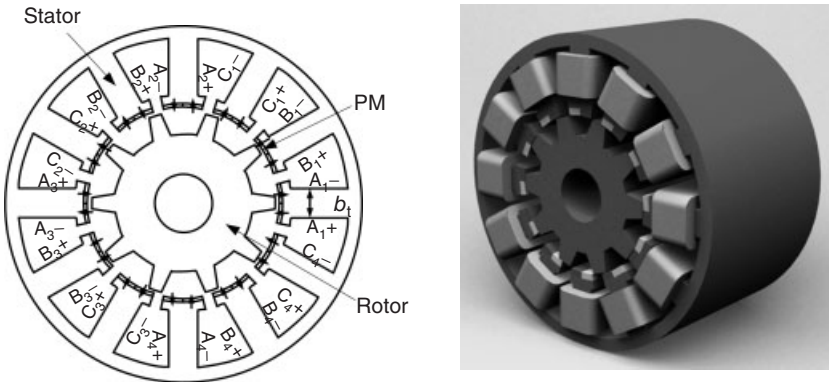


Figure 6. FRPM machine.

BLDC operation. However, a major disadvantage of the DSPM motor is relatively low torque density as compared to that of other PM brushless machines.

2. *Flux-Reversal PM Machine*. The FRPM machine has the magnets located on the surface of stator teeth and concentrated windings. Figure 6 illustrates a 12/10-pole FRPM machine topology. Each stator tooth has a pair of magnets of different polarity mounted at its surface. When a coil is excited, the field under one magnet is reduced while that under the other is increased, and the salient rotor pole rotates toward the stronger magnetic field. The flux linkage with each coil reverses polarity as the rotor rotates. Thus, the phase flux linkage variation is bipolar, whereas the phase back EMF waveform is, again, essentially trapezoidal. Thus, it is also suitable for BLDC operation mode. Additionally, the FRPM machine offers fault-tolerance capability due to its natural isolation between the phases. Such a machine topology exhibits a low winding inductance, while the magnets are more vulnerable to partial irreversible demagnetization. In addition, significant eddy current loss may be induced in the magnets, which also experience a significant radial magnetic force. Furthermore, as the air-gap flux density is limited by the magnet remanence, the torque density may be compromised.
3. *Flux-Switching PM Machine*. In this FSPM machine, the stator consists of U-shaped laminated segments between which circumferentially magnetized PMs are sandwiched, while the direction of magnetization is being reversed from one magnet to the next. Figure 7 shows a 12/10-pole FSPM machine topology. Each stator tooth consists of two adjacent laminated segments and a PM. Thus, flux-concentration can be readily incorporated, so that low cost ferrite magnets can be employed (Zhu and Howe, 2007). In addition, in contrast to conventional PM brushless machines, the influence of the armature reaction field on the working point of the magnets is minimal. As a consequence, the electric loading of FSPM machines can be very

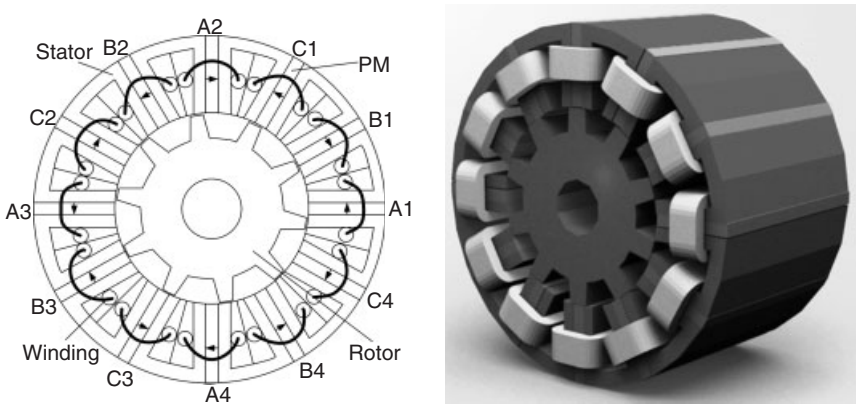


Figure 7. FSPM machine.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

high. Therefore, as the phase flux linkage waveform is bipolar, the torque capability is significantly higher than that of a DSPM machine (Hua *et al*., 2005). Due to the reluctance difference between the two pairs of coils composing a phase, the resultant phase EMF waveforms are essentially sinusoidal without any additional measures (Hua *et al*., 2007), making them more appropriate for BLAC operation. In addition, as a high per unit winding inductance can readily be achieved, such machines are eminently suitable for constant-power operation over a wide speed range.

2.4 PM hybrid motor

Although the PM brushless motors possess the highest efficiency and power density over the others, they suffer from a difficulty in flux control. Hence, the current phase angle has to be progressively advanced as the speed is increased above the base speed so that a demagnetizing d-axis current component is produced which reduces the flux linkage. Ultimately, however, this may cause partial irreversible demagnetization of the magnets. At the same time, due to the inverter voltage and current limits, the torque-producing q-axis current component has to be reduced correspondingly. Consequently, the torque and power capabilities are limited (Zhu and Howe, 2007). Thus, a compromise has to be made between the low speed torque capability and high speed power capability. Hybrid PM and field current excitation has been shown to be beneficial in improving the power capability in the extended speed range, enhancing the low speed torque capability, and improving the overall operational efficiency. Figure 8 shows PM hybrid motors with rotary and stationary PMs, respectively (Chan *et al*., 1996; Zhu and Cheng, 2010). The PM hybrid motor is a special type of PM brushless motors. In this motor, an auxiliary DC field winding is so incorporated that the air-gap flux is a resultant of the PM flux and field-winding flux. These PM hybrid motors offer many attractive features due to the presence of the hybrid field:

1. By changing the polarity and magnitude of the DC field current, the air-gap flux density can be easily controlled.
2. By realizing flux strengthening, the machine can offer the exceptionally high torque, which is very essential for cold cranking HEVs (hybrid electric vehicles) or providing temporary power for vehicular overtaking and hill climbing.

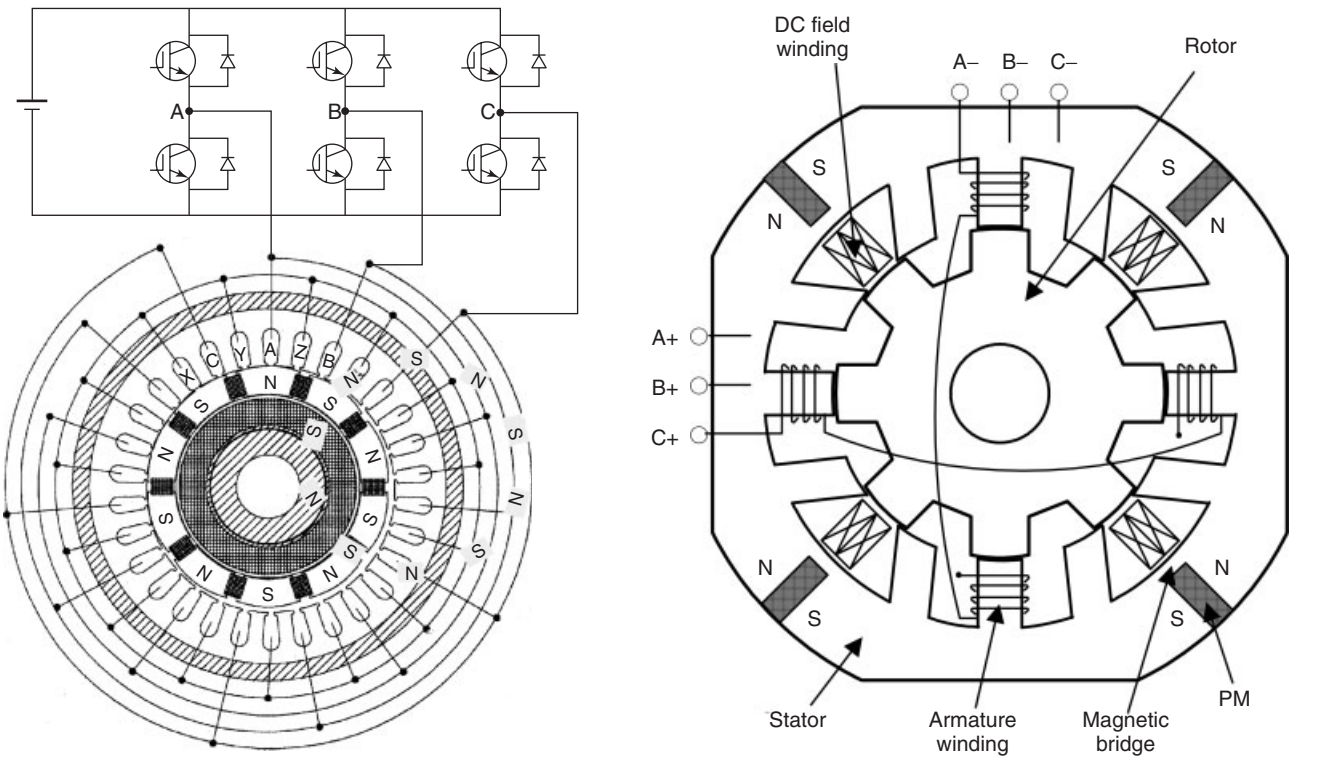


Figure 8. Hybrid PM machine. (a) Rotary PMs; (b) stationary PMs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3. By realizing flux weakening, the machine can offer the exceptionally wide-speed constant-power range, which is very essential for EV cruising.
4. By online tuning the air-gap flux density, the machine can maintain a constant voltage output under generation or regeneration over a very wide speed range, which is very essential for battery charging of various EVs.
5. By online tuning the air-gap flux density, the machine can also offer efficiency optimizing control (EOC), which is highly desirable for EVs.

2.5 Switched reluctance motor

The switched reluctance (SR) motors have been recognized to have considerable potential for EV applications. They have the definite advantages of simple construction, low manufacturing cost, inherent fault tolerance, and outstanding torque-speed characteristics for EV propulsion. Figure 9 shows the schematic of an 8/6-pole SR motor. Although they are simple in structure, it does not imply any simplicity of their design and control. Because of the heavy saturation of pole tips and the fringe effect of poles and slots, their design and control are difficult and subtle. Moreover, they usually exhibit relatively high acoustic noise, vibration, and torque ripple problems. Traditionally, the SR motors operate with shaft sensors to detect the relative position of the rotor to the stator. These sensors are usually vulnerable to mechanical shock and sensitive to temperature and dust, and thus reduce the reliability of the SR motors and constrain some applications. Recently, sensorless technologies have been developed for the SR motors.

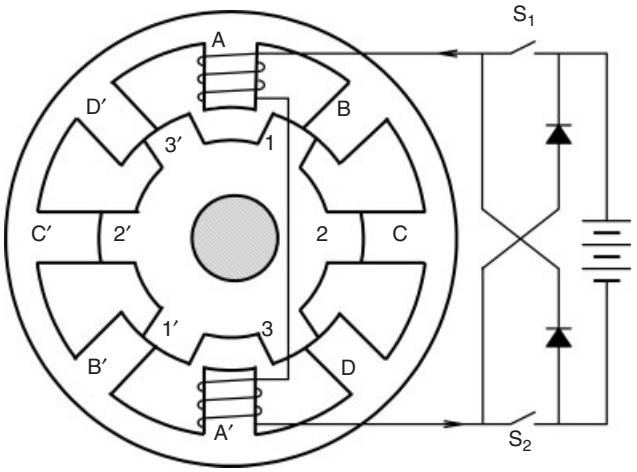


Figure 9. Basic structure of switched reluctance motor drive (only one phase winding shown).

Table 1. Applications of traction motors in EVs.

EV Models	EV Motors
Fiat Panda Elettra	Series DC motor
Mazda Bongo	Shunt DC motor
Conceptor G-Van	Separately excited DC motor
Suzuki Senior Tricycle	PM DC motor
Fiat Seicento Elettra	Induction motor
Ford Th!nk City	Induction motor
GM EV1	Induction motor
Honda EV Plus	PM synchronous motor
Nissan Altra	PM synchronous motor
Toyota RAV4	PM synchronous motor
Nissan Leaf	PM synchronous motor
Chloride Lucas	SR motor
Toyota Prius (2005)	PM BLDC motor
Honda Civic	PM BLDC motor

The motor types that have ever been adopted by recent EVs are indicated by shaded blocks in Figure 2. Table 1 also illustrates their recent applications in EVs.

In order to evaluate the aforementioned traction motor types, a point grading system is adopted. The grading system consists of six major characteristics and each of them is graded from 1 to 5 points. As listed in Table 2, this evaluation indicates that IMs and PM brushless motors are relatively most acceptable. When the cost of PM material has significant improvements, the PM brushless (including AC or DC) motors will be most attractive. Conventional DC motors seem to be losing their competitive edges, whereas both SR and PM hybrid motors have increasing potentials for EV propulsion.

3 DESIGN CONSIDERATION OF TRACTION MOTOR

3.1 Basic consideration

The basic consideration of motor design includes magnetic loading—the peak of fundamental component of radial flux density in the air-gap of the motor, and electric loading—the total rms current per unit length of periphery of the motor or ampere-turns per unit periphery; power per unit volume and weight; torque per unit volume and weight; flux density at each part of the magnetic circuit; speed, torque, and power; losses and efficiency; and thermal design and cooling.

The corresponding key issues are better utilization of steel, magnet, and copper; better electromagnetic coupling between stator and rotor; better geometry and topology; better thermal design and cooling; understanding the limits on the motor performance; and understanding the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Evaluation of traction motors.

	DC Motor	Induction Motor	PM Brushless Motor	SR Motor	PM Hybrid Motor
Power density	2.5	3.5	5	3.5	4
Efficiency	2.5	3.5	5	3.5	5
Controllability	5	4	4	4	4.5
Reliability	3	5	4	5	4
Maturity	5	5	5	4	3
Cost	4	5	3	4	3
Total	22	26	26	24	23.5

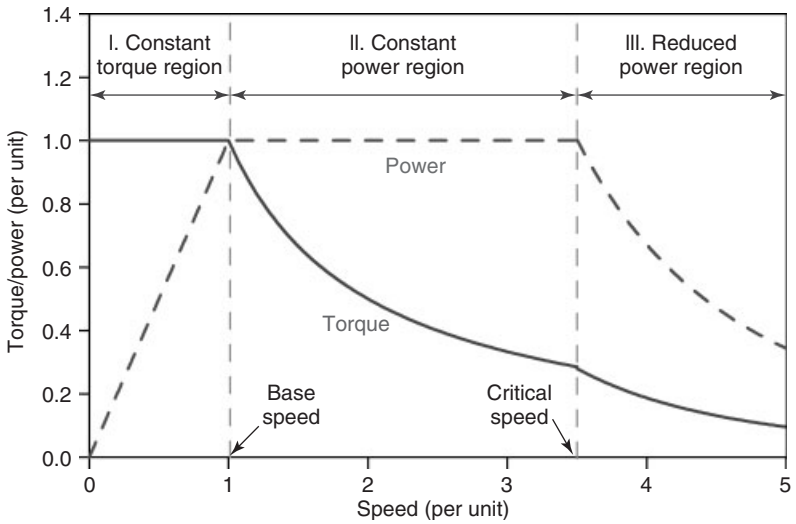


Figure 10. Ideal torque/power-speed characteristics of traction motor.

relationship among geometry, dimensions, parameters, and performance, thus to achieve higher power per unit weight, higher torque per unit weight, and better performance.

Traction motor drives for EVs should be designed, as close as possible, to the ideal torque/power-speed characteristics as shown in Figure 10. In the constant-torque region I, the maximum torque capability is determined by the current rating of the inverter, while in the constant-power region II, flux-weakening or commutation phase advance has to be employed due to the inverter voltage and current limits. In region III, the torque and power are reduced due to the increasing influence of the back EMF.

3.1.1 Sizing equation

The first important task in the design of a traction motor is to calculate the size of the motor. In the following, a PM motor will be taken as an example to illustrate the key points of design.

Neglecting the stator resistance, the input power of a PM motor can be expressed as

$$P_1 = mIV\cos\varphi = mIE_0\cos\delta \tag{1}$$

where φ is the power factor angle and δ is the inner power angle (the angle between the back EMF and the current). In addition, the back EMF can be expressed as

$$E_0 = \sqrt{2}\pi f K_w W \Phi \tag{2}$$

where W is the number of winding turns in series per phase, Φ is the total air-gap flux per pole, and K_w is the winding factor. Substituting Equation 2 into Equation 1 yields

$$P_1 = mE_0 I\cos\delta = mI\cos\delta\cdot\sqrt{2}\pi f K_w W\Phi \tag{3}$$

The output power of the traction motor can be obtained by multiplying the input power with efficiency η,

$$P = \eta P_1 = \sqrt{2}\pi\eta mf K_w WI\Phi\cos\delta \tag{4}$$

The electric loading (or linear current density along the inner stator surface) is A (A/m) and the inner diameter of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the stator is D. Then

$$2mWI = \pi DA \tag{5}$$

The total flux per pole can be expressed in terms of air-gap flux density B:

$$\Phi = \frac{\pi Dl}{2p} B\alpha \tag{6}$$

where p is the number of pole pairs, l is the stack length of the stator, α is the pole arc factor which equals to the ratio of pole enclosure to the pole pitch, and B is the air-gap flux density. To achieve the maximum power, the inner power angle δ can be set to 0. Substituting Equations 5, 6 and $f = pn/60$ into Equation 4 yields

$$P = m\eta \frac{\pi DA}{2mW} \sqrt{2}\pi K_{\mathrm{w}} \frac{pn}{60} W \frac{\pi Dl}{2p} B\alpha \tag{7}$$

Therefore, the sizing equation for PM motors can be obtained as

$$D^2 l = \frac{60}{\sqrt{2}\pi^3} \frac{4}{\alpha\eta K_{\mathrm{w}}} \frac{P}{AB} \frac{1}{n} \tag{8}$$

For other type of traction motors, similar sizing equation can be derived. As K_{w}, A, and B are in a relatively narrow range for all types of motors, Equation 8 shows that the effective volume of a motor is proportional to power P and inversely proportional to speed n. Considering that $P = \frac{2\pi n}{60} T$, we have

$$D^2 l \propto \frac{P}{n} \propto T \tag{9}$$

where T is the torque. In other words, the size of an electric motor is proportional to its torque rating.

3.1.2 Selection of A and B

In Equation 8, both A and B are experience-based selections. The magnetic loading B shows the utilization of magnetic material (silicon steel) and its value is limited by the magnetic saturation in teeth and yoke and the iron loss. A higher B means less magnetic material but higher magnetic losses. The electric loading A shows utilization of electric material (copper or aluminum). A higher A means less copper material but higher electric losses. Ambient temperature, operating frequency, and cooling method can impact the selection of A and B (Mi, Masrur, and Gao, 2011).

Typical range for A is 10 kA/m for small air-cooled motors, and up to 100 kA/m for liquid-cooled motors. The typical range for B is about 0.4 T for small motors, and up to 1.2 T for high density motors. Generally, large motors will have larger values of A and B.

3.1.3 Speed rating of the traction motor

It can be seen from Equation 9 that the motor volume is inversely proportional to rotor speed. Hence, a higher speed rating means a smaller motor size. However, a higher speed means a higher operating frequency, which results in more magnetic losses (eddy current and hysteresis losses). Smaller values of A and B may be necessary to limit the loss in high speed motors. For example, a four-pole 1500 r/min motor operates at 50 Hz, but a four-pole 15000 r/min motor operates at 500 Hz. As eddy current loss is proportional to f^2, and hysteresis loss is proportional to f^{α} $(1 < \alpha < 2)$, if the same magnetic flux density is chosen for the two motors, then the losses in the high speed motor will be many times that of the low speed one even if the size of the high speed motor is much smaller. This is because the eddy current loss increases 100 times, but size ($D^2 l$) reduces by only factor of 10 (Mi, Masrur, and Gao, 2011).

3.2 System consideration

Electrical machine design cannot be undertaken in isolation, but must account for the control strategy and the application requirements. Hence, a system-level design approach is essential for traction motors.

Vehicle operation consists of three main segments. They are (i) the initial acceleration; (ii) cruising at vehicle rated speed; and (iii) cruising at the maximum speed. These three operations set the basic design constraints for the EV and HEV drivetrain.

Apart from satisfying the aforementioned special requirements, the design of traction motors also depends on the system technology of EVs. From the technological point of view, the following key issues should be considered (Chan and Chau, 2001):

1. *Single- or Multiple-Motor Configurations*. One adopts a single motor to propel the driving wheels, while another uses multiple motors permanently coupled to individual driving wheels. The single-motor configuration has the merit of using only one motor with the minimum corresponding size, weight, and cost. On the other hand, the multiple-motor configuration takes the advantages of reducing the current/power ratings of individual motors and evenly distributes the total motor size and weight. Moreover, the multiple-motor one needs additional precaution to allow for fault tolerance

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

during the electronic differential action. The comparison between single- and dual-motor configurations is listed in Table 3. As these two configurations have their merits, both of them have been employed by modern EVs. For examples, the single-motor configuration has been adopted in the GM EV1, whereas the dual-motor configuration has been adopted in the NIES Luciole.

2. *Fixed- or Variable-Gearing Transmissions*. It is also classified as single-speed and multiple-speed transmissions. The former adopts single-speed fixed gearing, while the latter uses multiple-speed variable gearing together with the gearbox and clutch. On the basis of the fixed-gearing transmission, the motor should be so designed that it can provide both high instantaneous torque (3–5 times the rated value) in the constant-torque region and high operating speed (3–5 times the base speed) in the constant-power region. On the other hand, the variable-gearing transmission provides the advantage of using conventional motors to achieve high starting torque at low gear and high cruising speed at high gear. However, there are many drawbacks on the use of variable gearing such as the heavy weight, bulky size, high cost, less reliable, and more complex. Table 4 gives a comparison of fixed-gearing and variable-gearing transmissions (Chan and Chau, 2001). Currently, almost all the modern EVs adopt fixed-gearing transmission.
3. *Geared or Gearless*. The use of fixed-speed gearing with a high gear ratio allows traction motors to be designed for high speed operation, resulting in high power density. The maximum speed is limited by the friction and windage losses as well as the tolerance of drive axle. On the other hand, traction motors can directly drive the transmission axles or adopt the in-wheel drive without using any gearing (gearless operation). However, it results in the use of low speed outer-rotor motors, which generally suffer from relatively low power density. The breakeven point is whether this increase in motor size and weight can be outweighed by the reduction of gearing. Otherwise, the additional size and weight will cause suspension problems in EVs. Both of them have been employed by modern EVs. For examples, the high speed geared inner-rotor in-wheel motor has been adopted in the NIES Luciole while the low speed gearless outer-rotor in-wheel motor was adopted in the TEPCO IZA. Nevertheless, with the advent of compact planetary gearing, the use of high speed planetary-geared in-wheel motors is becoming more attractive than the use of low speed gearless in-wheel motors.
4. *System Voltage*. The design of traction motors is greatly influenced by the voltage level of the EV system. Reasonable high voltage motor design can be adopted to reduce the cost and size of inverters. As different types of EVs adopt different system voltage levels, the design of traction motors needs to cater for different EVs. Roughly, the system voltage is governed by the battery weight that is about 30% of the total vehicle weight. In practice, higher power motors adopt higher voltage levels. For examples, the GM EV1 adopts the 312 V voltage level for its 102 kW motor, the Reva EV adopts the 48 V voltage level for its 13 kW motor, whereas Nissan Leaf adopts 360 V voltage level for its 80 kW motor.
5. *Integration*. The integration of the motor with the converter, controller, transmission, and energy source is prime, important consideration. The traction motor designer should fully understand the characters of these components, thus to design the motor under these given environments. It is quite different from the normal standard motors under standard power source for normal industrial drives.

Table 3. Comparison of single- and dual-motor configurations.

	Single-Motor	Dual-Motor
Cost	Lower	Higher
Size	Lumped	Distributed
Weight	Lumped	Distributed
Efficiency	Lower	Higher
Differential	Mechanical	Electronic
Reliability	Higher	Lower
Failure modes	Better	Worse

Table 4. Comparison of fixed- and variable-gearing transmissions.

	Fixed-Gearing	Variable-Gearing
Motor rating	Higher	Lower
Inverter rating	Higher	Lower
Cost	Lower	Higher
Size	Smaller	Larger
Weight	Lower	Higher
Efficiency	Higher	Lower
Reliability	Higher	Lower

3.3 Efficiency

The efficiency may be classified into two types, namely energy efficiency and power efficiency. The energy efficiency η_e is the ratio of energy output E_{out} to energy input E_{in}, whereas the power efficiency η_p is the ratio of power output P_{out} to power input P_{in}. So, they can simply be

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

expressed as:

$$\eta_e = \frac{E_{out}}{E_{in}} \tag{10}$$

$$\eta_p = \frac{P_{out}}{P_{in}} \tag{11}$$

For industrial operation, these two efficiencies may not be necessarily distinguishable. On the contrary, for vehicular operation, there is a significant difference because the power efficiency varies continually during the operation of most vehicles. Thus, it is necessary to delineate the power efficiency associated with the speed and torque conditions. Instead of using a particular operating point (such as rated power at rated torque and rated speed) to describe the power efficiency of a vehicle subsystem or component, an efficiency map is generally adopted. Figure 11 shows typical efficiency maps of a three-phase IM and a PM BLDC motor for propelling an EV. Hence, the energy efficiency can be derived by summing powers over a given time period.

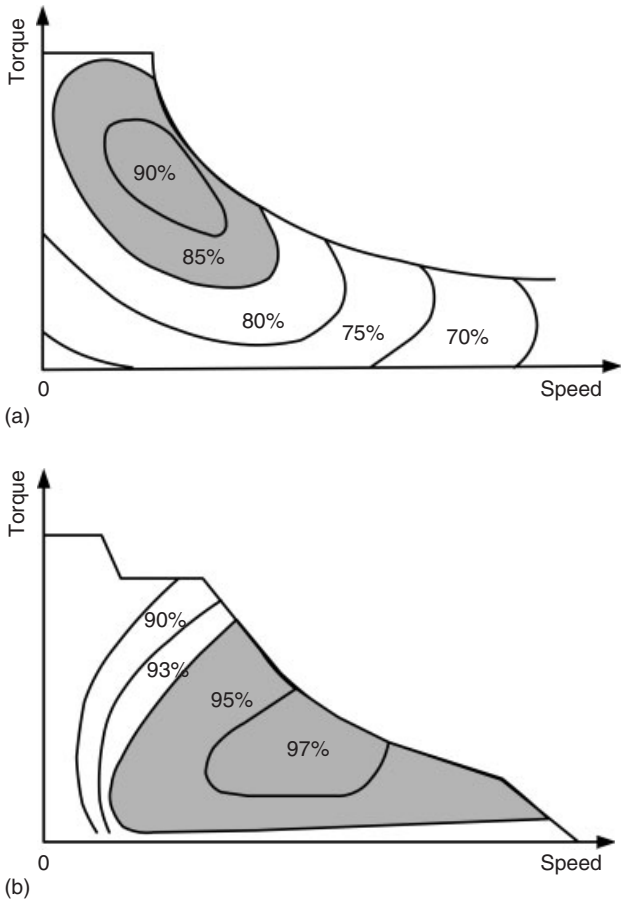


Figure 11. Typical power efficiency maps of EV traction motors. (a) Induction motor; (b) PM BLDC motor.

4 CONTROL CONSIDERATION OF TRACTION MOTOR DRIVE

4.1 Power electronics

4.1.1 Power devices

In the past decades, power semiconductor device technology has made tremendous progress. These power devices have grown in power rating and performance by an evolutionary process. Among existing power devices, the power diode behaves as an uncontrolled switch, whereas the others are externally controllable. Some of the more commonly used controllable devices are as follows:

- silicon-controlled rectifier (SCR), also known as *the thyristor*;
- gate turn off thyristor (GTO);
- bipolar junction transistor (BJT);
- metal-oxide semiconducting field effect transistor (MOSFET);
- insulated gate bipolar transistor (IGBT);
- static-induction transistor (SIT);
- static-induction thyristor (SITH);
- metal-oxide semiconducting-controlled thyristor (MCT).

In selection of power devices for traction motor drive, the following factors should be considered (Chan and Chau, 2001):

- *Ratings*. The voltage rating is based on the battery nominal voltage, maximum voltage during charging, and maximum voltage during regenerative braking. On the other hand, the current rating depends on the motor peak power rating and number of power devices connected in parallel. When paralleling these devices, on-state and switching characteristics have to be matched.
- *Switching Frequency*. Switching at higher frequencies can bring down the filter size and help meet the electromagnetic interference (EMI) limitation requirements. Over the switching frequency of 20 kHz, there is no acoustic noise problem.
- *Power Losses*. The on-state conduction drop or loss should be the minimum while the switching loss should be as low as possible. As higher switching frequencies increase the switching loss, switching the device at about 10 kHz seems to be an optimum for efficiency, power density, acoustic noise, and EMI considerations.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- *Base/Gate Drivability*. The device should allow for simple and secure base/gate driving. The corresponding driving signal may be either triggering voltage/current or linear voltage/current. The voltage-mode driving involves very little energy and is generally preferable.
- *Dynamic Characteristics*. The dynamic characteristics of the device should be good enough to allow for high dv/dt capability, high d*i*/d*t* capability, and easy paralleling. The internal antiparallel diode should have similar dynamic characteristics as the main device.
- *Ruggedness*. The device should be rugged to withstand a specific amount of avalanche energy during overvoltage and be protected by fast semiconductor fuses during overcurrent. It should operate with no or minimal use of snubber circuits. As EVs are frequently accelerated and decelerated, the device is subjected to thermal cycling at frequent intervals. It should reliably work under these conditions of thermal stress.
- *Maturity and Cost*. As the cost of power devices is one of the major parts in the total cost of traction motor drive, these devices should be economical.

Taking into account the above factors, the GTO, power BJT, power MOSFET, IGBT, and MCT are preferable for traction motor drive. The thyristor is not considered because it requires additional commutating components to turn off and its switching frequency is limited to 400 Hz. The SIT and SITH are also excluded because of their normally turn-on property and limited availability. In order to evaluate their suitability, a point grading system is adopted, which consists of eight major characteristics and each of them is graded from 1 to 5 points. From Table 5, the power MOSFET, IGBT, and MCT score high points indicating that they are particularly suitable for traction motor drive. Due to its highest score, the IGBT is almost exclusively used for modern traction motor drives. Nevertheless, the power MOSFET has also been accepted for those relatively low power electric tricycles and bikes.

Table 5. Evaluation of power devices for traction motor drive.

	GTO	BJT	MOSFET	IGBT	MCT
Ratings	5	4	2	5	3
Switching frequency	1	2	5	4	4
Power losses	2	3	4	4	4
Base/gate drivability	2	3	5	5	5
Dynamic characteristics	2	3	5	5	5
Ruggedness	3	3	5	5	5
Maturity	5	5	4	4	2
Cost	4	4	4	4	2
Total	24	27	34	36	30

4.1.2 Power converters

Power converters are usually classified by their input and output. As the input and output of a power converter can be either AC or DC, there can be four types of power converters:

- DC–DC converter
- DC–AC inverter
- AC–DC rectifier
- AC–AC cycloconverter.

The first three types of power converters are used in traction motor drives. The fourth type, AC–AC cycloconverters, is only used in high power AC–AC systems to control the voltage magnitude and frequency of large motors. However, AC–AC conversion involving an AC–DC circuit and a DC–AC circuit is not unusual. Depending on the power train configuration, a traction motor drive may involve one or more types of power converters.

A power converter typically consists of four parts: switching and peripheral circuits, filtering circuits, control circuits and feedback, and an optional user interface, as shown in Figure 12. The main circuit consists of power semiconductor devices (switches and diodes) and peripheral circuits. The semiconductor switches are controlled to turn on and turn off at a frequency ranging from a few kilohertz to a few tens of kilohertz for traction motor drives. Power converters usually involve LC low pass filters that will filter out the high frequency components of the output voltage and let the low frequency components or DC component pass to the load side. The control and feedback circuits typically involve the use of microcontrollers and sensors. Traction motor drive applications usually involve feedback torque control. Current feedback is usually necessary.

The DC–DC converters are also known as *DC choppers*, which are used for DC motor drives. Initially, DC choppers were introduced in the early 1960s using force-commutated thyristors that were constrained to operate at low switching frequency. Due to the advent of fast-switching power devices, they can now be operated at tens or hundreds of kilohertz. In electric propulsion applications, two-quadrant DC choppers are desirable because they convert battery DC voltage to variable DC voltage during the motoring mode and revert the power flow during regenerative braking. Furthermore, four-quadrant DC choppers are employed for reversible and regenerative speed control of DC motors. A four-quadrant DC chopper is shown in Figure 13.

The DC–AC inverters are generally classified into voltage-fed and current-fed types. Because of the need

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

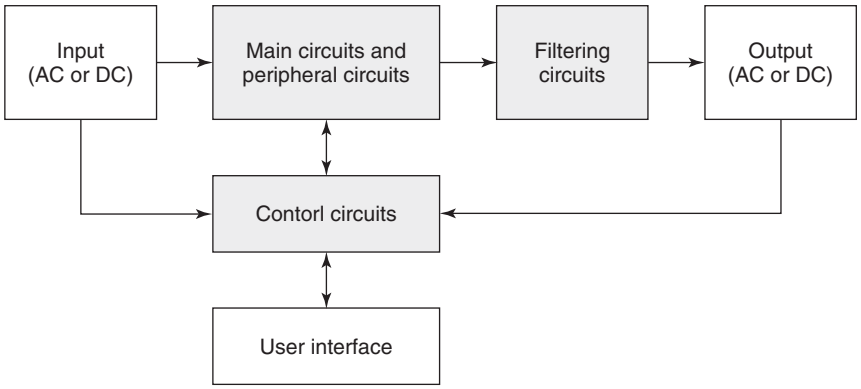


Figure 12. Schematics of power converter.

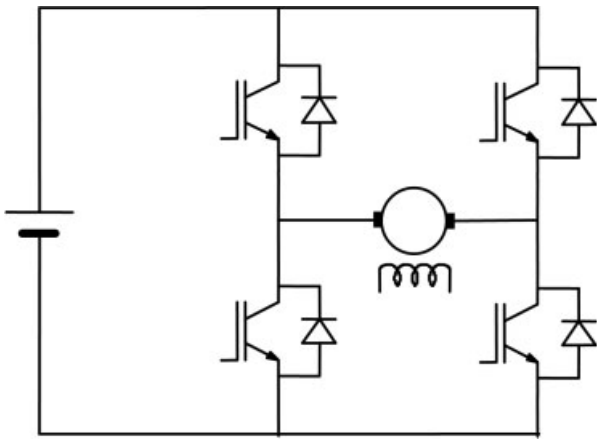


Figure 13. Four-quadrant DC chopper.

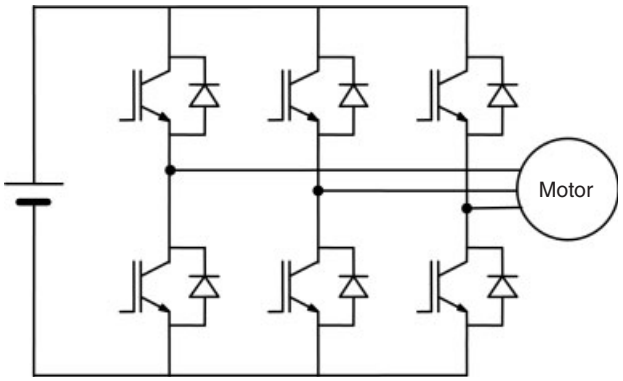


Figure 14. Three-phase full-bridge voltage-fed inverter.

of a large series inductance to emulate a current source, current-fed inverters are seldom used for traction motor drives. In fact, voltage-fed inverters are almost exclusively used because they are very simple and can have power flow in either direction. A typical three-phase full-bridge voltage-fed inverter is shown in Figure 14. Its output waveform may be rectangular, six-step, or PWM, depending on the switching strategy for different applications. For example, a rectangular output waveform is produced for a PM BLDC motor, whereas a six-step or PWM output waveform is produced for an IM. It should be noted that the six-step output is becoming obsolete because its amplitude cannot be directly controlled and its harmonics are rich. On the other hand, the PWM waveform is harmonically optimal and its fundamental magnitude and frequency can be smoothly varied for speed control.

In the last decades, numerous PWM switching schemes have been developed for voltage-fed inverters, focusing on the harmonic suppression, better utilization of DC voltage, tolerance of DC voltage fluctuation as well as suitability for real-time and microcontroller-based implementation. These schemes can be classified as voltage-controlled and current-controlled PWM. The state-of-the-art voltage-controlled PWM schemes are natural or sinusoidal PWM, regular or uniform PWM, harmonic elimination or optimal PWM, delta PWM, carrierless or random PWM, and equal-area PWM. On the other hand, the use of current control for voltage-fed inverters is particularly attractive for high performance motor drives because the motor torque and flux are directly related to the controlled current. The state-of-the-art current-controlled PWM schemes are hysteresis-band or band-band PWM, instantaneous current control with voltage PWM, and space vector PWM.

4.1.3 Emerging power electronic devices

The present silicon (Si) technology is reaching the material's theoretical limits and cannot meet all the requirements of vehicle applications in terms of compactness, light weight, high power density, high efficiency, and high

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

reliability under harsh conditions. The silicon carbide (SiC), new semiconductor material, with the potential increased power density and high temperature capability makes it an ideal candidate in traction motor drive applications (Kelley, Mazzola, and Bondarenko, 2006).

SiC power devices have much lower switching and conduction losses and can operate at much higher temperature than comparable Si power devices. Hence, a SiC-based power converter will have a much higher efficiency than that of a Si-based one at the same switching frequency. Alternatively, a higher switching frequency can be used to reduce the size of the magnetic components in a SiC-based power converter. In addition, because SiC power devices can be operated at much higher temperatures without much change in their electrical properties, ease of thermal management and high reliability can be achieved.

4.2 Control strategies

Conventional linear control, such as PID, can no longer satisfy the stringent requirements of high performance motor drives. In recent years, many modern control strategies have been developed. The state-of-the-art control strategies that have been proposed for motor drives are DTC, EOC, artificial intelligent control (AIC), position-sensorless control (PSLC), and so on (Chau, Chan, and Liu, 2008).

4.2.1 Direct torque control

The DTC is becoming attractive for traction motor drives, particularly for those equipped with dual-motor propulsion which desires fast torque response. It does not rely on current control and depends less on parameters. For the PM BLAC drives, the DTC controls both the torque and the flux linkage independently. The controller outputs provide proper voltage vectors via the inverter in such a way that these two variables are forced to predefined trajectories. The control block diagram of the DTC is shown in Figure 15a.

4.2.2 Efficiency-optimizing control

The EOC of motor drives is highly desirable for traction motor drives as their on-board energy storage is very limited. Different types of motor drives may employ different ways for efficiency optimization. For the rotor-PM BLAC drives, the EOC can be achieved by online tuning the input voltage or the *d*-axis armature current I_{2d} to minimize

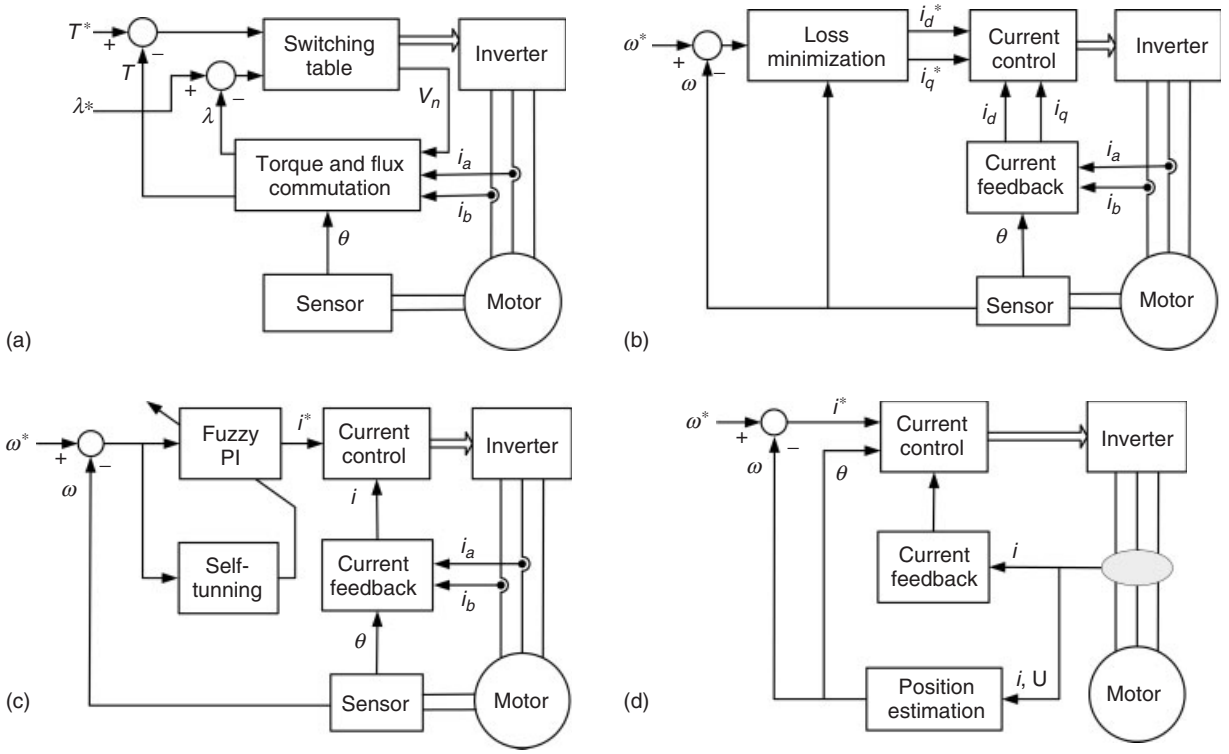


Figure 15. Control block diagrams. (a) DTC; (b) EOC; (c) AIC (fuzzy PI); (d) PSLC.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the total losses P_{loss} (Cavallaro *et al*., 2005)

$$P_{\text{loss}}(I_{2\text{d}}, T, \omega) = P_{\text{cu}}(I_{2\text{d}}, T, \omega) + P_{\text{Fe}}(I_{2\text{d}}, T, \omega) \qquad (12)$$

where P_{Cu} is the copper loss and P_{Fe} is the iron loss for the given torque T and speed ω. It can be found that there is a unique optimal operating point. In particular, the minimum total losses occur at a lower *d*-axis armature current than that of the minimum copper loss, hence illustrating that the maximum torque per ampere control cannot maximize the efficiency of the PM BLAC drives. Figure 15b shows the control block diagram of the EOC. For the hybrid PM BLAC drive incorporating with an additional DC field winding, the EOC can be easily achieved by tuning the polarity and magnitude of the DC field current (Shu, Cheng, and Kong, 2008).

4.2.3 *Artificial intelligent control*

All artificial-intelligence-based control strategies, such as fuzzy logic control, neural network control, neuro-fuzzy control, and genetic control, are classified as AIC. Among them, the fuzzy logic control and the neural network control are most mature and attractive as they can effectively handle the system's nonlinearities and sensitivities to parameter variations. Figure 15c shows the block diagram of the fuzzy PI (proportional-integral) control.

4.2.4 *Position-sensorless control*

In order to achieve high performance for traction motor drives, position feedback is almost mandatory. The position sensor is usually either a three-element Hall-effect sensor or an optical encoder, which are high cost, fragile elements. In order to get rid of the costly and bulky position sensor, PSLC is becoming attractive. Moreover, position-sensorless technology can effectively continue the operation of the system in case the position sensors lose their function. This is crucial in some applications, such as military vehicles.

There are various PSLC techniques. The majority of them are based on the voltage, current, and back EMF detection. These techniques can be primarily grouped into four categories (Ehsani, Gao, and Emadi, 2010):

1. Those using measured currents, voltages, fundamental machine equations, and algebraic manipulations.
2. Those using observers.
3. Those using back EMF methods.
4. Those with novel techniques not falling into the previous three categories.

It should be noted that the PSLC can be readily incorporated into other control strategies such as the EOC, the DTC, and the AIC.

4.2.5 *Comparison of control strategies*

As shown in Table 6, the aforementioned control strategies are compared in terms of their major advantages, major disadvantages, and typical techniques (Chau, Chan, and Liu, 2008). As there are many possible strategies for the AIC, the self-tuning fuzzy PI control (Cheng, Sun, and Zhou, 2006) is used for exemplification. Finally, some sample results of these control strategies have illustrated that the EOC can achieve the minimum total losses (Cavallaro *et al*., 2005), the DTC can provide direct bang–bang control of torque (Pascas and Weber, 2005), the AIC can achieve fast and accurate response, and the PSLC can offer accurate estimation of rotor position.

5 CONCLUSION

In this chapter, the general requirement for traction motor drives in EVs has been presented. The potential candidates

Table 6. Comparison of control strategies.

	Advantage	Disadvantage	Techniques
DTC	Fast torque response; no need for current control; less parameter dependence	Cause errors due to drift flux linkage estimation, and variation of stator resistance	Generate the voltage vectors using independent torque and flux computations
EOC	Minimize the overall losses; no need for accurate loss model; work for wide speed and torque range	Originate system oscillation or convergence problem	Control the input voltage or d-axis armature current; control DC field current
AIC	Flexible control algorithms; adapt nonlinearities and parameter variations	Require expert knowledge or intensive computation and sophisticated hardware	Incorporate fuzzy logic, neural network, and other AI into traditional controls
PSLC	Eliminate position sensor, hence reduce system size and cost; readily merge into other controls	Require intensive computation and sophisticated hardware	Estimate the position based on motional EMF, inductance variation, or flux linkage variation

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

for a traction motor for EVs have been evaluated according to the major requirements of an EV electric propulsion system. Design consideration of traction motors is described. Power devices and power converters for traction motor drives are discussed and evaluated. Control strategies for traction motor drives are presented and compared.

Currently, the PM motors and cage IMs present better comprehensive performance than others, hence are highly dominant in recently released EVs (de Santiago *et al*., 2012), whereas the DC motors are losing attraction though still in use in some small vehicles, and switched reluctance motors and stator-PM motors are gaining much interest.

Thanks to persistent hard work of both academic and industrial communities in the past years, the performance of traction motors for EVs has been improved greatly. With quick development of industry technology, motor drives in EVs would meet with new renovations. The development trends of the traction motor drive in EVs may include the following:

1. *High Speed Motors*. By increasing the speed, the size of electric motors may be reduced greatly, viz. higher power from smaller machines and redesigning for increased material utilization. Some companies have started to focus on high speed of 16,000 r/min PM motors that can achieve field weakening within the structure of the motor and eliminate the need for a DC–DC boost converter.
2. *Redundant and Fault-Tolerant Motor Structure*. Continued operation of motor drive is an essential requirement in EV application. Therefore, the need for high degree of reliability in motor drive system has inspired much research in the area. To achieve high reliability, redundant or conservative design techniques have been employed in many motor drives.
3. *Novel Manufacture Techniques*. To achieve high power density, high efficiency, and low cost motors for EVs, the manufacture technique of motors is being improved. The segmented stator and concentrated winding are examples. In addition, using flat wire to replace round wire in motor windings can increase slot filling factor, enabling both a higher torque constant and a lower copper loss.
4. *Novel Machine Topologies with Composite Structures and New Materials*. For traditional machines, each has its own merits and demerits. The composition of different machines may significantly improve the performance. Hence, the traction machines that consist of different structures may be noticed in the next step.

RELATED ARTICLES

Permanent Magnet Brushless Motor Drives
Switched Reluctance Motor Drives
Future Direction of Traction Motor Drives
DC Motor Drives
Induction Motor Drives

REFERENCES

Cavallaro, C., Tommaso, A.O.D., Miceli, R., *et al*. (2005) Efficiency enhancement of permanent-magnet synchronous motor drives by online loss minimization approaches *IEEE Transactions on Industrial Electronics*, **52** (4), 1153–1160.

Chan, C.C. and Chau, K.T. (2001) *Modern Electric Vehicle Technology*, Oxford University Press, Oxford.

Chan, C.C., Chau, K.T., Jiang, J.Z., *et al*. (1996) Novel permanent magnet motor drives for electric vehicles *IEEE Transactions on Industrial Electronics*, **43** (2), 331–339.

Chau, K.T., Chan, C.C., and Liu, C. (2008) Overview of permanent-magnet brushless drives for electric and hybrid electric vehicles *IEEE Transactions on Industrial Electronics*, **55** (6), 2246–2257.

Cheng, M., Chau, K.T., and Chan, C.C. (2001) Static characteristics of a new doubly salient permanent magnet motor *IEEE Transactions on Energy Conversion*, **16** (1), 20–25.

Cheng, M., Sun, Q., and Zhou, E. (2006) New self-tuning fuzzy PI control of a novel doubly salient permanent-magnet motor drive *IEEE Transactions on Industrial Electronics*, **53** (3), 814–821.

Cheng, M., Hua, W., Zhang, J., and Zhao, W. (2011) Overview of stator-permanent magnet brushless machines *IEEE Transactions on Industrial Electronics*, **58** (11), 5087–5101.

Deodhar, R.P., Andersson, S., Boldea, I., and Miller, T.J.E. (1996) The Flux-Reversal Machine: A New Blushless Doubly-Salient Permanent-Magnet Machine. *IEEE IAS Annual Meeting Record*, San Diego, pp. 786–793.

Ehsani, M., Gao, Y., and Emadi, A. (2010) *Modern Electric, Hybrid Electric and Fuel Cell Vehicles*, 2nd edn, CRC Press, Boca Raton.

Hoang, E., Ben-Ahmed, A.H., and Lucidarme, J. (1997) Switching Flux Permanent Magnet Polyphased Machines. *Proceedings of European Conference on Power Electronics and Applications*, Trondheim, pp. 903–908.

Hua, W., Zhu, Z.Q., Cheng, M., *et al*. (2005) Comparison of Flux-Switching and Doubly-Salient Permanent Magnet Brushless Machines. *Proceedings of International Conference on Electrical Machines and Systems*, Nanjing, pp. 165–170.

Hua, W., Cheng, M., Zhu, Z.Q., and Howe, D. (2007) Analysis and optimization of back EMF waveform of a flux-switching permanent magnet motor *IEEE Transactions on Energy Conversion*, **23** (3), 727–733.

Kelley, R., Mazzola, M.S., and Bondarenko, V. (2006) A scalable SiC device for DC/DC converters in future hybrid electric vehicles. *Proceedings of IEEE Applied Power Electronics Conference and Exposition*, Dallas, pp. 460–463.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Liao, Y., Liang, F., and Lipo, T.A. (1995) A novel permanent magnet motor with doubly salient structure *IEEE Transactions on Industry Applications*, **31** (5), 1069–1078.

Mi, C., Masrur, A., and Gao, D. (2011) *Modern Hybrid Electric Vehicles*, John Wiley & Sons, Ltd, Chichester.

Pascas, M. and Weber, J. (2005) Predictive direct torque control for the PM synchronous machine *IEEE Transactions on Industrial Electronics*, **52** (5), 1350–1356.

de Santiago, J., Bernhoff, H., Ekergård, B., *et al.* (2012) Electrical motor drivelines in commercial all-electric vehicles: a review *IEEE Transactions on Vehicular Technology*, **61** (2), 475–484.

Shu, Y., Cheng, M., and Kong, X. (2008) Online Efficiency Optimization of Stator-Doubly-Fed Doubly Salient Motor Based on a Loss Model. *Proceedings of 11th International Conference on Electrical Machines and Systems*, Wuhan, pp. 1174–1178.

Zhu, X. and Cheng, M. (2010) Design, analysis and control of hybrid excited doubly salient stator-permanent-magnet motor *Science China Technological Science*, **53** (1), 188–199.

Zhu, Z.Q. and Howe, D. (2007) Electrical machines and drives for electric, hybrid, and fuel cell vehicles *Proceedings of the IEEE*, **95** (4), 746–765.

FURTHER READING

Lipo, T.A. (2007) *Introduction to AC Machine Design*, University of Wisconsin, Madison.

Mohan, N., Undeland, T.M., and Robbins, W.P. (1995) *Power Electronics: Converters, Applications and Design*, 2nd edn, John Wiley & Sons, Inc, New York.

Novotny, D.W., Lipo, T.A., and Jahns, T.M. (2009) *Introduction to Electric Machines and Drives*, University of Wisconsin, Madison.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto041
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

EVT and E-CVT for Full Hybrid Electric Vehicles

Yuan Cheng[1] **and Ming Cheng**[2]

[1]*Harbin Institute of Technology, Harbin, China*
[2]*Southeast University, Nanjing, China*

1 Introduction 1
2 The E-CVT Propulsion Systems 1
3 The EVT Propulsion Systems 4
4 Conclusion 8
References 9

1 INTRODUCTION

With ever-increasing concerns on energy shortage and environmental protection, hybrid electric vehicles (HEVs) have become globally attractive. They take the distinct advantages of lower fuel consumption and emissions than internal combustion engine vehicles (ICEVs), whereas longer driving range and easier refuel than battery electric vehicles (BEVs) (Chau and Chan, 2007).

According to the level of electric power contribution in the total propulsion power, HEVs can be classified into micro, mild, and full hybrids. Among these hybrid vehicles, full hybrids have the greatest electric power fraction and possess all hybrid features, such as electric launch, idle stop, regenerative braking, and so on. These features enable ICE to operate at its efficient regions and achieve high fuel economy and low emissions.

The Toyota Prius HEV is a typical full hybrid vehicle, in which an ICE and a pair of electric machines (EMs) are adopted. As the core of the Toyota hybrid system (THS), a planetary gear (PG) is used to connect three power plants. By adopting the PG, electric driving force and ICE driving force can be effectively split and combined, giving more freedom to optimize power flows and reduce fuel consumption. The THS system is also known as *electronic-continuously variable transmission* (*E-CVT*) as the ICE power can be transferred to the vehicle seamlessly and smoothly. But different from the mechanical CVT, the E-CVT is implemented in the way of electric drives instead of mechanics, hydraulics, or hydromechanics.

Many different E-CVT concepts have been put forth through various combinations of PGs, EMs, ICE, and so on (Miller, 2006; Wang, Cheng, and Chau, 2009). But recently, a novel transmission, called *electric variable transmission* (*EVT*), has been proposed, which also realizes the power split function without the aid of PGs (Hoeijmakers and Ferreira, 2006). By integrating a double-rotor EM cascaded with another EM, the EVT not only implements a transmission of the ICE power but also opens a gate enabling the input and output of electric power. These features make EVT a competitive HEV solution with a highly integrated function of CVT, starter, and generator.

In this chapter, the system architectures and operating principles of both EVT and E-CVT propulsion systems have been introduced.

2 THE E-CVT PROPULSION SYSTEMS

According to different system architectures, the E-CVT can be sorted into two categories, input power split and compound power split (Miller, 2006).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto042
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2.1 Input split E-CVT

In the input split E-CVT, there exists a power split device (namely PG) at the transmission input. For example, the HEVs developed by Toyota and Ford fall into this category.

2.1.1 Toyota E-CVT system

Figure 1a shows the PG used in Toyota E-CVT propulsion system. The Toyota E-CVT mainly consists of a PG, a battery pack, two power converters, and two EMs. The PG consists of a set of gears, namely, the ring gear, the sun gear, the carrier gear, and several pinion gears. As shown in Figure 1b, the ICE is linked to the carrier gear, the EM1 is connected to the sun gear, and the EM2 is connected to the ring gear and then to the vehicle. The PG splits the ICE power into electric power flow and mechanic power flow. On one hand, a portion of the ICE power can be transferred from the carrier gear directly to the ring gear and then to the vehicle. On the other hand, the electric power flow converts the rest of the ICE power into the electrical form through the EM1. The electric power can either charge the battery or supply the EM2. The power converters share a common DC bus, to which the battery pack is connected. The battery can release/absorb the electric power to/from the DC bus in different operating modes. The Toyota E-CVT system is also well known as a *series and parallel hybrid electric vehicles* (*SP-HEV*). It involves the features of series and parallel hybrids, and many operating modes are possible in order to enhance the system performance and reduce the fuel consumption.

Due to the mechanical connection through the PG, the speed and torque relationships between different components in steady state can be expressed as:

$$T_{\mathrm{d}} = T_{\mathrm{EM2}} + \frac{\rho}{1+\rho} T_{\mathrm{ICE}} \tag{1}$$

$$T_{\mathrm{EM1}} = \frac{1}{1+\rho} T_{\mathrm{ICE}} \tag{2}$$

$$\Omega_{\mathrm{ICE}} = \frac{\rho}{1+\rho} \Omega_{\mathrm{EM2}} + \frac{1}{1+\rho} \Omega_{\mathrm{EM1}} \tag{3}$$

where T_{d} is the driveline torque and Ω_{EM1}, T_{EM1}, Ω_{EM2}, T_{EM2}, Ω_{ICE}, T_{ICE} are the speed and torque of EM1, EM2, and ICE respectively. ρ is the basic ratio of the PG.

From Equation 3, it can be seen that there are two independent variables among three speeds. As the EM2 is linked to the vehicle, its speed is proportional to the vehicle speed. Therefore, Ω_{EM1} can be chosen as an independent variable, which can be controlled to optimize the ICE speed. In the same manner, the EM2 torque can be controlled independently in order to optimize the ICE torque. Therefore, by properly controlling the speed and the torque of both EM1 and EM2, the ICE can be operated independent of the road load and at its optimal region.

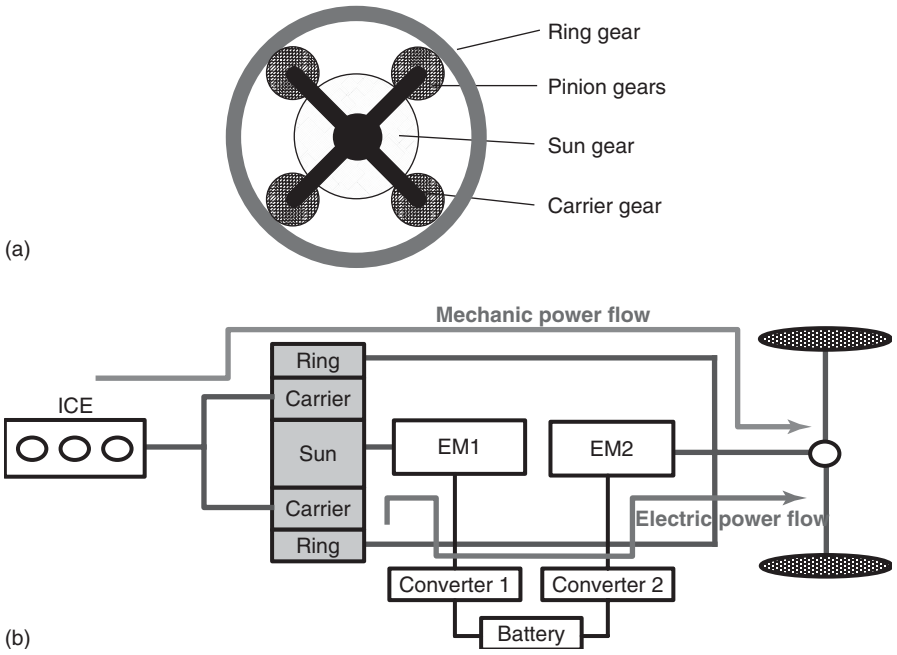


Figure 1. The (a) PG and (b) E-CVT propulsion system in Toyota Prius HEV.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto042
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.1.2 *Ford E-CVT system*

The architecture of Ford E-CVT propulsion system is shown in Figure 2. Besides the PG at the transmission input, an output torque multiplier gear set is added. The steady state relationship between the ICE, the EM2, and the driveline torque can be expressed as:

$$T_{\mathrm{d}} = \frac{N_2}{N_1} T_{\mathrm{EM2}} + \frac{\rho}{1+\rho} \frac{N_2}{N_3} T_{\mathrm{ICE}} \tag{4}$$

where N_1, N_2, and N_3 are the teeth numbers of output gear set. From Equation 4, it can be seen that the Ford E-CVT is very similar to the Toyota HEV. However, a boost output torque can be achieved due to the adoption of the output gear, which provides further advantages to reducing the ICE and EM2 torques.

2.2 Compound split E-CVT

Compound split E-CVT has also been presented in recent years. Such a compound split can be obtained by replacing the output gear in the Ford E-CVT system with a second PG. Because of variable structure control, such a compound split system has more flexibility of power distribution in addition to the capability to output higher torque. An example can be seen in the GM E-CVT system, which is originally developed by Allison for GM as a hybrid heavy-duty transmission.

The GM E-CVT propulsion system is shown in Figure 3, which is mainly composed of three clutches, two PGs, two EMs, two power converters, and a battery pack (Miller, 2006). The compound split E-CVT is also known as *two-mode system* defined in both low speed and high speed ranges. By means of engaging or disengaging different clutches, the E-CVT propulsion system can transform its architecture so that the output torque can meet the road load demand.

When vehicles run in low speed or in city driving mode, both Clutch 1 and Clutch 3 are engaged whereas the Clutch 2 is disengaged. The system becomes input split type in operation. The PG1 attached to the ICE works as an input power split type, whereas the PG2 coupled with the

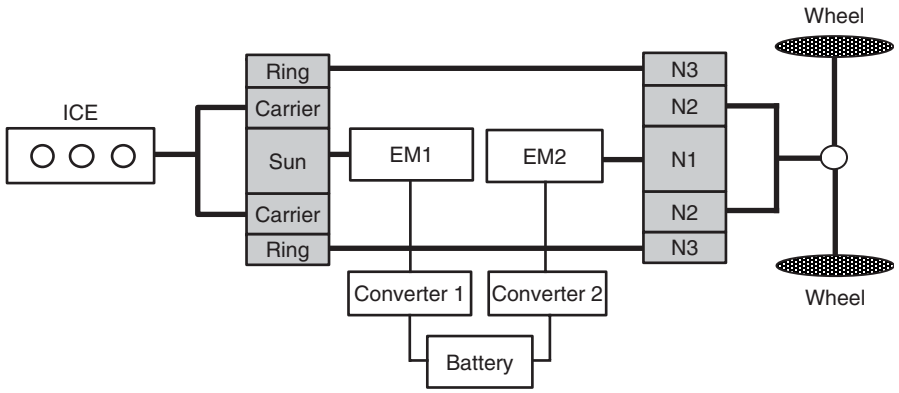


Figure 2. Ford E-CVT propulsion system.

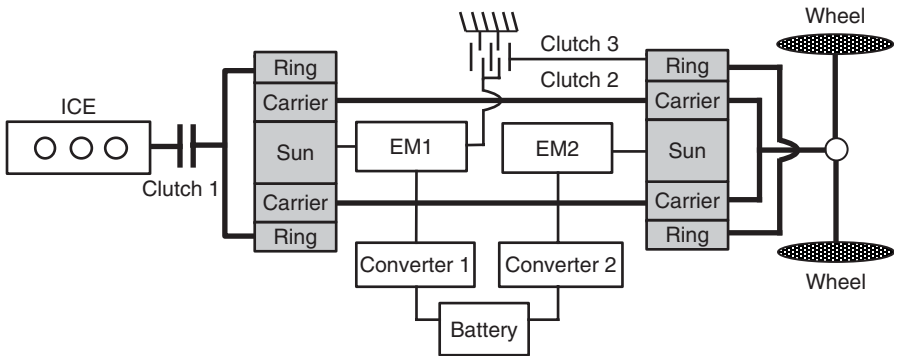


Figure 3. GM-Allison E-CVT propulsion system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto042
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

driveline is in charge of output torque coupling. The steady state relationship between the ICE torque and the driveline torque can be expressed as:

$$T_{\mathrm{d}} = \left(1 + \frac{1}{\rho_1}\right) T_{\mathrm{ICE}} + (1 + \rho_1) T_{\mathrm{EM2}} \tag{5}$$

where ρ_1 is the input PG ratio.

In highway driving mode, both Clutch 1 and Clutch 2 are engaged whereas Clutch 3 is disengaged. This kind of operating mode is called *compound split*, in which the input and output PGs perform the function of power split together. The steady state relationship between the ICE torque and the driveline torque can be expressed as:

$$T_{\mathrm{d}} = \frac{1}{1 + \rho_2} T_{\mathrm{EM2}} - \frac{1}{\rho_1 \rho_2 (1 + \rho_2)} T_{\mathrm{ICE}} \tag{6}$$

where ρ_2 is the output PG ratio.

Through combinations of PGs and clutches, various compound split systems can also be found, such as eVT from Timken, IVT from Renault, and the hybrid power train from Geely (Miller, 2006; Tenberge *et al*., 2010).

It should be noticed that the power split devices adopted in most of the E-CVT propulsion systems are simple PGs, as shown in Figure 1. Currently, an example of heavy-duty power train using Ravigneaux Geartrain (RG) can also be found (Syed *et al*., 2010). Shown in Figure 4, the RG is a compound geartrain consisting of two simple PGs, PG1 and PG2. The PG1 consists of a ring gear, a large sun gear, and outer pinion gears. The PG2 shares the common ring gear with PG1 and the inner pinion gears are meshed with a smaller sun gear. The inner and outer pinion gears are meshed with each other and connected independently with a planet carrier. In comparison to the E-CVT systems utilizing two individual simple PGs, the RG is more compact and has more degrees of freedom to optimize the operation of different power sources.

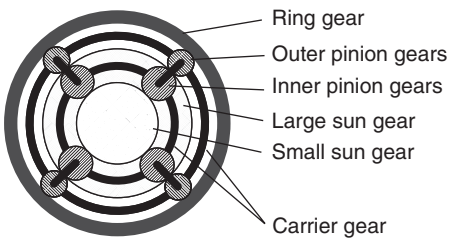


Figure 4. The ravigneaux gear.

3 THE EVT PROPULSION SYSTEMS

Although possessing both advantages of series and parallel HEVs, the structures of E-CVT are relatively complicated and costly. Recently, a family of EVT propulsion systems is being actively developed. Illustrated in Figure 5, the key to the EVT system is to make use of a double-rotor machine (DRM) to split the ICE power into electric power flow and mechanical power flow. Compared with the E-CVT system, it is also an input power split power train. However, due to the elimination of the PGs, the mechanical wear and audible noise associated with the EVT propulsion systems can be eliminated.

3.1 Configuration and operating principle

As shown in Figure 5, the EVT propulsion system is composed of a DRM, two power converters, and a battery pack. The DRM has three parts: the stator (S), the inner rotor (IR) with individual windings, and the outer rotor (OR) with permanent magnets (PMs) or squirrel cages. The DRM can be seen as an integrated electromechanical converter consisting of two concentrically arranged EMs. The inner machine EM1 is a DRM, which consists of the IR and the inner part of OR. The outer machine EM2 is a normal EM, which consists of the stator and the outer part of OR. The IR is attached to the ICE shaft, and the OR is connected to the driveline and then to the vehicle.

Owing to the two independent rotatory parts, the ICE power P_{ICE} can be divided into two parts at the EM1. One part is electrical power P_{elec}, which reaches the EM2 through the slip rings and the power converters. The other is mechanical power P_{mech}, which is directly coupled to the EM2 and used to propel the vehicle. In fact, the "mechanical" power is transmitted not through mechanical connection but through the action–reaction electromagnetic torque between the two rotors of the EM1. It should be noted that, if the EM1 is not in an active state, the "mechanical" power does not exist any longer.

In steady state, due to the action and reaction, the ICE torque is equal to the electromagnetic torque of the EM1, $T_{\mathrm{ICE}} = T_{\mathrm{EM1}}$. Therefore, without considering losses, the relationship between the ICE power P_{ICE} and the EM powers, P_{EM1} and P_{EM2}, can be expressed as:

$$P_{\mathrm{ICE}} = T_{\mathrm{ICE}} \Omega_{\mathrm{ICE}} = P_{\mathrm{mech}} + P_{\mathrm{elec}} \tag{7}$$

$$P_{\mathrm{mech}} = T_{\mathrm{EM1}} \Omega_{\mathrm{EM2}} = T_{\mathrm{ICE}} \Omega_{\mathrm{EM2}} \tag{8}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto042
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

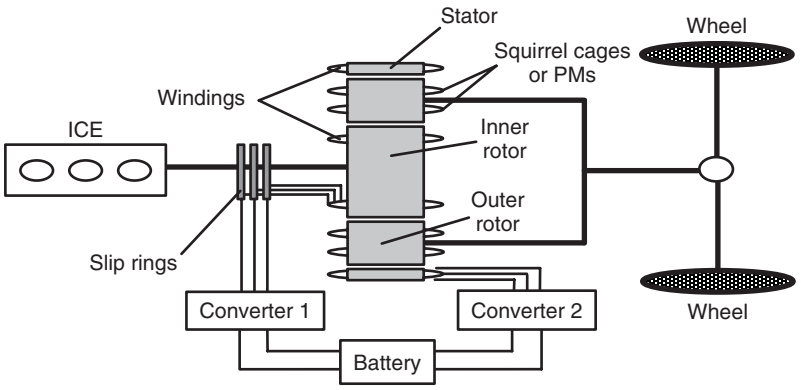


Figure 5. The EVT propulsion system.

$$P_{\text{elec}} = P_{\text{EM1}} = P_{\text{EM2}} = P_{\text{ICE}} - P_{\text{mech}} = T_{\text{ICE}}(\Omega_{\text{ICE}} - \Omega_{\text{EM2}}) \quad (9)$$

where Ω_{ICE} and Ω_{EM2} are the speeds of the ICE and the EM2, respectively. Due to the two rotator parts, the actual EM1 speed is the difference between the ICE and the EM2, $\Omega_{\text{EM1}} = \Omega_{\text{EM2}} - \Omega_{\text{ICE}}$. The generated electric power P_{elec} can produce the torque on the EM2 machine

$$T_{\text{EM2}} = \frac{P_{\text{elec}}}{\Omega_{\text{EM2}}} = \frac{T_{\text{ICE}}(\Omega_{\text{ICE}} - \Omega_{\text{EM2}})}{\Omega_{\text{EM2}}} \quad (10)$$

Therefore, the final driving torque is

$$T_{\text{d}} = T_{\text{EM1}} + T_{\text{EM2}} = T_{\text{ICE}} + \frac{T_{\text{ICE}}(\Omega_{\text{ICE}} - \Omega_{\text{EM2}})}{\Omega_{\text{EM2}}} = \frac{T_{\text{ICE}}\Omega_{\text{ICE}}}{\Omega_{\text{EM2}}} = iT_{\text{ICE}} \quad (11)$$

where i is the EVT speed ratio. It can be seen that the EVT can also work as a CVT. But different from the E-CVT, the EVT implements the CVT function by directly controlling the speed and the torque of the EM1 and EM2 machines without introducing the PG. More specifically, the EM1 adopts speed control to balance the difference between the required speed Ω_{d} at the driveline and the optimal speed Ω_{ICE} of the ICE. In the same manner, the EM2 adopts torque control to balance the difference between the required torque T_{d} at the driveline and the optimal torque T_{ICE} of the ICE. Figure 6 illustrates such a control strategy. The optimal operating line (OOL) shows the ICE optimal operating line, along which the ICE has the best efficiency in all operating points with the same powers. P_{d} shows the required operating point at the driveline shaft, and P_{ICE} is the actual operating point offered by the ICE with the same output power as P_{d}. Through controlling the EVT, the ICE

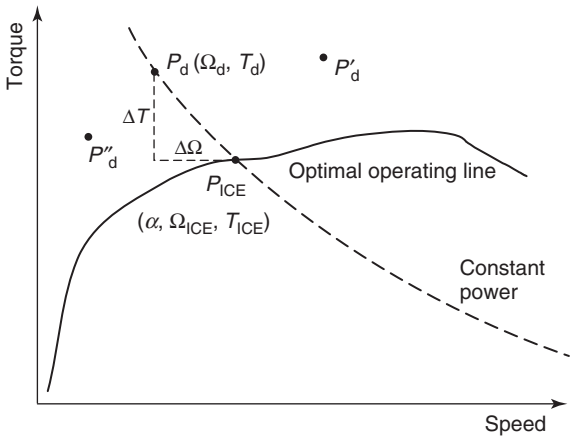


Figure 6. The EVT control strategy.

can be operated independent of the driving conditions and achieve the best efficiency (Cheng *et al*., 2007). This OOL operation principle is exactly the same as in CVTs where it is realized by continuously changing the transmission ratio between input and output shafts.

If the ICE power cannot fulfill the desired driving power, the battery can release power to boost the driving force. In this case, P_{d} moves to P'_{d}. Alternatively, the battery can absorb the regenerative braking energy when the vehicle decelerates, and P_{d} moves to P''_{d}. With this function, the EVT can work in hybrid mode or starter and generator mode. The power flows of EVT in different operating modes have been illustrated in Figure 7.

3.2 Different topologies of EVT

Although the name of the EVT was first presented by Prof. Hoeijmakers based on induction machines (IMs) (Hoeijmakers and Ferreira, 2006), various machine types can

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto042
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

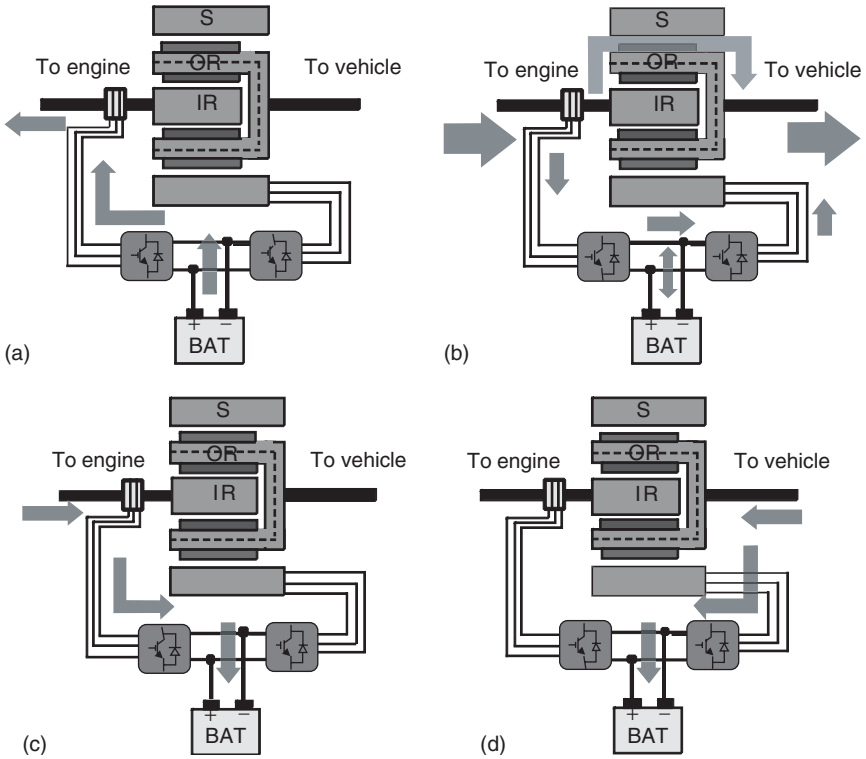


Figure 7. Power flows of EVT in different operating modes. (a) Starter mode, (b) hybrid mode, (c) generator mode, and (d) regenerative braking.

be conceptually introduced into the EVT concept. Similar designs can be found by using PM machines, switched reluctance machines (SRMs), and so on. Axial–axial field type or compound structures have also been studied in order to enhance the EVT performance as well as conventional topologies with radial–radial field distribution.

3.2.1 Radial–radial field topology

Currently, most of the EVTs adopt this topology, in which both the EM1 and the EM2 have radial air gap field distribution. For this topology, there are individual polyphase windings on the stator and the IR. According to the structures of the OR, the EVT can be based on IM machines, PM machines, or SRM machines, as shown in Figure 8.

Figure 8a shows an IM-based EVT, which has two separate layers of squirrel cages in the outer and inner part of the OR. Operating on the principle of electromagnetic induction, IMs are considered robust, low cost, and mature. Good dynamic performance can be achieved by adopting field-oriented control (FOC) or direct torque control (DTC). Replacing the squirrel cages with PMs leads to the PM-EVT with inherent advantages of high efficiency and high torque density, as shown in Figure 8b. PMs can be located on the surfaces or the interior of the OR according to performance requirements. In this case, the PM-EVT can be considered the combination of two separate permanent magnet synchronous machines (PMSMs) and be controlled independently according to the control strategy (Nordlund and Sadarangani, 2002). Instead of two layers, one-layer PM can also be adopted on the OR. In this case, PM-EVT has a uniform field in outer and inner air gaps and strong magnetic interference exists between its stator and IR. The EM1 and EM2 cannot be controlled independently. As an example of such a one-layer PM-EVT, dual mechanical ports machine (DMPM) is being developed (Xu, 2005). With structures of doubly salient poles in EM1 and EM2 machines, EVT can also have SRM versions, as shown in Figure 8c. The SRM-EVT runs through the magnetic reluctance torque rather than the electromagnetic torque. SRM-EVT offers the advantages of low cost, simple structure, high power density in low speed structure, and so on. However, due to the discontinuous torque and high ripple, SRM can cause vibration and also acoustic noise.

3.2.2 Axial–axial field topology

In order to save more space under the vehicle hood or enhance performances, the EVT could also transform into an axial–axial field version as shown Figure 9 (Zheng *et al*.,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto042
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

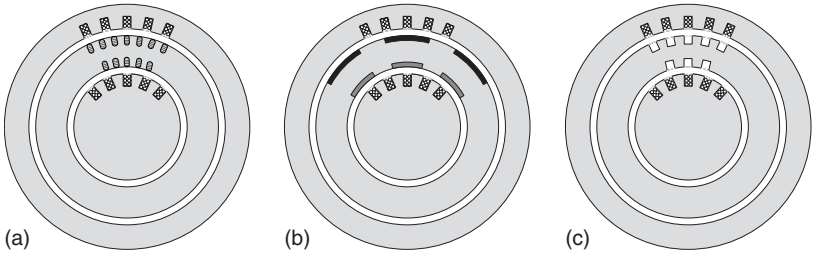


Figure 8. Different machine-based EVTs of radial–radial field topology. (a) IM-EVT, (b) PM-EVT, and (c) SRM-EVT.

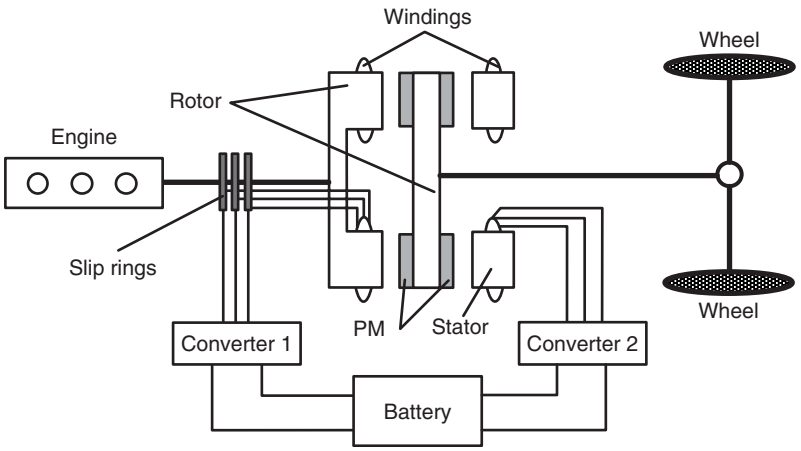


Figure 9. Axial–axial EVT propulsion system.

2008a). The left rotor, the middle rotor, and the right stator function exactly as the IR, the OR, and the stator in the radial–radial version. The EM1 and EM2 become axial flux disk machines. Obviously, different machine types can be adopted in the axial–axial EVT version.

3.2.3 Compound topology

No matter the EVT propulsion system is based on IMs, PM brushless machines, SRMs, or radial–radial and axial–axial topologies, the magnetic fields in the two air gaps are coupled to certain extent and interference with each other (Zheng *et al*., 2007). Such a magnetic coupling will degrade the independency of two EMs and consequently affect the controllability of the EVT system. By combining radial and axial field topologies, the field coupling can be completely ignored. An example of axial–radial field topology EVT is shown in Figure 10. It employs a DRM with axial flux in the first stage and another radial flux machine in the second stage (Zheng *et al*., 2008b). Due to its decoupling nature between the first stage and the second stage, the EVT propulsion system can provide better controllability.

Several topologies are also possible, such as radial–axial topology, axial/radial–radial topology, and so on.

3.2.4 Brushless EVT system

In order to feed electricity into the DRM, slip rings and carbon brushes have to be used in the EVT. However, the complication and maintenance associated with slip rings and brushes degrade the reliability and the efficiency of the EVT system. Therefore, one of the development trends of the EVT systems is to eliminate the slip rings and the brushes.

Very recently, a double-stator PM brushless EVT propulsion system has been proposed. As shown in Figure 11, a double-stator machine is used to split the ICE power. Two power flows are possible from the ICE to the wheels:

1. from the outer stator of the first-stage machine, Converter 1 and Converter 2 to the second-stage machine;
2. from the inner stator of the first-stage machine, Converter 3 and Converter 2 to the second-stage machine.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto042
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

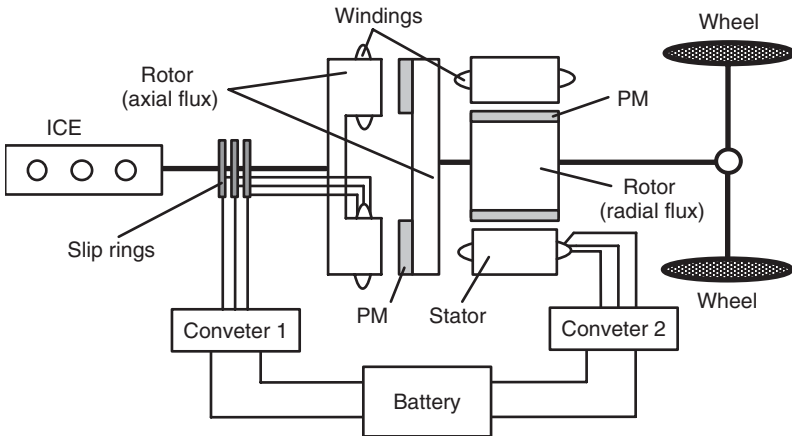


Figure 10. Axial–radial EVT propulsion system.

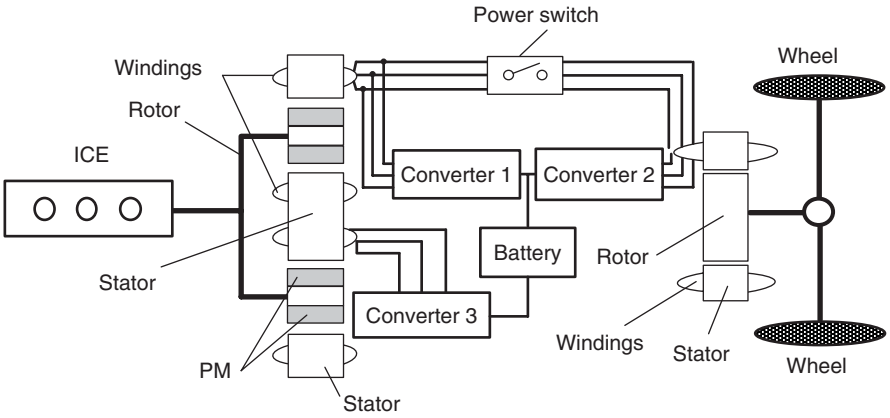


Figure 11. Double-stator PM brushless EVT propulsion system.

The first-stage machine is a double-stator PM machine, and the second-stage machine is an IM or a PM brushless machine. All the power converters are standard three-phase inverters. The outer stator transfers a major part of the ICE power and the inner stator performs power split control using the battery pack as a buffer (Wang *et al*., 2008, 2011).

By properly controlling the two power converters, the ICE can operate at constant speed while the wheel speed varies with the road load profile and driver's control, hence offering the merit of continuously variable gearing. When there is no need to perform power split, the system efficiency can be further improved by using a power switch to directly transfer the ICE power from the outer stator of the first machine to the second machine.

Compared with the double-rotor counterparts, this double-stator PM brushless EVT propulsion system possesses the advantages of improved reliability.

4 CONCLUSION

In this chapter, various EVT and E-CVT propulsion systems have been introduced. Both EVT and E-CVT can realize the power split function and optimize the vehicle power flows. Currently, with the aid of PG, the E-CVT is more mature. Successful cases such as Toyota Prius HEVs have appeared in markets for many years. In addition, many innovative E-CVT concepts are also in development. However, the E-CVT system is relatively complicated because of the power split devices.

As a promising HEV power train, the EVT eliminates mechanical transmission and the clutches, which eliminates the mechanical wear and noise and simplifies the system structure. From the overview of state of the art, PM-EVT has become more attractive than other machine types due to higher efficiency and torque density of PM machines. Moreover, the application of PMs in the EVT leads to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto042
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

various topologies as well as interesting issues of machine design and control. However, the existence of slip rings and brushes reduces the reliability and the efficiency of the EVT system.

REFERENCES

Chau, K.T. and Chan, C.C. (2007) Emerging energy-efficient technologies for hybrid electric vehicles. *Proceedings of the IEEE*, **95** (4), 821–835.

Cheng, Y., Cui, S., Song, L., and Chan, C.C. (2007) The study of the operation modes and control strategies of an advanced electromechanical converter for automobiles. *IEEE Transactions on Magnetics*, **43** (1), 430–433.

Hoeijmakers, M.J. and Ferreira, J.A. (2006) The electric variable transmission. *IEEE Transactions on Industry Applications*, **42** (4), 1092–1100.

Miller, J.M. (2006) Hybrid electric vehicle propulsion system architectures of the E-CVT type *IEEE Transactions on Power Electronics*, **21** (3), 756–767.

Nordlund, E. and Sadarangani, C. (2002) The Four Quadrant Energy Transducer. *Conference Record of 37th IEEE IAS Annual Meeting*, pp. 390–391.

Syed, S.A., Lhomme, W., Bouscayrol, A., *et al*. (2010) Modeling of Power Split Device for Heavy-Duty Vehicles. *Proceedings of IEEE VPPC*, pp. 1–6.

Tenberge, P., Gläser, K., and Zhang, T. (2010) Geelys New Full Hybrid Transmission with Two e-Motors. *Proceedings of CVT-Hybrid International Conference 2010*, pp. 200–207.

Wang, Y., Cheng, M., Fan, Y., and Chau, K.T. (2008) Design and Analysis of Double-Stator Permanent Magnet Brushless Motor for Hybrid Electric Vehicles. *Proceedings of the International Conference on Electrical Machines and System*, pp. 3241–3246.

Wang, Y., Cheng, M., and Chau, K.T. (2009) Review of electronic-continuously variable transmission propulsion system for full hybrid electric vehicles *Journal of Asian Electric Vehicles*, **7** (2), 1297–1302.

Wang, Y., Cheng, M., Chen, M., *et al*. (2011) Design of high-torque-density double-stator permanent magnet brushless motors *IET Electric Power Applications*, **5** (3), 317–323.

Xu, L. (2005) A New Breed of Electric Machines—Basic Analysis and Applications of Dual Mechanical Port Electric Machines. *Proceedings of the 8th International Conference Electrical Machine and System*, pp. 24–29.

Zheng, P., Liu, R., Wu, Q., *et al*. (2007) Magnetic coupling analysis of four-quadrant transducer used for hybrid electric vehicles *IEEE Transactions on Magnetics*, **43** (6), 2597–2599.

Zheng, P., Liu, R., Wu, Q., *et al*. (2008a) Compound-Structure Permanent-Magnet Synchronous Machine Used for Hybrid Electric Vehicles. *Proceedings of the International Conference on Electrical Machines and System*, pp. 2916–2920.

Zheng, P., Zhao, J., Wu, Q., *et al*. (2008b) Evaluation of the magnetic coupling degree and performance of an axial–axial flux compound-structure permanent-magnet synchronous machine used for hybrid electric vehicles *Journal of Applied Physics*, **103**(7), pp. 07F113–07F113-3.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto042
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Micro, Mild, and Full Hybrids

Thomas Pels and Carsten Kaup
AVL Schrick GmbH, Remscheid, Germany

1 Introduction 1
2 Classification of Micro, Mild, and Full Hybrids 1
3 Description of Basic Powertrain Functionalities 1
4 Examples of Micro, Mild, and Full Hybrid Architectures 3
5 Impact of Powertrain Architecture and Technical Solutions on the Classification of Hybrids 5
6 Summary and Conclusion 5
Related Articles 6

1 INTRODUCTION

At present, there is no common opinion about the classification of micro, mild, and full hybrids even among automotive experts. Sometimes, the proposed differentiation is the rated power of the electric drive or rated power versus vehicle weight. Other examples suggest to distinguish based on the power ratio between electric drive and combustion engine.

This article provides an explicit definition, based on hybrid powertrain functions. The basic functionalities and their typical characteristics will be explained in detail. Consequently, this article helps to avoid misinterpretations when discussing about hybrid electric vehicles (HEVs).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto046
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 CLASSIFICATION OF MICRO, MILD, AND FULL HYBRIDS

HEVs can be classified into micro, mild, and full hybrids. Table 1 shows the classification of HEVs, differentiated by powertrain functionality that can be categorized into three functional groups. Energy management contains functionalities to improve the efficiency of energy conversion or to reduce the waste of energy within the powertrain. Torque management implies the torque split in between the internal combustion engine (ICE) and the electric drive to operate the power sources within its optimal operation areas. The pure electric propulsion and its specific functionality are covered by the third group. The main motivation is the reduction of GHG emissions and the improvement of air quality, especially in urban areas.

On the basis of these classifications, range-extended electric vehicles (REEVs) can be categorized as full hybrids, whereas battery electric vehicles (BEVs) are actually not hybrids at all. The basic functions of all three groups are given in Table 1.

All additional functions and new functionalities, which might be relevant in the future, are referred to those functional groups. For instance, battery charging from the grid to enhance the all-electric driving range (plug-in hybrid functionality) is associated with electric propulsion. The basic functionalities will be explained more in detail within the next section.

3 DESCRIPTION OF BASIC POWERTRAIN FUNCTIONALITIES

3.1 Stop&start function

When a vehicle is stopped, for instance, at red traffic light or in traffic jam, a lot of energy is wasted to keep the engine

Table 1. Classification of hybrids.

Functional Groups	Basic Functionalities	Micro Hybrid	Mild Hybrid	Full Hybrid
Energy management	Stop&start Smart charging	×	×	×
Torque management	Regenerative brake Torque assist/boost	—	×	×
Electric propulsion	Pure electric drive Charge sustaining by ICE	—	—	×

The energy management functionalities of micro hybrids mentioned earlier do not require two power sources or two energy sources in all cases. Therefore, micro hybrids are not necessarily hybrid electric vehicles per definition.

running that can be significantly reduced or eliminated by implementing stop&start functionality. This function simply cuts off the engine when no propulsion torque is required, although it needs to be ensured that the engine stop is only executed after assessing the needs of auxiliary supply and engine conditions. Vehicle auxiliaries require energy whether it is mechanical or electrical, for example, for onboard power supply, air conditioning, power steering, or brake boosting. Concerning the combustion engine for instance cooling requirements and catalyst temperatures must also be considered.

The driver is the main actor for the stop&start functionality. Control units evaluate brake pedal and accelerator pedal positions as well as transmission states (such as clutch or gear positions) to decide whether it is appropriate to stop or restart the engine. The variety of hybrid electric vehicle architectures leads to a huge diversity in this functionality. Stop in gear versus stop in neutral and restart triggered by brake versus clutch triggered are typical examples for different strategies that significantly influence the fuel economy potential.

Apart from the drivers input, further system relevant signals contribute to the stop&start controls. Battery conditions such as state of charge (SOC) or state of health (SOH), electrical power demand, cabin heating, and cooling requirements are just some examples for further function initiators or inhibitors.

3.2 Smart charging

The operation of vehicle, independently of whether it is conventional or a hybrid electric vehicle, requires electrical power to supply the 12 V consumers. Unless the car is not a plug-in hybrid, the required energy must ultimately be generated from the combustion engine. In a conventional car, the alternator controller keeps the output voltage on a defined constant level. An intelligent control strategy (=smart charging) means not only to control the alternator on a constant set point but also to control the output voltage on different levels, depending on the driving condition (for instance, at full load, idle, and braking). This leads to a more efficient energy flow and helps to reduce fuel consumption.

Smart charging needs a battery monitoring system that helps to avoid the most common cause of vehicle breakdown—a flat battery. The battery monitoring system measures the battery voltage, current, and temperature and calculates from these all the information that reflects the condition of the 12 V battery (SOC and SOH). The smart charging function uses these values to guarantee a sufficient and continuous level of battery energy, so that the vehicle can be operated reliably under all conditions, even after a long stationary period. In hybrid vehicles, two different systems manage the smart charging control functionality. Some HEVs directly control electrical generators to convert mechanical energy to 12 V electrical energy, whereas others are using a DCDC converter to transform high voltage from the HV battery to the vehicle's electrical system voltage. In both cases, the hybrid master controller is responsible for arbitrating electrical energy transfer.

Besides the battery monitoring system mentioned earlier, it is mandatory to evaluate the vehicle operation mode to improve the overall fuel economy. The hybrid controller collects information about engine load, driver torque demand, and brake position to determine the target voltage, which is subsequently allocated to the 12 V battery and 12 V consumers.

3.3 Regenerative brake

During braking or vehicle coast down, some hybridized powertrain systems have the capability to recuperate energy by converting the vehicle's kinetic energy to electrical energy, which will be stored in an energy storage device (the most common energy storage device is an electrochemical battery). This functionality is called *recuperation* or *regenerative braking*. Energy is regained for free instead of being wasted to heat at the mechanical brakes or to the drag of the engine. The amount of energy and therefore the recuperation efficiency depends on the powertrain architecture and the brake system. Some brake systems provide

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto046
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

active blending between a mechanical and an electrical brake system that increases significantly the utility of recuperation where others allow only parallel braking and use only part of the overall potential.

Independent of the technical solution, the system needs to assure that regenerative braking is not jeopardizing the vehicle's braking ability on varying road surfaces. Thus, the hybrid controls have to ensure the driver input and ABS and ESP signals as well as the torque/power limitations that are reflected in the recuperation algorithm. The challenging task, especially for systems without brake blending capabilities, is to guarantee reproducible brake behavior independently of the charge acceptance condition of the HV battery. Mandatory for this functionality is an accurate battery monitoring without violating the battery limits in terms of energy and charging power thresholds.

3.4 Torque assist/boost

ICEs have typically low fuel economy at high engine speed, at moderate loads, or at high dynamic response. In hybrid electric vehicles, the additional power source can assist (boost) the combustion engine. Hence, the boost functionality gives the opportunity to avoid high dynamic engine torque response and therefore to operate more efficiently. In addition, NO*x* emission peaks of diesel engines in transient engine operation can be avoided or at least significantly reduced.

Similarly, with the aim to increase overall efficiency, powertrain design utilizes downsizing and downspeeding. Downsizing is a measure to increase the specific engine loads. Downspeeding helps to lower the engine speed and increases the loads at the same time. Both of which may negatively influence the drivability because these measures limit the engine torque reserve. As mentioned earlier, with the help of an electrical motor to support the combustion engine, this impact can be reduced or even eliminated. The boost function can also be used to realize alternative combustion systems, such as using the Atkinson or Miller cycle without degradation of either power or torque response at lower engine speeds. There are many more powertrain functionalities to increase comfort or efficiency where the electric boost of the second electric power source is acting as an "enabler." Electric four-wheel drive with the help of an electric axle or the electric compensation of torque interruptions of automated manual transmissions during gear shifting are just a few examples.

3.5 Pure electric drive

Pure electric driving means—vehicle propulsion and accessories powered by the battery with the combustion engine off. In a conventional powertrain, the combustion engine has to provide the power according to the driver's demand. Therefore, ideally the engine should have high efficiency in all operation modes such as idling or cruising at constant velocity or acceleration. Both gasoline and diesel engines have their peak efficiency typically at high loads and moderate engine speeds. In a conventional powertrain, the engine is working seldomly at the best specific operation point, which results in lower average efficiency, especially in city traffic. With this in mind, pure electric driving can help to increase the fuel efficiency in a hybrid electric vehicle. Furthermore, this functionality enables zero emission driving that may be needed for accessing dedicated restricted areas.

The hybrid supervisory controller determines the propulsion torque demands and enables the pure electric drive operation based on the accelerator pedal and brake pedal movement, selected driving mode and selected driving direction, and the torque/power limitations of the hybrid components.

Depending on the actual SOC of the battery, the supervisory controller decides between pure electric and charge sustaining mode, which will be described in the following section.

3.6 Charge sustaining mode

Once the electrical energy of the HV battery drops below a certain SOC threshold, the ICE provides energy, depending on the powertrain architecture, to a generator or direct to the wheels to extend the driving range. In most conditions, the power is primarily used for propulsion. On an average basis, the combustion engine will maintain the HV battery at a minimum SOC for extended range operation. Occasionally, the energy produced by the engine is used to recharge the battery. In either case, the distribution is controlled by the hybrid supervisory controller according to the SOC and health of the HV battery pack and the required traction power.

4 EXAMPLES OF MICRO, MILD, AND FULL HYBRID ARCHITECTURES

4.1 Micro and mild hybrid families

Hybrid powertrain functionalities can be realized with a wide variety of powertrain architectures and technical solutions, as mentioned earlier. In general, micro and mild hybrids have no ZEV range and use parallel hybrid architectures. In these configurations, the electric motor

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto046
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

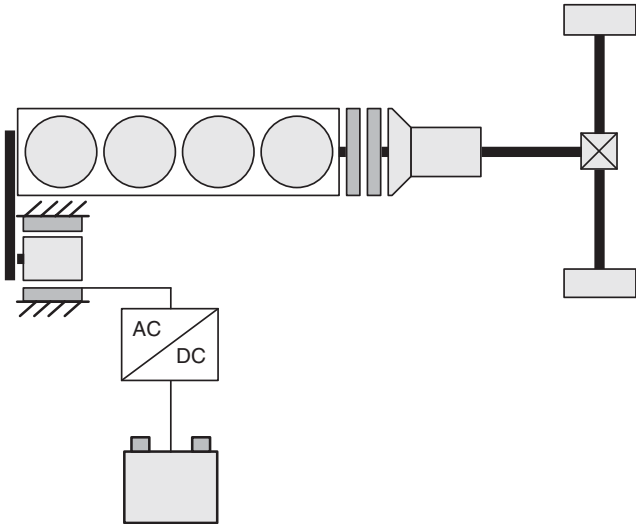


Figure 1. Micro hybrid in belt starter generator architecture.

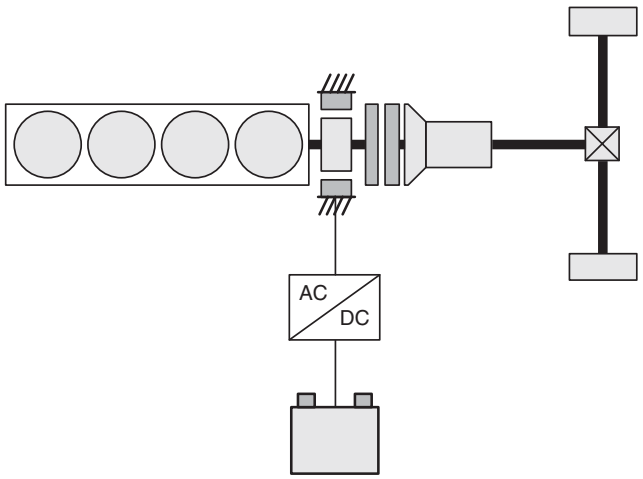


Figure 2. Mild hybrid in crankshaft starter generator architecture.

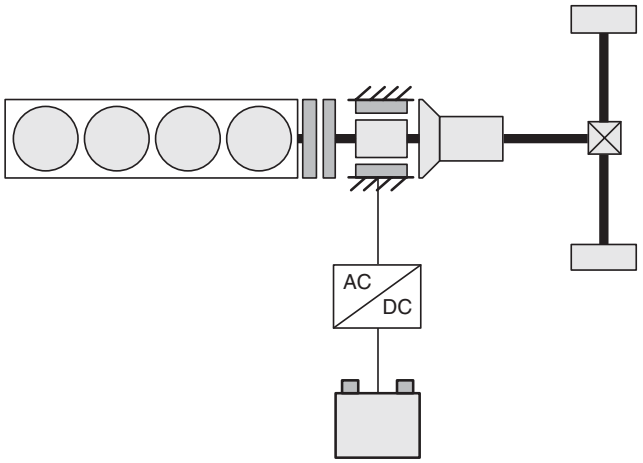


Figure 3. Mild or full hybrid with transmission-integrated E-motor.

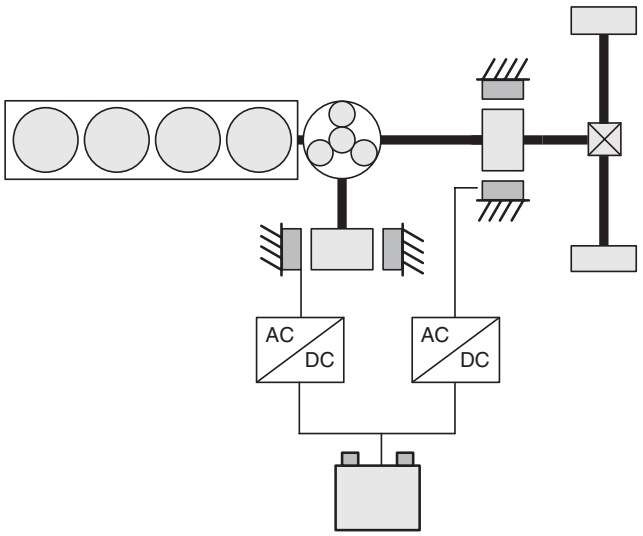


Figure 4. Full hybrid with power split configuration.

is usually a belt-driven starter generator or an integrated crankshaft starter generator (Figures 1 and 2).

Alternative to starter generators, Figure 3 shows a parallel hybrid configuration where the electric motor is integrated into the transmission. Especially, in this architecture, depending on the electric motor size, this powertrain offers also full hybrid functionality.

4.2 Full hybrid configurations

A power split concept (Figure 4) features a planetary gear set to split the power of the combustion engine into a mechanical path (directly to the wheels) and an electrical path (the generated electrical energy is either stored in a battery or is used to power a traction motor mechanically connected to the wheels).

An alternative concept for full hybrids is the parallel hybrid architecture (Figure 5). The parallel hybrid powertrain with an electric machine mounted directly to the crankshaft of the combustion engine with one or two clutches in line (before and/or after the electric motor) offers benefits when being added to an existing conventional drive train.

In the series hybrid configuration (Figure 6), the ICE is mechanically disconnected from the drive train and charges the battery through a generator to overcome the limitations of the battery's energy content.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto046
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

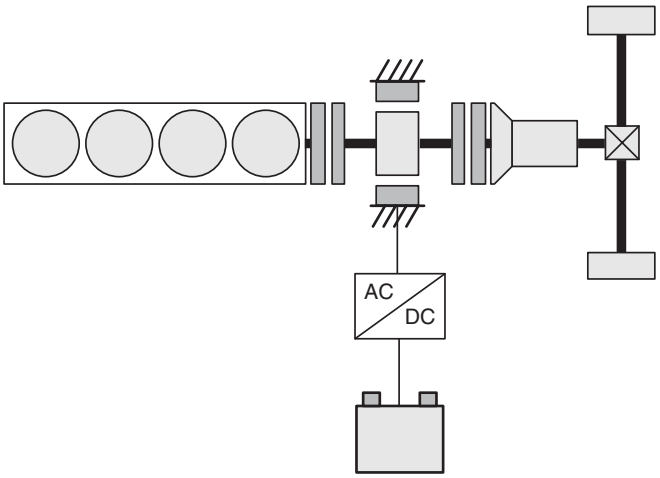


Figure 5. Full hybrid with parallel hybrid architecture.

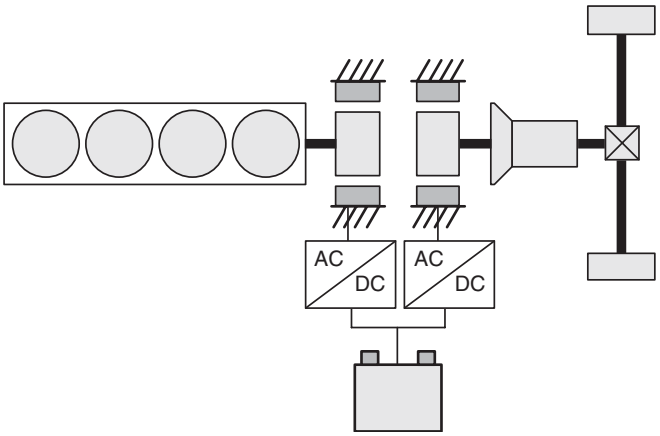


Figure 6. Full hybrid with series hybrid architecture.

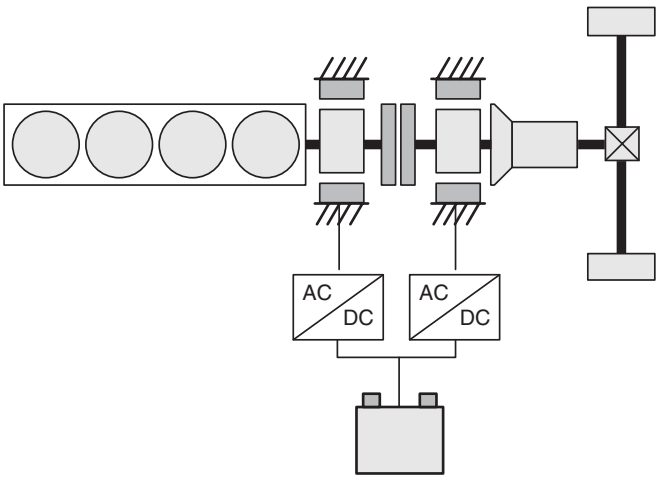


Figure 7. Full hybrid with series–parallel hybrid configuration.

The combined (series–parallel; Figure 7) configuration is equipped with two electric motors and a clutch in-between, which allows switching between series and parallel driving mode.

5 IMPACT OF POWERTRAIN ARCHITECTURE AND TECHNICAL SOLUTIONS ON THE CLASSIFICATION OF HYBRIDS

Hybrid powertrain functionalities can be realized with a wide variety of powertrain architectures and technical solutions, as mentioned earlier.

However, there is no impact of the (detailed) technical solution on the classification of a hybrid electric vehicle as the classification is purely related to system functionalities. Furthermore, it is not dependent on the specific realization of the functions. For example, a boost function can be implemented using a belt-driven or a crankshaft-mounted electric motor, alternatively using a transmission-integrated EM or a wheel hub motor. Moreover, the electric machine itself can be realized as an induction motor, permanent magnet synchronous machine, switched reluctance machine, transversal flux machine, and so on. Of course, there is a significant influence of technical features and component characteristics (e.g., efficiency) on the potential of the hybrid functions (e.g., fuel consumption), but this is not decisive for the classification.

This context applies also for different powertrain architectures, such as series, parallel, or complex structures. The drive train configuration and the position of the electric machine(s) might have an impact on the envisaged potentials of functionalities but not necessarily on its classification according to Table 1. For example, the fuel consumption reduction potential of a regenerative brake function depends (among others) on the ability to switch off the combustion engine while decelerating—characterized by the hybrid powertrain architecture.

It is obvious that there are exceptions as specific hybrid functionalities require dedicated attributes from the powertrain architecture, such as the ability to disconnect the combustion engine from the drive train to enable pure electric drive.

6 SUMMARY AND CONCLUSION

Many hybrid architectures and technical solution exist and many different approaches classify micro, mild, and full hybrids. Classification by rated power of the electric drive

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto046
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

or rated power versus vehicle weight is not really sufficient because it is impossible to define clear borders.

Currently, we do not have dedicated legislation prescribing specific component characteristics or properties, such as minimum rated electrical power to be installed or the like thought this might change in the future. The electrification of the powertrain offers new functionalities to improve fuel economy, reduce emissions, or enhance driveability. For this reason, the generic classification based on powertrain functions remains more flexible, especially if the associated statutory requirements are not harmonized worldwide.

RELATED ARTICLES

Basic Consideration
EV Powertrain Configurations
EVT and E-CVT for Full Hybrid Electric Vehicles
Overview of Electric, Hybrid and Fuel Cell Vehicles
Parallel Hybrid Electric Vehicles (Parallel HEVs)
Range Extender EV
Series Hybrid Electric Vehicles (SHEVs)
Series-Parallel Hybrid Electric Vehicles

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto046
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Range Extender Ev

Frank Beste, Günter Fraidl, Vincent Benda, and Martin Atzwanger
AVL List GmbH, Graz, Austria

1 Introduction 1
2 Reality of BEVs 1
3 Technical RE-Concept Selection 4
4 Range Extender as Highly Integrated Performance Module 7
5 RE Vehicle Integration 13
6 Real-Life Testing of the Range Extender 15
7 Summary and Conclusion 18
Related Articles 20
References 20

1 INTRODUCTION

Replacement of fossil fuels by regenerative and, therefore, CO_2-neutral energy sources is the long-term motivation for sustainable mobility. While hydrogen as energy carrier together with the fuel cell technology have been the focus during recent years, the focus now shifts toward battery electric mobility in pure electric or hybrid configuration. The enhanced development of the Li-ion battery technology started with consumer cells (laptop, mobile phone, etc.) many years ago. The Li-ion technology is now being adapted for the needs of high energy cells for automotive purposes. However, even with this development history, essential limitations of energy density and cost of automotive battery systems remain.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

First results from battery electric vehicle (BEV) fleet testing are available and influence the assessment of different drivetrain electrification approaches. Not only the reduced all-electric range (AER) under real-world conditions but also the restrictions to the recharging of the high voltage (HV)-battery result in a subjective "range anxiety." The recent efforts to develop local zero-emission mobility focus on plug-in hybrids with full purpose driving performance and BEVs as the only real zero-emission vehicle for mega city mobility. Now, developers increasingly consider the range extender (RE) approach as a solution for secure and cost-efficient range extension for BEVs (List, 2009).

2 REALITY OF BEVS

The simple comparison of published data of nominal BEV, AER, and HV-battery capacity answers the basic question about such vehicles—with increasing AER, the HV-battery capacity, weight, box dimension, and cost need to increase significantly (Figure 1).

In most cases, the published data does not consider real-world driving conditions. Published CO_2 emission specifications are generated with vehicles in minimum resistance and optimum weight configuration in standardized but not always practice-oriented test cycles. If these values are assessed with characteristic vehicle weight and drive resistances, as known from conventional vehicles, then the NEDC (New European Drive Cycle) requires increased battery capacities for the same AER.

This is even more valid if a realistic drive cycle for mega city application is considered. One example that better fits these BEV-use characteristics is the NYCC (New York City Cycle) with stop-and-go traffic, a much lower average speed, and high dynamic requirements. Even with the utilization of brake energy recuperation and without any air-conditioning or heating, the NYCC requires an

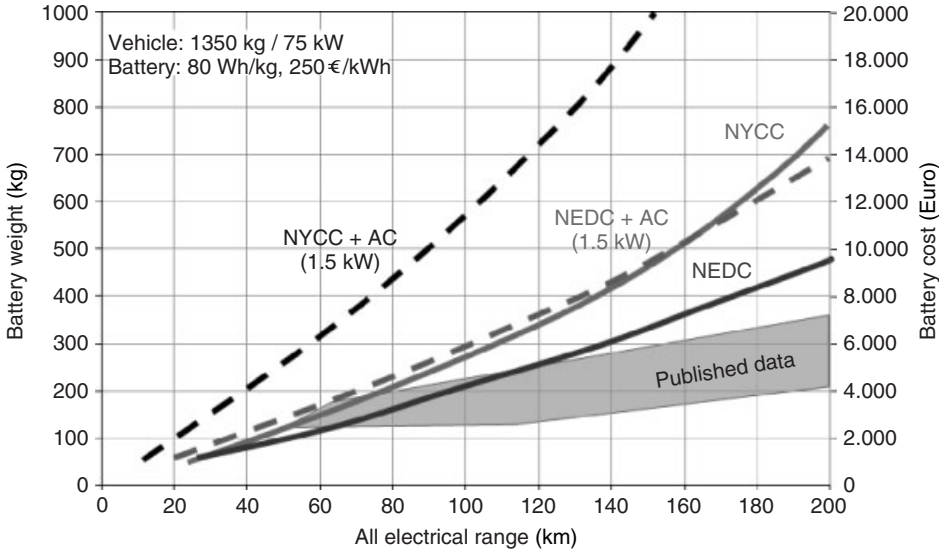


Figure 1. Influence of the AER on battery capacity, weight, and cost (simulation).

additional, approximately 40%, increase in the HV-battery energy content in comparison to the NEDC.

Furthermore, any air-conditioning or heating requirements for the passenger compartment have much higher effect on the AER than known from conventional vehicles. Assuming a vehicle operation at 32°C ambient temperature and 21°C in the passenger compartment, a reduction of the AER between 25% (NEDC) and 50% (NYCC) can be expected. A BEV capable of about 200-km AER in the NEDC without air-conditioning or heating reaches only about 60-km AER in the NYCC with worst-case consideration of auxiliary consumers such as air-conditioning. If a minimum range of 100 km must be reached under severe conditions, an HV-battery weight of about 500–600 kg results. With such a weight growth, the vehicle structure and the performance of the electric traction motor also need to be enhanced. As only 20–35% of the weight-driven additional traction energy can be recuperated, the increase in HV-battery size cannot fully be accounted for additional AER. Even with the production scale driven optimistic prognosis for a specific battery-cell cost reduction down to 250, €/kWh the HV-battery is still the most significant cost burden for the economic operation electric mega city vehicles (MCVs).

A more differentiating picture of advantages and disadvantages of BEVs results from a vehicle start at cold ambient condition. For conventional vehicles, the significantly increased fuel consumption at cold start primarily results from heating the internal combustion engine (ICE) and transmission and from the rich fuel mixture of gasoline engines. Nevertheless, for gasoline engines, and with some limitations for diesel engines, the ICE generates sufficient additional heat for the passenger compartment. Owing to the slightly extended ICE warm-up phase for the passenger compartment heating, the cold-start fuel consumption increases, but can be neglected in relation to the total energy consumption.

For BEVs, the rather high efficiency of the electric traction motor and inverter at low system temperatures does not generate sufficient additional waste heat to supply the passenger compartment. An additional electric heating device for the passenger compartment is necessary. For the BEV cold start, the increase in energy consumption does not result from efficiency losses of the traction system but from the heating of the passenger compartment (Figure 2).

The further reduction of the ambient cold-start temperatures below −10°C also effects the HV-battery's ability to deliver the required performance for the electric traction motor. The question "driving or heating" can become relevant for the driver. Such issues are unknown for conventional vehicles and not accepted by the customer. Furthermore, at such low temperatures, the HV-battery-charging ability also suffers. This must be considered for the vehicles recuperation strategy. Therefore, a volume production of BEVs seems feasible with two strategies:

1. Vehicle design with sufficient HV-battery capacity: for typical everyday AER requirements, the HV-battery is oversized. Together with weight, package, and cost limitations, no vehicle-use flexibility comparable to conventional vehicles seems feasible. A "change of mind" and "adaption for drive behavior" is required. Risks for the vehicle availability are simple user

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

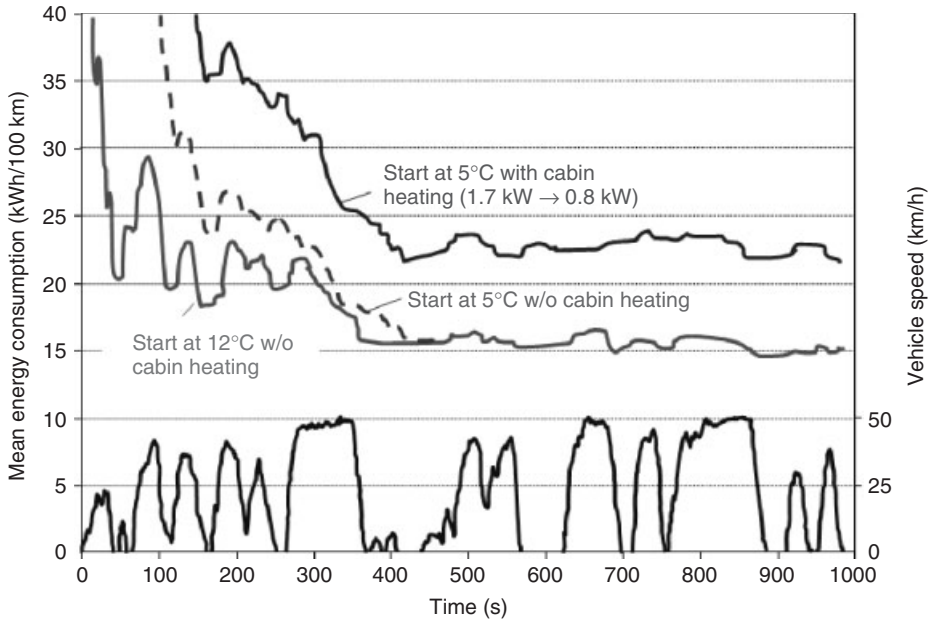


Figure 2. Specific energy consumption of BEVs at cold start.

mistakes such as insufficient charging or unexpected demands.

2. Design of the HV-battery for the typical everyday AER requirement and consideration of an additional RE system for the further energy requirements: for the definition of the typical daily AER requirement, different analyses exist. For an industrialized country with a dense population such as Germany, 70–80% of all daily driving range requirements are below 50–60 km (IVT, 2004). For MCVs, this percentage is expected to be even higher. Such AER requirements and real-world conditions can be solved at acceptable battery weight, cost, and box dimension. The RE is the enabler for a broad market acceptance for electric mobility (Figure 3).

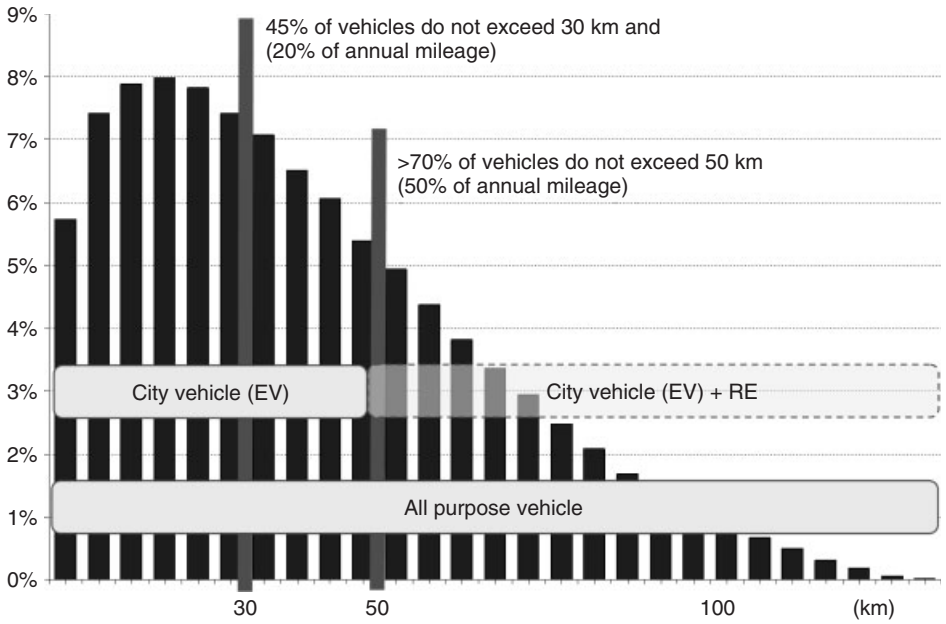


Figure 3. Statistics of all daily driving requirements in Germany (IVT, 2004). (Reproduced with permission from AVL.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

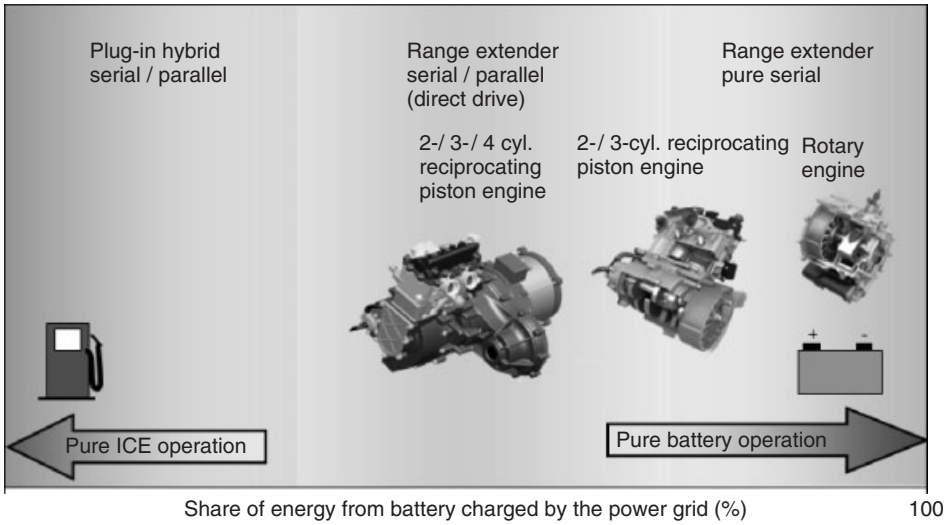


Figure 4. Dependence of different drivetrain configurations from the share of battery electric driving.

Conventional drivetrains with ICE meet a wide variety of customer usage profiles. The electrified drivetrain requires a stronger focus on the relevant use case and vehicle concept targets. For any hybrid configuration, this is especially true for the contribution of the plug-in-charged HV-battery in relation to the energy delivered from the ICE. The differentiation of technical concepts requires more variants than known in conventional drivetrains (Figure 4).

For "full purpose vehicles," wide parts of the usage profiles have to be propelled by the ICE, and high requirements on performance and the total vehicle range potential have to be fulfilled. The best technical approaches for these requirements are plug-in hybrid solutions with parallel configuration of ICE and electric traction motor. The required ICE performance level needs to be in the same range or even higher than that of the electric traction motor as battery electric and long-range ICE driving will be required.

With an increased share of battery electric driving, a series or serial hybrid RE system would allow to utilize the key advantages of the electric drivetrain for the wide majority of daily drive distances without losing the flexibility given by an ICE-driven energy source. The required performance level of the ICE in series or serial configuration needs to be lower than the peak performance of the electric traction motor by a factor 2–4, as only average energy demands need to be supplied. In comparison to parallel plug-in hybrids, the reduced RE performance allows a very compact and cost-efficient design and more flexible system integration in the vehicle. The disadvantage of additional efficiency losses by the double energy conversion (mechanical power from the ICE → electric power from the generator → mechanical power from the electric traction motor to drive the wheels) can only be accepted, if the intended use of RE system is limited to a small part (10–30%) of the main BEV operation.

In general, the term *range extender* is used for a variety of technical concepts, including series or serial hybrids with low electric range, parallel plug-in hybrids, and BEVs with an additional emergency or limp-home functionality. The best configuration depends on the share of battery electric driving. The consequent development of plug-in hybrids toward increased AER is given by a switchable parallel/series or serial hybrid solution, which avoids energy conversion losses at higher traveling speeds by the switchable parallel ICE configuration (Grebe, 2009; Steiger, 2009; Najork *et al*., 2009). For vehicles with clear focus on the pure battery electric driving, the series or serial configuration should be preferred (Fischer, 2009; Beste *et al*., 2009).

When considering these alternative technologies, the pure BEV seems to be limited to niche applications. It can be questioned, if broad market acceptance as required for volume production can be reached with the described restrictions for the driver and the rather high purchase price. It seems as if higher sales numbers for BEVs are only possible with significant governmental funding and/or city access restrictions for ICE-driven vehicles.

3 TECHNICAL RE-CONCEPT SELECTION

For a range-extended electric vehicle (REEV), the daily driving needs define the HV-battery size (e.g., 50-km range and $\leq$100 km/h). The required energy reserve is supplied

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

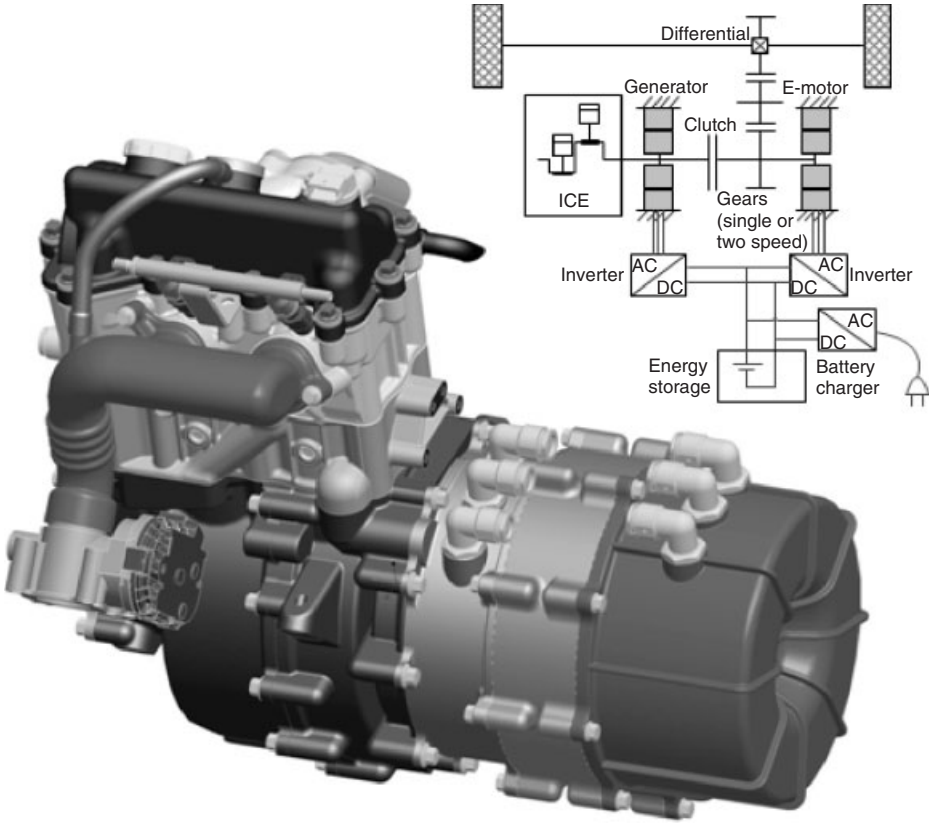


Figure 5. RE concept with switchable parallel/series or serial ICE configuration

by the ICE of the RE system. The battery cost of a REEV can become independent of the total vehicle range potential. The selection of the best technical approach for the RE system is driven by the intended vehicle usage profile, legislative requirements, and the continuation of existing production facilities and technologies of the system manufacturer. However, the main driver is the vehicle concept.

3.1 Vehicle concept starting from conventional vehicle architecture

The RE system is an extended hybridization level that enhances the vehicle's AER ability together with the plug-in functionality and an increased battery size. Because of the switchable series or serial/parallel configuration, the generator performance can be lower than that of the ICE [e.g., by 30% (Najork *et al*., 2009) to 60% (Steiger, 2009)] but requires a wider operation range. The ICE should primarily be activated at higher vehicle speeds and ambient noise. The ICE operation strategy should be defined in accordance with the vehicle speed and load-connected noise expectations of the driver to minimize acoustic problems. The ICE can be derived from existing production engines to limit development efforts and production cost (Fraidl *et al*., 2009; Steiger, 2009; Najork *et al*., 2009). Figure 5 shows a new specific ICE design that has been done for the needs of the switchable parallel/series or serial ICE configuration. The two-cylinder in-line four-stroke ICE has been optimized for production cost and efficiency.

3.2 Vehicle concept starting from a BEV

The RE system is a chemical-electric energy conversion device that supplies energy to the electric traction motor in parallel to the HV-battery. It is an extension of the HV-battery capacity and has a similar functionality as an emergency power generator for a BEV limp-home functionality. These boundary conditions for the RE system integration support the pure series or serial configuration of the ICE. With a vehicle operation strategy that considers the HV-battery to cover the dynamic peak-load demands, the RE performance can be defined by the factor 2–4 lower than the peak performance of the electric traction motor. The required RE performance is defined by the average performance demand at a predefined vehicle layout speed.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In comparison to conventional drivetrains, the following key requirements for the design of the RE system result:

- The RE will need to start automatically, when the HV-battery state of charge (SOC) drops below a predefined minimum HV-battery energy reserve or if required by the system operation strategy. Therefore, ICE operation is decoupled from the very silent driving experience of an electric vehicle and should not or only secondarily be noticeable by the customer. Excellent NVH (noise, vibration, and harshness) properties are the decisive key requirement for customer acceptance.
- For the most part of the vehicle usage, the battery electric driving, the RE reduces the vehicle efficiency by its additional weight. Lightweight design for the RE and its vehicle integration are essential.
- The operation strategy allows to run the RE system in only a few load points without any significant transient or idle requirements. Additional specific fuel efficiency advantages and unconventional ICE concepts become possible.

In addition to the mandatory safety and legislative requirements (emission targets, etc.), the priorities for a series or serial RE development can be summarized:

- Lowest acoustic radiation and excitation
- Dependable system availability even after a long stand-still
- Compact package and high power density
- Low system costs, good vehicle integration properties, and performance scalability
- High system efficiency.

Another design criterion for the RE is its power output. This depends on several factors. The maximum power of the vehicles electric propulsion motor is primarily defined by vehicle acceleration- and climbing requirements. These peak power requirements are significantly higher, than the continuous power requirements. The RE can have a smaller power level if the HV-battery energy reserve can cover the required dynamic power peaks. Figure 6 shows a dimensioning diagram. Plotted are driving resistance curves for several grades. The dotted lines represent the total demanded power including the use of auxiliaries. The selected vehicle (1.475 kg fully occupied) requires a constant velocity of 100 km/h and a propulsion power of 13 kW, correlating with an approximately 15 kW total electric power supply requirement. With an HV-battery energy reserve of 2 kWh, the total combined power allows a driving distance of 21 km at a 2% slope at 100 km/h. It also can be determined what constant speeds are possible at certain slopes and how much acceleration can be performed at a given vehicle speed (Fischer, 2009).

It is also essential to determine the vehicle's remaining abilities, if the HV-battery energy reserve has been depleted. For this "miss-use" of the originally intended

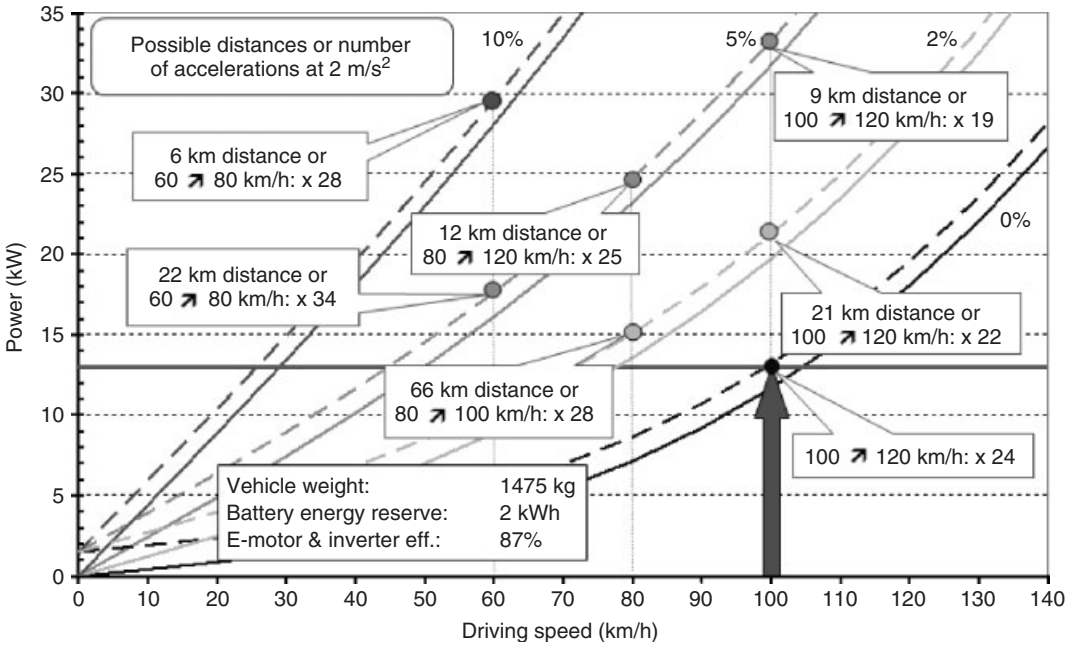


Figure 6. AVL pure range extender—performance requirements/benefit of energy reserve

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

operation range of the REEV, the RE gets into a limp-functionality with a reduced vehicle performance according to the performance definition of the RE. Potentially critical vehicle operations are:

- Remaining acceleration ability of the maximum-weight vehicle at very steep grades (>14%) and a vehicle speed of 30 km/h
- Maximum highway speed at typical highway grades (~5%) should be above 100 km/h to allow passing of trucks, and so on.

The "miss-use" situations with depleted HV-battery should be avoided by the vehicle status information provided to the driver. Nevertheless, an additional high power "emergency" or "highway" load point of about 25 kW for the described vehicle specification can help avoid critical vehicle situations, but would not meet the tight NVH and efficiency targets of the normal RE operation points.

4 RANGE EXTENDER AS HIGHLY INTEGRATED PERFORMANCE MODULE

The RE as a self-regulating auxiliary power supply unit for BEVs can be consequently optimized for operation at few load points by a specific system design of the generator-starter (GS) and the ICE. Reduced development effort, reduced system complexity, and optimized fuel efficiency at its predefined operation points are essential advantages. The complete functional, mechanical, thermal, and control-specific integration of the ICE and the GS and minimum interfaces with the vehicles (electric power connections, control signals, fuel supply, cooling and mechanical mountings) allow diverse vehicle applications at a low development effort.

An essential step of further concept definition is the comprehensive comparison of alternative concept approaches for the system core functionalities. Figure 7 shows an extract of the concept assessment for the most promising ICE concepts that have been derived from the first and very open assessment loop of known technologies. The boundary condition to meet the most stringent existing and future emission legislations requires significant effort and cost for two-cycle approaches with single cylinder or boxer configurations. Four-cycle gasoline concepts will meet the requirements with the lowest additional development effort and support the utilization of existing production facilities.

Fuel cell RE concepts are not mentioned in detail. Their key advantages are that no additional energy conversion from shaft power to electric power is required and that fuel cells reach best NVH and emission properties. Nevertheless, for the RE technology, key requirements are the reduced production cost compared to the alternative bigger HV-battery and the possibility to rapidly fill up the fuel tank at any gas station. The current fuel cell development status and the need for hydrogen do not support these targets. For these reasons, fuel cell REs are not discussed in detail.

The rotary piston engine (RPE) generally exhibits concept-specific advantages regarding power density, acoustics radiation, and compactness. In addition, the known dominant concept-specific disadvantages such as HC raw emissions, durability of the apex seals, and elevated fuel consumption can be avoided completely or

	Otto 2-cycle 1-cylinder piston-contr. balancer shaft	Otto 2-cycle 2-cylinder opposed piston piston contr.	Rotary engine single piston	Otto 4-cycle 1-cylinder balancer shaft	Otto 4-cycle 2-cyl. inline balancer shaft	Diesel 4-cycle 2-cyl. inline balancer shaft
I NVA	o	+	++	–	o	–
II Package	–	–	++	+	o	o
III Weight	o	o	++	+	o	–
IV Product cost	o	o	–	+	o	–
V Efficiency	–	–	–	o	o	++

Figure 7. Technology matrix for the ICE.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

at least considerably by the described special RE requirements. This presents the RPE as a concept alternative with good potential compared to rather conventional approaches. Besides the described development priorities, the ICE concept must also be assessed with regard to the share of RE operation for the intended vehicle concept.

For an RE system operation share of above 20–30% of the total vehicle operation, the RPE results in higher total CO_2 emissions because of its engine concept-related specific fuel consumption (SFC), which is about 6–10% worse. Figure 8 considers CO_2 emissions from ICE operation and from plug-in charging according to the German power-plant mix of 625 grCO_2/kWh and an assumed charging efficiency of 95%.

If the REEV is mainly operated in the original intended battery electric mode, the compactness and the weight advantage of the RPE approach compensates the disadvantage of the RPE SFC and results in a vehicle CO_2 advantage for this concept.

For the GS system, the permanent magnet synchronous machine (PMSM) technology is advantageous because of its compactness, high electric efficiency, and comparably low development risks (Figure 9).

Figure 10 summarizes the selected friction-optimized two-cylinder four-cycle in-line gasoline engine because of

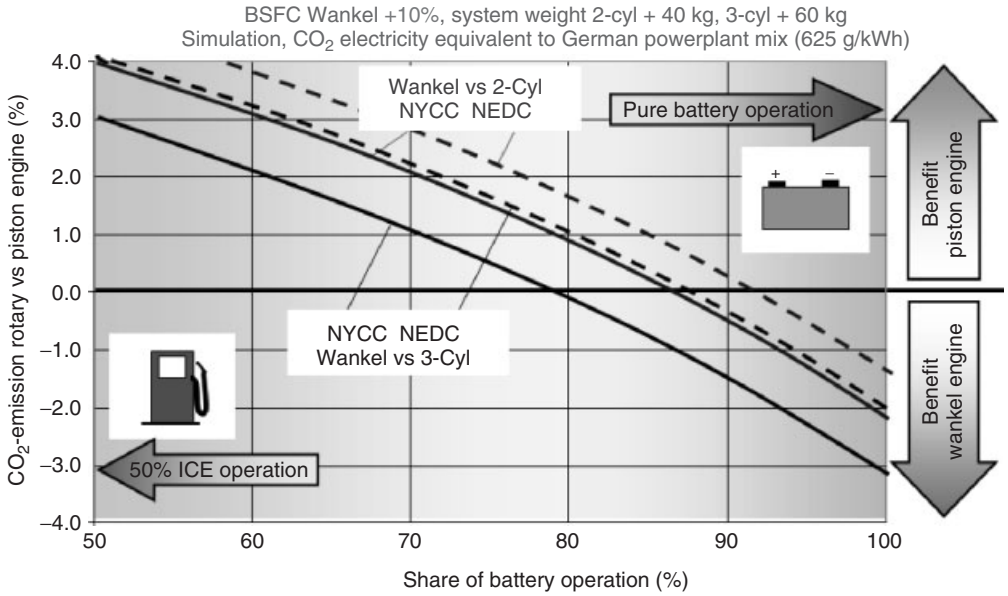


Figure 8. Effect of the RE operation share on the relative CO_2 emissions of different series or serial RE approaches.

	A-synchronous machine	Permanent magnet synchronous machine	Switched reluctance machine
I NVH	+	+	–
II Package	+	++	O
III Weight	O	+	O
IV Product cost	+	O	++
V Efficiency	+	++	O

Figure 9. Technology matrix for the GS.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

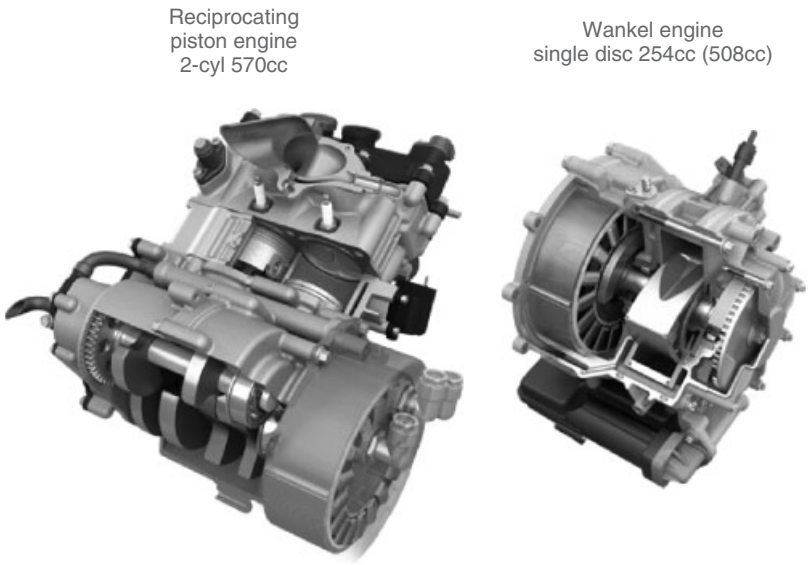


Figure 10. Alternative core modules of a series or serial hybrid RE system (15 kW electric power @ 5000 rpm)

its similarity to existing production engines and the single rotary piston RPE. The following design concepts represent the most compact integration of ICE and GS. Significant advantages in box size and weight can be achieved in comparison to RE systems with existing automotive engines.

4.1 ICE—two-cylinder in-line gasoline engine

The two-cylinder engine concept (Figure 11) considers the boundary conditions for excellent NVH, high efficiency, and engine friction as well as low production cost by the following main features (Atzwanger *et al*., 2010):

Excellent NVH:

- 90° crankpin offset and balancer shaft
- Integrated exhaust manifold in cylinder head.

High efficiency and low engine friction:

- Roller bearings on main and balancer shaft bearings as well as on the camshaft

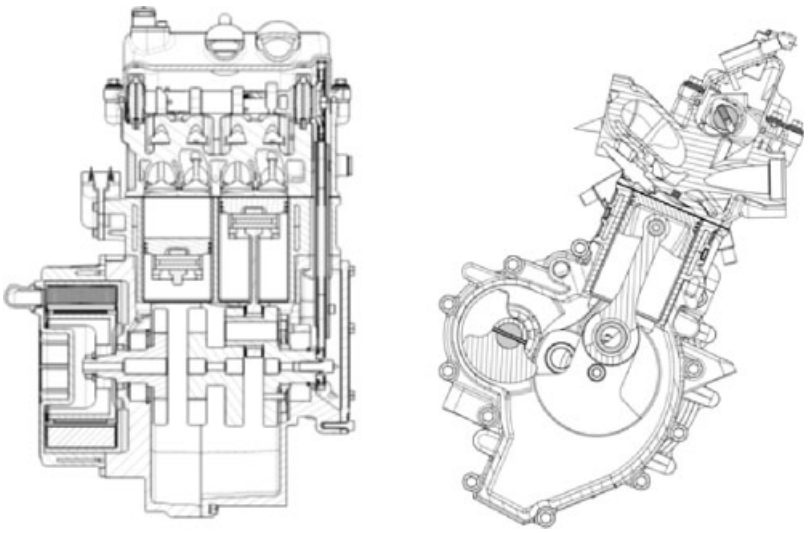


Figure 11. Design example for RE with two-cylinder in-line gasoline engine

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- Nonpressurized lubrication without oil pump
- Two-valve direct acting valvetrain.

Low production cost:

- Two-bearing crankshaft and camshaft design
- No oil pump
- Integrated high pressure die cast crankcase—generator housing
- 2 V direct acting valvetrain
- Integrated exhaust manifold.

To fit the very limited available space for an RE unit, the engine architecture allows installation angles of up to 45° (Table 1).

The thermodynamic layout of the engine considers the set power target of up to 18 kW. Using FMEP (friction mean effective pressure) values, which consider the friction-optimized engine architecture, the achieved SFC values for this engine are in the range 250–260 g/kWh @ 5000 rpm at a BMEP (break mean effective pressure) of 7 bar. For this engine load, it is possible to operate the engine at a stoichiometric air–fuel ratio ($\lambda = 1$) without exceeding the set limits of maximum 850°C exhaust gas temperature. The predicted peak firing pressure in the cylinder is 55 bar.

In contrast to well-established concepts such as bed plate or deep skirt crankcase design, a vertical split crankcase with separate cylinder unit was chosen, which is usually based on motorcycle applications. The crankcase consists mainly of two pieces, the right and the left sides with a split line in the middle, having a short offset to ensure a bolting possibility for the middle two-cylinder head bolts. The advantage of such a design is that a reduced number of parts are required because of the omission of the bearing caps and the resulting less complex production This layout also gives the opportunity for pressed-in main roller bearings, which reduces the required number of parts (e.g., circlips and the machining associated with axial fixation).

The crankshaft is designed as a built assembly that is also common in motorcycle applications. The inner two counterweights form one part that is connected to the outer counterweights via press-fit crankpins. This enables roller-supported connecting rods, which ensures low crank-train friction in combination with the omission of the pressure oil circuit and a simplification of crankshaft machining (no long thin pressure oil drillings) (Figure 12).

The 90° crankpin offset configuration shows a balancing of second-order mass forces, so only first-order mass forces remain as the first- and second-order moments can be neglected. For this engine concept, the free first-order mass forces are balanced by one balancer shaft driven by the crankshaft with engine speed (Figure 13). To reduce NVH, a zero-backlash gear is used. The balancer weights on the balancer shaft offset each other by 90°. The shafts roller bearings are lubricated by the oil mist in the crankcase.

Following the target of a simple friction-optimized engine concept, the cylinder head is a two-valve, single overhead camshaft concept. The camshaft is supported only by two needle bearings to further optimize engine friction. The camshaft operates mechanical tappets to actuate one intake and one exhaust valve per cylinder. In the camshaft, a simple decompression device as also known from motorcycle engines is integrated to reduce the required starting energy. This mechanism opens the exhaust valves during the compression stroke slightly at low speeds of the camshaft. At higher engine speeds, the mechanism is deactivated by centrifugal forces. The exhaust manifold is integrated into the cylinder head and is cooled by the cylinder head coolant jacket. This design leads to a very compact design of the top end of the engine. In addition, the temperature of the exhaust gas is reduced (Figure 14).

The GS, considered a PMSM, is integrated into the engine housing. The stator is pressed-fit into the engine crankcase and the cooling circuit is shared with the cooling circuit of the ICE. The rotor is mounted on the rear end of

Table 1. Core two-cylinder engine specifications.

Key Specifications of the RE Unit with the Two-Cylinder Engine	
Engine configuration	Inline two-cylinder, gasoline, four-stroke, MPI, NA
Displacement	570 ccm
Bore	70 mm
Stroke	74 mm
Power ICE	18 kW (up to 25 kW) @ 5000 rpm
Compression ratio	11 : 1
Specific fuel consumption	250–260 g/kWh at design operation point (15 kW electrical output of RE unit)
Generator concept	Permanent magnet synchronous machine
Thermal management	Single circuit liquid cooling, integrated ICE and GS cooling
Electric output	15 kW @ 320–420 V

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Inline 2 cylinder configurations						
Crankpin offset	Firing order	1st order free mass forces	1st order free moments	2nd order free mass force	2nd order free moments around crank axis	2nd order free moments rectangular to crank axis
180°	180°/360°	No	Yes	Yes	Yes	No
360°	360°	Yes	No	Yes	Yes	No
90°	270°/450°	Yes	Low	No	Low	Low

Figure 12. Potential crank-train configurations for two-cylinder in-line engines.

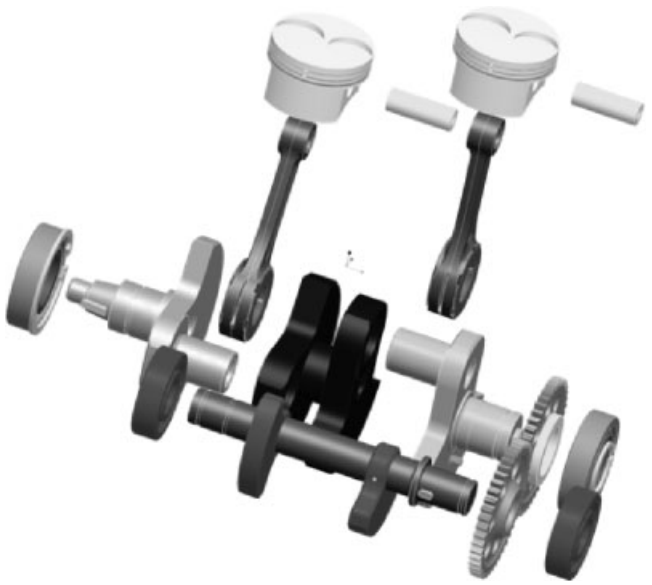


Figure 13. Crank-train and balancer shaft.

the crankshaft. This gives the possibility to further reduce the number of parts and assembly effort.

4.2 ICE—single-rotor RPE

To fulfill the unique RE requirements, a new and dedicated RPE design is required. Figure 15 shows the RPE in a compact common-shaft assembly with the GS. The thermodynamic layout was carried out for the quasi-stationary operation at the demanded electric output power of 15 kW. The defined engine speed of 5000 rpm represents an optimum regarding efficiency and NVH. The specific displacement of 254 ccm represents a compromise among the boundary conditions, low fuel consumption, little wear, beneficial airborne sound emissions, and required potential for a further performance increase. The thermodynamic simulation also shows that different levels of performance in the range 15–25 kW can be achieved with a unified geometry by increasing mean effective pressure and engine speed. By extending the rotary piston width by 20 mm, the displacement increases to 357 ccm and the potential electric output to approximately 36 kW (Figure 16, Table 2).

The centrally positioned common shaft of the RPE and the GS allows a compact design of the core unit. The high degree of shaft bending stiffness due to the short shaft length allows the elimination of the GS-sided third roller bearing. The low friction roller bearing design reduces the oil pump requirements down to the needs for the pressureless rotary piston cooling and lubrication of rotary piston seals. An integrated unbalance in the rotor carrier of the GS and a second balancing weight at the other shaft end eliminate the unbalance in the rotary piston and so fulfill the high NVH demands.

The joint cooling circuit of the ICE and the GS is integrated in a tubeless common housing. Owing to the low acceptable temperature level of the GS and the power electronics, the coolant temperature level is considerably lower ($<75^{\circ}$C) than that for typical conventional ICE configurations. This requires an adequate design of the friction-relevant components. In the RPE segment of the housing, only the "hot" part of the trochoid that experiences the heat transfer from the combustion is cooled. The temperature of the "cold" part is regulated by material heat conduction. The trochoid surface and the oil temperature are important for the wear and friction of the sealing components. The oil temperature lies at a higher level compared to the cooling medium temperature as oil is used for rotor-piston cooling.

The limited number of engine operation points for the pure series or serial hybrid configuration allows optimizing the thermodynamic layout. For the RPE, for example, the Atkinson cycle can be adapted. The late exhaust-port opening and the extended expansion phase not only reduces

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

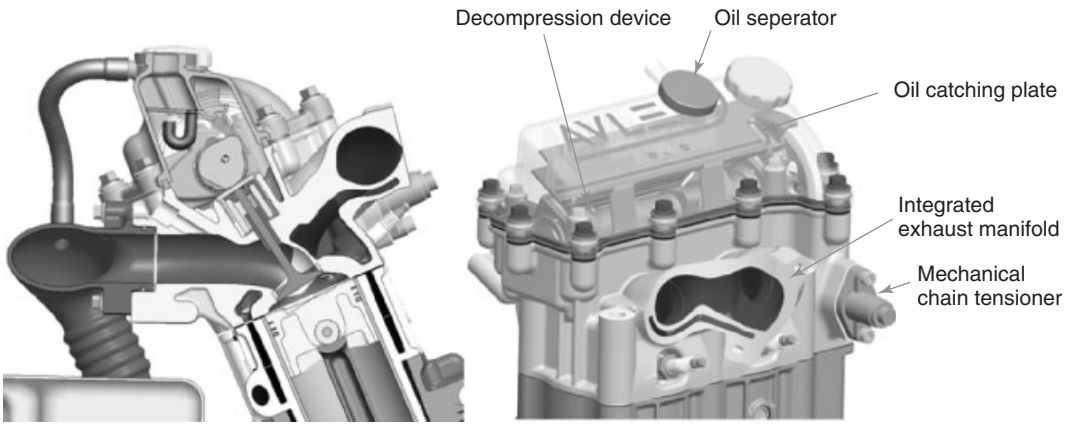


Figure 14. Cylinder head assembly.

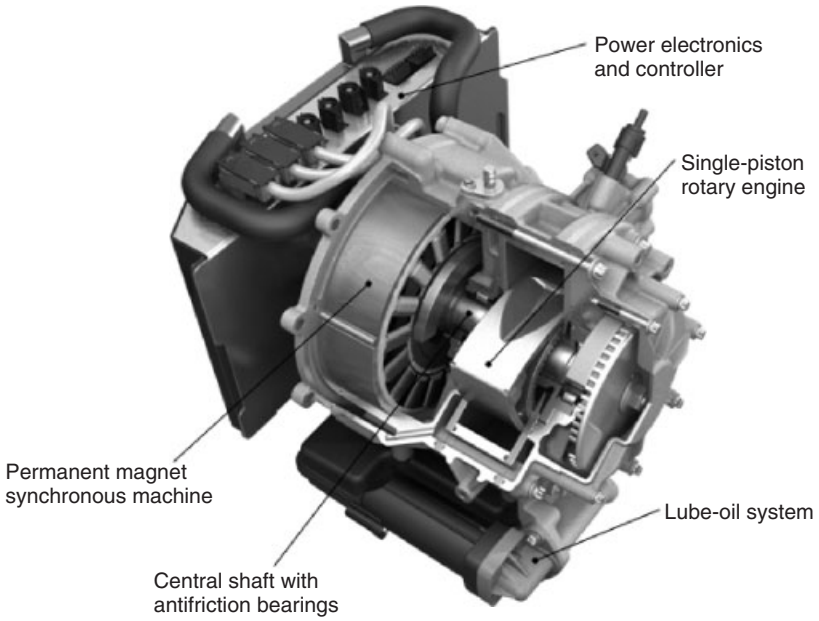


Figure 15. Design example for RE with single rotor-piston RPE.

exhaust gas temperatures and improves SFC and HC emissions but also reduces the acoustic relevant pressure pulses in the exhaust system. The achievement of future legislative emission targets by the definition of an optimum engine start procedure and the engine calibration requires comprehensive model-based simulation and test-bed validation.

4.3 Generator-starter and control system

The GS used in the RE has to fulfill requirements different from those of the electric propulsion motor of the vehicle. The required functions are basically limited to a fast and vibration-poor start of the ICE and the generation of electrical energy in a very limited operation range. This is possible with known technologies. However, the central demand for a very compact, efficient RE module with minimum production costs, which is optimized for the vehicle operation and achieves the demanded robustness, requires intensive development work. The system optimization regarding the reduction of overall package volume, weight, and costs has to be accomplished by an adequate detail design and simulation. The GS has to be designed for the required electrical power output in the HV-battery voltage range, for example, 15 kW in a voltage range

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Core RPE specifications.

Key Specifications of RE Unit with the RPE	
Combustion engine	Single-piston rotary engine (RE)
Displacement	254 ccm
Engine speed	5000 rpm
Specific fuel consumption	280 gr/kWh at design operation point (15 kW electrical output of RE unit)
Generator—starter machine	Permanent magnet synchronous machine
Electric output performance	15 kW at 320–420 V
Cooling system	Single circuit liquid cooling, integrated ICE and GS cooling
Communication	CAN bus
Sound emission (outside vehicle)	65 dBA (lm averaged sound pressure)
System weight (ICE + GS core module)	70 kg (35 kg)
Interfaces range extender—vehicle	HV and 12 V connector, CAN bus, fuel supply, Cooling, acoustic decoupled system mounts

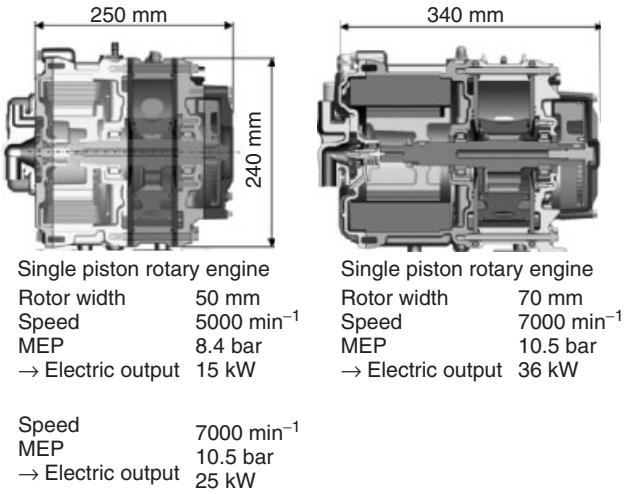


Figure 16. Core module of range extender system for 36-kW electric power @ 7000 rpm

320–420 V (12 kW at 250 V). The GS achieves a machine efficiency of up to 96%. Further performance classes can be obtained by scaling and adaptation of the GS design. Special attention is required for the additional thermal and vibration loads by the ICE on the electric components. For optimum cooling conditions, the cold coolant first enters the power electronics module and the integrated GS cooling jacket and is then guided to the directly connected ICE housing.

The integrated module control of the RE merges the functionality of the power electronics, the GS controller, the ECU, and the hybrid logics in a common device. Expensive and sensitive sensors such as the ICE shaft position determination can be avoided by the fixed coupling of the ICE and the GS. Figure 17 shows the control system layout for the RE.

5 RE VEHICLE INTEGRATION

Targets for the integration of an RE system into the vehicle are:

- No restriction to passenger compartment and trunk
- Package flexibility with alternative drivetrain versions (pure BEV, conventional drivetrain, etc.)
- Utilization of stiff chassis areas to support a low acoustic chassis excitation by the RE system
- Reduction of auxiliary components such as mounts, HV cables, connectors, cooling lines, and so on
- Low impact on vehicle chassis design to avoid expensive additional tooling
- Integration of system assembly process with assembly of drivetrain alternatives
- Vehicle crash requirements
- Vehicle system safety and electromagnetic radiation (EMR) requirements
- Axle load distribution requirements.

The compact box dimensions of the RE system support the system integration even with compact cars. Figure 18 shows two alternative integration concepts. The definition of the best RE integration depends on the vehicle architecture and the intended strategy:

System integration in separate and independent RE module for flexible vehicle packaging:

- flexibility BEV versus REEV
- vehicle package flexibility.

System integration in engine compartment together with electric traction motor system:

- assembly similar to conventional vehicle
- axle load distribution rather similar to conventional drivetrain system

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

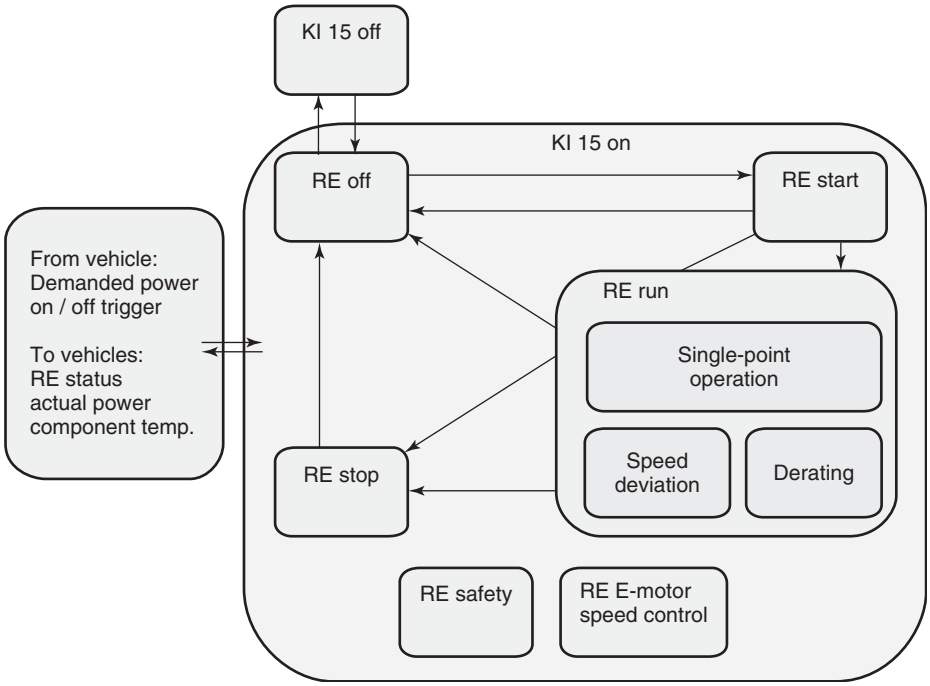


Figure 17. Concept of system control structure.

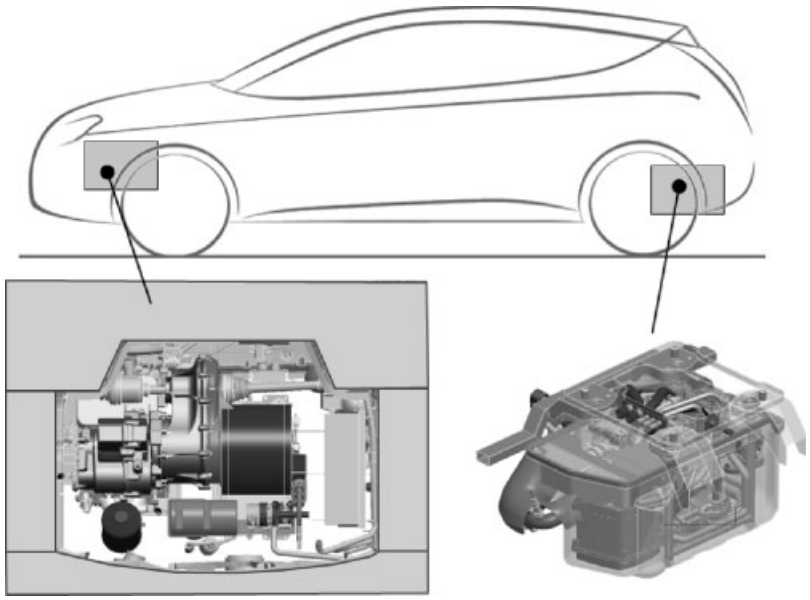


Figure 18. Concepts for series or serial hybrid RE system integration.

- joint auxiliary systems of RE and electric traction motor:
  - power electronic module for generator and traction motor system
  - joint liquid cooling system
  - joint acoustic encapsulation
  - joint mounting system

Figure 19 shows a design example for the integration of the RE as a separate and independent RE module that

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

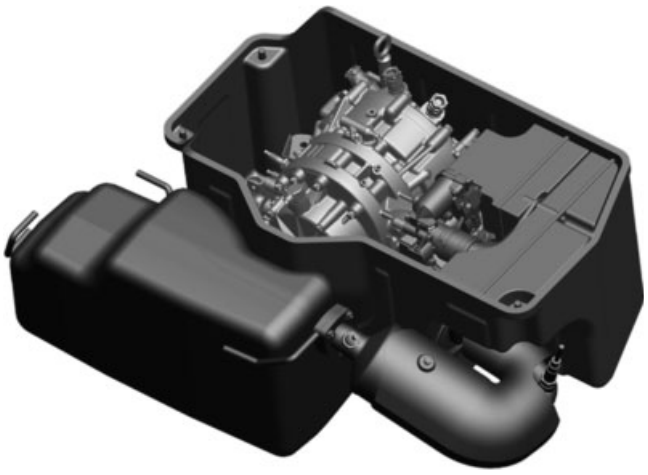


Figure 19. Design example for RE module.

is intended for the assembly underneath the vehicle trunk. This integration includes:

- RE core module (ICE, GS, mounting frame, and engine-mounted auxiliaries)
- Intake system
- ICE and GS control unit
- GS inverter system
- Connectors for the RE/vehicle interface (HV cables, cooling, 12 V supply, CAN, system mounts, fuel line, and fuel evaporation control system)
- Acoustic sealed RE encapsulation
- Exhaust system with three-way catalyst.

The intake system, the exhaust system, and the damping elements of the module mounts are optimized by detailed acoustic simulation to accomplish an averaged outside 1-m sound pressure of 65 dBA at the back of the vehicle and an interior level of 58 dBA at the codrivers ear. For the early development of the acoustic damping of the RE enclosure, the interior of the box is excited via loudspeaker using a measured sound sample of the ICE. The transfer functions of the relevant components have to be optimized.

Encapsulation of RE module:

- RPE or reciprocating piston ICE with low mechanical excitation levels
- Sealed RE encapsulation as acoustic shield to reduce airborne sound radiation
- Elastic mounting of RE module in vehicle.

Acoustic damping of intake system:

- Three resonators tuned for first three orders at the low noise operation point
- Intake system integrated in RE module
- Decoupled mounting of intake system within RE module.

Acoustic damping exhaust system:

- Integration of reflection damper
- Four resonators (tuned for first four orders at the low noise operation point)
- Absorption damper (for high frequency range).

Acoustic optimization of vehicle integration:

- Additional absorption and damping elements for weak vehicle structures
- Sealed cover of RE module lid
- Isolated mounting of auxiliaries (cooling pump, fan, etc.).

6 REAL-LIFE TESTING OF THE RANGE EXTENDER

For real-life testing, an RE system with RPE and 15 kW electric power at 5000 rpm was integrated in a demonstrator vehicle that has been built on the BMW Mini basis. This electric vehicle is designed as an MCV and has been equipped with a 12-kWh Li-ion HV-battery, which allows a battery electric driving range of 50 km city driving. The described 15 kW RE module and a fuel tank of about 12 l guarantees independency for an additional range of at least 200 km. Propelled by a permanent magnet synchronous motor with 75 kW peak power, the vehicle accelerates from 0 to 100 km/h in about 12 s (0–60 km/h in 5.4 s) and accomplishes a continuous maximum speed of 100 km/h with a peak vehicle speed exceeding 130 km/h.

6.1 Acoustic properties

Besides the absolute acoustic sound pressure by the RE, a good customer acceptance also requires the consideration of psychoacoustic effects. Subjective feedback from test-drives indicates that a continuous single-point operation of the RE system at specific transient vehicle situations such as deceleration can irritate the driver. This is also indicated by the interior sound analysis (Figure 20). With a strict continuation of the 15 kW@5000 rpm load point below vehicle speeds of 50–60 km/h, the RE becomes the dominant issues for acoustic convenience. The subjective noise perception can be improved by some degree of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

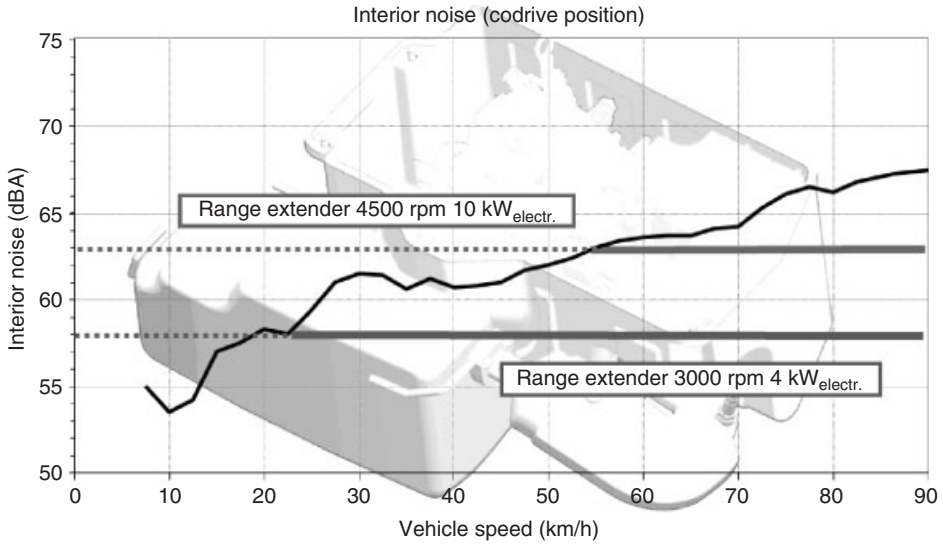


Figure 20. Acoustic analysis in the passenger compartment for pure battery electric driving with different RE load points.

tracking the vehicle speed with the RE speed or by the definition of different RE load points according to vehicle speed and energy requirement of the vehicle. For high speed driving, the general vehicle noise level allows an additional high power load point of the RE by increasing engine speed and mean effective pressure to avoid HV-battery depletion, even at highway driving speeds. Relevant RE load points for an MCV could be:

- Stop-and-go traffic; 4–5 kW @ 3000 rpm
- City driving; 10 kW @ 4000 rpm
- Extra-urban driving; 15 kW @ 5000 rpm
- Highway driving; 25 kW @ 6500 rpm.



Figure 21. Vehicle cold-start calibration

6.2 Vehicle cold start and defrosting

Known issues of BEVs at very cold ambient conditions can also be improved by adapted RE operation strategies to provide the required electric power to operate the vehicle and heat for passenger cabin and HV-battery heating. Because of the reduced charging ability of the HV-battery at low system temperature, the RE is operated at a low load operation point. The fuel consumption is higher than for best-point operation of the RE, but the acoustic properties are acceptable even at vehicle standstill. As the RE operation strategy allows to operate the RE at vehicle standstill only for extreme situations, the fuel consumption disadvantage can be accepted (Figure 21).

6.3 Energy management and fuel consumption

The differentiated 55-km AVL test-drive cycle with city, extra-urban, and highway driving has been used to compare the REEV technology with conventional drivetrains. Uphill gradients with an altitude difference of about 150 m for the highway part and 200 m for the extra-urban part have also been considered.

To evaluate the REEV concept, a limited HV-battery capacity of 7 kWh usable energy and a reduced RE output of 10 kW electric power have been considered. With these limitations and at winter time conditions, the test cycle and especially its uphill gradients require the development

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

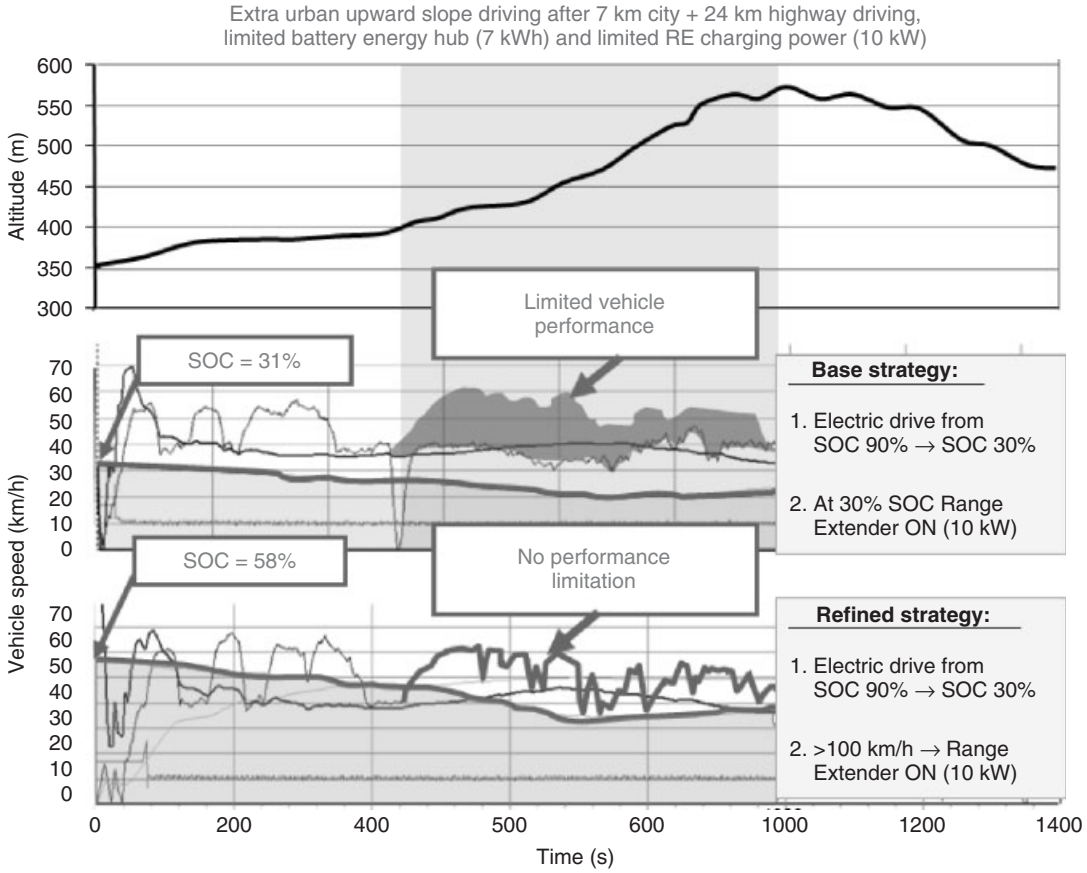


Figure 22. Analysis of the minimum required RE output and HV-battery capacity in dependence of the RE operation strategy.

of an effective RE operation strategy. For a simple HV-battery SOC-driven RE operation strategy, the available limited 10 kW RE output would not be sufficient to manage the uphill gradients with standard traffic speeds when the battery is depleted (Figure 22).

An advanced RE total-range-oriented operation strategy that also considers vehicle speed and energy consumption would activate the RE much earlier (in this case, during highway driving) and avoids noticeable power restrictions for the driver even with the predefined RE output and HV-battery capacity limitations. Further refinement of intelligent operation strategies is one of the key engineering challenges and could also consider GPS (global positioning system) data to support customer-accepted REEV concepts with minimum battery capacity and RE output.

For the direct comparison with conventional drivetrain concepts, an identical BMW Mini with 1.4 l NA engine with fully variable valvetrain, start–stop strategy, and intelligent battery management to recuperate brake energy has been defined. The REEV concept results in an estimated add-on vehicle weight of about 150 kg. The prototype status of the tested car did not allow full weight optimization and resulted in a prototype add-on weight of about 300 kg in comparison to the conventional reference car. To compare the CO_2 emissions of both concepts, the electric energy of the plug-in charging has been considered according to the German power-plant mix with 625 grCO_2/kWh.

For cold start at ambient temperatures of 0°C and city driving conditions, the REEV shows significant CO_2-emission advantages even with battery electric cabin heating (Figure 23). In contrast to the conventional vehicle, a high proportion of the electric energy consumption of the REEV is due to the cabin heating. This part could be reduced by utilizing the RE waste heat. Advantage of the REEV is its ability for recuperating the brake energy. For city driving, 20–25% can be achieved.

If the test vehicles are preconditioned at 20°C and without cabin heating, the CO_2-emission advantage is even higher (Figure 24). Even for an assumed empty HV-battery condition and a continuously running RE (= "miss-use"), the CO_2 emissions are on a similar level for both concepts.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

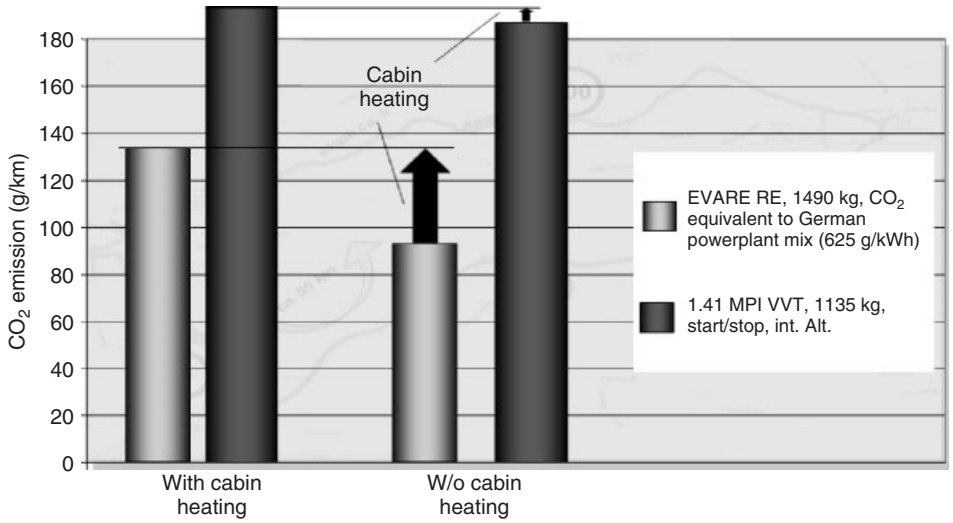


Figure 23. City test-drive at low ambient temperature—cold start.

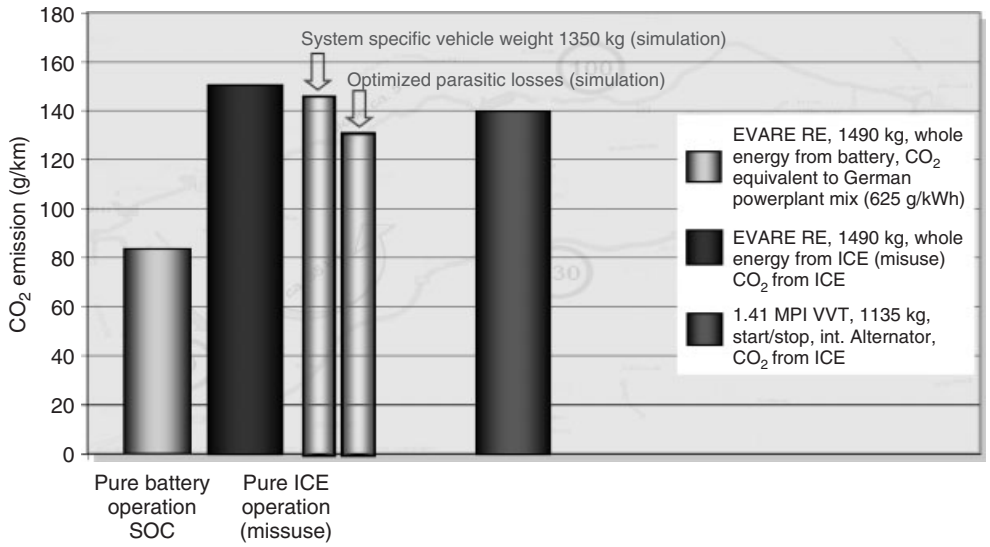


Figure 24. City test-drive with 20°C preconditioned vehicles—hot start.

The effect of further improvement by reducing the vehicle weight and parasitic losses has been simulated.

These circumstances are different in the extra-urban and highway sections of the AVL test cycle. Owing to the higher average load level according to the required traction motor performances (Figure 25), the ICE typical low load fuel consumption disadvantage is of less relevance. These sections of the test cycle also have less potential for recuperation of brake energy. These effects result in a reduced CO_2-emission advantage for the battery electric operation. Correspondingly, the CO_2 emission of the REEV with all energy generated by the RE (=miss-use) is above the emission level of the conventional vehicle. The good SFC property of the conventional vehicle, efficiency disadvantages by the double energy conversion of the series or serial hybrid architecture, and the high REEV weight are the dominant effects (Figure 26).

7 SUMMARY AND CONCLUSION

In summary, the test-drive results show that pure series or serial hybrid REEV architecture can be a cost-efficient concept for MCV applications that focus on the battery

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

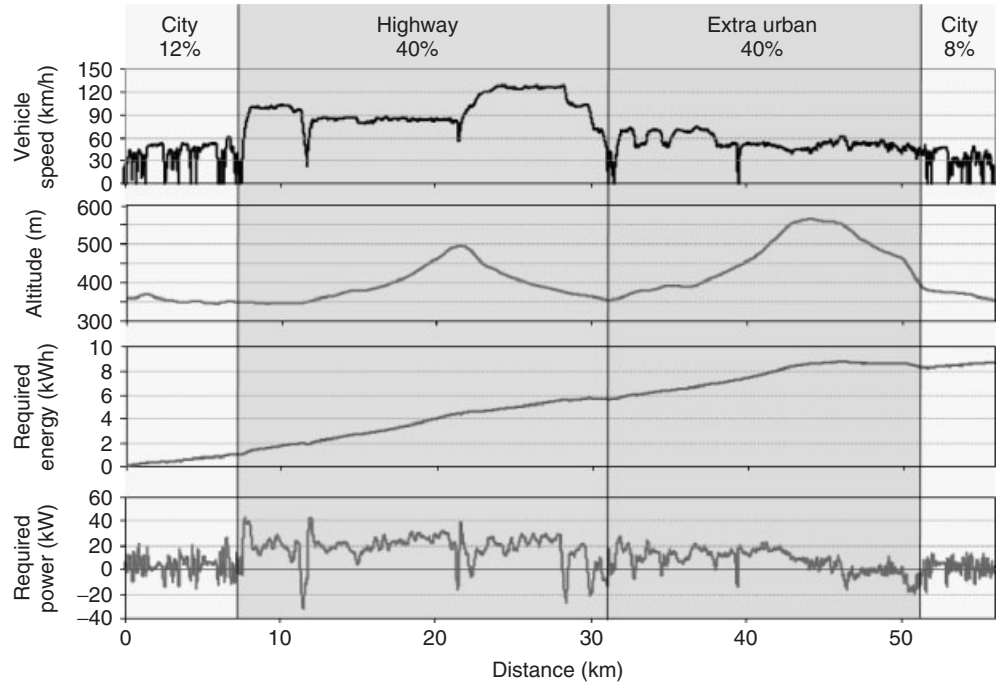


Figure 25. AVL standard test-drive cycle: profiles of speed, altitude, and power requirement.

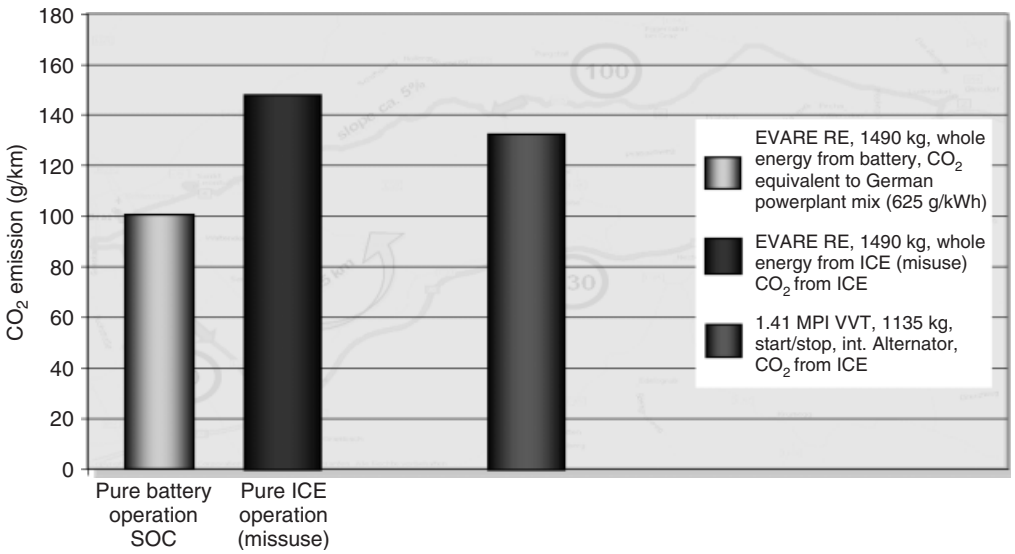


Figure 26. Extra-urban test-drive

electric driving range. For vehicle concepts with more focus on extra-urban and highway driving, an RE concept with switchable direct drive should be preferred to reduce efficiency losses.

Of significant influence on the overall REEV efficiency are also the parasitic losses by auxiliary component efficiencies and the weak thermal isolation of today's vehicle and battery concepts. Improvements and the utilization of synergies such as the reuse of waste heat of the electric traction system are cost-efficient and directly improve the AER.

A further significant contribution to the success of future REEVs is less connected to the hardware but is given by an intelligent system vehicle operation strategy. By

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a predictive and driver-demand-oriented operation of the RE, the requirements for the HV-battery capacity and the RE output performance can be reduced for maximum cost efficiency and system weight efficiency. This requires a precise prediction of the energy requirement for the intended route and direct information to the driver about remaining vehicle range and the effect of the recent driving habits. Beside GPS and ambient temperature data, the energy consumption and vehicle range prediction should also consider "learned" data about the driver's preferences and usual routes.

With a well-developed combination of RE, HV-battery, and operation strategy, the REEV approach can be the enabler for affordable and locally emission-free electric mobility.

RELATED ARTICLES

Overview of Electric, Hybrid and Fuel Cell Vehicles
EV Powertrain Configurations
Series Hybrid Electric Vehicles (SHEVs)
Series-Parallel Hybrid Electric Vehicles
Range extender EV
Energy Management Systems of EVs
Drive Train Noise, Vibration and Harshness
EV auxilaries
Basic Consideration
Power and Energy Requirements for Electric and Hybrid Vehicles

REFERENCES

Atzwanger, M., Hubmann, C., Kometter, B., *et al*. (2010) Two-cylinder gasoline engine concept for highly integrated range extender and hybrid powertrain applications. SAE 2010-32-0130, JSAE 20109130.

Beste, F., Fischer, R., Ellinger, R., and Pels, T. (2009) *The pure range extender as enabler for electric vehicles*. 21st International Conference Engine & Environment, September 10–11, 2009.

Fischer, R. (2009) *The electrification of the powertrain—from turbohybrid to range extender*. 30th Vienna Engine Symposium, May 7–8, 2009, Volume 2, pp. 1–23.

Fraidl, G.K., Ebner, P., Geiger, U., *et al*. (2009) *Impact of electrification on the internal combustion engine*. 21st International Conference Engine & Environment, September 10–11, 2009.

Grebe U.D. (2009) *GM's Voltec Antriebssystem- Elektrifizierung der Fahrzeuge auf neuem Niveau*. 21st International Conference Engine & Environment, September 10–11, 2009.

IVT (2004) Analyse von Änderungen des Mobilitätsverhaltens – insbesondere der PKW-Fahrleistung – als Reaktion auf geänderte Kraft-stoffpreise. im Auftrag des Bundesminis-teriums für Verkehr, Bau- und Wohnungswesen, Bonn.

List, H. (2009) *Future powertrains in a fast evolving global environment*. 30th Vienna Engine Symposium, May 7–8, 2009, Volume 1, Supplement.

Najork, R., Steinberg, I., Leibbrant, M., and Strube, A. (2009) *BOOSTED RANGE EXTENDER—GETRAG's concept for a highly flexible electric powertrain with ultra low CO_2 emission*. 21st International Conference Engine & Environment, September 10–11, 2009.

Steiger, W. (2009) *twinDrive®—concept for electrification of powertrains*. 21st International Conference Engine & Environment, September 10–11, 2009.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto047
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Batteries Indication and Management

Bernhard Kortschak
AVL List GmbH, Graz, Austria

1 Introduction 1
2 Hardware Architecture 3
3 Software Architecture 9
4 Functional Safety 13
5 Conclusion 14
References 14
Further Reading 15

1 INTRODUCTION

The traction battery for a full electric vehicle (EV) is the primary energy storage device, and as such, it is one of the key components of the possible future propulsion systems. Even for hybrid electric vehicles (HEVs) or plug-in EVs, the battery plays a key role in the improvement of the fuel economy. The precise indication and management is mandatory for the fulfillment of its application-specific function and purpose. In addition, only proper management ensures that the expected lifetime can be achieved as inappropriate use can lead to significant battery degradation or even to safety critical conditions.

The following subsections introduce the main indicators for the state of the battery and the management functionalities of the battery. In addition, the hardware and software system architecture is presented.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

1.1 Battery indication

The main battery characteristics, which are used in the context of electrification, are the state-of-charge (SoC), the state-of-health (SoH), and the state-of-function (SoF), and these states are explained briefly in the following subsections. Battery indication, in a wider sense, also includes the monitoring of the components in the battery pack and of the safety relevant subsystems, such as interlock circuit or isolation monitoring.

1.1.1 State-of-charge

The SoC is defined as the percentage of the maximum possible charge that is present in the battery.

This value is one of the main measures of the state of the battery and it is used for the control of the energy distribution in an HEV or full EV. Its precise determination was one of the key research topics in the past and there are still some open aspects that need to be addressed when it comes to a reliable and robust estimation under the boundary condition of a traction battery.

1.1.2 State-of-health

The SoH is a measure of the condition of the battery compared to a fresh battery.

This measure reflects the state of the battery in terms of capacity loss or in terms of increase in resistance over the battery lifetime. This value shows how much of the lifetime has been spent or is still remaining and indicates if the battery must be replaced by a new one. This measure can be expressed either in terms of years or as a percentage of

battery degradation. Actually, this indicator and its precise definition are depending on the actual type of application because the end-of-life (EoL) criterion of the battery can be defined only in the context of its application.

1.1.3 State-of-function

The SoF consists of measures for the ability to fulfill the application-specific function of the battery.

Depending on the application, the battery must fulfill a specific functionality. The measures of fulfillment can be summarized as SoF. Again, these measures are application specific but examples of common indicators are as follows: peak power capability (e.g., 2, 10, and 30 s charging/discharging power and current limits), cold cranking performance, and remaining energy content of the battery (measured in terms of watthours). These measures are often a prediction of future function fulfillment under predefined boundary conditions. For instance, the peak power capability is a prediction of the battery performance under the assumption of a constant battery load for the next few seconds.

1.2 Battery management

The term *battery management* summarizes all control relevant aspects. This chapter focuses on the ability of directly influencing the battery operation by the battery main controller. On the basis of the indicators defined earlier, the battery controller has an indirect influence on the battery operation, for example, in which SoC window the battery is operated. However, the decision on the actual energy distribution is commonly carried out by the main powertrain controller (Section 2), which is responsible for controlling the propulsion system (its discussion is out of the scope of this chapter, see Energy Management Systems of EVs and Energy Management Systems of HEVs for details). The battery main controller can directly influence the following three actuators.

1.2.1 Activating the charge equalization circuit

The charge equalization circuit is responsible for an equalized and balanced state of the individual cells in a battery pack. This kind of circuit is usually present in lithium-ion-based battery packs and packs consisting of electrical double-layer capacitors (EDLCs). For these types of packs, the SoC of the cells will drift apart over time. The equalization of the cell states must be managed by an external electric circuit and its activation is usually triggered by an external command. The equalization can be realized by two different hardware concepts, namely the active balancing or passive balancing systems.

1.2.2 Controlling the contactors

Two contactors are used for a physical disconnection of the battery pack during the time when the vehicle is parked or under service. In addition, the disconnection can be triggered by the battery controller to avoid safety critical situations for the battery. This emergency switch off must be the last measure for the battery main controller to influence the operation of the battery pack as this heavily influences the operation of the vehicle. Before the contactors are closed, the direct current (DC) link capacitor must be charged up to the battery voltage. This is done by a precharge unit activated by an additional contactor.

1.2.3 Thermal management

Depending on the system architecture, the battery enclosure contains the actuators for the cooling and heating of the battery. As such, the battery main controller can activate a fan for air cooling or a pump for liquid cooling, respectively.

1.3 System architecture of the battery management system

The battery management system (BMS) is responsible for carrying out the functions for the battery indication and management as defined earlier. This system consists of hardware and software and can be split into several layers as visualized in Figure 1. The battery controller is placed on top of the BMS hardware. Commonly used terms for this controller are battery control unit (BCU) and battery module/management controller (BMC). For a distributed system, a subcontroller is present that is responsible for the measurement and charge equalization of a subset of cells (modules). This subcontroller is often called *module control unit* (MCU), *cell supervising circuit* (CSC), or *cell supervising electronics* (CSE). These subcontrollers are usually mounted close to the cells and are used to measure the cell voltages and the temperatures and incorporate safety functions such as over- and undervoltage protections of cells, and they include the charge equalization circuit. Other possible hardware architectures are discussed in Section 2.2.

From a software point of view (see the upper part of Figure 1), the BCU software can be split into two main

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

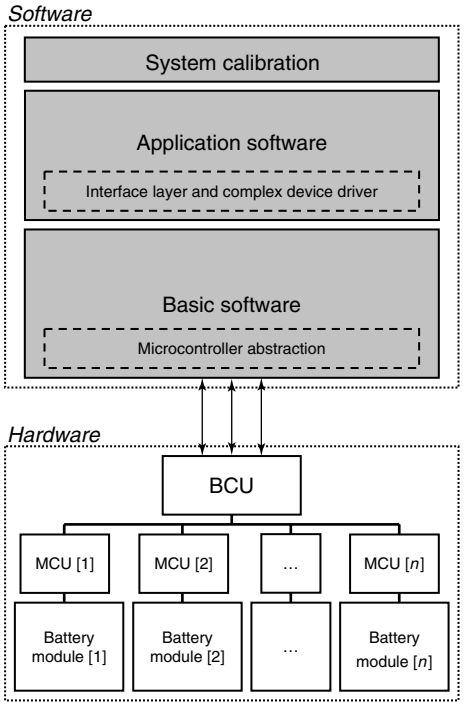


Figure 1. Battery management system—system architecture. (Reproduced by permission of AVL List GmbH.)

layers. The lower layer consists of the basic software that includes the real-time operating system, the low level device drivers for the direct inputs and outputs, and the communication stack, for example, for the CAN (controller area network) bus. Within this layer, an abstraction of the hardware is performed so that the upper layer can be designed independently from the hardware. The upper software layer contains the application software. This part of the software is responsible for the implementation of the high level functions, such as the estimation of the SoC. BMSs according to this structure are highly modular and offer the flexibility that these systems can be applied to different applications. The application software, for instance, can be easily extended by customer-specific code. At last, the software must be calibrated for the cells used in the battery and for the specified vehicle.

The total system costs can be reduced if functions of the application software are embedded into the powertrain controller (see Section 2.2.4 for details). As a result, the size and the complexity of the controller on the battery side can be reduced. In addition, the calibration is easier if all application-specific functions are placed together in a single control unit.

2 HARDWARE ARCHITECTURE

The BCU, as the main controller of the battery, has two major interfaces (see Hybrid Systems and High Voltage Components). Firstly, the BCU communicates with the vehicle control architecture. A CAN interface is usually offered by the BCU and data is exchanged with the powertrain control unit as shown in Figure 2. In the case of an HEV, this controller is often referred to hybrid control unit (HCU). Secondly, the BCU communicates with subsystems inside the battery or is directly connected to battery components (Figure 3).

This section concetrates on the interface to the components inside the battery enclosure. Section 2.1 presents these components followed by a subsection, which describes the possible architectures of the control system. Section 2.3 compares the possible hardware solutions for the charge equalization circuit.

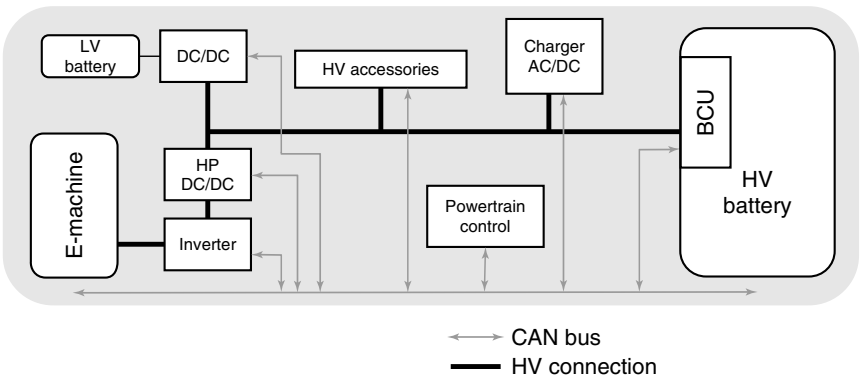


Figure 2. Interface between battery control unit and other control units of the powertrain. (Reproduced by permission of AVL List GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

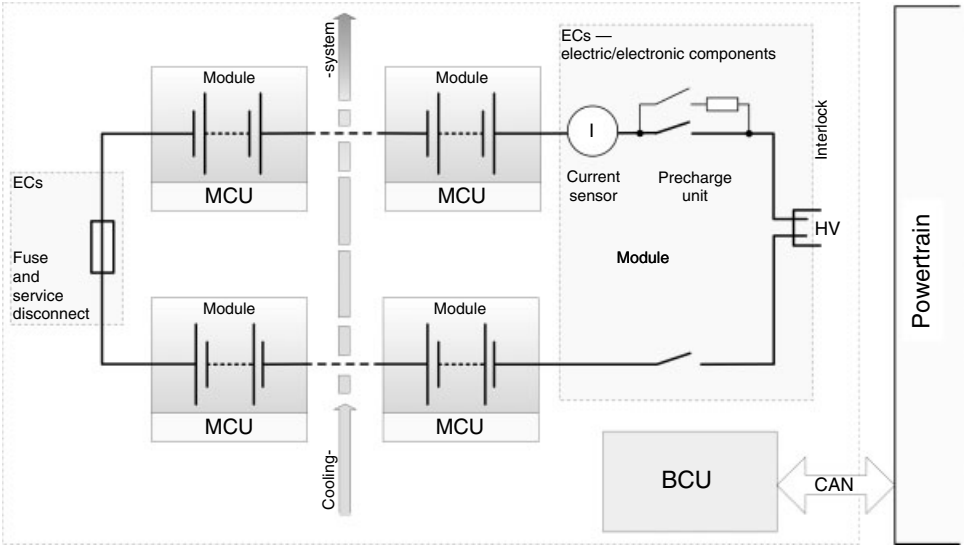


Figure 3. Battery control unit and the battery interior components and subsystems. (Reproduced by permission of AVL List GmbH.)

2.1 Components of battery management systems

The following components are part of the BMS (see, e.g., Pichler *et al*., 2007). In general, they are placed inside the enclosure of the battery pack (Figure 3).

Some components are part of the power system and must carry the main battery current. In this case, the components must be designed to withstand the continuous and peak currents. For a passenger HEV or EV, a typical value for the continuous battery current is 100 A, whereas the peak current can reach values of 250–350 A (both values depend on the specific application).

2.1.1 Current sensor

The pack is equipped with either one or two current sensors. They have to measure the pack current in a wide measurement range. In addition, the integration of the current must be accurate over a long period of time, because an offset would directly influence the accuracy of the SoC calculation. For this reason, the current sensors for traction batteries often offer two or more measurement ranges. Typical sensors are based on a shunt and provide very precise measurements. Other types of sensors, such as hall-effect sensors, can be used as backup devices.

2.1.2 Main contactors

The main contactors are used for a physical disconnection of the battery pack and therefore provide a galvanic isolation. In addition, they offer the possibility to force a shutdown of the battery for emergency reasons. It is important to note that this emergency shutdown under load is not the main function of these components and, as such, they can usually only withstand a few switching cycles under these conditions. On the other hand, the contactors for traction batteries must be small and light weight compared to their counterparts in other applications. Protection measures must be included so that these contactors can safely disconnect the battery from the system under all operating conditions (e.g., for very high peak currents). In order to ensure a precise diagnosis, the voltage across these contactors can be measured so that an increase in the resistance can be detected. Another effect, which should be monitored, is the decrease in the insulation resistance over time because of evaporated metal inside the contactor.

2.1.3 Precharge unit

A precharge unit must be activated before the main contactors can be closed. It consists of a precharge relay and an appropriate power resistor. The purpose of the precharge unit is the charging of the DC link capacitor of the electrical system and to reduce the inrush current. The main contactors can be safely closed if the voltage difference between battery and DC link capacitor is below a predefined threshold. This procedure ensures that the lifetime of the main contactors can be achieved.

2.1.4 Fuse

The main contactors can be used for an emergency power down of the system. These contactors can provide this

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

functionality only up to a certain current limit (e.g., a few 100 A). Above this limit, the contactors must not be opened as a disconnection can lead to a destruction of these components. As such, additional measures must be applied and a fuse can be used for currents, for example, during a short circuit. The short-circuit current of a lithium-ion battery can reach values of 3000–7000 A and it is of importance that the contactors can withstand this overcurrent situation for some time because the fuse needs a specific amount of time unless it can break the connection.

2.1.5 Service disconnect

The battery can be equipped with a service disconnect. This switch can be removed for vehicles under maintenance in order to break the electrical connection in the middle of the battery pack. Thus, only two strings of cells remain with half the system voltage.

2.1.6 Cell bypass device

Currently, bypass devices on the cell or module level are under research. A cell with a failure can be shorted by this device. This concept makes sense under the assumption that common cell failures will lead to a cell internal disconnection (an open circuit). Under this condition, the pack can still be used with only limited effects on the overall performance of the battery. High costs are currently one of the challenging aspects of this technology.

2.1.7 Isolation monitoring system

The battery terminals and the cells will be electrically isolated to the vehicle conductive structure because of safety requirements for voltages above 60 V DC. The isolation monitoring system is part of the overall safety system and it will prevent the usage of the battery in case of isolation failures. The monitoring system periodically measures the insulation resistance (leakage resistance). In typical applications, the measurement range for this resistance is between several 10 kΩ up to several 1 MΩ. Another term, which is commonly used for this system, is *insulation resistance monitoring system*.

2.1.8 Connector and interlock circuit

The connectors from the battery to the other electrical components can include additional signal wires so that an interlock system can be implemented. These signal wires are used for monitoring the correct connection and wiring between components. The battery is switched off as soon as a failure is detected, for instance in case of a disconnection of the battery or in case of a failure in the power cables. The electronic circuit can be part of the BMS or it can be placed in other parts of the electrical system.

2.1.9 Thermal subsystem

For cooling and heating, the battery can be equipped with a thermal system for air or liquid cooling. For air cooling, the BMS can often control a fan; however, for liquid cooling, the system is often connected to the air-conditioning system. In this case, the activation of the pump is out of the scope of the battery controller. In each case, the battery is equipped with a number of temperature sensors, such as a temperature sensor for the cells, for the inlet/outlet temperatures of the cooling system, and for temperature measurements on the printed circuit board (PCB). In addition, the battery can be equipped with a PTC (positive thermal coefficient) heater or similar heating devices for cold temperatures.

2.1.10 Subcontrollers on module level

Depending on the controller architecture, the BMS can consist of the main controller and subcontrollers, which are responsible for the monitoring and controlling a subset of cells. The different possibilities are discussed in Section 2.2.

2.2 Centralized versus distributed controller architecture

In traction batteries consisting of cells based on lithium-ion technology or EDLCs, all cells must be monitored individually and the charge balance between these cells must be managed by an external circuit. For this reason, all cells are connected to the BMS for cell voltage measurements and for the purpose of charge equalization. In addition to the cell connection, a large number of temperature sensors, for example, one sensor for each cell, are located in the pack.

In order to reach the required voltage level, the number of cells in series connection can exceed 100 pieces for an HEV, plug-in HEV, or EV application. In case of a centralized system, all wires are connected to a single PCB, which introduces a high complexity for the wiring harness. On the other hand, the wiring effort can be reduced in a distributed system of subcontrollers, where each subcontroller is responsible for a subset of cell voltage and cell temperature sensors. Different degrees of centralized and distributed systems are visualized in Figure 4, which are explained in detail in the following subsections.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

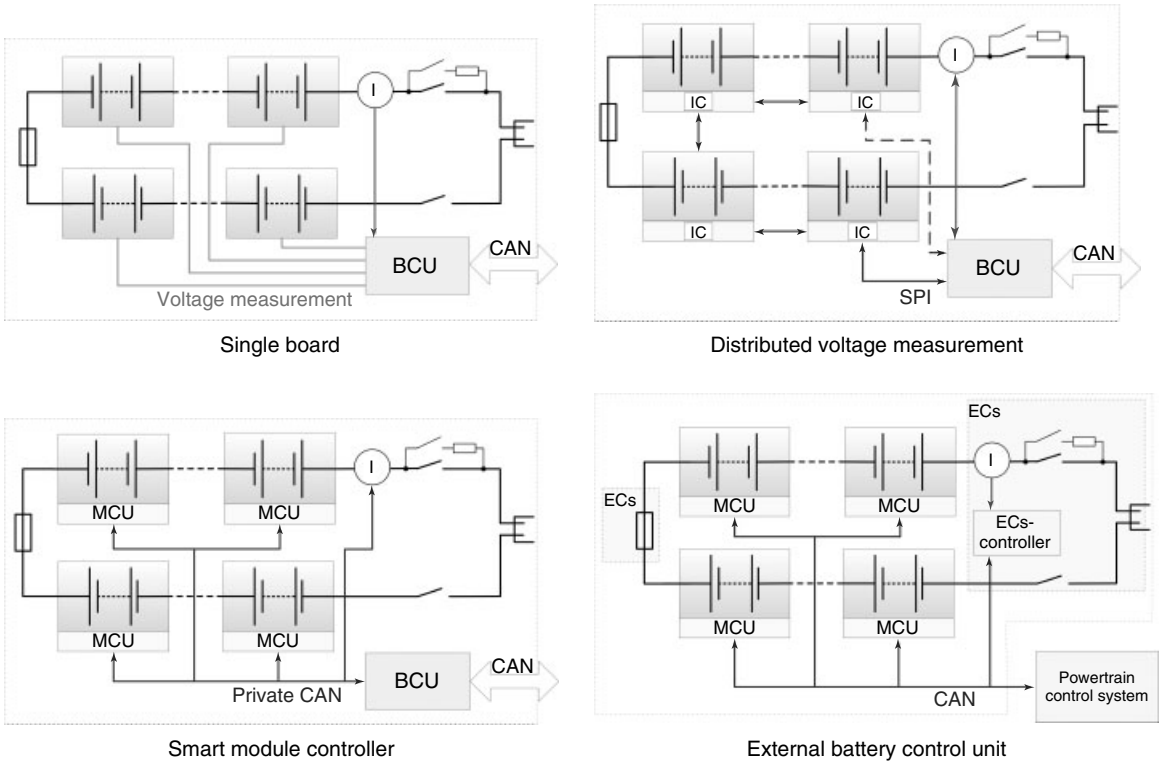


Figure 4. Different complexities of the controller architecture. (Reproduced by permission of AVL List GmbH.)

In all cases, specialized integrated circuits (ICs) can be utilized for the purpose of monitoring the cell voltages and for the charge equalization. The use of ICs for cell monitoring significantly reduces the total system costs and increases the reliability as the number of components is decreased. Several manufacturers offer possible solutions for the monitoring of traction batteries. A bus system is required for the information transfer between the different PCBs for a distributed system. The simple SPI (serial peripheral interface) communication is not appropriate for long wires in a harsh environment. Despite proprietary communication systems, CAN offers a robust communication between the different boards with the drawback of higher costs.

2.2.1 *Single board*

All relevant electronics are placed on a single board. All wires from the cells and temperature sensors must be routed to this board. This solution is attractive for small and highly integrated battery packs and the onboard communication between the BCU and the cell monitoring ICs can be solved by SPI. However, the wiring harness is getting more and more complex as the size of the battery pack increases with the number of cells. In addition, measurements over long cables are prone to disturbances and active balancing systems over long cables can introduce problems with respect to electromagnetic interference. The board itself must be designed for voltages up to several 100 V, and the computational power of the BCU must be sufficient to process all input data.

2.2.2 *Distributed voltage measurement*

For larger packs, the cell monitoring ICs can be placed locally close to the modules and this reduces the number of connections to the main PCB. These modules can consist of 6–16 cells in series connection and some space must be reserved on these modules for the additional electronics. A smaller number of cells per module and a reduced voltage simplify the handling of these modules, if they are built as physically separated units.

In this solution, the module electronic is responsible for data acquisition and execution of the charge equalization. The BCU contains all required software and processes and manages all information. The development of a robust communication system is one of the major issues. Proprietary solutions exist, which connect the individual module controllers with the BCU by a daisy-chain-like connection. An SPI similar interface is used, but it has to be extended to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

be able to provide the required galvanic isolation between the different modules. The communication system must support a high bandwidth as all voltages and temperatures are to be transmitted over this line and so this system is sensible to disturbances. Despite this issue, this architecture offers a very flexible and cheap solution suitable for all different applications.

2.2.3 *Smart module controllers*

This solution consists of a master/slave controller architecture. Small microcontrollers are placed on the module boards, which are communicating with the BCU by CAN bus. The utilization of the CAN bus offers a robust solution to the bandwidth problem. The controller on the module side can be kept small, for example, an 8-bit controller can be utilized in order to minimize the additional costs. As such, the BCU is still responsible for all data processing, but the controller on the module level can take over some small tasks, for example, the protection of the cells against over- and undervoltage. These controllers can also monitor the main controller and a high level monitoring concept can be realized without the utilization of dual-core processors for the BCU. For this monitoring purpose, an additional hardwired line is required so that the module controllers are able to force the contactors to open in a case of a serious system failure.

Even more software functionalities can be moved from the BCU to the module controllers, but in this case more computation power is required, for example, 16-bit microcontrollers can be used. For this option, high level functions such as the estimation of the SoC are performed on a module level. In addition, self-diagnostic of the modules is possible and, from a software point of view, the modules can be exchanged in service easily. The BCU can be equipped with a smaller and cheaper processor but this hardly pays of the additional costs on the module side. A smart solution is also required for software updates and for calibration so that it is not required to flash and calibrate each module controller individually.

2.2.4 *External battery control unit*

The BCU in the options above is responsible for the direct control of the battery electric. Thus, the PCB consists of two areas, a lower voltage part (microcontroller and related components) and a higher voltage part (battery voltage measurement and isolation monitoring). Both areas must be separated by an isolation barrier. The BCU is therefore a battery-specific controller board.

On the other hand, it is possible to design a small controller, which is responsible for the battery-related functions, that is, this controller is directly connected to the electric and electronic battery components. The BCU itself is reduced to a standard controller, which communicates with the battery by CAN bus and which can be placed outside of the battery housing. A dedicated BCU is not even required, if enough processor resources are available on other controllers in the vehicle, for example, in the main powertrain controller.

2.3 Passive and active balancing systems

The SoC levels of the cells within a battery pack will drift apart over time. The main reasons for this drift are as follows: different self-discharge rates, different cell capacity values, and different Coulomb efficiencies. The self-discharge rate is relevant during standstill periods, whereas the two other effects apply during the usage of the battery. In addition, the cells age at different rates because of the individual initial conditions of the cells, for example, impurities in the cell production, or because of different operating conditions during cell usage, for example, temperature gradients inside the battery. This leads to a further increasing spread in the cell characteristics over lifetime.

A balancing system consisting of a charge equalization circuit can offset these differences unless to a certain point, where the spread is significant and the balancing current is not sufficient anymore. The resulting SoC differences will lead to limitations during the usage of the battery pack. For instance, during charging, the cell with the highest SoC will reach the upper voltage limitation before all other cells are completely charged up. A similar situation will occur during discharging and both effects together will reduce the available energy content of the battery.

The charge equalization between the cells can be activated by the BMS. Despite this activation, different hardware concepts for the charge equalization circuit exist. An overview can be found in Daowd *et al*. (2011), which provides an overview of advantages and disadvantages of 18 possible hardware topologies. Principally, two different concepts are possible, namely the passive balancing system and the active balancing system. Passive systems are only capable to discharge individual cells, whereas the cells in active systems can be discharged and charged depending on the operating mode of the system.

Figure 5 shows the effect of this principle difference between the two concepts. In an unbalanced state, the cells show a different level of SoC. In this example, only three cells are displayed for the sake of simplification, whereas the first cell has the highest SoC compared to the third cell with the lowest SoC level. With a passive balancing

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

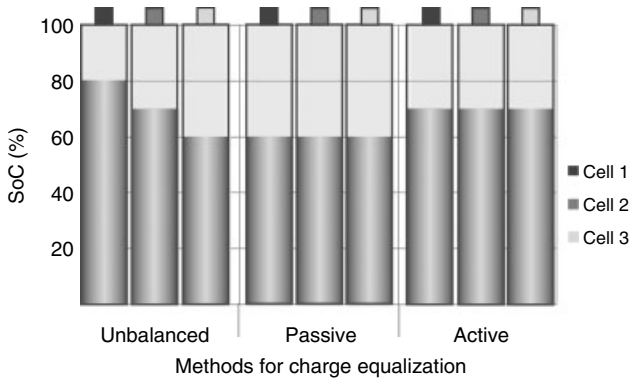


Figure 5. Impact of passive and active balancing methods on the cell charge level. (Reproduced by permission of AVL List GmbH.)

system, it is only possible to discharge cells. As such, the first and second cell must be discharged unless they reach the SoC level of the worst cell. In this example, the worst cell is the cell number 3. The power during discharging is dissipated and this reduces the overall efficiency of the system. With an active balancing system, it is possible to discharge cell number 1 and to charge cell number 3, such that all cells are matched to an average SoC level. The active balancing circuit will transfer the charge between the cells with a specific efficiency and the average SoC after balancing will be less than the average before balancing. Nevertheless, this system is more effective compared to passive balancing.

Balancing systems can be designed not only for currents of several 100 mA but also for currents up to several 1 A. Balancing boards for higher currents yield to complex electronic circuits compared to small balancing systems and therefore to higher cost. The optimal balancing current depends mainly on the capacity of the cells in use and of the expected differences in the cell-to-cell characteristics at EoL.

Passive balancing systems offer the opportunity to activate the balancing for each cell independently, whereas active systems are usually restricted to certain operating modes. In other words, on active balancing systems, it is usually not possible to apply the balancing to several cells of a module in question at the same time. In this case, balancing of several cells must be performed in consecutive steps. This must be considered for the sizing with respect to the maximum current of the balancing circuit. In addition, active balancing systems offer good efficiencies for transferring charge from one to another cell. However, owing to restriction in the operating modes, it can occur that the charge must be transferred over several cells as described earlier. In this case, the efficiencies of each balancing activity must be multiplied yielding significantly lower system efficiency under real operating conditions.

The different balancing concepts are briefly discussed as follows.

2.3.1 *Passive balancing—dissipative resistors*

For passive systems, shunts, which are placed in parallel to the cells, can be activated. Depending on the SoC level of the cells, these shunts are switched on or off. Several cells can be discharged at once, but the overall heat generated on the PCB must be dissipated properly.

2.3.2 *Active balancing—charge shuttling*

One possible concept of shuttling the charge between the cells is the concept of a *flying capacitor*. A capacitor can be consecutively switched in parallel to the cells (see Figure 6 for a possible schematic). Cells with higher SoC, and thus with a higher voltage level, will charge up the capacitor, whereas cells with lower voltage will be charged up by the capacitor. The switching pattern between the cells can be fixed in advance or an intelligent logic can pick out the cells that need to be balanced. High peak currents can occur during charging and discharging, if the voltage levels are significantly different. On the other hand, hardly any balancing takes place if the voltages of the cells are similar. This can lead to problems for technologies, where the SoC

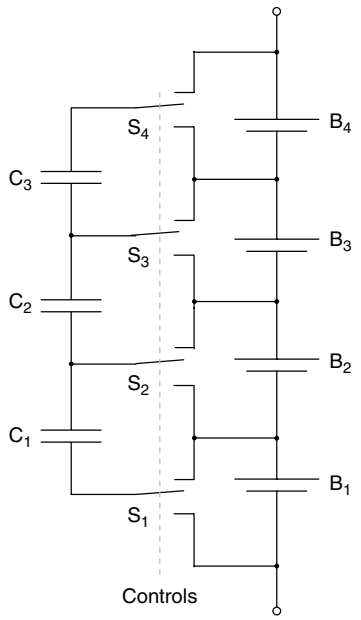


Figure 6. Sketch of a balancing concept with flying capacitors C_1–C_3 for four cells B_1–B_4. Note: Switches S_1–S_4 can stay in three states (upper/lower connection, respectively, and open). (Reproduced by permission of AVL List GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

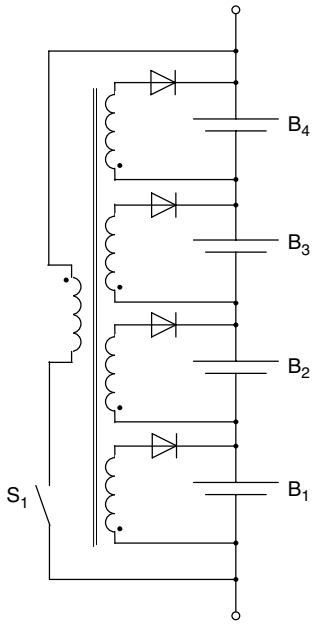


Figure 7. Sketch of balancing concept for four cells B_1-B_4 with a shared transformer with a switch S_1 on the primary side. (Reproduced by permission of AVL List GmbH.)

has only a small impact on the open circuit voltage (OCV) especially if the battery is used within a small SoC window like in an HEV. For EDLCs, this method can be applied successfully.

2.3.3 Active balancing—energy converters

Active balancing systems can be realized using inductors or transformers. Several concepts using transformers can be considered. For the *switched transformer* concept, a flyback converter is used consisting of a transformer where the primary side can be connected to the module or pack terminals, respectively, and the secondary side of the transformer can be connected to one of the single cells. The primary and secondary sides are switched on and off allowing an appropriate charge transfer between the cells. For the *shared transformer* concept, a similar principle is used. In this case, as many secondary windings are used as cells, which need to be balanced (Figure 7).

3 SOFTWARE ARCHITECTURE

The BCU is the main device for the management and evaluation of the battery measurements in order to provide the information about the state of the battery to the vehicle. In addition, it controls the battery contactors, the charge equalization, and fans and pumps for the thermal management. One of its major functions is the estimation of the SoC and the SoH and the prediction of the battery performance, that is, the SoF.

The main functions of the software can be summarized as follows (Pichler *et al*., 2007):

- state estimation of the battery modules (SoC and SoH);
- prediction of battery performance (SoF);
- control of the charge equalization;
- measurement signal processing (mainly current, voltage, and temperature);
- safety relevant monitoring (interlock circuit monitoring and isolation monitoring);
- fault detection and error management;
- communication with vehicle and module controller boards.

3.1 Software modules

The software of the controller can be split into different layers (Figure 8), whereas the first layer includes the basic software with the hardware-dependent functionalities, for instance the device drivers, the transfer layers for different communication systems such as LIN (local interconnect network) or CAN, the memory management, and the operating systems. On top of the basic software is the applications software with the high level functions. The high level functions are structured in software modules with defined inputs and outputs. These modules can be used in a very flexible way, for example, they can be exchanged by application-specific software modules, or they can be placed in different controllers if required. A possible structuring into software modules could be

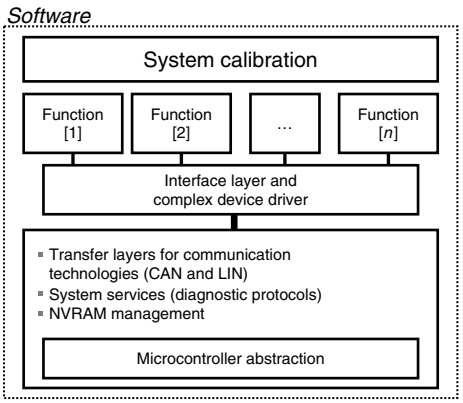


Figure 8. Basic software and application software in the BCU. (Reproduced by permission of AVL List GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

as follows (compare with Languang *et al*., 2013, with a slightly different structuring).

3.1.1 BCU control

This module consists of the main state machine of the BCU. The BCU state can switch between different operating modes, such as initialization, running, after-run, or shutdown.

3.1.2 Error management

A diagnostic function monitors all input and output ports of the BCU. An error management is responsible for proper reactions in case of any signal errors.

3.1.3 Safety monitoring

A safety concept can be required depending on the required ASIL (automotive safety integrity level) classification (Section 4). Additional monitoring levels exist, which are double checking the results of the other software modules.

3.1.4 Communication

The communication module is responsible for the CAN communication to the vehicle and, if required, to the subcontrollers in the battery.

3.1.5 Contactor control

This software module is responsible for the opening and closing of the main battery contactors. In addition, this function will control the precharge unit.

3.1.6 Battery protection

The battery must be operated within a safe operation area. This function prevents an inappropriate use of the battery with respect to current, voltage, power, temperature, or SoC limitations.

3.1.7 Electrical hazard protection

This module provides protection against electrical hazards and handles the isolation monitoring, the monitoring of interlock system, and service disconnect switch.

3.1.8 Battery state calculation

One of the main functions is the observation of the state of the battery and the calculation of the SoC, SoF, and SoH. A detailed explanation of this function is given in Section 3.2 to 3.5.

3.1.9 Balancing control

The charge equalization circuit is activated depending on the actual state of the cells. The balancing activity can be triggered based on the relative comparison of cell voltage levels or by comparing cell SoC levels (Section 3.2). By considering the differences in SoC and in the actual capacity of the cells, the balancing strategy can be even further improved to balance according to a uniformity of the remaining (usable) cell capacity.

3.1.10 Thermal management

The thermal management includes the monitoring of the cell and module temperatures and, if available, the activation of fans and/or pumps for the cooling and heating of the battery.

3.2 Methods for state estimation (SoC)

The SoC is calculated by dividing the charge present in the battery Q by the maximum possible charge $Q_{\max}$. This relationship is described by Equation 1 and the SoC is expressed as a value between 0% and 100%.

$$\mathrm{SoC} = \frac{Q}{Q_{\max}} \cdot 100\% \tag{1}$$

It is important to note that the SoC of lithium-ion batteries cannot be measured directly and that the correct value also depends on the measurement procedure, for example, by the test temperature.

In the past, several methods have been developed for the purpose of SoC calculation. A good overview of the possible methods can be found in Languang *et al*. (2013), which includes a brief discussion of the advantages and disadvantages of the different methods and the resulting estimation errors. Another overview can be found in Zhang and Lee (2011) with additional subsections for the capacity and the lifetime prediction. The possible SoC estimation methods can be classified as follows.

- discharge test;
- Coulomb counting (ampere-hour counting);
- OCV measurement;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- impedance measurements (with electrical impedance spectroscopy, EIS);
- algorithms based on neural networks;
- heuristic interpretation of measurements, for example, by fuzzy logic;
- model-based estimation methods, for example, by Kalman filter, or other state observers.

In literature, different approaches are described based on ampere-hour counting or by utilization of the relationship between SoC and OCV. However, these methods are not reliable for a precise determination of the SoC in the case of traction batteries because of the following facts:

- The voltage and current measurements are inaccurate and noisy. An offset in the current measurement will lead to significant errors in the calculation of the battery charge.
- The battery characteristics change significantly especially at low temperatures. However, the cell and the sensor temperature will be different even if the sensor is mounted close to the cells.
- Batteries in HEV application are usually operated in a small SoC range only. Thus, it is not possible to reset the charge counter because the battery is possibly never fully charged.
- For some lithium-ion technologies, such as LFP (lithium iron phosphate) or LTO (lithium titanate oxide), the cells show a very good performance over a wide range of SoCs. In other words, the battery characteristics over SoC are very similar and the measurements contain only limited information about the actual SoC (Figure 9).
- Some technologies, such as LFP, show hysteresis effects on the relationship between the OCV over the SoC (the hysteresis is not shown in Figure 9).

Promising methods rely on a battery model and utilize an observer design for estimating the internal states of this model (e.g., see Plett (2006) for a sigma-point Kalman filter or Kim (2005) for a sliding mode observer). Section 3.3 provides further details about the advantages of a model-based estimation of the SoC.

On the other hand, model-based algorithms are computationally very demanding. Thus, simplifications are required for an SoC estimation for all cells. For example, these SoC values are used for the correct activation of the charge equalization circuit. However, the precise estimation of the SoC in the cells is difficult especially while operating the vehicle. During operation, a voltage drop occurs during charging and discharging and differences in the cell resistances yield to inaccurate predictions of the cell states, if the predication is based on the cell voltage level only. Thus, the algorithms for controlling the charge equalization circuit must consider this effect in order to avoid an inappropriate activation of the balancing.

3.3 Model-based SoC algorithms in detail

In order to achieve high accuracy, model-based algorithms can be used, for example, Kalman filter or nonlinear adaptations of this filter, for example, sigma-point Kalman filter or extended Kalman filter (EKF). Kalman filters have played an important role in controls of various systems and this technique has proved to be extremely adequate for many HEV and EV applications.

In order to adapt the internal states of the model, the outputs of the battery model are compared with the measurement results. The observer iteratively updates the battery states and parameters such that the model outputs fit to the measurements.

The model, used for this approach, consists of an electrical battery model and a thermal model. An initial (offline) parameter identification technique is required. The parameters, which are often considered in these models, depend on SoC, temperature, current rates, life cycle, or duty cycle. The battery model should fulfil the following requirements:

- The model must be suitable for online calculation.
- A simple model is preferred because of restrictions of computation time and memory.
- The model must cover the whole battery operating range, that is,
  - temperature range,
  - SoC range,
  - current range, and

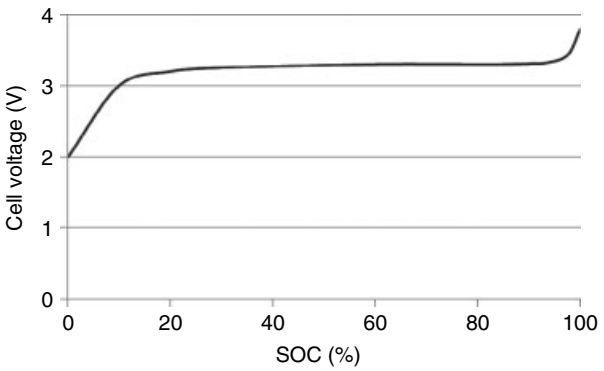


Figure 9. Relationship between state-of-charge and open circuit voltage for LFP cells. (Reproduced by permission of AVL List GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

  - dynamic range of battery operation (from pulses of seconds to constant current over minutes).
- The model parameters must be adapted over lifetime such that the changing characteristics over battery life are considered.

The Kalman filter performs two different calculation steps. In the prediction step, the battery voltage and its uncertainty are predicted based on the measured current. In the update step, the state estimates are updated according to the difference between the predicted voltage and the measured battery voltage.

Despite of having advantages, the Kalman filter has some disadvantages too. The disadvantages can be summarized as follows:

- High memory consumption and computation time.
- Floating point operations are required (a fixed-point implementation is challenging for nonlinear systems).
- The joint estimation of states and parameters can lead to stability problems (Hu *et al*., 2012, and Andre *et al*., 2013).
- It is difficult to quantify the accuracy.
- The battery dynamics cannot be described perfectly and thus model errors cannot be avoided.

Especially the accuracy of the SoC method is difficult to define. First, a definition and measurement procedure for the accuracy is required. In addition, the accuracy depends on many influencing factors, such as

- sensor quality (sensor noise and offset);
- quality of signal processing (antialiasing filter, digital signal processing, and time jitter between voltage and current signals);
- quality of battery model (model errors);
- calibration of the battery model (calibration errors);
- choice of estimation method (approximation method for nonlinear system);
- calibration parameters of the observer.

On the other hand, these model-based estimators are usually robust and quite efficient. In addition, some estimators, such as the Kalman filter, can also provide error bounds on the estimates.

3.4 State-of-function

The definition of the SoF depends on the application. In HEV, the battery is used for boosting and recuperation of brake energy. As such, the power of the battery is one of the major measures for its functionality. Thus, the SoF includes indicators for

- peak charging/discharging current (for 2, 10, 30 s, or similar);
- peak charging/discharging power (for 2, 10, 30 s, or similar);
- cold cranking performance (if required);
- maximum and minimum pack voltages;
- indicators for derating due to low or high battery temperature.

For EVs, additional measures are important. Thus, the SoF can include indicators such as

- remaining charge assuming a specific load profile and/or
- remaining energy assuming a specific load profile.

The SoF can be calculated using predefined maps, for example, by storing the maximum charge power in a map over SoC, temperature, and SoH. On the other hand, the SoF can be predicted using model-based techniques, which utilize the state estimation and battery model of the SoC calculation as described in Section 3.3.

3.5 State-of-health

Similar to the definition of the SoF, the SoH can be defined depending on the usage of the battery in an HEV or EV. In the HEV, the SoH is usually an indicator for the increase in the battery resistance. The EoL of the battery is reached, if the battery power is limited by its resistance value. The calculation of the SoH can be performed as defined in Equation 2, whereas the SoH depends on the actual resistance value R, on the resistance at beginning-of-life (BoL) R_{BoL} and at EoL R_{EoL}, respectively.

$$\mathrm{SoH} = \frac{R_{\mathrm{EoL}} - R}{R_{\mathrm{EoL}} - R_{\mathrm{BoL}}} \cdot 100\% \tag{2}$$

For this calculation, the SoH is defined as 100% for a new battery and reaches 0% at EoL. The calculation assumes a linear relationship between increase in battery resistance and SoH. Other solutions incorporate nonlinear relationships between the battery characteristics and the battery life, which is usually a more realistic assumption of real-world battery degradation.

In EVs, on the other hand, the main indicator for the SoH is the capacity loss, which can be between 20% and 30% at EoL. For this kind of applications, the SoH can be calculated, if the actual battery capacity C is known and when this value is compared with the capacity at

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

BoL and EoL (C_{BoL} and C_{BoL}, respectively), as defined in Equation 3.

$$\mathrm{SoH} = \frac{C - C_{\mathrm{EoL}}}{C_{\mathrm{BoL}} - C_{\mathrm{EoL}}} \cdot 100\% \tag{3}$$

It is important to notice that additional measures for the SoH exist. For instance, the cells in the battery will degrade differently. In addition, the weakest cell will limit the minimum/maximum values for the battery voltage, current, and temperature. Thus, the SoH can be defined as the relative difference between the individual cell characteristics of the battery. Another measure for the health of the battery is the increase in the self-discharge rate of the battery over time. A corresponding SoH value can be derived in a similar approach as shown in the calculations earlier.

3.6 Thermal management

The thermal management of batteries is the base for an effective operation in all climates. The BMS calculates the thermal energy transport and the need for heating or cooling including all measurements and diagnostic functions, for example, a temperature measurement plausibility check. The system of the battery thermal management subdivides in the following main functions:

- Sensor value conditioning
  - Convert voltage signals to temperature or pressure signals
  - Electrical check of sensor values
  - Filtering of physical values
  - Plausibility check
- Thermal modeling
  - Calculation of cell/module temperature values
  - Calculation of heat transfer between cells and coolant
  - Calculation of coolant temperature values
- Thermal controlling
  - Components switch on and off (e.g. fan)
  - Plausibility check of requests
  - Compare and reach temperature targets.

In detail, the software module acquires the temperatures via thermal sensors and calculates the coolant temperature at specific points. Moreover, a detailed evaluation can estimate the cell internal temperatures at critical points. The fan and pump control values are calculated and checked for plausibility.

The placement of the temperature sensors is very much depending on the type of cell (pouch cell, cylindrical or prismatic cell) and the design of the battery pack (see Rechargeable Battery Basics). Therefore, the placement is supported by results of thermal simulations with finite element analysis (FEA). Especially for pouch type cells, the placement is critical as the thermal conductivity of the pouch surface is worse than for other types of cells.

Bad cell-to-cell connections can lead to critical hot spots at the cell terminals. These hot spots are difficult to detect, if the battery pack is not equipped with a high number of temperature sensors. In order to improve the reliability of the cell diagnostic and the response time in case of connection failures, a monitoring of the cell resistance values could be applied in addition to temperature measurements.

4 FUNCTIONAL SAFETY

Each function has the potential risk to fail and might cause harm (see Battery safety for lithium batteries in vehicle applications). In case of batteries, the lithium technology is especially very sensible to over-/undervoltage, over-/undercharge, and over-temperature conditions. Therefore, the battery control system must be made safe with respect to malfunctions as an incorrect usage of the battery can lead to safety critical situations.

The impact of the conditions mentioned earlier is depending on the cells, which are applied in the battery. The safety properties of the cells itself can be classified into hazard levels, for example, defined by European Council for Automotive R&D (EUCAR). These levels describe the impact and the mechanical, thermal, and chemical hazards on a cell level. It is clear that safety critical cells need more efforts on the control side than BMSs, which are used for very mature and safe cells. As such, the technical solution itself may vary between the different suppliers.

Section 4.1 focuses on the concept phase in the development process, whereas Section 4.2 describes the possible technical implications on the BMS.

4.1 Functional safety in the BMS concept phase

The standard ISO26262 can be used during the development of a BMS. Despite mandatory processes during the development, this standard also covers the life cycle of the product including the production and the usage phase. A detailed discussion is out of the scope of this work, but three development steps during the concept phase are mentioned in this subsection, namely the hazard analysis and risk assessment, the determination of the ASIL, and the development of the function safety concept.

The aim of the hazard analysis and risk assessment is the identification and the categorization of potential

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

hazards. During the assessment, different risk parameters are defined, namely the severity, the exposure, and the controllability. In a consecutive step, safety goals and their assigned ASIL are determined by a systematic evaluation of these hazardous situations. Finally in the last step, a functional safety concept for the BMS can be derived.

Depending on the ASIL of the system, different requirements of the control system can be derived. One of three basic technical solutions can be selected according to the ASIL, but the detailed concept must be appropriately developed for the specific application. For ASIL A, the first level, a simple hardware (HW) watchdog could be sufficient. For ASIL B and C, a high level monitoring concept could be implemented, whereas for ASIL D, a dual-processor system could be the proper solution for the high safety demands.

4.2 Possible technical implications on the battery management system

It seems to be reasonable that a BMS is classified as safety relevant system. As such, a three-level monitoring concept could be implemented, for example, by implementing safety critical functions and diagnosis on the first level, by implementing a process monitoring on the second level, and by processor monitoring on the third level (see, e.g., Schwertfuehrer (2003) for a further description).

Possible technical solutions for the monitoring on the second level could include the following functionalities:

- CAN diagnostic monitoring (e.g., via checksum and alive counter),
- pack current monitoring,
- cell voltage monitoring,
- cell temperature monitoring,
- isolation monitoring, and
- fault reaction monitoring.

For the third monitoring level, a diagnosis of the main processor of the BCU can be implemented. For distributed controller architectures, a safety communication between the controllers can be implemented such that the different controllers monitor their activity and the correctness of their calculations. For instance, an additional monitoring can be implemented to ensure the correct program flow and the validity of the measured cell voltages. For this reason, the safety communication could include a question–answer game between module controller and BCU, for example, over the CAN bus. In addition, module controller can include redundant safety relevant functions, for example, for the opening of the main contactors in case of a severe overvoltage condition. The correct execution of this emergency switch-off functionality itself can be tested during the initialization phase of the battery. For a centralized system, the BCU can be supported by an additional microcontroller or by an application-specific integrated circuit (ASIC), which is responsible for the processor monitoring.

5 CONCLUSION

This chapter presented the concepts for the realization of a battery indication and management system. In the first chapter, the main indicators for the state of the battery have been defined, namely the SoC, the SoH, and the SoF, and the main tasks for the battery management have been identified.

The hardware was explained by discussing the different components, which are part of the BMS. In addition, different concepts for the controller architecture have been presented and the advantages and disadvantages of centralized and distributed systems have been examined. For the required charge equalization between the cells, concepts based on active and passive balancing systems have been introduced.

The presentation of the software used a structuring into different software modules. The focus of this section was the explanation of methods for the battery state calculation and the explanation of the thermal management. For the SoC, different possible algorithms have been listed and the model-based algorithm has been discussed in detail. A brief explanation of the implications of a functional safety concept concluded this chapter.

REFERENCES

Andre, D., Appel, C., Soczka-Guth, T., and Sauer, D.U. (2013) Advanced mathematical methods of SOC and SOH estimation for lithium-ion batteries. *Journal of Power Sources*, **224**, 20–27.

Daowd, M., Omar, N., Van Den Bossche, P., and Van Mierlo, J. (2011) Passive and Active Battery Balancing Comparison Based on MATLAB Simulation. *IEEE Vehicle Power and Propulsion Conference 2011 (VPPC 2011)*.

Hu, C., Youn, B.D., and Chung, J. (2012) A multiscale framework with extended Kalman filter for lithium-ion battery SOC and capacity estimation. *Applied Energy*, **92**, 694–704.

Kim, I.-S. (2005) The novel state of charge estimation method for lithium battery using sliding mode observer. *Journal of Power Sources*, **163** (1), 584–590.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Languang, L., Xuebing, H., and Jianqiu, L., *et al*. (2013) A review on the key issues for lithium-ion battery management in electric. *Journal of Power Sources*, **226**, 272–288.

Pichler, P., Heidenbauer, O., Lind, R., *et al*. (2007), *Development of Li-ion Battery Systems for HEV Applications at MAGNA STEYR*. Electric Vehicle Symposium EVS.

Plett, G.L. (2006) Sigma-point Kalman filtering for battery management systems of LiPB-based HEV battery packs: Part 1 and 2. *Journal of Power Sources*, **161** (2), 1356–1384.

Schwertfuehrer, G. (2003) Method for Monitoring Distributed Software. Patent application, DE10331873.

Zhang, J. and Lee, J. (2011) A review on prognostics and health monitoring of Li-ion battery. *Journal of Power Sources*, **196**, 6007–6014.

FURTHER READING

Andrea, D. (2010) Battery Management Systems for Large Lithium Ion Battery Packs, 1st edn, Artech House, Norwood, MA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto048
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

EV Powertrain Configurations

K.T. Chau
The University of Hong Kong, Pokfulam, Hong Kong

1 Introduction 1
2 Motor Topology Variations 2
3 System Topology Variations 6
Acknowledgments 10
Related Articles 11
References 11

1 INTRODUCTION

Previously, the electric vehicle (EV) was mainly converted from the internal combustion engine vehicle (ICEV), simply replacing the combustion engine by the electric motor while retaining all the other components. This converted EV has been faded out because of the drawback of heavy weight, loss of flexibility, and degradation of performance. At present, the modern EV is purposely built. This purpose-built EV is based on original body and frame designs to satisfy the structural requirements unique to EVs and to make use of the greater flexibility of EV powertrain.

Compared with the ICEV, the EV enjoys a much more flexible configuration. This flexibility is due to several factors that are provided by the EV. First, the power flow in the EV is mainly via flexible electrical wires rather than bolted flanges or rigid shafts. Thus, the concept of distributed electrical subsystems in the EV is really achievable. Second, different EV powertrain configurations such as single-motor drives and multiple-motor drives involve a significant difference in both hardware count and control complexity. Third, different electric motors such as the direct current (DC) and alternating current (AC) motors work with different power converters for electric propulsion (Chau and Wang, 2005). Fourth, different energy sources such as the batteries, fuel cells, capacitors, and flywheels (Chau, Wong, and Chan, 1999; Chau and Wong, 2001) interact with different power converters. The corresponding refueling systems also involve different hardware and mechanism. For example, the batteries can be electrically recharged via conductive or inductive means, or can be mechanically swapped and then recharged centrally.

Figure 1 shows the general system configuration of the EV, consisting of three major systems—EV powertrain, EV energy system, and EV auxiliary system, where a mechanical link is represented by a double line, an electrical link is represented by a thick line, a control link is represented by a thin line, and the arrow on each line denotes the direction of electrical power flow or control signal flow. The EV powertrain comprises the electronic controller, the power converter, the electric motor, and the transmission. There are bidirectional control links among the electronic controller, the power converter, and the electric motor, indicating that the electronic controller functions to control both the power converter and the electric motor and to import their feedback signals such as voltage, current, speed, and temperature. On the other hand, the electronic controller generally does not control the transmission, but collects its feedback signals such as speed and temperature for control purposes. It should be noted that the transmission may be totally removed, so-called direct-drive, if the electric motor adopts low speed high torque design.

The EV energy system involves the rechargeable energy source, the energy management unit, and the energy refueling unit. The energy management unit functions to manage and monitor the energy source, and then to communicate internally with the energy refueling unit

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

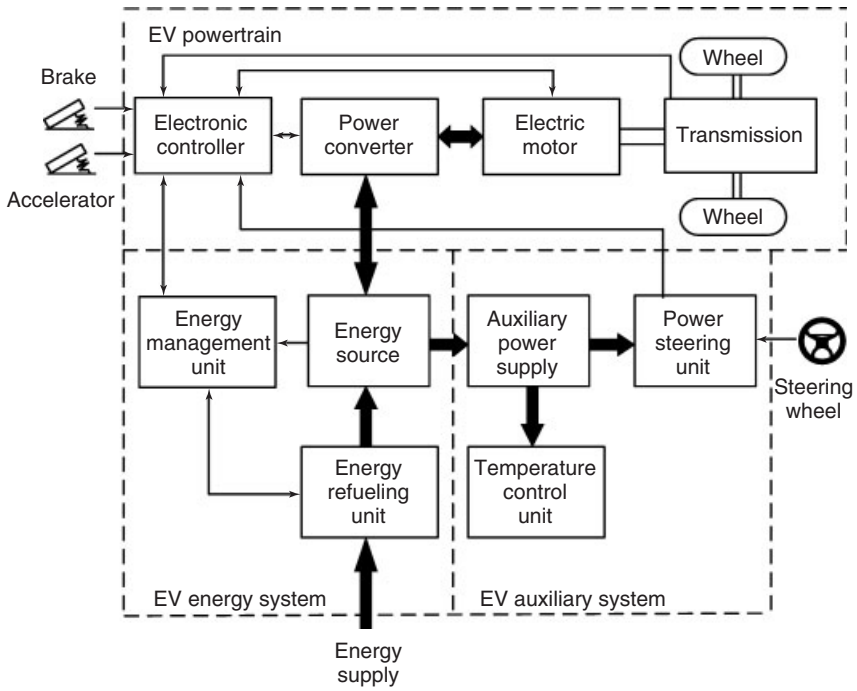


Figure 1. General EV system configuration.

and externally with the electronic controller of the EV powertrain. The energy refueling unit functions to regulate the external energy supply for refueling the energy source. In terms of power flow, the energy source provides the necessary power to the power converter of the EV powertrain for electric propulsion, and the required power for the EV auxiliary system. Typically, a lithium-ion battery is selected as the energy source. The corresponding energy management unit and refueling unit becomes the battery management unit and battery charger, respectively.

The EV auxiliary system consists of the auxiliary power supply, the power steering unit, and the temperature control unit. The auxiliary power supply provides the necessary power with different voltage levels for all EV auxiliaries, especially for power steering and temperature control. The power steering unit functions to determine how sharply the vehicle should turn for cornering in accordance with the angular position of the steering wheel. Moreover, it sends the steering information to the electronic controller of the EV powertrain to perform necessary speed control of the electric motor. The temperature control unit, which generally consists of a cooler and a heater, functions to control the temperature of the vehicle compartment. More importantly, the cooler and heater need to keep the battery pack working at a proper temperature range. If the battery is too hot, it can deteriorate the battery life or even cause the EV to stop working. At very hot temperatures, the battery pack may explode. If the battery is too cold, the EV can be sluggish or its driving range may drop significantly. Moreover, at very low temperatures, the battery may not be able to accept recharging.

On the basis of the control inputs from the brake and accelerator pedals, the electronic controller provides proper control signals to switch on or off the power devices of the power converter that functions to regulate power flow between the electric motor and the energy source. The backward power flow is due to regenerative braking of the EV and this regenerative energy can be stored in the rechargeable energy source. Notice that most available EV batteries as well as capacitors and flywheels readily accept regenerative energy. The energy management unit cooperates with the electronic controller to control regenerative braking and its energy recovery.

In this chapter, viable EV powertrain configurations will be classified into two groups—one is based on the variations in motor topologies while another is based on the variations in system topologies. The corresponding discussion will be focused on their basic principle and qualitative comparison.

2 MOTOR TOPOLOGY VARIATIONS

The group of viable EV powertrain configurations due to the variations in motor topologies (Chau, Chan, and Liu, 2008;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Chau, 2009) will be split into the DC, AC, and switched reluctance (SR) types. All these electric motors have been accepted for the EV powertrain.

2.1 DC powertrain

The DC powertrain has ever been widely used for EVs. As shown in Figure 2, it consists of the electronic controller, the DC motor, the DC–DC converter, and the transmission. The key components are the DC motor and the DC–DC converter.

Figure 3 shows the basic topology of the DC motor. On the basis of the methods of field excitation, it can be further split into the self-excited DC and separately excited DC types. On the basis of the source of field excitation, it can also be grouped as the wound-field DC and permanent magnet (PM) DC types. As determined by the mutual interconnection between the field winding and the armature winding or the use of PM excitation, the whole family consists of the separately excited DC, shunt DC, series DC, and PM DC types. All types of the DC motor suffer from the same problem due to the use of commutators and brushes. Commutators cause torque ripples and limit the motor speed, whereas brushes are responsible for friction and radio-frequency interference. Moreover, due to the wear and tear, periodic maintenance of commutators and brushes is always required. These drawbacks make them less reliable and unsuitable for maintenance-free operation. The major advantages of the DC motor are the maturity and simplicity.

When the DC–DC converter adopts the chopping mode of operation, it is usually named as the DC chopper. This DC chopper can be classified as the first-, second-, two-, and four-quadrant versions. The first-quadrant DC chopper is suitable for motoring and the power flow is from the source to the load, whereas the second-quadrant one is for regenerative braking and the power flow is out from the load into the source. As regenerative braking is very essential for EVs that can significantly extend the vehicle driving range,

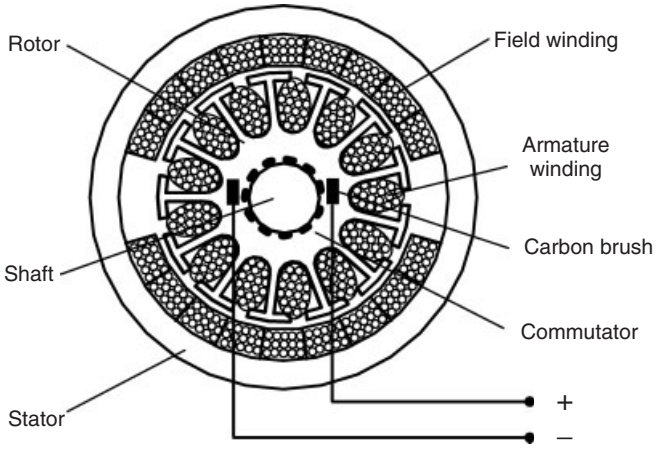


Figure 3. DC motor topology.

the two-quadrant DC chopper is preferred as it is suitable for both motoring and regenerative braking. Moreover, instead of using mechanical contactors to achieve reversible operation, the four-quadrant DC chopper shown in Figure 4 can be employed so that motoring and regenerative braking in both forward and reversible operations are controlled electronically.

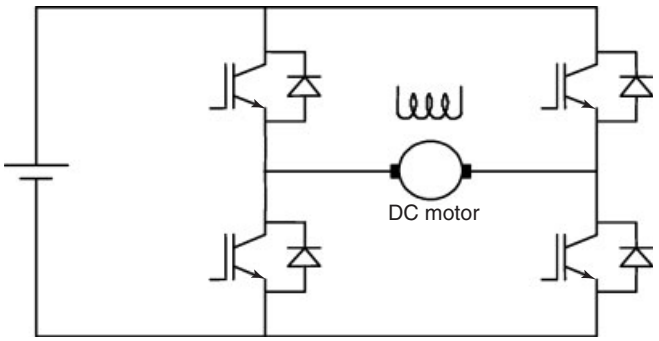


Figure 4. Four-quadrant DC chopper topology.

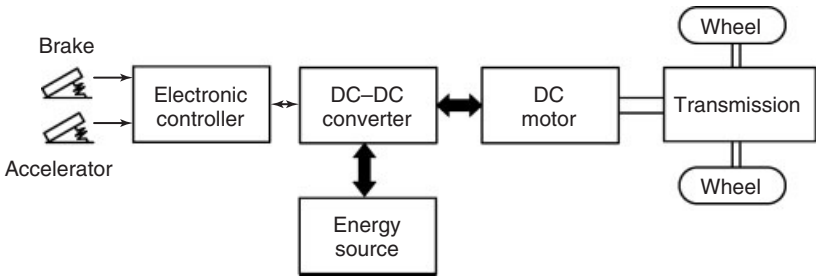


Figure 2. DC powertrain configuration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In short, the DC powertrain takes the merits of being mature and simple, but suffers from the demerits of relatively low efficiency and need for maintenance. So, this powertrain is becoming obsolete for modern EVs.

2.2 AC powertrain

The AC powertrain is widely adopted for EVs. As shown in Figure 5, it consists of the electronic controller, the AC motor, the pulse-width modulation (PWM) inverter, and the transmission. The corresponding AC motor includes two major types—namely the induction motor and the PM brushless motor.

At present, the induction motor powertrain is the most mature technology among all AC powertrains. There are two types of induction motors, namely the wound-rotor and the cage-rotor. Because of high cost, need for maintenance, and lack of sturdiness, the wound-rotor induction motor is less attractive than the cage-rotor counterpart. As shown in Figure 6, the cage-rotor induction motor is loosely named as the induction motor for EV powertrain. Apart from the common advantages of AC machines such as the brushless and hence maintenance-free operation, the induction motor possesses the definite advantages of low cost and ruggedness.

The PM brushless motor is becoming more and more attractive for EV powertrain. It possesses the definite advantages of higher efficiency and higher power density than the induction motor. Nevertheless, it suffers from the drawbacks of relatively high PM material cost and uncontrollable PM flux. On the basis of the back electromotive force (EMF) and current waveforms, the PM brushless motor can be divided into two main types (Pillay and Krishnan, 1988)—the PM brushless alternating current (BLAC) and the PM brushless direct current (BLDC). The PM BLAC motor is fed by sinusoidal AC current while the back EMF waveform is sinusoidal, whereas the PM BLDC motor is fed by rectangular AC current while the back EMF waveform is trapezoidal. Actually, the PM BLAC motor is usually called the *permanent magnet synchronous machine* (*PMSM*) (Chan and Chau, 1996). As the interaction between trapezoidal EMF and rectangular current in the motor can produce higher torque product than that produced by sinusoidal EMF and sinusoidal current, the PM BLDC motor possesses higher power density than the PMSM (Gan *et al*., 2000). Meanwhile, the PM BLDC motor has a significant torque pulsation (Kim, Kook, and Ko, 1997), whereas the PMSM produces an essentially constant instantaneous torque or so-called smooth torque like a wound-rotor synchronous motor. According to the position of PMs in the rotor, PM brushless motors have many possible topologies such as the surface-mounted, surface-inset, interior-radial, and interior-circumferential. Figure 7 shows a typical interior-radial PM brushless motor topology. Each PM brushless motor topology can operate at both the BLAC and the BLDC modes if the torque density, torque smoothness, and efficiency are not of great concern.

The voltage-fed PWM inverter shown in Figure 8 is almost exclusively used for AC powertrain. The inverter topology highly depends on the technology of power

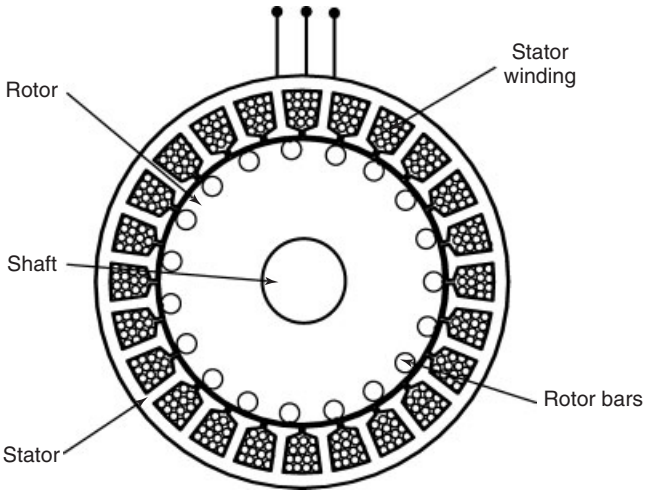


Figure 6. Induction motor topology.

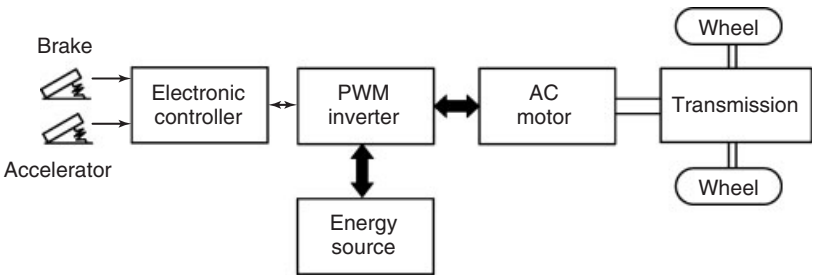


Figure 5. AC powertrain configuration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

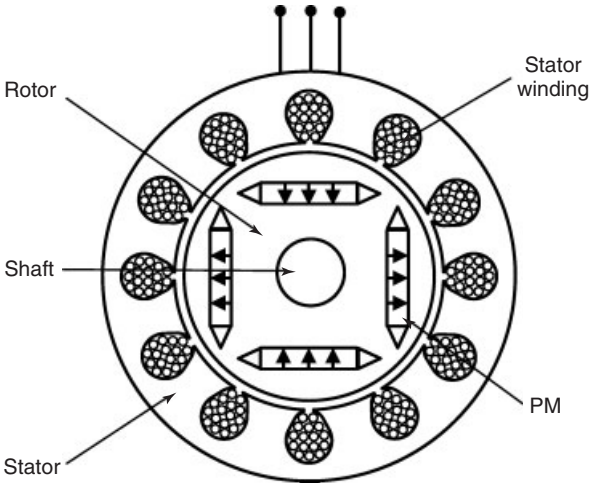


Figure 7. PM brushless motor topology.

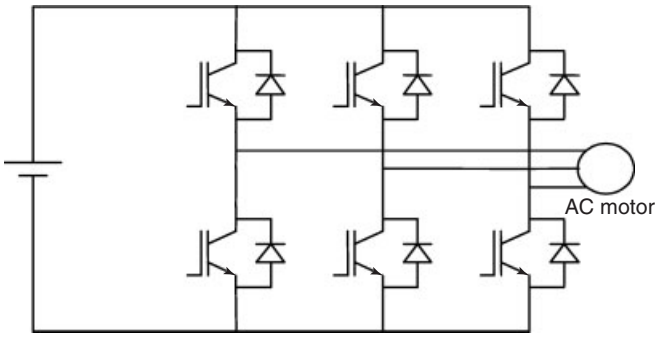


Figure 8. PWM inverter topology.

devices. At present, the insulated gate bipolar transistor (IGBT)-based inverter is most attractive. The inverter design is generally governed by the ratings of power devices and the selection of switching schemes. The ratings of power devices are based on three criteria: first, the voltage rating is at least twice the nominal supply voltage because of the voltage surge during switching; second, the current rating is large enough so that there is no need to connect several power devices in parallel; and third, the switching speed rating is sufficiently high to suppress motor harmonics and acoustic noise levels. The power module is normally a two-in-one or even six-in-one type, namely two or six devices are internally connected with an antiparallel diode across each device, to minimize wiring and stray impedance. On the other hand, the selection of switching schemes depends on the motor types. For the induction motor and the PM BLAC motor, the corresponding switching schemes are similar, aiming to provide a near-sinusoidal AC current with the minimum switching loss and acceptable harmonic distortion. For the PM BLDC motor, the corresponding switching scheme aims to provide a rectangular AC current under the two-phase 120° conduction mode or the three-phase 180° conduction mode (Zhu and Howe, 2007).

2.3 SR powertrain

The SR powertrain is occasionally adopted for EVs. As shown in Figure 9, it consists of the electronic controller, the SR motor, the SR converter, and the transmission. Although the concept of variable reluctance was adopted for electric motors over a century, the SR motor could not realize their full potential until the advent of power electronics. Figure 10 shows a typical three-phase 6/4-pole SR motor, where there are six salient poles in the stator and four salient poles in the rotor. Because of the salient nature of both the stator and rotor poles, the inductance of each phase varies with the rotor position. The operating principle of the SR motor is based on the "minimum reluctance" rule.

Differing from the PWM inverter for AC motors, the power converter for the SR motor does not need to provide bipolar operation. Many converter circuits have been developed in attempts to reduce the number of power devices and take full advantage of unipolar operation. However, when the device count is reduced, there is a penalty in the form of lower controllability, lower

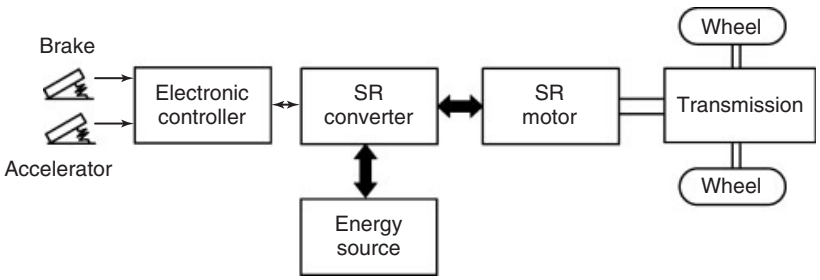


Figure 9. SR powertrain configuration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

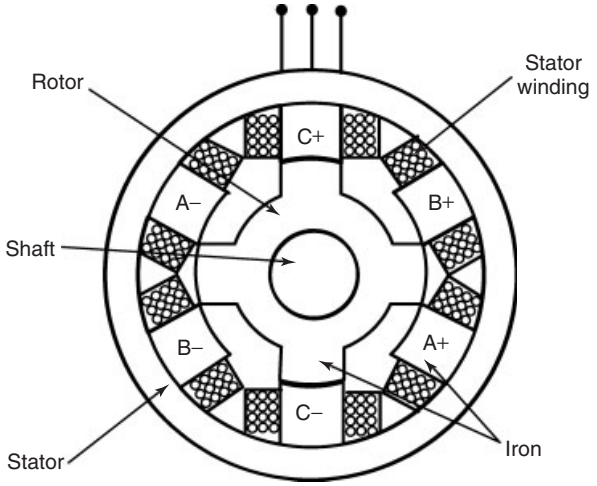


Figure 10. SR motor topology.

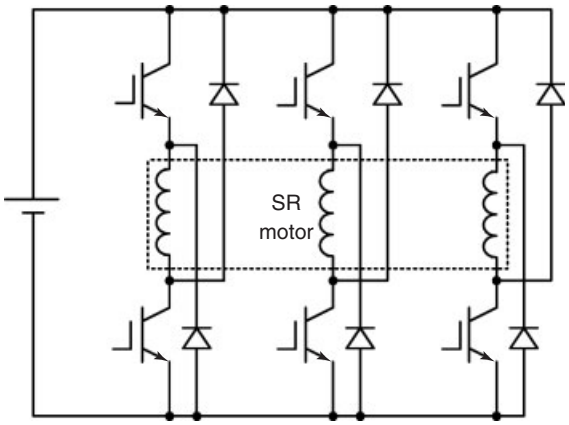


Figure 11. SR converter topology.

reliability, lower operating performance, or extra passive components. The SR converter shown in Figure 11 is well suited for the SR powertrain. It utilizes two power devices to independently control the current of each phase and two freewheeling diodes to return any stored magnetic energy to the energy source. As this circuit topology needs two power devices per phase, the converter cost is relatively higher than that with less power devices. However, this bridge arrangement allows control of each phase winding independent of the state of other phase windings. Thus, it is possible to allow for phase overlapping so as to increase the torque production and to extend the constant-power operating range of EVs.

3 SYSTEM TOPOLOGY VARIATIONS

The group of viable EV powertrain configurations due to the variations in system topologies (Chan and Chau, 2001) will be split into the converted EV and purpose-built EV types. The purpose-built EV type is further divided into the single-motor and multiple-motor types. All these system topologies have been adopted for the EV powertrain.

3.1 Converted powertrain

Figure 12 shows the earliest EV powertrain configuration which is a direct conversion from the existing ICEV adopting longitudinal front-engine front-wheel drive. It consists of an electric motor, a clutch, a gearbox, and a differential (Unnewehr and Nasar, 1982). The clutch is a mechanical device that is used to connect or disconnect power flow from the electric motor to the wheels. The gearbox is another mechanical device that consists of a set of gears with different gear ratios and a gear shifting mechanism. By incorporating both the clutch and the gearbox, the driver can shift the gear ratios and hence amplify the torque going to the wheels. The wheels have the high torque low speed features in the lower gears and the high speed low torque features in the higher gears. The differential is a mechanical device that enables the wheels to be driven at different speeds when cornering—the outer wheel covering a greater distance than the inner wheel.

The key component of this converted powertrain configuration is the gearbox. By using a combination of clutch and gearbox, variable gearing can be accomplished. The purpose of variable gearing is to provide multiple-speed transmission, namely achieving wide ranges of speed and torque using different gear ratios. Generally, four-speed transmission is used for passenger cars. When the clutch is engaged, the electric motor and the gearbox are coupled

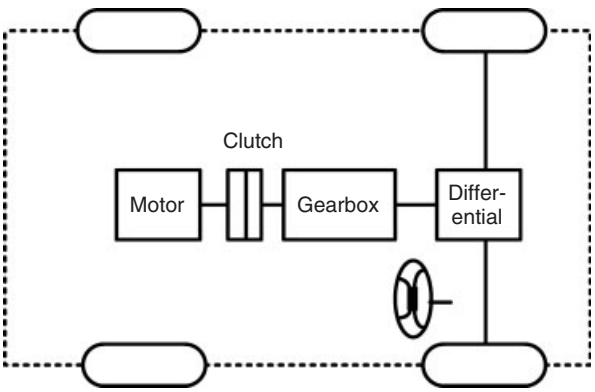


Figure 12. Converted powertrain configuration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

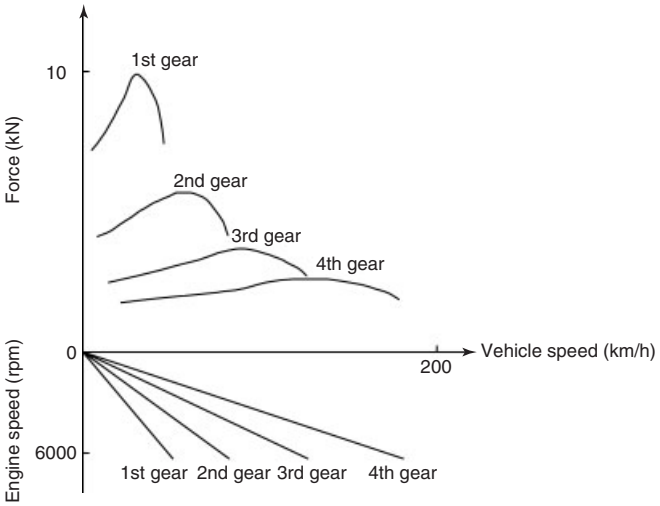


Figure 13. ICEV force-speed characteristics with four-speed transmission.

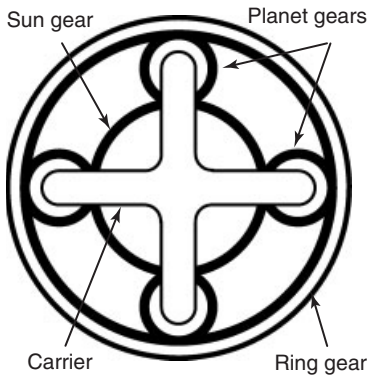


Figure 14. Planetary gear.

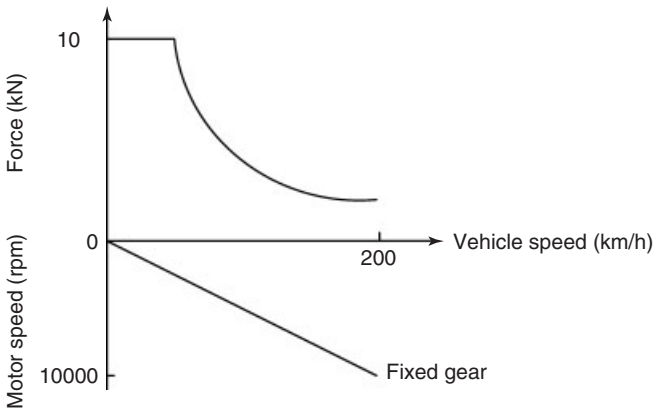


Figure 15. EV force-speed characteristics with fixed gearing.

together and power transmission is enabled. When it is disengaged manually or automatically, the power transmission is interrupted so that the gear ratio in the gearbox can be shifted. For the ICEV, there is no alternative to the use of variable gearing as the combustion engine cannot offer the desired torque-speed characteristics such as high torque for hill climbing and high speed for cruising without using multiple-speed transmission. Figure 13 shows the typical force-speed characteristics of a combustion engine with four-speed transmission. For the EV, the employment of variable gearing to achieve multiple-speed transmission used to be controversial. For the EV converted from the ICEV, the use of variable gearing was claimed to be natural because both gearbox and clutch are already present and their maintenance costs are minor. However, the concept of converted EV is obsolete as it cannot fully utilize the flexibility and potentiality offered by the EV. It was also claimed that the use of variable gearing can enhance the electric motor achieving regenerative braking and high efficiency operation over a wide speed range. With the advances of power electronics and control algorithms, both regenerative braking and high efficiency operation of electric motors can be handled by electronic means rather than mechanical means.

3.2 Single-motor-geared powertrain

Fixed gearing means that there is a fixed gear ratio between the electric motor and the driving wheels. Practically, fixed-gearing transmission is usually based on planetary gearing. As shown in Figure 14, a planetary gear consists of a sun gear, several planet gears, a planet gear carrier, and a ring gear. It takes the advantages of strong, compact, high efficiency, high speed-reduction ratio, and in-line arrangement of input and output shafts over the conventional parallel-shaft variable gearset. Moreover, modern electric motors with the use of fixed gearing can readily offer the desired torque-speed characteristics for vehicular operation. Figure 15 shows typical force-speed characteristics of an EV with fixed gearing, consisting of constant-torque operation for acceleration and hill climbing as well as constant-power operation for high speed cruising.

By replacing the gearbox with fixed gearing and hence removing the clutch, the longitudinal single-motor-geared powertrain is resulted. Figure 16 shows this arrangement which consists of an electric motor, a fixed gear, and a differential. Notice that this configuration is not suitable for the ICEV as the engine by itself, without the clutch and gearbox, cannot offer the desired torque-speed characteristics. The removal of variable gearing can significantly reduce the overall complexity, size, weight, and cost of the transmission. Moreover, the absence of gear changing, irrespective of whether it is manual or automatic, can greatly

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

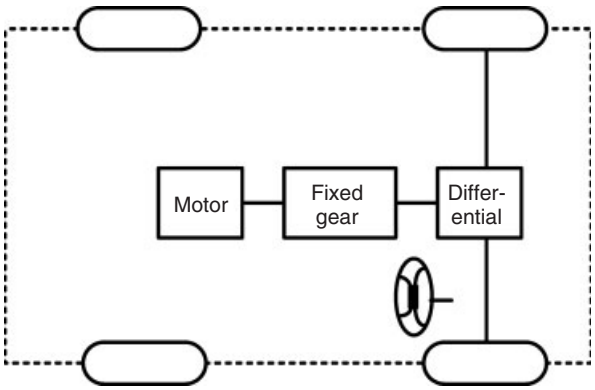


Figure 16. Longitudinal single-motor-geared powertrain configuration.

enhance smooth driving and transmission efficiency. Therefore, modern EVs almost exclusively adopt fixed gearing rather than variable gearing.

Borrowing the concept from the transverse front-engine front-wheel drive of the existing ICEV, the electric motor, fixed gear, and differential can be integrated into a single assembly while both axles are connected to the driving wheels. Figure 17 shows this transverse single-motor-geared powertrain configuration which takes the advantages of more compact size and higher transmission efficiency than its longitudinal counterpart. In fact, this configuration is the most commonly adopted powertrain by modern EVs.

3.3 Multiple-motor-geared powertrain

A differential is a standard component for conventional ICEVs and this technology can be carried forward to the EV field. When a vehicle is rounding a curved road, the outer wheel needs to travel on a larger radius than the inner wheel. Thus, the differential adjusts the relative speeds of the wheels; otherwise, the wheels will slip which causes tire wear, steering difficulties, and poor road holding. For all ICEVs, whether front- or rear-wheel drive, a mechanical differential is mandatory. Notice that this mechanical differential is not only bulky and heavy but also complicated and lossy.

For EVs, it is possible to dispense with a mechanical differential. By separately coupling two or even four electric motors to the driving wheels, the torque of each motor can be independently controlled in such a way that the differential action can be electronically achieved when cornering. Figure 18 shows the principle of electronic differential in which two electric motors, so-called dual motors, are employed. This arrangement can totally eliminate the mechanical differential, hence reducing the overall size and weight as well as improving the cornering control and transmission efficiency.

Figure 19 shows the dual-motor-geared powertrain configuration in which dual motors separately drive the driving wheels via fixed gearing. The differential action when cornering is electronically provided by the dual motors operating at different speeds. Unlike the choice between variable gearing and fixed gearing, the selection of

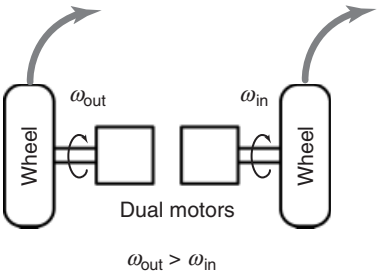


Figure 18. Principle of electronic differential.

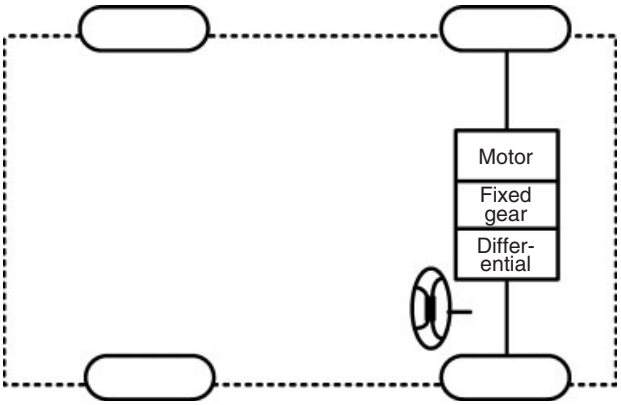


Figure 17. Transverse single-motor-geared powertrain configuration.

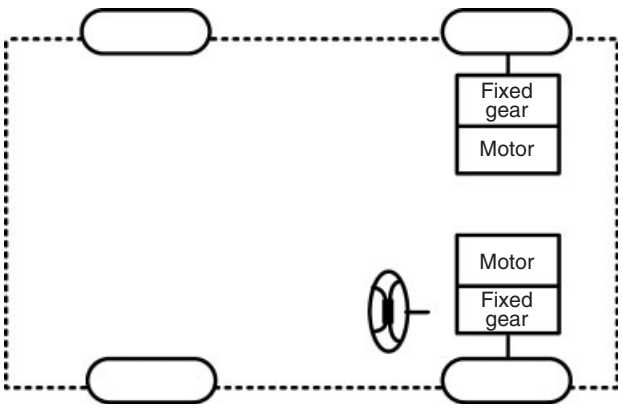


Figure 19. Dual-motor-geared powertrain configuration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

either a single-motor drive with a mechanical differential or a dual-motor drive using electronic differential is still controversial. Positively, the removal of a mechanical differential can reduce the overall size and weight while the electronic differential can accurately control the motor torques so as to achieve better performance during cornering. Negatively, the use of an additional motor and power converter causes an increase in the initial cost while the reliability of the electronic controller to accurately control the dual motors at various driving conditions is a concern. In recent years, the reliability of this electronic controller has been greatly improved by incorporating the capability of fault tolerance. For instance, the electronic controller can utilize three independent microprocessors. Two of them are used to separately control the motor torques for the left and right wheels while the remaining one is used for coordination and fault-tolerant control. All of them watch one another to improve the reliability.

3.4 Multiple-motor-geared in-wheel powertrain

In order to further shorten the mechanical transmission path from the electric motor to the driving wheel, the electric motor can be placed inside a wheel. This arrangement is the so-called in-wheel motor. Figure 20 shows the dual-motor-geared in-wheel powertrain configuration in which fixed planetary gearing is employed to reduce the motor speed to the desired wheel speed. It should be noted that planetary gearing offers the advantages of a high speed-reduction ratio as well as an in-line arrangement of input and output shafts.

In general, the electric motor for this configuration is a high speed inner-rotor motor while the high speed-reduction planetary gear is adopted which is mounted between the motor shaft and the wheel hub. Typically, this motor is purposely designed to operate over 10,000 rpm so as to give a high power density. This operating speed is mainly limited by the friction and windage losses as well as the transmission tolerance. Thus, the corresponding planetary gear ratio is of about 10 : 1 (Mellor, Allen, and Howe, 1996) to provide the wheel speed range from zero to about 1000 rpm. Apart from adopting a single-stage high reduction ratio which is demanding in construction, a two-stage planetary gear can also be employed. Inevitably, the use of mechanical gearing for speed reduction involves the transmission loss, wear-and-tear problem, and regular lubrication. It should be noted that for the high speed design, the operating frequency can be up to a kilohertz. The operating voltage should be sufficiently large to maintain the back EMF at the high rated speed.

3.5 Multiple-motor gearless in-wheel powertrain

By fully abandoning any mechanical gearing, the in-wheel drive can be realized by installing a low speed outer-rotor electric motor inside a wheel. Figure 21 shows the dual-motor gearless in-wheel powertrain configuration in which the outer rotor is directly mounted on the wheel rim. Thus, speed control of the electric motor is equivalent to the control of the wheel speed and hence the vehicle speed.

Figure 22 shows the schematic of this in-wheel motor which is a low speed outer-rotor motor. This low speed outer-rotor motor has the definite advantages of simplicity and gearless, hence offering high transmission efficiency and free from maintenance. However, because of the inherent low speed requirement of wheel rotation, typically up to 1000 rpm, the electric motor has to adopt

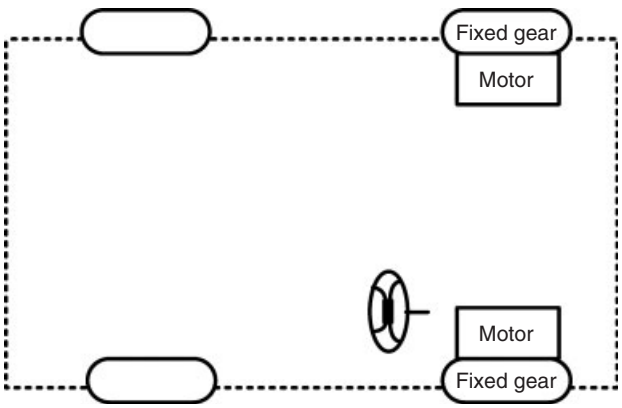


Figure 20. Dual-motor-geared in-wheel powertrain configuration.

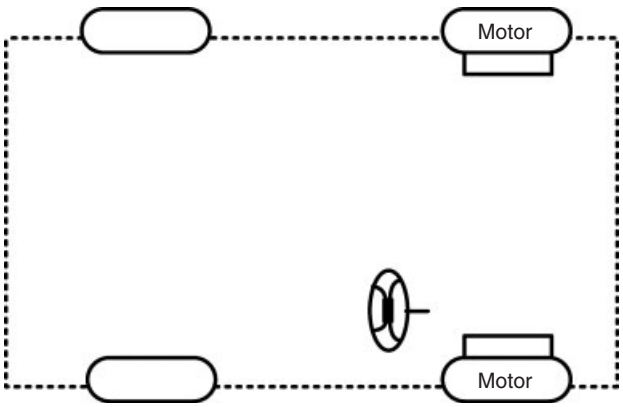


Figure 21. Dual-motor gearless in-wheel powertrain configuration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

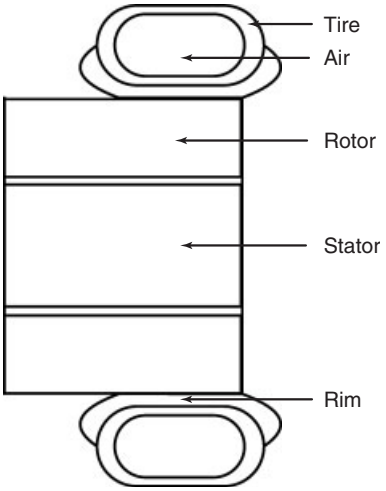


Figure 22. Gearless in-wheel motor.

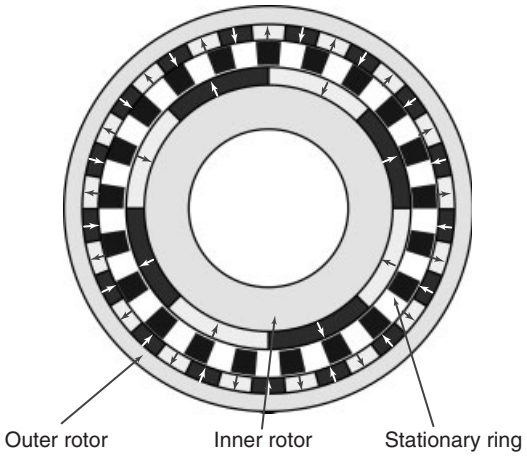


Figure 23. Coaxial magnetic gear.

the low speed design—namely, low operating frequency or large number of poles. In general, such low speed design suffers from the drawbacks of increased size, weight, and cost.

In recent years, the concept of gearless has been extended to magnetic gears because they fundamentally differ from the mechanical gears and solve their mechanical problems. Namely, they inherently offer the merits of high transmission efficiency, no wear-and-tear problem, free from maintenance, and physical isolation between input and output (Atallah and Howe, 2001). Figure 23 shows the schematic of a coaxial magnetic gear in which the inner rotor is mounted with PMs for high speed input, the outer rotor is also mounted with PMs for low speed output, and the torque transmission between two rotors is based on PM flux modulation via the stationary

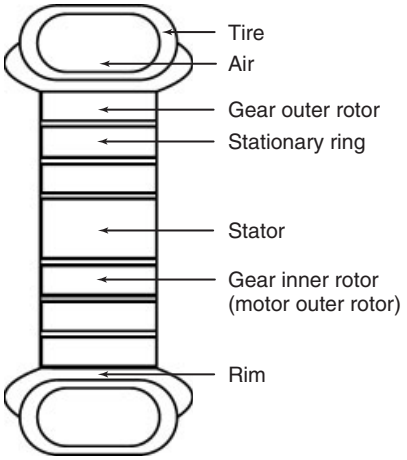


Figure 24. Magnetic-geared in-wheel motor.

ferromagnetic ring (Jian *et al*., 2009). By artfully integrating this magnetic gear into a PM brushless motor, the magnetic-geared in-wheel motor is resulted in which the low speed output requirement for direct-drive and the high speed rotating field requirement for motor design can be achieved simultaneously (Chau *et al*., 2007). Figure 24 gives the schematic of this magnetic-geared in-wheel motor. Compared with the planetary-geared in-wheel motor, this in-wheel motor offers the outer-rotor topology with reduced size and weight, while eliminating all the drawbacks due to the planetary gear. The artfulness is the share of a common PM rotor, namely the outer rotor of the PM brushless motor and the inner rotor of the coaxial magnetic gear. The operating principle of this magnetic-geared in-wheel motor is actually similar to that of the planetary-geared in-wheel motor. For instance, the stator is initially fed by high frequency power supply to produce high speed rotating field and hence creating high speed rotation at the inner rotor; then the magnetic gear steps down the inner-rotor speed, and hence boosts up the outer-rotor torque for direct-driving of the wheel.

ACKNOWLEDGMENTS

The author would like to express heartfelt thanks to all group members of the International Research Center for EVs for their contributions to this chapter. He expresses his indebtedness to Joan and Aten for their support all the way.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

RELATED ARTICLES

Overview of Electric, Hybrid and Fuel Cell Vehicles
EV powertrain parameters
General Requirement of Traction Motor Drives
DC Motor Drives
Induction Motor Drives
Permanent Magnet Brushless Motor Drives
Switched Reluctance Motor Drives
Future Direction of Traction Motor Drives

REFERENCES

Atallah, K. and Howe, D. (2001) A novel high performance magnetic gear. *IEEE Transactions on Magnetics*, **37** (4), 2844–2846.

Chan, C.C. and Chau, K.T. (1996) An advanced permanent magnet motor drive system for battery-powered electric vehicles. *IEEE Transactions on Vehicular Technology*, **45** (1), 180–188.

Chan, C.C. and Chau, K.T. (2001) *Modern Electric Vehicle Technology*, Oxford University Press, Oxford.

Chau, K.T. (2009) Electric motor drives for battery, hybrid and fuel cell vehicles in *Electric Vehicles: Technology, Research and Development* (ed. Raines, G.B.), Nova Science Publishers, Hauppauge, NY, 1–40.

Chau, K.T. and Wang, Z. (2005) Overview of power electronic drives for electric vehicles. *HAIT Journal of Science and Engineering—B: Applied Sciences and Engineering*, **2** (5–6), 737–761.

Chau, K.T. and Wong, Y.S. (2001) Hybridization of energy sources in electric vehicles. *Energy Conversion and Management*, **42** (9), 1059–1069.

Chau, K.T., Wong, Y.S., and Chan, C.C. (1999) An overview of energy sources for electric vehicles. *Energy Conversion and Management*, **40** (10), 1021–1039.

Chau, K.T., Zhang, D., Jiang, J.Z., *et al*. (2007) Design of a magnetic-geared outer-rotor permanent-magnet brushless motor for electric vehicles. *IEEE Transactions on Magnetics*, **43** (6), 2504–2506.

Chau, K.T., Chan, C.C., and Liu, C. (2008) Overview of permanent-magnet brushless drives for electric and hybrid electric vehicles. *IEEE Transactions on Industrial Electronics*, **55** (6), 2246–2257.

Gan, J., Chau, K.T., Chan, C.C., and Jiang, J.Z. (2000) A new surface-inset, permanent-magnet, brushless DC motor drive for electric vehicles. *IEEE Transactions on Magnetics*, **36** (5), 3810–3818.

Jian, L., Chau, K.T., Gong, Y., *et al*. (2009) Comparison of coaxial magnetic gears with different topologies. *IEEE Transactions on Magnetics*, **45** (10), 4526–4529.

Kim, Y., Kook, Y., and Ko, Y. (1997) A new technique of reducing torque ripples for BDCM drives. *IEEE Transactions on Industrial Electronics*, **44** (5), 735–739.

Mellor, P.H., Allen, T. and Howe, D. (1996) *Hub-mounted electric drive-train for a high performance all-electric racing vehicle*. IEE Colloquium on Machines and Drives for Electric and Hybrid Vehicles, 3/1–3/6.

Pillay, P. and Krishnan, R. (1988) Modeling of permanent magnet motor drives. *IEEE Transactions on Industrial Electronics*, **35** (4), 537–541.

Unnewehr, L.E. and Nasar, S.A. (1982) *Electric Vehicle Technology*, John Wiley & Sons, Inc, New York.

Zhu, Z.Q. and Howe, D. (2007) Electrical machines and drives for electric, hybrid and fuel cell vehicles. *IEEE Proceedings*, **95** (4), 746–765.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto049
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

EV Powertrain Parameters

K.T. Chau
The University of Hong Kong, Pokfulam, Hong Kong

1 Introduction 1
2 Mass and Size Parameters 1
3 Force and Torque Parameters 2
4 Power and Energy Parameters 5
5 Driving-Cycle Performances 7
Acknowledgments 11
Related Articles 11
References 11

1 INTRODUCTION

The electric vehicle (EV) powertrain consists of the electronic controller, the electric motor, the power converter, and the mechanical transmission. In contrast, the internal combustion engine vehicle (ICEV) powertrain consists of the electronic controller, the combustion engine, and the transmission. The core difference between them is the electric motor versus the combustion engine. Although both of them serve to produce the driving torque, their principles of operation and hence parameters are fundamentally different. For instance, the combustion engine power is normally described by the engine displacement in cubic centimeter (cc) or liter (L), whereas the electric motor power is always in kilowatt (kW). Thus, the EV powertrain parameters are mainly extended from the electric motor (Chan and Chau, 2001), rather than borrowed from the well-established definitions used for the ICEV powertrain (Bosch, 2007).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 MASS AND SIZE PARAMETERS

Among various components of the EV powertrain, the electronic controller has virtually no influence on the overall mass and size of the powertrain. On the contrary, the transmission dominates the overall mass and size especially when adopting the mechanical differential (Duffy, Stockel, and Stockel, 1988). Notice that the transmission is optional, depending on the type of electric motor and the number of electric motors adopted. In addition, the power converter essentially exhibits the same mass and occupies the same size for the same type of powertrain, again depending on the type of electric motor used (Chau and Wang, 2005). For instance, the converter for the direct current (DC) motor powertrain is lighter and smaller than the inverter for the alternating current (AC) motor powertrain, whereas various inverters have similar mass and size for the AC motor powertrain. Therefore, the electric motor plays a very important role on the overall mass and size of the powertrain.

The electric motor aims to be as light and small as possible for a given output power, hence offering high gravimetric power density and high volumetric power density, respectively. The former one is usually termed the *specific power* in kilowatt per kilogram, whereas the latter one is loosely termed the *power density* in kilowatt per liter. For the EV powertrain, the specific power is more important as it directly affects the driving range of the EV, which is much shorter than that of the ICEV. The corresponding specific power depends on three main factors: the motor type, the cooling method, and the motor speed.

There are three major types of electric motors that are adopted by the EV powertrain, namely, the DC motor, the AC motor, and the switched reluctance (SR) motor (Chau, 2009). The AC motor includes the induction motor and the permanent magnet (PM) brushless motor

(Chau, Chan, and Liu, 2008). The DC motor needs to incorporate the commutator and carbon brushes, suffering from heavy weight and bulky size. The AC motor and the SR motor take the definite advantages of higher specific power and higher power density. Because of the use of high energy PM materials, the PM brushless motor offers the highest specific power. Figure 1 gives an indicative comparison of the specific power of different types of electric motors with respect to the motor power.

Basically, there are two types of cooling methods for the electric motor—air cooling and liquid cooling. The liquid cooling includes the water cooling and the oil cooling. The advantages of using water cooling over air cooling are its higher specific heat capacity and higher thermal conductivity. On the other hand, the oil cooling outperforms the water cooling due to the fact that oil offers higher boiling point than water so as to be able to cool parts with temperature higher than 100°C, and oil is an electrical insulator that can be in direct contact with electrical parts. When the electric motor is liquid cooled, the heat dissipation can be effectively removed from the motor. Hence, the corresponding electric loading and magnetic loading can be significantly increased; thus, even taking into account the additional weight of the cooling system, the overall specific power can be improved. Figure 2 gives an indicative comparison of the specific power of the electric motor using different cooling methods with respect to the motor power.

The mass and size of the electric motor are strongly influenced by the motor speed. In general, the higher the motor speed, the lighter the motor mass and the smaller the motor size are resulted. It is because of the fact that the motor speed depends on the rate of change of electromagnetic field in the electric motor, and the use of fast-changing field can reduce both the copper material for armature windings and the iron material for magnetic flux paths. Thus, the high speed motor design is particularly attractive for the EV powertrain. Even taking into account the additional reduction gear to scale down the motor speed to the wheel speed, the overall mass and size of the high speed motor plus its built-in reduction gear can be lighter and smaller than that of the low speed motor. Table 1 gives an indicative comparison of the mass and size of an induction motor at the same power under different speeds.

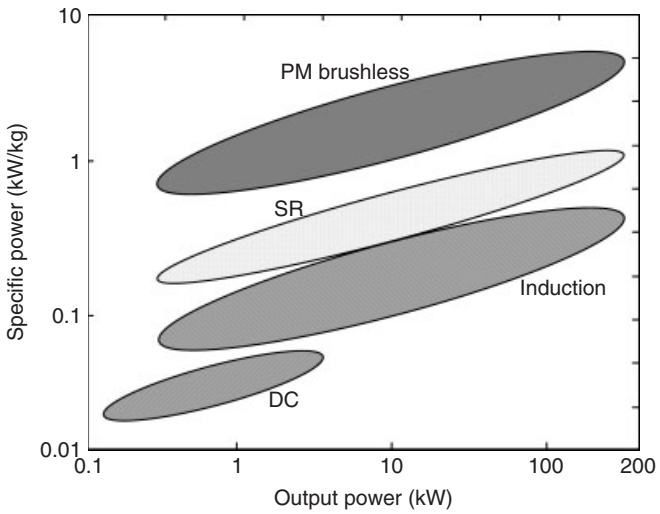


Figure 1. Comparison of specific power of electric motor using different types.

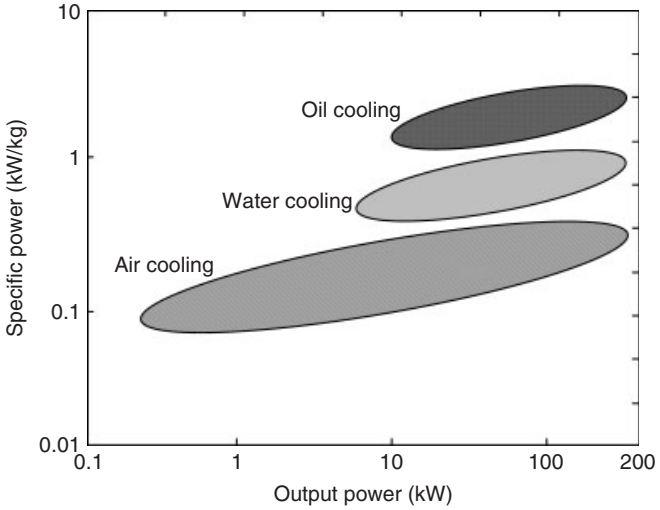


Figure 2. Comparison of specific power of electric motor using different cooling methods.

Table 1. Comparison of mass and size of induction motor using different speeds.

Speed (rpm)	Mass (kg)	Volume (L)
750	570	240
1000	415	165
1500	310	121
3000	270	92

3 FORCE AND TORQUE PARAMETERS

In general, the force and torque parameters of the EV powertrain are borrowed from the well-established definitions used for the ICEV powertrain (Fenton, 1996; Lucas, 1996; Newton, Steeds, and Garrett, 1996). Figure 3 shows the force components that the EV powertrain must provide for the vehicle to travel, which include the road load F_{load}

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

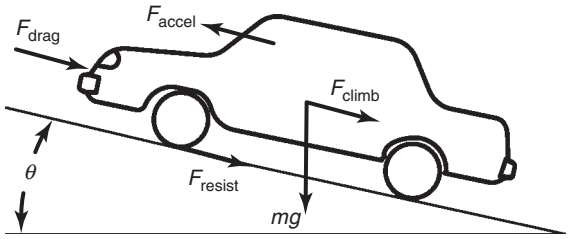


Figure 3. EV powertrain force.

and the acceleration force F_{accel}. This road load consists of three main components as described by Equation 1—the aerodynamic drag force F_{drag}, the rolling resistance force F_{resist}, and the climbing force F_{climb} (Unnewehr and Nasar, 1982):

$$F_{load} = F_{drag} + F_{resist} + F_{climb} \quad (1)$$

The aerodynamic drag force is due to the drag upon the vehicle body when moving through the air. Its composition is due to three aerodynamic effects: namely, the skin friction drag due to the air flow in the boundary layer; the induced drag due to the downwash of the trailing vortices behind the vehicle; and the normal pressure drag that is proportional to the vehicle frontal area and speed. In general, the skin friction drag and the induced drag are usually much smaller than the normal pressure drag. Thus, this aerodynamic drag force can be described by Equation 2 as:

$$F_{drag} = \frac{1}{2}\rho_{air} C_{drag} A_{front} (v + v_{wind})^2 \quad (2)$$

where C_{drag} is the aerodynamic drag coefficient (dimensionless), ρ_{air} is the air density in kilogram per cubic meter, A_{front} is the vehicle frontal area in square meter, v is the vehicle speed in meter per second, and v_{wind} is the head wind speed in meter per second. In general, ρ_{air} is taken as 1.23 kg/m^3 although it is dependent on the altitude. On the other hand, C_{drag} varies significantly, ranging from 0.2 to 1.5. For instance, a sport car has C_{drag} from 0.2 to 0.3, a sedan from 0.3 to 0.5, a van from 0.5 to 0.6, a bus from 0.6 to 0.7, and a truck from 0.8 to 1.5. Figure 4 shows the force and hence power requirements to overcome different aerodynamic drags (from 0.2 to 0.5 with step size 0.1) of a typical passenger car in which ρ_{air} is 1.23 kg/m^3, A_{front} is 2 m^2, and v_{wind} is zero. It can be observed that the aerodynamic drag squarely increases with the vehicle speed, and the corresponding power consumption becomes very significant at high speeds.

The rolling resistance force is due to the work of deformation on the tire and the road surface. The deformation on the tire heavily dominates the rolling resistance, whereas the deformation on the road surface is generally insignificant. Factors that affect the rolling resistance are the tire type, the tire pressure, the tire temperature, the vehicle speed, the tread thickness, the number of plies, the mix of the rubber, and the level of torque transmitted. Among them, the tire type and the tire pressure are relatively more dominant. This rolling resistance force can generally be described by Equation 3 as:

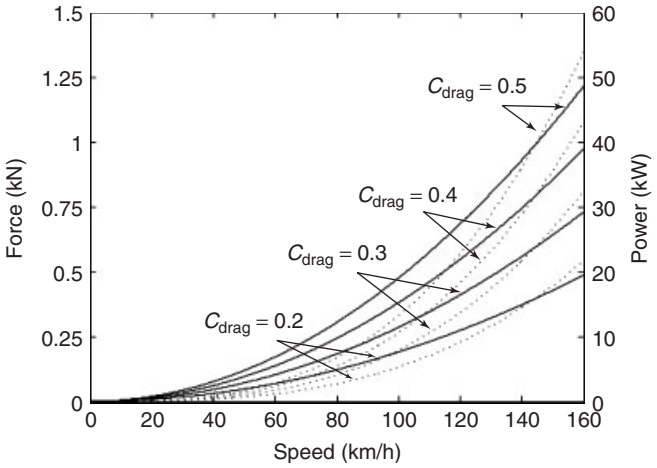


Figure 4. Force and power requirements to overcome different aerodynamic drags.

$$F_{resist} = mg\, C_{resist} \cos\theta \quad (3)$$

where m is the vehicle mass in kilogram, g is the gravitational acceleration equal to 9.81 m/s^2, C_{resist} is the rolling resistance coefficient (dimensionless), and θ is the angle of incline in degree or radian (Tabbache, Kheloui, and Benbouzid, 2010). In general, the C_{resist} of radial-ply tires is about 0.013, which is lower than that of cross-ply tires, about 0.018. It also varies inversely with the tire pressure—namely, the higher the tire pressure, the lower the C_{resist}. Figure 5 shows the force and hence power requirements to overcome different rolling resistances (from 0.005 to 0.02 with step size 0.005) of a typical passenger car in which m is 1000 kg and θ is zero degree. It can be observed that the rolling resistance is independent of the vehicle speed, and the corresponding power consumption is generally less significant than that of the aerodynamic drag especially during high speed operation.

The climbing force is simply the climbing resistance or downward force for a vehicle to climb up an incline. This force can be represented by Equation 4 as:

$$F_{climb} = mg \sin\theta \quad (4)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

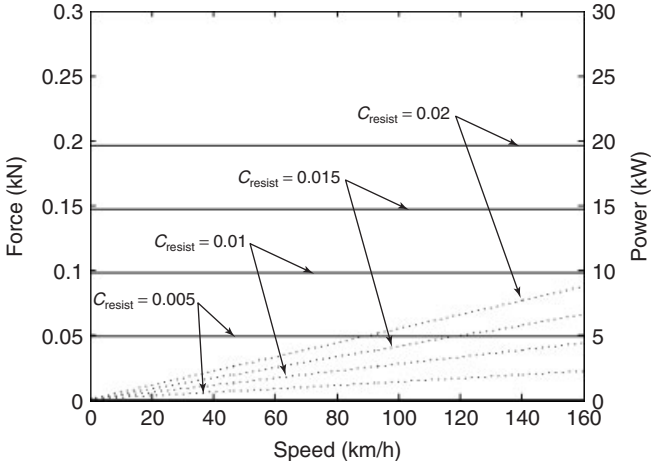


Figure 5. Force and power requirements to overcome different rolling resistances.

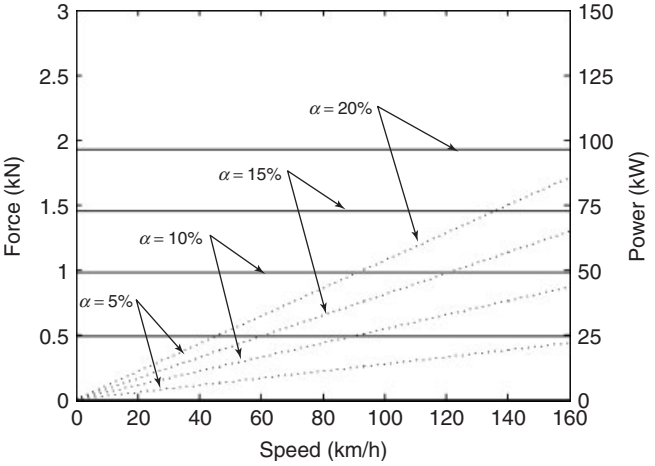


Figure 6. Force and power requirements to overcome different gradeabilities.

Usually, the incline is expressed as a percentage gradeability α, which is given by:

$$\alpha = \frac{H}{D} \times 100\% \tag{5}$$

where H is the vertical height of the road over the horizontal distance D. From Equation 5, θ and α are related by:

$$\theta = \tan^{-1}\left(\frac{\alpha}{100}\right) \tag{6}$$

For example, making use of Equation 6, the gradeability of 20% is equivalent to the angle of incline of 11.3°. The maximum gradeability denotes the maximum incline that a vehicle can climb at essentially zero speed. Figure 6 shows the force and hence power requirements of a typical passenger car to overcome different gradeabilities (from 5% to 20% with step size 5%) in which m is 1000 kg. It can be observed that the climbing force increases remarkably with the gradeability, and the corresponding power consumption is generally much higher than that of the aerodynamic drag and the rolling resistance. Notice that the climbing force will be negative when the vehicle is going download.

When the EV is accelerating or decelerating, the acceleration force F_{accel} can be expressed as:

$$F_{accel} = k_m ma \tag{7}$$

where a is the acceleration of the vehicle and k_m denotes a correction factor that there is an apparent increase in vehicle mass due to the inertia of rotational masses. Typically, k_m is about 1.05 which is equivalent to increase in the mass by 5%. When it is under deceleration or braking, a becomes negative. Figure 7 shows the force and hence power requirements of a typical passenger car to offer different acceleration rates (0–80 km/h from 6 to 12 s with step size 2 s) in which m is 1000 kg and k_m is 1.05. It can be observed that the acceleration force is essentially independent of the vehicle speed, and the corresponding power consumption becomes significant when high acceleration rate is demanded. Notice that the acceleration rate can also be represented by the required time from zero speed to 40 or 60 km/h.

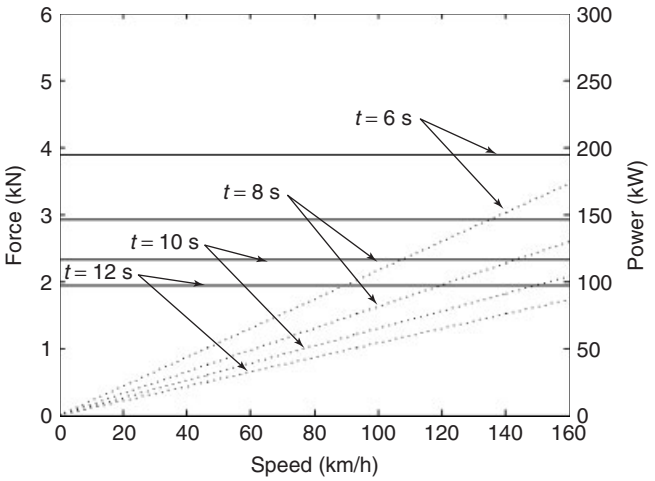


Figure 7. Force and power requirements to provide different acceleration rates.

By using Equations 1–4 and 7, the motive force F that the EV powertrain must provide at the wheel can be obtained as:

$$F = F_{load} + F_{accel} = (F_{drag} + F_{resist} + F_{climb}) + F_{accel} \tag{8}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

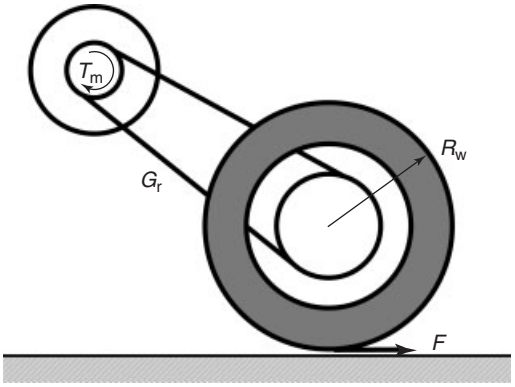


Figure 8. EV powertrain torque.

In general, the electric motor of the EV powertrain is rotational, rather than linear. The above linear motive force at the wheel needs to be translated as the rotational torque at the electric motor. Figure 8 shows the model to describe the transmission between the electric motor and the wheel. Thus, the motor torque T_m can be expressed in terms of Equation 8 as:

$$T_m = \frac{F\ R_w}{G_r} \tag{9}$$

where R_w is the radius of the wheel and G_r is the overall gear ratio of the EV powertrain from the electric motor to the wheel axle.

4 POWER AND ENERGY PARAMETERS

In transportation, the unit of energy is usually in kilowatt-hour (kWh) which is a non-SI unit rather than the SI unit joule (J), because the latter is too small for such application. To assess the energy consumption of a vehicle, the energy per unit distance in kilowatt-hour per kilometer is generally used. This unit can be applied to both the ICEV and the EV. However, an ICEV driver generally has no idea about the kWh and prefers a physical unit of fuel volume such as liter. So, the energy consumption unit of the ICEV is usually in liter per kilometer. On the other hand, the energy parameter of the ICEV can also be expressed as the distance per unit volume of fuel, the so-called fuel economy. The corresponding SI unit is in kilometer per liter. For the EV, the original energy consumption unit in kilowatt-hour per kilometer becomes suitable because the fuel for recharging the battery is electricity which can directly be measured in kilowatt-hour. The corresponding fuel economy is expressed in kilometer per kilowatt-hour. When the EV is fed by fuel cells, the corresponding fuel may be compressed gaseous hydrogen, liquid hydrogen, or even liquid methanol; hence, the energy consumption unit in liter per kilometer and the fuel economy unit in kilometer per liter become applicable.

4.1 Efficiency parameters

The energy efficiency η_e is the ratio of energy output E_{out} to energy input E_{in} as defined in Equation 10, whereas the power efficiency η_p is the ratio of power output P_{out} to power input P_{in} as defined in Equation 11:

$$\eta_e = \frac{E_{out}}{E_{in}} \tag{10}$$

$$\eta_p = \frac{P_{out}}{P_{in}} \tag{11}$$

For industrial operation, these two efficiencies may not be necessarily distinguishable. On the contrary, for the EV powertrain, there is a significant difference between these two efficiencies because the power efficiency varies continually during the operation of the vehicle. Thus, it is necessary to delineate the power efficiency associated with the speed and torque conditions. Instead of using a particular operating point such as the rated power at the rated torque and the rated speed to describe the power efficiency of the EV powertrain, an efficiency map is generally adopted in which the power efficiency, loosely termed the *efficiency*, is plotted as various contours on the torque–speed plane. Consequently, the energy efficiency can be deduced by summing powers over a given time period.

Among the three major types of electric motors for the EV powertrain, namely the DC motor, the induction motor, and the PM brushless motor, the DC motor has the worst efficiency map, the PM brushless motor has the best one, whereas the induction motor has the one probably in between them. Figures 9–11 show their typical efficiency maps for the EV powertrain in which the limits are under continuous load. It can be observed that the PM brushless motor offers not only the highest efficiency but also high efficiencies over a wide range of speed and torque.

Regenerative braking is a definite advantage of the EV over the ICEV. During braking, the motor operates in the regenerative mode, which converts the reduction in kinetic energy during braking into electrical energy, hence recharging the battery. On average, the amount of convertible energy is only up to 40% (Larminie and Lowry, 2003). Assuming that the overall efficiency of the EV powertrain is about 70%, the amount of energy actually stored in the battery is only up to 28%. This is known as the *regenerative braking efficiency*. Similar to the motoring case, in order to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

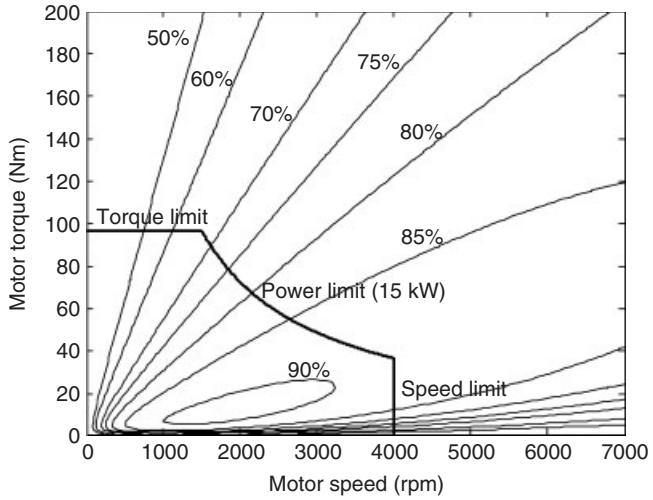


Figure 9. Typical efficiency map of DC motor.

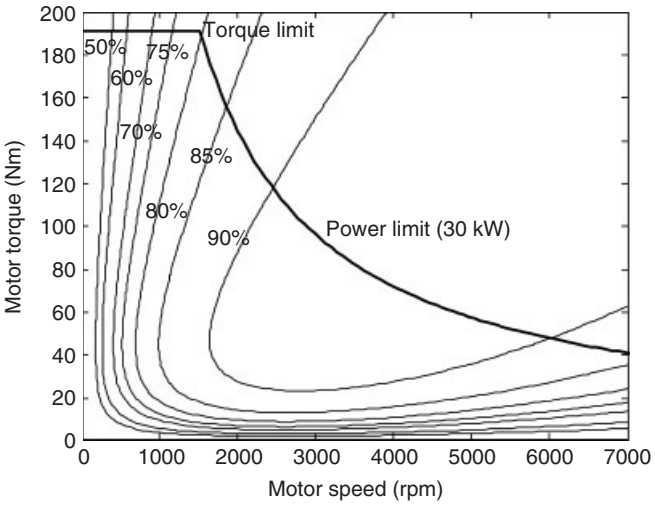


Figure 11. Typical efficiency map of PM brushless motor.

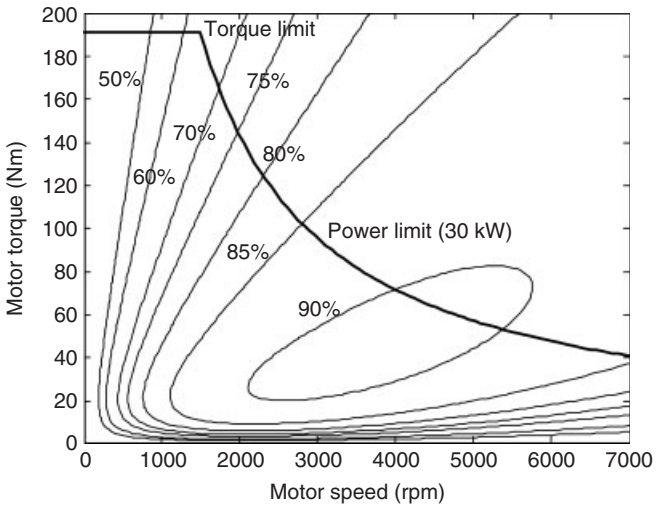


Figure 10. Typical efficiency map of induction motor.

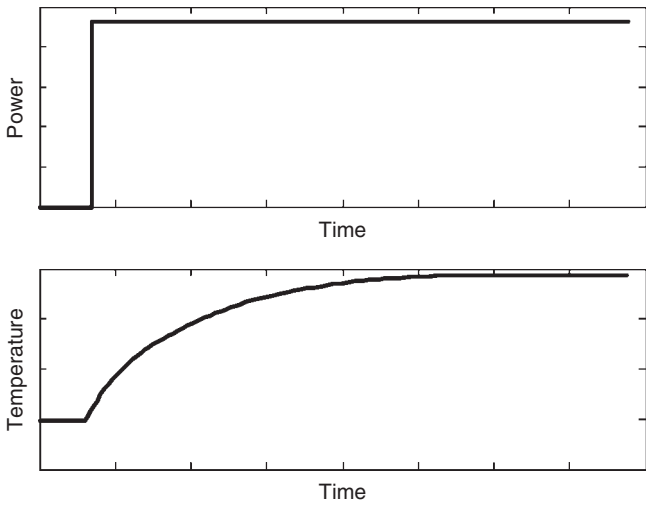


Figure 12. Continuous load of electric motor.

accurately depict the value at different torques and speeds, a regenerative braking efficiency map should be adopted.

4.2 Load parameters

The power rating of electric motors depends on the load condition, which in turn depends on the specific application. For the EV powertrain, the corresponding power rating is generally specified by three types of load duty: continuous, intermittent, and temporary. For the same motor, the power ratings under various load duties are different. Typically, the power rating under the continuous load is the lowest among the three types of load duties, whereas the rating under the temporary load is the highest.

The continuous load of the electric motor means that the specified power can be maintained continuously or with a practical duration that is much longer than the thermal time constant. As depicted in Figure 12, it can be seen that the temperature of the electric motor rises from the initial temperature and reaches the thermal equilibrium within the practical duration of the continuous load.

The temporary load of the electric motor means that the specified power under the specified duration does not cause the motor temperature to reach the thermal equilibrium, and then with a subsequent duration that lasts until the motor temperature is no more than 2 K from the initial temperature

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

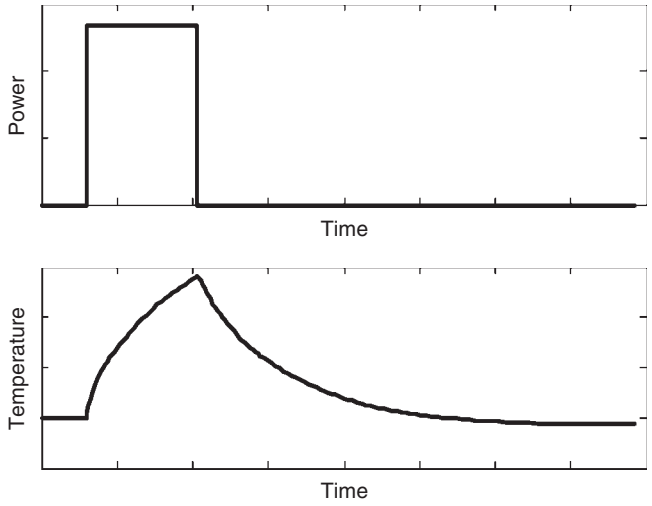


Figure 13. Temporary load of electric motor.

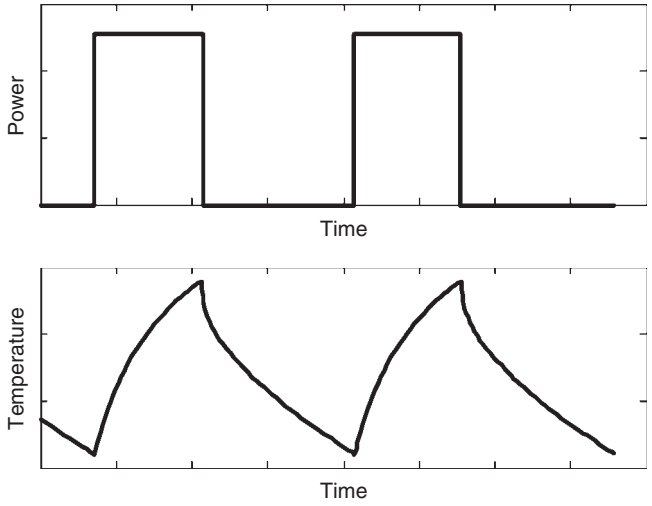


Figure 14. Intermittent load of electric motor.

(Rockwell Automation, 1996). As depicted in Figure 13, it can be seen that the motor temperature rises abruptly in the presence of high temporary power and then gradually cools down to almost the initial temperature. Typically, the temporary power of the electric motor is quoted together with the operating time period of 10, 30, 60, or 90 min. Compared to the continuous power rating, the temporary power rating is usually higher. Moreover, the shorter the operating time period, the higher the temporary power of the electric motor.

The intermittent load of the electric motor means that the specified power under the specified duration does not cause the motor temperature to reach the thermal equilibrium, and then with a subsequent duration which lasts until the motor temperature returns to the original cyclic temperature. As depicted in Figure 14, it can be seen that the motor temperature cyclically swings between two temperature limits. The intermittent power of the electric motor is quoted together with the load period and the cyclic period, or with the percentage duty cycle and the cyclic period. Typically, the percentage duty cycle quoted is 15%, 25%, 40%, or 60%. Compared to the continuous power rating, the intermittent power rating is usually higher. On the contrary, as the electric motor does not have sufficient time to gradually cool down to the initial temperature, the intermittent power rating is generally lower than the temporary power rating.

Figure 15 shows the torque–speed capabilities of a typical induction motor under different loads—continuous, temporary, and intermittent. Firstly, the continuous capability is the most conservative one which is particularly essential for the powertrain to provide the EV long-term hill climbing. Secondly, the temporary capability is the most aggressive one which is particularly worthwhile for the powertrain to describe the overtaking capability of the EV. Thirdly, the intermittent capability is the most useful one for the powertrain to satisfy the operation of urban driving which is essentially in periodic nature.

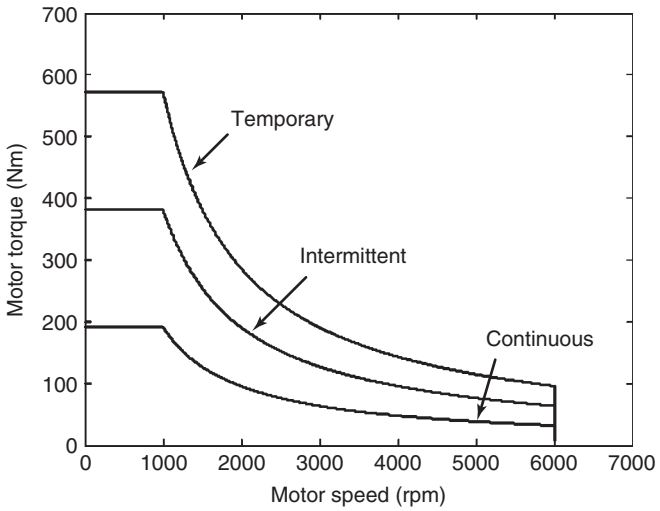


Figure 15. Torque–speed capabilities at different loads of electric motor.

5 DRIVING-CYCLE PERFORMANCES

For the EV, the driving range per charge is the most important parameter to assess the vehicle performance. The reason is simply that the range of a 1500-cc ICEV can be about 500 km per refuel, whereas the range of an EV is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

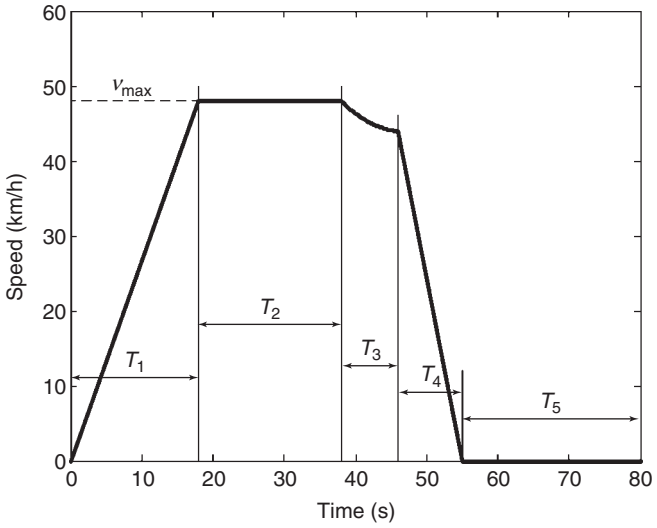


Figure 16. SAE J227a cycles.

only about 100–200 km per charge, which may not satisfy a normal driver's expectation. There are two main types of assessment that can be performed. The first type is based on the constant-speed test, whereas the second type is based on the driving-cycle test. These two tests are not only to assess the driving range per charge but also to evaluate the performance of the EV powertrain.

The constant-speed test normally assumes that the EV operates at certain constant speed such as 40 or 60 km/h, without considering any start, stop, acceleration, or deceleration. In reality, no vehicle is really driven at constant speed. Because of this ideal condition, the driving range per charge and the performance of the EV powertrain are too optimistic, if not unrealistic. The driving-cycle test is developed, which can provide a more realistic and practical condition to assess the EV and its powertrain. So far, there is no universal driving cycle that has been accepted by different automakers in different countries. In fact, different countries or even cities may have their own driving cycles for testing.

A well-known driving cycle that was specifically developed for the EV in the 1970s is the Society of Automotive Engineers (SAE) J227a driving cycle. It has four versions with different power requirements and maximum speed requirements, which are named as the J227a-A, -B, -C, and -D cycles. Figure 16 shows the general speed profile of this SAE J227a driving cycle. It consists of the acceleration stage, the constant-speed stage, the coasting stage, and the braking stage. Notice that the motive force is set to zero during the coasting stage. The details of all four versions are summarized in Table 2. The most commonly used one is the SAE J227a-C cycle, which is particularly useful for the design of the EV and its powertrain.

The federal urban driving schedule (FUDS) is the most common driving cycle used in the United States. It is also called the *federal test procedure 72* (*FTP-72*). It is a standard driving cycle lasting 1500 s, and for each second, there is an individual speed as depicted in Figure 17. This FUDS was developed originally to evaluate the noxious emissions of the ICEV and was based on a profile derived from the statistical flow of traffic patterns in Los Angeles. It simulates an urban route of 12.07 km with frequent starts and stops, the maximum speed of 91.2 km/h, and an average speed of 31.5 km/h. Thus, it is widely used to

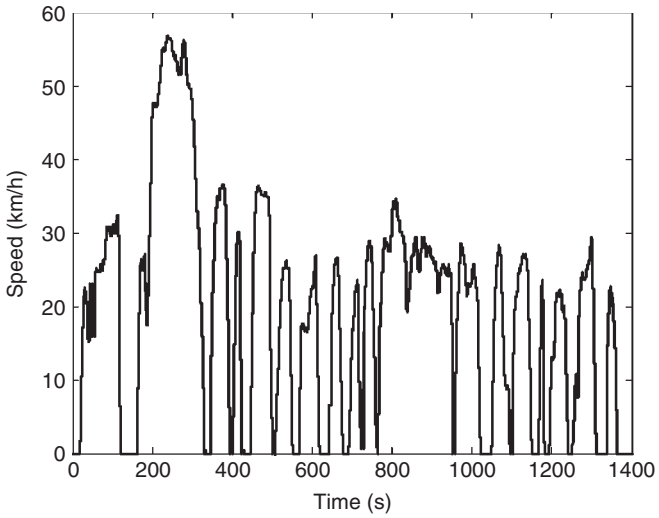


Figure 17. Federal urban driving schedule.

Table 2. Parameters of SAE J227a cycles.

	Cycle A	Cycle B	Cycle C	Cycle D
Maximum speed v_{max} (km/h)	16	32	48	72
Acceleration time T_1 (s)	4	19	18	28
Cruising time T_2 (s)	0	19	20	50
Coasting time T_3 (s)	2	4	8	10
Braking time T_4 (s)	3	5	9	9
Idling time T_5 (s)	30	25	25	25

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

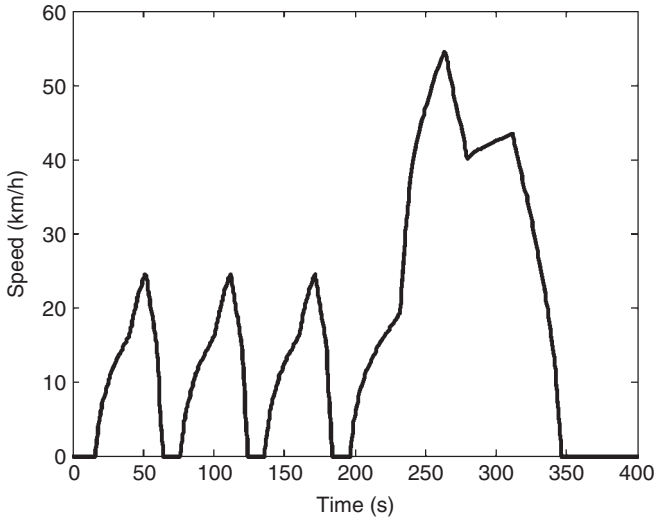


Figure 18. Simplified federal urban driving schedule.

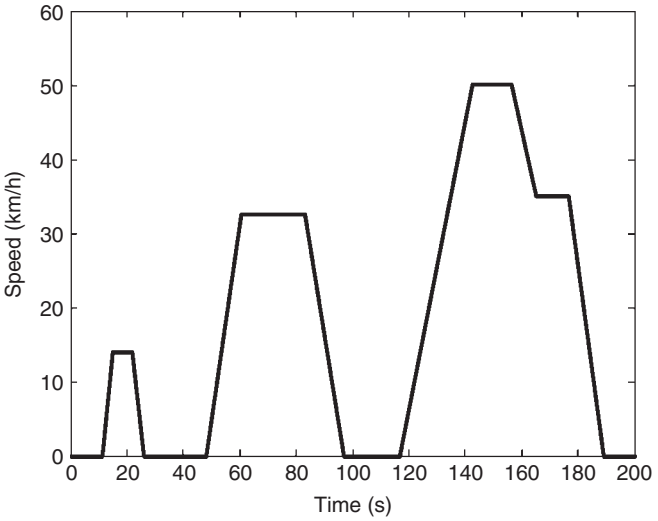


Figure 20. ECE-15.

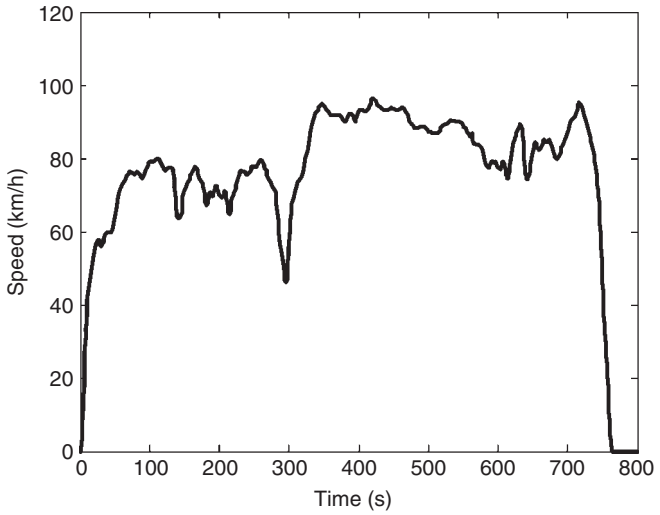


Figure 19. Federal highway driving schedule.

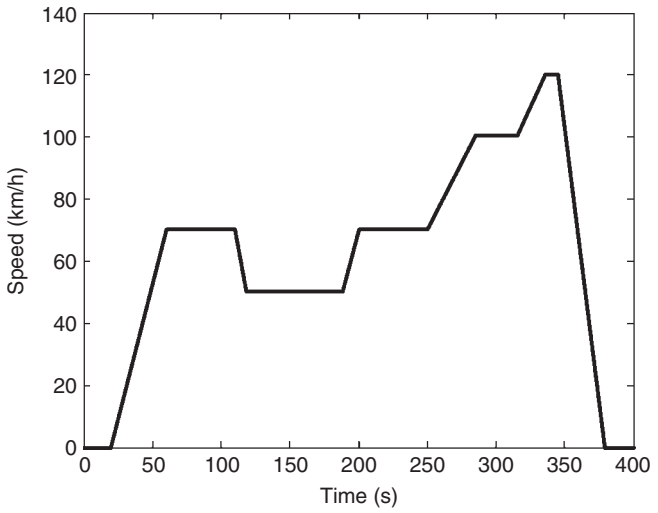


Figure 21. Extra-urban driving cycle.

evaluate the fuel economy of urban or city driving. There are some simplified versions of the FUDS (Larminie and Lowry, 2003), such as the simplified federal urban driving schedule (SFUDS) as depicted in Figure 18. This SFUDS only lasts 360 s and has only 360 data points, while offers the same average speed, the same proportion of stationary, the same maximum acceleration, and the same maximum braking. Thus, the SFUDS takes the advantage that it is less computational demanding than the FUDS while giving a similar result when it is used for simulating the driving range. On the other hand, the federal highway driving schedule (FHDS) is the driving cycle developed to typify rural or cross-country driving in the United States. As depicted in Figure 19, the FHDS lasts 765 s and maintains nonstop high speed operation for highway driving.

In the European Union, the vehicle is commonly tested using two kinds of driving cycles. The first one is the urban driving cycle known as the *ECE-15* (Economic Commission for Europe), which was introduced in 1999. It was devised to represent city driving conditions in Paris or Rome. As shown in Figure 20, it simulates an urban trip of 4052 m at an average speed of 18.7 km/h and the maximum speed of 50 km/h. It is particularly useful for testing urban driving performance of the EV. The second one is the extra-urban driving cycle (EUDC). As shown in Figure 21, it simulates a suburban trip that lasts 400 s at an average

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

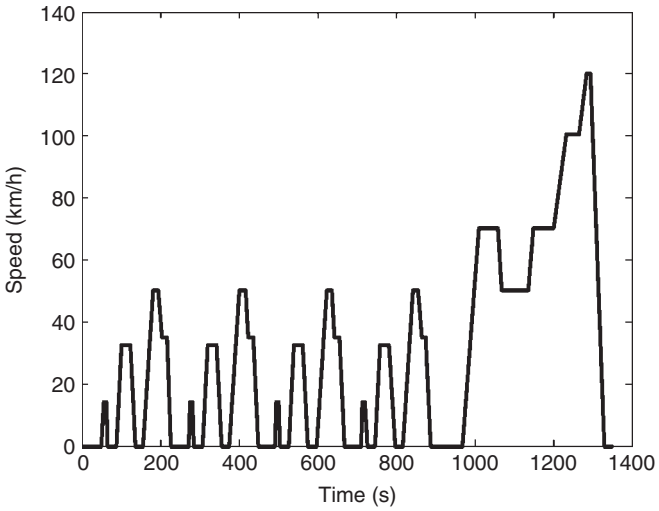


Figure 22. New European driving cycle.

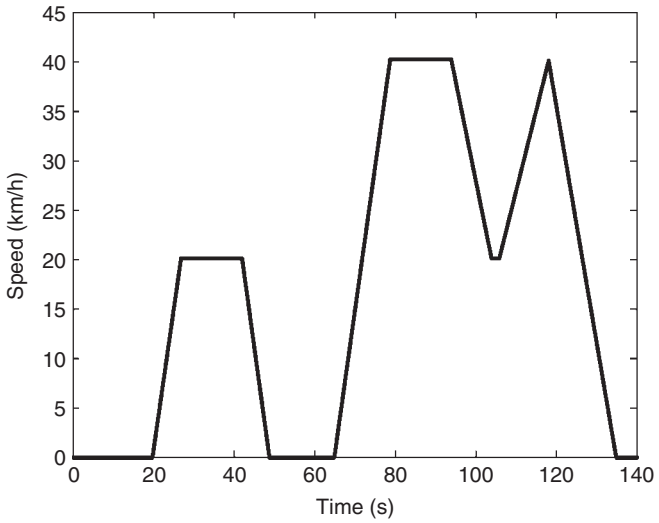


Figure 23. Japan 10 mode.

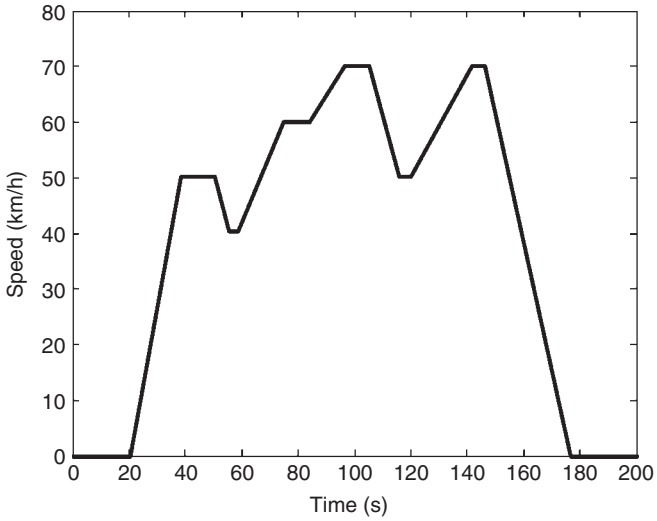


Figure 24. Japan 15 mode.

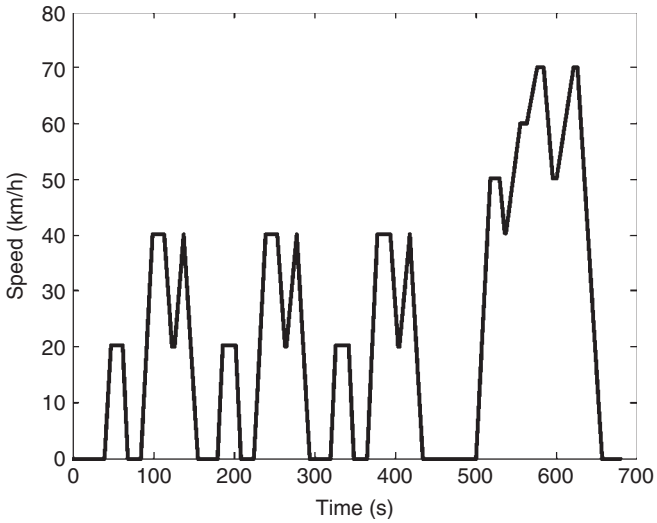


Figure 25. Japan 10–15 mode.

speed of 62.6 km/h and the maximum speed of 120 km/h. Furthermore, the new European driving cycle (NEDC) is the driving cycle that incorporates four repeated ECE-15 driving cycles and an EUDC. As shown in Figure 22, the NEDC is supposed to represent the typical usage of a vehicle in Europe.

In Asia, the most widely used standard driving cycles are developed in Japan. As depicted in Figure 23, the Japan 10 mode represents an urban driving cycle that covers a distance of 0.664 km at an average speed of 17.7 km/h, and lasts 135 s with the maximum speed of 40 km/h. On the other hand, as depicted in Figure 24, the Japan 15 mode denotes a combination of the urban and extra-urban route with the maximum speed of 70 km/h. Similar to the NEDC for the European Union, the Japan 10–15 mode incorporates five cycles—first the 15 mode, then three times the 10 mode, and at last again the 15 mode. The assessment is normally based on the last four segments (three 10 mode + one 15 mode) as depicted in Figure 25, in which the distance is 4.16 km, the average speed is 22.7 km/h, and the duration is 660 s. As a regulation, the driving range of EVs in Japan has to be evaluated by using this driving cycle.

In addition to the above homologated driving cycles for EVs, it is also important to take into account vehicle use

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

in realistic driving conditions. It is particularly important that the energy consumption of EVs greatly depends on the driving profiles and conditions. EV manufacturers are actually facing with the challenge of convincing their potential customers over the limited driving range per charge. Thus, it is necessary to consider other real-world driving cycles to assess the EV performances, including the urban, extra-urban, and highway driving conditions. There are many representative cycles for real-world driving conditions (Barlow *et al*., 2009). Among them, the Artemis driving cycles that were developed within the European 5th Framework project, Assessment and Reliability of Transport Emission Models and Inventory Systems (ARTEMIS), are widely accepted in Europe, which well describe various real-world driving conditions, such as the Artemis Traffic Jam for congested urban traffic, the Artemis Urban for urban traffic, the Artemis Road for road conditions at extra-urban traffic, and the Artemis Motorway for highway traffic (Zaccardi and Le Berr, 2012). Considering this variety of driving cycles, it is not straightforward to select the most relevant cycles.

The EV powertrain needs to satisfy both the torque demand and the power demand of the EV imposed by the driving cycle. For instance, based on the driving cycle with speed versus time, the required motive force and hence the required motor torque can be deduced by using Equations 2, 3, 7, and 9:

$$T_{\mathrm{m}} = \frac{R_{\mathrm{w}}}{G_{\mathrm{r}}}(F_{\mathrm{drag}} + F_{\mathrm{resist}} + F_{\mathrm{accel}}) \tag{12}$$

Consequently, the required average power P_{av} that the EV powertrain needs to produce is given by:

$$P_{\mathrm{av}} = \frac{1}{T}\int_0^T (F_{\mathrm{drag}} + F_{\mathrm{resist}} + F_{\mathrm{accel}})\, v \, \mathrm{d}t \tag{13}$$

where T is the duration of the driving cycle.

ACKNOWLEDGMENTS

The author would like to express heartfelt thanks to all group members of the International Research Center for Electric Vehicles for their contributions to this chapter. He expresses his indebtedness to Joan and Aten for their support all the way.

RELATED ARTICLES

Overview of Electric, Hybrid and Fuel Cell Vehicles
EV Powertrain Configurations
General Requirement of Traction Motor Drives
DC Motor Drives
Induction Motor Drives
Permanent Magnet Brushless Motor Drives
Switched Reluctance Motor Drives
Future Direction of Traction Motor Drives

REFERENCES

Barlow, T.J., Latham, S., McCrae, I.S., and Boulder, P.G. (2009) *A Reference Book of Driving Cycles for Use in the Measurement of Road Vehicle Emissions*. Project Report 354, TRL Limited, Wokingham.

Bosch, R. (2007) *Bosch Automotive Handbook*, 7th edn, John Wiley & Sons, Plochingen.

Chan, C.C. and Chau, K.T. (2001) *Modern Electric Vehicle Technology*, Oxford University Press, Oxford.

Chau, K.T. (2009) Electric motor drives for battery, hybrid and fuel cell vehicles in *Electric Vehicles: Technology, Research and Development* (ed. G.B. Raines), Nova Science Publishers, New York, pp. 1–40.

Chau, K.T. and Wang, Z. (2005) Overview of power electronic drives for electric vehicles *HAIT Journal of Science and Engineering—B: Applied Sciences and Engineering*, **2** (5–6), 737–761.

Chau, K.T., Chan, C.C., and Liu, C. (2008) Overview of permanent-magnet brushless drives for electric and hybrid electric vehicles *IEEE Transactions on Industrial Electronics*, **55** (6), 2246–2257.

Duffy, J.E., Stockel, M.T., and Stockel, M.W. (1988) *Automotive Mechanics Fundamentals: How and Why of the Design, Construction, and Operation of Modern Automotive Systems and Units*, Gregory's Automotive Publications, Sydney.

Fenton, J. (1996) *Handbook of Vehicle Design Analysis*, Society of Automotive Engineers, Warrendale.

Larminie, J. and Lowry, J. (2003) *Electric Vehicle Technology Explained*, John Wiley & Sons, Ltd, Chichester.

Lucas, G.G. (1996) *Road Vehicle Performance: Methods of Measurement and Calculation*, Gorden and Breach, New York.

Newton, K., Steeds, W., and Garrett, T.K. (1996) *The Motor Vehicle*, Butterworth-Heinemann, Oxford.

Rockwell Automation (1996) *Application Basics of Operation of Three-Phase Induction Motors*, Sprecher + Schuh AG Rockwell Automation, Aarau.

Tabbache, B., Kheloui, A., and Benbouzid, M.E.H. (2010) Design and Control of the Induction Motor Propulsion of an Electric Vehicle. *Proceedings of IEEE Vehicle Power and Propulsion Conference*, Lille, pp. 1–6.

Unnewehr, L.E. and Nasar, S.A. (1982) *Electric Vehicle Technology*, John Wiley & Sons, Inc, New York.

Zaccardi, J.-M. and Le Berr, F. (2012) Analysis and choice of representative drive cycles for light duty vehicles - case study for electric vehicles *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, **227** (4), 605–616.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto050
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Basic Consideration

Y.S. Wong[1], C.C. Chan[2], and Samir Nazir[1]

[1]*National University of Singapore, Singapore*
[2]*The University of Hong Kong, Pokfulam, Hong Kong*

1 Introduction 1
2 Features of Engine and Electrical Powertrains 2
3 Drivetrain Topologies of HEVs 7
4 Functionality of HEVs 12
5 Conclusions 15
Related Articles 15
References 15

1 INTRODUCTION

Electric vehicles (EVs) offer high energy efficiency, allow diversification of energy resources, enable load equalization of power systems, deliver zero local and minimal global exhaust emissions, and operate quietly. However, there are two major barriers hindering the commercialization of EVs, namely, short driving range and high upfront cost. These barriers cannot be easily solved by the available EV energy source technologies, including batteries, fuel cells, capacitors, and flywheels. The hybrid electric vehicle (HEV), incorporating the engine and electric motor, has been introduced as an interim solution.

The HEV greatly extends the limited EV driving range by two to four times and offers rapid refueling by liquid gasoline or diesel. Moreover, the HEV requires no change in the energy supply infrastructure. The HEV cannot deliver zero-emission driving such as the battery electric vehicle (BEV) does and has a more complex architecture. Nevertheless, the HEV offers more efficiency and correspondingly lower emissions than the internal combustion engine vehicle (ICEV) while having comparable driving range. With the assist of electric motors, the engine in an HEV can operate in its most efficient mode, yielding low emissions and low fuel consumption. Some HEVs may also be purposely operated as an EV for a short period of time in the zero-emission zone. The HEV is not only an interim solution for implementation of zero-emission vehicles but also a practical solution for commercialization of super ultralow emission vehicles.

1.1 Definition of HEV

The demand for vehicles with substantially higher fuel economy and lower vehicular emissions has motivated the development of HEVs, fuel cell electric vehicles (FCEVs), and BEVs for a number of years. The first production HEV available to the public was the Toyota Prius. The Prius was sold successfully in Japan since late 1997 and in the United States since 1999. Nowadays, HEVs are taking center stage, whereas BEVs are used in some niche areas where shorter distances are traveled. The available HEV models encompass light-duty cars and trucks, sports utility vehicles, delivery trucks, transit buses, and line-haul trucks.

The available definition of HEV is not clear. As proposed by Technical Committee 69 (Electric Road Vehicles) of the International Electrotechnical Commission, an HEV is a vehicle in which propulsion energy is available from two or more kinds or types of energy stores, sources, or converters, and at least one of them can deliver electrical energy. On the basis of this general definition, there are many types of HEVs, such as the engine and battery, battery and fuel cell,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

battery and capacitor, battery and flywheel, and battery and battery hybrids. However, the above-mentioned definition is not well accepted. To avoid confusing readers, specialists prefer using the HEV to represent a vehicle only adopting the energy source combination of an engine and battery. For example, a battery and fuel cell HEV is simply referred to as a *fuel cell EV*. The term HEV in this chapter refers only to the vehicle adopting both an engine and electric motor in the drivetrain, whereas the petroleum fuel (gasoline and diesel) and battery both are the energy source (Kabza, 2009; Chau and Wong, 2002).

The basic principle of HEV design is the coordination of the electric propulsion system and the internal combustion engine (ICE) system. An HEV is designed to meet the performance, including acceleration, gradeability, and minimum range, of a comparable baseline ICE vehicle (Chan *et al*., 2009).

This chapter gives an overview of features of ICE and electrical powertrains in an HEV and the main functionality of HEVs. Consideration of drivetrain design and operating models of HEVs are discussed to elaborate novelties of different HEVs.

1.2 History of hybrid electric vehicles

In 1801, Richard Trevithick built a steam-powered carriage, opening the era of horseless transportation. The first battery-powered electric bicycle was built by Thomas Davenport in 1834. It was powered by a nonrechargeable battery and used on a short track. In 1838, Robert Davidson built a nonrechargeable battery-powered electric locomotive. After the invention of the lead-acid battery in 1859, David Salomons built a rechargeable battery-powered electric vehicle in 1874. The first gasoline-powered ICE vehicle was built in 1885 and the first HEV was presented by J. Lohner and F. Porsche in 1901 (Chan and Chau, 2001). By 1920, Henry Ford's assembly line and the arrival of the self-starting petrol engine resulted in the rapid decline of hybrid cars (Chan, 2007).

The ICE vehicle outperformed the EV and HEV in the automotive century because there was no high performance battery for EVs and HEVs to overcome four major barriers to commercialization, short driving range, long charge time, long recharge time, and high lifecycle cost (Chan and Chau, 2001). The rekindling of interests in EVs started at the outbreak of energy crisis and oil shortage in the 1970s. The actual revival of EVs was due to the ever-increasing concerns on energy conservation and environmental protection throughout the world in the 1990s.

The first production HEV was commercialized in 1997. The world's first modern hybrid vehicle was sold in Japan in 1997, the Toyota Prius. In 1999, the Honda Insight became the first hybrid vehicle sold in the United States. Combined with the release of the Honda Civic Hybrid, vehicles that offered some of the benefits of battery electric vehicles to conventional ICEs were available to the public. Since then, many of the major automakers have offered hybrid vehicles in their lineup (Chan, 2007). Development of the plug-in hybrid electric vehicle (PHEV) accelerated in 2006 because of its potential to reduce oil dependence. The PHEV can deliver pure electric operations by the motor and batteries and can also be recharged from the power grid such as a BEV. The PHEV operates in pure electric mode when the battery's state of charge (SoC) is high and switches to HEV mode when the SoC is lower than a threshold value at 10–50% dependent on specification and battery capacity.

2 FEATURES OF ENGINE AND ELECTRICAL POWERTRAINS

The ICE has relatively low thermal and mechanical efficiencies, and much of the energy of the fuel is dissipated as heat. Figure 1 illustrates that in gasoline-powered vehicles about 62.4% of the fuel's energy is lost in the ICE (USDOE, 2011a; Taylor, 1998). The top three losses, accounted for 85.4% of total energy dissipation, are engine losses (62.4%), engine idling loss (17.2%), and energy dissipation in acceleration and braking (5.8%).

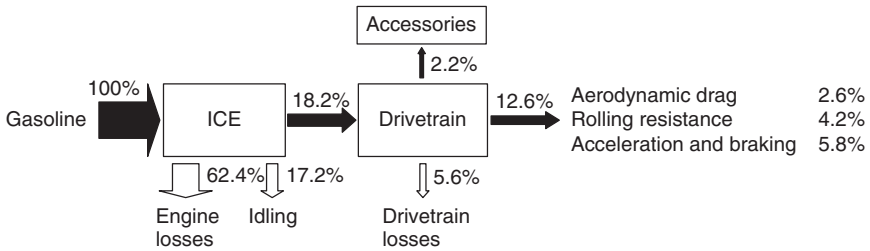


Figure 1. Energy flow in an ICE car.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

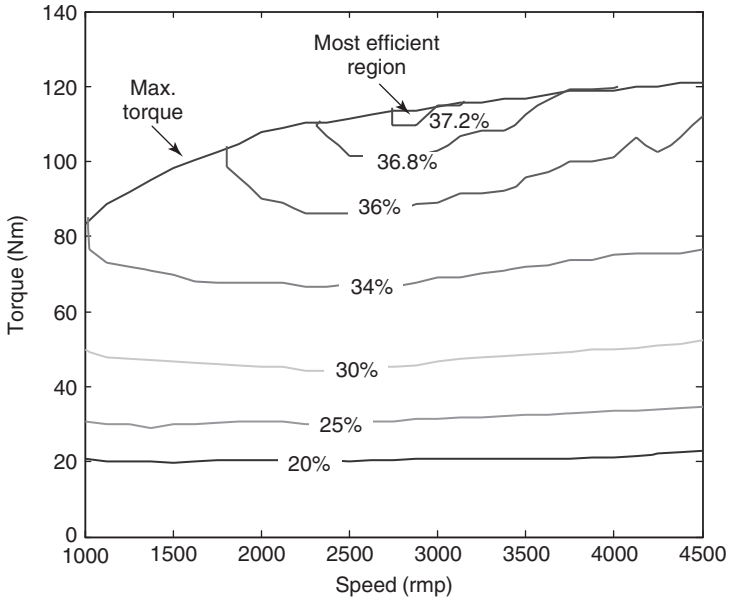


Figure 2. Engine performance and efficiency.

2.1 Efficiencies of ICE and electric motor

The ICE is intrinsically very inefficient at converting the fuel's chemical energy to mechanical energy, losing energy to engine friction, pumping air into and out of the engine, and wasted heat. Figure 2 shows efficiency contours of a typical gasoline engine. The thermal efficiency of this engine is low with a maximum efficiency of 37.2% and it can only operate in the most efficient area occasionally. In addition, energy is lost during idling. The fuel economy of ICEVs in city driving is lower than that in highway driving.

Technologies such as using lighter materials in the powertrain, reducing the vehicle exterior's drag coefficient, optimizing the engine cooling system, low rolling resistance tires, lower friction lubricants, locking torque converters in automatic transmissions to reduce slip and power losses, continuously variable transmissions (CVTs), overdrive gears, engine combustion optimization, and cylinder deactivation can reduce the engine losses. Moreover, diesel engines have inherently lower losses and are generally one-third more efficient than their gasoline counterparts. This is largely due to the higher compression ratios of diesel engines and higher density of diesel fuel. Diesel engines typically have lower power to weight ratios than comparable spark ignition engines. Recent advances in diesel technologies and fuels are making diesels very competitive.

Addition of an electric motor boosts the vehicle's overall efficiency to higher than that of the diesel engine. The electric motor does not consume energy at standstill and it has higher efficiency than the ICE. Figure 3 shows the efficiency contour profiles of a typical induction motor for vehicular applications. Comparing with the efficiency contour profiles of a typical gasoline engine in Figure 2, the electric motor can achieve higher efficiencies. The electric motor also has maximum torque at zero and low speeds and it allows bidirectional energy flow such that mechanical energy can be converted into electrical energy and stored in the traction batteries by regenerative braking. Hence, the hybridization of electric motors and the ICE in an HEV can significantly boost efficiency of the vehicle drivetrain and thus reduce oil consumption and emissions.

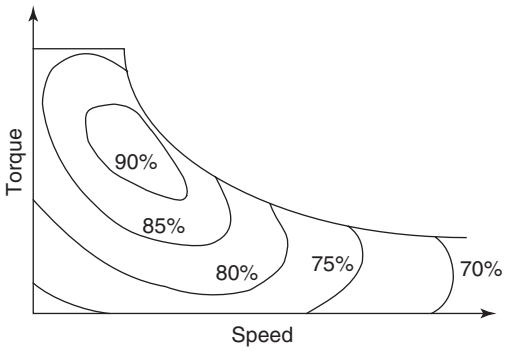


Figure 3. Motor performance and efficiency.

2.2 Regenerative braking

Regenerative braking is a unique feature of an electrical powertrain that allows the motor drive to convert a vehicle's

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

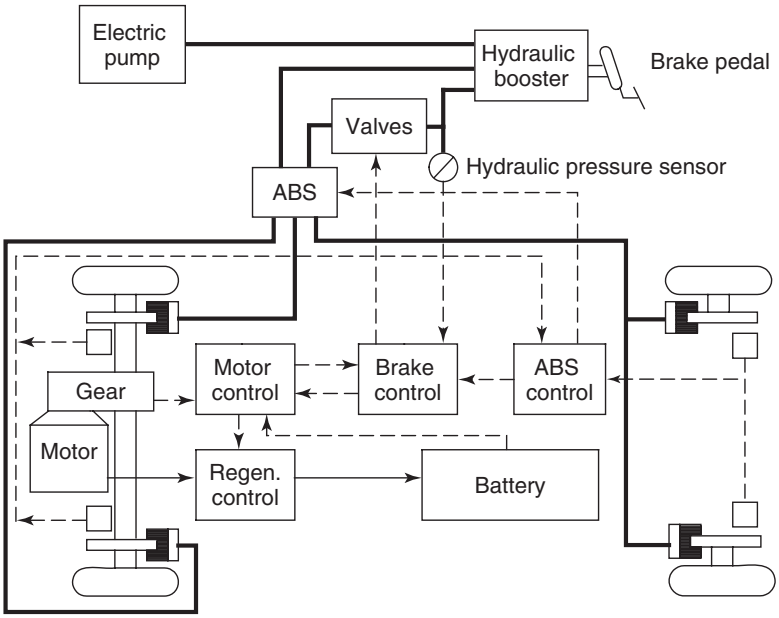


Figure 4. Configuration of a regenerative-hydraulic braking system in the electrical drivetrain.

kinetic energy to electrical energy during braking and in replacing engine braking. The converted electrical energy is stored in receptive energy sources such as batteries or ultracapacitors to supplement the ICE or extend all-electric driving range. If the rechargeable sources are fully charged, regenerative braking can no longer be applied and the vehicle can be slowed by the conventional friction braking system.

In general, the regenerative braking system is activated when the vehicle throttles off deceleration at highways or its brake pedal is pressed during braking. The total braking torque is the sum of the regenerative braking torque and the friction braking torque. The control of their distributions aims to provide the driver the same braking feel as that in conventional ICEVs while maintaining maximum regenerative braking.

In order to realize the above-mentioned requirements, a regenerative-hydraulic hybrid braking system can be configured as shown in Figure 4. By employing the electric pump, the hydraulic booster produces the desired hydraulic braking pressure that is activated by the driver's brake pedal operation. The brake control is used to collaborate with the motor control to produce the regenerative braking torque as well as the desired hydraulic braking torque on the front and rear wheels. During regenerative braking, the regenerative control recovers the kinetic energy for battery charging. The antilock braking system (ABS) control and its valves are the same as that in conventional ICEVs.

2.3 Power boost

HEVs allow both the engine and the electric motor to supplement each other in delivering power to drive the vehicle. The engine and electric motor can be coupled to the drive shaft of the wheel mechanically or electrically, such that the propulsion power may be supplied by the engine, the motor, or both. If the HEV is propelled by both the engine and the motor, the motor assists the engine during vehicle acceleration for maximization of engine fuel economy by shifting engine's torque and speed to the most efficient region as shown in Figure 2.

2.4 Energy saving potential

The electrical drivetrain boosts an HEV's fuel economy in the following ways:

- assist vehicle launch and acceleration when the vehicle is propelled by both the engine and the motor;
- assist load enhancement for the ICE for faster warm up and better efficiency;
- undertake regenerative braking; and
- deliver pure electric driving for certain period of time.

The energy saving potential of HEVs achieved by an electrical powertrain is shown in Figure 5. The HEV can boost the fuel economy from its ICE counterpart by about 6% by reducing the energy losses during idling and stop and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

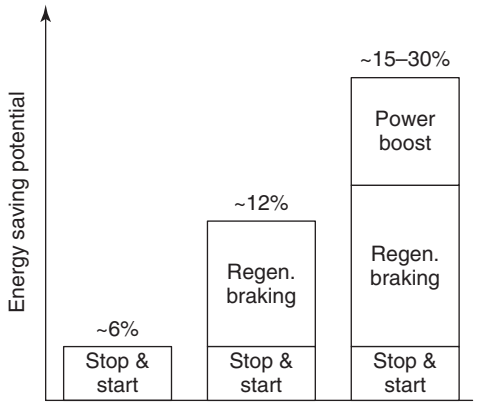


Figure 5. Energy saving potential in HEVs.

start in city driving. An additional 6% of energy savings can be achieved by regenerative braking. By optimizing efficiency of the ICE, an HEV can achieve an aggregated savings potential of 15–30% in fuel consumption relative to the ICE vehicle, subject to the vehicle design and energy management strategies. Table 1 lists fuel consumption of HEVs and their comparable ICE models in the United States (USDOE, 2011b). The HEVs achieved a reduction of 8.4–46.7% in fuel consumption in the urban diving cycle and a lower reduction of 3.9–20.3% in the highway driving cycle. In testing the fuel consumption, the vehicle is tested on a dynamometer. When the engine or motor and transmission drive the wheels, the momentum is passed to the dynamometer; as a result, the vehicle does not move. A professional driver runs the vehicle through two standardized driving schedules, one each to simulate city and highway driving cycles. The driver has to maintain the mandated pace via a real-time computer display.

2.5 Electrical drivetrain design

The design of energy management strategies for HEVs involves different considerations. Some key considerations are summarized in the following:

- *Optimal Engine Operating Line*. In case, the engine needs to deliver different power demands, the corresponding optimal operating points constitute an optimal operating line. Figure 6 shows a typical optimal operating line of an engine, in which the optimization is

Table 1. Fuel consumption of HEVs in the market and their comparable ICE models.

Year	Make	Model	Displacement (L)	Fuel Consumption in Liter per 100 km (Urban)	Fuel Consumption in Liter per 100 km (Highway)	Fuel Consumption Reduction (Urban) (%)	Fuel Consumption Reduction (Highway) (%)
2007	Toyota	Camry	2.4	11.2	7.6	−36.6	−9.2
2007	Toyota	Camry hybrid	2.4	7.1	6.9		
2008	Chevrolet	Malibu	2.4	10.7	7.8	−8.4	−5.1
2008	Chevrolet	Malibu hybrid	2.4	9.8	7.4		
2008	GMC	Yukon 1500 2WD	6.2	19.6	12.4	−42.9	−13.7
2008	GMC	Yukon 1500 hybrid	6	11.2	10.7		
2010	Ford	Fusion	2.5	10.7	8.1	−46.7	−19.8
2010	Ford	Fusion hybrid	2.5	5.7	6.5		
2010	Toyota	Highlander	3.5	13.8	10.2	−37.0	−7.8
2010	Toyota	Highlander hybrid	3.3	8.7	9.4		
2010	Mercury	Milan	2.5	10.7	7.6	−46.7	−14.5
2010	Mercury	Milan hybrid	2.5	5.7	6.5		
2011	Porsche	Cayenne	3.6	14.7	10.2	−19.7	−3.9
2011	Porsche	Cayenne S	3	11.8	9.8		
2011	Honda	Civic	1.8	9	6.9	−34.4	−20.3
2011	Honda	Civic hybrid	1.3	5.9	5.5		
2011	Hyundai	Sonata	2.4	10.7	6.7	−37.4	−11.9
2011	Hyundai	Sonata hybrid	2.4	6.7	5.9		

Source: Official US government source for fuel economy information http://fueleconomy.gov/.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

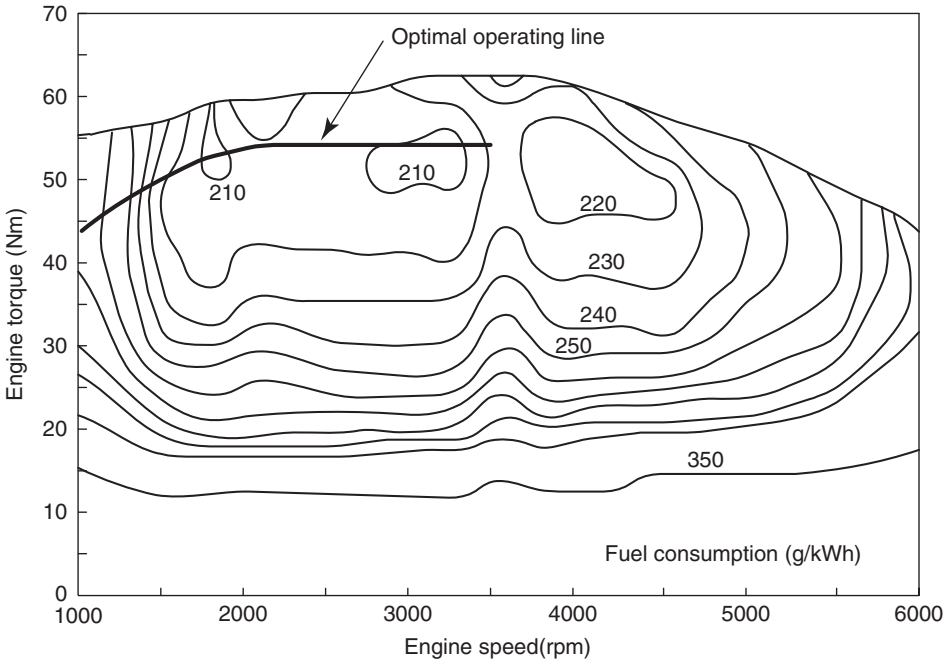


Figure 6. Optimal operating line in an engine fuel consumption map.

based on the minimum fuel consumption, which is equivalent to maximum fuel economy.

- *Optimal Engine Operating Region*. The engine has a preferred operating region on the torque–speed plane, in which the fuel efficiency remains optimum.
- *Minimum Engine Dynamics*. The engine operating speed needs to be regulated in such a way that any fast fluctuations are avoided, hence minimizing the engine dynamics.
- *Minimum Engine Speed*. When the engine operates at low speeds, the fuel efficiency is very low. The engine should be cutoff when its speed is below a threshold value.
- *Minimum Engine Turn-On Time*. The engine should not be turned on and off frequently; otherwise, it results in additional fuel consumption and emissions. A minimum turn-on time should be set to avoid such drawbacks.
- *Proper Battery Level*. The onboard battery level should be sized properly so as to provide sufficient power for acceleration and accept regenerative power during braking or going downhill.
- *Safe Battery Voltage*. The battery voltage may be significantly altered during discharging, generator charging, or regenerative charging. This battery voltage should not be too high or too low; otherwise, the battery may be permanently damaged.
- *Geographical Policy*. In certain cities or areas, the HEV needs to be operated in the pure electric mode. The changeover could be controlled manually or automatically.

The mature electric motors utilized in HEVs include the permanent magnet (PM) motor and induction motors. Main considerations of motors for vehicular applications are torque and power densities, wide range of speed (including constant torque and constant power operations), efficiency over a wide range of speeds, reliability, and robustness. PM motors possess high efficiency, high torque, and high power density; however, they have a short constant power range and their stator winding generates a back electromotive force (EMF). Induction motors offer simplicity, robustness, and a wide range of speeds; however, owing to rotor loss, these motors have lower efficiencies and must be larger than PM motors of similar capability.

The battery is a critical part in any HEVs. The battery has to be intrinsically tolerant of abusive conditions such as overcharge, short circuit, crush, fire exposure, mechanical shock, and vibration. A typical battery pack may contain cells in different series or parallel combination in a vehicle battery system; thus, cells' SoCs have to be balanced to prevent undercharge and overcharge. The key requirements for vehicle batteries are high specific energy and high specific power, long cycle life, high efficiency, wide operating temperature, and low cost for commercialization. The viable batteries for vehicle applications consist of the valve-regulated lead acid (VRLA), nickel–cadmium (Ni–Cd),

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

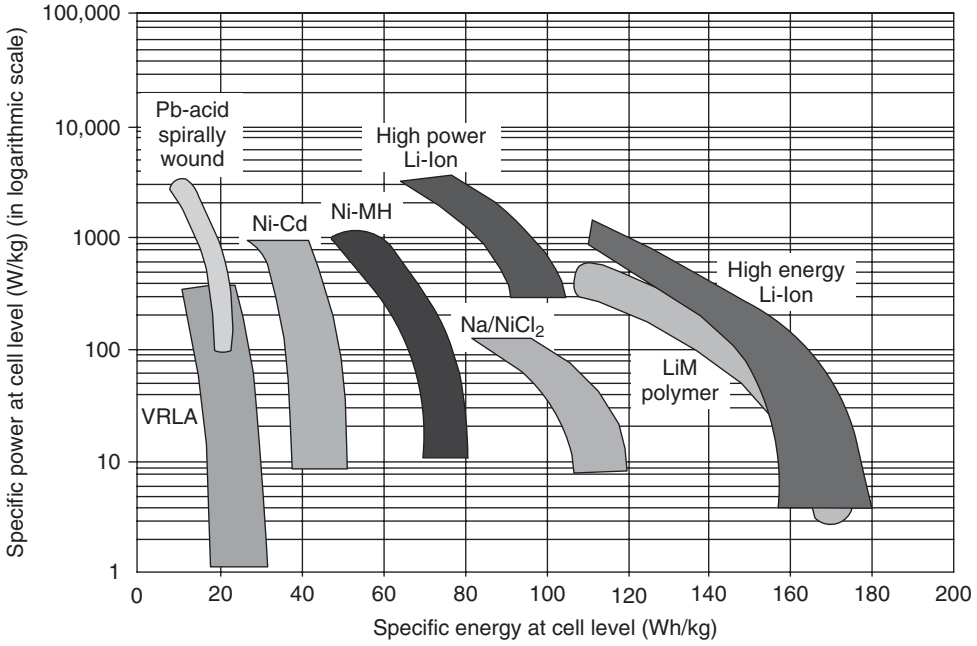


Figure 7. Specific energy and specific power of vehicle batteries.

nickel–zinc (Ni–Zn), nickel-metal hydride (Ni-MH), zinc/air (Zn/Air), aluminum/air (Al/Air), sodium/sulfur (Na/S), sodium/nickel chloride ($Na/NiCl_2$), lithium metal-polymer (LiM-polymer) and lithium-ion (Li-ion) batteries. The Ni-MH battery has been widely equipped in most commercial HEVs. The specific energy and specific power of these batteries are shown in Figure 7.

The Li-ion battery, with advanced positive electrode and negative electrode materials, is a promising battery for HEVs. The advanced positive electrode materials are $LiCoO_2$, $LiMn_2O_4$, $LiFePO_4$, lithium nickel manganese cobalt (NMC) oxide ($LiNiMnCoO_2$), and lithium nickel cobalt aluminum (NCA) oxide ($LiNiCoAlO_4$). The mature negative electrode materials are graphite and titanate. Table 2 shows the potential Li-ion batteries for vehicle applications.

3 DRIVETRAIN TOPOLOGIES OF HEVS

The HEV can be divided into series, parallel, series–parallel, and complex hybrids with reference to their drivetrain topologies. Figure 8 shows the corresponding functional block diagrams, in which the electrical link is bidirectional; the fuel link is unidirectional and the mechanical link (including clutches and gears) is also bidirectional. The series hybrid couples the engine with the generator to produce electricity for pure electric propulsion, whereas the parallel hybrid couples both the engine and the electric motor with the transmission via the same drive shaft to propel the wheels. The series–parallel hybrid is a direct combination of both the series and parallel hybrids. On top of the series–parallel hybrid operation, the complex hybrid can offer additional and versatile operating modes by adding one more electric motor to deliver four-wheel drive (4WD) operations.

Table 2. Advanced Li-ion batteries for vehicle application.

Positive Electrode	Negative Electrode	Manufacturers	Key Feature
LiCoO2	Graphite	Sony	Mature
LiMn2O4	Graphite	NEC, GS, Yuasa, LG	High power
NCA/NMC	Graphite	SAFT, Samsung, Sanyo, Evonik	High energy
LiFePO4	Graphite	A123, Valence Tech, BYD	Highly stable
LiMn2O4	Titanate	Toshiba, Enerdel	High discharge rate

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

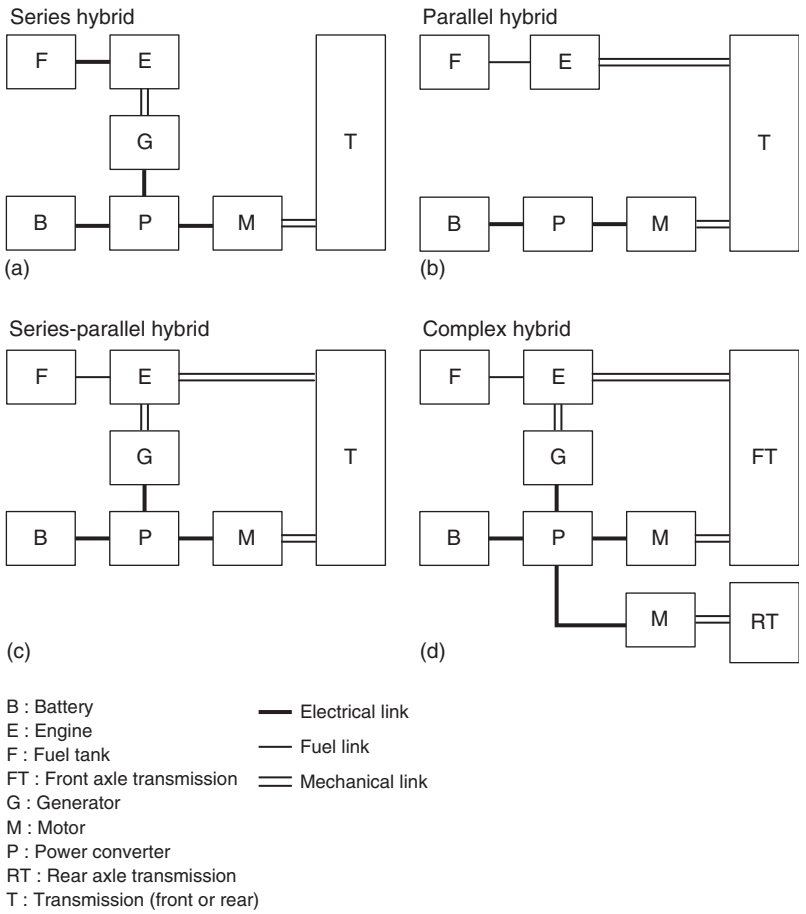


Figure 8. Classification of HEVs.

3.1 Series hybrid system

The series hybrid is the simplest kind of HEV. The engine's mechanical output is first converted into electricity using a generator. The converted electricity either charges the battery or bypasses the battery to drive the motor so as to propel the wheels. Conceptually, it is an engine-assisted EV, aiming to extend the driving range comparable with that of the ICEV. The generator is connected with the electrical drivertrain by wire such that it has the definite advantage of flexibility for locating the engine generator set. Although it has an added advantage of simplicity of its drivetrain, the series HEV is solely propelled by the electric motor, whereas the electrical energy is provided by the onboard battery or by the engine via the generator. Another disadvantage is that all these propulsion devices must be sized for maximum sustained power if the series HEV is designed to climb a long grade.

In the series hybrid system, the energy flow can be illustrated by four operating modes as shown in Figure 9. During startup, normal driving or acceleration of the series HEV, both the engine (via the generator) and the battery deliver electrical energy to the power converter that drives the electric motor and then the wheels via the transmission. It can also operate in an EV mode in the charge-depleting mode. At light loads, the engine output is greater than the output required to drive the vehicle so that the generated electrical energy is also used to charge the battery until the battery capacity reaches a proper level. During braking or deceleration, the electric motor acts as a generator that transforms the kinetic energy of the wheels into electricity, hence charging the battery via the power converter. In addition, the battery can be charged by the engine via the generator and power converter, even when the vehicle is standstill.

3.2 Parallel hybrid system

As opposed to the series hybrid, the parallel HEV allows both the engine and the electric motor to deliver power

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

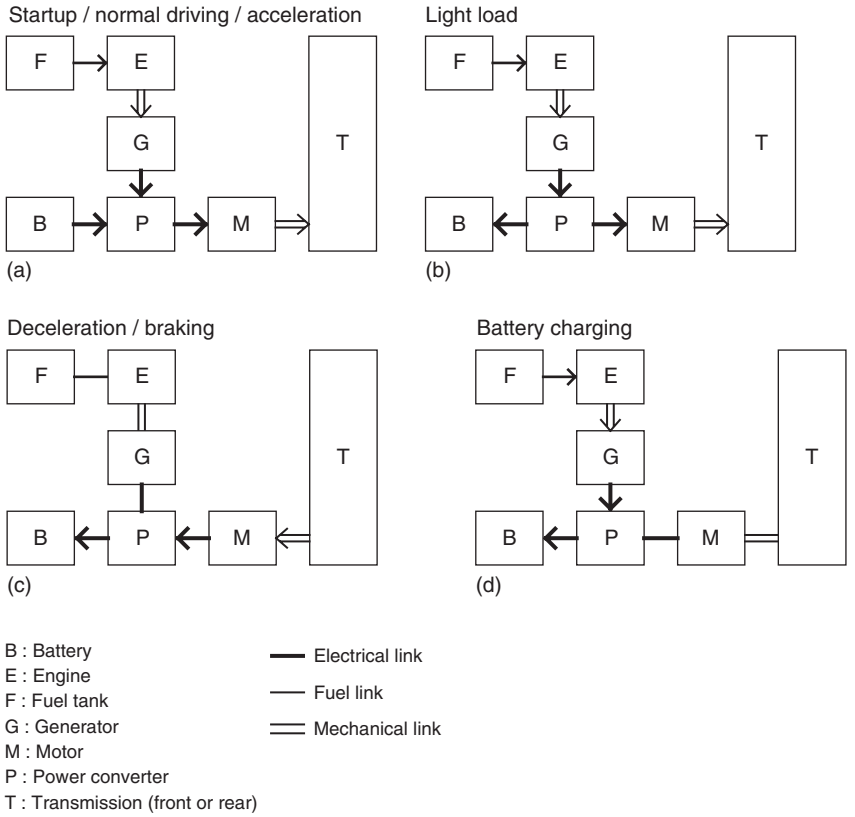


Figure 9. Operating models of series hybrids.

in parallel to drive the vehicle. Both the engine and the electric motor are generally coupled to the drive shaft of the wheels via two clutches such that the propulsion power may be supplied by the engine, the electric motor, or both. Conceptually, it is an electrically assisted ICEV for achieving lower emissions and fuel consumption. The electric motor can be used as a generator to charge the battery by regenerative braking or absorbing power from the engine when its output is greater than that vehicle demand. The parallel hybrid needs only two propulsion devices—the engine and the electric motor. Another advantage over the series case is that a smaller engine and a smaller electric motor can be used to get the same performance until the battery is depleted. The Honda Civic Hybrid is a parallel HEV, which is propelled by a 1.3 L ICE and a 12 kW electric motor integrated into the powertrain to boost efficiency of the engine.

Operating modes of a parallel HEV are illustrated in Figure 10. During startup or full-throttle acceleration, both the engine and the electric motor propel the vehicle. Typically, the power distribution between the engine and the electric motor is 80% to 20%. During normal driving, the engine solely supplies the necessary power to propel the vehicle, whereas the electric motor remains in the off mode. The parallel HEV can also run in EV mode in the charge-depleting mode. During braking or deceleration, the electric motor acts as a generator to charge the battery via the power converter. In addition, as both the engine and the electric motor are coupled to the same drive shaft, the battery can be charged by the engine via the electric motor when the vehicle is at light load.

3.3 Series–parallel hybrid system

In the series–parallel hybrid, the configuration incorporates the features of both the series and parallel hybrids. It has advantageous features of both the series and parallel hybrids, but it is relatively more complicated and costly. The Toyota Prius is the first series–parallel HEV on the market. The series–parallel hybrid has more freedom to boost the system efficiency of the vehicle. Figure 11 shows a typical series–parallel hybrid system and six operating modes. During startup and driving at light load, the battery solely feeds the electric motor to propel the vehicle while the engine is in the off mode. For full-throttle acceleration and normal driving, both the engine and the electric motor

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

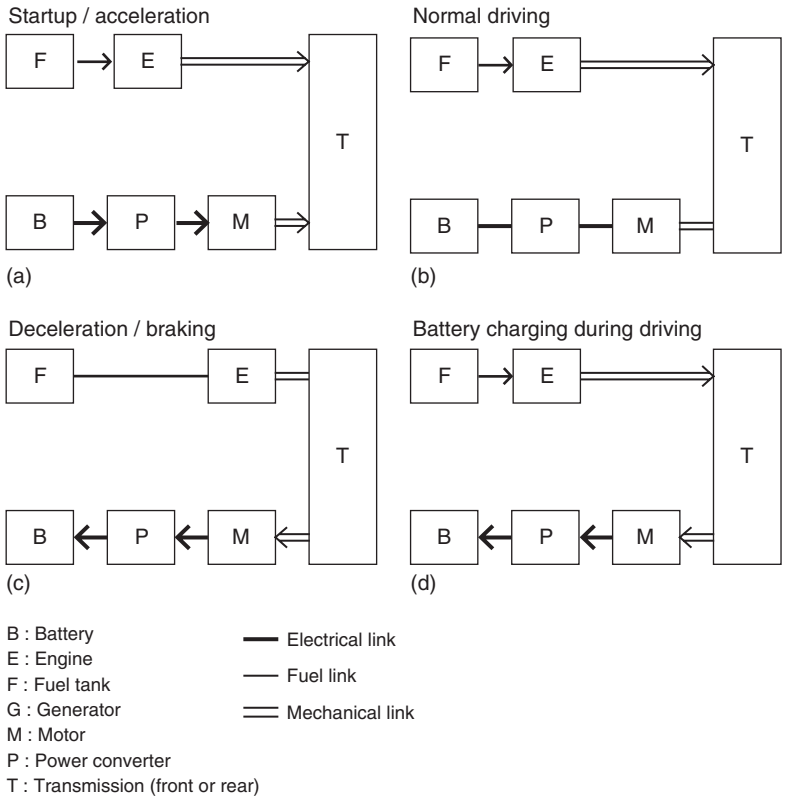


Figure 10. Operating modes of parallel hybrids.

work together to propel the vehicle. The electrical energy used for full-throttle acceleration comes from both the generator and the battery, whereas the energy used for normal driving is solely from the generator driven by the engine. Notice that a planetary gear is usually employed to split the engine output for propelling the vehicle and driving the generator. During braking or deceleration, the electric motor acts as a generator to charge the battery via the power converter. Moreover, for battery charging during driving, the engine not only drives the vehicle but also the generator to charge the battery. The engine can also charge the battery at standstill.

3.4 Complex hybrid system

As reflected by its name, this system cannot be classified into the above-mentioned three hybrids. As shown in Figure 8, the complex hybrid seems to be similar to the series–parallel hybrid; however, the key difference is the bidirectional power flow of the electric motor in the complex hybrid, comparing to the unidirectional power flow of the generator in the series–parallel hybrid. This is also known as a *through the road* hybrid when there is no propeller shaft connection. This bidirectional power flow can allow for versatile operating modes, especially the three propulsion power (owing to the engine and two electric motors) operating mode that cannot be offered by the series–parallel hybrid. Similar to the series–parallel HEV, the complex hybrid suffers from higher complexity and costliness. The Lexus RX 400 h was built with this topology. The front wheels of the RX 400 h are propelled by both the 3.3 L ICE and the 123 kW electric motor in the series–hybrid mode, whereas the rear wheels are propelled by a 50 kW electric motor. There is no direct mechanical coupling between the ICE and the 50 kW rear motor but they are electrically connected by the 82 kW generator.

The energy management system of the complex hybrid is focused on the dual-axle propulsion system. In this system, the front-wheel and rear-wheel axles are separately driven. There is no propeller shaft to connect the front and rear wheels, so it enables a more lightweight propulsion system and increases the vehicle packaging flexibility. Moreover, regenerative braking on all four wheels can significantly improve the vehicle fuel efficiency.

Figure 12 shows a dual-axle complex hybrid system, where the front-wheel axle is propelled by a hybrid drive-train and the rear-wheel axle is driven by an electric motor. There are six operating modes. During startup, the battery

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

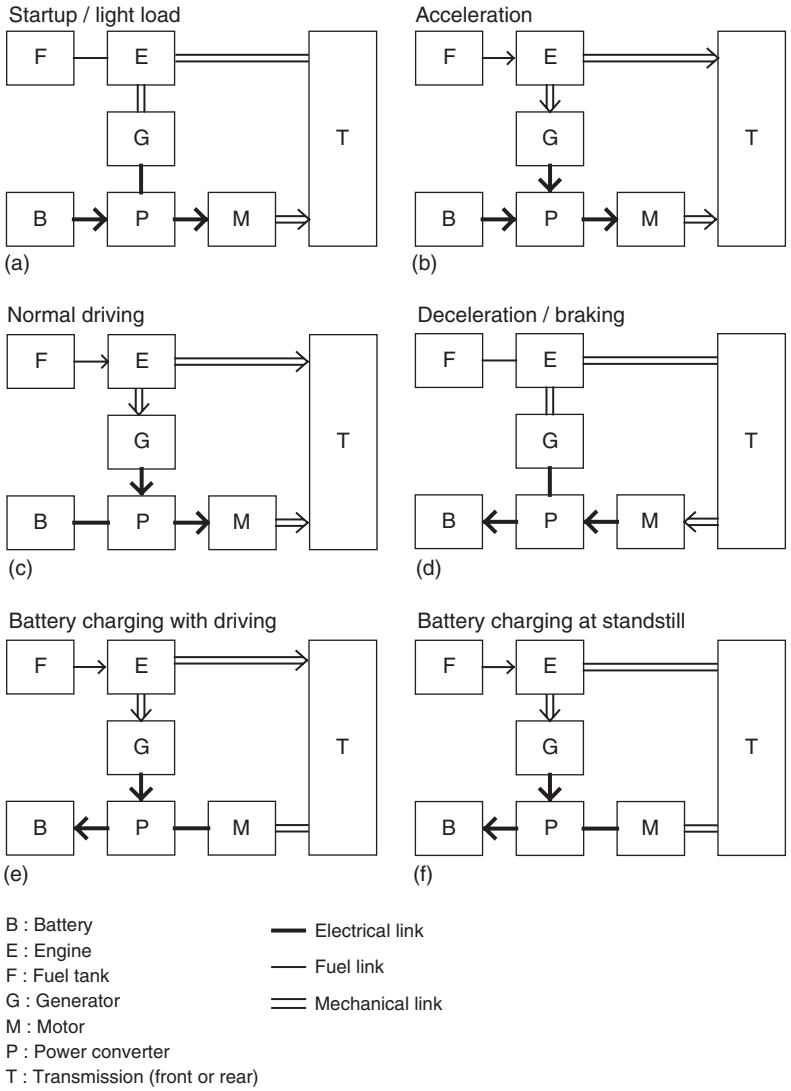


Figure 11. Operating models of series–parallel hybrids.

delivers electrical energy to feed both the front and rear electric motors to individually power the front and rear axles of the vehicle while the engine is off. For full-throttle acceleration, both the engine and front electric motor power the front axle, whereas the rear electric motor drives the rear axle. Notice that this operating mode involves three propulsion devices (one engine and two electric motors) to simultaneously propel the vehicle. During normal driving and/or battery charging, the engine output is split to power the front axle and to drive the electric motor, which works as a generator to charge the battery. The engine, front electric motor, and front axle can be mechanically coupled by planetary gear sets. When driving at light load, the battery delivers electrical energy to the front electric motor only to drive the front axle, whereas both the engine and rear electric motor are off. During braking or deceleration, both the front and rear electric motors act as generators to simultaneously charge the battery. A unique feature of this dual-axle system is the capability of axle balancing. In case the front wheels slip, the front electric motor works as a generator to absorb the change of engine output power. Through the battery, this power difference is then used to drive the rear wheels to achieve axle balancing.

Figure 13 shows another dual-axle complex hybrid system, where the front-wheel axle is driven by an electric motor and the rear-wheel axle is propelled by a hybrid drivetrain. Focusing on vehicle propulsion, there are six operating modes. During startup, the battery delivers electrical energy only to the front electric motor that in turn drives the front axle of the vehicle, whereas both the engine

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

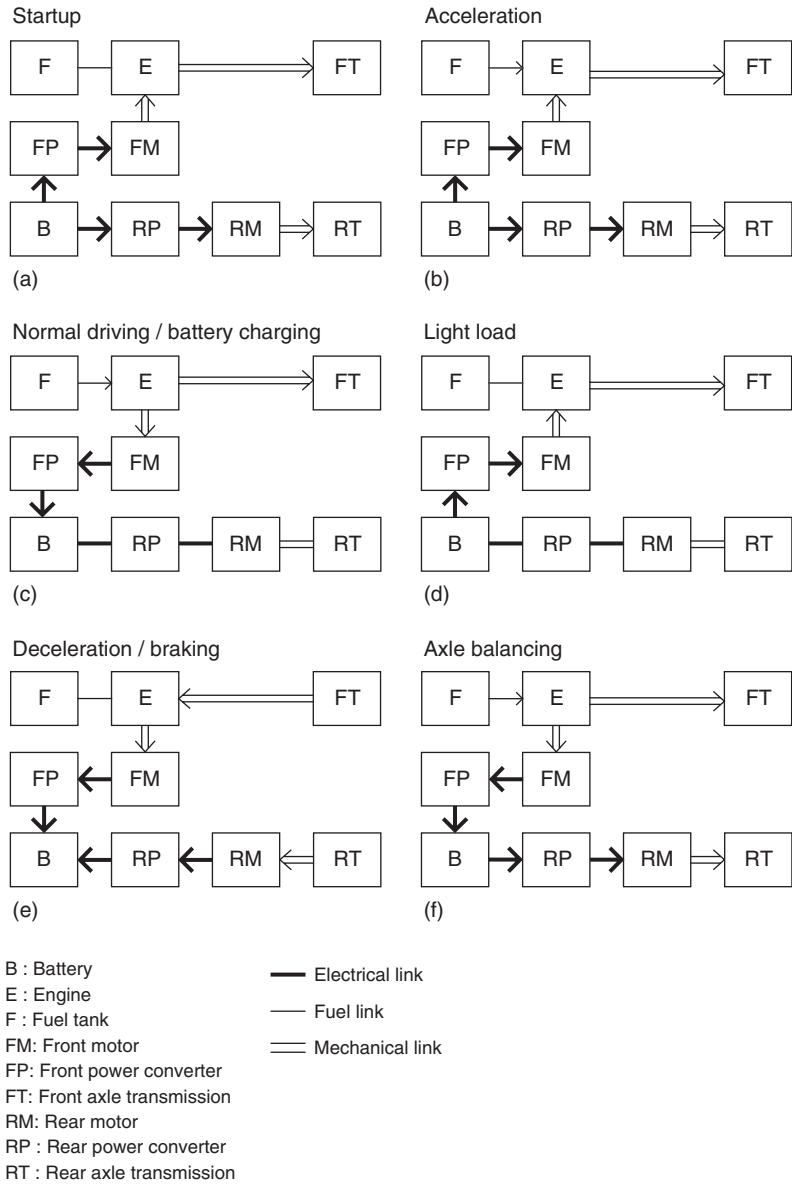


Figure 12. Operating modes of dual-axle (front-hybrid rear-electric) complex hybrids.

and rear electric motor are off. Once the vehicle moves forward, the battery also delivers electrical energy to the rear electric motor raising the engine speed, thus starting the engine. For full-throttle acceleration, the front electric motor drives the front axle, whereas both the engine and rear electric motor work together to propel the rear axle. Therefore, the three propulsion devices (one engine and two electric motors) propel the vehicle simultaneously. During normal driving, the engine works alone to propel the rear axle of the vehicle. During braking or deceleration, both the front and rear electric motors act as generators to simultaneously charge the battery. For battery charging during driving, the engine output is split up to propel the rear axle and to drive the rear electric motor (which works as a generator) for charging the battery.

4 FUNCTIONALITY OF HEVS

According to the level of electric power contribution and functionalities of the electrical powertrain, HEVs can be classified into micro hybrid vehicle (MHV), mild hybrid electric vehicle (MHEV), full hybrid electric vehicle (FHEV), and PHEV. The battery is also an important

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

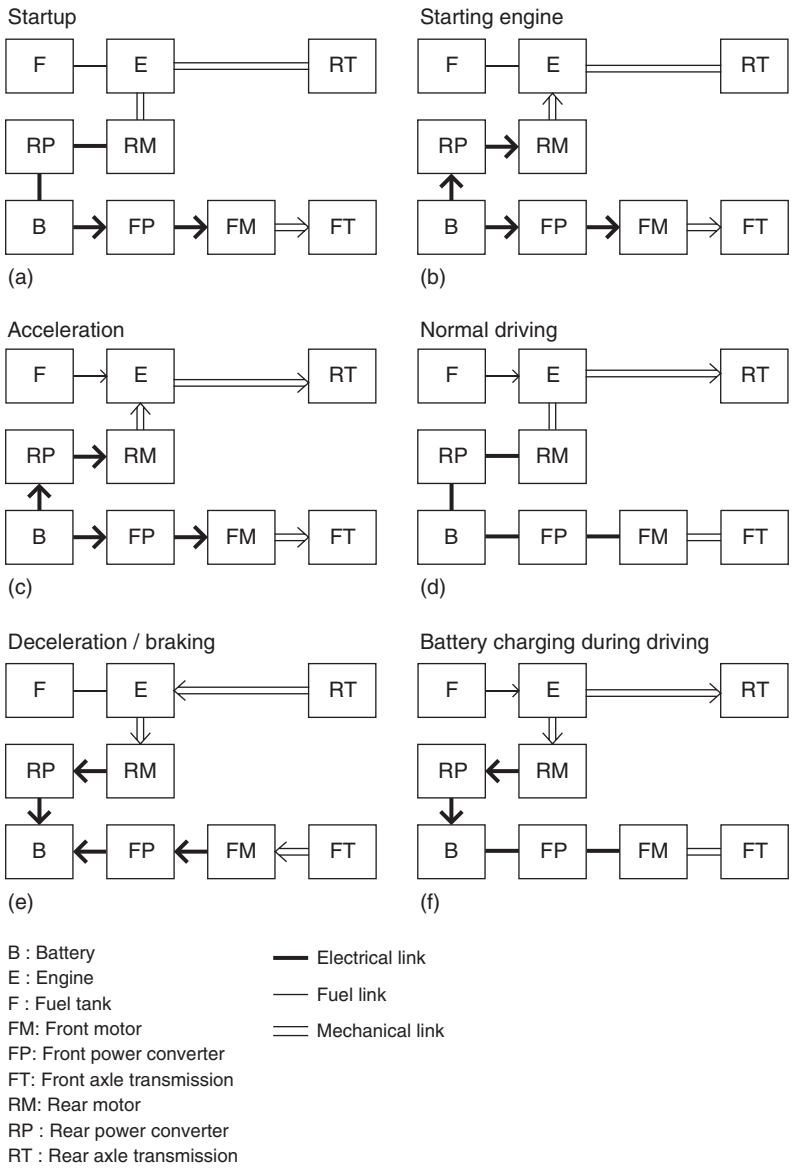


Figure 13. Operating modes of dual-axle (front-electric rear-hybrid) complex hybrids.

component of these HEVs but the requirements for power, energy, cycle life, and system voltage are different. Degree of hybridization of the electrical powertrains and the favorable battery voltages are shown in Figure 14. The operating voltage is raised to increase the discharge power of the battery and to minimize the current and conductor weight.

4.1 Micro hybrid vehicle

The MHV has an electric motor with peak power of about 2.5 kW. The electrical powertrain is driven by a battery system at 12–42 V. The motor is small and simple in structure and serves a function similar to the starter and alternator in an ICE vehicle. The electrical and engine powertrains in an MHV are governed by an automatic stop–start mechanism, in which the engine shuts down under vehicle braking and rest. The battery is recharged with the engine with a conventional generator. The MHV is favorable for city driving, where there are frequent stops and starts. An MHV's fuel economy can be 5–10% higher than that of an ICE vehicle in city driving. The Citroen C3 is an MHV using the Valeo motor system. The battery discharges frequently in cranking the engine in MHVs. Thus, there is a demand for high cycle life for batteries in MHVs. When the functionality of a "micro hybrid" is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

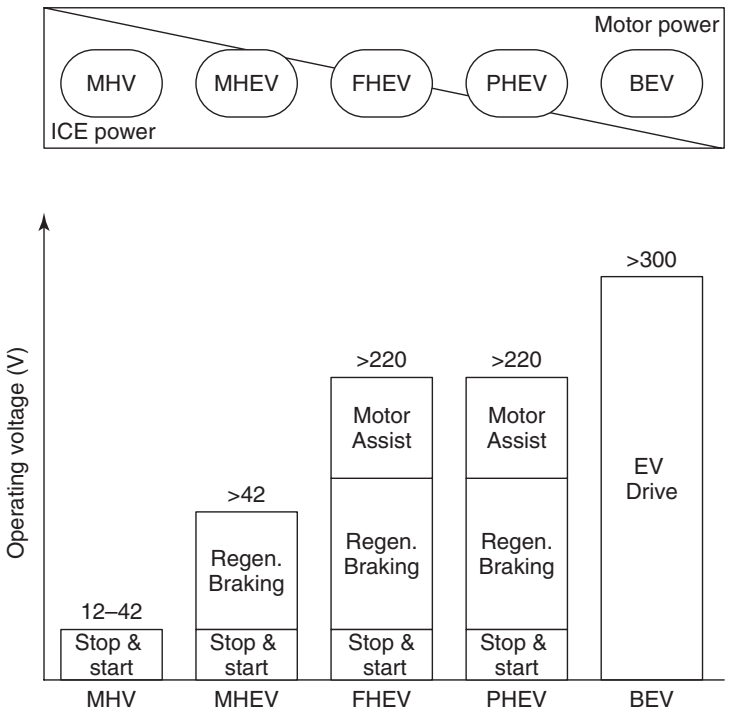


Figure 14. Operating voltages of electrical drivetrains in HEVs.

simply to provide stop–start, it is not really a hybrid as there is no contribution to propulsion from the motor.

4.2 Mild hybrid electric vehicle

The MHEV has a more powerful electrical powertrain than an MHV's. The typical electric motor power of a sedan MHEV is about 10–20 kW at 100–200 V. The motor is directly operated with the engine. The motor has a large inertia such that it can replace the original flywheel of the engine. The motor and the engine are generally coupled in parallel. The electrical powertrain is designed to crank the engine and offer regenerative braking during braking. There are demands of high specific power and long service life for batteries in MHEVs. The battery's charge and discharge power depend on its SoC. The battery's discharge power decreases with its SoC. The minimum operating SoC is around 40–50% to uphold sufficient power for launch and acceleration support. On the other hand, the battery's recharging power drops when the SoC is high; thus, the maximum operating SoC is regulated at around 70–80% to maintain sufficient recharge power for regenerative braking. Typically, the batteries operate in an SoC window between 40% and 70%. Comparing with an ICE vehicle, the MHEV can boost the fuel economy by 20–30% in city driving. Examples of MHEVs are Honda Insight Hybrid and Civic Hybrid.

4.3 Full hybrid electric vehicle

The FHEV has a high power electrical powertrain to drive the vehicle purely by electricity in a short driving range. The typical electric motor power for sedan FHEV is about 50 kW at 200–350 V. Generally, the motor, generator, and engine are coupled in series–parallel configuration. With the aid of power split devices, which are mainly built by planetary gear sets and clutches, the energy management system of the engine, motor, and generator is designed to maximize energy efficiency and minimize emissions.

The FHEV can be driven in pure EV mode and hybrid mode. The electrical powertrain assists the engine, not only at the starting, but also during acceleration in the hybrid model, which is also called *charge sustaining mode*. In the charge-sustaining mode, the battery is recharged not only by regenerative braking but also by the engine to maintain the SoC in a high and narrow window. The FHEV can achieve higher fuel economy than that of the ICE vehicle by 30–50% in city driving. Examples of FHEVs are the Ford Escape Hybrid, GM Volt, Toyota Prius, Toyota Highlander, and Lexus RX 400h.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.4 Plug-in hybrid electric vehicle

The PHEV is similar to that of an FHEV. The key differences are the additional battery pack and the functionality of grid recharging. In addition to the charge-sustaining mode, the PHEV can also operate in the charge-depletion mode, in which the PHEV operates in pure EV mode. Thus, the battery SoC drops in the charge-depletion mode.

The electrical drivetrain of a PHEV works in a high voltage at 220–350 V. The battery energy capacity in PHEVs is the largest among all HEVs, and it is determined by the designated pure electric driving range. The PHEV operates in the charge-depletion mode first and then the charge-sustaining mode. In the charge-depletion mode, the battery SoC decreases from 100% to a threshold SoC (typically 20–30%), which triggers the operation mode change. In the charge-sustaining mode, the battery SoC oscillates around the threshold SoC. The battery is recharged from the grid at the end of the trip. Similar to the EV, the PHEV suffers from complexity and costliness. However, the PHEV delivers longer driving range than the EV's that is comparable to conventional ICEs.

The BYD F3DM is the world's first mass production PHEV, which went on sale to the government agencies and corporations in China in December 2008. Toyota also worked on a plug-in version of the Prius. The plug-in Prius was converted from the Prius by adding additional 1.3 kWh battery pack into the car and a charging unit. A PHEV can also be implemented in a series hybrid topology. The GM Chevrolet Volt is a series hybrid PHEV, which is also called *extended-range electric vehicle* (*EREV*). The EREV is driven by one sole electrical powertrain, powered by the battery and a small engine.

5 CONCLUSIONS

The chief advantage of HEV technology is that it reduces fuel consumption relative to standard ICEVs and subsequently reduces emissions associated with fuel combustion. HEV technology can take advantage of benefits resulting from the coupling of electric motors with an ICE.

The electric motor is more efficient than an engine, does not consume energy at standstill, and delivers maximum torque at zero and low speeds. With the electric motor, the HEV can further reduce fuel consumption by shutting off the engine during idle and restarting the engine in a start–stop system. Regenerative braking can further reduce fuel consumption by converting the vehicle's kinetic energy, which is usually dissipated as heat in the brakes, into electrical energy that can recharge the vehicle's traction battery. The presence of an electric motor also allows for reducing the size of the ICE that can be tuned and controlled to operate at more efficient levels. Optimized ICE operation results in reduced maintenance costs by way of less mechanical wear. The presence of sophisticated electrical components such as energy storage devices, motors, and power converters unavoidably increases the cost of HEVs.

HEVs are designed for functionality and the vehicle market; therefore, there are MHs, MHEVs, FHEVs, and PHEVs. All these HEVs can have a series, parallel, series–parallel, or complex drivetrain. HEV development is driven by advances in ICEs improving combustion with the aid of automotive electronics and by the development of an advanced electric drive and its optimized integration with ICE.

RELATED ARTICLES

EVT and E-CVT for Full Hybrid Electric Vehicles
EV Powertrain Configurations
Micro, Mild and Full Hybrid
Range Extender EV
Overview of Electric, Hybrid and Fuel Cell Vehicles
Power and Energy Requirements for Electric and Hybrid Vehicles

REFERENCES

Chan, C.C. and Chau, K.T. (2001) *Modern Electric Vehicle Technology*, Oxford University Press.

Chan, C.C. (2007) The state of the art of electric, hybrid, and fuel cell vehicles. *Proceedings of the IEEE*, **95** (4), 704–718.

Chan, C.C., Wong, Y.S., Bouscayrol, A., and Chen, K. (2009) Powering sustainable mobility: roadmaps of electric, hybrid, and fuel cell vehicles. *Proceedings of the IEEE*, **97** (4), 603–607.

Chau, K.T. and Wong, Y.S. (2002) Overview of power management in hybrid electric vehicles. *Energy Conversion and Management*, **43**, 1953–1968.

Kabza, H. (2009) Hybrid Electric Vehicles: Overview in *Encyclopedia of Electrochemical Power Sources* (ed J. Garche), Elsevier, Amsterdam, pp. 249–268.

Taylor, C. (1998) Automobile engine tribology—design considerations for efficiency and durability. *Wear*, **221** (1), 1–8.

USDOE (2011a). Fuel Economy: Where the Energy Goes, http://www.fueleconomy.gov/feg/atv.shtml (accessed 21 October 2013).

USDOE (2011b). Find a Car. http://www.fueleconomy.gov/feg/findacar.htm (accessed 21 October 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto051
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Regenerative Braking Systems

Akira Sakai, Tetsuya Miyazaki, Takahiro Okano, Kazunori Nimura, and Daisuke Nakata

Toyota Motor Corporation, Toyota, Japan

1 Introduction 1
2 Regenerative Braking in HEVs and EVs 1
3 Braking Systems of HEVs and EVs 6
4 Regenerative Braking and Friction Braking Control 11
Related Articles 14
References 14

1 INTRODUCTION

In response to growing demands to reduce CO_2 emissions, recent years have seen rapid progress in the development of next-generation vehicles such as hybrid electric vehicles (HEVs), plug-in hybrid electric vehicles (PHEVs), electric vehicles (EVs), and fuel cell vehicles (FCVs). Most next-generation vehicles use electrical energy, and one of their primary features is regenerative braking that generates braking force by using motors as generators during deceleration.

The function of the novel braking systems installed in these vehicles is to improve fuel efficiency through regenerative braking and to secure both braking and dynamic performance.

Furthermore, regenerative braking has also been adopted in vehicle categories other than HEVs, EVs, and the like. These include vehicles with front-wheel drive (FWD), rear-wheel drive (RWD), and all-wheel drive (AWD) systems (AWD systems include mechanically coupled and independent front/rear wheel types). Consequently, the amount of regenerative braking force with respect to the motors and battery capacity and the types of axles where regenerative braking can be applied have increased. In response to these trends, braking systems are becoming more diverse.

Moreover, to further improve fuel efficiency, HEVs use a wide range of technology for stopping the engine when it is not needed as a source of motive power, such as when the driver is not operating the accelerator pedal. In addition, as EVs and FCVs do not have an engine, the braking systems in these types of vehicles must be developed to secure the necessary braking performance without using engine intake vacuum.

This chapter details the requirements of braking systems in accordance with the purpose of the vehicle and describes the configuration and controls of braking systems to achieve these requirements.

2 REGENERATIVE BRAKING IN HEVs AND EVs

2.1 Regenerative braking and its uses

The term *regenerative braking* refers to the functions of braking devices that convert kinetic energy to another form of energy during deceleration or the like, recover and store that energy, and use the resistance generated during this process as braking force. Recovering and reusing kinetic energy, which is normally wasted in conventional friction braking as heat, enables more efficient energy usage. Regenerative braking is already in widespread use

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

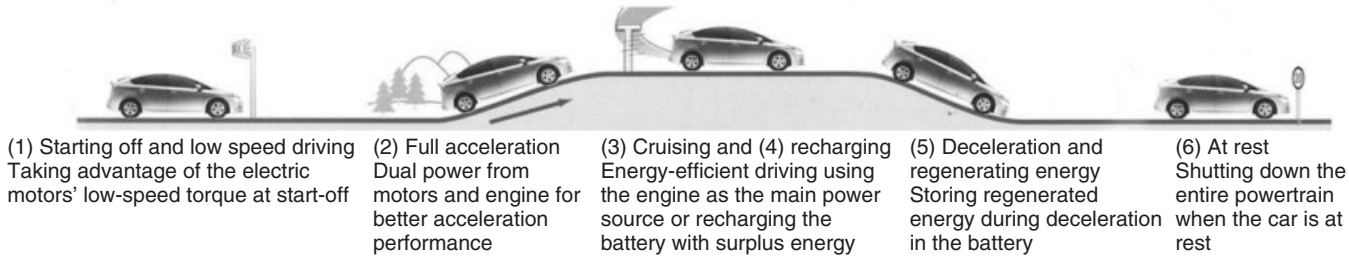


Figure 1. Examples of HEV driving patterns.

in trains (bullet trains, conventional railways, subways, and streetcars), elevators, electrically assisted bicycles, and so on. These systems recover kinetic energy as electrical energy by using motors to generate power during braking and achieve deceleration using the generation resistance.

Since Toyota Motor Corporation released the Prius HEV in 1997, various PHEVs and EVs have also entered mass production. Most of these vehicles are equipped with motors and a battery for storing electrical energy. Figure 1 shows some examples of HEV driving patterns (Hano and Hakiai, 2011).

Typically, regenerative braking in a vehicle uses the motors (or generators) to convert the vehicle's kinetic energy into electrical energy, which is then stored in the battery and used to assist driving force or the like (Figures 2 and 3). In some cases, capacitors may also be used to store energy. In addition to driving force assistance, the stored electrical energy can also be supplied to other auxiliary electrical equipment.

Regenerative braking helps improve fuel efficiency and extend the cruising range of the vehicle. The frequency of deceleration events that waste kinetic energy as heat is particularly high in urban driving and other driving patterns with repeated acceleration and deceleration. The merits of regenerative braking are therefore easy to obtain in these patterns as there are many opportunities to recover energy that is normally wasted.

Regenerative braking is reported to have a 10–25% fuel efficiency improvement effect in current mass-produced HEVs and EVs (Nakamura *et al*., 2002; Nakata *et al*., 2009).

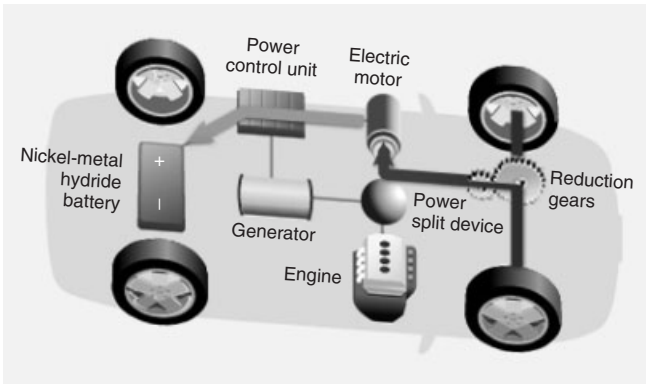


Figure 2. Example of energy flow during regenerative braking.

2.2 Characteristics of regenerative braking in vehicles

The motor torque during regenerative braking can be expressed using the following equation.

$$\text{Motor torque (Nm)}(\propto \text{braking force}) = \frac{\text{generation efficiency} \times \text{generated power (W)}}{[\text{motor speed (1/sec)}\ (\propto \text{vehicle speed})]}$$

In general terms, the restrictions when adopting regenerative braking in a vehicle are as follows.

2.2.1 *Dependence on vehicle speed*

The possible regenerative braking force differs depending on the vehicle speed (line (a) in Figure 4). In general terms, the possible regenerative braking force is relatively lower in higher speed driving. In addition to the output characteristics of the motors, the power generated by the motors during regenerative braking is restricted by the capacity of the battery to accept charge, the processing capacity of the inverter, and so on.

2.2.2 *Braking force while vehicle stopped*

Regenerative braking requires the generation motors to be rotating (line (b) in Figure 4). As the vehicle cannot generate power while stopped, regenerative braking cannot be used.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

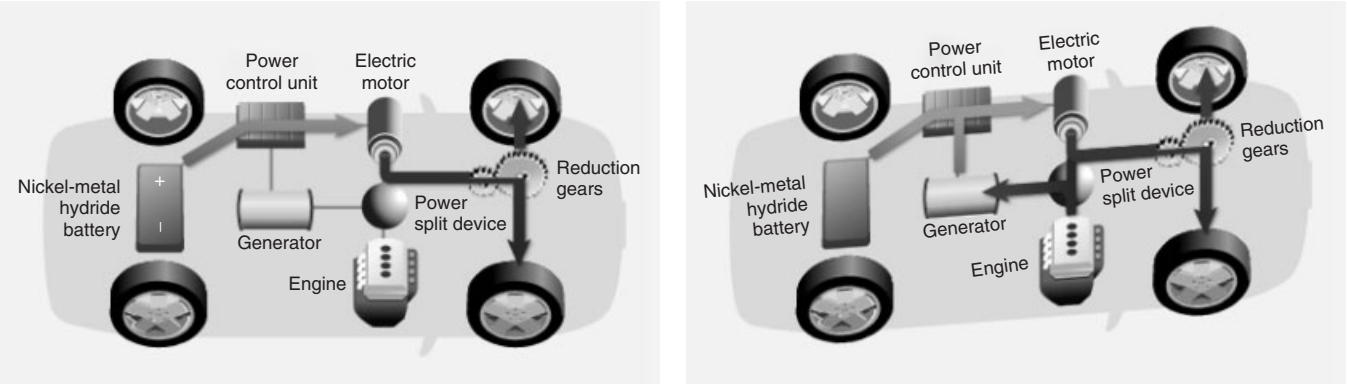


Figure 3. Example of energy flow during driving force assistance. (a) Starting off and low speed driving. (b) Full acceleration.

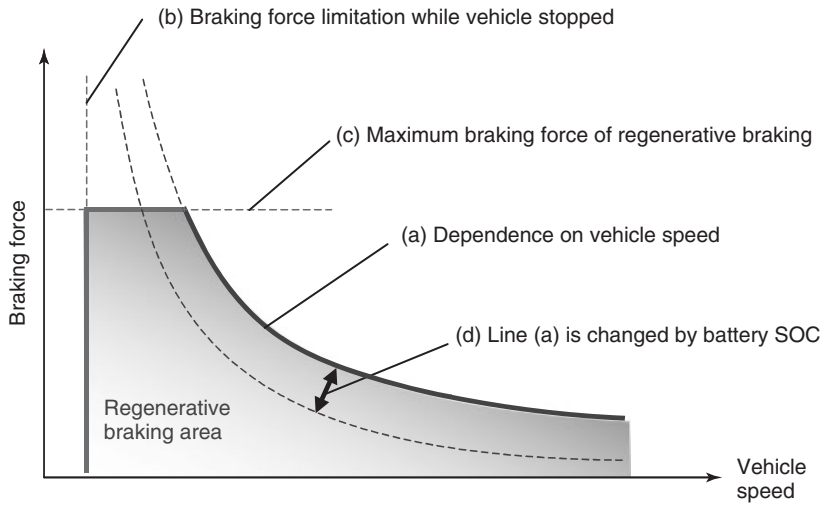


Figure 4. Illustration of possible regenerative braking force.

2.2.3 Maximum braking force of regenerative braking

Conventional friction braking is generally capable of generating a maximum deceleration of approximately 1 G (line (c) in Figure 4). In contrast, regenerative braking force is transferred as torque from the generation motors via parts such as the driveshaft, transmission, and the like. The regenerative braking force is often restricted from the standpoints of the characteristics of these transmission systems and energy efficiency.

2.2.4 Battery state of charge (SOC)

Most HEVs and EVs control the charging process to maintain the battery state of charge (SOC) within a certain range (line (d) in Figure 4). For this reason, the amount of charge is restricted if the SOC is close to the upper limit of control. In this case, the regenerative braking force is restricted as the amount of power generated from regenerative braking that can be accepted by the battery is reduced.

2.3 Characteristics of vehicles with regenerative braking

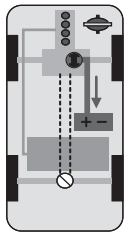
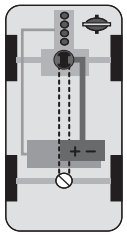
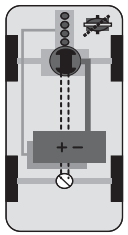
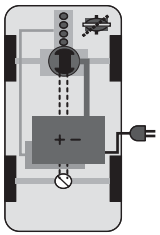
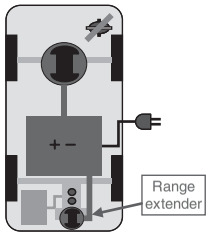
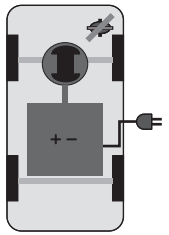
Table 1 shows a list of HEVs and EVs, categorized by drive system (engine and motors), battery capacity, and EV driving capability.

2.3.1 Hybrid electric vehicles

There are many examples of HEVs that can be categorized as either mild or strong HEVs depending on the capacity of the motors and battery.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Categories of HEVs and EVs.

	Micro HEV (usually NOT categorized as HEV)	HEV		PHEV		EV
		Mild HEV	Strong HEV			
Example powertrain						
Idling stop	Possible				Not applicable	
Battery capacity	Small → Large					
Cruising distance as EV	EV operation not possible	Short → Long				
Deceleration force by regenerative braking	Small → Large					
Driving force assistance by motor	Not possible	Low → High			Not applicable (motor drive only)	
Example of past and present commercially available vehicles (passenger vehicles only)	BMW Mini E MAZDA 3	TOYOTA Crown mild Hybrid HONDA Civic Hybrid HONDA Insight HONDA CR-Z HONDA Fit Hybrid BENZ S400 BlueHybrid BMW ActiveHybrid 7	TOYOTA Prius TOYOTA Estima Hybrid TOYOTA Alphard Hybrid TOYOTA Highlander (Kluger) Hybrid TOYOTA Camry Hybrid LEXUS LS600h/hL LEXUS RX400h/450h LEXUS HS250h LEXUS CT200h NISSAN Fuga Hybrid NISSAN Altima HYUNDAI Sonata Hybrid GM Chevrolet Suburban GM Chevrolet Tahoe FORD Fusion Porsche Cayenne VW Touareg	TOYOTA Prius Plug-in Hybrid BYD F3DM	GM Chevrolet Volt	MITSUBISHI i-Miev NISSAN Leaf TOYOTA Rav4 EV TESLA Roadster

Combustion engine | Fuel tank | Vacuum/no vacuum | E-machine | Battery

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Strong HEVs are equipped with comparatively large motors and have a large-capacity battery to allow the vehicle to be driven as an EV for a certain distance. Strong HEVs recover large amounts of energy through regenerative braking using the motors and achieve large improvements in fuel efficiency by reusing this energy while driving.

Mild HEVs are equipped with a comparatively small motor, which is used to restart the engine when the vehicle starts moving and to provide driving force assistance. Mild HEVs have a relatively small-capacity battery, which limits the amount of energy recoverable by regenerative braking.

Micro HEVs do not have a motor for assisting vehicle drive. These vehicles only have an idling stop function and regenerative braking using a generator. Regenerative braking in most micro HEVs uses a conventional generator (i.e., the alternator) to charge the auxiliary battery (12 V). The recovered energy is used to improve fuel efficiency by restarting the engine when the vehicle starts to move and to power auxiliary electrical devices such as the air conditioning. Micro HEVs are usually not categorized as HEVs due to the lack of a drive motor. The energy recoverable by regenerative braking is further restricted as a result.

2.3.2 *Plug-in hybrid electric vehicles*

Unlike strong HEVs, PHEVs are capable of being charged from external sources. Consequently, PHEVs are equipped with a larger capacity battery and have a much longer EV cruising range. The regenerative braking force may be restricted when the battery SOC is close to its upper limits due to external charging.

2.3.3 *Electric vehicles*

EVs have an even larger battery capacity than PHEVs. EVs are not generally equipped with an engine, but one may be provided as a range extender. Regenerative braking plays an extremely important role in EVs due to their short cruising range caused by the lower energy density of batteries compared with fossil fuels.

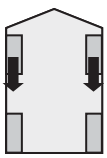
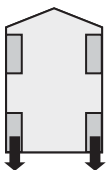
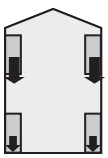
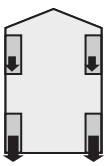
2.4 Characteristics of regenerative braking depending on drive system

Similar to conventional vehicles equipped with an engine, HEVs and EVs have been developed with various types of drive systems. Table 2 shows a list of some typical examples.

2.4.1 *Front-wheel drive*

In a FWD HEV or EV, the motor is connected only to the front wheels. As a result, regenerative braking acts on the front wheels only. However, regenerative braking cannot provide sufficient braking force in every case. Therefore, when regenerative braking is used to provide some of the service braking force, friction braking is used to compensate

Table 2. Categorization by drive system

		FWD	RWD	AWD	
Braking force distribution when braking using regenerative braking only				 When independent control of braking force to the front and rear wheels is possible	 When the front and rear wheels are mechanically connected
Hydraulic braking force				All four wheels	
Regenerative braking force	Front wheels	Yes	No	Braking force distribution as desired to front or rear wheels	Braking force distribution to front and rear wheels depends on torque distribution
	Rear wheels	No	Yes		
Example of past and present commercially available vehicles (passenger vehicles only)		Toyota Prius Nissan Leaf Honda Insight	Toyota Crown HV Toyota GS450h Nissan Fuga Tesla Roadster Mitsubishi i-Miev	Lexus LS600h Toyota Estima/Alphard Hybrid Toyota Highlander/Kluger Hybrid Porsche Cayenne VW Touareg	

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

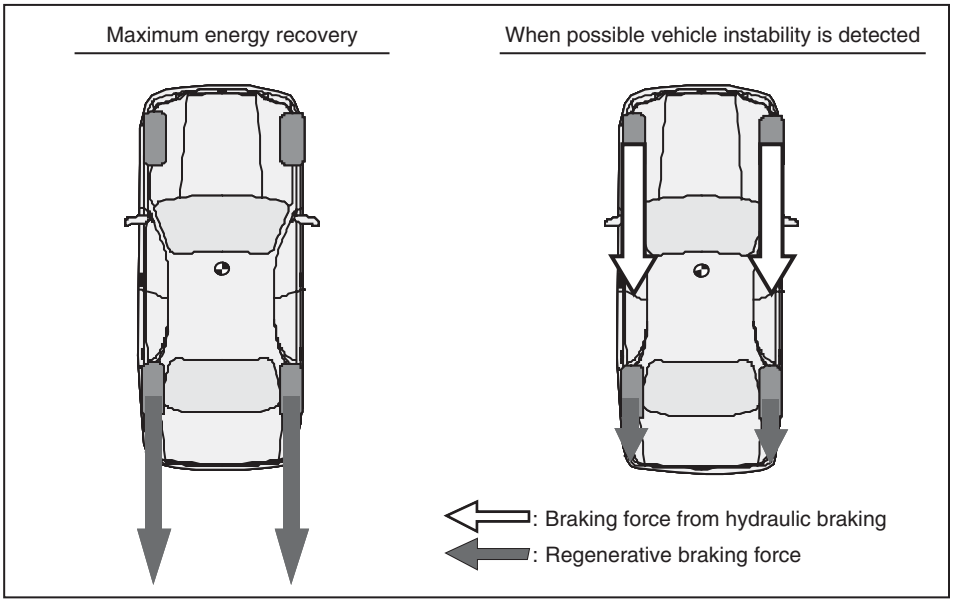


Figure 5. Example when vehicle stability is affected during regenerative braking (RWD vehicle.)

for any insufficiency from regenerative braking at the front wheels and to provide the braking force for the rear wheels.

2.4.2 Rear-wheel drive

In a RWD HEV or EV, the motor is connected only to the rear wheels. As a result, regenerative braking acts on the rear wheels only. In the same way as a FWD HEV or EV, regenerative braking cannot provide sufficient braking force by itself in every case. Therefore, when regenerative braking is used to provide some of the service braking force, friction braking is used to compensate for any insufficiency from regenerative braking at the rear wheels and to provide the braking force for the front wheels.

2.4.3 All-wheel drive

In an AWD HEV or EV, motors are connected to all four wheels, which enable the application of regenerative braking to each wheel. As a result, larger amounts of energy can be recovered by regenerative braking than in a FWD or RWD system. There are various types of AWD systems. When separate motors are provided for the front and rear wheels, the regenerative braking force can be adjusted independently, thereby allowing braking force to be distributed optimally between the front and rear wheels. In contrast, in AWD systems in which the front and rear wheels are mechanically coupled, the regenerative braking distribution is restricted by the torque distribution of the front and rear wheels.

When braking is performed by regenerative braking alone, the braking force distribution to the front and rear wheels differs from that in a conventional vehicle depending on the drive system. This has an effect on vehicle stability. To help resolve this issue, vehicle behavior is detected by various sensors. In situations where vehicle stability might be affected, the desired braking force distribution to the front and rear wheels is achieved by utilizing friction braking (Figure 5).

The following section describes the technology used in vehicles with regenerative braking systems.

3 BRAKING SYSTEMS OF HEVs AND EVs

This section describes the requirements of braking systems in HEVs and EVs. The fundamental requirement of the braking system in an HEV is to guarantee the same level of safety and comfort as the braking system in a gasoline vehicle (referred to below as a *conventional braking system*). At the same time, it is preferable that the braking system also includes a function to help improve fuel efficiency through the use of regenerative braking and the reduction of fuel consumption by enabling the engine to be stopped whenever possible. Therefore, the following sections discuss the requirements for a braking system from the standpoints of safety, comfort, and fuel efficiency improvement.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 6. Securing boost vacuum by electric pump.

3.1 Types of functions to boost pedal effort in braking systems

Braking systems require a boost function to supplement the effort applied to the brake pedal by the driver. Most gasoline vehicles generally achieve this function through a vacuum booster using vacuum generated by the engine as the engine is always in operation in a conventional vehicle.

However, HEVs do not require constant operation of the engine to power the vehicle. In these cases, the engine is stopped to improve fuel efficiency. In a braking system that relies on the engine to generate vacuum for boost, engine operation is required if there is insufficient vacuum to provide the assist in accordance with the driver's brake pedal operation. This is not preferable in an HEV from the standpoint of fuel efficiency. Furthermore, this vacuum is not available in EVs, which have no engine.

Consequently, it is important for the braking systems in HEVs and EVs to achieve a boost function without relying on engine intake vacuum while also improving fuel efficiency. These two aims must be achieved to a high level of performance. There are the following two approaches for achieving a boost function without relying on engine operation.

3.1.1 Using electric pump to secure boost vacuum

In this method, the vacuum for boost is generated by an electric pump instead of the engine. A vacuum sensor in the braking system detects the boost vacuum level and drives the electric pump when the level is below a certain threshold. Figure 6 shows an example of this pump (von Albrichsfeld and Karner, 2009).

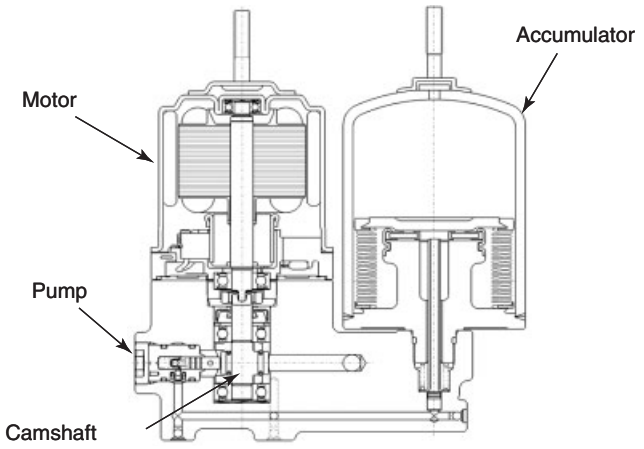


Figure 7. securing high-pressure brake fluid for boost by electric pump.

3.1.2 Using electric pump and accumulator pressure for boost

In this method, high pressure brake fluid is stored in an accumulator by an electric pump and used for generating boost. Figure 7 shows a typical example of this system, which consists of a motor, pump, and accumulator. The motor turns a camshaft to drive a pump, which dispenses high pressure brake fluid into an accumulator for storage (Nakata *et al*., 2009).

3.1.3 Using electric motor thrust as boost

Unlike the two methods described above, this method drives an electric motor when assist is required without relying on a unique braking power source. Instead, this method provides assist directly from the driver's brake pedal operation. In other words, electrical energy stored in the battery is used as the source of the boost (Fujiki *et al*., 2011) (Figure 8).

3.1.4 Using engine drive when engine vacuum is insufficient

In this method, a vacuum sensor such as that described in Section 3.1.1 is added to a conventional braking system. Then, the engine is driven when the boost vacuum falls below a certain threshold. Although this is the lightest method, it requires engine operation to secure the boost vacuum.

If regenerative braking torque is used independently of service braking in an HEV, the requirements for a braking system can be satisfied by adding a boost function using either of the above four methods to a conventional braking system. However, the requirements for an EV or

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

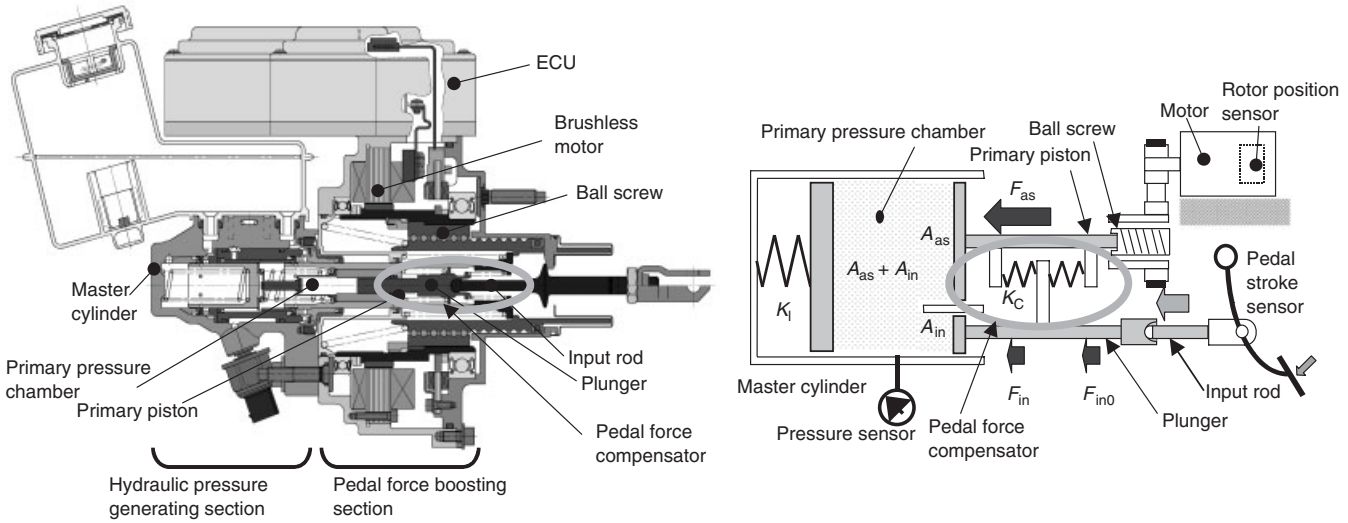


Figure 8. Direct boost using thrust of electric motor.

HEV braking system are greater when regenerative braking torque is used to provide some of the service braking force to further increase the amount of kinetic energy recovered.

3.2 Regenerative-friction brake coordination system

As described in Section 2, regenerative braking torque from electric motors has the following characteristics.

1. The generated braking torque fluctuates in accordance with the battery SOC and vehicle speed.
2. Sufficient braking performance cannot be achieved only from the braking torque generated by electric motors.
3. Regenerative braking requires the generation motors to be rotating. As the vehicle cannot generate power while stopped, regenerative braking cannot be used.
4. The wheels at which regenerative braking torque is applied differ depending on the drive system.

For these reasons, it is not realistic to achieve vehicle braking torque using only electric motors. Therefore, when using regenerative braking torque to provide some of the service braking force, the friction braking torque must be adjusted in accordance with the regenerative braking torque. The following section describes the methods of adjusting the friction braking torque.

3.2.1 Methods of adjusting friction braking torque

There are four methods of adjusting the friction braking torque, which are closely related to the methods of achieving the boost function described in Sections 3.1.1–3.1.2.

3.2.1.1 Wheel cylinder hydraulic pressure control using active vacuum booster. In this method, an active vacuum booster is driven in accordance with the required friction braking torque to adjust the hydraulic pressure inside the master cylinder. The friction braking torque is then adjusted by introducing the desired hydraulic pressure into the wheel cylinder (von Albrichsfeld and Karner, 2009) (Figure 9).

3.2.1.2 Wheel cylinder hydraulic pressure control using electric motor. In this method, the electric motor on the master cylinder is driven in accordance with the required friction braking torque to adjust the hydraulic pressure inside the master cylinder. The friction braking

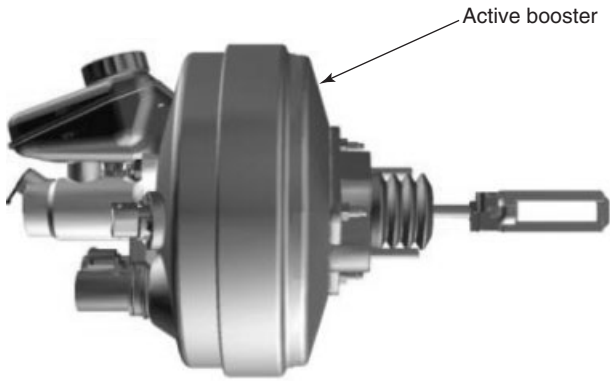


Figure 9. Adjusting friction braking torque by boost vacuum control.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

torque is then adjusted by introducing the desired hydraulic pressure into the wheel cylinder (Hano and Hakiai, 2011; Obata *et al*., 2011).

3.2.1.3 Wheel cylinder hydraulic pressure control using linear solenoid valve. In this method, a linear solenoid valve is driven in accordance with the required hydraulic pressure. The friction braking torque is then adjusted by directly introducing accumulated high pressure brake fluid into the wheel cylinder (Nakata *et al*., 2009) (Figure 10).

3.2.1.4 Direct friction braking torque control using electric motor. In this method, the friction braking torque is adjusted by an electric motor directly installed on the wheel cylinder, which is driven in accordance with the degree of required friction braking torque.

Another requirement for a regenerative-friction brake coordination system is a natural braking feel. The braking feel is enhanced by adjusting the friction braking torque in accordance with the action of regenerative braking system to achieve the vehicle deceleration intended by the driver. In addition, the brake pedal stroke and effort must not fluctuate depending on the state of regenerative braking torque generation or the friction braking torque adjustment. The following section describes how fluctuations in brake pedal stroke and effort are suppressed.

3.2.2 Suppression of fluctuations in brake pedal stroke and effort using pedal stroke simulator

The common approach to generate the targeted brake pedal feeling is to provide a separate pedal stroke simulator to make sure that the adjustment of the friction braking torque does not affect the brake pedal (Nakamura *et al*., 2002; Nakata *et al*., 2009). This separates the friction braking torque adjustment from the brake pedal feeling so that fluctuations in the pedal stroke and effort can be suppressed.

The brake pedal feeling in a conventional braking system is the result of the wheel cylinder in each wheel consuming brake fluid in accordance with hydraulic pressure. The role of the pedal stroke simulator is to realize this feeling without consuming brake fluid. The main method for accomplishing this is to adopt a pedal stroke simulator comprising springs and a piston that reproduces reaction force characteristics equivalent to a wheel cylinder (Figure 11).

3.2.3 Securing regenerative braking and vehicle stability

Sections 3.2.1 and 3.2.2 described how the regenerative braking torque and friction braking torque are controlled independently from the driver's brake pedal operation. However, further considerations are required depending on the drive system of the vehicle as this alters the wheels at

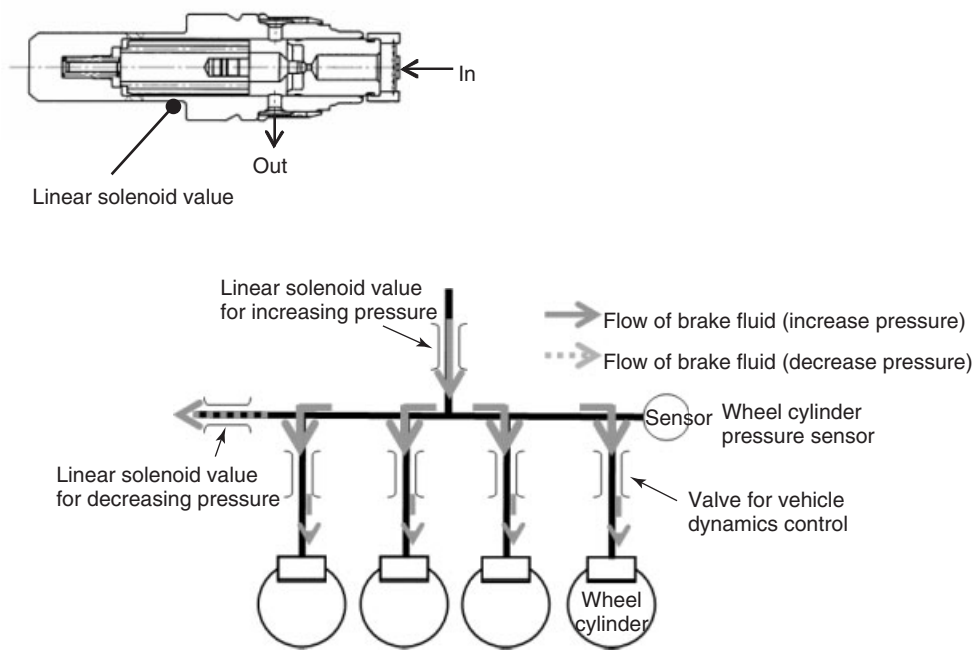


Figure 10. Adjusting friction braking torque by linear solenoid valve control.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

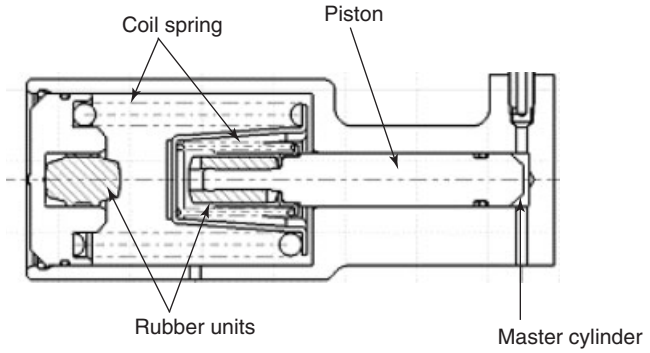


Figure 11. Example of pedal stroke simulator using springs and piston.

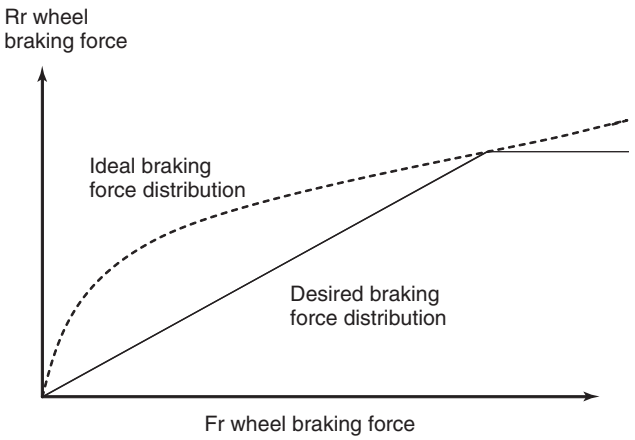


Figure 12. Front/rear braking force distribution.

which regenerative braking torque is applied (see point 4 in Section 3.2). This point is described in detail below.

Figure 12 shows an example of an ideal front/rear wheel braking force distribution. With conventional braking, the desired front/rear braking force distribution can be achieved by designing brake specifications for each wheel. However, regenerative braking torque can only be applied to the drive wheels (only the front wheels in a FWD vehicle and only the rear wheels in a RWD vehicle). Therefore, many vehicles with regenerative braking systems have a different front/rear braking force distribution from the desired distribution for a vehicle with conventional braking.

3.2.3.1 Braking systems capable of independently controlling braking force generated by front and rear friction braking. A braking system configuration that adjusts the wheel cylinder hydraulic pressure independently for each wheel or each braking circuit enables coordinated control between regenerative braking and friction braking with a free front/rear wheel braking force distribution. More specifically, the braking force distributions shown in examples (i) and (ii) in Figure 13 can be achieved even in a FWD vehicle where regenerative braking torque is applied to the front wheels only (Nakamura *et al*., 2002).

3.2.3.2 Braking systems that control braking force generated by friction braking at all four wheels simultaneously. This configuration performs coordinated control between regenerative braking and friction braking. More specifically, this configuration can achieve the braking force distributions shown in examples (iii) and (iv) in Figure 13. In practice, the braking force distribution is determined in consideration of the improvement in fuel efficiency due to the amount of regenerative braking and the vehicle characteristics on braking (Hano and Hakiai, 2011; Nakata *et al*., 2009; Obata *et al*., 2011).

3.3 Braking systems that can perform vehicle dynamic control function

The growing popularity of HEVs and EVs is also helping advance the growth of safety functions such as antilock brake systems (ABSs), traction control systems (TCSs), and electronic stability control (ESC) systems. The most common method of achieving these ABS, TCS, and ESC dynamic control functions is to install a conventional ESC unit inside the braking system (von Albrichsfeld and Karner, 2009; Fujiki *et al*., 2011; Hano and Hakiai, 2011; Obata *et al*., 2011).

These systems have been developed in compact and lightweight forms by reconfiguring the braking mechanism for achieving coordinated control between the components of the ESC unit and regenerative braking (Nakamura *et al*., 2002; Nakata *et al*., 2009).

3.4 Relevant regulations

Regulations for braking systems that include regenerative braking include FMVSS 125 (Federal Motor Vehicle Safety Standard) and ECE-R13H (Economic Commission for Europe). Various regulations also include references to regenerative braking, which require compliance as well. ECE-R13H divides regenerative braking systems into categories A and B according to the system characteristics. Category A applies when regenerative braking is applied independently of service braking and category B applies when regenerative braking is utilized to provide some of the service braking force.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

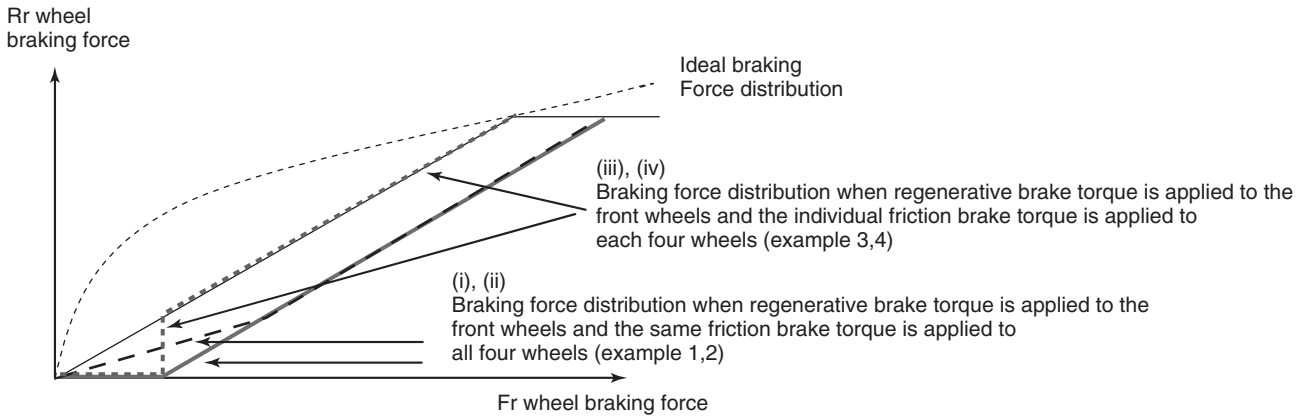


Figure 13. Examples of front/rear braking force distribution when regenerative braking torque is applied to the front wheels and friction braking torque at the same wheel cylinder hydraulic pressure is applied to all four wheels.

4 REGENERATIVE BRAKING AND FRICTION BRAKING CONTROL

Sections 2 and 3 focused on the characteristics of regenerative braking and methods for adjusting braking force through braking systems that utilize regenerative braking and friction braking. This section examines the technologies for securing vehicle dynamic performance and improving fuel efficiency using regenerative braking and friction braking.

There are two categories of regenerative braking systems. With the first, regenerative braking force is applied independently of service braking. With the second, regenerative braking force is applied as part of service braking. In the first system, regenerative braking is performed by an accelerator pedal operation, shifting, or switch operation by the driver, that is, in the same way as engine braking in a gasoline-powered vehicle. In the second system, the vehicle coordinates the regenerative braking control with conventional service braking to generate the required braking force in accordance with the driver's brake pedal operation. The characteristics of these systems are described below.

4.1 Regenerative braking applied independently of service braking

The explanation in this section uses a method that generates braking force equivalent to engine braking in a conventional vehicle and then recovers the kinetic energy from this process (Figure 14). This method is characterized by the fact that the regenerative braking that recovers kinetic energy is isolated from the service braking. For example, the driver can trigger regenerative braking in this case by easing off the accelerator pedal, lifting off the pedal, or pressing a switch while driving (it depends on operation system of vehicle). This is similar to engine braking in

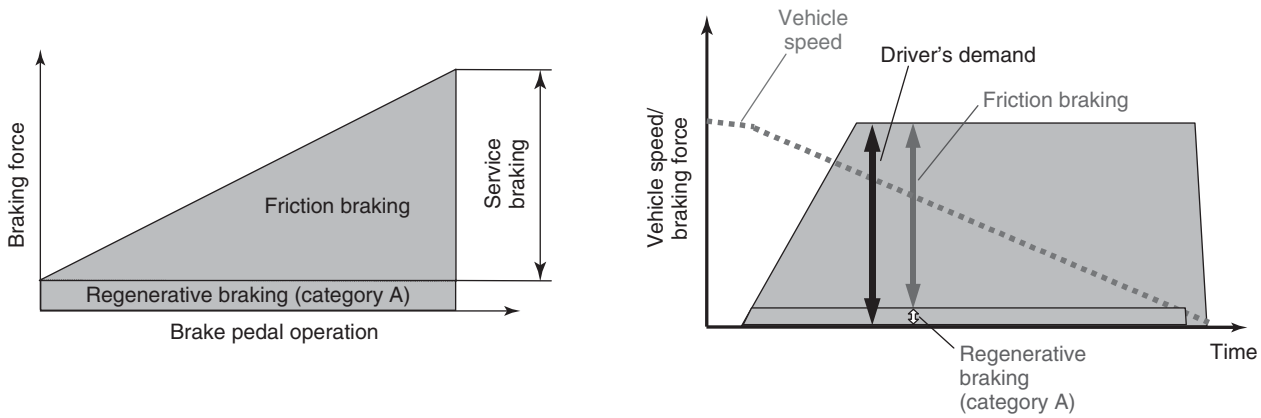


Figure 14. Illustration of regenerative braking applied independently of service braking.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

conventional vehicles. Some HEVs also use technology to supplement braking force by a powertrain control such as engine braking to reduce changes in braking force caused by fluctuations in regenerative braking.

4.2 Regenerative braking applied as part of service braking

The explanation in this section uses a method that expands the application of regenerative braking as part of service braking to recover a larger amount of kinetic energy (Figure 15). More kinetic energy can be recovered by allowing regenerative braking to provide some of the service braking force generated by operating the brake pedal. In other words, this method optimizes the use of regenerative braking within the range of braking force demanded by the driver. Compared to the method described in Section 4.1, the larger amount of recovered kinetic energy allows fuel efficiency to be improved.

The required deceleration for service braking reaches approximately 1 G. However, there are various factors that limit the available regenerative braking force, such as the battery SOC, vehicle speed, and so on. Therefore, regenerative braking alone cannot generate the braking force required by the driver and friction braking must be used to compensate for any insufficiency. At the same time, the regenerative-friction braking coordination control must also achieve the same natural braking feel as a conventional gasoline vehicle, control friction braking independently of the driver's brake pedal effort, and accurately coordinate friction braking in accordance with the fluctuations in regenerative braking.

Figure 16 shows an example of regenerative-friction braking coordination control. As described in Section 2, regenerative braking is not available when the vehicle is not moving. For this reason, the following control is performed to improve fuel efficiency and secure the braking force demanded by the driver. In the initial period of braking, the driver's braking force demand is achieved

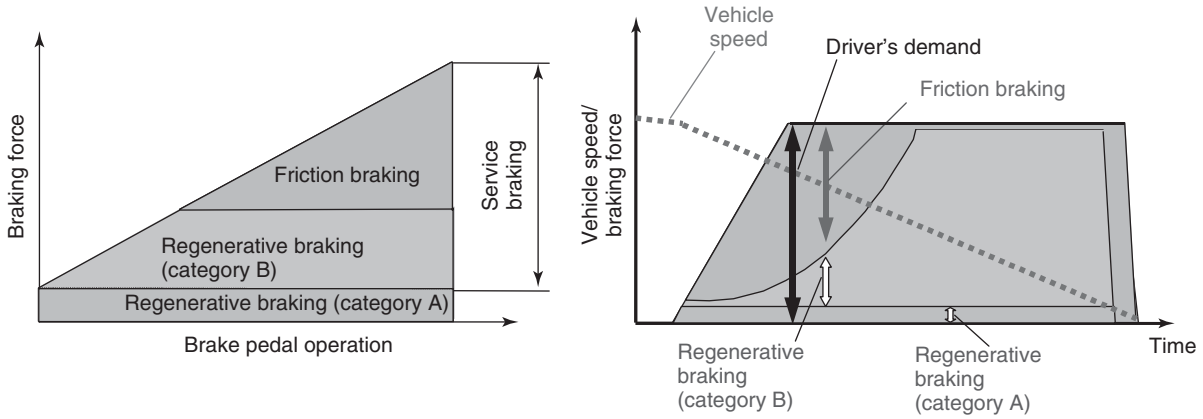


Figure 15. Illustration of regenerative braking applied as part of service braking.

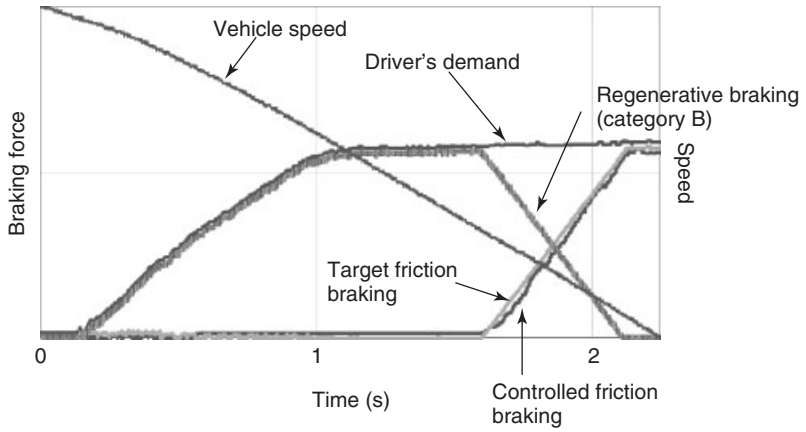


Figure 16. Illustration of regenerative-friction braking coordination control.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

mainly by regenerative braking alone to maximize the kinetic energy recovered during braking. Then, immediately before the vehicle stops, the distribution of braking force from friction braking is increased as regenerative braking is decreased.

Regenerative braking is also reduced in situations such as when the battery has a full charge or a failure has occurred, or when possible vehicle instability is detected. Consequently, the regenerative-friction braking coordination control must be highly accurate to constantly change the distribution between regenerative braking and friction braking. Technological development is advancing to reduce changes in the degree of deceleration while this control is performed (Nakata *et al*., 2009).

4.3 Front/rear braking force distribution depending on drive system

The front/rear braking force distribution is designed to maximize braking force on any road surface. It is generally understood that too much braking force distributed to the front or rear wheels will adversely affect vehicle stability.

It is preferable to increase the proportion of regenerative braking compared to the braking force from friction braking to recover as much energy as possible. However, as the drive system of the vehicle determines the wheels on which regenerative braking is applied, the use of regenerative braking alone to generate braking force will result in over-distribution of braking force to the front wheels in the case of a FWD vehicle or to the rear wheels in the case of a RWD vehicle. Consequently, it will not be possible to achieve the desired front/rear braking force distribution. In other words, regenerative braking has the potential to affect vehicle stability. Therefore, the application of regenerative braking must consider the front/rear braking force distribution as well as coordination with friction braking. The following sections describe the approaches to the front/rear braking force distribution in consideration of typical vehicle drive systems.

4.3.1 Front-wheel drive

In a FWD HEV or EV, the motor is only connected to the front wheels. Therefore, braking the vehicle using regenerative braking alone will result in a disproportionate distribution of braking force to the front wheels. Consequently, the application of regenerative braking in a FWD vehicle may have the following adverse effects on vehicle handling.

4.3.1.1 When regenerative braking is applied independently of service braking. Generally, the application of regenerative braking in this case creates a disproportionate distribution of braking force to the front wheels. When a situation that might affect vehicle handling occurs (e.g., when the front tire slip ratio is high), regenerative braking is reduced to secure vehicle stability.

4.3.1.2 When regenerative braking is applied as part of service braking. The front/rear braking force distribution is generally created by the following types of coordination between regenerative braking and friction braking (Hano and Hakiai, 2011; Nakamura *et al*., 2002).

4.3.1.2.1 Braking systems capable of independently controlling braking force generated by front and rear friction braking. The desired front/rear braking force distribution can be achieved by applying friction braking to the rear wheels while applying regenerative braking to the front wheels (line a in Figure 17). However, when the braking force demanded by the driver is divided between regenerative braking and friction braking, the application rate of regenerative braking will be reduced by the amount of friction braking application. For this reason, it is difficult to maximize energy recovery. As shown by line b in Figure 17, energy recovery can be increased by generating braking force using regenerative braking at the front wheels up to a certain braking force region. Then, once that region is exceeded, braking force is applied to the rear wheels as well by friction braking (Obata *et al*., 2011).

4.3.1.2.2 Braking systems that control braking force generated by friction braking at all four wheels simultaneously. When the braking force demanded by the driver is divided between regenerative braking and friction braking, braking force is generated by regenerative braking at the front wheels up to a certain braking force region (line c in Figure 17). Then, once that region is exceeded, braking force is applied to all four wheels

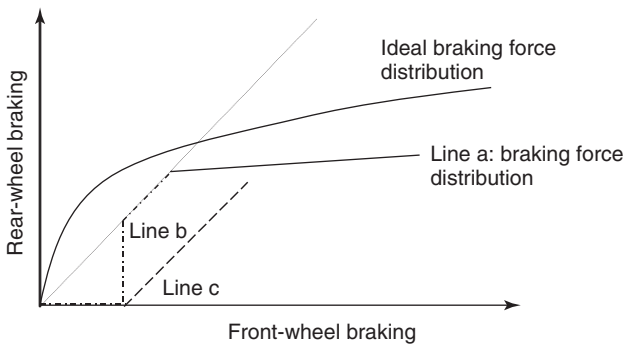


Figure 17. Front/rear braking force distribution.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

as well by friction braking. Alternatively, regenerative braking and friction braking may also be applied at the same time.

For both systems with or without coordinated control between regenerative braking and service braking, the front/rear braking force will be distributed more to the front wheels than in a conventional vehicle. Moreover, in comparison to a conventional vehicle, this may cause a vehicle with regenerative braking to have worse handling or increase the frequency of unexpected intervention of ABS at the front wheels. Consequently, as the front tire slip ratio is high, the general approach is to determine the regenerative braking force from the standpoints of both improved fuel efficiency and vehicle stability.

4.3.2 Rear-wheel drive

In a RWD HEV or EV, the motor is only connected to the rear wheels. Therefore, braking the vehicle using regenerative braking alone will result in a disproportionate distribution of braking force to the rear wheels. Consequently, the application of regenerative braking in a RWD vehicle may make the rear wheels more susceptible to skidding. In the same way as a FWD vehicle, the regenerative braking force and front/rear wheel braking force distribution must be determined in consideration of vehicle stability and dynamic performance.

4.3.3 All-wheel drive

4.3.3.1 Drive system with separate motors connected to the front and rear wheels. As a separate motor is attached to the front and rear wheels, the front/rear wheel braking force distribution can be determined freely by instructing regenerative braking separately to each motor. Therefore, the desired front/rear wheel braking force distribution can be achieved by using the motors to adjust the distribution of regenerative braking force at the front and rear wheels. Additionally, more energy can be recovered than in a FWD or RWD vehicle without disrupting the desired front/rear wheel braking force distribution.

4.3.3.2 Drive system with one motor connected to the front and rear wheels. In this system, the front/rear wheel regenerative braking force distribution is determined by the characteristics of the center differential gear or the like. Normally, as the front/rear wheel braking force distribution becomes weighted toward the front or rear wheels depending on the situation, the amount of energy recovery from regenerative braking and the use of friction braking are generally determined in consideration of vehicle stability and dynamic performance in the same way as a FWD or RWD vehicle.

4.4 Dynamic control and regenerative-friction braking coordination functions

4.4.1 Applications to driving assist systems

Cruise control is one example of a recently developed driving assist system that is starting to become more popular in HEVs and EVs. Regenerative braking is also starting to be applied to cruise control systems so that regenerative braking force can be generated during deceleration while these systems are in operation. In this situation, ways are being found to increase opportunities to recover energy and improve fuel efficiency even when driving assist systems are active.

4.4.2 Applications to safety functions

A growing number of vehicles are being provided with ABS and ESC functions to enhance safety performance. These functions require braking force to be adjusted independently at all four wheels. This makes it difficult to operate ABS and ESC functions using regenerative braking that can only generate braking force at the wheels to which the motors are connected.

Therefore, when ABS or ESC is in operation, the general approach is to reduce regenerative braking and supplement the braking force using friction braking. In this case, friction braking can be controlled as required using information from wheel speed sensors, accelerometers, and the like. This control can use conventional technology for adjusting braking force independently at all four wheels. However, as described in Section 4.2, the accuracy of the regenerative-friction braking coordination control technology must be enhanced to vary the distribution between regenerative braking and friction braking. Technological research has made progress in recent years to actively adopt regenerative braking in ABS and ESC functions by making use of the excellent response of regenerative braking and the ability to accurately identify output torque (Fujimoto, 2011; Hori, 2004; Murata, 2011; Okano *et al*., 2002).

RELATED ARTICLES

Chassis ECU (Vehicle dynamics, ABS)

REFERENCES

von Albrichsfeld, C. and Karner, J. (2009) Brake System for Hybrid and Electric Vehicles. *Proceedings of SAE World Congress*, Detroit, USA.

Fujiki, N., Koike, Y., Itou, Y., *et al*. (2011) Development of an Electrically-Driven Intelligent Brake System for EV. *Proceedings of EVTeC*, Yokohama, Japan.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fujimoto, H. (2011) Regenerative Brake and Slip Angle Control of Electric Vehicle with In-Wheel Motor and Active Front Steering. *Proceedings of EVTeC*, Yokohama, Japan.

Hano, S. and Hakiai, M. (2011) New challenges for brake and modulation systems in hybrid electric vehicles (HEVs) and electric vehicles (EVs). SAE Technical Paper 2011-39-7211.

Hori, Y. (2004) Future vehicle driven by electricity and control—research on 4-wheel motored UOT March II *IEEE Transactions on Industrial Electronics*, **51**(5), 954–962.

Murata, S. (2011) Vehicle Dynamics Innovation with In-Wheel Motor. *Proceedings of EVTeC*, Yokohama, Japan.

Nakamura, E., Soga, M., Sakai, A., *et al*. (2002) Development of electronically controlled brake system for hybrid vehicle. *SAE International*, Warrendale, USA.

Nakata, D., Nakamura E., Fukasawa, T., and Ohya, K. (2009) Development of the spread type electronically controlled brake system for hybrid vehicle. JSAE Technical Paper 20095631.

Obata, T., Ohtani, Y., Shirakawa, N., *et al*. (2011) Development of electronically-driven intelligent brake actuator with regenerative braking system. JSAE Technical Paper 20115172.

Okano, T., Sakai, S., Uchida, T., and Hori, Y. (2002) Braking Performance Improvement for Hybrid Electric Vehicle Based on Electric Motor's Quick Torque Response. *Proceedings of 19th Electric Vehicle Symposium (EVS19)*, Pusan, South Korea.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto053
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Energy Management System of EVs

Masayuki Komatsu

Toyota Motor Corporation, Toyota, Japan

1 Introduction 1
2 EV Systems 1
3 Battery System Management 3
4 Charging Systems 6
5 Auxiliary System Energy Management 9
Related Articles 10
References 10

1 INTRODUCTION

Electric vehicles (EVs) are powered by rechargeable onboard batteries. Although conventional EV batteries are still chemical-based (e.g., lead storage batteries), alternative means have also been developed, including flywheel batteries that store mechanical energy and capacitors that store electrical energy. These are commonly known as *rechargeable electrical energy storage systems* (*RESSs*).

The main performance requirements for onboard EV batteries are (i) single charge driving range, (ii) a short charging time, (iii) state-of-charge (SOC) display accuracy, (iv) lifetime, and (v) safety of handling. Owing to these extensive battery requirements, energy management is a considerably more important design issue for an EV than for a gasoline or hybrid electric vehicle (HEV). Energy management in an EV is performed by the battery management system.

The most fundamental way to extend the single charge driving range is to increase the number of onboard batteries. In this case, the key point is to balance the trade-off relationship between the increased battery weight and the subsequent reduction in usability caused by lower dynamic performance (slower acceleration and the like) and longer charging time. For these reasons, manufacturers are working to increase the popularity of EVs by minimizing the effects on usability by enabling the battery to be switched over or by providing rapid-charging infrastructure, even when the EV has limited battery capacity. Therefore, in a broad sense, the method energy management adopted in an EV is not determined simply by the design of the vehicle and battery systems and also depends on the state of the social infrastructure. Research has also been started into the use of EVs themselves as a RESS to supply power to homes. However, at this stage, the establishment of technology for energy management is awaiting further research for practical application.

The main object of energy management is to calculate the battery SOC. As the design requirements for detecting, calculating, and managing the battery state differ depending on the type of battery and the configuration of the EV propulsion system, this chapter focuses on the configuration and outline of onboard EV systems.

2 EV SYSTEMS

Figure 1 shows the configuration of a general EV system. Its main components are the RESS (i.e., the battery system for storing electrical energy), the charging system, the electric propulsion system which treat high power, the auxiliary system that includes the temperature control unit, power steering, and the like, and the energy management unit that integrates and controls these systems. The flow of signals between these component systems are shown by the bold lines. The basic flow of energy is as follows. AC from

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto054
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

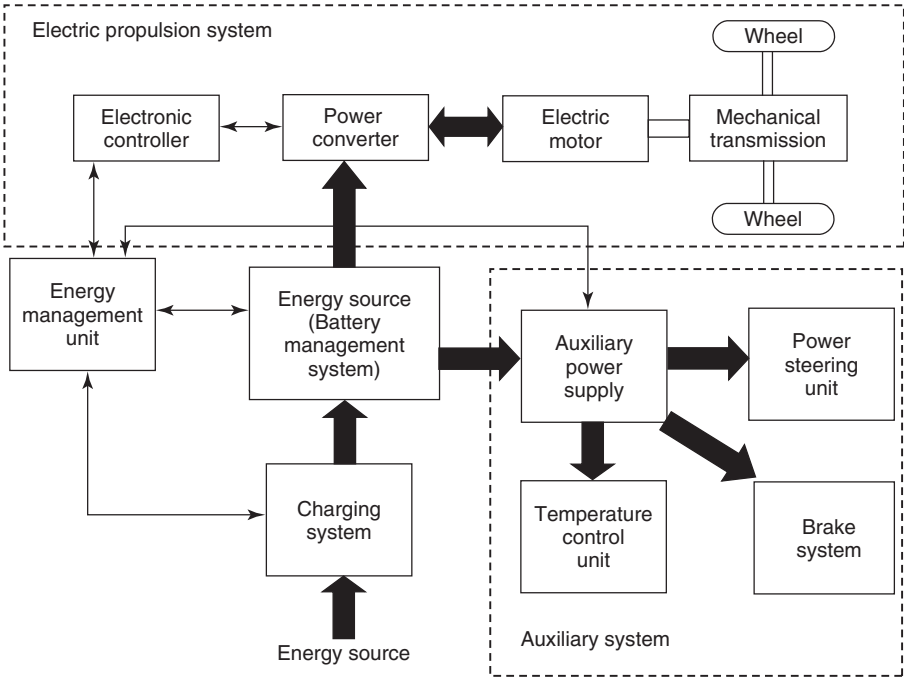


Figure 1. General EV system configuration. (Reproduced from Chan and Chau (2001). By permission of Oxford University Press (www.oup.com).)

an external energy source is passed through the charging system, converted to DC, and used to charge the RESS. The energy required to propel the vehicle is then supplied from the RESS to the electric propulsion system. In turn, kinetic energy generated when the vehicle brake is used to recharge the RESS as regenerated energy through the drive motor that acts as a generator. Furthermore, as the RESS supplies the energy required by the temperature control unit, power steering, and the like, reducing the power demand from the auxiliary system is a key part of extending the single charge driving range.

The energy management unit in Figure 1 functions as the vehicle control unit and battery management unit (BMU). However, the system design may combine the vehicle control unit with the motor system control unit. Other system configurations may place the vehicle control unit with the battery control or motor drive control units, or even by itself. In these systems, as the RESS is typically a lithium ion (Li-ion) or other rechargeable secondary battery, the following sections describe this type of battery system (Chan and Chau, 2001).

2.1 Energy management unit

The energy management unit functions as the vehicle control unit. It detects the driver's operations and the state of the component EV systems (i.e., the battery system, charging system, electric propulsion system, and auxiliary system including the temperature control unit) and manages the SOC of the battery system and the power used to propel the vehicle. The vehicle control unit may also be provided separately. Its configuration can be designed freely depending on the vehicle.

2.2 Rechargeable electrical energy storage system

The RESS is the source of propulsion energy. It recovers and stores the regenerative kinetic energy generated by the vehicle on braking and deceleration. Although some types of RESS use flywheel batteries or capacitors, most are rechargeable secondary batteries. Conventional lead-acid, nickel–cadmium, and nickel-metal hydride batteries have begun to be replaced by Li-ion batteries that have higher volumetric and gravimetric energy densities. The most important function of the RESS is to detect the state of the battery (i.e., the charge and discharge volumes, temperature, and the like) and to measure the SOC.

2.3 Charging system

The charging system supplies the power to be used by the EV from an external power source to the secondary battery.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto054
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

There are two basic types of charging system: one that converts commercially available AC to DC and the other that directly charges the battery with DC from a stationary charger. The former can use household power sources, whereas the latter is used in DC rapid charging systems.

The charging system consists of a connector on the power supply side called a *charging coupler* and an *inlet* on the vehicle side. The coupler may use a conductive or noncontact inductive electrical connection system. Both AC and DC conductive systems are available, and there are also differences in the number and layout of the conductive connecting pins. International efforts are being made to standardize both the charging system and the coupler. In addition, rather than charging the battery onboard the vehicle, methods have been developed in which the electrical system itself is removed, charged outside the vehicle, and replaced.

2.4 Electric propulsion system

This system converts power from the battery to power for driving the vehicle. It generally comprises a motor. The motor used may be categorized in accordance with the type of induction voltage applied to the input terminals, its excitation system, or its field system. Common examples include the DC motor, permanent magnet synchronous motor, induction motor, and switched reluctance (SR) motor. Typically, an AC motor is used with inverter control to increase the efficiency and controllability of the motor drive.

2.5 Auxiliary system

In conventional internal combustion engine (ICE) vehicles, the ICE provides the source of power and heat for the auxiliary units. As this is not available in an EV, all auxiliary units have to be powered by electricity. Typical examples of auxiliary units include the air conditioning, electronic power steering, power brake booster, and the like.

3 BATTERY SYSTEM MANAGEMENT

3.1 Configuration

In an EV or HEV, the battery management system refers to the battery used for driving the vehicle and its associated monitoring units. The main components of the battery management system are generally as follows:

- the battery pack;
- the battery assemblies;
- the battery monitoring system; and
- the battery cooling system.

The battery management system performs the following functions:

- storage of energy for driving the vehicle;
- control and maintenance of battery properties in the optimum state for driving; and
- maintenance of battery safety and reliability.

This section describes the battery monitoring system, which ensures the optimal function of the driving battery.

The main function of the battery monitoring system is to identify the state of the battery pack and to transmit control data based on this information. Figure 2 shows an example of a battery monitoring system for an EV that uses a Li-ion battery. This system measures various battery parameters in the battery modules inside the battery pack. It includes cell monitoring units (CMUs) that send this data to the BMU. The CMUs measure the voltage of each cell and the temperature inside the battery modules. At the same time, the BMU also collects information such as the current of the main high voltage circuits and the insulation resistance of the high voltage system from sensors installed inside the battery pack.

The BMU monitors the battery state based on this information. It then sends commands to maintain optimal battery function to the overall vehicle control unit, the battery cooling system controller, and the CMUs.

The specific primary functions of the battery monitoring system are as follows:

- battery protection;
- remaining power calculation;
- battery lifetime calculation; and
- fault diagnosis.

Of these, battery protection by monitoring the battery voltage and temperature and battery fault diagnosis are the most important functions. For this reason, the BMU at the core of the battery monitoring system is preinstalled with various control information in accordance with the chemical system of the battery (The Society of Automotive Engineers of Japan, 2011).

The following sections describe the main items related to the battery protection function.

3.1.1 Over-charging prevention function

Over-charging occurs when the upper limit voltage of a cell is exceeded and may be a cause of heat or smoke

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto054
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

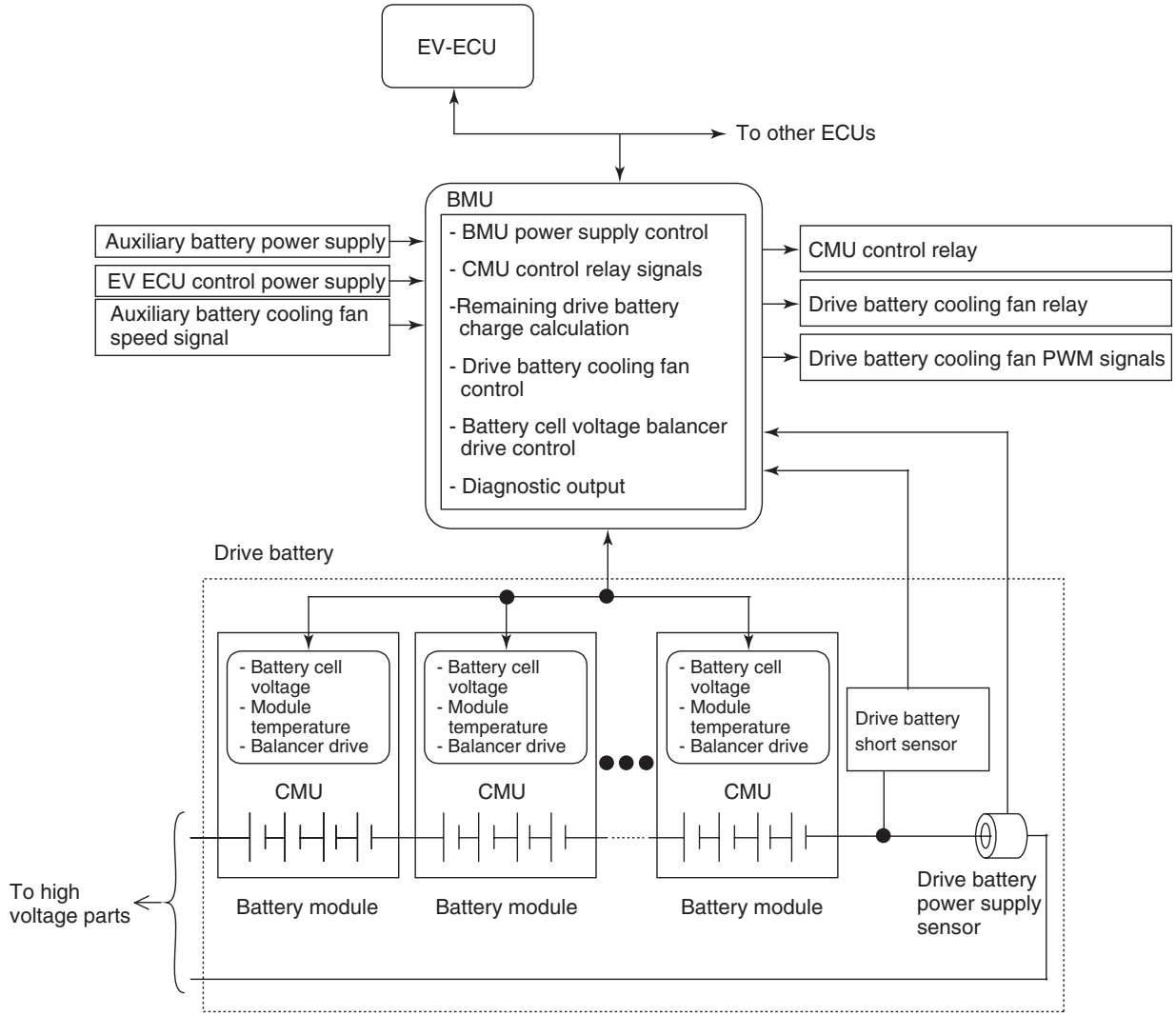


Figure 2. Diagram of battery monitoring system. (Reproduced with permission from Society of Automotive Engineers Japan (2011). © Society of Automotive Engineers Japan.)

generation in addition to battery performance deterioration. The voltage of each cell is monitored to prevent this, and the charge and regeneration currents are controlled to make sure that the upper limit voltage is not exceeded (Society of Automotive Engineers of Japan, 2011).

3.1.2 Over-discharging prevention function

The chemicals used in the cells have a particular lower limit voltage. Over-discharging occurs when the voltage falls below this level. Over-discharging causes irregular chemical reactions to occur inside the battery, which alters the internal composition of the battery and may prevent the battery from being used if allowed to continue. For this reason, the output current is controlled to prevent the cell voltage falling below its lower limit when the vehicle is being driven (Society of Automotive Engineers of Japan, 2011).

3.1.3 Voltage equalization function

As discussed earlier, an EV is generally propelled by a multiple number of cells that are connected in series. If the voltage of each cell is different, the cell with the lowest voltage will affect the performance of all the cells, and the battery pack will not function as designed. This is usually prevented by a voltage equalization circuit (a balancer circuit) provided in the CMU or BMU. Resistance absorption or power transfer type circuits are most common (Society of Automotive Engineers of Japan, 2011).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto054
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.1.4 Over-heating prevention function

This function prevents each cell from exceeding its recommended operating temperature range. Continuous driving or rapid charging at maximum power generates heat because of the internal resistance of the cells. Exceeding the upper limit temperature may affect battery capacity or output performance or lead to issues such as blistering of the battery. The CMUs measure and monitor the temperature of each cell or the battery modules in a typical unit. The upper limit temperature is protected by reducing the output or charging current and by using the battery cooling system to forcibly bring the temperature of the battery down (Society of Automotive Engineers of Japan, 2011).

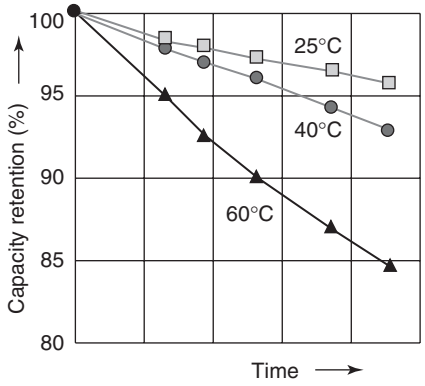


Figure 3. Example of battery deterioration temperature dependence test results. (Reproduced with permission from Society of Automotive Engineers Japan (2011). © Society of Automotive Engineers Japan.)

3.2 Battery cooling system

As mentioned earlier, the usage conditions of the battery may cause heat generation. In addition, the battery has a predetermined operating temperature range. Figure 3 shows the correlation between capacity reduction and temperature as a battery is used. When the cell temperature rises, the chemical substances inside the cells undergo changes, which clearly reduce the battery capacity. For this reason, an upper limit temperature is set and a battery cooling system is generally provided to ensure that this is not exceeded.

Figure 4 shows the configuration of a commonly adopted air cooling system. Air from the vehicle interior is blown onto the battery pack through cooling ducts. This air mainly flows through passages provided between each battery assembly and it is expelled through an outlet at the rear. In this configuration, the battery assemblies use this cooling air to expel heat to the rear. Some structures also include air passages between the cells in the battery assemblies to enable more efficient cooling.

The key points in designing a battery cooling system are the control timing to ensure that the cells do not exceed the upper limit temperature (e.g., the timing of input/output

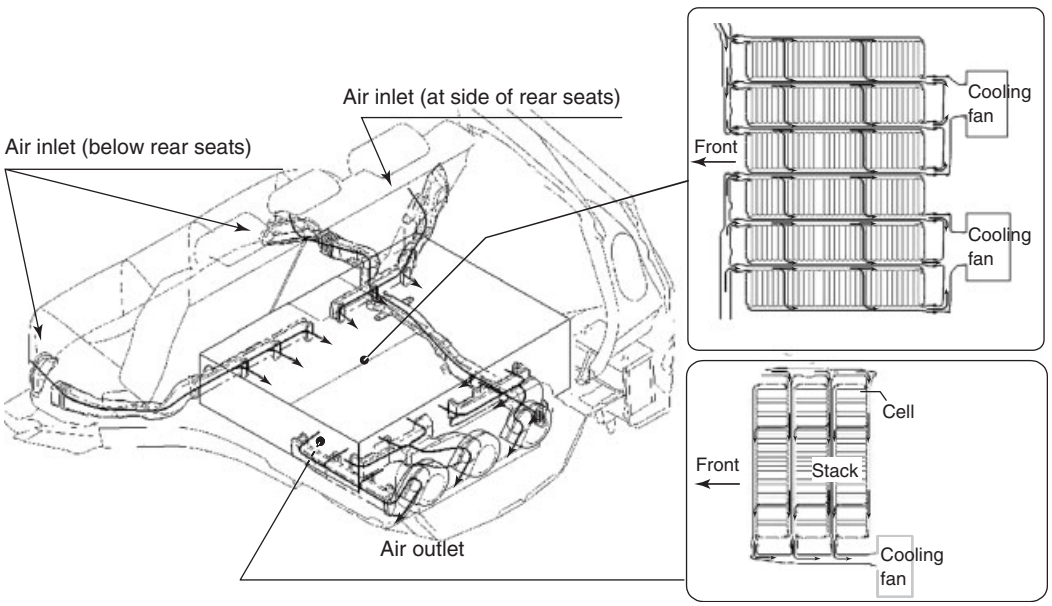


Figure 4. Example of air-cooling structure configuration inside battery pack. (Reproduced with permission from Onomura *et al*. (2010). © Toyota Motor Corporation, Japan.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto054
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

voltage reduction, cooling air introduction timing, and the like) and preventing temperature differences between the cells. As cell performance is strongly dependent on temperature, cooling may generate differences in cell voltage if it leads to the creation of different temperatures between the cells.

Consequently, cell temperature equalization is important. It relies on the basic battery pack structure related to the air distribution inside the pack (particularly the cooling air duct structure), the layout of the battery assemblies, the way that cooling air is supplied inside the battery assemblies, and the like (Society of Automotive Engineers of Japan, 2011).

3.3 Battery charge/discharge management

The battery charge/discharge state management technology plays the most important role in ensuring that the battery is used safely and in extending the charge/discharge lifetime. The characteristics of charge/discharge state estimation differs depending on the battery type, the control method of the battery management system, and the cooling method. The SOC is a critically important parameter in battery control, and its identification methods differ greatly depending on the battery. The characteristics of some typical batteries and their SOC management technologies are described later.

3.3.1 Lead-storage batteries

Lead batteries are comparatively stable with low energy density, high power density, and low self-discharge. The state of lead batteries (including the state of deterioration) is difficult to detect as it requires measurement of the specific gravity of the electrolyte. To counter this issue, technology was developed that estimates the battery SOC and deterioration from premeasured maps of the relationship between the open-circuit voltage and internal battery resistance (EV Handbook, 2001).

3.3.2 Nickel-metal hydride batteries

Nickel-metal hydride batteries have extremely well-balanced basic properties from the standpoint of energy density, power characteristics, regenerability, and lifetime, as well as from the standpoints of safety and recyclability. As the characteristics of nickel-metal hydride batteries change depending on the temperature, SOC is generally detected by measuring the temperature and voltage of the battery. SOC detection methods that also incorporate energy calculation have also been developed (EV Handbook, 2001).

3.3.3 Li-ion batteries

Li-ion batteries have the highest energy density, voltage, and long-term capacity retention characteristics of the various types of secondary batteries in practical use. Although the SOC of Li-ion batteries is generally calculated from the open-circuit voltage, voltage plateaus may be generated depending on the electrode material, which makes SOC estimation difficult. In these cases, the SOC is derived from fixed equivalent circuit parameters or the modeling of state variables (EV Handbook, 2001).

4 CHARGING SYSTEMS

Chargers and vehicle connection formats can be broadly categorized into direct-contact (conductive) and noncontact (inductive) types. Although ZEV rules in the United States include both types, conductive charging has been adopted for the creation of charging standards.

External large-capacity rapid chargers directly charge the vehicle battery with DC current by converting the supplied power to DC outside the vehicle. The most common proposed capacity for rapid charging is approximately 50 kW, in consideration of the ability of the battery to accept charge. Large-capacity household charging requires harmonization between the electrical specifications of the vehicle and the wiring of the building. Creating the necessary specifications for power supply, electrical distribution, and vehicle charging (i.e., the degree of electrical power reception by the vehicle) is an issue that transcends vehicle design and also includes the question of how far an EV can be made compatible with the power supply situation in each country.

In the latest developments related to inductive charging, feasibility studies are under way using large vehicles such as buses that drive over fixed routes. Inductive charging involves installing coils in the vehicle and infrastructure. In this way, charging takes place automatically and the required power is provided on the route of the vehicle. Verification tests are taking place and show promise for use with new forms of public transport such as trolley buses with no overhead cables.

In contrast, battery switching stations have also been suggested as an alternative to expanding the charging infrastructure. Shortening the time required to change the battery is a promising way of providing the same utility as rapidly charging a large-capacity battery. Feasibility studies carried out in Japan have discussed the difficulty of establishing these stations as a self-sustaining market depending on how customers perceive the added economic value of this service. However, as the reduction of CO_2

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto054
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

emissions and the adoption of alternative energies have emerged as pressing social issues, the number of battery switching stations may increase in certain local markets in accordance with the political needs of the country or region.

4.1 EV charging systems

Charging systems have been identified earlier as one of the key issues of EVs. The charging system of EVs is the same as that of plug-in hybrid electric vehicles (PHEVs). However, as the installed battery capacity of a PHEV is lower than that of an EV, it has a smaller need for large-capacity rapid charging. PHEVs are also designed to be mainly charged from normal charging sources (i.e., household outlets) from the standpoint of the battery life. On the basis of these characteristics and usability, normal charging is a significant merit from the standpoints of being able charge a vehicle easily using conventional existing household outlets.

Consequently, the specifications of chargers depend on the infrastructure available to normal households. In Japan, normal household outlets are rated at 100 V/15 A or 200 V/maximum 30 A. In addition, the maximum current under Japan's Meter-Rate Lighting B power supply contract for general households is 60 A. For this reason, charger output is likely to be several kilowatts. A small-output power charger like this is both compact and light, which means it can be used anywhere. Taking advantage of this fact, the charger can also be installed on the vehicle side. Figure 5 is a block diagram of a normal charging system featuring an in-vehicle charger. Usual charging configurations can be categorized into types that insulate or do not insulate the charging system through an electrical circuit by providing a transformer or the like between the charger side and the high voltage vehicle parts. The configuration in Figure 5 is the commonly used insulated type (Ishikawa, 2010).

Examples of the noninsulated type include the charger for Tesla Motor's Roadster EV, which was integrated with the main traction inverter circuit, and the charger for Toyota Motor Corporation's e-com EV, which was integrated with the circuits for the drive motor coil and traction inverter (Ishikawa *et al*., 1997).

In addition, Figure 6 shows some of the components of a charging system, such as the lid of the vehicle charging socket and the charging cable that includes a connector and charging circuit interruption device (CCID, that is, a ground fault breaker or the like).

The following sections describe the components in Figure 6 in more detail.

4.1.1 Charger

Normally, the charger is provided on the vehicle side. The charger charges the battery by converting external AC input to DC output through a high frequency transformer by semiconductor switching. Although the charging method depends on the battery characteristics, forced charging by constant power (CP) control (i.e., CP charging) or at a constant current/constant voltage (CC/CV charging) can also be performed by adopting a multistep CC charging method or the like. Figure 7 shows an example charging circuit configuration and its specifications and characteristics. Similar to a normal AC/DC converter, the circuit

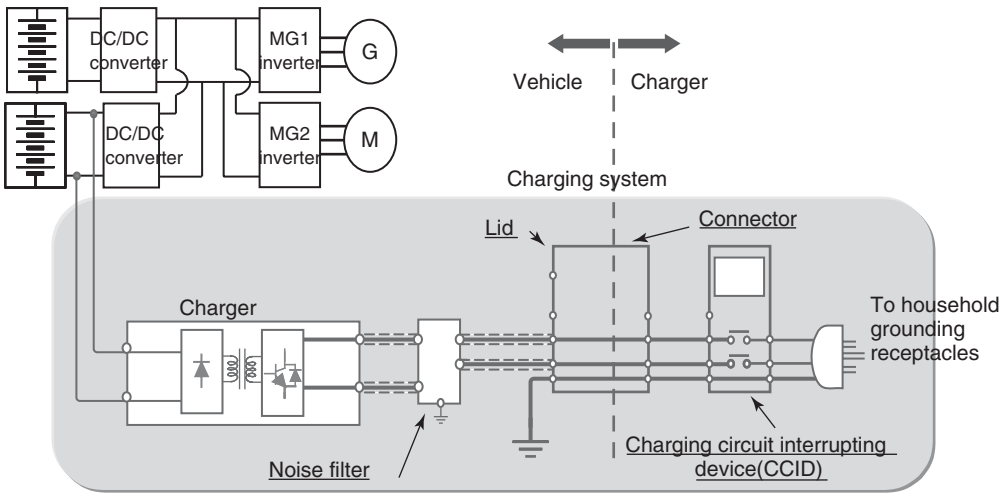


Figure 5. Configuration of commonly used charging system. (Reproduced with permission from Society of Automotive Engineers Japan (2011). © Society of Automotive Engineers Japan.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto054
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

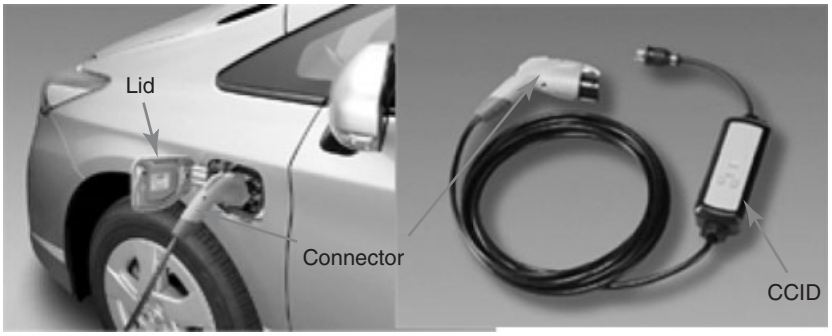
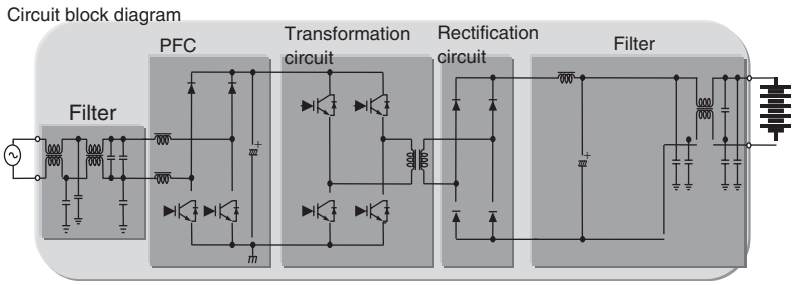


Figure 6. Charging system components. (Reproduced with permission from Society of Automotive Engineers Japan (2011). © Society of Automotive Engineers Japan.)



Specifications and characteristics

Input	AC 100 V/AC 200 V
Max. output	2 kVA
Max. efficiency	88% (total efficiency 83%)
Carrier frequency	21/50kHz (PFC/full bridge)

Figure 7. Example charger circuit configuration and specifications. (Reproduced with permission from Society of Automotive Engineers Japan (2011). © Society of Automotive Engineers Japan.)

consists of a filter, power factor corrector (PFC), bridge circuit, high frequency transformation circuit, and rectifying circuit.

4.1.2 Noise filter

A normal multilayer LC filter is provided to regulate harmonics and reduce line noise.

4.1.3 Vehicle charging socket lid and connector

The charging system is provided with a connector standardized for use in Japan, the United States, and Europe.

4.1.4 Charging circuit interruption device (ground fault breaker or the like)

Regulations in the United States require the use of a ground fault breaker and predetermined signal function (see SAE J1772) to guarantee safety when high voltages are applied.

4.2 Key points of normal charging systems

As mentioned earlier, the key point of a normal charging system is the extremely small power available. Assuming that the vehicle will be charged at a normal household, the available power will be 1.5 kVA for a 100 V system and a maximum of 6 kVA for a 200 V system, even in the case of commercial-use power sources. Consequently, ECUs and other auxiliary equipments consume a large proportion of the power during charging. For this reason, the total charging efficiency concept shown in Figure 8 becomes important. In specific terms, key issues are reducing the electricity consumption of auxiliary vehicle equipment during charging and improving the efficiency of the charger.

4.3 Public charging

Figure 5 shows the configuration of a normal household charging system (also referred to as *Mode 2 charging*). However, design of the charging system must also consider charging of the vehicle at public charging stations (Mode 3 charging). Figure 9 shows an example connection configuration and specifications of a public charging station. As the illustrated station complies with the requirements of SAE J1772, the charging system on the vehicle side can follow the same control sequence as that shown in Figure 5. However, it is also likely that some charging stations will not comply with this standard. For this reason,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto054
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

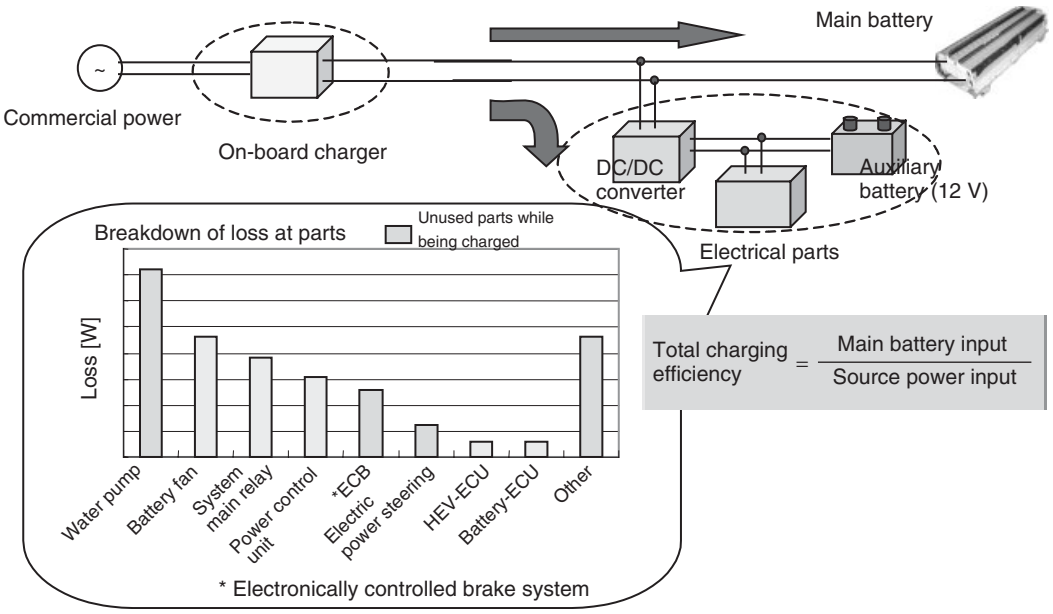
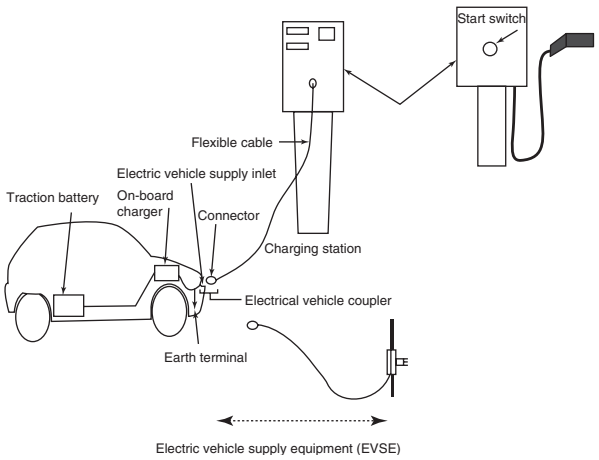


Figure 8. Total charging efficiency. (Reproduced with permission from Society of Automotive Engineers Japan (2011). © Society of Automotive Engineers Japan.)



	Item		Specifications
Basic specifications	Electric	Power supply	Single phase AC 200 V 20A
		Rated frequency	50/60 Hz
		Over current protection	Mounted on the main control board (MCB) (20 A)
		Electric shock protection	Ground fault breaker (current sensitivity 15 mA)
		Pilot signal (CPLT) function	CPLT function compliant with SAE J1772
	Connector		Connector with cable recommended by Japan Automobile Research Institute (JARI)
	Condition		Indoor/outdoor operating temperature : –10 to 40°C

Figure 9. Example public charging station configuration and specifications. (Reproduced with permission from Society of Automotive Engineers Japan (2011). © Society of Automotive Engineers Japan.)

the circumstances of the infrastructure, particularly Mode 3 public charging stations, must also be investigated when designing the vehicle side charging system.

5 AUXILIARY SYSTEM ENERGY MANAGEMENT

In the same way as conventional ICE vehicles, EVs also require a large number of auxiliary devices and systems to enhance the functionality of the vehicle and the comfort of the driver. Some of these systems, including the battery charger, BMS, regenerative brake system, and the like, are not required for ICE vehicles. In contrast, many are common between the two vehicle types, including the heating, ventilation, and air-conditioning (HVAC) system, auxiliary power supply, power steering system, navigation system, lighting, defroster, wipers, and radio. However, even when the purpose of the device is the same, its configuration often has to be designed especially for use in an EV.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto054
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

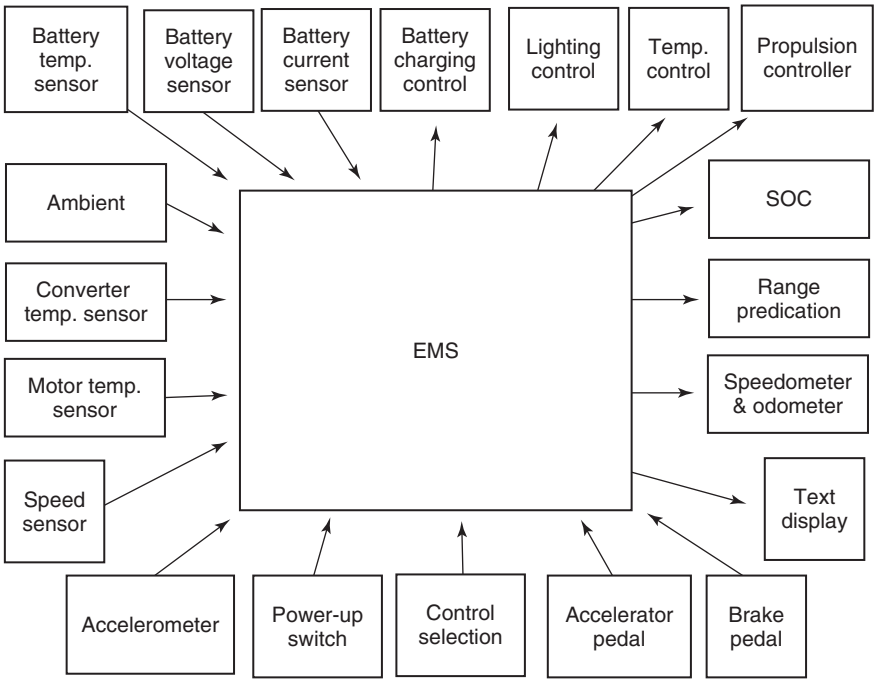


Figure 10. Energy management system. (Reproduced from Chan and Chau (2001). By permission of Oxford University Press (www.oup.com).)

As many auxiliary systems used in current EVs affect energy management, it is extremely important to improve their efficiency. The energy management system (EMS) plays a major role in systematically controlling the energy used by these systems.

As shown in Figure 10, the EMS is connected to the BMS, charging system, and the auxiliary device subsystems. It functions to efficiently manage the consumption of the limited available energy by identifying the state of these connected systems. The typical functions of the EMS are as follows (Chan and Chau, 2001).

- adjustment of power supplied to auxiliary systems;
- regulation of internal temperature control functions of EV components;
- display of available driving range; and
- limitation of energy used for interior HVAC.

RELATED ARTICLES

Regenerative Braking Systems
Energy Management System of HEVs
Battery Charging Standards
Power and Energy Requirements for Electric and Hybrid Vehicles
Fuel Cell Powered Vehicles

REFERENCES

Chan, C.C. and Chau, K.T. (2001) *Modern Electric Vehicle Technology*, Oxford University Press, Oxford, UK.

Editorial Committee of Electric Vehicle Handbook. 2001) *Electric Vehicle Handbook*, Maruzen Publishing Co., Ltd., Tokyo.

Ishikawa, T. (2010). Symposium documents of Eco-car Conference in Tottori Prefecture.

Ishikawa, T., Sekimori, T., Suzuki, A., and Hotta, T. (1997). Development of a traction inverter with charging function. Electric Vehicle Symposium **14**.

Onomura, Y., Inazu, M., Ito, M., *et al*. (2010) Secondary Battery Development for Hybrid Vehicle in TOYOTA in *Toyota Technical Review*, Toyota Motor Corporation, Technical Administration Division, Japan.

Society of Automotive Engineers of Japan (2011) *The Handbook of Automotive Engineering No.10: Design (EV & Hybrid Vehicles)*, Society of Automotive Engineers of Japan, Japan.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto054
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Energy Management System of HEVs

Shinichi Abe
Toyota Motor Corporation, Toyota, Japan

1 Introduction 1
2 Improvement in Fuel Efficiency Through Full Hybridization 1
3 Energy Management in Parallel Type Mild Hybrid System (IMA) 6
4 Thermal Energy Management in HEVs 8
5 PHEV System Configuration 10
6 Features of PHEVs 11
Related Articles 12
References 12

1 INTRODUCTION

The basic configuration of a hybrid system can be categorized as either a series or a parallel type. A series-type HEV only uses the engine to generate electrical power. The generated electrical power is then used to drive electric motors that propel the vehicle. Surplus generated electrical power is stored in a secondary battery. The secondary battery provides additional power when the demand for power from the vehicle is large; for example, when the vehicle starts moving from a stop or when it is accelerating.

In a parallel type HEV, both the internal combustion engine (ICE) and the electric motor are mechanically coupled to the drive train. The ICE can provide mechanical power directly to the drive train. In addition, the electric motor can also provide mechanical power to the drive train using electrical power from a secondary battery. The ICE and electric motor can either work together or independently to provide mechanical power to the drive train.

The hybrid system in the Toyota Prius splits the power from the engine into two parts using a planetary gear set. A portion of the engine power is use to power a generator, whereas the rest of the engine power is use to directly propel the vehicle. This is sometimes called a *power split hybrid system*. This system is primarily known for its application in the Toyota Hybrid System (THS), but it can also be found in the two-mode hybrid system developed by the alliance between GM, Daimler, Chrysler, and BMW. This two-mode system features an extra planetary gear set compared to the THS, which gives it a wider control range for distributing mechanical and electrical power.

HEVs can also be classified as having *strong* or *mild* hybrid systems, depending on the functions of the system. Figure 1 illustrates the differences between a strong and a mild hybrid system.

Figure 2 shows the operating principle of a general strong hybrid system.

Energy management in an HEV refers to how the vehicle is controlled to optimize fuel efficiency while responding to the demands of the driver.

2 IMPROVEMENT IN FUEL EFFICIENCY THROUGH FULL HYBRIDIZATION

Full hybridization improves the fuel efficiency of a vehicle in the following four ways:

(a) At times of low engine efficiency when the engine is running at light loads (including when the engine is idling), full hybridization allows the engine to be

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

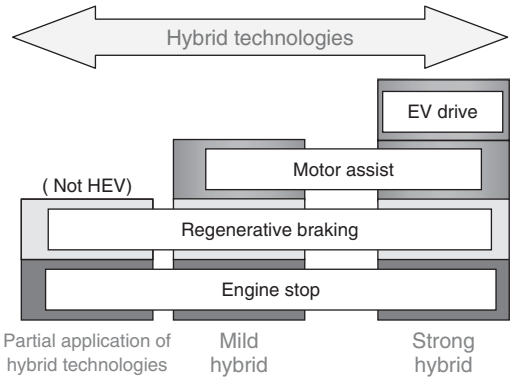


Figure 1. Classification of hybrid systems.

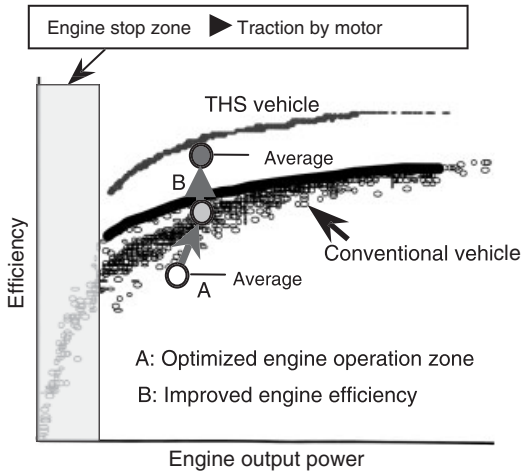


Figure 3. Efficiency improvement by full hybridization.

stopped and the vehicle to be operated as an electric vehicle (EV).

(b) Enhanced continuously variable transmission (CVT) functionality (i.e., the series/parallel hybrid configuration) and EV functionality [above-mentioned point (a)] enable the engine to be operated with optimum efficiency.

(c) A highly efficient engine that is sized for steady state operation can be used as power assist from the battery is available to meet peak power requirements. For example, full hybridization allows the adoption of an engine that uses a high expansion ratio cycle called the *Atkinson cycle*, which can be designed for optimum efficiency. It may also be possible to use a smaller engine depending on performance requirements.

(d) The electric motor can be operated as an electric generator, which converts the vehicle's kinetic energy to electric energy during braking. The electric energy is then stored in the battery to be reused later as kinetic energy. The vehicle's kinetic energy is usually dissipated as heat into the atmosphere by the mechanical friction service brakes.

The key point of a hybrid system is combining the above four merits to achieve optimal fuel efficiency. It is also critically important to use highly efficient and optimized individual components, such as the engine and motors. Figure 3 shows the differences in thermal efficiency between a vehicle with a conventional gasoline engine and one installed with the THS in the urban test cycle used in the United States.

2.1 Basic configuration of THS

Figure 4 shows the configuration of the THS. This system controls the distribution of power from the engine and the motor through a planetary gear set. The resultant driving force is composed of directly transmitted torque from the

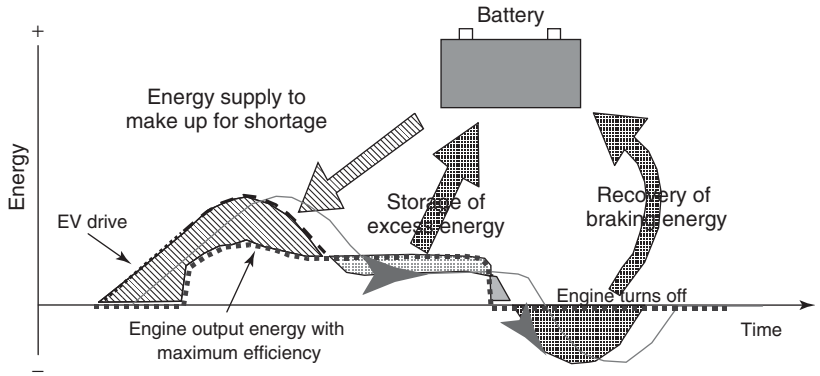


Figure 2. Energy management of strong hybrid.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

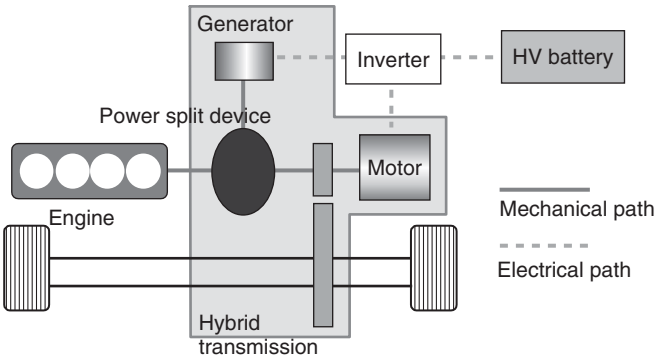


Figure 4. Configuration of THS.

engine and torque from the motor. The THS is also capable of stopping the engine and running the vehicle on motor power alone. Battery power also assists engine power under sharp acceleration.

2.2 Engine

A fuel-efficient engine is a prerequisite for maximizing system efficiency. High efficiency is the aim of an HEV, and the displacement of the engine can be selected freely within the limits of the required engine power and installation conditions. This allows friction losses to be reduced by adopting a high expansion ratio cycle with late closing of the intake valves and by lowering the maximum engine speed. The greater the displacement, the more it is possible to lower the maximum engine speed. Furthermore, friction losses can be reduced at the same power level by lowering the spring forces in the valve train and the piston ring tension. On the basis of these principles, Figure 5 shows the relationship between displacement and fuel efficiency. In high power regions, thermal efficiency increases in accordance with displacement. In contrast, in low power regions, thermal efficiency increases as displacement decreases. This demonstrates the fact that both indicated thermal efficiency and mechanical efficiency (i.e., friction losses) improve as displacement increases but that thermal efficiency improves as displacement decreases in low power regions because of the effect of pumping losses caused by the transition to partial load.

2.3 Relationship between planetary gear set, engine power, and axle output

Figure 6 shows the kinematic relationship between the three elements of a planetary gear set: the sun, carrier, and ring gears. It also illustrates the relationship between the torque split by the sun and ring gears with respect to the torque input through the carrier gear. The nomenclature is defined in Table 1.

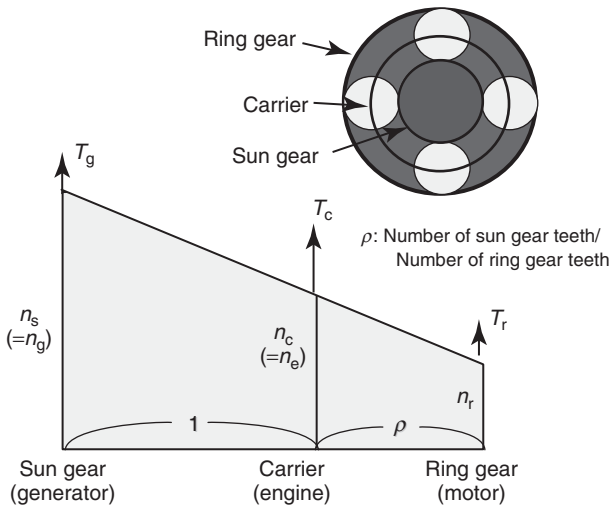


Figure 6. Colinear graph of planetary gear set.

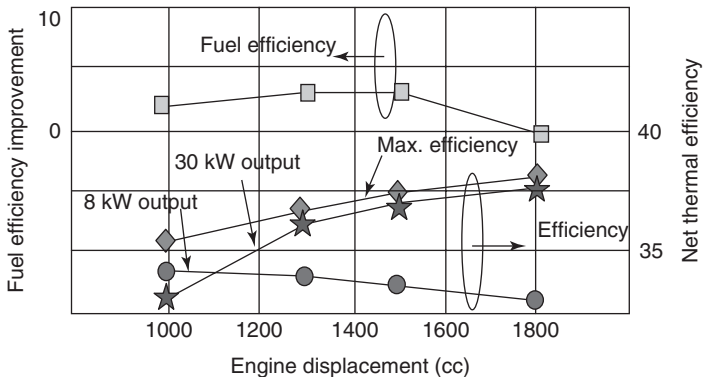


Figure 5. Engine displacement and fuel efficiency.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Nomenclature.

Powertrain Component	Engine	Generator		Motor	Axle
Planet element	Carrier	Sun	Ring		
Speed	n_e	n_g	n_r	n_m	n_a
Torque	T_e or T_c	T_g	T_r	T_m	T_a

The planetary gear set has two degrees of rotational freedom and three inputs. Equation 1 describes the relationship between the rotational velocities of each element.

$$\rho \times n_g + n_r = (1 + \rho) \times n_e \tag{1}$$

where ρ is the ratio of the number of sun gear teeth to the number of ring gear teeth.

Equations 2 and 3 describe the torque relationships between the three elements of the planetary gear set. These equations show that the input torque from the engine (coupled to the carrier) is split into two output torques: the generator (coupled to the sun gear) and the ring gear.

$$T_g = \frac{\rho}{1 + \rho} \times T_e \tag{2}$$

$$T_r = \frac{1}{1 + \rho} \times T_e \tag{3}$$

The output from the ring gear is directly transmitted to the axles via the reduction gear. The output from the sun gear is converted into electrical energy by the generator. This electrical energy is then supplied to the motor and converted into mechanical energy, which is mechanically coupled to the ring gear. The motor converts the electrically energy back into mechanical energy.

Equation 4 describes the total power at the axles as output from the generator is supplied to the motor. For simplicity, the conversion efficiency at the generator and motor and the efficiency of each gear are assumed to be 1.

$$n_a \times T_a = n_r \times T_r + n_m \times T_m \tag{4}$$

Equations 5 and 6 describe the right-hand side of Equation 4 in a different way using Equations 2 and 3.

$$n_r \times T_r = n_r \times \frac{1}{1 + \rho} \times T_e \tag{5}$$

$$\begin{aligned} n_m \times T_m &= n_g \times T_g \\ &= n_g \times \frac{\rho}{1 + \rho} \times T_e \end{aligned} \tag{6}$$

To summarize the above-mentioned explanation, Equation 4 becomes Equation 7 based on Equations 1, 5, and 6.

$$\begin{aligned} n_a \times T_a &= (1 + \rho) \times n_e \times \frac{1}{1 + \rho} \times T_e \\ &= n_e \times T_e \end{aligned} \tag{7}$$

As shown in Equation 7, power output from the engine is conserved and converted to power at the axles. In addition, Equation 1 can be rewritten as Equation 8 to describe the relationship between the generator rotational speed (n_g), the axle rotational speed (n_a), and the engine speed (n_e). Here, G_r is the gear reduction ratio, determined by dividing the rotational speed of the ring gear by the rotational speed of the axles.

$$G_r \times n_a = (1 + \rho) \times n_e - \rho \times n_g \tag{8}$$

In other words, Equation 8 shows that the ratio between the rotational speed of the axles and the engine speed can be constantly varied by changing the rotational speed of the generator. Therefore, in conjunction with the power-saving relationship shown in Equation 7, the planetary gear system can be considered as having CVT functionality. As a result of this function, an efficient engine operating point can be selected regardless of the vehicle speed, thus reducing fuel consumption and improving efficiency.

2.4 HEV control strategy

The HEV control strategy for the engine, motor/generator, inverter, battery, and other main systems plays an extremely important role in maximizing fuel efficiency and reducing emissions, while achieving the excellent dynamic performance of HEVs.

2.4.1 HEV driving control

The purpose of the HEV driving control is to optimize fuel efficiency while simultaneously controlling the driving force and performing power and energy management.

The power required at the axles is calculated based on the information from the driver (i.e., the accelerator angle and shift position) and the vehicle speed. In addition, the required engine power is determined in consideration of the battery state of charge (SOC).

The engine speed cannot be controlled independently by the engine. However, as described in Equation 8, this system has CVT functionality that allows the engine speed to be controlled by the generator to achieve optimum fuel efficiency.

The vehicle must be able to produce the commanded drive force regardless of the engine operating point.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Therefore, it is necessary to estimate the output torque at the ring gear, which is split from the engine torque. In addition, the ring gear is mechanically coupled to the axles. Therefore, the ring gear torque is directly related to the drive force. Assuming steady state conditions, Equation 2 can be substituted into Equation 3 to give the estimated ring gear torque as a function of the generator torque.

$$T_r = \frac{T_g}{\rho} \tag{9}$$

The commanded drive force can be achieved by applying the appropriate amount of motor torque to the ring gear. The sum of the ring gear torque and the motor torque is the commanded drive force. Figure 7 illustrates the block diagram of driving control and Figure 8 illustrates the schematic diagram of the THS-II control.

2.4.2 *Engine control*

The following engine parameters are controlled by the engine ECU to generate the engine power commanded by the drive control algorithm: throttle position, ignition timing, and variable valve timing (VVT). The engine speed is controlled by the generator.

The opening and closing timing of the intake valves is optimized for efficiency and to suppress vibration when the engine starts and stops.

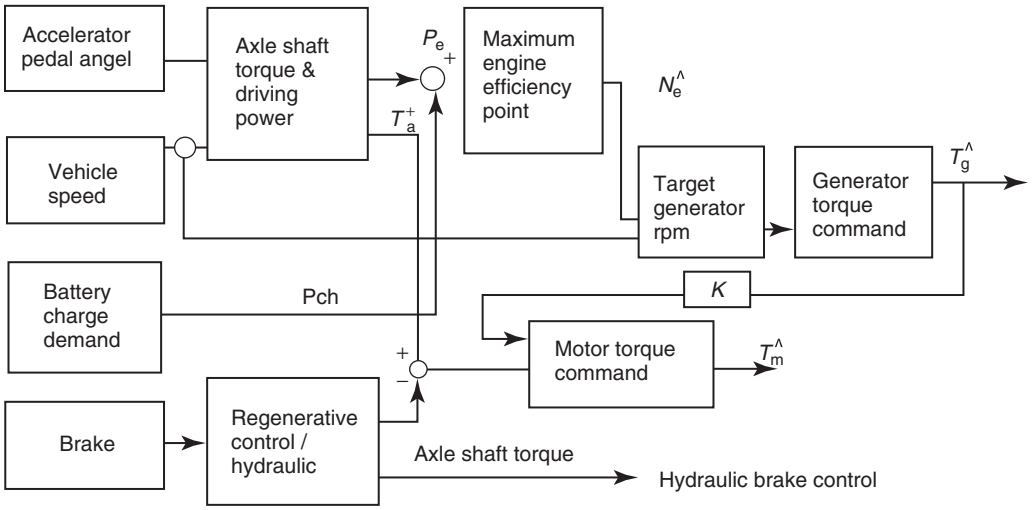


Figure 7. Block diagram of driving control.

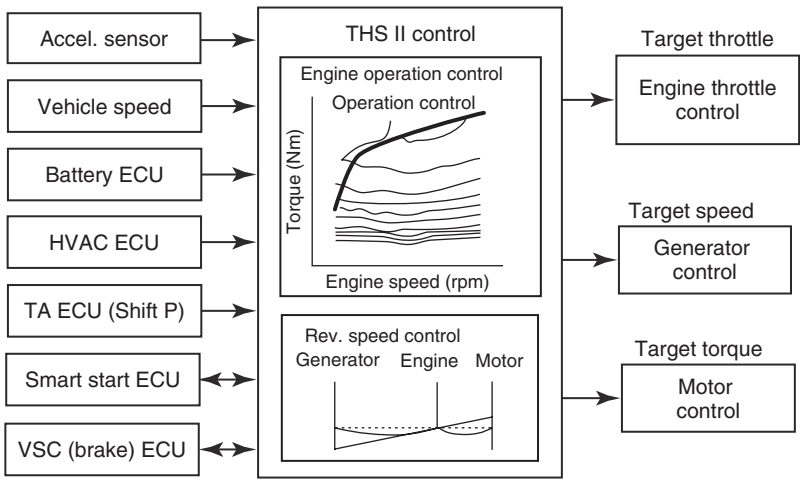


Figure 8. Schematic diagram of THS II control.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

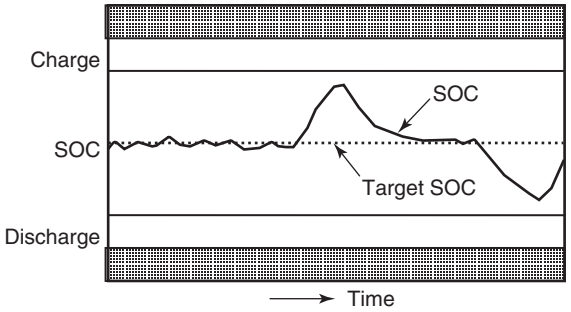


Figure 9. SOC control.

2.5 Battery control

Battery control involves calculating the available input and output power, and SOC at the current battery state using current, voltage, and temperature sensors. It also entails controlling the airflow of the battery cooling fan. The graph in Figure 9 illustrates that the target SOC is an average value during driving based on charge and discharge requests (Pch in the block diagram of driving control in Figure 7) from the battery.

2.6 Regenerative-hydraulic brake coordination system

The electronically controlled braking (ECB) system performs coordinated control between the hydraulic service brakes and regenerative brakes. The regenerative brakes improve fuel efficiency and HEV energy management using the electric motor as a generator to convert the vehicle's kinetic energy to electrical energy and storing it in the battery during braking. In a conventional vehicle, the kinetic energy is lost as heat during braking. Figure 7 illustrates the regenerative braking torque as "axle shaft torque." At the same time, increasing the maximum regenerated power by adopting a high power density battery improves regeneration efficiency. Figure 10 shows the distribution of hydraulic and regenerative braking in the THS (from 1997 to 2003) and THS II (from 2003) systems in response to braking demands. This improvement was completed by the refinement of the ECB system and the battery charge control. The use of regenerated energy has also been increased by improved aerodynamics and reducing various rotational losses, such as through brake friction and bearings. This increases fuel efficiency not only in low speed urban driving where there are many stops but also in higher speed suburban driving pattern and on the expressway.

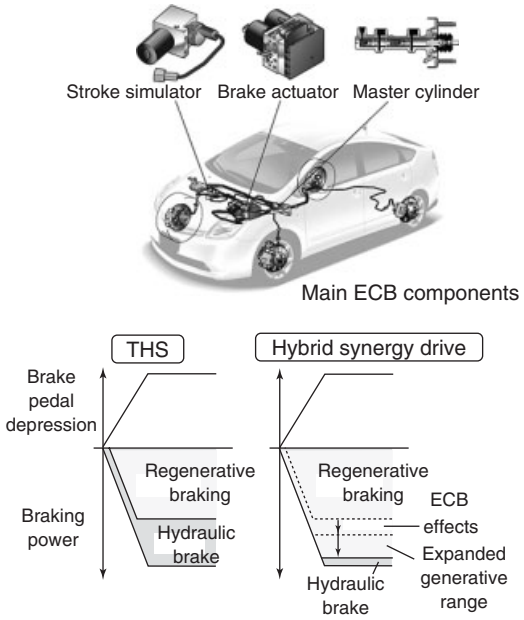


Figure 10. Improved regenerative braking.

3 ENERGY MANAGEMENT IN PARALLEL TYPE MILD HYBRID SYSTEM (IMA)

3.1 Purpose and aims of system

The integrated motor assist (IMA) system developed by Honda Motor Co., Ltd. is a parallel type hybrid system in which a motor provides assistance to the engine as the main source of driving power. A general hybrid system improves the efficiency of energy usage by recovering the vehicle's kinetic energy on deceleration and increasing the frequency that the engine operates in the most efficient operating regions. However, these gains are offset to some degree by a reduction in fuel efficiency due to the mass and volume of the system. The IMA system was developed with an emphasis on maintaining a balance between efficient energy usage and the mass and volume of the system. Its simple structure reduces mass and volume by directly connecting the motor to the engine crankshaft.

3.2 Engine

The engine in the IMA system can be characterized by two main points for improving fuel efficiency: reduced displacement and the adoption of a cylinder deactivation system. The following section describes these points in more detail.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

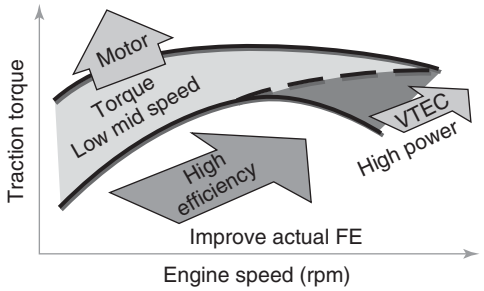


Figure 11. Dynamic performance of IMA system. (Reproduced with permission from Society of Automotive Engineers of Japan. © Society of Automotive Engineers of Japan.)

Improving fuel efficiency is the critical element for reducing engine displacement. However, the engine in a passenger vehicle is required to operate over a wide range of operating regions. In a vehicle installed with a conventional engine, this requirement is usually met by increasing displacement to ensure the necessary basic dynamic performance. Consequently, conventional vehicles generally have a substantial margin for driving power in low load regions in which actual operation is frequent, thereby deliberately reinforcing engine operation in relatively inefficient states. In contrast, reducing engine displacement to increase the engine operation frequency in high load regions, where engine efficiency is high, has an adverse effect on the dynamic performance of the vehicle. The IMA system increases engine efficiency in these frequently used low load regions by reducing engine displacement. At the same time, it ensures the required dynamic performance using the motor to assist the engine in low engine torque and speed regions. Figure 11 illustrates the dynamic performance of the IMA system.

3.3 Outline of IMA system operation

Figure 12 illustrates the basic energy management of the IMA system in the following five stages. (i) Normal engine start and engine start by the idling stop system are performed by the IMA motor. (ii) Acceleration after engine start is mainly performed by the engine, with motor assistance when required. (iii) When the vehicle is cruising in low load regions, the vehicle operates as an EV on motor power alone to avoid operating the engine in the inefficient regions. This stage represents highly efficient operation by the IMA system. When the vehicle is operating as an EV, the intake and the exhaust valves of the engine are closed to minimize pumping losses, and power from the motor is used to drive the vehicle efficiently. (iv) The vehicle is powered by the engine when cruising in high load regions. However, power is also generated by the motor in these regions, depending on the load region and the battery SOC (this operation is not described in Figure 12). (v) During deceleration, the motor acts as a generator to recover kinetic energy for charging the battery. In the same way as when the vehicle is operating as an EV, the intake and the exhaust valves of the engine are closed to minimize pumping losses and increase the amount of recovered energy. Finally, when the vehicle is stopped, the idling stop system ensures that the energy is not consumed needlessly.

In the IMA system, the engine and motor are directly coupled. However, a parallel type strong hybrid system is also being produced that allows the vehicle to be driven as an EV by providing a clutch to decouple the engine and motor. This ensures that the engine does not operate needlessly.

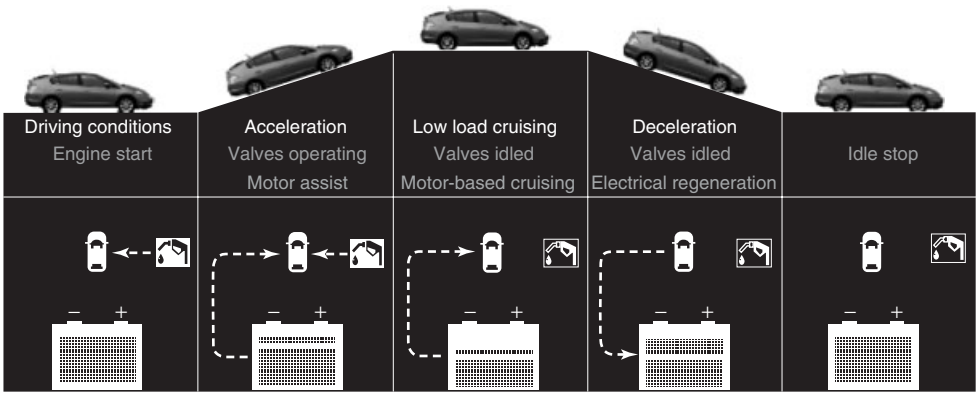


Figure 12. Driving condition-based energy management of IMA. (Reproduced with permission from Society of Automotive Engineers of Japan. © Society of Automotive Engineers of Japan.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4 THERMAL ENERGY MANAGEMENT IN HEVS

Figure 13 shows the changes in fuel economy by season for the first-generation (1997) Japanese market Prius, second-generation (2000) Japanese market Prius, and an average conventional vehicle with an automatic transmission (AT). This data is normalized using the best month for fuel efficiency.

Figure 13 can be used to assess differences in fuel consumption between seasons. Both the conventional vehicle and the Prius models show a decline in fuel efficiency during the summer and winter months. The Prius shows a larger decline in fuel efficiency over the winter months than the conventional vehicle. The reasons for this are as follows: cold starts in cold weather require more energy to warm up the engine and catalyst, the heater uses more energy to warm up the vehicle's interior, and the duty cycle of the engine is higher to ensure that the required energy is available. A probable factor in the decline in fuel efficiency during the summer is higher energy consumption due to the use of air conditioning. Therefore, the efficiency of the hybrid system can be improved by focusing on the areas where such declines in fuel efficiency occur.

Figure 14 shows the energy distribution when the Prius is driven under the US urban test cycle in various heating conditions. With an outside temperature of 5°C, it can be seen that when the heater is turned on, more energy is used to heat the car than to drive the car.

4.1 Adoption of electrical air conditioning

As shown in Figure 15, the current Prius does not use a conventional engine-driven air conditioner compressor. Instead, it uses a high efficiency inverter-driven compressor

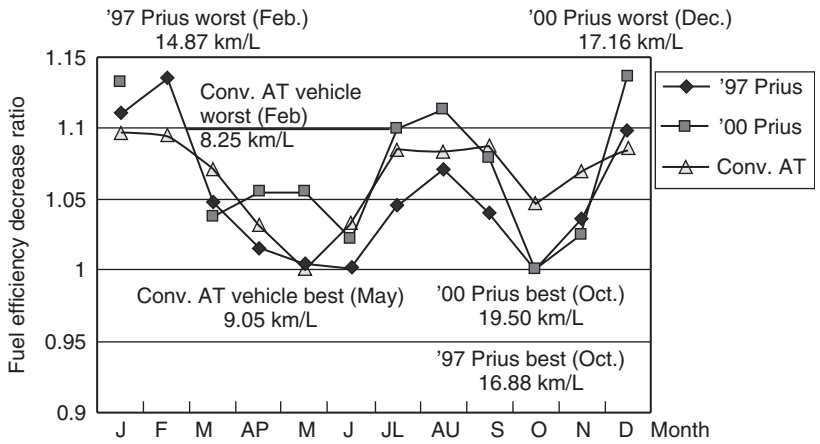


Figure 13. Fuel efficiency by season.

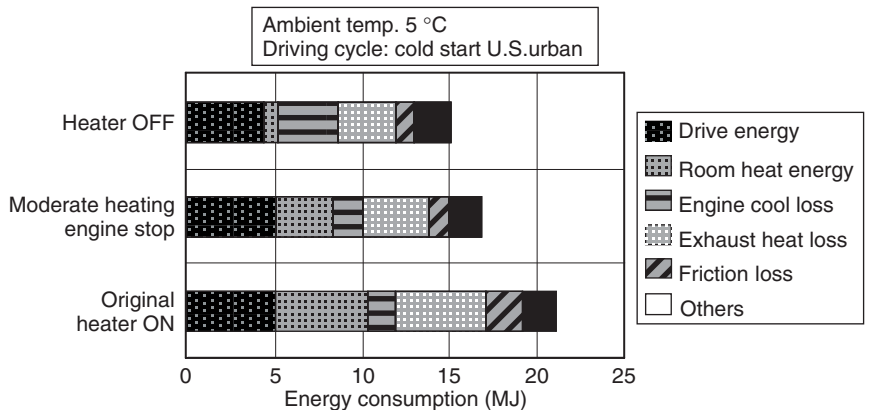


Figure 14. Prius energy distribution.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Electric air conditioning compressor

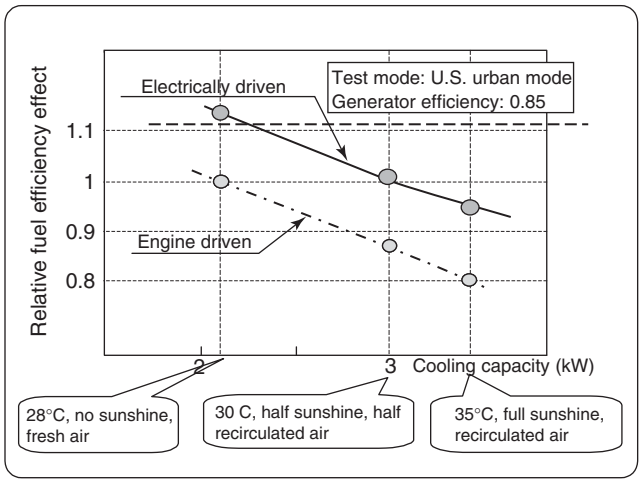


Figure 15. Fuel efficiency improvement by electric air conditioner compressor.

that takes advantage of the high voltage from the hybrid system battery. This contributes to improved air conditioning performance when the engine is stopped. In addition, it also improves fuel efficiency when the air conditioner is operating because the electrical system reduces the number of times the engine is required to start to drive the air conditioner compressor.

4.2 Engine thermal energy management

4.2.1 Electric water pump (EWP)

To further improve fuel efficiency, an enhanced engine cooling system is used. It consists of an electric water pump and exhaust heat recirculation device, which were specifically developed to improve fuel efficiency.

The flow rate of a conventional water pump is dependent on the engine speed. This uses excessive power even in conditions that require lower flow rates, such as in low temperatures or in low load, high engine speed conditions. However, using an electric water pump, the flow rate is no longer dependent on the engine speed. Now the flow rate can be varied and optimized according to the actual cooling demand.

Figure 16 shows a schematic drawing of a cooling system and compares the structure of a conventional water pump to the structure of an electric water pump. To achieve a size comparable to a conventional water pump, the electric water pump adopted a flattened compact motor design with an integrated motor driver. Unlike the conventional water pump, the impeller rotates about a center shaft, which is fixed to the water pump housing. This impeller is not powered by a traditional impeller shaft that passes through the housing to a pulley. Instead, the magnetic impeller is rotated by a rotating magnetic field situated in the water pump housing, which means that the water chamber is isolated from the outside. This eliminates the need for a mechanical seal and results in lower friction loss and no risk of water leakage.

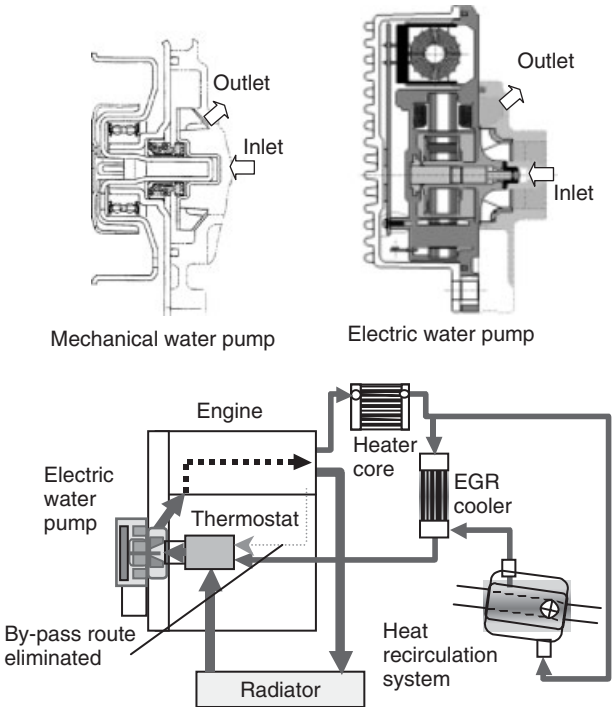


Figure 16. Cooling system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

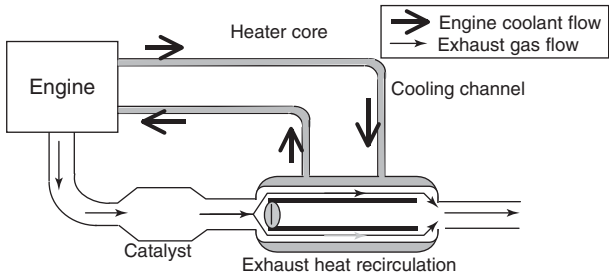


Figure 17. Exhaust heat recirculation system.

4.2.2 Improving engine warm up

The electric water pump is able to operate at a lower speed (lower flow rate) after a cold start-up, shortening the engine warm-up period.

Figure 17 shows a schematic drawing of the exhaust heat recirculation system. This was adopted to improve fuel efficiency in cold weather conditions. The warm-up performance of the engine and engine coolant is improved using the exhaust heat, which is normally wasted in conventional systems.

As a result, friction is reduced during the engine warm-up period. In addition, owing to the characteristics of the hybrid system, fuel efficiency is further improved by allowing the engine to stop earlier. Figure 18 shows the improvement in vehicle fuel efficiency during the winter. The fuel efficiency of the third-generation Prius, equipped with the EWP and the exhaust heat recirculation system, is 19% higher than the previous generation under the same test conditions.

5 PHEV SYSTEM CONFIGURATION

The various types of plug-in hybrid electric vehicles (PHEVs) can be categorized into two main types based on the way the vehicle uses its energy sources. The first is the all electric range (AER) type, which operates on battery power alone as long as it has remaining energy. When the battery is depleted of energy, the engine charges the battery. The second is the blended type, which "blends" power from both a battery and an ICE. The engine is not only used for charging the battery but also used to supplement the battery power when the requested drive force exceeds the available drive force from the battery power alone. Figure 19 shows an example for each configuration. Table 2 describes the characteristics of each type. Toyota's PHEV has a blended type system because of its lower cost and smaller size compared to an AER type with equal vehicle performance. Figure 20 shows the results from a Toyota PHEV demonstration project. The graph shows the percentage of users in the project versus the output power of the PHEV system. While the maximum output was rarely used, 30 kW or less was needed by the drivers 93% of the time. This demonstrates that a blended PHEV with a battery power of 30 kW can be driven on electricity alone 93% of the time, as long as battery energy remains. Furthermore, only small changes are required to the existing Toyota THS to upgrade to a blended type PHEV. This means that the

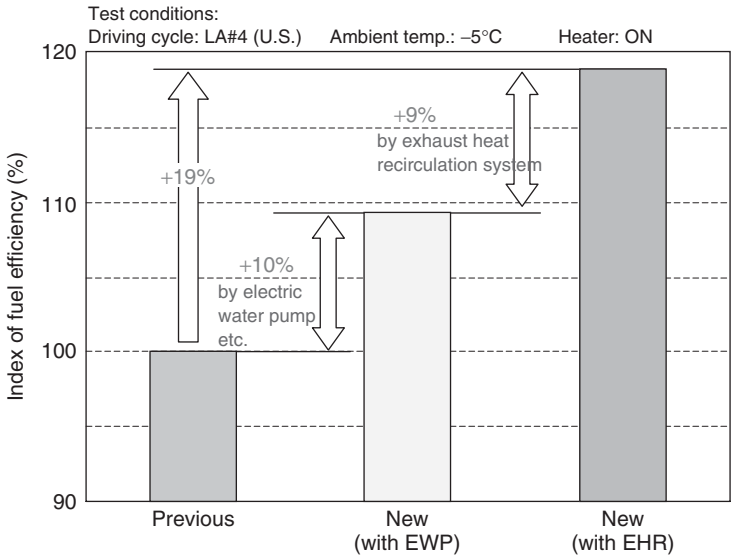
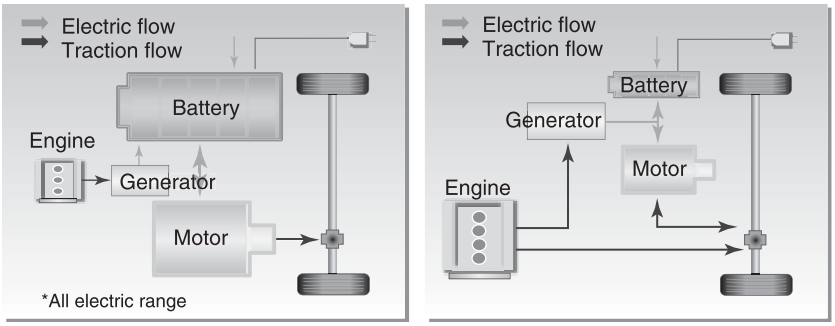


Figure 18. Improvement of fuel efficiency in winter.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Mode	AER type	Blended type
Charge depleting	Battery only	Mainly battery + engine if needed
Charge sustaining	Mainly engine	Mainly engine + battery if needed

Figure 19. PHEV system comparison.

Table 2. Vehicle comparison.

	EV	PHEV (AER, EV Base)	PHEV (Blended, HEV Base)	HEV
CO_2 reduction	++	+/++	+/++	+
Driving distance	−	+	++	++
Charge time	−	±	+	++
Special infrastructure	− (required)	− (required)	+ (on demand)	++
Cost	−−	−	+	++

++, excellent; +, good; ±, fair; and −, poor (compared to conventional vehicle).

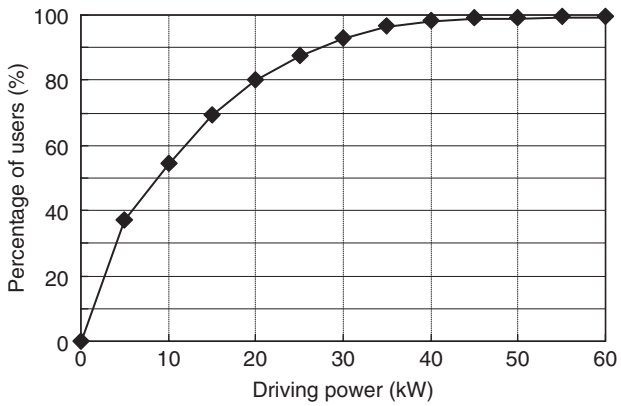


Figure 20. EV power usage.

benefits of driving a vehicle as an EV can be achieved at low cost with the THS.

6 FEATURES OF PHEVS

On average, the daily mileage of personal passenger vehicles is relatively short. PHEVs take advantage of this fact in their application of electrical energy. PHEVs have the advantages of both HEVs and EVs. The charging system of a PHEV differs from that of an HEV in the following manner: it has a larger energy capacity battery and it can use external electricity (e.g., household outlets) to charge the battery. PHEVs can operate as an EV over short distances. After the electrical energy in the battery is consumed, it operates as an HEV, propelled by both the engine and the motor, while consuming less fuel than conventional vehicles. In typical passenger vehicle usage, a PHEV can be driven almost fully as an EV during weekdays because trips are usually relatively short. In addition, it can be driven as an HEV after the energy in the battery is consumed for long distance trips that usually occur on weekends. This PHEV concept combines the advantage of an EV of using clean electrical energy with the advantages of an HEV, such as longer range, high fuel efficiency, and sufficient luggage space. In this way, a PHEV can be a substitute for a conventional vehicle.

Figure 21 shows a comparison of the battery SOC control strategies for HEVs and PHEVs. The left figure shows the SOC control strategy in an HEV. The SOC fluctuates in a small range but is controlled to maintain a set point

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

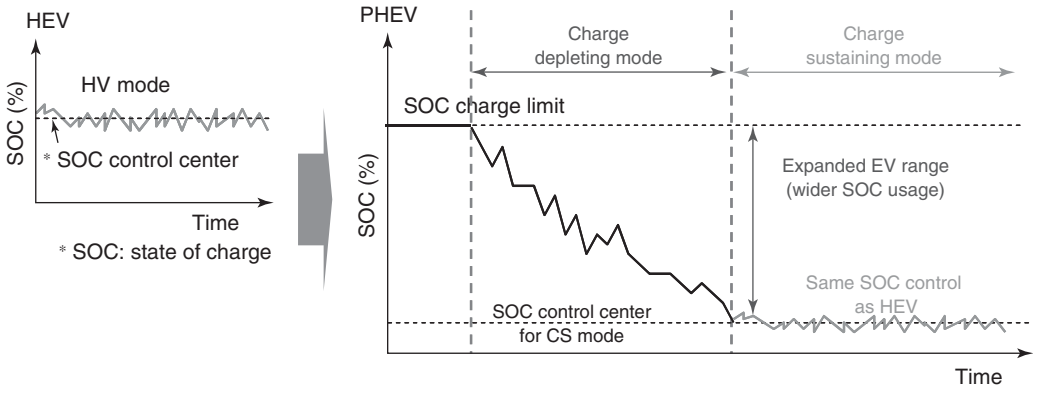


Figure 21. Comparison of HEV and PHEV operation.

or target SOC, because an HEV has no external charging system.

PHEVs have two driving modes: the first is charge depletion (CD) mode, which mainly uses the stored electrical energy that was externally supplied to the battery. After consuming the stored external electrical energy in the battery, the PHEV goes into charge-sustaining (CS) mode and operates like an HEV. In CD mode, the PHEV prioritizes EV operation, where energy is supplied from the battery. However, in the same manner as an HEV, it is still possible to start the engine when rapid acceleration is required. In CS mode, the PHEV operates like an HEV and it has the same characteristics of a long driving range and lower fuel consumption than a conventional vehicle. Accordingly, it is effective to design a PHEV based on an HEV, where the advantages of an EV and an HEV can be combined by simply expanding the capacity of the battery and adding battery charging control logic for SOC management.

RELATED ARTICLES

Range extender EV
Rechargeable Battery Basics
Batteries indication and management
EVT and E-CVT for Full Hybrid Electric Vehicles

REFERENCES

Abe, S. (2008) "10 years after Prius: Current and Future of Hybrid Vehicles" P.78 KURUMA NAVI III. News Digest.

Abe, S., *et al*. (2011) Descriptions of THS and IMA in *EV/HV Handbook of Automotive Technology*, JSAE, Japan.

Kawamoto, N., Naiki, K.,Kawai, T., *et al*. (2009) Development of new 1.8-Liter engine for hybrid vehicles. SAE Technical Paper 2009-01-1061, Society of Automotive Engineers: USA.

Society of Automotive Engineers Japan (2011) *The Handbook of Automotive Engineering No.10: Design (EV & Hybrid Vehicles)*, The Society of Automotive Engineers of Japan, Japan.

Yaegashi, T., Abe, S., and Hermance, D. (2004) Future Automotive Powertrain – Does Hybridization Enable ICE Vehicles to Strive Towards Sustainable Development?' Paper number 2004-21-0082, presented at *Convergence International Congress & Exposition On Transportation Electronics*, 18 October 2004, Detroit, Michigan, United States.

Yamamoto, M., Takaoka, T., Masayuki Komatsu, M., and Gotoda, Y. (2010) Development of a Toyota plug-in hybrid vehicle. SAE Technical Paper 2010-01-0839, Society of Automotive Engineers: USA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto055
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Overview of Electric, Hybrid, and Fuel Cell Vehicles

C.C. Chan
The University of Hong Kong, Pokfulam, Hong Kong

1 Introduction 1
2 Overview of BEV, HEV, and FCV 1
3 Architecture of Powertrains 3
4 Energy Management of BEVs, HEVs, and FCVs 6
5 Commercialization Road Map 8
6 Conclusion 11
Acknowledgment 14
References 14

1 INTRODUCTION

This chapter provides an overview of the state of the art of battery powered electric vehicles (BEVs), hybrid electric vehicles (HEVs), and fuel cell electric vehicles (FCVs) (Chan, 2007; Chan, Bouscayrol, and Chen, 2010; Chan and Chau, 2001; Yamada, Hanada, and Sasaki, 2006; Burke, 2002; Lai and Nelson, 2007; Eshani *et al*., 2005a; Gao, Mi, and Emadi, 2007) with focus on their technical features and commercial road map. This chapter begins with the distinct features of BEV, HEV, and FCV. As electric vehicle technology is the integration of automotive technology and electrical technology, system integration and optimization approach is essential, the overall design of electric vehicle system including its infrastructure and business model should be guided by proper engineering philosophy. Powertrain is the heart of electric vehicle; therefore, this chapter analyzes the architecture and functionality of various types of powertrains. Energy management is the key to achieve fuel economy and emission reduction; hence, this chapter explores the challenge of energy management, discusses the function and interaction of global control and local control, and suggests the possible types of appropriate control models.

The success of promotion of electric vehicles relies on the availability of good product, good infrastructure, and good business model as well as the integration of electric vehicles with smart grid and with information communication technology (ICT). Our goal is gradually to achieve four zeros: zero emission, zero gasoline, zero traffic accident, and zero traffic jams.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 OVERVIEW OF BEV, HEV, AND FCV

2.1 Features of BEV, HEV, and FCV

Since the oil crisis in 1973, the world started paying attention to the balance of oil supply and consumption. The amount of fossil fuel resources is limited but the oil demand has increased significantly. The transportation sector is one of the largest energy sectors. In addition, it has the highest growth rate of oil consumption in recent decades. This growth has largely come from new demands for personal use vehicles powered by conventional internal combustion engine (ICE).

Some environmental problems, such as the greenhouse effect, the acid deposition, and the air pollution, are directly related to the vehicle emission. There have been increased tensions in part of the world because of the energy crisis.

Government agencies and organizations therefore have developed more stringent standards for the fuel consumption and emissions. BEVs look like ideal solutions to tackle

the energy crisis and global warming, as they have zero oil consumption and zero emission *in situ*. However, the high initial cost, short driving range, and long charging time have caused BEV only suitable for certain applications unless providing good infrastructure such as battery leasing or swapping, combination of slow charge, medium charge and quick charge, and integration with smart grid.

HEVs were developed to overcome the disadvantages of ICE vehicles and BEVs. An HEV combines a conventional propulsion system with an electric energy storage system and an electric machine (EM). It has a longer range than BEV. It shows improved fuel economy as compared with those of conventional ICE vehicles. The ICE can be stopped if the vehicle is at a stop. The electric drive system can optimize the efficiency of the ICE and thus reduce the oil consumption and emission. The kinetic energy can be recovered during braking and down slope driving. A certain range of silent operation with zero emission is possible when HEVs are driven in electric mode. In addition, this operating range may be extended if the battery can be recharged by connecting a plug to an electric power source, such as the electricity grid. This kind of HEV is called *plug-in hybrid electric vehicle* (*PHEV*). Furthermore, the onboard EMs of HEVs provide more flexibility and controllability to the vehicle control, such as antilock braking system (ABS) and vehicle stability control (VSC), and thus offer improved performance. Although HEVs can contribute to meeting the challenges in road transport regarding energy crisis and pollution, it is still somewhat difficult to be widely accepted by general public. Three main hurdles exist for vehicle buyers: the first one is their high purchase price; the second one is reliability and warranty related to the lack of electrician in the car shops; the third one is a lack of confidence in electric powered vehicle, concerning the introduction of high voltage, electromagnetic interference caused by high frequency high current switching and so on.

Fuel cell electric vehicles (FCVs) use fuel cell to generate electricity from hydrogen and air. The electricity is either used to drive the vehicle or stored in an energy storage device, such as battery pack or supercapacitors. They only emit water vapor, and they have the potential to have high efficiency. The major issues related to FCV are as follows: firstly, the high price and the life cycle problem of FCs; secondly, hydrogen onboard storage needs improvement of energy density; and thirdly, hydrogen distribution and refueling infrastructure needs to be constructed (International Energy Agency Implementing Agreement on Hybrid and Electric Vehicles, 2008).

Table 1 shows the comparison of the characteristics of EV, HEV, and FCV.

2.2 Engineering philosophy of EV, HEV, and FCV

The overall EV engineering philosophy essentially is the integration of automobile engineering and electrical engineering. Thus, system integration and optimization are major considerations to achieve good EV performances at affordable cost. As the characteristics of electric propulsion are fundamentally different from those of engine

Table 1. Characteristics of BEV, HEV, and FCV.

Types of vehicles	BEV	HEV	FCV
Propulsion	• Electric motor drives	• Electric motor drives • Internal combustion engines	• Electric motor drives
Energy storage subsystem (ESS)	• Battery • Ultracapacitor	• Battery • Ultracapacitor • ICE generating unit	• Fuel cells • Need battery/ultracapacitor to enhance power density.
Energy source and infrastructure	• Electric grid charging facilities	• Gasoline stations • Electric grid charging facilities (for plug-in hybrid)	• Hydrogen • Hydrogen production and transportation infrastructure
Characteristics	• Zero emission • High energy efficiency • Independence on crude oils • Relatively short range • High initial cost • Commercially available	• Very low emission • High fuel economy • Long driving range • Dependence on crude oil • Higher cost than ICE vehicles • Commercially available	• Zero or ultralow emission • High energy efficiency • Independence of crude oil (if not using gasoline to produce H_2) • High cost • Under development
Major issues	• Battery and battery management • Charging facilities • Cost	• Multiple energy sources control, optimization, and management. • Battery sizing and management	• Fuel cell cost, cycle life, and reliability • Hydrogen infrastructure

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

propulsion, a novel design approach is essential for EV engineering. Moreover, advanced energy sources and intelligent energy management are key factors to enable EVs competing with ICEVs (internal combustion engine vehicles). Of course, the overall cost effectiveness is the fundamental factor for the marketability of EVs.

The design approach of modern EVs should include state-of-the-art technologies from automobile engineering, electrical and electronic engineering, and chemical engineering, should adopt unique designs that particularly suitable for EVs, and should develop special manufacturing technology that particularly suitable for EVs. Every effort should be made to optimize the energy utilization of EVs.

The EV engineering philosophy is the marriage of automotive engineering and electrical engineering that includes motor, power electronic converter, controller, battery, or other energy storage device and energy management system. Marriage implies that the bride and the groom have fully understood the character of the partner and able to cope together harmoniously and best perform to achieve the required driveability at maximum energy efficiency and minimum emission.

The HEV engineering philosophy is $1 + 1 > 2$. This implies the added value gained from the integration of engine propulsion and motor propulsion, fully sizes the advantage and flexibility of electrical, electronic, and control technologies, not only to increase energy efficiency and reduce emission, but also become more intelligent, driving comfort, and safety. Just like mule is the hybrid of horse and moke, mule possesses the best DNA of horse and moke and hence more powerful and endurance. In HEV, the prime key technology is the control algorithm and optimization for energy management.

The FCV engineering philosophy is the integration of automotive engineering, electrical engineering, and fuel cell engineering. As fuel cell is a new kind of energy device that quite different with gasoline and batteries, every effort should ensure that the overall system of fuel cell is efficient, reliable, optimum, and last long at reasonable cost. Other high power density device such as lithium-ion battery or ultracapacitor may be used in conjunction with fuel cell to improve the starting performance of the vehicle. The electric propulsion system and fuel cell system must cope very well to achieve the required driveability at maximum energy efficiency and minimum emission.

In summary, the core of engineering philosophy is system integration and optimization. The principles of integrated system design can be summarized into the following six principles (Elliot and Deasley, 2007):

1. Debate, define, revise, and pursue the purpose/objective

   The system exists to deliver capability, and the end justifies the means. The statement of a requirement must define how it is to be tested. Requirements reflect the constraints of technology and budgets.

2. Think holistic

   The whole is more than the sum of the parts—and each part is more than a fraction of the whole.

3. Be creative

   See the wood before the trees.

4. Follow a disciplined procedure

   Divide and conquer, combine and rule.

5. Take account of the people

   To err is human; Ergonomics; Ethics and Trust.

6. Manage the project and the relationships

   All for one, one for all.

3 ARCHITECTURE OF POWERTRAINS

ICEVs are propelled by fuels and ICEs. BEVs are propelled by batteries and EMs. HEVs are propelled by the combination of the two powertrains. In which, the ICE provides the hybrid vehicle an extended driving range, whereas the EM increases efficiency and fuel economy by regenerating energy during braking and storing excess energy from the ICE during coasting. According to the way of how the two powertrains are integrated, generally, there are three basic architectures of HEVs, namely series hybrid, parallel hybrid, and series–parallel hybrid (Ehsani, Gao, and Miller, 2007; Gao and Ehsani, 2006; Miller, 2006).

Among these diverse architectures, a series–parallel hybrid vehicle with a planetary gear (Figure 1) has a "maximal" architecture, which can maximize the optimization. Therefore, this architecture is chosen as our base for discussion and comparison, and other architectures are derived from this basic architecture scheme. In these architectures, batteries are expressed by BAT, fuel tank by fuel, voltage source inverter by VSI, electric machine by EM, internal combustion engine by ICE, and Transmission by Trans respectively. Black lines mean electric coupling and orange lines mean mechanical coupling.

It is noted that the transmission could be a discrete gearbox with clutch, continuously variable transmission

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

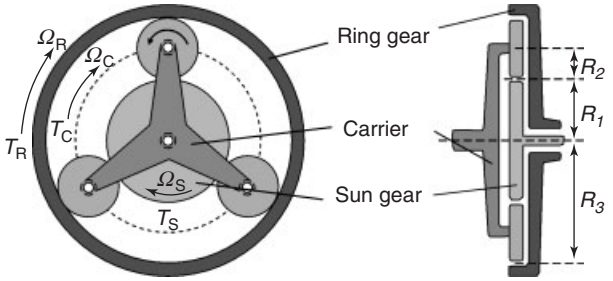


Figure 1. Planetary gear unit.

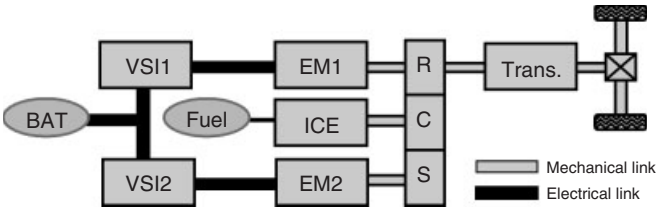


Figure 2. Series–parallel hybrid vehicle using a planetary gear unit.

(CVT), or fixed reduction gear. Series–parallel HEV in a series–parallel hybrid vehicle, a planetary gear (Figure 1) is used (Eshani, *et al*., 2005b; Syed *et al*., 2006). The electric machine 1 (EM1) and the transmission shaft (Trans.) are connected to the planetary set ring gear (R), whereas the ICE is connected to the carrier (C) and the EM2 is connected to the sun gear (S), respectively.

Thanks to the DC bus and the planetary gear, a series–parallel hybrid can operate as a series hybrid or a parallel hybrid, respectively. Because of the planetary gear, the ICE speed is the sum of the EM1 speed and the EM2 speed. The EM1 speed is proportional to the vehicle speed. At a given vehicle speed (or given the EM1 speed), the EM2 speed can be adjusted in order to adjust the ICE speed. The ICE can thus operate in an optimal region by controlling the EM2. Despite possessing these features of both the series and parallel HEVs, two machines and a planetary unit are necessary, which makes the drivetrain somewhat complicated and costly. Moreover, the control of this architecture is quite complex. This architecture (Figure 2) is depicted in such a way in order to deduce the other classical architectures.

3.1 Series HEV

From this series–parallel hybrid architecture, if the connection between the EM1 and the ICE is removed, a series hybrid vehicle is obtained. In this series hybrid vehicle (Figure 3), the energetic node among the power sources

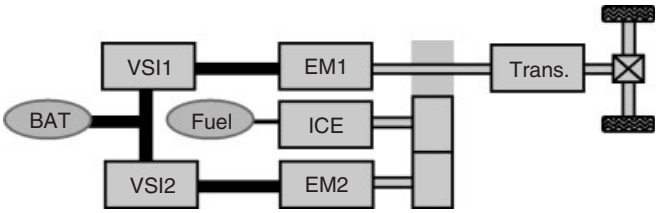


Figure 3. Series hybrid vehicle.

and transmission is occurred at the DC bus. Please note that Figures 2 and 3 are only schematic diagram, where the split path epicyclic cannot give true series operation.

In series HEV, the ICE mechanical output is first converted into electricity using EM2. The converted electricity either charges the battery or can bypass the battery to propel the wheels via EM1 and the transmission. Owing to the decoupling between the ICE and the driving wheels, it has the definite advantage of flexibility for locating the ICE generator set. Because of the same reason, the ICE can operate at its very narrow optimal region independently from the vehicle speed. Its control is simple because of a single torque source (EM1) for the transmission. Owing to the inherent high performance of torque–speed characteristic of the EM drive, multigear transmission and clutch would be unnecessary. However, such cascade structure leads to a relatively low efficiency and three machines (ICE, motor, and generator) are required. Another disadvantage is that all these propulsion devices need to be sized for the maximum sustained power, making series HEV expensive. On the other hand, when it is only needed to serve such as short trips commuting to work and shopping, a lower rating of the corresponding ICE generator set can be adopted.

More recently, EVT (electronic variable transmission) has been developed to replace the planetary gear. EVT is an electromechanical converter with two mechanical ports and one electrical port (consisting of two electrical machines and two inverters) (Hoeijmakers and Ferreira, 2006). An EVT (Cheng *et al*., 2007; Chen *et al*., 2008; Cheng *et al*., 2008) could be considered as a combination of two induction machines EM1 and EM2: the stator of EM2 rotates, and the rotor of EM2 is connected with the EM1 rotor. In order to reduce the system weight and size, the two machines can be integrated into a single machine, as a double-rotor induction machine (Hoeijmakers and Ferreira, 2006; Chau and Chan, 2007) or double-rotor permanent magnet machine (Chau, Chan, and Liu, 2007).

3.2 Parallel HEV

From the series–parallel hybrid architecture, if the EM2-based powertrain is removed, then a parallel hybrid vehicle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

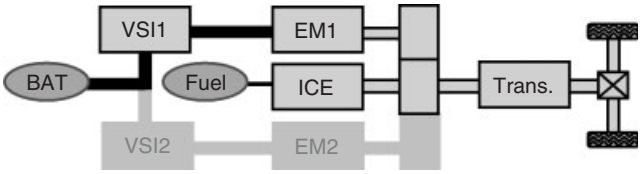


Figure 4. Parallel hybrid vehicle.

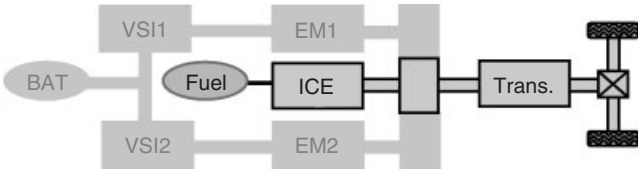


Figure 5. ICE vehicle.

is obtained (Figure 4). In a parallel hybrid drivetrain, the energetic node is occurred at a mechanical coupling. This mechanical coupling may be considered as one common shaft or connection of two shafts by gears, pulley-belt unit, and so on.

The propulsion power may be supplied by ICE alone, by EM1, or by both. EM1 can be used as a generator to charge the battery by regenerative braking or absorbing power from ICE when its output is greater than that required power to drive the wheels. Better than the series HEV, the parallel hybrid needs only two propulsion devices, the ICE and the EM. Moreover, smaller devices can be used to get the same dynamic performance. However, owing to the mechanical coupling between the ICE and the transmission, the ICE cannot always operate in its optimal region and hence clutches are necessary. Another drawback is the relatively complex control.

3.3 ICE vehicle

An ICE vehicle (Figure 5) is obtained when only the ICE-based drivetrain is remained from the series–parallel hybrid architecture. ICE vehicles have a long driving range and a short refueling time but are facing the challenge of the pressure of pollution problems and oil demand.

3.4 Battery powered electric vehicle

A BEV (Figure 6) is obtained when only the EM1-based drivetrain is kept. Because only batteries or electric energy-storage-based power sources drive the vehicle, zero emission can be achieved. However, its high initial cost, short driving range, and long charging (refueling) time have caused its limitation.

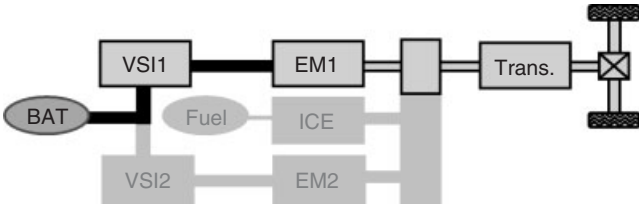


Figure 6. Battery powered electric vehicle.

3.5 Fuel cell vehicle

From the structural point of view, a fuel cell vehicle can be considered as BEV type. A fuel cell vehicle can also be equipped with a battery or a supercapacitor (Lai and Nelson, 2007). This fuel cell vehicle can be thus considered as series hybrid vehicle type with the fuel cell as an electrical generator from hydrogen (Hissel, Candusso, and Harel, 2007). The onboard fuel cell produces electricity, which is either used to provide power to the propulsion motor EM1 or stored in the battery or supercapacitor for future use.

3.6 Architecture versus functionality

Different functions can be achieved using the previous HEV architectures. They can be classified by their power ratio between the ICE and electrical machines. The more of using electrical power propulsion, the more of fuel economy improvement is gained.

3.6.1 Micro-hybrid (stop-and-go function)

In the case of micro-hybrid vehicle, a small power electrical machine is used as starter alternator (Sepe *et al.*, 2001). The ICE ensures the propulsion of the vehicle. The electrical machine helps the ICE to get a better operation point at the start-up. Because of the fast dynamics of EM, micro-hybrid HEVs ensure a "stop-and-go" mode: the ICE can be stopped when the vehicle is at standstill (e.g., at the traffic lights). A fuel economy improvement is estimated from 2% to 10% for urban drive cycles. Note that strictly speaking, micro-hybrid is not really a hybrid as the EM does not contribute to propulsion.

3.6.2 Mild hybrid (boost function)

In this case, the EM is used to boost the ICE during acceleration or braking by a supplementary torque. Of course, the stop-and-go function is also ensured. However, the electrical machine alone cannot propel the vehicle. A regenerative braking function is also achieved. A fuel economy improvement is estimated from 10% to 20%.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.6.3 *Full hybrid (electric traction)*

In this case, the EM is able to ensure the whole propulsion of the vehicle for a zero emission vehicle (ZEV) mode. This mode can be used in urban centre. The propulsion of the vehicle can also be made by the ICE or by the combination of ICE and the electrical machine. A fuel economy improvement is estimated from 20% to 50%.

3.6.4 *Plug-in hybrid (and range extender)*

In plug-in HEV, the battery can be externally charged using an electric plug. By this way, more energy can be extracted from the batteries and the electric grid; while the ICE is not required to maintain the SOC of the battery in a light range. In some cases, the plug-in vehicle can just be a BEV with a small power ICE in order to extend the driving range by charging the batteries from the fuel, ICE, and generator, such case is also called *range extender* The fuel economy of plug-in hybrid can be improved 100% if the ICE is not used to charge the battery (e.g., in urban drive cycles).

Architectures and HEV functionality have to be differentiated. However, some architectures are dedicated to some HEV's functions (Table 2). BEV and FCV are not considered in this table because they use only electric power and all the functions are ensured by electrical machines.

Figure 7 shows two ways of powertrain technology road map. One way is continuous improvement of ICE and another way is development of advanced motor drives. While HEV needs the combination of both. The key technologies of EVs are motor drives and batteries. Figure 8 shows the evolution of motor drives technology, and Figure 9 shows the evolution of batteries technology.

Table 2. Functionality versus architectures for HEVs (main industrial solutions).

	Micro-HEV	Mild-HEV	Full-HEV	Plug-in HEV
Series–parallel	—	x	x	x
Parallel	x	x	x	—
Series	—	—	x	x

Note that Figures 8 and 9 just show how things have developed in the past and should continue in likely more rapid developments. Such as special structure EMs with better materials that suit better the higher performance of electric vehicles; advanced power electronic devices with higher power density and higher reliability; and advanced lithium batteries with new materials for positive electrode, negative electrode, electrolyte, and separator. There are claims of revolutionary next generation battery technologies including lithium-air, lithium-sulfur, solid state, and aqueous lithium.

4 ENERGY MANAGEMENT OF BEVs, HEVs, AND FCVs

Energy management of BEVs, HEVs, and FCVs is a real challenge (Kessels, 2007). Firstly, they are composed of various multiphysical subsystems, which have their own control constraints. Secondly, the aim of BEVs, HEVs, and FCVs is to replace classical thermal vehicles. In order to achieve this goal, the same (or better) dynamical performance must be obtained and the best energy management is required to reduce fuel consumption and pollutant emissions.

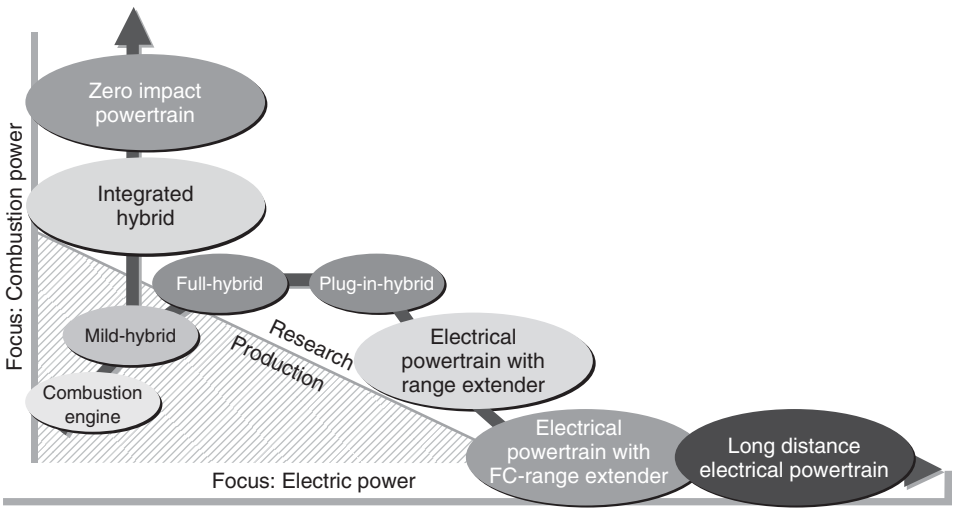


Figure 7. Two-way powertrain road map.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

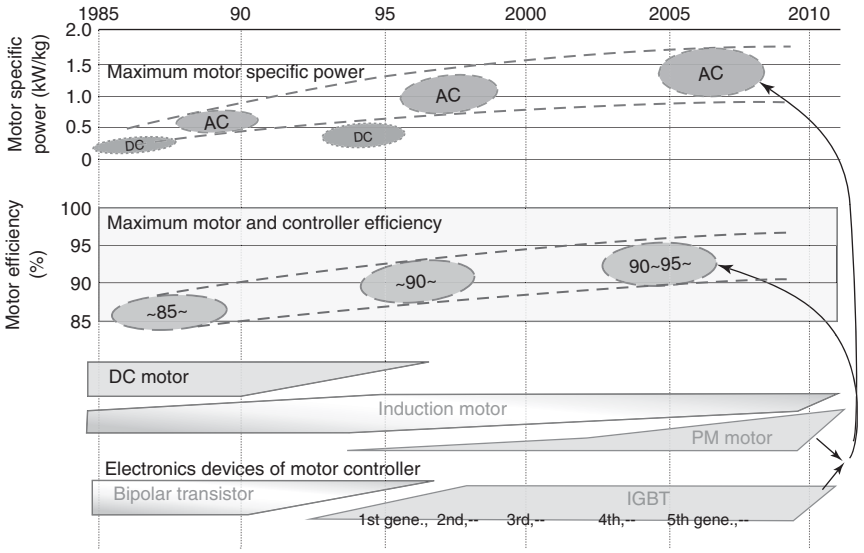


Figure 8. Evolution of motor drives technology.

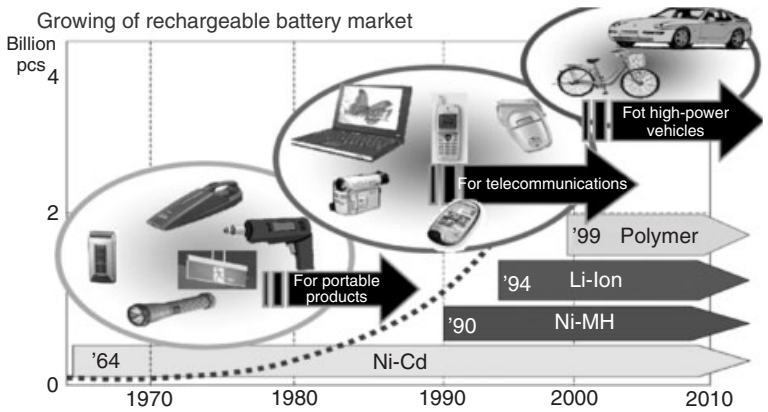


Figure 9. Evolution of battery technology.

It is essential to appreciate different control level (Moskowa, Munns, and Rubin, 1997). Because of the complexity of such systems, two control levels can be defined. The first one is associated with the local control of subsystems. The second control level is associated with the energy management of the whole system in order to coordinate all the subsystems.

Most of the papers on energy management of HEVs are only focused on the global control level. Nevertheless, appropriate local controls should be considered, and some constraints have to be considered on the subsystems.

1. *Local Control of Subsystems.* —The local control of subsystems is often neglected in study of energy management of HEVs. However, an efficient strategy cannot be achieved if the subsystems are not controlled in the best way. Most of the time the subsystems are considered as independent. From this assumption, classical controls of these devices are often used. However, from a systemic point of view, all the subsystems are interconnected. Their interactions have thus to be considered. Moreover, some associations of subsystem lead to modify their functional model (holistic approach). For example, when two electric motors and an IC engine are connected through a planetary gear, the three shafts have to be represented by a two dimension equivalent inertia (Bouscayrol, Hissel, and Trigui, 2008): only two independent rotation speeds have to be considered, in function of the three input torques.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Two specific speed controllers (and not three) are required to manage both speeds from the three torques. Moreover, the local control schemes have their own dynamics and limitations that impact on the global control.

2. *Global Control of the Whole System*. The global control has to co-ordinate the local subsystems in function of an energy management strategy. This supervision level generally leads to references for local control loops. For example, torque references of an electrical machine and ICE is often given from an energy management in parallel HEV. Because this global control has to manage the whole vehicle, it is often based on a simplified model. Backwards, noncausal, or steady-state models can be used to reduce the computation time.
3. *Local and Global Interactions*. Because this global control has to give reference to local control, its dynamics must be lower than the local closed-loop dynamics. This fact justifies the uses of quasi-static model to develop the global control. Moreover, limitations are often imposed during transient states (i.e., torque limitation). If the global control level does not consider limitations of local controls, the desired energy management cannot be achieved during these phases. It has been shown that, for systems with energy sharing, the energy nodes are a key issue (Bouscayrol, Hissel, and Trigui, 2008). The way to distribute energy is often multiple and several solutions can be used. These degrees of freedom have to be highlighted. In some works, this degree of freedom is used to connect the local controls and the energy management of vehicles (Guzzella and Amstuz, 1999; Bouscayrol *et al*., 2006; Won and Langari, 2005). In fact, the border of both control levels is not so obvious. The first step consists to define what can be considered in a local control and what can be considered in a global control.

In summary, energy management of HEVs, EVs, and FCVs using several energy sources are generally organized in two levels. The control of each subsystem has obviously to ensure to obtain the reference requested by the supervision level. Fast dynamics are required to respond quickly to the different power demand. The supervision of the whole system has to ensure the best energy management to reduce fuel consumption and pollutant emission. This global level has to distribute local references to the subsystem controls.

The decomposition between both control levels is not so obvious, when multiple subsystems strongly interact through energetic nodes. Functional causal descriptions can be very useful in these cases to analyze the constraints of the associations of local subsystems. Moreover, if one subsystem is in limitation, this information must be considered in the supervision level in order to distribute the power demand in a more appropriate way.

The controls of local subsystems are generally derived from expertise, such as ICE or electrical machine controls. For new components, such as fuel cell, inversion-based control is useful in order to exploit the subsystem in the best way. Dynamical causal models are thus required to develop these controls in real time.

The energy management (supervision) of the whole system has a lower dynamics than the controls of subsystems. Moreover, they have to consider the overall system from a global point of view.

5 COMMERCIALIZATION ROAD MAP

Cost reduction, reduction in size and weight, high performance, support of all stakeholders are the major issues for successful market penetration of EVs, HEVs, and FCVs. The support from government agencies, academic institutions, consumers, major components suppliers, and oil, gas, and electricity utilities is particularly critical. The key issues are as follows: What economical or other benefit that the users can get? What environmental benefit that the society can enjoy? Is it fun to drive without negative point as compared with conventional vehicles? Suitable legislative measures and incentives are essential in order to effectively reduce CO_2 and reduce dependence on oil. Levers that impact the penetration of electric and hybrid vehicles include fuel price, regulation and taxes, local legislation, purchase incentive, ownership intangibles such as green image and fun to drive, and public education.

The success of penetration of electric vehicles relies on availability of good product at reasonable cost; good infrastructure that is economy, efficient, and convenient; and innovative business model that can leverage the initial cost of batteries. The major criterion of the success of commercialization of electric vehicles lies on cost, usage convenience, energy consumption, and emission level. To reduce the cost, we should exploit the residual value of batteries. The roles of batteries is not only to provide energy and power while the vehicle is in driving mode but also to serve as energy storage to interact with the grid when the vehicle is not in driving mode. Thus, the electric power industry is in the position to exploit the value of the batteries during their service life and hence to globally optimize the usage of batteries. This needs the shake hand between two giant industries, namely automotive industry and electric power industry.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

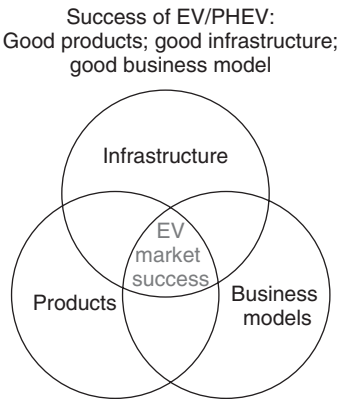


Figure 10. Three good factors.

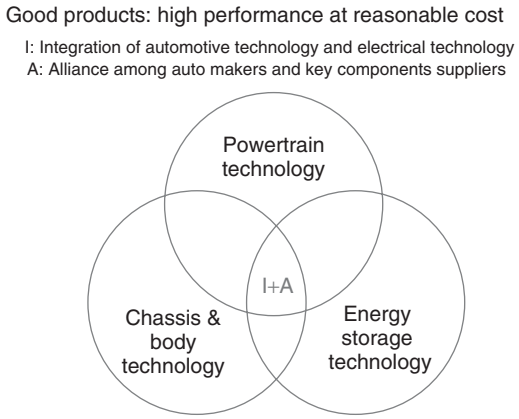


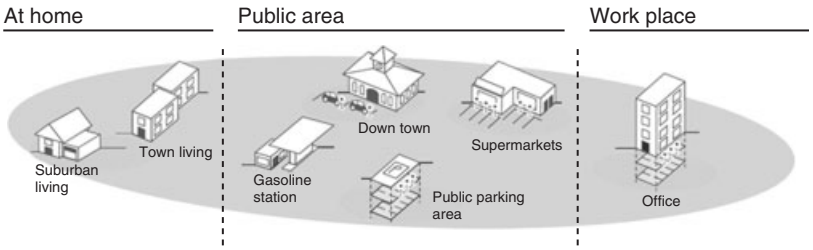
Figure 11. Good product approach.

Figures 10–14 illustrate these concepts.

In order to maximize the added value of electric vehicles, the integration of electric vehicles with smart grid (Ota *et al*., 2010) and also with information and communication technology is essential. Figures 15 and 16 explain these advantages. The batteries of an EV have two roles: it acts as energy and power storage for EV, and it also acts as a spinning reserve of power grid. While the grid is strong, the grid will charge the batteries, but when the grid is weak, the batteries deliver power to the grid. Thus, it can achieve three advantages, namely 1 to enhance the stability and load factor of the grid; 2 to enhance the life of the batteries, as the batteries are properly charge and discharge by controlling the SOC of the batteries; and 3 to reduce the loss of wind power and solar power that connected to the grid at serious weather condition, as the batteries can serve as storage to accept the dynamic power of wind and solar.

Figure 17 shows the EV commercialization road map. It can be seen that the first step is driven by government. This is essential in order to gain experience and enhance the quality and reliability. It seems that the promotion in public transportation will have distinct advantages as it can reduce pollution in urban area, whereas the route of public transportation is fixed so that the batteries operation can be well designed according to the duty cycles of the buses. In addition, it seems that the promotion of mini battery electric vehicle will have distinct advantages as the required energy and power of batteries is relatively small. Moreover, the structure topology of mini-EVs can be optimized, such as using light materials and innovative body design.

Figure 18 shows the forecast of EV penetration. It is difficult to make accurate forecast, since now the development



Parking durations	14 h per day	2 h per day	7 h per day
Charging points	1 charging point per vehicle	< 0.5 charging point per vehicle	1 charging point per vehicle
Power & charging time requirements	Low power and normal charging (e.g., 3 kW, 10 h)	High power and quick charging (e.g., 22 kW, 2 h)	Low power and normal charging (e.g., 3 kW, 7 h)

Figure 12. Infrastructure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

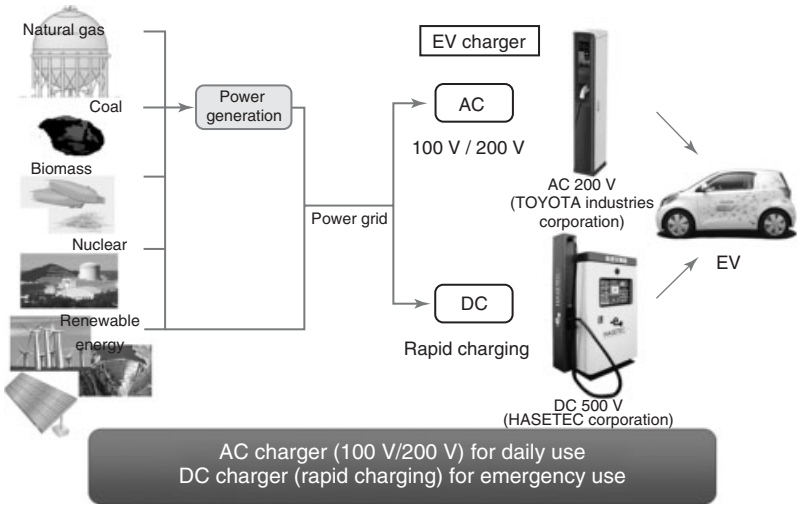


Figure 13. EV charger.

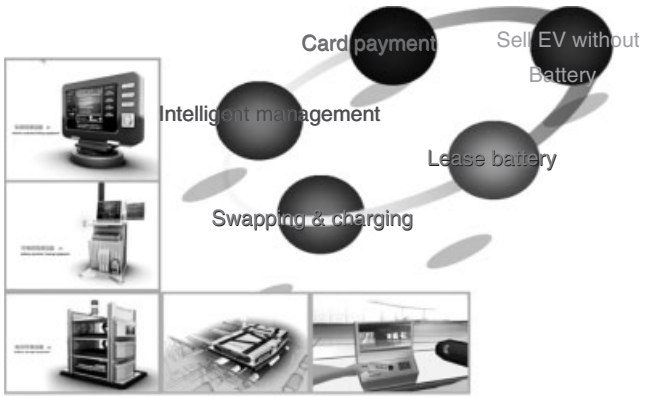


Figure 14. Possible business model.

of EV is in crucial and dynamic phase. Holistic approach of forecast is essential.

Electric vehicle industry is a disruptive industry and hence we should open mind, innovative design, and manufacturing as well as commercialization approach should be adopted. Figure 19 shows the distinct difference between traditional automotive industry and EV industry. It is expected that in addition to tradition automotive OEM, new EV manufacturers that only produce EV will appear. These new EV OEM will have no burden of ICE associated investment and hence will be able to create new value chain.

The standardization of electric vehicles and their infrastructure architectures will certainly enhance the safety, compatibility, and performance. Electric vehicle should be able to be charged safely everywhere. Figure 20 shows the related EV standards.

It is important that the company has proper commercialization road map. The company CEO should take lead in drawing the commercialization road map and the technical road map. This task should not just dedicated to the R&D department or sales department, as this is a major project, which will have major impact to the society and the company. In addition to have clear objectives, the senior management should also have holistic and creative thinking to oversee the progress of the project. The following is a success experience of a major automobile company in the development and commercialization of hybrid vehicles:

1. To have correct strategic plan, including near term, medium term, and long term.
2. To have sufficient funding to support the development plan.
3. To have innovative core technology, particularly innovative technology in integration of automotive technology and electric drive technology and energy storage technology.
4. To have correct technical road map. Understand the state of the art of technology, the trade-off among technology, cost, and market. Understand the different situation in the performance benefit versus cost characteristics in electric vehicles, hybrid vehicles, and fuel cell vehicles. A company may opt to produce whole spectrum of hybrids from micro to full hybrid for various types and sizes of vehicles. Another company may opt to focus on micro or mild hybrids for certain type and size of vehicles. A new OEM may focus only on the development of battery electric vehicles.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

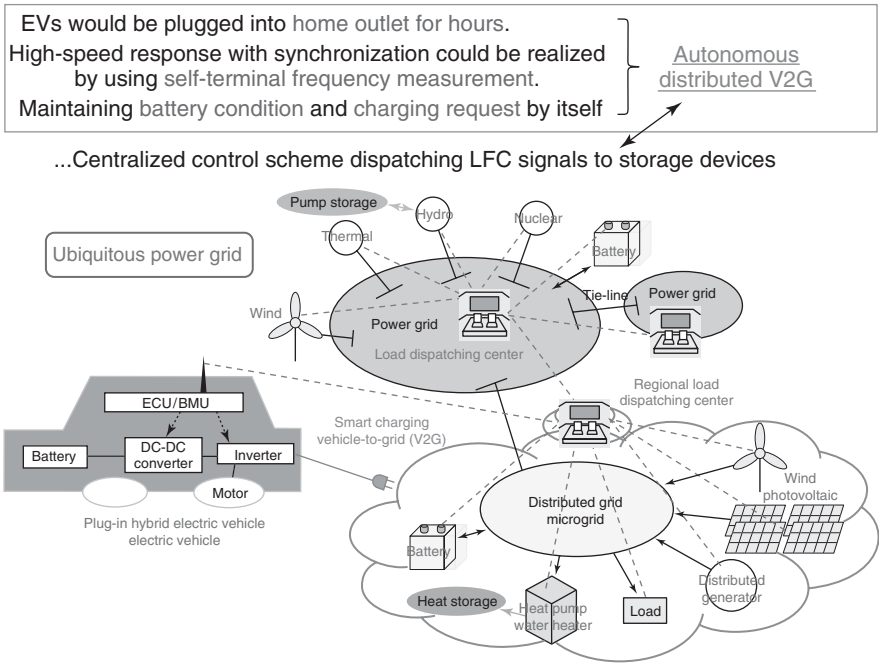


Figure 15. Integration of EV with smart grid.

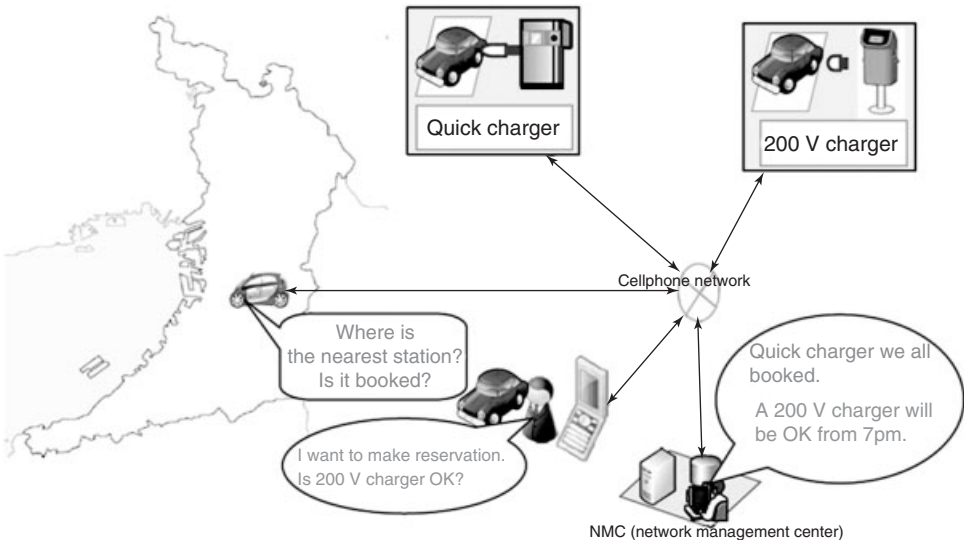


Figure 16. Integration of EV with information and communication technology.

5. Thoroughly understand the market demand and the required infrastructure and services.

6 CONCLUSION

In a world where environmental protection and energy conservation are growing concerns, the development of electric, hybrid, and fuel cell vehicles has taken on an accelerated pace. The dream of having commercially viable electric and hybrid vehicles is becoming a reality. Electric and hybrid vehicles are now available in the market. This chapter provides timely systematic review on the state of the art of electric, hybrid, and fuel cell vehicles with focus on technical features, architecture, energy management, and commercialization road map. The success of penetration of electric vehicles relies on the availability of good product

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

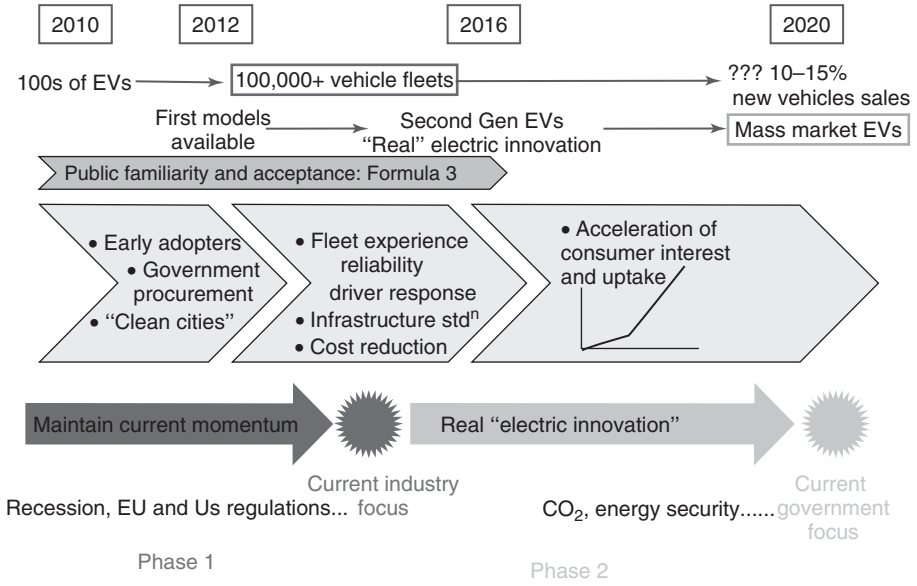
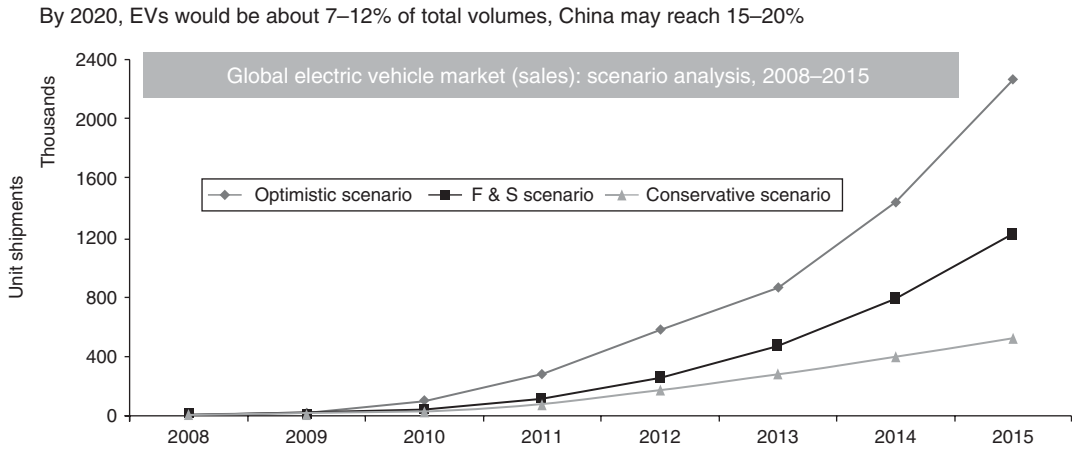


Figure 17. Electric vehicle commercialization road map.



Scenario's	2008	2009	2015	2020 (% of total car sales)
Optimistic scenario	5,103	17,475	2,266,450	12%
Frost & Sullivan scenario	5,103	8,911	1,226,607	7%
Conservative scenario	5,103	7,550	520,953	4%

Figure 18. Forecast of EVs penetration.

at reasonable cost; good infrastructure that is economy, efficient, and convenient; and innovative business model that can leverage the initial cost of batteries. The integration of electric vehicles with smart grid and with information and communication technology is essential. The standardization of electric vehicles and their infrastructure architectures will certainly enhance the safety, compatibility, and performance. Electric vehicle should be able to be charged safely everywhere.

We are excited to be involved in the development of electric vehicles that will have impacts to the welfare of our future generation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

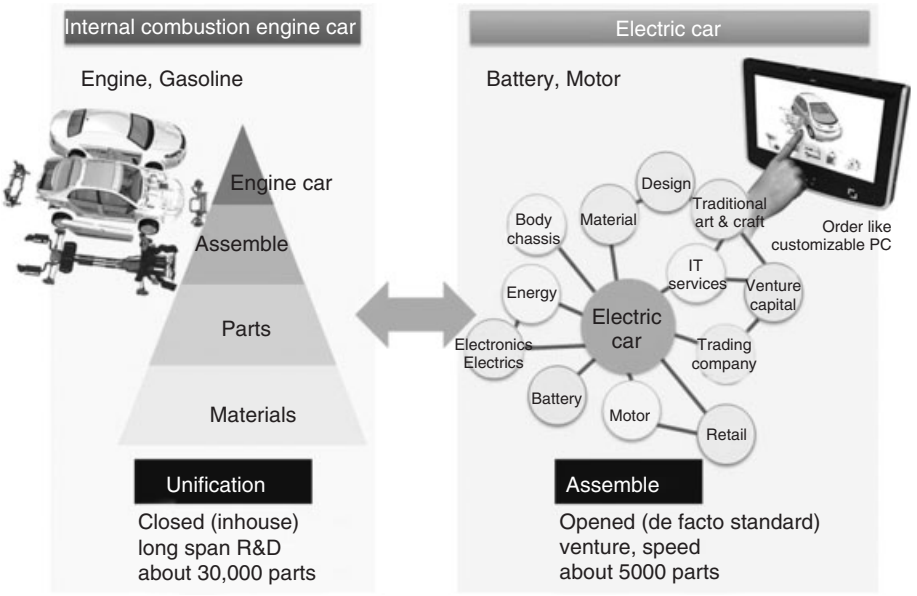


Figure 19. Changes in automotive industry.

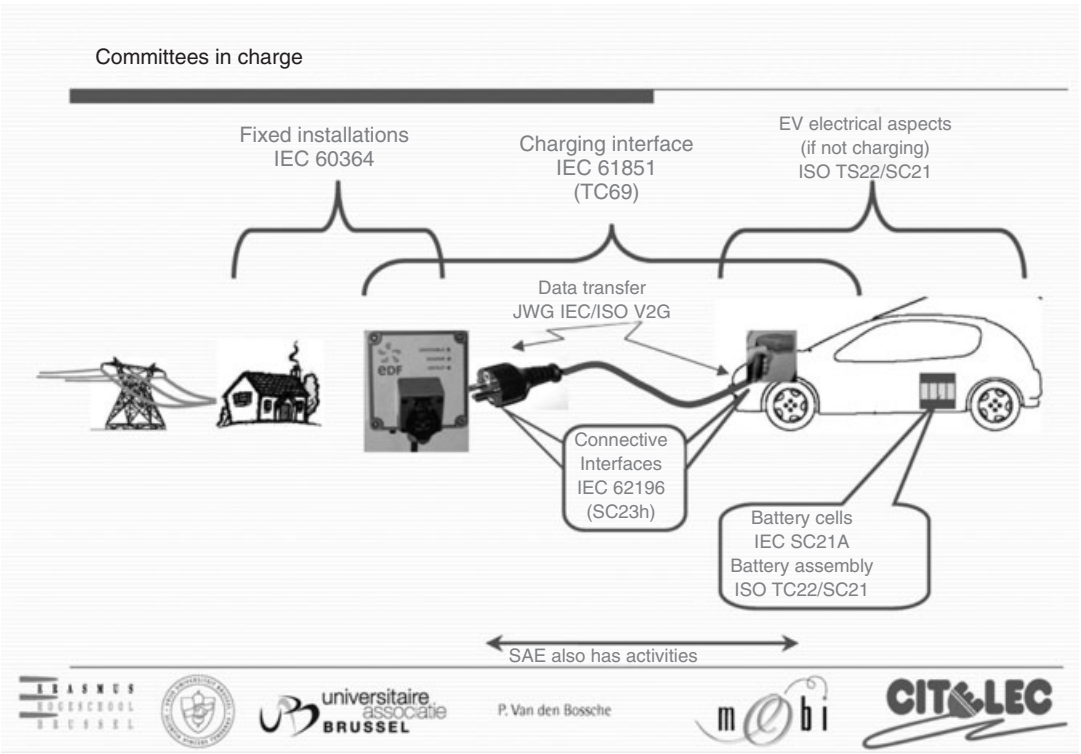


Figure 20. EV related standards.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ACKNOWLEDGMENT

Sections 3 and 4 of this chapter are based on the discussion with Prof. Alain Bouscayrol and Dr. Keyu Chen of University of Science and Technology Lille France, when the author delivered short course and seminar there in December 2007 and December 2008, respectively.

REFERENCES

Bouscayrol, A., Hissel, D., Trigui, R. (2008) Graphical Description for Inversion-Based Control of Hybrid Electric Vehicles. *In EET IAMF*, Geneva.

Bouscayrol, A., Lhomme, W., Delarue, P., *et al*. (2006) Hardware-in-the-Loop Simulation of Electric Vehicle Traction Systems Using Energetic Macroscopic Representation. *In Proceedings of IEEE-IECON*, Paris; pp. 4999–5004.

Burke, A.F. (2002) Cost-Effective Combinations of Ultracapacitors and Batteries for Vehicle Applications. *In Proceedings of the Second International Advanced Automotive Battery Conference*, Las Vegas.

Chan, C.C. (2007) The state of the art of electric, hybrid, and fuel cell vehicles. *Proceedings of the IEEE*, **95** (4), 704–718.

Chan, C.C., Bouscayrol, A., and Chen, K. (2010) Electric, hybrid and fuel cell vehicles: Architectures and modeling. *IEEE Transactions on Vehicular Technology*, **59** (2), 589–598.

Chan, C.C. and Chau, K.T. (2001) *Modern Electric Vehicle Technology*, Oxford University Press, London.

Chen, K., Cheng, Y., Bouscayrol, A., *et al*. (2008) Inversion-Based Control of a Hybrid Electric Vehicle Using a Split Electric Variable Transmission. In IEEE VPPC, Harbin; 1–6.

Cheng, Y., Chen, K., Chan, C.C., *et al*. (2008) Global Modeling and Control Strategy Simulation for a Hybrid Electric Vehicle Using Electrical Variable Transmission. In IEEE VPPC, Harbin; 1–5.

Cheng, Y., Cui, S., Song, L., *et al*. (2007) The study of the operation modes and control strategies of an advanced electromechanical converter for automobiles. *IEEE Transactions on Magnetics*, **43** (1), 430–433.

Chau, K.T. and Chan, C.C. (2007) Emerging energy-efficient technologies for hybrid electric vehicles. *Proceedings of the IEEE*, **5** (4), 821–835.

Chau, K.T., Chan, C.C., and Liu, C. (2007) Overview of permanent-magnet brushless drives for electric and hybrid electric vehicles. *IEEE Transactions on Industrial Electronics*, **55** (6), 2246–2257.

Ehsani, M., Gao, Y., and Miller, J.M. (2007) Hybrid electric vehicles: Architecture and motor drives. *Proceedings of the IEEE*, **95** (4), 719–728.

Eshani, M., Gao, Y., Gay, S.E., *et al*. (2005a) *Modern Electric, Hybrid Electric and Fuel Cell Vehicles*, CRC Press, New York.

Eshani, M., Gao, Y., Gay, S.E., *et al*. (2005b) *Modern electric, hybrid electric and fuel cell vehicles*, CRC Press, New York.

Elliot, C. and Deasley, P. (2007) *The Royal Academy of Engineering Report: Creating Systems that Works*, The Royal Academy of Engineering, London.

Gao, D.W., Mi, C., and Emadi, A. (2007) Modeling and simulation of electric and hybrid vehicles. *Proceedings of the IEEE*, **95** (4), 729–745.

Gao, Y. and Ehsani, M. (2006) A torque and speed coupling hybrid drivetrain-architecture, control, and simulation. *IEEE Transactions on Power Electronics*, **21** (3), 741–748.

Guzzella, L. and Amstuz, A. (1999) CAE tools for quasi-static modeling and optimization of hybrid powertrains. *IEEE Transaction on Vehicular Technology*, **48** (6), 1762–1769.

Hissel, D., Candusso, D., and Harel, F. (2007) Fuzzy-clustering durability diagnosis of polymer electrolyte fuel cells dedicated to transportation applications. *IEEE Trans on Vehicular Technology*, **56** (5), 2414–2420.

Hoeijmakers, M.J. and Ferreira, J.A. (2006) The electric variable transmission. *IEEE Transactions on Industry Applications*, **42** (4), 1092–1100.

International Energy Agency Implementing Agreement on Hybrid and Electric Vehicles. (2008) Outlook for hybrid and electric vehicles [EB/OL]. (2008-06-01), http:// www.ieahev.org/pdfs/ia-hev_outlook_2008 (accessed 9 October 2013).

Kessels, J. (2007) *Energy Management for Automotive Power Net*, TU Endhoven, The Netherlands.

Lai, J.S. and Nelson, D.J. (2007) Energy management power converters in hybrid electric and fuel cell vehicles. *Proceedings of The IEEE*, **95** (4), 766–777.

Miller, J.M. (2006) Hybrid electric vehicle propulsion system architectures of the e-CVT type. *IEEE Transactions on Power Electronics*, **21** (3), 756–767.

Moskowa, J.J., Munns, S.A. and Rubin, Z.J. (1997) The development of vehicular powertrain system modeling methodologies: philosophy and implementation. In *SAE'97*, Detroid; Paper 971089.

Ota, Y., Taniguchi, H., Nakajima, T., *et al*. (2010) Autonomous Distributed Vehicle-to-Grid (V2G) for Ubiquitous Power Grid and Its Effect as a Spinning Reserve. *In Proceedings of International Electrical Engineering Conference (ICEE 2010)*, Busan.

Sepe, R.B., Morrison, C.M., Miller, J.M., *et al*. (2001) High Efficiency Operation of a Hybrid Electric Vehicle Starter/Generator over Road Profile. *In IEEE Industry Application Society Annual Meeting*, Chicago; 921–925.

Syed, F.U., Kuang, M.L., Czubay, J., *et al*. (2006) Derivation and experimental validation of a power-split hybrid electric vehicle model. *IEEE Transactions on Vehicular Technology*, **55** (6), 1731–1747.

Won, J.S. and Langari, R. (2005) Intelligent energy management agent for a parallel hybrid vehicle; part II: torque distribution, charge sustenance strategies and performance results. *IEEE Transactions on Vehicular Technology*, **54** (3), 935–952.

Yamada, K., Hanada, H. and Sasaki, S. (2006) The Motor Control Technologies for GS450h Hybrid System. In EVS-22, Yokohama, 827.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto061
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Rechargeable Battery Basics

Andrew Burke
University of California, Davis, CA, USA

1 Introduction 1
2 Battery Nomenclature and Parameters 1
3 Basic Concepts 2
4 Battery Performance 5
5 Battery Life 10
6 Summary and Conclusions 15
Related Articles 15
References 15

1 INTRODUCTION

This chapter is concerned with various aspects of a basic understanding of rechargeable batteries for automotive applications, including cell construction and operation, testing/characterization, and battery performance in hybrid and electric vehicles (EVs). Emphasis is placed on lithium-ion batteries, but some consideration is also given to lead–acid and nickel metal hydride batteries. After summarizing the nomenclature and parameters to be used in subsequent sections, the first section of the chapter deals with basic concepts of cell design and testing for various battery chemistries. Next the performance—energy storage and power capability—of the batteries is considered including a discussion of installation in vehicles, battery charging, and battery management system (BMS) to mitigate safety issues with lithium batteries. The final sections of this chapter are concerned with battery life and cost. Both calendar and cycle life and factors that effect life for the various chemistries are considered. Battery cost estimates are given and life cycle and break-even fuel costs for vehicle applications are presented.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 BATTERY NOMENCLATURE AND PARAMETERS

In this section, the general nomenclature to be used in subsequent sections of this chapter is discussed and the key parameters to be used to describe the batteries considered are introduced.

2.1 Nomenclature

2.1.1 Battery cell

The cell is the basic building block of the battery (Figure 1). It consists of two electrodes and a separator. The open-circuit voltage of the cell is uniquely set by the cell chemistry.

2.1.2 Battery pack

The battery pack is the combination of the cells connected in series and parallel to achieve the system voltage and energy storage (kWh) required by the vehicle.

2.1.3 Battery chemistry

Battery chemistry refers to the electrochemical reactions that occur in the cell. These reactions are dependent on the active materials used in the electrodes and the electrolyte used in the cell (Sections 3.1 and 3.2). The battery types are named to indicate their chemistry (e.g., lead–acid, nickel metal hydride, and lithium iron phosphate).

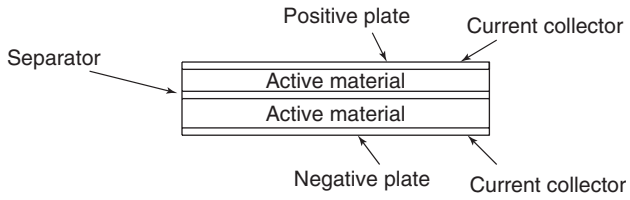


Figure 1. Basic cell geometry and construction.

2.1.4 Performance characteristics

The performance characteristics indicate the capability of the cells/pack to store electrical energy and provide electrical power for the vehicle. These characteristics are given in terms of the energy density (Wh/L) and power density (W/L) of the battery (Sections 4.1 and 4.2).

2.1.5 Battery management

Battery management refers to maintaining the cell voltages and temperatures during charging and discharging of the battery in specified ranges required for its safe and long life operation.

2.2 Battery parameters

2.2.1 Cell parameters

The key parameters describing the cell are the cell voltage V_{cell} and the ampere-hour capacity (Ah) of the cell. The cell voltage depends on the cell chemistry and the Ah capacity depends on the size of the cell. The energy (Wh) stored in the cell is given by $V_{\text{cell}} \times \text{Ah}$.

2.2.2 Energy density

The energy density of the cell or pack is determined by dividing the energy stored by the weight or volume of the cell or pack and is expressed as Wh/kg and Wh/L. The energy density is one of the key characteristics of the cells and pack that is determined from laboratory testing. The energy density is dependent on the discharge time of the test (Section 4.1).

2.2.3 Power density

The power density of the cell or pack is determined by dividing the maximum power that can be provided by the weight or volume of the cell or pack and is expressed as W/kg and W/L. The maximum power that can be provided depends markedly on the time period over which the power is sustained and the acceptable voltage drop (efficiency at which the power is provided) (Section 4.2).

2.2.4 Charge and discharge rates (n*C rate)*

The charge and discharge rate is often expressed as nC where n is defined as charge/discharge current divided by the Ah capacity of the cell ($n = I/\text{Ah}$). High power batteries can be discharged at nC rates >20C.

2.2.5 Cycle life

Cycle life for a battery is the number of times that a cell or pack can be cycled before a specified degradation in energy or power capacity occurs. This characteristic is determined from laboratory testing. The cycle life depends markedly on the temperature, the current or power profiles, and voltage limits of the tests. The calendar life of the battery depends on the calendar time over which the battery can be cycled before its performance degrades by a specified factor (Sections 5.1 and 5.2).

2.2.6 Battery cost

The battery cost is usually expressed as \$/kWh, which is simply the cost of the battery divided by the rated kWh energy storage capacity. For vehicle applications, the battery cost should be calculated for the complete pack with all accessory needed to operate it (Section 5.3).

3 BASIC CONCEPTS

3.1 Cell construction and operation

All batteries regardless of the detailed chemistry function in essentially the same way. As indicated in Figure 1, all batteries consist of positive and negative terminals/electrodes with a separator between the electrodes that conduct the ions exchanged and the blocks conduction of electrons that must pass through an external circuit. The electrodes and separator are porous and contain an electrolyte (usually a liquid or a gel), which has low resistivity for ion transport. The electrode materials are applied to a thin metallic foil or grid that collects/distributes the electrons to the electrodes from the positive and negative tabs of the cell. For the high power batteries used in vehicles, the electrodes are relatively thin being in the range of 100 μm.

The performance of the battery depends primarily on the materials used in the electrodes and the electrochemical reactions that occur between the ions in the electrolyte and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the solid material particles in the electrode. The mechanisms of these reactions and the ion transport in the porous electrodes are complex (Reddy, 2011; Yoshio, Brodd, and Kozawa, 2009; Nazri and Pistoia, 2003; Huggins, 2009), but the reactions in the electrodes can often be written in rather simple terms. For example, consider a nickel metal hydride cell in discharge, for which the electrochemical reactions are both electrodes are shown in Equations 1–3, where E_0 is the standard reduction potential. For charging, the reactions shown take place in the opposite direction.

Positive electrode:

$$\mathrm{NiOOH} + \mathrm{H_2O} + \mathrm{e} \rightarrow \mathrm{Ni(OH)_2} + \mathrm{OH^-}\ E_0 = 0.52\ \mathrm{V} \quad (1)$$

Negative electrode:

$$\mathrm{MH} + \mathrm{OH^-} \rightarrow \mathrm{M} + \mathrm{H_2O} + \mathrm{e}\ E_0 = -0.83\ \mathrm{V} \quad (2)$$

Total cell reaction:

$$\mathrm{MH} + \mathrm{NiOOH} \rightarrow \mathrm{M} + \mathrm{Ni(OH)_2} V_{\mathrm{cell}} = 1.35\ \mathrm{V} \quad (3)$$

MH is a mixed-metal oxide (Reddy, 2011; Linden and Reddy, 2002; Bennett and Sakai, 1994) that can store the hydrogen when the cell is charged. The electrolyte is 30% potassium hydroxide (KOH) in water. Note that the cell functions via the conduction (diffusion) of the hydroxide ion OH^- between the electrodes. In the case of the lithium batteries (Reddy, 2011; Yoshio, Brodd, and Kozawa, 2009; Nazri and Pistoia, 2003; Huggins, 2009), the Li^+ ion diffuses between the electrodes. Both battery chemistries are referred to as *swing* batteries; in both cases, the ion concentration in the electrolyte remains unchanged as the cell is charged/discharged. All the reactants are provided by the electrode materials and not the electrolyte. Experience has shown that "swing" batteries typically have long cycle life.

The voltages shown for each of the electrode reactions are the standard reduction potentials at 25°C (Reddy, 2011), which can be calculated from the differences in the free energy of the reactants and products of the reactions written in the "reduction" direction. The cell voltage is given by

$$V_{\mathrm{cell}} = (E_0)_+ - (E_0)_- \quad (4)$$

The theoretical charge transfer (Ah/g) in each half-reaction can be calculated from the following relationship:

$$\mathrm{Ah/g} = \frac{nF}{\mathrm{MW}}, F = 26.8\ \mathrm{Ah/mole} \quad (5)$$

where F is the Faraday constant, n the number of electrons involved in the half-reaction, and MW the molecular weight (gram) of the reactant.

Table 1. Electrochemical characteristics of various battery reactants.

Reactant	Valence	Standard Potential (V)	Charge Density (Ah/g)	Ah/cm^3
Hydrogen (H^+)	1	0.0	26.6	—
Lithium (Li^+)	1	−3.01	3.86	2.06
Manganese (Mn^{++})	2	−1.25	0.976	7.34
Iron (Fe^{++})	2	−0.44	0.960	7.54
Zinc (Zn^{++})	2	−0.76	0.82	5.82
Magnesium (Mg^{++})	2	−2.38	2.2	3.83
Lead (Pb^{++})	2	−0.13	0.259	2.94
Silver (Ag^+)	1	0.80	0.248	2.60
Copper (Cu^{++})	2	0.34	0.843	7.49
Nickel oxide (NiOOH)	1	0.45	0.29	2.16
Lead dioxide (PbO_2)	2	1.69	0.22	2.11
Manganese dioxide (MnO_2)	1	1.23	0.31	1.54
Lithium cobalt oxide ($LiCoO_2$)	0.5	−0.70	0.137	—

For a particular cell, the g/Ah for the total reaction (sum of the half-reactions) is given by

$$(\mathrm{g/Ah})_{\mathrm{cell}} = \left(\frac{\mathrm{MW}}{nF}\right)_- + \left(\frac{\mathrm{MW}}{nF}\right)_+ \quad (6)$$

The theoretical energy density of the cell is given by

$$(\mathrm{Wh/kg})_{\mathrm{theo.}} = V_{\mathrm{cell}}/(\mathrm{g/Ah})_{\mathrm{cell}} \times 1000 \quad (7)$$

For the nickel metal hydride cell cited above,

$$(\mathrm{Ah/g})_- = 0.45, (\mathrm{Ah/g})_+ = 0.292, (\mathrm{g/Ah})_{\mathrm{cell}} = 5.64$$

$$(\mathrm{Wh/kg})_{\mathrm{theo}} = 1000 \times 1.35/5.64 = 239$$

These theoretical calculations can be done for any electrochemical cell after the reactants and the electrolyte are postulated and their properties are known. The electrochemical properties of most reactants used in batteries are given in the study by Reddy (2011). Those for selected reactants are given in Table 1.

3.2 Battery chemistries

There are a number of battery types/chemistries being developed that can be used in vehicles. The chemistries of the various battery types are summarized in Table 2. The potential (theoretical limit) performance of each type is shown in Table 3 along with a practical performance to be

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Battery chemistries for vehicle applications.

Battery Type	Negative Electrode	Positive Electrode	Electrolyte
Lead–acid	Pb	PbO_2	Sulfuric acid
Nickel cadmium	Cd	NiOOH	Potassium hydroxide (KOH)
Nickel metal hydride	Mixed metal hydrides	NiOOH	Potassium hydroxide (KOH)
Lithium–nickel cobalt	graphite	Nickel cobalt oxide	Organic
Lithium manganese	graphite	Lithium manganese oxide	Organic
Lithium iron phosphate	graphite	Iron phosphate	Organic
Lithium titanate oxide	Lithium titanate oxide	Lithium manganese oxide	Organic
Sodium metal chloride (300°C)	Sodium (Na)	Ni chloride	Ceramic β-alumina
Zinc–air	Zinc (Zn)	Rechargeable air (O_2)	Potassium hydroxide (KOH)
Magnesium–air	Magnesium (Mg)	Rechargeable air (O_2)	Potassium hydroxide (KOH)
Lithium–air	Lithium metal (Li)	Rechargeable air (O_2)	Organic

Table 3. Performance characteristics of battery chemistries for vehicle applications.

	Theoretical			Practical		
Battery Type	Maximum Voltage (V)	Charge Transfer (Ah/g)	Energy Density (Wh/kg)	Nominal Voltage (V)	Energy Density (Wh/kg)	(Wh/L)
Lead–acid	2.1	0.12	252	2.0	35	100
Nickel cadmium	1.35	0.181	244	1.2	40	90
Nickel metal hydride	1.35	0.178	240	1.2	80	220
Lithium–nickel cobalt	4.1	0.109	448	3.8	200	420
Lithium manganese	3.5	0.122	426	3.0	150	400
Lithium iron phosphate	3.65	0.111	405	3.4	115	255
Lithium titanate oxide	2.8	0.090	252	2.5	75	150
Lithium–sulfur	2.5	0.341	950	2.15	400	365
Sodium metal chloride	2.6	0.22	572	2.5	120	190
Zinc–air	1.6	0.82	1312	1.2	400	900
Magnesium–air	3.1	2.2	6820	1.4	800	1390
Lithium–air	3.4	3.68	13124	2.7	2500	3750

expected based on past experience in battery development. Experience has shown that in most cases, the operating voltage of the cell is reasonably close to the theoretical value, but the energy density (Wh/kg and Wh/L) is much less than the theoretical value by a factor at least 2–3. In the cases of the metal-air batteries, the reduction factor is much larger as the weight of the positive electrode is not included in the theoretical calculation because the active material (air) for the positive electrode comes from the environment and not the battery itself. Regardless, the practical energy densities of the metal-air batteries are projected to be much higher than the other battery chemistries. Those battery types are presently in an early stage of development and their future performance is very uncertain at the present time.

The cell theoretical performance values shown in Table 3 can be calculated using Equations 4–8 with pertinent inputs taken from the study by Reddy (2011). The practical performance values were taken from the study by Reddy (2011); Sion Power (2008), Lithium Sulfur Rechargeable Battery Data Sheet; and Burke and Miller (2009aa,b) based on test data for the batteries of various types and projections of future development of the battery technologies.

In the near term (5–10 years), the lithium batteries of various types are likely to be the technology used in most electric and hybrid vehicles. These batteries have relatively high energy and power densities and long cycle life as will be discussed in later sections of this chapter. In the long term (10–20 years), the metal-air batteries might be available for use in plug-in hybrid electric vehicles (PHEVs) making feasible long ranges (>150 miles).

3.3 Cell characterization and testing

Detailed data are needed concerning the performance of cells before they can be incorporated in the design of an electric or hybrid vehicle. The performance of cells can be characterized in terms of the following parameters:

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

(a) Charge stored (Ah) The *Peukert curve* (Ah vs discharge time) shows how the charge stored varies with the discharge time.
(b) Energy stored (Wh) The *Ragone curve* (Wh/kg vs W/kg) shows how the energy stored varies with discharge power *W*.
(c) Resistance (*R*) The *pulse power density* (W/kg or W/L) can be calculated from the measured resistance *R* and open-circuit voltage V_{oc} as a function of state of charge (SOC).

Determination of these parameters requires extensive testing of the cells. Detailed procedures for the testing have been developed by the United States Advanced Battery Consortium (USABC) and other organizations (Electric vehicle battery test procedures manual, 1996; Battery test manual for plug-in hybrid electric vehicles, 2008; FreedomCAR battery test manual for power-assist hybrid electric vehicles, 2003). These procedures require testing the cell over a range of charge and discharge currents and discharge powers at both steady and pulsed conditions. The series of tests for lithium cells should include at least the following:

(a) constant current at C/3, C/2, 1C, 2C;
(b) constant power at W/kg corresponding to C/3, C/2, C/1, 2C;
(c) pulse tests to determine the resistance at 2C, 4C, 6C; and
(d) charging at C/2, 1C, 2C, 3C, 4C.

When modules are available, similar tests should be performed on the modules. Examples of test data from this type of testing will be presented in the next section.

4 BATTERY PERFORMANCE

4.1 Performance data for lithium cells

Test data are available (Burke and Miller, 2009a, b) for most of the lithium cell types listed in Tables 2 and 3. A summary of test data are given in Table 4 for cells from a number of manufacturers. It is clear from Table 4 that there are significant differences in both the energy density and power capability of cells using the different lithium chemistries and both the Ah capacity and resistance of cells can vary over a wide range. The cell characteristics for the various lithium cell chemistries are summarized in Table 5.

Detailed test data similar that shown in Tables 6–8 are needed to characterize the performance of batteries for use in a vehicle. Such data are seldom available from the battery manufacture and almost never included on the spec sheets for the cells. A complete set of performance data is given in Table 6 for a Kokam 30 Ah cell. From this data, both the Peukert and Ragone curves for the cell can be plotted and the power capability calculated (Section 3.2). In addition, the open-circuit voltage and the resistance of the cell as a function of SOC can be determined for use in vehicle simulations.

Table 4. Summary of the performance characteristics of lithium-ion batteries for various chemistries.

Battery Developer/Cell Type	Electrode Chemistry	Voltage Range (V)	Capacity (Ah)	Resistance (mΩ)	Energy Density (Wh/kg)	Power Density (W/kg) 90% efficiency[a]	Power Density (W/kg) Match. Imped.	Weight (kg)	Density (gm/cm^3)
Enerdel HEV	Graphite/Ni MnO_2	4.1–2.5	15	1.4	115	2010	6420	0.445	—
Enerdel EV/PHEV	Graphite/Ni MnO_2	4.1–2.5	15	2.7	127	1076	3494	0.424	—
Kokam prismatic	Graphite/$NiCoMnO_2$	4.1–3.2	30	1.5	140	1220	3388	0.787	2.4
Saft Cylinder	Graphite/NiCoAl	4.0–2.5	6.5	3.2	63	1225	3571	0.35	2.1
GAIA Cylinder	Graphite/$NiCoMnO_2$	4.1–2.5	40	0.48	96	2063	5446	1.53	3.22
			7	3.6	78		3472	0.32	—
A123 Cylinder	Graphite/Iron Phosphate	3.6–2.0	2.2	12	90	1393	3857	0.07	2.2
Altairnano prismatic	LiTiO/$NiMnO_2$	2.8–1.5	11	2.2	70	990	2620	0.34	1.83
Altairnano prismatic	LiTiO/$NiMnO_2$	2.8–1.5	3.8	1.15	35	2460	6555	0.26	1.91
Quallion Cylinder	Graphite/NiCo	4.2–2.7	1.8	60	144	577	1550	0.043	2.6
Quallion Cylinder	Graphite/NiCo	4.2–2.7	2.3	72	170	445	1182	0.047	2.8
EIG prismatic	Graphite/$NiCoMnO_2$	4.2–3.0	20	3.1	165	1278	3147	0.41	—
EIG prismatic	Graphite/iron Phosphate	3.65–2.0	15	2.5	113	1100	3085	0.42	—
Panasonic EV prismatic	Ni metal hydride	7.2–5.4	6.5	11.4	46	395	1093	1.04	1.8

[a] Power density P = Efficiency*(1−Efficiency) $V_{oc}{}^2/R$, $P_{\text{match. imped.}} = V^2/4R$.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 5. Characteristics of lithium-ion batteries using various chemistries.

Chemistry Anode/Cathode	Cell Voltage (V) Maximum/Nom.	Charge Transfer (Ah/g) Anode/Cathode	Energy Density (Wh/kg)	Cycle Life (Deep Cycle)	Thermal Stability
Graphite/$NiCoMnO_2$	4.2/3.6	0.36/0.18	100–170	2000–3000	Fairly stable
Graphite/Mn spinel	4.0/3.6	0.36/0.11	100–120	1000	Fairly stable
Graphite/$NiCoAlO_2$	4.2/3.6	0.36/0.18	100–150	2000–3000	Least stable
Graphite/iron phosphate	3.65/3.25	0.36/0.16	90–115	>3000	Stable
Lithium titanate/Mn spinel	2.8/2.4	0.18/0.11	60–75	>5000	Most stable

4.2 Power capability and cell design trade-offs

4.2.1 Power capability of batteries

In discussing the power capability of batteries (Burke and Miller, 2011), it is necessary to specify the time of the charge or discharge and the conditions under which the energy transfer takes place. By this is meant, the rate at which the energy stored in the device is transferred and at what SOC and/or voltage the transfer process is started and ended. The simplest process is a constant power discharge or charge of a device, which is customarily done to determine the energy density (Wh/kg or Wh/L). This test is usually started at full charge and is terminated at a specified cutoff or final voltage. These voltages are device chemistry dependent. In this test, the usable energy is measured for different power densities (W/kg) and the power density at which the usable energy begins to decrease markedly (e.g., has decreased by 20%) is determined. This power density is termed the $(W/kg)_{const.,\ max}$. It can be easily determined from constant power testing and there is little reason why the constant power capability should be unclear.

The definition and subsequent determination of the pulse power capability is not as straightforward as that of constant power capability. This is because the pulse power capability is highly dependent on the voltage drop permitted during the pulse and the duration (s) of the pulse. In general, the power capability is higher if a larger voltage drop is permitted and the duration of the pulse is shorter. The power capability of a battery is SOC dependent (lower at low SOCs). Hence, the test procedures should clearly state the SOC of the device and the voltage drop and duration of the pulse. Further the power capabilities of devices should be compared only for equivalent voltage drops/ranges, pulse times, and SOCs.

In general, there are two approaches to setting a limit to the power that can be taken from a battery (Burke and Miller, 2011). The first approach is to set a minimum voltage for a discharge pulse and a maximum voltage for a charge pulse that the device can experience during the pulse. This is the approach proposed by the United States Advanced Battery Consortium (USABC) (1996, 2003, 2008) in Electric vehicle battery test procedures manual, FreedomCAR battery test manual for power-assist hybrid electric vehicles, and Battery test manual for PHEVs. The initial voltage before the pulse is the open-circuit voltage (V_{oc}) at the stated SOC of the device. The maximum power then occurs at the current for which

$$V_{ch,max} - V_{oc} = I_{ch}R, \qquad P_{ch,max} = I_{ch}V_{ch,max}$$
$$V_{oc} - V_{disch,min} = I_{disch}R, P_{disch,max} = I_{disch}V_{disch,min}$$

R, which is the resistance of the device, varies with SOC and is determined from pulse tests of the device. The equations shown are the simple expressions for the ohmic voltage change due to the pulse current.

The second approach is concerned with the efficiency (EF) of the pulse or the fraction of the energy transferred from the device that is electrical energy rather than heat. In simplest terms using Ohm's law for a DC device,

$$P = VI, I = P/V, V = V_{oc} - IR$$

The efficiency is given by

$$\mathrm{EF} = P/(P + I^2R) = 1/(1 + IR/V) = V/V_{oc} \quad (8)$$

and the maximum power of the pulse becomes

$$P_{bat,max} = \mathrm{EF}(1 - \mathrm{EF}){V_{oc}}^2/R \quad (9)$$

The derivation of Equations 8–9 relating efficiency and maximum power to the voltage range of the pulse neglects the change in resistance during the pulse, but the equations show the direct relationship in principle between the power, open-circuit voltage, and resistance of the device. The efficiency at which the power capability is a maximum can be determined by differentiating Equation 9 with respect to EF. One finds that the power is a maximum for $EF = 1/2$ resulting in a maximum power of

$$P_{max,EF} = {V_{oc}}^2/4R \quad (10)$$

which is the well-known relationship for the matched impedance power of a device.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 6. UC Davis test data for the Kokam high power cells.

Constant current discharges 4.1–3.2 V			
Current (A)	Time (s)	Ah	*n*C
15	7417	30.9	0.485
30	3611	30.1	1.0
60	1728	28.8	2.08
100	976	27.1	3.69
150	603	25.1	5.97

Charge at 31 A to 4.2 V and taper to 1.5 A

Constant power discharges 4.1–3.2 V				
Power (W)	Time (s)	Wh	Wh/kg	W/kg
43	9806	117.1	148.8	55
82	4835	110.1	139.8	104
162	2355	106.0	137.7	206
242	1509	101.4	128.9	308
321	1097	97.8	124.3	408
402	854	95.4	121.2	511
482	674	90.2	114.6	612

Cell weight 787 g

Cell resistance based on pulse tests 5 s pulses, V at 2 s		
State of charge	Current (A)	Resistance (mΩ)
80%	150 A discharge	1.6
$V_{oc} = 3.94$	200 A discharge	1.6
	310 A discharge	1.55
	50 A charge	1.6
	100 A charge	1.6
	150 A charge	1.53
60%		
$V_{oc} = 3.75$	150 A discharge	1.53
	200 A discharge	1.5
	310 A discharge	1.45
	50 A charge	1.6
	100 A charge	1.5
	150 A charge	1.53
40%		
$V_{oc} = 3.64$	150 A discharge	1.53
	200 A discharge	1.5
	310 A discharge	1.42
	50 A charge	1.6
	100 A charge	1.6
	150 A charge	1.53
20%		
$V_{oc} = 3.55$	150 A discharge	1.80
	50 A charge	2.0

Pulse Power Capability
Matched impedance power $P = V^2/4R = 16/4 \times 0.00155 = 2580$ W, 3280 W/kg
Power at specified pulse efficiency $P = \text{EF}\ (1-\text{EF})\ V^2/R$
95% $P = 0.95 \times 0.05 \times 16/0.00155 = 490$ W, 623 W/kg
90% $P = 0.9 \times 0.10 \times 16/0.00155 = 929$ W, 1180 W/kg
80% $P = 0.8 \times 0.20 \times 16/0.00155 = 1651$ W, 2099 W/kg

Equations 9–10 were used to calculate the power capability values given in Table 3 for the various batteries. As indicated in Table 9, the power capabilities (W/kg) projected for the various batteries vary greatly depending on the approach used to calculate the values. The matched impedance and the USABC max/min approaches clearly over estimate the power capability for most electric and hybrid vehicle applications. For hybrid applications in which a significant fraction of the energy is stored before it is used to power the vehicle, the efficient pulse EF = 95% is the appropriate value and for EVs, it seems reasonable to use the EF = 80% value.

4.2.2 Battery design trade-offs

The battery for an electric or hybrid vehicle must satisfy both an energy storage and power requirement for the vehicle to function properly. Unfortunately, batteries designed to optimize power having a low resistance will not be optimized for energy density because low resistance requires thin electrodes and large electrode area. These latter requirements also result in shorter cycle life and higher cost than for batteries optimized for high energy density. These design trade-offs are unavoidable and mean that complete knowledge of the battery design and performance are needed, including cycle life data, before the selection of the battery can be completed for a given vehicle application.

4.3 Battery pack design and testing

The electric driveline in most vehicles utilizes relatively high voltage −150 to 400 V in order to reduce the currents to reasonable levels. This means that the battery pack in the vehicle will consist of many cells in series. If large Ah cells are not available, the pack can also consist of a large number of cells in parallel or several strings of cells in series. In most cases, the cells are assembled into modules having a voltage many times the cell voltage. The module voltage can be in the range 12–75 V. Photographs of typical modules are shown in Figure 2. These modules would be tested using much the same procedures as outlined in the previous section for cells. The currents and powers for the testing would be based on the Ah and voltage ratings of the module. It is best to test the cells used in the module first because knowing the cell characteristics, one can predict the performance of the module with good confidence. If the measured characteristics (Ah, R, Wh, etc.) of the module do not follow directly from the cell characteristics either the module data are not correct or the connections of the cells in the module are faulty. It is also advisable to include interior

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 7. Test data for an EIG 15 Ah iron phosphate Lithium cell.

Iron Phosphate				
FO 15A	Weight 0.424 kg	3.65–2.0 V		
Power (W)	Power density (W/kg)	Time (s)	Discharged energy (Wh)	Energy density (Wh/kg)
62	142	2854	49.5	117
102	240	1694	48.0	113
202	476	803	45.1	106
302	712	519	43.5	103
401	945	374	41.7	98
Current (A)	Time (s)	Discharged Capacity (Ah)	Discharge Rate (*n*C)	Resistance (mΩ)
15	3776	15.7	0.95	—
30	1847	15.4	1.95	2.5
100	548	15.2	6.6	2.45
200	272	15.1	13.2	—
300	177	14.8	20.3	—

Table 8. Test data for an Altairnano 11 Ah cells.

Constant current test data (2.8–1.5 V)				
Current (A)	Discharge Rate (*n*C)	Time (s)	Discharge Capacity (Ah)	Resistance (mΩ)
10	0.8	4244	11.8	—
20	1.7	2133	11.9	—
50	4.5	806	11.2	2.2
100	9.2	393	10.9	2.1
150	15.3	235	9.8	—
200	—	116	6.4	—

Resistance based on 5 s pulse tests

Constant power test data (2.8–1.5 V)					
Power (W)	Power Density (W/kg)	Time (s)	Discharge Rate (*n*C)	Discharge Energy (Wh)	Energy Density (Wh/kg)
30	88	2904	1.2	24.2	71.2
50	147	1730	2.1	24.0	70.7
70	206	1243	2.9	24.2	71.0
100	294	853	4.2	23.7	69.7
150	441	521	6.9	21.7	63.8
170	500	457	7.9	21.6	63.5
260	764	255	14	18.4	54.2
340	1000	103	35.0	9.7	28.6

Mass: 0.34 kg.

Table 9. Comparisons of the power capabilities (W/kg) of various batteries using the different methods for the calculation of the power capability.

Device	Measurement Approach			
Lithium batteries 60% SOC	Matched impedance	USABC Minimum/ maximum	Efficient pulse EF = 95%	Efficient pulse EF = 80%
Kokam NCM 30Ah	2893	2502	550	1848
Enerdel HEV NCM 15 Ah	5491	4750	1044	3507
Enerdel EV NCM 15 Ah	2988	2584	568	1908
EIG NCM 20 Ah	2688	2325	511	1721
EIG Fe Phosphate 15 Ah	2141	2035	458	1540
Altairnano LiTiO 11 Ah	1841	1750	350	1180
Altairnano LiTiO 3.8 Ah	4613	4385	992	3341

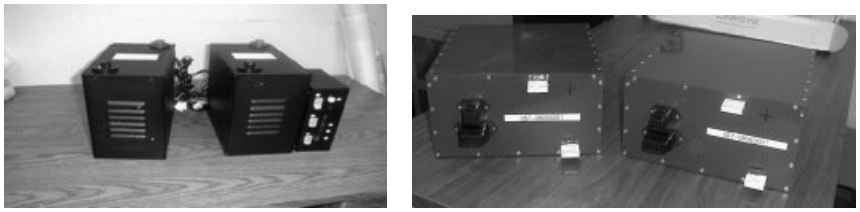
temperature measurements (Keyser and Smith, 2011) as part of the testing of modules. As discussed in a later section, thermal management of the modules and battery pack are critically important especially for lithium battery chemistries.

Before a battery pack is installed and tested in vehicles, it should be thoroughly tested in the laboratory. The test conditions for the pack should be determined from those to be experienced in the vehicle in which it will be installed. These conditions can be taken from simulation results of the vehicle operation. The battery pack should be tested using the same thermal management system (cooling), charger and charging algorithms, and battery management unit (BMU) as will be used when the battery pack is in the vehicle. The expected performance of the pack can be predicted with confidence from the test data for the cells and modules.

4.4 Battery charging

The proper charging of batteries (cells, modules, or pack) is dependent on information from the battery manufacturer.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) 70 V modules from EIG (b) 24 V, 50 Ah modules from Altairnano

Figure 2. (a,b) Photographs of lithium battery modules.

The key issue is the clamp (or maximum) voltage of the charge, which is dependent on the chemistry of the battery. The clamp voltage is close to the maximum voltage listed in Table 2 for each of the battery chemistries. In most cases, the battery is charged to the clamp voltage at a specified constant current and the current is then tapered to a specified fraction of the charge current. For example, a 20 Ah lithium iron phosphate cell could be charged at 20 A (1C rate) to 3.65 V and the current tapered to 1 A. For lithium batteries, regardless of the charging rate and battery age, the Ah put into the battery during charge (up to the clamp voltage and during taper) should be essentially equal to the Ah taken from the battery during the previous discharge. If this is not the case, either there is an error in the data or there are unwanted side reactions occurring in the cells. In either case, this is reason for concern.

When charging lithium modules or packs, it is advisable to track the voltages and temperatures of the individual cells and to maintain uniformity of the cells to a high degree. This should be accomplished by the BMU of the modules or pack if it is functioning properly. Part of the initial testing of the module or pack should be to establish that the BMS is functioning properly and to assess the magnitude of the variations between the cells. This will assure safe charging of the module or pack.

There is presently considerable interest in fast charging lithium batteries. By fast charging is meant replacing a significant fraction (at least 60–80%) of the Ah capacity of the battery in significantly <1 h. Laboratory testing of cells (Burke, 2009; Yvkoff, 2010) of lithium cells has indicated that with some cooling, charging rates up to 6C are possible for most of the lithium chemistries. The key issues are that the maximum voltage not be exceeded for any of the cells and that the interior temperature of the cells be limited to that specified by the battery manufacturer. Both of these conditions should be monitored by the BMS and the charge current reduced and/or charging terminated if one of the conditions is violated. Test results for the fast charging of the Altairnano Lithium Titanate cells are shown in Figures 3 and 4. The charging was done at the 6C rate with air cooling provided by a fan. Between the fast charges, the cells were discharged at the C/2 rate. As expected, the batteries heated

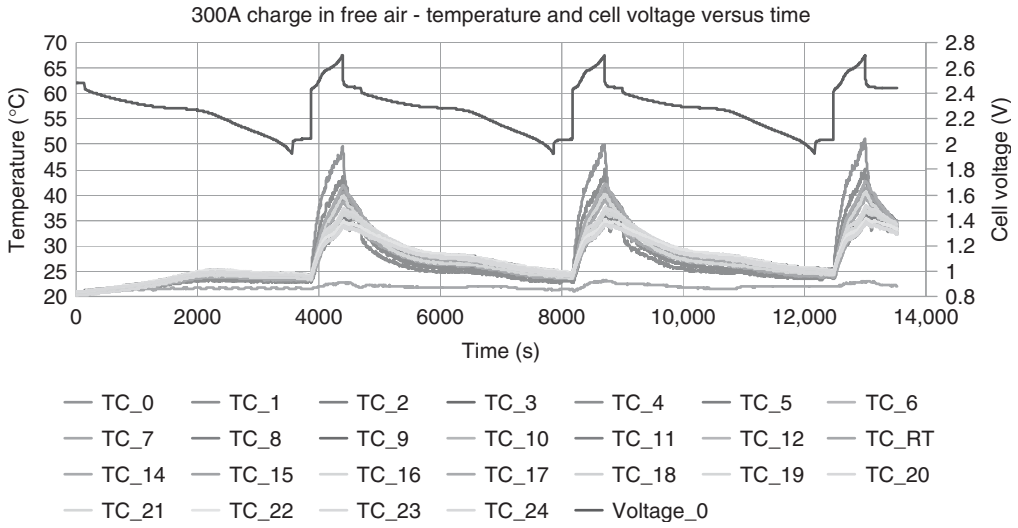


Figure 3. Fast charging (6C) of the 50 Ah Altairnano cell.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

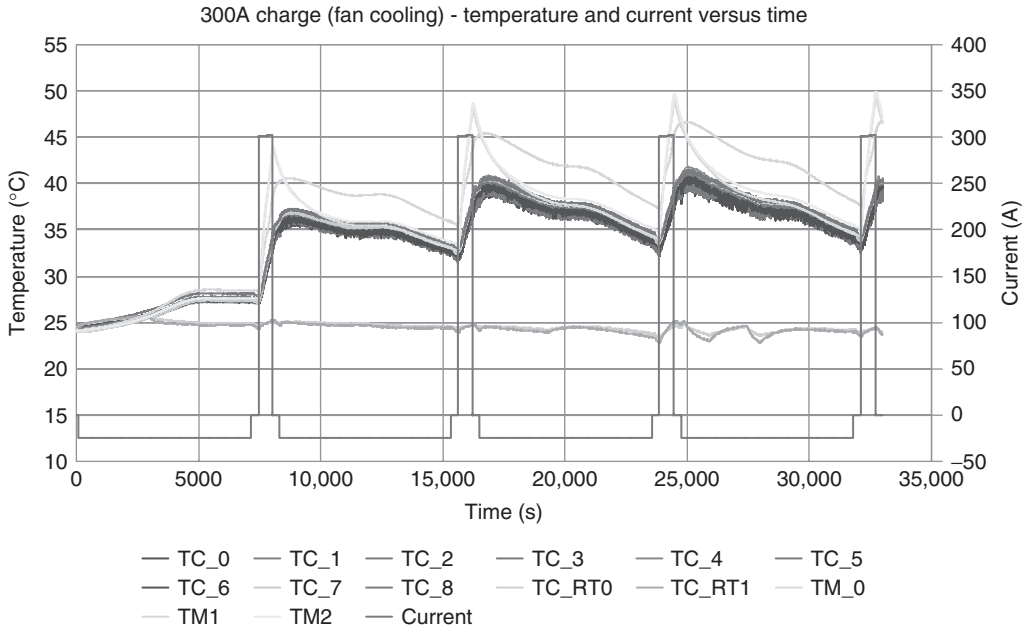


Figure 4. Fast charging (6C) of a 24 V module of the 50 Ah Altairnano cells.

up during the charging and cooled during the discharge reaching essentially a steady-state condition after about two cycles. For this cell, only small cooling is required to permit charging at 6C. The maximum temperatures measured on the exterior cell surface and the interior of the module were 45–50°C. In these fast charge tests, the batteries were completely charged and discharged.

4.5 Battery safety and management

There are safety concerns (Electrochemical Storage System Abuse Test Procedure Manual, 1999; Roth, 2004; Doughty, 2006; Orendorff, Roth, and Nagasubramanian, 2011; Keyser and Kim, 2011; Kim, Pesaran, and Smith, 2008) with all of the lithium battery chemistries. These concerns stem primarily with experience with cylindrical 18,650 cells in laptop computers. For this reason, lithium batteries have been used with caution in vehicle applications. The primary concerns are the possibility of thermal runaway because of an internal short in the battery caused by a manufacturing flaw (Orendorff, Roth, and Nagasubramanian, 2011; Keyser and Kim, 2011; Kim, Pesaran, and Smith, 2008) or abnormally high currents or temperature because of a failure of the battery control or a vehicle accident. These concerns have been addressed in two ways. First extensive abuse testing (Electrochemical Storage System Abuse Test Procedure Manual, 1999; Roth, 2004; Doughty, 2006) of lithium batteries has been done to show that thermal runaway does not occur even if the cells or modules are subject to acts of sudden compression, intrusion of sharp objects, dropping, fire, etc. Prevention of thermal runaway is dependent on both battery chemistry and design. The second approach to battery safety is to provide a BMU that monitors the cell/module voltages and temperatures and alerts the vehicle control system/computer if any of the cell voltages or temperatures is outside the normal range. Development of a BMU for use with their battery is common practice for battery manufacturers (Andrea, 2010; Pop *et al*., 2008). As noted earlier, validation of the functioning of the BMU should be done during testing of the battery modules and packs.

5 BATTERY LIFE

5.1 Factors affecting calendar and cycle life

In evaluating battery technologies for hybrid and EVs, it is important to consider battery life as well as battery performance. The usable life of a battery is a key issue in evaluating the economic viability of a particular battery technology. End-of-service life is defined as the time (calendar time or number of cycles) over which the energy capacity decreases by about 20% or the resistance increases by about 50%. Estimating cycle life for a particular vehicle design and application is not a simple matter (Battery

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

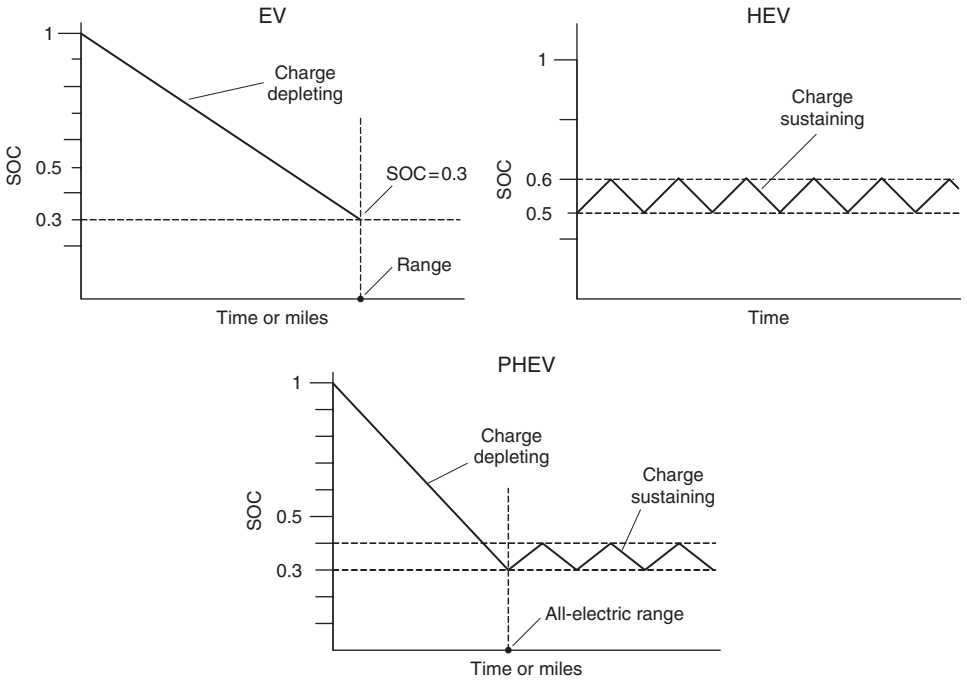


Figure 5. Battery state-of-charge history for EVs, HEVs, and PHEVs.

Technology Life Verification Test Manual, 2005). The cycle life depends on the rate of discharge, how the battery is recharged and the average depth of discharge before recharge, and the temperature of operation of the battery. The use of the battery in charge-sustaining hybrid electric vehicles (HEVs) and PHEVs and EVs is very different. As illustrated in Figure 5, the operating range of the battery's SOC in EV, HEV, and PHEV is different.

In the case of the hybrid vehicles, the battery experiences a very large number (300,000–500,000) shallow discharge cycles (about 5%) at an intermediate SOC (50–60%). The battery is charged and discharged at high rates requiring a high power design. It is never fully charged and the opportunities for cell equalization are limited. In the case of the plug-in vehicles, both plug-in hybrids and battery-powered EVs, the batteries experience deep discharge and long periods of charging from the wall plug. For plug-in hybrids, the energy stored in the battery is relatively small (<10 kWh in many vehicle designs) and the battery experiences its minimum SOC almost daily. In addition, the battery will experience long periods of shallow cycling near its minimum SOC when the plug-in hybrid is driven for long distances. Hence, life cycle testing of batteries for plug-in hybrid vehicles must include both deep and shallow cycling (Gaillac, 2008).

The batteries for EVs are sized to provide a relatively long vehicle range at least 75–100 miles in most cases and, as a result, will not be deep discharged nearly every day. The batteries are relatively large—store at least 25 kWh of energy—and experience high current pulses only during fast accelerations or quick decelerations (braking) of the vehicle. For most vehicle designs, the battery can be designed to maximize energy density and cycle life without power being a prime consideration. Except for fast charging, the EV batteries are expected to be charged in hours rather than minutes. For EVs, a battery cycle life of 1000–1500 deep discharge cycles is required, and for PHEVs, the cycle life is 2000–3000 cycles depending on the all-electric range of the vehicle. The calendar life of the batteries is intended to be at least 10 years. In order to reach this goal, the cell temperatures should be maintained below 50–55°C or even lower if specified by the battery manufacturer. This will require cooling of the battery pack in most cases.

In estimating the cycle life for a battery in a particular application, the depth of discharge before recharge and the fraction of the stored energy to be used are key factors. As shown in Figure 6, the cycle life increases rapidly if the usable depth of discharge of the cycles is <50%. Whether the battery is fully charged before each discharge cycle is also important. Both of these factors influence directly the usable energy and the effective energy density of the battery and the vehicle range for a given weight and cost of the battery.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

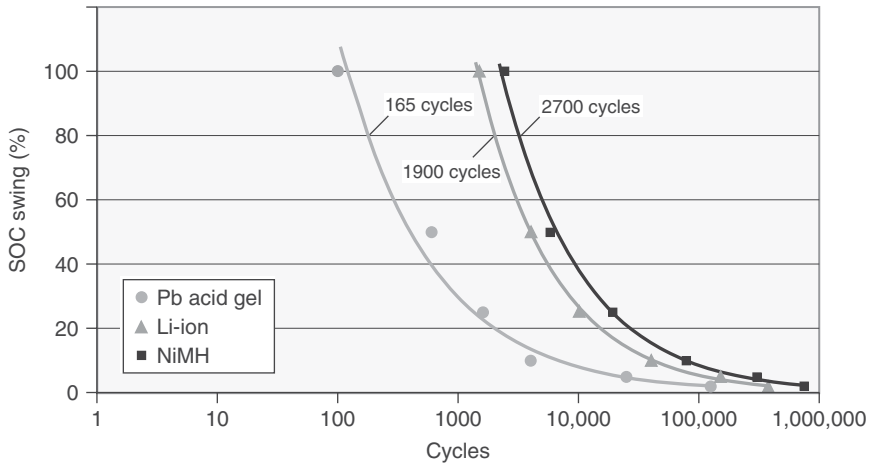


Figure 6. Cycle life correlations (Battery Technology Life Verification Test Manual, 2005) for various types of batteries.

5.2 Life cycle testing

Life cycle testing of batteries for vehicle applications is a long and complex process because of the diverse use and environmental conditions expected to be encountered by the battery during its lifetime. In addition, a vehicle battery pack consists of 100–200 cells in series with significant variability of the cells. These factors make both the planning of life cycle tests and analysis of data to estimate battery life with confidence difficult and expensive. As in the case of determining battery performance, life cycle testing is done first with cells and then with modules and packs. It is possible to do testing with large numbers of cells to get statistical data on cell degradation, but it is unlikely that equally large numbers of modules and packs can be tested. Another complication of life cycle testing is that one must select an appropriate test cycle (power vs time) and temperature environment for the testing.

Test procedures for testing batteries for HEVs, PHEVs, and EVs have been developed and published Electric vehicle battery test procedures manual, FreedomCAR battery test manual for power-assist hybrid electric vehicles, and Battery test manual for plug-in hybrid electric vehicles in the United States by the United States Advanced Battery Consortium (USABC) (1996, 2003, 2008). These procedures are intended to be followed by battery developers and National Laboratories and will yield a complete set of life cycle data. Study of the test procedures is informative, but little of the data from battery developers or National Laboratories will ever be published because it is proprietary to the battery developers. This is especially true for the advanced chemistry batteries—nickel metal hydride and lithium—of interest for vehicle applications.

In life cycle testing, the initial interest is in how the Ah capacity of the battery is degrading with calendar time and number of cycles and second how the resistance is increasing over the same period. The life of the battery is often defined as the number of cycles for which the Ah capacity decreases by 20% and/or the resistance increases by 50%. Often, the lithium cells are charged and discharged at the 1C rate. Test data (Shelburne, Manev, and Hanauer, 2009) for the Altairnano 11 Ah lithium titanate cell are shown in Figure 7. This confirms the expected long cycle life of the lithium titanate chemistry. Test data for the 20 Ah lithium manganese spinel chemistry are shown in Table 10. The data indicate that Ah capacity of two of the cells has decreased by 10% in 1100 cycles and the resistance of the cells has increased by 50–60%. Hence, the cell manufacturer has concluded the cell life of the cells is about 1000 cycles. It is not uncommon that the cell cycle life is determined by an increase in resistance and not the degradation in Ah capacity.

Life cycle testing of the Saft-Johnson Control lithium batteries used in the Daimler Chrysler Sprinter Van is reported in detail in the study by Gaillac (2008). The vehicle is a plug-in hybrid so the batteries were tested in both the charge-depleting mode (such as an EV) and the charge-sustaining mode (such as an HEV). The battery is discharged to a depth of discharge of 75% in the charge-depleting mode followed by testing in the charge-sustaining mode until the total cycle test time is 2.6 h corresponding to a vehicle range of 50 miles. The batteries are recharged at about the C/3 rate. The life cycle testing was done using three 21.6 V modules at room temperature (25°C) over a period of about 3 years. During that period, the test modules experienced 2100 cycles corresponding to 105,000 miles for the vehicle. Calibration testing of the modules at 240

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

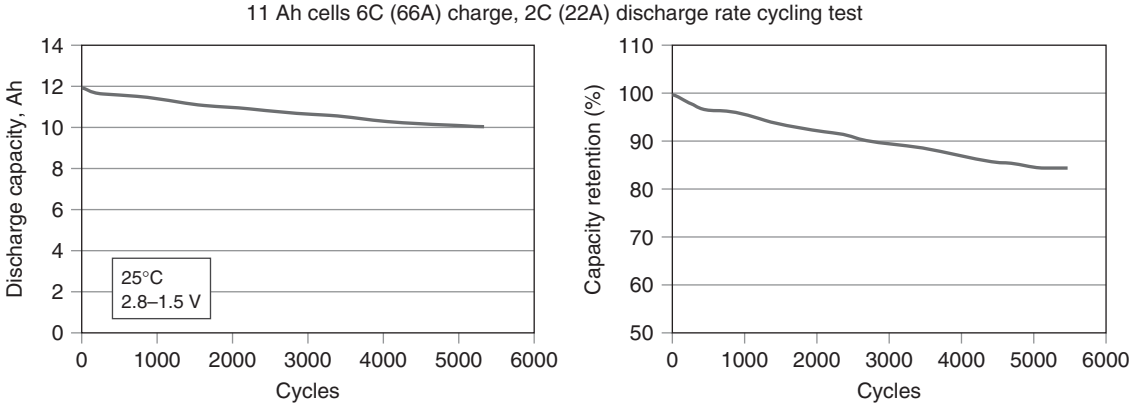


Figure 7. Life cycle data from Altairnano (Shelburne, Manev, and Hanauer, 2009) for the 11 Ah cell under fast charging (6C) conditions.

Table 10. Life characteristics of 20 Ah lithium manganese cells.

Cell Property[a]	Cell No. 1	Cell No. 2	Cell No. 3
Initial capacity (Ah)	19.72	19.57	19.74
Ah capacity at 1100 cycles (Ah)	17.83	18.8	18.6
% Capacity	90	91	94
Initial resistance (mΩ)	4.87	4.59	4.92
Resistance at 1100 cycles (mΩ)	7.87	6.94	7.56
% Increase of resistance	62	51	54

[a] Charge and discharge at 1°C at room temperature.

cycle intervals indicated that after the 2100 cycles, the C/1 and C/3 Ah capacities of the modules have decreased by 5% and 8%, respectively. The pulse power of the module has decreased by 7% over the same time period. During the testing, a BMS was used and the cell voltages were balanced if the BMS indicated that balancing was required after the charging and before the next discharge. These tests indicate that if the lithium batteries are maintained near room temperature and the cells are kept in good voltage balance, cycle life in excess of 2000 cycles can be expected with <10% degradation in capacity or pulse power for combined charge-depleting and charge-sustaining operation in a plug-in hybrid vehicle.

5.3 Battery cost considerations

The relative cost ($/kWh) of lithium-ion batteries of the different chemistries can be estimated using available computer programs. Projection of the cost of batteries requires inputs on the material costs as well as the costs of manufacturing equipment and processes. It is difficult to get good information on the costs of the various materials used in the electrodes of batteries. When such information is available, it is straightforward to estimate the differences in the electrode material costs for the different chemistries assuming ideal use of the materials in the electrodes. In terms of $/Wh, the following equation can be used:

$$(\$/\mathrm{Wh})_{\mathrm{materials}} = \{[((\$/\mathrm{gm}) + (\$/\mathrm{cm}^3)_{\mathrm{electrolyte}}/\rho/\varepsilon)/\mathrm{Ah/gm}]_{\mathrm{anode}} + [((\$/\mathrm{gm}) + (\$/\mathrm{cm}^3)_{\mathrm{electrolyte}}/\rho/\varepsilon)/\mathrm{Ah/gm}]_{\mathrm{cathode}}\}/V_{\mathrm{nom}} \quad (10)$$

The values for the Ah/gm and V_{oc} are given in Table 3. Calculated values for the electrode material costs ($/kWh) are shown in Table 11 for assumed unit costs of the various materials. The material unit costs used in the calculations are based on those used in a recent Argonne National Laboratory study (Nelson, 2008; Nelson, Santini, and Barnes, 2009). The results shown in Table 11 indicate that there is not a large difference in the electrode material costs of the various chemistries and also that electrode material costs should not dominate the total battery cost. Note that, in general, the higher cost lithium battery chemistries have the potential for longer cycle life which on a life cycle cost basis can compensate for the higher initial cost of those chemistries. This is especially true of the lithium titanate chemistry.

Researchers at Argonne National Laboratory (ANL) have developed a detailed lithium battery cost model (Nelson, 2008; Nelson, Santini, and Barnes, 2009) that is applicable to the various electrode chemistries. The model and results obtained at ANL are discussed in detail in the study by Nelson (2008); Nelson, Santini, and Barnes (2009); and Burke and Zhao (2010). Results obtained at UC Davis with the model for plug-in hybrid vehicle applications are summarized in Table 12 for the three electrode chemistries. The performance parameter values (Wh/kg and W/kg) for

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 11. Relative electrode material costs for various lithium battery chemistries.

Chemistry Anode/Cathode	Cell Voltage (V) maximum/nom.	Electrode Material ($/kg) Anode/Cathode[a]	Electrode Material Cost ($/kWh)	Cycle Life (Deep Cycle)
Graphite/$NiCoMnO_2$	4.2/3.6	19/19	44	2000–3000
Graphite/Mn spinel	4.0/3.6	19/8	35	1000
Graphite/$NiCoAlO_2$	4.2/3.6	19/19	44	2000–3000
Graphite/iron phosphate	3.65/3.25	19/16	47	>3000
Lithium titanate/Mn spinel	2.8/2.4	12/8	58	>5000

[a]The contribution of the electrolyte ($16/L) to the material costs was small partly because the porosity of the electrodes was only about 30%.

Table 12. Summary of battery performance and cost projections for various lithium battery chemistries using the Argonne National Laboratory cost model.

Battery	Energy Capacity (kWh)	Peak Power (kW)	Weight[a] (kg)	Energy Density (Wh/kg)	Power Density (W/kg)	Cost ($) Cell Mat.	Cost ($) Battery	Cost ($/kWh) Cell Mat.	Cost ($/kWh) Battery
NiCoAl									
Available	5.1	50	44	116	1136	716	1890	140	371
Energy	10.1	50	74	136	676	1156	2820	114	279
60%	20.2	76	143	141	531	2163	4143	107	205
LiFePhos.									
Available	4.8	50	47	102	1064	742	1943	155	405
Energy	9.4	50	80	118	625	1148	2838	122	302
65%	18.7	76	149	126	510	2132	4147	114	222
LiTitanate									
Available	3.6	50	55	65	909	668	1855	186	515
Energy	7.2	50	103	70	485	1196	2901	166	403
85%	14.4	76	201	72	378	2352	4458	163	310

[a]Unless noted otherwise, all values are the battery pack.

the battery pack given in the table are consistent with the cell test data presented previously in Table 4. The peak power in Table 12 corresponds to a pulse voltage of 80% of V_{oc}, which is an efficiency of 80%. The power densities (W/kg) for the modeled batteries are lower than those given in Table 8 because they are based on the pack weight not the weight of the cells.

Another aspect of the battery cost model that should be noted is that it accounts for the differences in the fraction of the stored energy expected to be available using the three chemistries—60% from NiCoAl, 65% from LiFe phosphate, and 85% from Li titanate oxide. This is the reason that the stored energy (kWh) is different for the three chemistries. This is also the reason that the battery costs for the different batteries are nearly the same even though the energy densities are quite different. Note also that on a $/kWh basis, the Lithium titanate batteries are significantly more expensive than the other two chemistries, but a significant part of the unit cost difference is negated by its higher energy use fraction. The battery costs ($/kWh) are sensitive to the unit material costs ($/kg), but it seems unlikely that the relative costs of the three chemistries will be much different than that shown in Table 12.

Vehicle cost and break-even gasoline price projections for mass marketed PHEVs and EVs in 2030 for mid-sized passenger cars are given in Burke and Zhao (2010); Ogden and Anderson (2011) for a range of battery costs. The results of the study are summarized in Table 13. Note that the lower end battery costs assumed in the study are consistent with the battery cost projections given in Table 12. The break-even gasoline price calculations were made assuming a 5-year payback period for the PHEV-20 and a 10-year period for the PHEV-40 and BEV-100. For the vehicles with the longer battery-only range, the cost calculations were done for the assumed life time of the battery. Note from Table 13 that the break-even gasoline prices are in a reasonable range when the battery costs approach those projected using the Argonne National Laboratory battery cost model. Hence, if the projected battery costs can be met in future years, it appears that both PHEVs and BEVs would have total ownership costs close to those of ICE vehicles in 2010. However, the results of the study discussed by Ogden and Anderson (2011) indicate

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 13. The effect of battery cost ($/kWh) on break-even gasoline price for PHEVs and EVs.

Vehicle Design	Battery Capacity (kWh)	Battery Cost[a] Table 12 ($/kWh)	Assumed Battery Cost[b] ($/kWh)	Differential Vehicle Cost[c] ($)	Breakeven Gasoline Cost ($/gal)
PHEV-20	5	370–400	800	6400	3.64[d]
	—	—	600	5600	3.19
	—	—	400	4800	2.73
PHEV-40	10	280–300	700	10,200	4.77[e]
	—	—	500	8200	3.83
	—	—	300	6200	2.89
BEV-100	28	200–220	700	20,300	8.09[e]
	—	—	500	14,700	6.04
	—	—	300	9095	3.99

[a] Battery cost to the auto manufacturer.
[b] Battery cost in the show room.
[c] Difference in show room vehicle cost compared to a conventional ICE car.
[d] 5 years and 4% discount rate, 12,000 miles/yr.
[e] 10 years and 6% discount rate, 12,000 miles/yr.

that the PHEVs and BEVs would have higher ownership costs than advanced ICEs and HEVs unless the price of gasoline was >$7–8/gal.

6 SUMMARY AND CONCLUSIONS

This chapter discusses the basic concepts related to batteries for automotive applications including HEVs and EVs. Emphasis is placed on lithium-ion batteries, but some consideration is also given to lead–acid and nickel metal hydride batteries. The performance—energy storage and power capability—of the batteries is considered including a discussion of installation in vehicles, battery charging, and BMS to mitigate safety issues with lithium batteries. The construction and operation of cells are discussed in detail as well as the laboratory testing of the cells and modules. Test data for lithium batteries of several chemistries are given from which it can be seen that the performance of lithium batteries is now suitable for HEV and EV applications. Test data indicate that fast charging (>4C) of lithium batteries, especially lithium titanate oxide, is feasible if the proper charging infrastructure is available.

Both calendar and cycle life and factors that effect life for the various chemistries are discussed. There is still considerable uncertainty concerning the cycle and calendar life of lithium batteries in vehicle applications, but the data shown indicate that long cycle life (at least 10 years) is a reasonable expectation. Cost estimates ($/kWh) are given for the various lithium battery chemistries based on a cost modeling program from Argonne National Laboratory. The projected future battery costs are in the range $300–$400/kWh. Using these cost estimates, it was found that the effective break-even gasoline prices for PHEVs and BEVs are $3–$4/gal when the battery cost is recovered over its 10-year life. Lithium battery safety is strongly dependent on the development of dependable BMU that monitor cell voltages and battery temperatures.

RELATED ARTICLES

General Requirement of Traction Motor Drives
Micro, Mild and Full Hybrid
Range Extender EV
Basic Consideration
Battery Charging Standards
Generators and Charging Control

REFERENCES

Andrea, D. (2010) *Battery Management Systems for Large Lithium-ion Battery Packs*, Artech House Publishers.

Battery Technology Life Verification Test Manual. (2005), FreedomCAR Report INEEL/EXT-04-01986, Idaho National Laboratory, February.

Burke, A.F. (2009) *Performance, charging, and second-use considerations for lithium batteries for plug-in electric vehicles*. ITS-Davis Report UCD-ITS-RR-09-17, July.

Burke, A.F. and Miller, M. (2009a) *Performance characteristics of lithium-ion batteries of various chemistries for plug-in hybrid vehicles*, EVS-24, Stavanger, Norway, May (paper on the CD of the meeting).

Burke, A.F. and Miller, M. (2009) *The UC Davis emerging lithium battery test project*. Report UCD-ITS-RR-09-18, July.

Burke, A.F. and Miller, M. (2011) The power capability of ultracapacitors and lithium batteries for electric and hybrid vehicle applications *Journal of Power Sources*, **196** (1, January), 514–522.

Burke, A.F. and Zhao, H. (2010) Projected fuel consumption characteristics of hybrid and fuel cell vehicles for 2015-2045. Paper presented at the Electric Vehicle Symposium 25, Shenzhen, China, November.

Doughty, D.H.. (2006) *LiIon battery abuse tolerance testing-an overview*. Presentation to the AQMD, July 12.

Gaillac, L. (2008) Accelerated testing of advanced battery technologies in PHEV applications *World Electric Vehicle Journal*, **2** (2)

Linden, D.B. and Reddy, T.B. (eds), Chapter 29, (2002) *Handbook of Batteries*, 3rd edn, McGraw-Hill.

Huggins, R.A. (2009) *Advanced Batteries-Material Science Aspects*, Springer.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Keyser, M. and Kim, G.H. (2011) *Numerical and experimental investigation of internal short circuits in a Li-ion cell*. Presentation at the 2011 U.S. DOE Hydrogen Program and Vehicle Technologies Program Annual Merit Review and Peer Evaluation Meeting, May 9-13, Arlington, Va.

Keyser, M. and Smith, K. (2011) Battery thermal modeling and testing at NREL. Presentation at the 2011 U.S. DOE Hydrogen Program and Vehicle Technologies Program Annual Merit Review and Peer Evaluation Meeting, May 9-13, Arlington, Va.

Kim, G.H., Pesaran, A., and Smith, K. (2008) *Thermal abuse modeling of Li-ion cells and propagation in modules*. Presentation at the 8th Advanced Automotive Battery Conference, Tampa, Florida, May 13-16.

Nazri, G.A. and Pistoia, G. (2003) *Lithium Batteries, Science and Technology*, Springer.

Nelson, P.A. (2008) Interim report on the cost study for plug-in hybrid vehicle batteries. Argonne National Laboratory report, April.

Nelson, P.A., Santini, D.J., and Barnes, J. (2009) Factors determining the manufacturing costs of lithium-ion batteries for PHEVs, EVS-24, Stavanger, Norway, May (paper on the CD of the meeting).

Orendorff, C.J., Roth, E.P., and Nagasubramanian, G. (2011) Experimental triggers for internal short circuits in lithium-ion cells *Journal of Power Sources*, **196**, 6554–6558.

Pop, V., Bergveld, H.J., Danilov, D., *et al*. (2008) *Battery Management Systems-Accurate State-of-charge Indication for Battery-powered Applications*, Springer.

Reddy, T.B. (2011) *Linden's Handbook of Batteries*, 4th edn, McGraw-Hill, USA.

Roth, P. (2004) Thermal abuse performance of high power 18650 lithium-ion cells *Journal of Power Sources*, **128** (2), 308–318.

Bennett, P.D. and Sakai, T. (1994) Hydrogen and Metal Hydride Batteries. *Proceedings Volume* 94-27, The Electrochemical Society, Inc., Pennington, NJ.

Shelburne, J., Manev, V., and Hanauer, B. (2009) *Large format Li-ion batteries for automotive and stationary applications*. 26th International Battery Seminar, March 2009, Fort Lauderdale, Florida (paper on the CD of the meeting).

Sion Power, *Lithium Sulfur Rechargeable Battery Data Sheet* (2008), October 3, www.sionpower.com (accessed 23 October 2013).

Ogden, J. and Anderson, L. (eds), Chapter 4 (2011) Comparing Fuel Economies and Costs of advanced vs. conventional vehicles in *Sustainable Transportation Energy Pathways*, Institute of Transportation Studies, University of California-Davis, August, Davis.

United States Advanced Battery Consortium (USABC) (1996), *Electric vehicle battery test procedures manual*, published (available on the USABC website).

United States Advanced Battery Consortium (USABC) (1999), Electrochemical Storage System Abuse Test Procedure Manual, Report SAND99-0497, published July, (available on the USABC website).

United States Advanced Battery Consortium (USABC) (2003), FreedomCAR battery test manual for power-assist hybrid electric vehicles, Report DOE/ID-11069, published October (available on the USABC website).

United States Advanced Battery Consortium (USABC) (2008), *Battery test manual for plug-in hybrid electric vehicles* Report INL/EXT-07-12536, published March (available on the USABC website).

Yoshio, M., Brodd, R.J., and Kozawa, A.K. (2009) *Lithium-ion Batteries-Science and Technologies*, Springer.

Yvkoff, L. (2010) *Will DC fast charging harm electric car batteries?*, web article, July 28.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto062
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Power and Energy Requirements for Electric and Hybrid Vehicles

Andrew Burke
University of California, Davis, CA, USA

1 Introduction 1
2 Vehicle Design and Performance Parameters 1
3 Vehicle/Powertrain Requirements 2
4 Vehicle Simulation Results 10
5 Summary and Conclusions 17
Symbols and parameters 17
Nomenclature 18
Related Articles 18
References 18

1 INTRODUCTION

This chapter is concerned with the assessment of the power and energy requirements for battery-powered electric vehicle (EV) and hybrid electric vehicle (HEV), including plug-in hybrid electric vehicles (PHEVs). Section 2 deals with the vehicle design parameters (weight, road load, and powertrain component powers) and their relationship to vehicle performance (range, speed, and acceleration). Section 3 is concerned with how the power and energy requirements are affected by the operating strategies of the vehicle powertrain especially in the case of plug-in hybrid vehicles. Section 4 includes vehicle simulation results for electric and hybrid vehicle designs that incorporate the design features discussed in the initial sections.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 VEHICLE DESIGN AND PERFORMANCE PARAMETERS

2.1 Design parameters

The key vehicle design parameters are the curb weight, the aerodynamic drag coefficient and frontal area, and the tire rolling resistance. When these parameters are known, the road load can be calculated and the powertrain power requirements for specified acceleration and top speed performance for the vehicle can be determined. The force Equation 1 for the vehicle operation is the following:

$$W_v \frac{dV}{dt} = F_{wheels,powertrain} - F_{aero} - F_{tires} \quad (1)$$

where $F_{aero} = {}^1/_2\ \rho C_d A_f\ V^2, F_{tires} = W_v\ f_{rolling}$

The driving force on the vehicle applied through the tires on the road is due to the torque of the electric motor and/or the engine on the drive shaft of the vehicle. The energy required to power the vehicle is stored onboard the vehicle in the fuel tank and/or in the battery. The force required at the wheels by the vehicle for any driving pattern can be reduced by decreasing the vehicle weight and road load parameters (C_D and $f_{rolling}$). The energy required to provide the force at the wheels depends on the efficiencies at which the electric motor and/or engine are operating. The primary objective of the hybrid vehicle operating strategy is to significantly increase the average efficiency of engine operation for typical vehicle driving patterns.

The weight and frontal area of the vehicle are highly dependent on the size and class of the vehicle and its utilization. Typical curb weights and frontal areas for

Table 1. Light-duty vehicle road load parameters.

Vehicle Class	Weight (kg)		Drag Coefficient		Rolling Resistance	
	2010	2030	2010	2030	2010	2030
Cars						
Compact	1270	1025	0.27	0.22	0.008	0.006
Mid-size	1479	1188	0.30	0.22	0.008	0.006
Full	1660	1330	0.30	0.25	0.008	0.006
SUVs						
Small	1614	1361	0.40	0.35	0.009	0.007
Mid-size	2050	1740	0.40	0.35	0.009	0.007

various classes of light-duty vehicles in 2010 are given in Table 1. The curb weight can be reduced by utilizing lightweight materials (Mallick, 2010; SAE International, 2004; Husain, 2005) in the construction of the vehicle. Weight reductions of 10–20% seem likely in the next 10–20 years (Assessment of Fuel Economy Technologies for Light-duty Vehicles, 2010; Schafer *et al*., 2009). The drag coefficients of most automobiles are about 0.3 in 2010 and could be decreased to 0.25 or lower in future years (Hucho, 1998). Little change in the frontal area is expected because it is set primarily by seating comfort of the passengers in the vehicle. The rolling resistance of tires (Tire and Passenger Vehicle fuel Economy, 2006; Barrand and Bokar, 2008) has decreased in recent years being about 0.8% in 2010. It may be further reduced to 0.6% if that can be done without sacrifice in tire traction on slippery roads. These road load parameters and the expected ranges of values in future are shown in Table 1.

Reducing the energy use of a vehicle is dependent on increasing the efficiency of the operation of the powertrain as well as the reducing its weight and road load. This aspect of the design and operation of the vehicle is discussed in later sections of this chapter.

2.2 Performance parameters

The design of a vehicle is highly dependent on the performance requirements set for the vehicle. In the case of a conventional internal combustion engine (ICE) vehicle, its performance is expressed in terms of the acceleration characteristics, top speed, and fuel consumption (L/100 km or mpg). In the case of a battery-powered EV, its performance is expressed in terms of its acceleration characteristics, top speed, range (miles), and electricity consumption (Wh/km). Two types of hybrid vehicles are of interest—a charge-sustaining HEV that utilizes only liquid fuel and a PHEV that utilizes both a liquid fuel and wall-plug electricity. In the case of the HEV, the performance parameters are the same as for a conventional ICE vehicle. In the case of the PHEV, the performance parameters are a combination of those of the EV and the HEV. In that case, the electric range (miles) and electricity consumption (Wh/km) using the battery in the EV mode and fuel consumption as an HEV are of interest as well as the acceleration characteristics in both the electric and HEV modes of operation. The performance parameters for ICE, EV, HEV, and PHEV are summarized in Table 2.

Table 2. Summary of the performance parameters for the various vehicle designs.

Performance Parameter	ICE Vehicle	Electric Vehicle (EV)	Hybrid Vehicle (HEV)	Plug-in Hybrid (PHEV)
Acceleration times				
0–48 kmph (s)	×	×	×	×
0–96 kmph (s)	×	×	×	×
Top speed (kmph)	×	×	×	×
Fuel consumption (L/100 km or mpg)	×		×	×
All-electric range (km)		×		×
Energy consumption from battery (Wh/km)		×		×

3 VEHICLE/POWERTRAIN REQUIREMENTS

3.1 Requirements for electric vehicles (EVs)

The simplest of the vehicles with an electric driveline is the battery-powered vehicle (EV). The driveline (Figure 1) of this vehicle consists of an electric motor, power electronics, and a large energy storage battery. All the energy to operate the vehicle is provided by the battery, which is recharged from the wall-plug. The range of the vehicle is determined by the energy storage capacity (kWh) of the battery and the vehicle's energy consumption (Wh/km). The acceleration and top speed characteristics are dependent on the power rating (kW) of the electric motor. As indicated

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

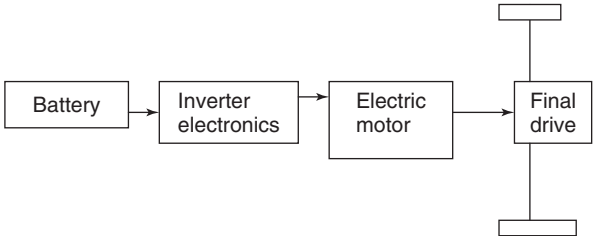


Figure 1. Battery electric vehicle driveline schematic.

in Table 3, battery electric vehicles have been designed and built in recent years by a number of auto companies. The size, functionality, electric motor power, acceleration, and range of these EVs vary greatly. Some of the vehicles are full-function vehicles intended for use on freeways at highway speeds, whereas others are intended for use on neighborhood streets at relatively low speeds.

The acceleration characteristics (acceleration times) are dependent primarily on the peak electric motor power. The acceleration times (simulation results) are given in Table 4 for a number of combinations of vehicle weight (kg) and motor power (kW). The low speed acceleration (0–48 kmph) is good for all the EVs, but the high speed acceleration (0–96 kmph) varies significantly with motor size. Good acceleration (0–96 kmph in <9 s) requires a power-to-weight ratio of 0.055–0.07 kW/kg.

Considerable care should be taken in the design of EVs to reduce their weight and road load compared to conventional ICE vehicles of the same size and type in order to attain maximum range for a reasonable size battery. The range of the vehicle is given by

$$\text{Range (km)} = (\text{kWh})_{\text{usable battery}} \times \frac{1000}{(\text{Wh/km})_{\text{veh}}}$$

Table 4. Relationship between electric motor power and acceleration characteristics for electric vehicles.

Vehicle Test Weight (kg)[a]	Peak Motor Power (kW)	Power-to-Weight Ratio (kW/kg)	Acceleration Time, 0–48 kmph (s)	Acceleration Time, 0–96 kmph (s)
1136	50	0.044	3.2	10.7
	60	0.053	2.9	9.1
	75	0.066	2.7	7.7
1386	75	0.054	2.9	8.8
	100	0.072	2.7	7.1
	125	0.090	2.7	6.3
1636	75	0.046	3.1	10
	100	0.061	2.8	7.9
	125	0.076	2.7	6.8

[a] For all the vehicles, $C_d = 0.25$, $A_f = 2.2\,m^2$, and $f_r = 0.007$.

where $(\text{Wh/km})_{\text{veh}}$ is the energy from the battery to operate the vehicle. The energy use of the vehicle depends on its weight, drag (C_DA), and rolling resistance coefficient (f_r) of the tires and the driving conditions (driving cycle). The range is dependent on the usable energy (kWh) from the battery—not the total energy stored. In most EVs, the fraction of the stored energy that can be used is only 60–80%. This is done to prolong the life of the battery.

Some of the trade-offs in energy use for light-duty vehicles are shown in Table 5 based on simulation results for the Federal City and Highway driving cycles. The battery storage requirement is given by

$$(\text{kWh})_{\text{bat}} = (\text{Range}) \times \frac{(\text{Wh/km})_{\text{battery}}}{(1000 \times \text{fr}_{\text{bat}})}$$

where fr_{bat} is the usable fraction of battery-stored energy.

Table 3. Characteristics of selected battery electric vehicles.

Model/Manufacturer	Type	Curb Weight (kg)	Length/Width/Height (cm)	Battery Type/Capacity (kWh)	Electric Motor (kW)	Range (km)	Max. Speed (kmph)
Leaf/Nissan	Full	1530	444/177/155	Lithium-ion/24	80	140	145
MiEV	Full	1080	340/147/160	Lithium-ion/16	47	120	130
EV1/GM	Full	1350	432/178/130	NiMthyd/29	104	224	>120
EV Plus/Honda	Full	1634	405/174/162	NimtHyd/30	49	160	>120
EV Fit/Honda	Full	1480	411/172/158	Lithium-ion/20	92	150	>120
RAV4/Toyota	Full	1560	398/169/167	NiMtHyd/28	50	150	>120
Altra/Nissan	Full	2080	487/177/169	Lithium-ion/32	62	192	>100
Smart/Mercedes	Full	1380	357/172/160	NaAlCl/30	50	200	>120
Think/Ford	NEV	960	300/160/155	NiCad/12	12	80	40
EV Focus/Ford	Full	1680	435/206/147	Lithium-ion/ 23	107	140	>120
E-com/Toyota	CEV	790	279/148/160	NiMtHyd/8	20	80	80
Hypermini/Nissan	CEV	840	266/148/155	Lithium-ion/10.5	24	96	80
Zenn/Feel good cars	NEV	510	258/138/139	VRLA lead-acid/7	8	56	40

Full-all roads and speeds, NEV neighborhood EV, and CEV City EV.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 5. Performance characteristics of battery-powered electric vehicles of various weight, drag, and tire rolling resistance.

Test Weight[a] (kg)	Drag Coefficient (C_d)[b]	Rolling Resistance Coefficient (f_r)	City Electricity Consumption[c] (Wh/km)	City Range (km)	Highway Electricity Consumption (Wh/mi)	Highway Range (km)	Acceleration 0–96 kmph (s)
1200	0.25	0.008	128	125	122	131	8.0
	0.30	0.008	130	122	131	123	8.0
	0.20	0.008	124	130	113	142	7.9
	0.25	0.006	121	132	115	139	7.9
	0.30	0.006	125	128	124	130	8.0
	0.20	0.006	118	136	106	152	7.9
1500	0.25	0.008	138	177	130	123	9.4
	0.30	0.008	142	112	139	115	9.5
	0.20	0.008	133	118	121	133	9.4
	0.25	0.006	130	123	121	133	9.3
	0.30	0.006	138	120	130	123	9.4
	0.20	0.006	125	126	111	142	9.3
1800	0.25	0.008	150	107	138	117	11.1
	0.30	0.008	154	104	147	109	11.2
	0.20	0.008	145	110	129	125	11.0
	0.25	0.006	140	114	128	125	11.0
	0.30	0.006	144	112	136	44	11.1
	0.20	0.006	136	117	118	136	10.9

[a] All vehicles used a 75 kW AC induction motor, regenerative braking.
[b] The frontal area is 2 m^2.
[c] All vehicles used lithium-ion batteries, 150 Wh/kg, 20 kWh, 80% usable.

The corresponding battery weight is given by

$$(\text{Weight})_{\text{battery}} = (\text{kWh})_{\text{bat}} \times \frac{1000}{(\text{Wh/kg})_{\text{battery}}}$$

For light-duty vehicles, usable energy storage of 25–30 kWh is needed to attain a range of 150 miles depending on vehicle size and weight. This appears to be practical only for high energy density batteries such as lithium-ion or other chemistries having a usable energy density greater than 100 Wh/kg. Ranges in excess of 200 miles will require usable energy densities greater than 150 Wh/kg. The United States Advanced Battery Consortium (USABC) has set a minimum goal of 150 Wh/kg for long-term commercialization of EVs and a longer term goal of 200 Wh/kg (USABC) Goals for Advanced Batteries for EVs.

The power and energy requirements for various sizes of EVs (cars and sport utility vehicles (SUVs)) are given in Table 6 for vehicles having a range of 100 miles. The

Table 6. Characteristics of battery-powered electric vehicles (EV) of various types.

Vehicle Type	Vehicle Test Weight (kg)	Battery Weight[a] (kg)	Battery kWh Stored[b] (kWh)	Electric Motor[c] (kW)	Required Battery Pulse Power[d] (W/kg)	Wh/km From Battery[e] (Wh/kg)	0–96 kmph (s)
Cars							
Compact	1373	168	20.2	65	387	126	11.3
Mid-size	1695	208	24.9	102	490	157	8.9
Full	1949	238	28.5	122	513	178	8.6
SUV							
Small	2103	266	31.9	128	481	199	9.6
Mid-size	2243	278	33.3	143	514	208	9.3
Full	2701	317	38.0	160	501	238	9.6

[a] Lithium-ion battery with an energy density of 120 Wh/kg.
[b] All vehicles have a range of 160 km.
[c] Peak motor power.
[d] Peak pulsed power required from the battery at 90% efficiency of the electric motor and electronics.
[e] Average energy consumption on the FUDS and FHWAY drive cycles.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

power and energy requirements are strongly dependent on the vehicle type and size. The power required from the battery is the electric motor power divided by the efficiency of the motor/electronics. It will be assumed in this chapter that the electric drive efficiency is 90%.

3.2 Requirements for hybrid electric vehicles (HEVs)

There are several design approaches for HEVs. In all cases, the vehicle has the capability to generate electricity onboard the vehicle from liquid or gaseous fuel and an electric motor provides at least part of the torque to propel the vehicle. The electricity can be generated using either an engine/generator or a fuel cell. The vehicle is not designed to be plugged into the wall.

The engine can be connected directly to the wheels or only to the generator. The latter arrangement is referred to as a *series hybrid* and the former as a *parallel hybrid*. These powertrain arrangements are illustrated in Figure 2. The battery is charged off the engine or fuel cell and maintained in a relatively narrow range of state of charge (SOC). This hybrid vehicle is referred to as a *charge-sustaining hybrid* and all the energy to operate the vehicle is provided by the liquid or gaseous fuel. Refueling such a hybrid vehicle is not much different from a conventional ICE vehicle. The intent of the HEV design is to improve the efficiency of the engine operation and thus the fuel economy of the vehicle. All the HEVs offered for sale by the auto companies before 2011 are of the charge-sustaining type. As with EVs, the energy consumption of the HEV can be reduced by decreasing the vehicle weight, aerodynamic drag, and tire rolling resistance.

3.2.1 Parallel hybrid drivelines

The powertrain characteristics of selected hybrids are shown in Table 7. From the driver's point of view, these vehicles operate essentially the same as a conventional ICE vehicle but get much higher fuel economy. Markets for HEVs are growing as more models are becoming available from more auto companies (Hybrid vehicle sales data (current)). The power requirements and operating strategies of parallel and series hybrids are discussed in the following sections.

3.2.2 Parallel hybrids

A schematic for a single-shaft parallel hybrid powertrain is shown in Figure 2a. The key feature of the parallel configuration is that the engine is connected directly to the wheels of the vehicle and the total torque and power to the wheels is the sum of the outputs of the electric motor and the engine. Figure 2b shows the more complex planetary arrangement in which there are two electric machines—one primarily used as a traction motor to power the vehicle and the other as a generator to utilize part of the engine

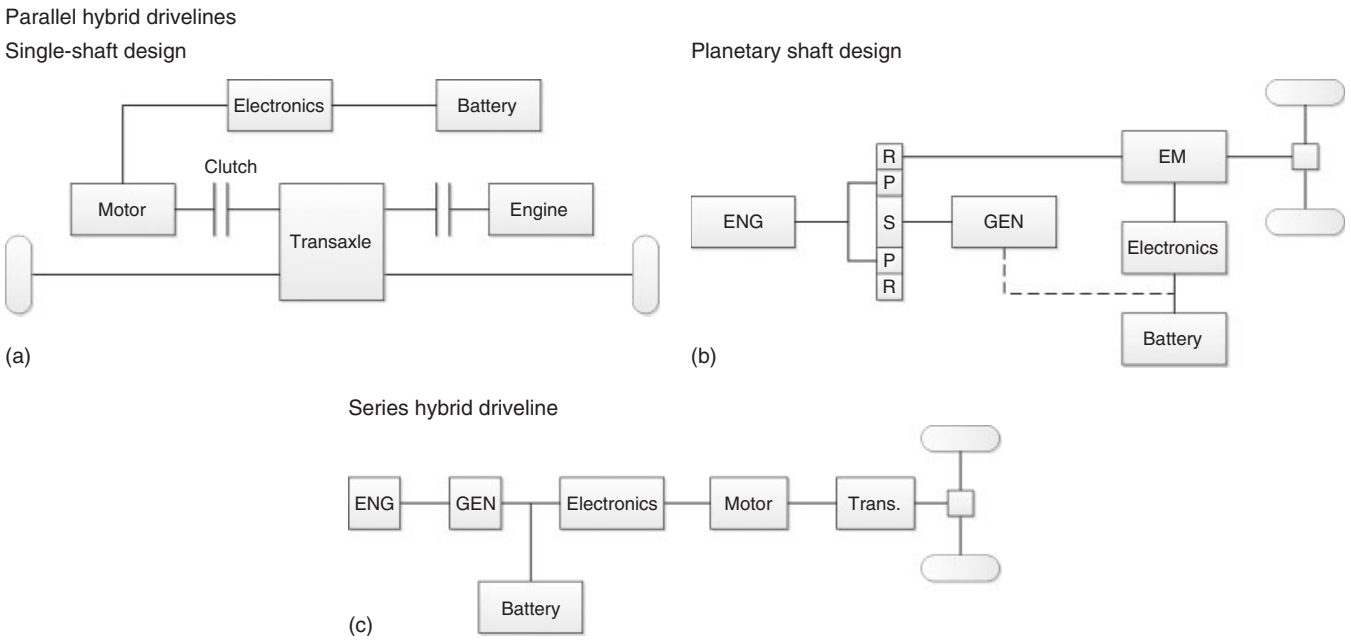


Figure 2. (a–c) Driveline schematics for parallel and series hybrid vehicle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 7. Fuel consumption and emissions of the Toyota and Honda Hybrid Cars (2006).

Vehicle	Transmission/Year	Electric Motor (kW)	0–60 mph Acceleration (s)	Emissions	Unadjusted L/100 km;mpg (City)	Unadjusted L/100 km;mpg (Hwy)
Honda Insight	M5	10	11.2	ULEV	3.55; 67	2.74; 87
	CVT	10	—	SULEV	3.78; 63	3.31; 72
Honda Civic (2002)	M5	10	—	ULEV	4.67; 51	3.66; 65
	CVT	10	12.0	SULEV	4.41; 54 56%[a]	3.90: 61 22%[a]
Honda Civic (2004)	CVT	15	10	SULEV	4.33; 55 62%[a]	3.84; 62 29%[a]
Toyota Prius	Planetary/2000	33	12.6	SULEV	4.18; 57 58%[a]	4.10; 58 18%[a]
	Planetary/2004	50	10.1	SULEV	3.55; 67 86%[a]	4.41; 64 26%[a]
Toyota Camry	Planetary/2006	105	9.0	SULEV	5.05; 47 81%[a]	5.06; 47 15%[a]

[a] % improvement in fuel economy for the HEV compared to an ICE car L/100 km = 238/mpg.

output to generate electricity onboard the vehicle. This is the arrangement used by Toyota in the Prius. It has some of the characteristics of a series hybrid as it is possible to recharge the batteries even when the traction motor and engine are being used to provide torque to the wheels. In this more complex arrangement, the combination of the engine, traction motor, and generator act as a continuously variable, electromechanical transmission whose effective gear ratio depends on the ratios of the powers of the three components. The traction motor is also used as a generator to recover energy during braking of the vehicle.

In the simple parallel arrangement (Figure 2a), the engine output can be used to power the vehicle and to generate electricity using the traction motor as a generator to recharge the battery. A mechanical transmission is required for efficient operation of the engine even with the electric motor. The optimum choice of transmission for this single-shaft parallel hybrid is a continuously variable, mechanical unit (Frank, 2004; Bradley, Huff, and Frank, 2005) consisting of a variable geometry steel belt or linked chain operating between two adjustable pulleys (Figure 3). The continuously variable transmission and utilization of part of the engine output to recharge the battery permits the engine to operate most of the time at high efficiency resulting in a large improvement in the fuel consumption of the vehicle. The continuously variable transmission also results in regenerative braking energy recovery down to near zero vehicle speed. Both parallel arrangements (Figure 2b and 3) utilize in essence a continuously variable transmission. It is likely that the system efficiency gains from the planetary arrangement will be slightly greater than with the single-shaft CVT arrangement (Burke and Van Gelder, 2008), but it also seems likely that the cost of the planetary arrangement will be higher.

One of the key issues in designing a parallel hybrid is the size (kW) of the engine and the resultant peak power requirement of the electric motor(s). One approach is to minimize the power of the electric motor and to downsize

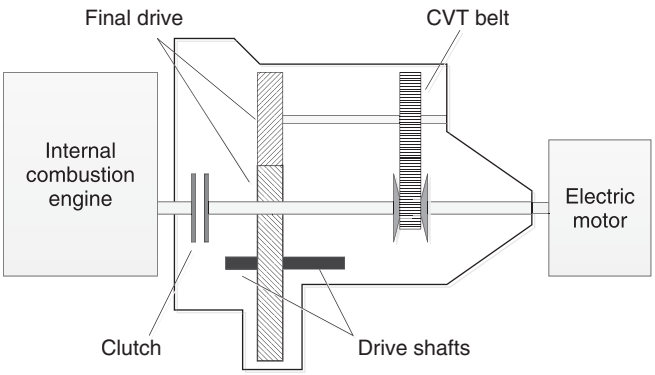


Figure 3. Schematic of a CVT in a hybrid driveline.

the engine only slightly. With this approach (Burke, 2009; Burke, Zhao, and Van Gelder, 2009), the power of the electric motor is selected such that the vehicle can be operated in the electric mode whenever operation of the engine would be inefficient. In this case, the peak power of the electric motor is only 15–20% of that of the engine and the vehicle is termed a *mild hybrid*. This approach also results in a relatively small, low cost energy storage unit that could utilize either batteries or ultracapacitors (Burke, Zhao, and Van Gelder, 2009). These single-shaft designs have been marketed by Honda (Table 5) and testing of the vehicles has shown that city cycle fuel economy improvements of 30–40% can be achieved using this approach.

Parallel hybrids can also be designed in which the engine is significantly downsized and a relatively large electric motor is used to augment the power of the engine. In this approach, which is termed a *full hybrid*, the electric motor and engine are of comparable size (kW) and the energy storage unit is significantly larger than for the "mild hybrid" in terms of both kilowatt-hours and power rating (kW). This approach is used by Toyota in the Prius (Table 5). The fuel economy improvement achieved with the "full

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 8. Characteristics of batteries for HEV vehicles.

Battery Developer/ Cell Type	Electrode Chemistry	Voltage Range (V)	Capacity (Ah)	Resist. (mΩ)	Energy Density (Wh/kg)	Power Density 90% Efficiency[a] (W/kg)	Power Density Match Impedance (W/kg)	Weight (kg)	Density (gm/cm^3)
Enerdel HEV	Graphite/Ni MnO_2	4.1–2.5	15	1.4	115	2010	6420	0.445	—
Kokam prismatic	Graphite/$NiCoMnO_2$	4.1–3.2	30	1.5	140	1220	3388	0.787	2.4
Saft Cylinder	Graphite/NiCoAl	4.0–2.5	6.5	3.2	63	1225	3571	0.35	2.1
A123 Cylinder	Graphite/iron phosphate	3.6–2.0	2.2	12	90	1393	3857	0.07	2.2
Altairnano prismatic	LiTiO/$NiMnO_2$	2.8–1.5	11	2.2	70	990	2620	0.34	1.83
Altairnano prismatic	LiTiO/ $NiMnO_2$	2.8–1.5	3.8	1.15	35	2460	6555	0.26	1.91
EIG prismatic	Graphite/$NiCoMnO_2$	4.2–3.0	20	3.1	165	1278	3147	0.41	—
EIG prismatic	Graphite/iron phosphate	3.65–2.0	15	2.5	113	1100	3085	0.42	—
Panasonic EV prismatic	Ni metal hydride	7.2–5.4	6.5	11.4	46	395	1093	1.04	1.8

[a] Power density $P = \text{Eff}(1-\text{Eff.})\ V_{oc}{}^2/R$, $P_{\text{match. imped.}} = V^2/4R$.

hybrid" approach will be somewhat larger than with the "mild hybrid" approach, but the incremental vehicle cost will also be significantly higher.

The battery used in the HEV whether in the single-shaft or in the planetary shaft designs is relatively small storing 0.75–1.2 kWh of energy. However, the power requirements for the battery depend on the size (kW) of the electric motor being used in the powertrain. Hence, the power requirement for a battery in the "full" hybrid (40–60 kW) is significantly greater than for the "mild" hybrid (10–20 kW). In both cases, the battery is maintained at an intermediate SOC (50–60%) in order that it can provide the high power in both discharge and charge. In addition, maintaining the battery in a narrow range of SOC results in long battery life of several hundred thousand shallow charge/discharges. The batteries used in HEVs must be high power batteries with power densities of 1000–1500 W/kg. The characteristics of a number of batteries (Burke and Miller, 2009; Burke and Miller, 2011) suitable for use in HEVs are given in Table 8.

3.2.3 Series hybrids

As shown in Figure 2c, the series hybrid is essentially an electric, battery-powered vehicle with an engine/generator onboard to generate electricity as the vehicle is driven. When the power output capacity of the engine/generator is equal to or greater than the average power required for a particular use (speed and grade) of the vehicle, the range of the vehicle for that use is set by the size of the fuel tank and not the battery. In most cases, the engine/generator rating is selected for range extension on some specified driving cycle. The advantage of the series hybrid is that the engine can be controlled to operate near its maximum efficiency independent of vehicle speed and power demand. The series approach is most suited for applications in which the peak electrical power demand is high compared to the average power demand and periods of high power demand are relatively short so that they can be met using energy from the battery. Otherwise, the power rating of the engine, generator, and electric motor have to be nearly the same, which results in the powertrain being heavy, large, and high cost.

The design of a series hybrid starts with the design of a battery-powered vehicle with a battery sized to a relatively short range. As for an EV, the electric motor is sized (kW) so that the vehicle has a specified acceleration performance (0–60 mph time). The engine/generator is sized to meet a specified maximum electrical output (kW), which is often set by maintaining the vehicle at a specified constant speed on a grade. The battery energy storage capacity (kWh) is determined from the range (km) requirement and the expected energy consumption (kWh/km) of the vehicle.

The series hybrid can be operated either as a charge-depleting hybrid such as an EV with the battery being charged from the wall-plug or as a charge-sustaining (HEV) hybrid using only a liquid fuel. In the latter case, the primary objective of the hybridization is to achieve a large improvement in fuel consumption. In the former case, the primary objective is range extension of an EV and the substitution of electrical energy for the liquid fuel. A combination of the two operating modes yields a PHEV like the GM volt.

3.3 Requirements for plug-in hybrid electric vehicles (PHEVs)

"Full hybrids" can be designed as either charge-sustaining (HEV) or PHEVs. All hybrids marketed by the auto companies before 2011 were HEVs, but some Toyota

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Prius were converted to PHEVs by several small vehicle engineering companies (Duoba *et al*., 2009; Ghorbani, Bibeau, and Filizadeh, 2010). In order to convert an HEV to a PHEV, it is necessary to increase the onboard energy storage from 1–1.5 to 5–10 kWh. In addition, the system control software must be altered to permit the batteries to be depleted as the vehicle is operated as an EV over some range of speed and power demand. Provision must also be made to recharge the battery from the wall-plug. Several hundred Prius have been converted to PHEVs with good success in that their fuel economy on gasoline have been increased to about 100 mpg for city driving. The cost of the conversions was high being $10,000–$15,000 in 2007–2009.

In 2011, General Motors began marketing the *Volt*, which is a plug-in hybrid. The Volt has a 110 kW electric motor and a lithium-ion battery that stores 16 kWh. It operates as an EV before the battery is depleted and as series hybrid after the battery is depleted. The range of the ***Volt*** on battery-stored energy operating as an EV is about 64 km. The electrical power capability of the engine/generator on the Volt is 54 kW. Hence, the *Volt* is essentially an EV with a range of 64 km, but it has extended range on gasoline comparable to a conventional ICE vehicle for most driving conditions.

Most of the auto companies that have marketed HEVs are in the process of developing plug-in hybrids based on their parallel HEV designs. The main change from the HEV design is to increase the size (kWh) of the battery. The selection of the battery for plug-in vehicles can be a complicated process depending on several factors. In simplest terms, the battery should meet the energy storage (kWh) and peak power (kW) requirements of the vehicle in the all-electric mode of operation. In addition, the battery must satisfy cycle life requirements for both deep discharge in the charge-depleting all-electric mode and shallow cycling in the charge-sustaining mode of operation. The final considerations are concerned with the initial and life cycle costs of the battery.

The battery size and cost will vary markedly depending on the all-electric range (AER) of the vehicle and whether all-electric operation means that the engine is not used under any circumstances when the battery's SOC is greater than a minimum specified value. This control strategy would limit the vehicle acceleration that is possible using only the electric motor even though greater acceleration performance would be possible by turning on the engine. Another control strategy would minimize engine use when the battery SOC is high, but not forbid it, and only recharge the battery from regenerative braking until its SOC has reached the minimum specified value. In this approach, high power demand is met by a combination (blending) of the engine and electric motor torques. This approach results in a near maximum substitution of electricity for petroleum and at the same time offers the driver the maximum vehicle acceleration performance. The blended approach would also lead to a lower cost hybrid powertrain because the power of the electric driveline (motor and batteries) could be smaller than would be the case if AER meant no engine operation under any circumstances. After the batteries are discharged to the minimum SOC, the control strategy would be similar to that for a charge-sustaining hybrid with the intent of maximizing the engine operating efficiency and vehicle fuel economy.

In the case of PHEVs, there is much design flexibility in selecting the battery size and the electric motor and engine powers because the AER is a design variable and the power demand of the vehicle can be met by a combination (blending) of motor and engine output even while the battery is being depleted. Typical design combinations for AERs between 10 and 40 miles are shown in Table 9 for a mid-sized passenger car. The increasing weight and decreasing power density requirement of the battery with increasing all-electric (battery depletion) range of the vehicle is typical for plug-in hybrid designs. The battery in

Table 9. Battery sizing and power density for plug-in hybrid vehicles for various all-electric range and electric motor power (mid-sized passenger car).

Range (miles)	Electric Motor (kW)	Engine Power (kW)	Battery Energy Needed[a] (kWh)	Battery Energy Stored[b] (kWh)	Battery Weight[c] (kg)	Battery Power Density[d] (kW/kg)
10	50	100	2.52	3.6	30	1.84
15	55	100	3.78	5.4	45	1.36
20	60	75	5.04	7.2	60	1.11
30	75	60	7.56	10.8	90	0.92
40	100	50	10.1	14.4	120	0.92

[a] Vehicle energy usage from the battery: 156 Wh/km.
[b] Usable state-of-charge for batteries: 70%, weights shown are for cells only.
[c] Battery energy density 120 Wh/kg.
[d] Electric driveline efficiency 90%.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a plug-in hybrid vehicle with a short AER (<50 km) will experience a deep discharge cycle almost every day and hence must be designed for more deep discharge cycles than the battery in a vehicle with a longer AER. It is clear from Table 9 that the requirements for batteries used in vehicles with short AER are more demanding than those in other hybrid vehicles. This will result in those batteries being more expensive on a $/kWh basis than batteries in vehicles with longer AER.

3.4 Summary of battery requirements for various electric drive vehicles

On the basis of the discussions of the previous sections, the battery requirements for mid-sized passenger cars and mid-sized SUVs are summarized in Table 10. It was assumed that lithium-ion batteries were used in all the vehicles but that the batteries were optimized for power and energy appropriate for the application. For example, the batteries used in the HEVs were optimized for power and had lower energy density than the batteries used in the EVs and PHEVs. It was also assumed that the PHEVs operated in a blended engine/electric mode for high power demands.

All the battery packs shown in Table 10 can be assembled using cell technology available in 2010 if the battery container and thermal management system designs yield packaging factors of 0.6–0.7.

Further development of lithium battery technology and technologies beyond lithium can be expected. Some of that development will be required to meet the battery pack goals set by the USABC and DOE for EVs and PHEVs (Snyder, 2012). These goals are summarized in Tables 11 and 12.

Table 10. Battery requirements for electric drive vehicles.

Vehicle Type	Vehicle Test Weight (kg)	Battery Pack Weight[a] (kg)	Battery Energy Stored[b] (kWh)	Electric Motor[c] (kW)	Required Battery Pulse Power[d] (W/kg)
Mid-sized cars					
EVs	1695	240	25	102	475
HEVs	1400	25	1	30	1330
PHEVs					
30 mi	1550	95	8	60	725
Mid-size SUV					
EVs	2350	335	35	145	480
HEVs	2050	40	2	60	1650
PHEVs					
30 mi	2175	140	12	110	875

[a] Lithium-ion cells with an energy density of 150 Wh/kg.
[b] All electric vehicles have a range of 160 km.
[c] Peak motor power.
[d] Peak pulsed power required from the battery at 90% efficiency of the electric motor and electronics.

Table 11. USABC goals for batteries in EV applications.[a]

Parameters	Unit	EV Commercialization Goals	EV Long-Term Goals
Discharge specific power at 80% DoD for 30 s	W/kg	300	400
Regenerative specific power at 20% DoD for 10 s	W/kg	150	200
Power density	W/L	460	600
Onboard energy capacity	kWh	40[b]	40[b]
Specific energy at C/3 discharge rate	Wh/k[a]	150	200
Energy density at C/3 discharge rate	Wh/L[a]	230	300
Calendar life	yr	10	10
Cycle life to 80% DoD	Cycle	NA	1000
Operating temperature	°C	−40 to +50	−40 to +85
Selling price	USD/kWh	<150	<100
Normal recharge time	h	6	3–6
High recharge rate	h	0.5 (20–70% SOC)	0.25 (40–80% SOC)

NA, not applicable.
[a] The goals are for the battery pack including container and all management systems.
[b] Vehicle range of 240–320 km (approximately 140 Wh/km).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 12. USABC goals for batteries in PHEV applications.

Parameters	Unit	High P/E Ratio	High E/P Ratio
Reference equivalent electric range	Mile	10	40
Peak pulse discharge power for 2 s	kW	50	46
Peak regenerative power for 10 s	kW	30	25
Available energy for charge-depleting mode at 10 kW discharge rate	kWh	3.40	11.60
Available energy for charge-sustaining mode	kWh	0.50	0.30
Cold cranking power at −30°C	kW	7	7
Calendar life at 35°C	yr	15	15
Maximum system weight	kg	60	120
Maximum system volume	L	40	80
Specific energy at C/3 discharge rate	Wh/kg	56	97
Energy density at C/3 discharge rate	Wh/L[a]	85	145
Specific power	W/kg[a]	833	383
Maximum operating voltage	V	400	400
Operating temperature	°C	−30 to +52	−30 to +52
Deep discharge cycles	Cycle	5000	5000
Shallow HEV cycles	Cycle	300,000	300,000

[a]The goals are for the battery pack including container and all management systems.

4 VEHICLE SIMULATION RESULTS

The previous sections have discussed in general terms various aspects of the design and operation of electric and hybrid vehicles. The discussions focused on the energy and power requirements and the trade-offs between those requirements, vehicle design parameters, and component performance characteristics. In this section, the results of detailed vehicle simulations are presented that show how vehicles meeting specified power and energy design requirements can be expected to perform on various driving cycles. Simulation results will be discussed separately for EVs and HEVs and PHEVs. The simulations were obtained using the Advisor computer program (Wipke *et al*., 1999; Gonder *et al*., 2009) as modified at the University of California-Davis.

4.1 Electric vehicle designs and performance

The design of EVs has been discussed in general terms in Section 3.1. The simulation of EV operation is quite simple because all the torque to the wheels is applied by the electric motor and all the energy to power the vehicle comes from the battery, which is charged from the wall-plug. The block diagram of the driveline schematic used in Advisor to simulate EVs is shown in Figure 4.

The input screen for the Advisor simulation is shown in Figure 5 and typical output graphics for a simulation are given in Figures 6 and 7. Outputs are shown for the FUDS and US06 driving cycles. The higher power demands and battery currents for the US06 driving cycle are clearly evident from the figures.

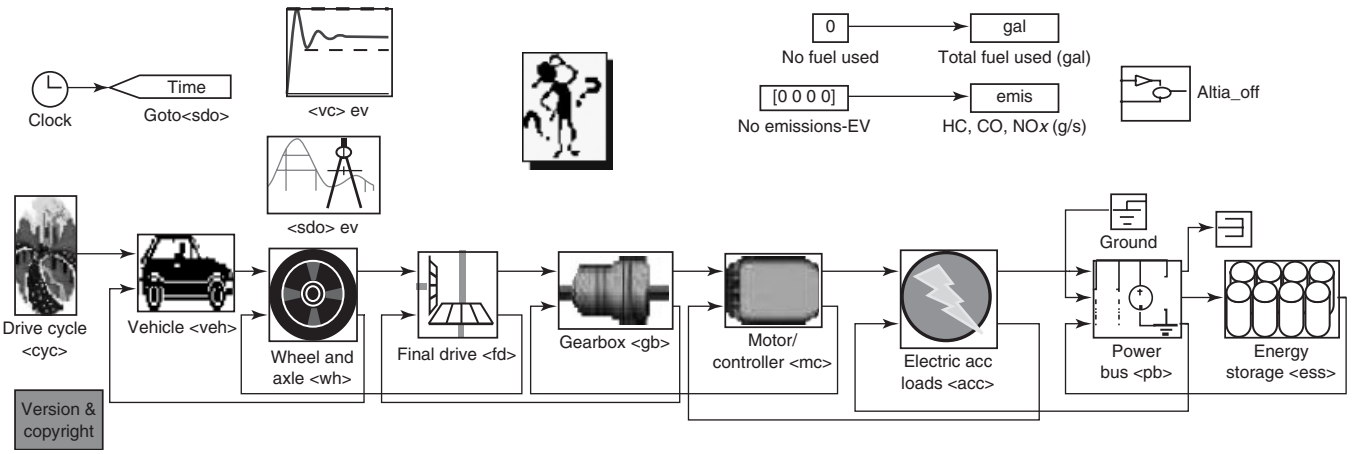


Figure 4. Advisor block diagram of the EV driveline.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

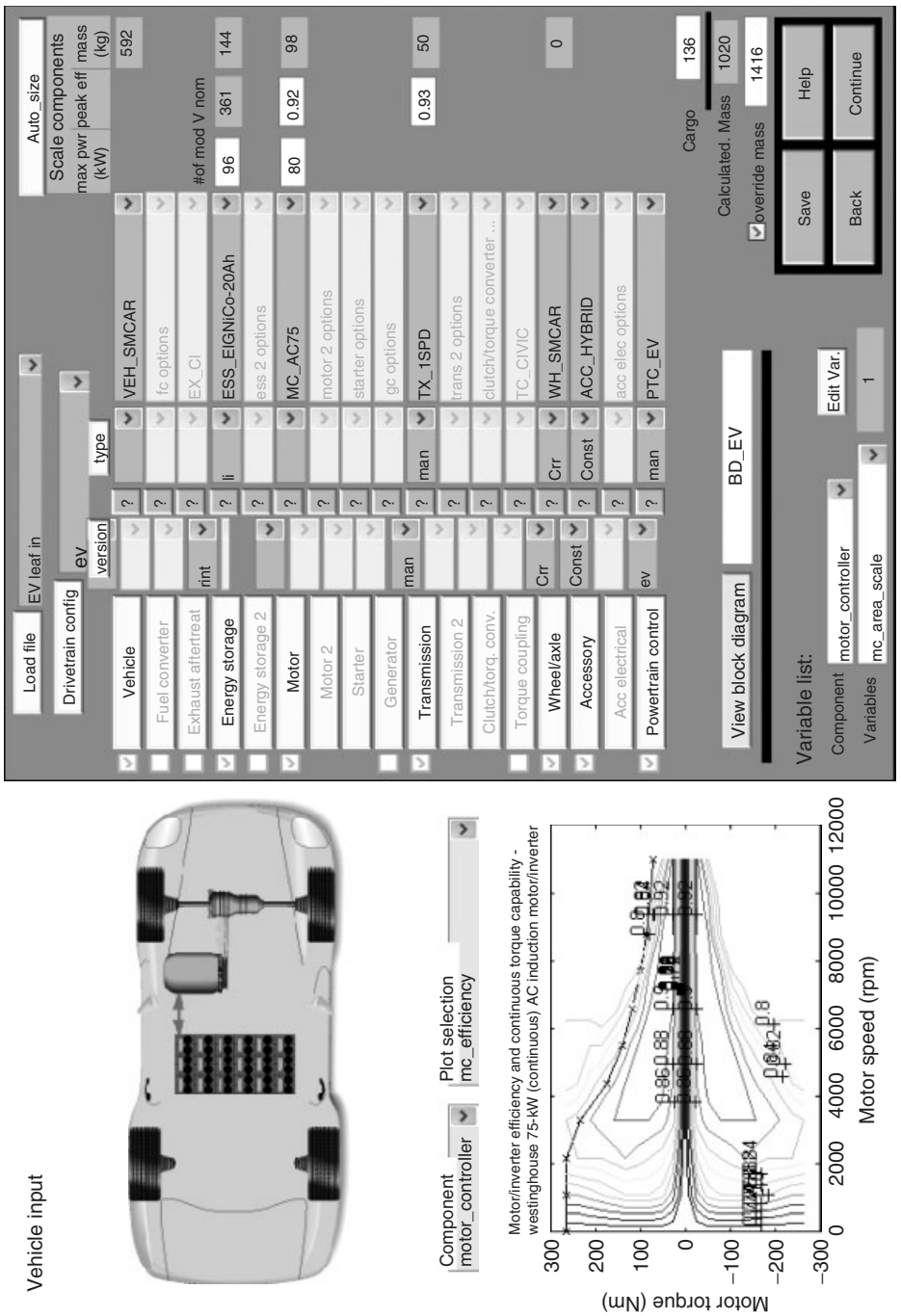


Figure 5. Input graphic for an electric vehicle simulation using Advisor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

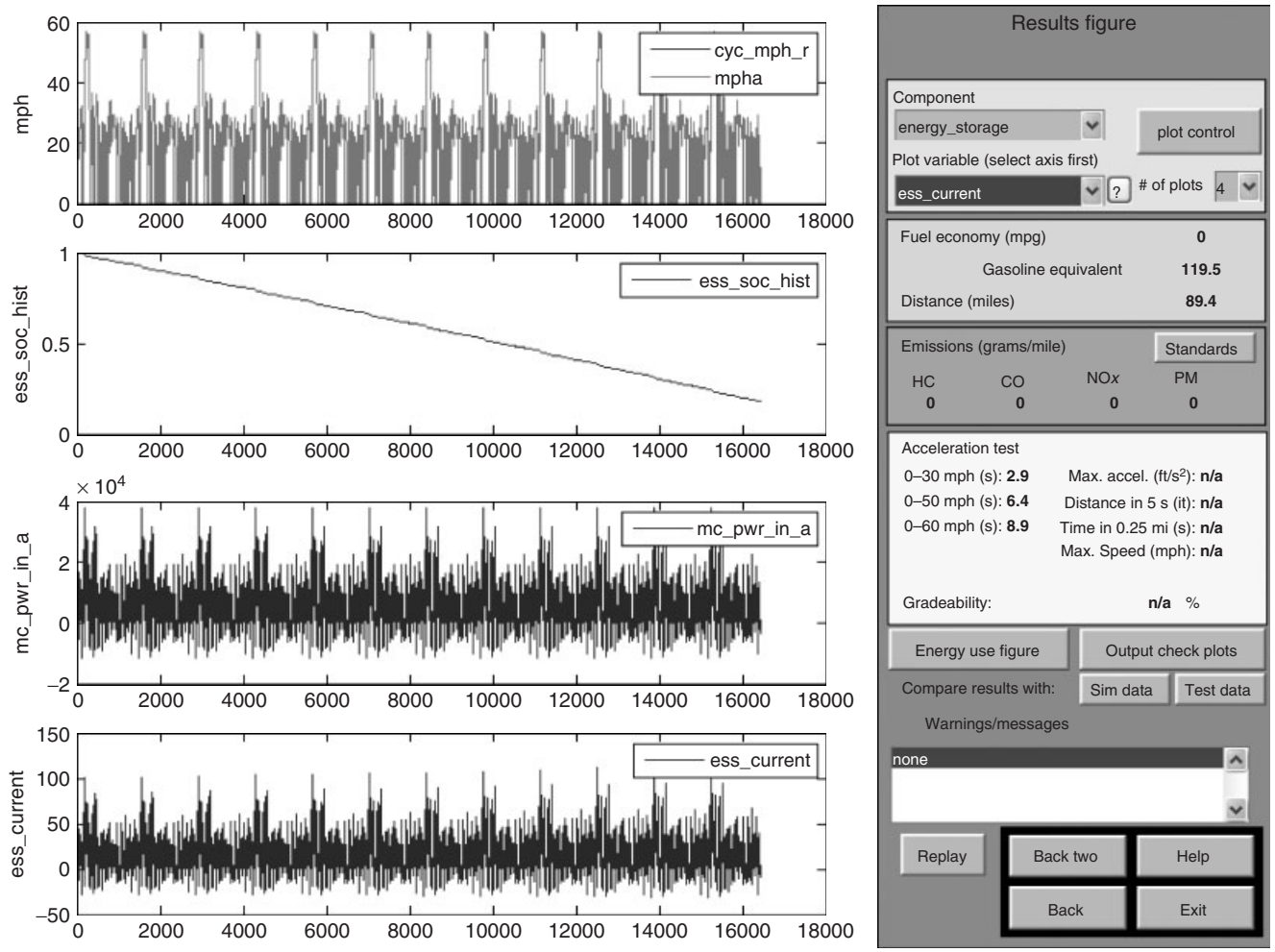


Figure 6. Output graphic for an electric vehicle simulation using Advisor—FUDS driving cycle.

Simulation results for various vehicle designs and driving cycles are summarized in Table 10. The results are given for the energy use (Wh/km) from the battery, range (km) using 80% of the energy stored in the battery, and the times (s) for 0–48 and 0–96 kmph accelerations. The influence of changes in vehicle weight, drag, rolling resistance, and peak motor power on the vehicle performance is evident in the table. The energy use and range of an EV are strongly dependent on the driving cycle and vehicle speed. Hence, specifying a single value for the range can lead to an inaccurate description of the EV range capability. The acceleration capability of an EV depends on the maximum power of the electric motor. As indicated in Table 10, the acceleration performance of the EV can be significantly increased with only a small decrease in vehicle range using the same battery.

Battery-powered EVs are recharged with electricity from the wall-plug. The energy use of the EVs is given as Wh/mi from the battery. The gasoline equivalent can be calculated from $(\text{gal/mi})_{\text{gas.equiv.}} = (\text{kWh/mi})/33.7$. The energy saved depends on the battery charging efficiency and the efficiency of the powerplant generating the electricity. For an EV using 245 Wh/mi from the wall-plug (90% charging efficiency), the gasoline energy equivalent saved is 75% from the wall-plug and 38% at a 40% efficient powerplant compared to the 2007 baseline ICE mid-sized car (34 mpg). Compared to a 2030 HEV (85 mpg), the gasoline equivalent saved is only 38% from the wall-plug and there are no savings at the powerplant until the efficiency of the powerplant exceeds about 60%. Hence, the energy saving potential of EVs can be expected to decrease markedly in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

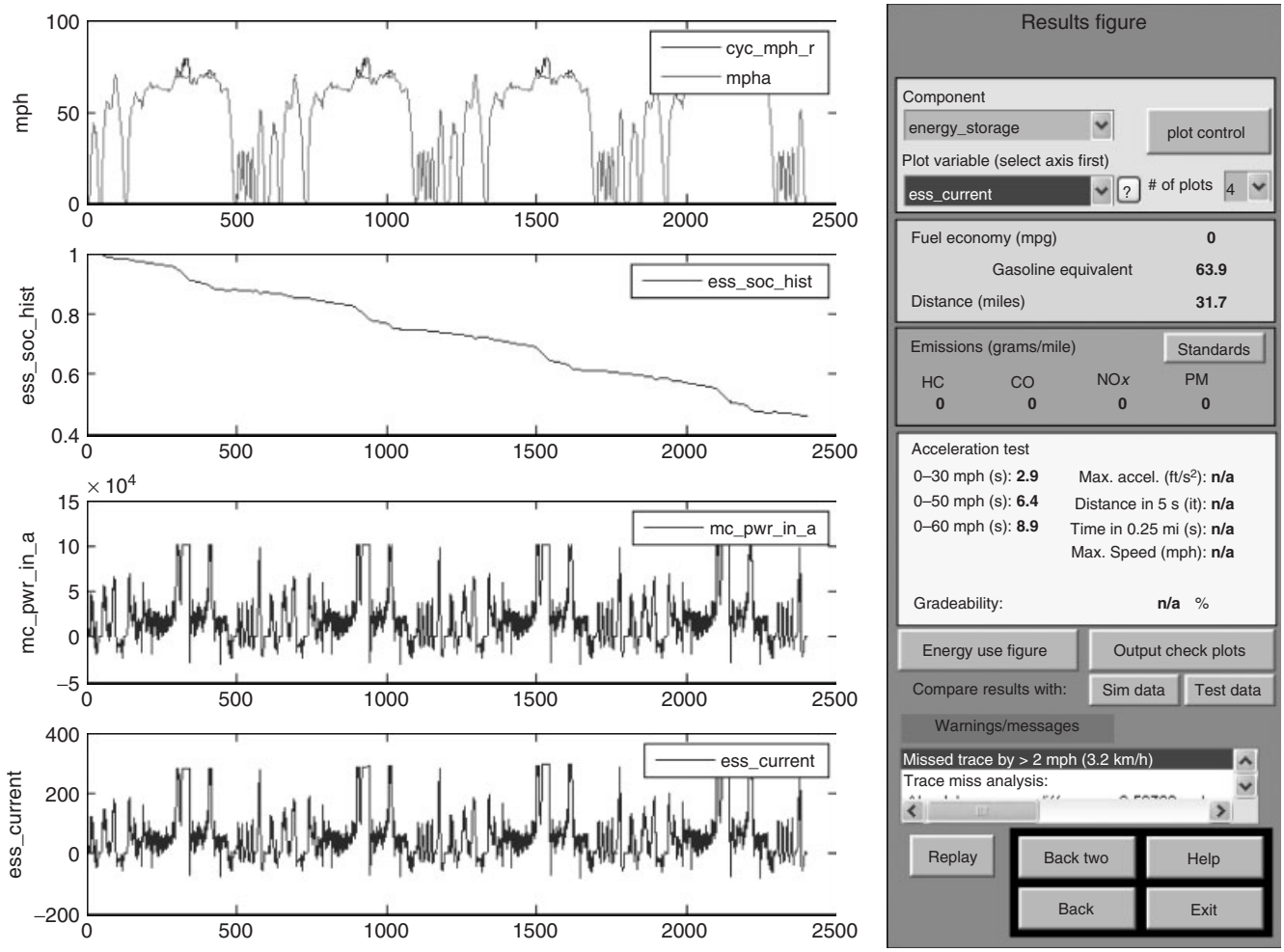


Figure 7. Output graphic for an electric vehicle simulation using Advisor—US06 driving cycle.

future as the efficiency of hybrid vehicles continue to increase.

4.2 Hybrid vehicle (HEV) designs and performance

The design of charge-sustaining HEVs has been discussed in general terms in Section 3.2. Detailed simulations of HEVs are discussed in this section. The Advisor program can be used to simulate the operation of hybrid vehicles (Burke, 2007; Burke and Miller, 2010). As indicated in Figure 8, the block diagram of the hybrid driveline in Advisor is considerably more complex than that of the EV (Figure 4).

The control strategy for the HEV driveline that is implemented via the block diagram is complex because its intent is to share the power demanded between the engine and the electric motor such that the engine operates only in the most efficient regions of the engine map. As will be discussed later in this section, this strategy leads to large improvements in the fuel economy of the HEVs compared to conventional ICE vehicles.

The Advisor vehicle simulation program has been utilized to calculate the fuel economy of a number of conventional ICE and HEV vehicles being marketed in 2010. The results of the simulations are summarized in Table 13. In nearly all the cases, the EPA dynamometer test data (U.S. Department of Energy and U.S. Environmental Protection Agency) and the simulation results are in good agreement. These comparisons serve to validate the Advisor program for evaluating HEV technologies.

Projections of fuel economy and fuel savings of HEVs have been made for future (Ogden and Anderson, 2011) using Advisor. The inputs used in the simulations are given in Table 14 and the results of the simulations are summarized in Table 15. A typical output graphic of the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

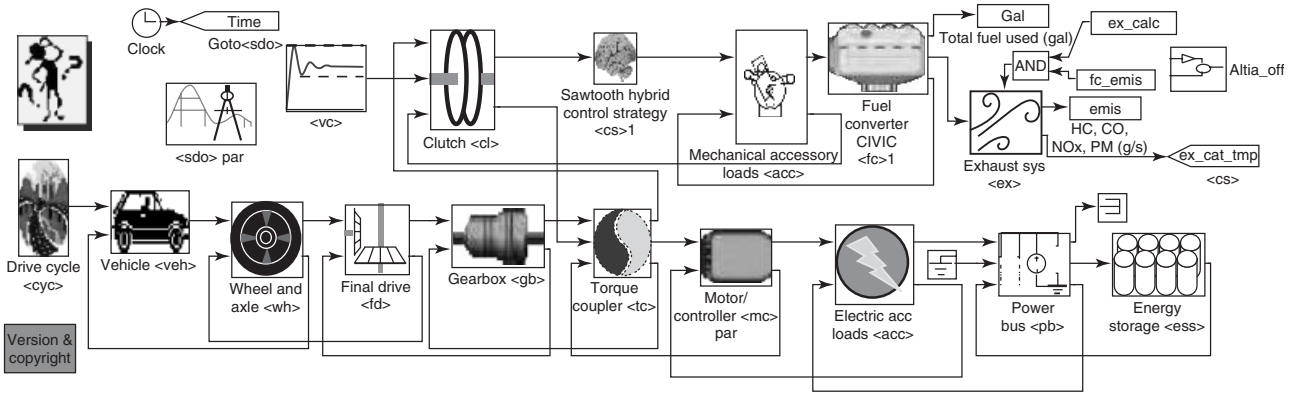


Figure 8. Advisor block diagram of the HEV driveline.

Table 13. Summary of fuel consumption simulation results for HEVs marketed in 2010.

Model/Year	Engine	Driveline Type	L/100 km; mpg City[a]	L/100 km; mpg Highway[a]
Ford Focus/2010 simulation	Focus	Conventional	8.5; 28	5.41; 44
EPA test 2007/Ford Focus	Focus	Conventional	7.93; 30	5.41; 44
Honda Civic simulation	i-VTEC	Conventional	7.21; 33	5.29; 45
EPA test 2007/ Honda Civic	i-VTEC	Conventional	7.21; 33	50
Honda Civic simulation	i-VTEC	Hybrid	4.21; 56.5	3.81; 62.5
EPA test 2007/ Honda Civic	i-VTEC	Hybrid	4.38; 54.4	3.64; 65.4
Toyota Prius simulation	Atkinson	Hybrid	3.5; 68	3.53; 67.5
EPA test 2007/ Toyota Prius	Atkinson	Hybrid	3.60; 66.6	3.64; 65.4
EPA test 2007/ Honda Accord	4 cylinder 140 kW	Conventional	8.95; 26.6	5.46; 43.6
EPA test 2007/ Toyota Camry	4 cylinder 140 kW	Conventional	8.95; 26.6	5.63; 42.3

[a] L/100 km = 238/mpg.

Table 14. Input parameters for the advanced ICE and HEV vehicle simulations of mid-sized automobiles.

Vehicle Configuration	Parameter	2015	2030	2045
	C_D	0.25	0.22	0.20
	A_F m^2	2.2	2.2	2.2
	F_r	0.007	0.006	0.006
Advanced ICE	Engine kW	105	97	97
	Maximum engine efficiency (%)	39	40	41
	Vehicle test weight (kg)	1403	1299	1299
HEV	Engine kW	73	67	67
	Maximum engine efficiency (%)	39	40	41
	Motor kW	26	24	24
	Lithium battery kWh	1.0	0.9	0.9
	Vehicle test weight (kg)	1434	1324	1324

calculations is shown in Figure 9. The large improvements in fuel economy shown in Table 14 are consistent with those found in previous studies by other groups (Assessment of Fuel Economy Technologies for Light-duty Vehicles, 2010; Plotkin and Singh, 2009; Kasseris and Heywood, 2007). These improvements are due largely to the efficient operation of the engine made possible by the addition of the electric motor to the driveline. As indicated in Figure 9, the engine is operated in an on/off mode and only consumes fuel when it is needed to power the vehicle and recharge the battery to maintain it in a narrow range of SOC. For the example simulation shown in Figure 9, the average engine efficiency was 37%. The fuel and energy savings are 60% and 29% compared with the 2007 ICE and 2030 ICE vehicle, respectively.

4.3 Plug-in hybrid vehicle (PHEV) designs and performance

The design of PHEVs has been discussed in general terms in Section 3.3. Detailed simulations of HEVs are discussed in this section. The Advisor program can be used to simulate the operation of PHEVs using essentially the same block diagram shown in Figure 8. In the case of the PHEVs, the vehicle operates when possible as an EV when the battery SOC is above a minimum specified

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

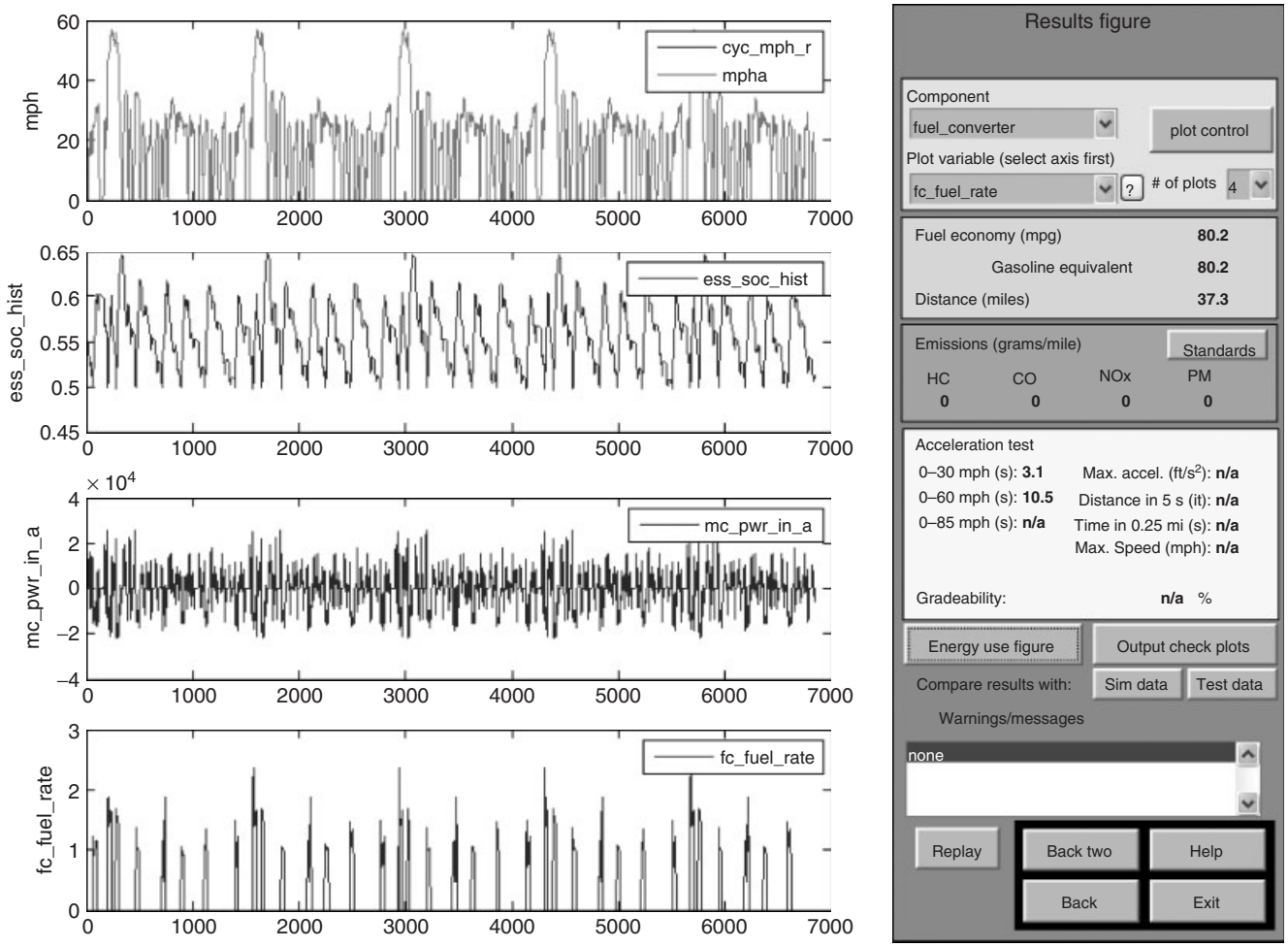


Figure 9. Output graphic for HEV-2030 on the FUDS driving cycle.

value and as a charge-sustaining HEV after the battery has been depleted to the minimum value. When the electric motor is large enough (kW) to meet the power demand of the vehicle, the engine is not turned on and the PHEV operates as an EV. If the power demand exceeds that of the electric motor, the engine is turned on even though the battery SOC is above the specified minimum and the vehicle operates in the blended mode using both the

Table 15. Simulation results for advanced ICE and HEV vehicles.

Year	FUDS L/100 km; mpg[a]	FHWDS L/100 km; mpg	% Fuel Savings	US06 L/100 km; mpg	Acceleration 0–48/ 0–96 kmph
Baseline 2007	9.15; 26	5.67; 42	0		
Advanced ICE					
2015	5.75; 41.4	3.82; 62.3	33.5	6.35; 37.5	4.3/9.7
2030	5.02; 47.4	3.25; 73.3	42.8	5.41; 44	4.7/10.3
2045	4.87; 48.9	3.09; 77.1	45.2	5.16; 46.1	4.6/10.3
HEV					
2015	3.25; 73.3	3.21; 74.1	53.1	5.12; 46.5	4.3/9.7
2030	2.78; 85.7	2.83; 84	59.3	4.43; 53.7	4.7/10.3
2045	2.71; 87.9	2.67; 89.2	61.0	4.27; 55.8	4.6/10.3

[a] L/100 km = 238/mpg (gasoline).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 16. Characteristics of the batteries used in the Advisor simulations.

Vehicle Configuration	2015				2030–2045			
	Battery Type	Ah	Wh/kg	Resistance(mOhm)	BatteryType	Ah	Wh/kg	Resistance (mΩ)
HEV	Li Titanate	4	35	1.1	Li Titanate	4	42	0.9
PHEV-20	Ni MnO_2	15	120	1.5	Ni MnO_2	15	135	1.3
PHEV-40	Ni MnO_2	50	140	0.8	Ni MnO_2	50	170	0.65

Table 17. Inputs for the Advisor simulations of mid-sized PHEV passenger cars.

Vehicle Configuration	Parameter	2015	2030	2045
	C_D	0.25	0.22	0.20
	A_F m^2	2.2	2.2	2.2
	F_r	0.007	0.006	0.006
PHEV-20	Engine kW	75	69	68
	Motor kW	61	57	57
	Lithium battery kWh	4.0	3.6	3.6
	Vehicle test weight (kg)	1475	1361	1354
PHEV-40	Engine kW	77	71	67
	Motor kW	63	59	59
	Lithium battery kWh	11.1	9.8	9.4
	Vehicle test weight (kg)	1535	1415	1407

Table 18. Summary of the simulation results for mid-sized PHEV passenger cars.

Year	Driving Cycle	Electric Range (km)	Charge-Depleting mpg	Charge-Depleting (Wh/km)	Charge-Sustaining L/100 km; mpg
PHEV-20					
2015	FUDS	27	All-elec	102	3.4; 70.0
	FHWDS	27	All-elec	103	3.42; 69.6
	US06	16	1570	175	5.29: 45
2030	FUDS	27	3333	89	3.09; 77
	FHWDS	27	7500	91	2.83; 84
	US06	18	1500	146	4.49; 53
2045	FUDS	29	All-elec	88	2.78; 85.6
	FHWDS	30	All-elec	84	2.71; 87.8
	US06	18	1400	146	4.51; 52.8
PHEV-40					
2015	FUDS	74	All-elec	167	3.44; 69.1
	FHWDS	72	All-elec	171	3.32; 71.7
	US06	50	800	251	5.15; 46.2
2030	FUDS	78	All-elec	141	2.81; 84.6
	FHWDS	77	All-elec	143	2.77; 86.0
	US06	51	1495	218	4.37; 54.5
2045	FUDS	78	All-elec	135	2.71; 87.8
	FHWDS	78	All-elec	134	2.57; 92.5
	US06	51	1731	205	4.03; 59

Table 19. Comparisons of the fuel (gasoline) and energy savings with various advanced vehicle technologies using different baseline vehicles.

Driveline	2007 ICE Baseline		2030 ICE Baseline		2030 HEV Baseline	
	% Fuel	% Energy[a]	% Fuel	% Energy[a]	% Fuel	% Energy[a]
PHEV-20	66	58	40	26	16	—
PHEV-40	86	60	75	29	66	—
BEV-100[b]		47		6		—
2030 HEV	60	60	29	29		

[a] Powerplant efficiency 40%.
[b] 211 Wh/mi from wall-plug.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

electric motor and the engine. This will be the case for most PHEVs for high power driving cycles such as the US06 even though the same vehicles could operate as EVs on the FUDS driving cycle. One of the key design parameters for the PHEVs is their AER. This is the driving distance for the vehicle when the battery becomes depleted to the minimum specified SOC. For driving distances shorter than AER, the vehicle will use mostly electricity from the battery and the fuel economy based on gasoline used will be very high (often in excess of 100 mpg). The characteristics of the batteries used in the Advisor simulations are given in Table 16. In all cases, lithium batteries were used.

Simulations of PHEVs have been performed using the Advisor program for a number of vehicle designs (Ogden and Anderson, 2011). The inputs for these designs are given in Table 17 for nominal AER values of 20 and 40 miles. The simulation results are summarized in Table 18. Results are given for operation of the vehicles in both the charge-depleting and charge-sustaining modes. For these designs, the vehicles are essentially EVs for driving distances shorter than the AER. The fuel economy in the charge-sustaining mode is essentially the same as an HEV (Table 15). For a specified driving range in excess of AER, the energy use (electricity and gasoline) can be calculated from the values given in Table 18. The acceleration characteristics of the PHEVs on the helectric motor only will be modest (0–96 kmph in about 10 s) based on a power-to-weight ratio of 0.04 kW/kg but will be very good using both the electric motor and engine (blended operation) with a power-to-weight ratio of 0.092 kW/kg.

The split between electricity and gasoline for PHEV vehicle will depend on its usage pattern (average miles driven per day and number of long trips taken). Assuming for the PHEV-20 and PHEV-40 mid-sized car that 20% and 65% of the total annual kilometers (miles) (city plus highway), respectively, are driven on electricity, one can calculate the wall-plug electricity and gasoline used and the total energy (gasoline plus energy needed to generate the electricity) savings. Assuming 24,000 annual kilometers (15,000 miles), a battery charger efficiency of 90%, and a powerplant efficiency of 40%, one can calculate the gasoline (fuel) and energy savings of the two PHEVs. The results for the PHEVs are given in Table 19 compared with similar results for other advanced technologies using different baselines for the comparisons. In terms of saving energy, there is not much difference between the various hybrid technologies. However, the energy savings with the BEV-100 are much less than with the hybrids regardless of the baseline used. The energy savings with the EV would be greater for a more efficient powerplant than the 40% plant assumed in the calculations. Clearly, the PHEV-40 and the BEV-100 result in more fuel (gasoline) savings because of the substitution of electricity to power them for most or all of their operation.

5 SUMMARY AND CONCLUSIONS

In this chapter, the power and energy requirements (both electricity and gasoline) for battery-powered EVs and HEVs, including PHEVs, are assessed. The influence of the vehicle design parameters (engine and motor power, weight, frontal area, drag coefficient, and tire rolling resistance) on vehicle performance (range, speed, and acceleration) is determined based on simulations of several classes of vehicles. The battery must be sized to meet the energy storage (kWh) and power (kW) requirements of the vehicles. It was found that lithium batteries of various chemistries are available that satisfy the requirements for both HEVs and PHEVs.

The power and energy requirements were calculated for both parallel and series powertrain arrangements using the Advisor vehicle simulation computer program. The simulations were performed for several driving cycles (FUDS, Highway, and the US06) for vehicle and component characteristics expected to be appropriate for 2015–2045. The simulation results indicated that large reductions in fuel consumption and energy use can be expected in the future with the various advanced technologies (Table 19). The magnitudes of the reductions depend on the baseline used. The energy savings (gasoline plus electricity including powerplant efficiency) are about 60% for all the hybrid technologies considered using the 2007 ICE vehicle baseline and about 30% using the 2030 ICE baseline. The energy savings (electricity) for the EV are smaller than for the hybrids especially compared with the advanced ICE and HEV technologies. The primary advantage of the EV is the substitution of electricity for a liquid fuel.

SYMBOLS AND PARAMETERS

C_d	drag coefficient of the vehicle
A_f	frontal area of the vehicle (m^2)
$f_{rolling}$	rolling resistance of the tires (kg/kg)
W_V	inertia weight of the vehicle (kg)
L/100 km	fuel consumption in terms of liters per 100 km
mpg	fuel consumption in terms of miles per gallon

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Wh/km	electrical energy from the battery per kilometer
$(kWh)_{usable}$	usable energy (kWh) from the battery
Wh/kg	energy density of the battery
W/kg	power density of the battery
Ah	Amp-hour capacity of a cell in the battery
Driving cycles (FUDS, Highway, and US06)	The *velocity versus time* profiles used in the vehicle simulations that are intended to model driving in the real world. These same driving cycles are used in the United States to test vehicles for emissions and fuel economy.

NOMENCLATURE

ICE	Conventional vehicle with engine and transmission
EV	Electric vehicle (battery powered)
HEV	Hybrid electric vehicle (charge sustaining with electric motor and engine)
PHEV	Plug-in hybrid electric vehicle with a rechargeable battery
PHEV-XX	PHEV vehicle with an all-electric range of XX miles
Parallel hybrid	Hybrid vehicle in which both the electric motor and engine apply a torque to the vehicle drive shaft
Series hybrid	Hybrid vehicle in which the electric motor is connected to the wheels and the engine drives a generator to produce electricity
Single-shaft hybrid	Hybrid vehicle in which the electric motor and engine are on the same shaft with or without a clutch to decouple the engine from the shaft
Planetary shaft hybrid	Hybrid vehicle in which the electric motor, generator, and engine is connected using a planetary gear arrangement (Toyota Prius)
CVT	Continuously variable transmission (mechanical component)
All	electric range—the range for a PHEV in which the vehicle is operated in the charge-depleting mode and the fuel usage is minimal
Advisor	The vehicle simulation computer program used to simulate the operation of the ICE, EV, HEV, and PHEV vehicles

RELATED ARTICLES

Fundamentals, Basic Principals in Road Vehicle Aerodynamics & Design
Wind Tunnel Aerodynamic Measurements: Aerodynamic Detailing, Wind Noise, etc.
Physics of Car Crashes: Design Concepts for Safer Cars
Body Design, Overview, Targeting a Good Balance Between all Vehicle Functionalities
Lightweighting Approach: A Historical Perspective
Automotive Applications for Magnesium
Automotive Applications for Titanium
Structure and Properties of Polymeric Composites

REFERENCES

(2004) *Achieving Lightweight Vehicles*. SAE Publication SP-1846, March, SAE International.

Assessment of Fuel Economy Technologies for Light-duty Vehicles 2010, National Research Council Report.

Barrand, J. and Bokar, J. (2008) Reducing tire rolling resistance to save fuel and lower emissions. SAE paper 2008-01-0154.

Bradley, T., Huff, B., and Frank, A. (2005) Energy consumption test methods and results of servo-pump continuously variable transmission control system. SAE Paper 2005-01-3782.

Burke, A.F. (2007) Batteries and ultracapacitors for electric, hybrid, and fuel cell vehicles. *IEEE Journal*, special issue on Electric Powertrains, April, **95** (4), 806–821.

Burke, A.F. (2009) Ultracapacitor technologies and applications in hybrid and electric vehicles. *International Journal of Energy Research* (Wiley), **34** (2), 133–151.

Burke, A.F. and Miller, M. (2009) Performance characteristics of lithium-ion batteries of various chemistries for plug-in hybrid vehicles, EVS-24, Stavanger, Norway, May (paper on the CD of the meeting).

Burke, A. and Miller, M. (2010) Lithium batteries and ultracapacitors alone and in combination in hybrid vehicles: fuel economy and battery stress reduction advantages. Paper presented at the Electric Vehicle Symposium 25, Shenzhen, China, November.

Burke, A.F. and Miller, M. (2011) The power capability of ultracapacitors and lithium batteries for electric and hybrid vehicle applications. *Journal of Power Sources*, **196** (1, January), 514–522.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Burke, A.F. and Van Gelder, E. (2008) Plug-in hybrid-electric vehicle powertrain design and control strategy options and simulation results with lithium-ion batteries. Paper presented at EET-2008 European Ele-Drive Conference, Geneva, Switzerland, March 12 (paper on CD of proceedings).

Burke, A.F., Zhao, H., and Van Gelder, E. (2009) Simulated performance of alternative hybrid-electric powertrains in vehicles on various driving cycles, EVS-24, Stavanger, Norway, May (paper on the CD of the meeting).

Duoba, M., Carlson, R., Jehlik, F., *et al*. (2009) Correlating dynamometer testing to in-use fleet results of plug-in hybrid electric vehicles, EVS-24, Stavanger, Norway, May 13-16.

Frank, A. (2004) Engine Optimization Concepts for CVT-Hybrid Systems to obtain the Best Performance and Fuel Efficiency, *Proceedings of the 2004 UC Davis Continuously Variable Transmission Conference*.

Ghorbani, R., Bibeau, E., and Filizadeh, S. (2010) On conversion of hybrid electric vehicles to plug-in. *IEEE Transactions on Vehicular Technology*, **50** (4), 2016–2020.

Gonder, J. Pesaran, A., Lustbader, J., and Tataria, H. (2009) *Fuel economy and performance of mild hybrids with ultracapacitors – simulations and vehicle test results*. Presented at the 5th International Symposium on Large EC Capacitor Technology and Applications, Long Beach, California, June 9-10.

Hucho, W.-H. (ed.) (1998) *Aerodynamics of Road Vehicles*, 4th edn, SAE International.

Husain, I. (2005) *Electric and Hybrid Vehicles: Design Fundamentals*, 2nd edn, CRC Press.

Hybrid vehicle sales data (current), hybridcar.com.

Kasseris, E. and Heywood, J. (2007) Comparative analysis of automotive powertrain choices for the next 25 Years. SAE paper 2007-01-1605.

Mallick, P.K. (ed.) (2010) *Materials Design and Manufacturing for Lightweight Vehicles*, CRC Press, Oxford.

Plotkin, S. and Singh, M. (2009) Multi-path transportation futures study: vehicle characterization and scenarios, Argonne Lab and DOE Report (draft), March 5.

Schafer, A., Heywood, J., Jacoby, D., and Waitz, I.A. (2009) *Transportation in a Climate-Constrained World*, Chapter 4, Road Vehicle Technology, MIT Press.

Snyder, K. (2012) Overview and Progress of United States Advanced Battery Consortium (USABC) Activity. Presentation given May 15.

Ogden, J. and Anderson, L. (eds), Chapter 4 (Comparing Fuel Economies and Costs of advanced vs. conventional vehicles), August (2011) *Sustainable Transportation Energy Pathways*. University of California-Davis, Institute of Transportation Studies.

Tire and Passenger Vehicle fuel Economy (2006) Transportation Research Board Special Report 286.

USABC Goals for Advanced Batteries for EVs, USCAR home web page, Energy Storage TLC/Publications.

U.S. Department of Energy and U.S. Environmental Protection Agency, Fuel Economy Guide-2007, DOE/EE-0314.

Wipke, K., Cuddy, M., Bharathan, D., *et al*. (1999) Advisor 2.0: A Second-Generation Advanced Vehicle Simulator for Systems Analysis, NREL Technical Report NR EL/TP-540-25928; National Rene wable Energy Laboratory, last modified: April 30, 2002 [version 2002].

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto063
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Advanced Batteries for Vehicle Applications

Andrew F. Burke
University of California, Davis, CA, USA

1 Introduction 1
2 Battery Requirements for Future Electric and Plug-in Hybrid Vehicle Development 2
3 Battery Chemistries Being Developed 3
4 Test Data and the Status of Advanced Batteries of Various Battery Chemistries 12
5 Vehicle Calculations Using Advanced Batteries 15
6 Summary and Conclusions 18
Nomenclature and Abbreviations 19
Related Articles 19
References 19

1 INTRODUCTION

This chapter is concerned with the research and development on advanced batteries for plug-in electric vehicles (PEVs) in the next 10–20 years. The PEVs include both electric vehicles (EVs) and plug-in hybrid electric vehicles (PHEVs). The primary motivation for the development of the advanced batteries is to permit the design of electric vehicles having ranges comparable to conventional internal combustion engine (ICE) vehicles—300–500 miles. By advanced batteries are meant cell chemistries having energy densities significantly higher than those presently available, hopefully leading to the batteries needed for 300–500-mile range electric cars. This chapter is not concerned with batteries for charge sustaining HEVs. These vehicles require batteries with high power capability, and energy density is a secondary concern. For this reason, it is unlikely that any of the battery technologies considered in this chapter will be appropriate for HEVs. Present lithium battery technologies seem to be adequate for the HEV application as all the auto manufacturers that are presently marketing HEVs have begun to use lithium batteries in place of the nickel metal hydride batteries used in their earlier designs.

Early research is underway and laboratory cells are being prepared for most of the chemistries considered in this chapter, but in nearly all cases, it is anticipated that many years of R & D will be needed before the most advanced batteries could be marketed for use in vehicles. Most of the research is presently concerned with new lithium chemistry formulations and metal–air in electrodes and small cells. It is expected that some of the new lithium chemistries will be the first of the advanced battery technologies to be marketed in the next 5–10 years. The metal–air batteries are much further in the future and likely will be required to approach the 300–500-mile range goal.

The first sections of the chapter will deal with defining the requirements for the advanced batteries if they are to be used in long-range electric vehicles and a review of the chemistries that are likely to lead to high energy density, rechargeable batteries suitable for use in these vehicles. In the next sections, the present status of research on advanced batteries is discussed and available data for the various chemistries presented. In the final section, computer simulation results are given for vehicles using projected advanced batteries. These simulations include combinations of the advanced batteries with ultracapacitors or high power lithium batteries that would be needed if the advanced batteries do not have sufficiently high power capability to meet vehicle acceleration and braking requirements.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 BATTERY REQUIREMENTS FOR FUTURE ELECTRIC AND PLUG-IN HYBRID VEHICLE DEVELOPMENT

Using lithium batteries available in 2012, the range of PEVs is limited to 75–100 miles in order that the weight and volume of the battery pack does not significantly affect the utility/functionality of the vehicle. For example, the Nissan Leaf utilizing lithium batteries has a useful range of about 75 miles with a battery pack weighing about 300 kg and a volume of about 300 L. The battery stores 24 kWh with about 80% useable on a regular basis. The energy consumption of the Leaf is about 250 Wh/mi. The energy densities of the cells in the Leaf battery are 140 Wh/kg and 300 Wh/L. In order to achieve a range of 300 miles, a vehicle such as the Leaf would need to store about 100 kWh and require cells with energy densities of 600 Wh/kg and 1200 Wh/L to maintain the same size battery pack. The maximum power requirement of the battery in the Leaf is 100 kW assuming a 90-kW electric motor with an efficiency of 90%. For a 300-mile vehicle, the power requirement for the cells would be 600 W/kg, which is the same as the battery in the Leaf. This is a modest power requirement for the lithium chemistries, but may not be easily met with the metal–air chemistries.

As shown in Table 1, the vehicle energy and power requirements can be reduced if the vehicle weight and road load characteristics are improved. For example, if the vehicle energy consumption is reduced to 180 Wh/mi, the battery pack energy storage requirement for a 300-mile range Leaf-type EV becomes 72 kWh and the power requirement is 80 kW. The corresponding cell performance requirements are 433 Wh/kg, 880 Wh/L, and 480 W/kg for the same size battery as in the 2011 Leaf.

It seems clear that to achieve 300-mile range in a mid-size passenger car will require a significant reduction in the energy consumption (Wh/mi) of the vehicle and large increases in the energy density of the battery. Reasonable design targets for the advanced batteries are 500–600 Wh/kg and 1100–1400 Wh/L for energy density and 500–600 W/kg for power density. Successful development of these batteries would permit the design of electric vehicles with ranges comparable to conventional ICE vehicles. However, the refueling (recharge) time for the batteries would be much longer than filling the gasoline tank of the conventional vehicle. The power density (W/kg and W/L) of the advanced batteries must be sufficiently high to permit the desired acceleration performance of the vehicle without the need to size the battery based on power requirements. High energy density batteries can be used to either extend the range of the EV or reduce the size and cost of the battery for the same range. In either case, the development of high energy density batteries will improve the prospect for marketing electric vehicles.

Table 1. Performance characteristics of battery-powered electric vehicles of various weight, drag, and tire rolling resistance.

Test Weight[a] kg	Drag Coefficent[b] C_d	Rolling Resistance Coefficient f_r	City Energy Consumption[c] (Wh/mi)	City range[d] (miles)	Highway Energy Consumption (Wh/mi)	Highway Range (miles)	Acceleration 0–60 mph (s)
1200	0.25	0.008	204	78	195	82	8.0
	0.30	0.008	210	76	209	77	8.0
	0.20	0.008	198	81	180	89	7.9
	0.25	0.006	194	83	184	87	7.9
	0.30	0.006	200	80	198	81	8.0
	0.20	0.006	188	85	169	95	7.9
1500	0.25	0.008	221	72	208	77	9.4
	0.30	0.008	228	70	222	72	9.5
	0.20	0.008	215	74	193	83	9.4
	0.25	0.006	208	77	194	83	9.3
	0.30	0.006	214	75	208	77	9.4
	0.20	0.006	202	79	179	89	9.3
1800	0.25	0.008	239	67	220	73	11.1
	0.30	0.008	246	65	235	68	11.2
	0.20	0.008	233	69	206	78	11.0
	0.25	0.006	224	71	204	78	11.0
	0.30	0.006	230	70	218	73	11.1
	0.20	0.006	218	73	189	85	10.9

[a] All vehicles used a 75-kW AC induction motor, regenerative braking.
[b] The frontal area is 2 m^2.
[c] All vehicles used lithium-ion batteries, 150 Wh/kg, 20 kWh, 80% usable.
[d] 16kWh useable energy in all the battery packs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The advanced batteries can also be used in PHEVs, permitting the design of vehicles with all-electric ranges of 50–75 miles, which would satisfy the needs of most drivers on most days. The result would be vehicles that would use gasoline only for long trips and electricity for a large fraction (until 75%) of the total miles per year. The size (weight and volume) of the battery would be reduced proportional to the increase in the energy density of the battery. For example, in the case of the General Motor's Volt, the weight of the cells in the battery using 300 Wh/kg cells would be 75 kg for an all-electric range of 75 miles if 80% of the stored energy (22.5 kWh) was useable. This weight compares with 110 kg for 40 miles using the cells available in 2011. The power density of the advanced batteries used in the 75-mile range PHEVs would need to be about 1500 W/kg to meet the 115-kW power requirement of the Volt. Hence, the power density of batteries to be used in PHEVs must be higher than those used in EVs.

If the power capability of the advanced battery is not sufficient to meet the acceleration requirements of either the EV or the PHEV, it would be necessary to combine the advanced battery with an ultracapacitor or high power lithium battery. These high power energy storage units are similar to those used in HEVs (charge sustaining hybrids) and should be readily available for use with the advanced batteries.

The previous paragraphs have been concerned with determining the performance requirements of the advanced battery cells. The advanced batteries must also meet essentially the same cycle and calendar life and safety requirements as the lower performance batteries presently available. That is at least a 10-year life in the vehicle application and safe operation with minimal concern for dangerous failure modes such thermal runaway. If the EV using the advanced batteries is to be cost competitive with a conventional engine-powered vehicle, the battery unit cost ($/kWh) must be at least a factor 3–4 less than present lithium batteries. This will be possible only if the advanced batteries have a weight-based cost ($/kg) equal to or less than present lithium batteries. This will preclude the use expensive materials in the advanced batteries.

3 BATTERY CHEMISTRIES BEING DEVELOPED

3.1 Introduction

There are a number of battery types/chemistries being developed that could be used in vehicles. The potential (theoretical limit) performance of the various chemistries is shown in Table 2 along with a practical performance to be expected based on battery developments as of 2012. Experience has shown that in most cases, the operating voltage of the cell is reasonably close to the theoretical value, but the energy density (Wh/kg and Wh/L) are much less than the theoretical value by a factor at least 2–3. Table 2 indicates the chemistries that are the prime candidates for the development of advanced batteries. These candidates are lithium–sulfur, zinc–air, magnesium–air, and lithium–air. In addition, various combinations of lithium chemistries show promise of considerably higher energy density than that in presently available cells even if it is not likely they can approach the 500–600 Wh/kg needed for the 300-mile range EVs.

The metal–air chemistries have the highest theoretical potential, but the reduction factors to practical designs are larger than that for the other chemistries because the weight of the positive air electrode is not included in the theoretical calculations. The active material (air) for the positive electrode comes from the environment and

Table 2. Performance characteristics of various battery chemistries.

Battery Type	Theoretical			Practical (2011)		
	Maximum Voltage (V)	Charge Capacity (Ah/g)	Energy Density (Wh/kg)	Nominal Voltage (V)	Energy Density (Wh/kg)	Energy Density (Wh/L)
Lead–acid	2.1	0.12	252	2.0	35	100
Nickel cadmium	1.35	0.181	244	1.2	40	90
Nickel metal hydride	1.35	0.178	240	1.2	80	220
Lithium–nickel cobalt	4.1	0.109	448	3.8	200	420
Lithium manganese	3.5	0.122	426	3.0	150	400
Lithium iron phosphate	3.65	0.111	405	3.4	115	255
Lithium titanate oxide	2.8	0.090	252	2.5	75	150
Lithium–sulfur	2.5	1.16	2900	2.15	400	365
Sodium metal chloride	2.6	0.22	572	2.5	120	190
Zinc–air	1.6	0.82	1312	1.2	400	900
Magnesium–Air	3.1	2.2	6820	1.4	800	1390
Lithium–air	3.4	3.68	13124	2.7	2500	3750

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

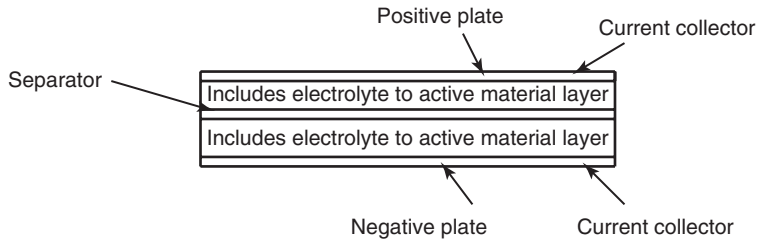


Figure 1. Cell schematic.

not from the battery itself. Regardless, the practical energy densities of the metal–air batteries are projected to be much higher than that of the other battery chemistries. Those battery types are presently in an early stage of development and their future performance is uncertain.

The steady improvement in the energy density of batteries for PEVs over the last 40 years (1970–2010) is evident in Table 2. In the 1970s, most of the EVs available used lead–acid batteries and had a useable range of <80 km (50 miles) and acceleration capability less than that required to follow the Federal Urban Driving Schedule (FUDS). The energy density of the lead–acid batteries (on a cell basis) was about 30 Wh/kg. The next generation of electric vehicles in the 1990s used nickel metal hydride batteries that had an energy density of 70–80 Wh/kg and a usable range of 120–160 km (80–100 miles). These electric vehicles had sufficient power to follow the FUDS driving cycle. The present generation (2010 and beyond) of electric vehicles use lithium-ion batteries and can have ranges of 160–240 km (100–150 miles). The lithium batteries have an energy density of 150–200 Wh/kg. The present generation of electric vehicles has excellent acceleration characteristics that can exceed that of conventional ICE vehicles. To a significant extent, both the range and acceleration characteristics of the EVs being marketed in 2012 are limited not by the performance of available batteries but by their high cost (>500$/kWh). Hence, the information in Table 2 show the steady progress in battery technology for EVs since 1970, and the projected development of the advanced batteries would only continue that trend, leading to longer-range vehicles.

3.2 Development of near-term advanced lithium-ion batteries

By near-term batteries are meant advanced batteries that are likely to being marketed for vehicle applications in 5–10 years. The chemistry of these batteries will evolve from the presently available lithium batteries and likely have energy densities <300 Wh/kg for cells.

There is much research (Yoshio, Brodd, and Kozawa, 2009; Reddy, 2011a; Schalkwijk and Scrosati, 2002; Thackeray *et al*., 2007; Marom *et al*., 2011) currently underway to increase the energy density of lithium-ion batteries. The research is being done in three areas: (i) increase the specific charge capacity (mAh/g) of the cathode (positive electrode), (ii) increase the specific charge capacity of the anode (negative electrode), and (iii) increase the cell voltage. The performance of the cells must be improved maintaining high cycle life and sufficient power capability for vehicle applications.

A schematic of a cell is shown in Figure 1. The specific charge capacity of the electrodes depends primarily on the materials used and the cell voltage depends primarily on the electrolyte utilized. Hence, the research cited above (Yoshio, Brodd, and Kozawa, 2009; Reddy, 2011a; Schalkwijk and Scrosati, 2002; Thackeray *et al*., 2007; Marom *et al*., 2011) is concerned with the electrochemistry of electrode materials and organic electrolytes for reversible intercalation of lithium ions in those materials. The objectives of the research are to find materials that have significantly higher mAh/g capacity than graphite currently used in the anode and the combinations of metal oxides used in the cathodes in presently available lithium batteries. The cell voltage depends both on the characteristics of the electrolytes and their compatibility with the electrode materials. Presently, cell voltages for lithium batteries vary between 3 and 4 V. Present research is targeting cells with voltages up to 5 V.

A summary of the characteristics of advanced electrode materials being studied is given in Table 3. For each electrode, the general approach is to prepare material composites that have specific charge capacities (mAh/g) that are significantly higher than that in the present baseline. In the case of the anode, the composites consist of mixtures of graphite/carbon and silicones. For the cathodes, the composites consist of layered atomic structures of various metal oxides—Mn, Ni, Co, and Al. As indicated in Table 3, present lithium cells utilize materials that have charge capacity values of 100–150 mAh/g, and advanced materials having charge capacities up to 250 mAh/g are being tested

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Properties of electrode materials for advanced lithium cells.

Electrode Material	Charge Capacity (mAh/g) Reversible	Density (g/cc)	Average Voltage	Resistivity (Ohm-cm)
Cathode				
Lithium Manganese Oxide (LMO)	100	5.0	4.0	20
Lithium Cobalt Oxide (LCO)	150	5.0	3.97	10
Nickel Manganese Cobalt (NMC)	160	4.8	3.8	50
NCA	180	4.8	3.8	50
Layered–layered NMC	210	4.4	3.85	50
Layered–layered NMC	230	4.4	3.48	50
Anode				
Graphite	330	1.7	0.1	0.01
Carbon–silicone (39%)	1050	1.19	0.60	0.01
Carbon–silicone (26%)	745	1.09	0.40	0.01
Electrolytes				
EC–EMC		1.16		103
PC-DEC (propylene carbonate-diethyl carbonate)		1.09		137

in the laboratory. In the case of the anodes, most cells presently use graphites having specific charge capacities of 330–350 mAh/g. As indicated above, the most promising of the new anode materials appear to be mixtures of carbon and silicone (Si/C). Research to date seems to indicate that mass fractions of silcone of 25–40% are possible consistent with reasonable cycle life. In that case, the mAh/g of the anodes would be in the range of 800–1000 (Table 3).

The key considerations in developing an electrolyte for lithium batteries are the following: (i) the extent to which it forms an SEI (solid electrolyte interface) layer on the anode and cathode materials and whether that layer is stable for several thousand charge/discharge cycles, ii) the voltage at which the cell using the advanced electrode materials is stable, (iii) high solubility of the lithium salts and high conductivity of the lithium ion, and (iv) stable cell operation at high temperatures (until at least 65–70°C). Most lithium cells having graphite anodes use 50 : 50 mixtures of EC–EMC and 1 mol LiPF6 as the lithium salt. The ionic conductivity of the electrolyte is about 10 mS/cm and the maximum cell voltage is 4.0–4.2 V. Considerable research effort (Zhang, Zhang, and Amine, 2011a; Horino *et al*., 2010; Smart *et al*., 2011; Hassoun *et al*., 2010) is underway to increase the cell voltage to 4.4–4.6 V and later to 4.8–5.0 V. Much of this research is directed to the development of additives that will increase the stability of the SEI layer at the higher voltages.

It is not a simple matter to translate the electrode material properties into projected values for the energy density and power characteristics of the advanced lithium cells. Most of the available references (Yoshio, Brodd, and Kozawa, 2009; Reddy, 2011a; Schalkwijk and Scrosati, 2002; Thackeray *et al*., 2007; Marom *et al*., 2011; Zhang, Zhang, and Amine, 2011a; Horino *et al*., 2010; Smart *et al*., 2011; Hassoun *et al*., 2010) deal with the characteristics of the electrodes alone. However, the presentations shown in Nelson *et al*. (2002) and Barnett (2011) compare the performance of cells using the advanced materials to a baseline cell using graphite and LCO. Further comparisons can be made by estimating the energy density and resistance of the cells for the various material combinations as follows.

The charge stored in the electrodes can be calculated from the mAh/g characteristic of the material and its density (ρ), porosity (ε), and thickness (δ).

$$(\text{mAh})_+ = (\text{mAh/g})_+ \rho_+ \delta_+ (1 - \varepsilon_+)$$

$$(\text{mAh})_- = (\text{mAh/g})_- \rho_- \delta_- (1 - \varepsilon_-)$$

The weight of the electrodes is given by

$$W_+ = 2.7\, t_+ + \rho_+ \delta_+ (1 - \varepsilon_+) + \rho_+ \delta_+ \varepsilon_+ \rho_{\text{electrolyte}}$$

$$W_- = 8.9\, t_- + \rho_- \delta_- (1 - \varepsilon_-) + \rho_- \delta_- \varepsilon_- \rho_{\text{electrolyte}}$$

t_+ and t_- are the thicknesses of the current collectors, which are copper foil for the negative and aluminum for the positive electrodes, respectively. $\rho_{\text{electrolyte}}$ is the density of the electrolyte.

The thicknesses of the negative and positive electrodes are related by the relative charge storage capacities of the electrodes.

$$\text{NPR} = (\text{mAh})_-/(\text{mAh})_+, \text{NPR} = \text{negative–positive ratio}$$

In most cases, NPR is slightly >1 meaning that the cell charge capacity is limited by that of the positive electrode. This is done to prevent overcharging of the negative electrode (plating of lithium). Hence, in the analysis, the thickness of the positive electrode will be specified and the thickness of the negative will be calculated. Hence, the energy storage (Wh) of the cell is given by

$$\text{Wh} = (\text{mAh})_+ V_{\text{cell, av}}/1000$$

$$V_{\text{cell, av}} = V_{\text{cell, av+}} - V_{\text{cell, av–}}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and the energy density

$$\text{Wh/kg} = [(\text{mAh})_{+} V_{\text{cell, av}}]/(W_{-} + W_{+}) \quad (1)$$

The cell resistance is difficult to estimate in a simple way, because it includes the ionic resistance of the porous electrodes, the contact resistances and the effective resistances of the current collection, the electronic resistance of the electrode material, and the solid diffusion resistance of the lithium ions into the nanoparticles of the electrode material. It is possible to calculate rather simply the ionic resistance of the electrolyte in the porous electrodes ($R_{\text{electrolyte}}$) and the electronic resistance of the electrode material ($R_{\text{elec.mat}}$), but estimating the other resistances is complex and requires detailed models of the cell.

$$R_{\text{electrolyte}} = \left[\left(\delta_{+}\varepsilon_{+}{}^{-1.5} + \delta_{-}\varepsilon_{-}{}^{-1.5}\right)/2 + \delta_{\text{sep}}\varepsilon_{\text{sep}}{}^{-1.5}\right] \kappa_{\text{electrolyte}}$$

$$R_{\text{elec.mat}} = \left[\delta_{+}\kappa_{\text{elec.mat}+}\left(1-\varepsilon_{+}\right)^{-1.5} + \delta_{-}\kappa_{\text{elec.mat}-}\left(1-\varepsilon_{-}\right)^{-1.5}\right]/2$$

The total resistance is given by

$$R_{\text{cell}} = R_{\text{electrolyte}} + R_{\text{elec.mat}} + R_{\text{contact}} + R_{\text{collect}} + R_{\text{particlediffus}}$$

An estimation of an upper bound for the power capability of the cell can be made using only the resistance R_{mat} of the electrode materials

$$R_{\text{mat}} = R_{\text{electrolyte}} + R_{\text{elec.mat}}$$

The pulse power capability of the cell can be calculated from the relationship

$$P = \text{EF}\ (1 - \text{EF})\ V^{2}{}_{\text{cell, av}}/R_{\text{mat}} \quad (2)$$

where EF is the efficiency of the pulse (Burke and Miller, 2011).

The above equations were incorporated into a spreadsheet model in which the material characteristics could be easily varied. The model combined with the material characteristics shown in Table 3 was then used to estimate the energy density and power capability of cells using the various material technologies. The following assumptions were made concerning the thickness of the various cell components:

Electrodes $\delta_{+} = 30$, 50, and 75 μ, $\varepsilon = 0.25$
Separator $\delta_{\text{sep}} = 20\,\mu$, $\varepsilon_{\text{sep}} = 0.40$

Current collector foil 25 μ (active material coating on both sides).

The results of the cell calculations for energy density and power capability are shown in Table 4. The values shown in the table do not include packaging weight and volume.

The first results in Table 4 are for lithium cells presently available (2010) and represent the baseline for the cells using the advanced materials. Values are given for cathode thicknesses of 30, 50, and 75 μ. Depending on the cathode material and the electrode thickness, the baseline cells have energy densities of 150–250 Wh/kg and 350–650 Wh/L. These cell performance values are consistent with test data (Burke and Miller, 2009a, b).

If graphite anodes are combined with layered–layered composite metal oxide cathodes, significant increases in energy densities are projected—240–320 Wh/kg and 500–710 Wh/L. The comparisons with the baseline cells should be made at the same electrode thickness. The results indicate that the cells with the composite layered cathodes are likely to have lower power capability than the baseline cells.

The next set of results shown in Table 4 is for the silicone–carbon anodes combined with cathodes using the various materials. The results indicate that the use of the silicone–carbon anodes results in very little change in cell performance for all the electrode thicknesses and cathode materials compared to the corresponding performance using a graphite anode. There also seems to be very small change in the power potential using the silicone–carbon anodes. These surprising comparisons likely are owing to the higher voltage of the silicone–carbon anodes compared to the voltage of the graphite anodes.

One of the research targets of present research is to increase the voltage of the cathode. In the last set of results in Table 4, the cathode voltage was increased by 0.5 V. The benefit of increasing the cathode and thus the cell voltage is shown in the results for all the combinations of anode and cathode materials. For example, for a cell having a 50-μ thick cathode, the energy densities using graphite and a layered-composite cathode are 334 Wh/kg and 710 Wh/L compared to 289 Wh/kg and 620 Wh/L at the lower voltage. The power potential using the higher cathode voltage is also significantly higher.

The projected cell results using the various chemistries indicate that the largest improvements in cell performance will result from significant increases in cathode charge capacity mAh/g and increases in the cathode voltage. Using silicone–carbon anodes is not projected to have a significant improvement in cell performance compared to using graphite in the anode. Large differences in cell performance and power capability result from changes in the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

electrode thicknesses, and comparisons in cell performance for different electrode chemistries should be done at the same electrode thickness.

3.3 Metal–air battery potential and development

In this section, the design of metal–air batteries is discussed. As indicated in Table 2, the theoretical potential energy density of metal–air batteries is much greater than that of lithium batteries. For this reason, there is much interest in the development of these battery chemistries for vehicle applications. As discussed in Reddy (2011b), Kowalczk, Read, and Salomom (2007), (Dahn), Lee *et al*. (2011), Girishkumar *et al*. (2010), Kraytsberg and Ein-Eli (2011), Zhang *et al*. (2011b), metal–air battery technology is in a very early stage of development especially for chemistries suitable for vehicle applications that require electrically rechargeable and high power cells. More R & D has been done on primary (non-rechargeable) metal–air cells, especially Zn–air, but research is now underway on rechargeable cells, especially Li–air.

A cell schematic is shown in Figure 1. In the case of the metal–air cell, the negative (anode) electrode is the metal (lithium or zinc) and the positive (cathode) is an air electrode consisting of porous carbon with an added catalyst. The reactant (oxygen) at the cathode is taken from the ambient air and thus it does not add to the weight of the cell. This is one of the reasons that the metal–air batteries have such high theoretical energy densities. As discussed later, one of the reasons that development of the metal–air batteries has been so difficult for vehicles applications is that they require a bidirectional air cathode that must be capable of reasonable high rates for both charge and discharge of the cell. The development of the air electrode is much less difficult if it must function only in the discharge direction (chemical reduction of oxygen) as in a hydrogen fuel cell.

For vehicle applications, the most attractive metals are lithium (Li), zinc (Zn), and magnesium (Mg). A summary of selected characteristics (Reddy, 2011b) of these metals for metal–air cells are given in Table 5. Lithium is the most attractive of the metals, but all of them have good potential for developing high energy density batteries. In the case of lithium, neglecting the weight and volume of the air cathode, the energy densities (13124 Wh/kg and 7008 Wh/L) are comparable to those of gasoline (12,222 Wh/kg and 9533 Wh/L). However, as discussed in later paragraphs, the practical potential energy density of lithium–air batteries is much less than that of liquid fuels.

It is of interest to develop a simple method of evaluating the practical potential of metal–air batteries, which includes the weight and performance of the air cathode. This can be

Table 4. Performance characteristics of advanced lithium-ion batteries using various electrode technologies.

Cell Technology	Energy Density (Wh/kg)[a]	Wh/L	Power Density (kW/L)[b]
Anode/cathode			
Graphite/LMO	146/178/200	340/440/510	46/24/13
Graphite/LCO	197/237/264	450/560/640	36/18/9
Graphite/NMC	200/241/268	450/560/640	31/15/8
Graphite/layered–layered 210 Ah/g	241/289/321	570/630/710	27/13/7
Graphite/layered–layered 245 Ah/g	242/289/319	510/620/690	19/9/5
SiliconeC 745 Ah/g/LCO	202/249/282	450/580/680	39/20/11
SiliconeC 745 Ah/g/NMC	206/254/288	450/580/670	33/17/9
SiliconeC 745 Ah/g/layered–layered 245 Ah/g	258/317/359	530/660/750	21/10/6
SiliconeC 1050 Ah/g/layered–layered 245 Ah/g	254/316/360	550/700/820	25/13/7
graphite/layered–layered 245 Ah/g, $\Delta V = 0.5$ V	280/334/368	590/710/800	26/12/6
SiliconeC 745 Ah/g/layered–layered 245 Ah/g, $\Delta V = 0.5$ V	302/371/419	620/770/870	30/14/8
SiliconeC 1050 Ah/g/layered–layered 245 Ah/g, $\Delta V = 0.5$ V	300/373/425	650/830/970	35/18/10

[a] a/b/c correspond to cathode electrode thicknesses of 30, 50, and 75 μ.
[b] The power density corresponds to ionic and electronic resistance of the electrode material only and does not include resistance to particles. The power density shown is for an 85% efficient pulse (Equation 1).

Table 5. Summary of the characteristics of the metals in a metal–air cell.

Metal	Atomic Weight (g)	Density (g/cm^3)	Charge Density (Ah/g)	Charge Density (Ah/cm^3)	Valence Change	Cell Voltage with Air Electrode (V)	Theoretical Energy Density (kWh/kg)
Lithium	7	0.534	3.86	2.06	1	3.4	13
Magnesium	24	1.74	2.2	3.83	2	3.1	6.8
Zinc	65	7.1	0.82	5.82	2	1.6	1.3

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

done using an approach similar to that discussed previously for lithium-ion cells.

The energy stored (Wh/cm^2) in the cell is that resulting from the electrochemical oxidation of the metal at the anode and can be calculated from the charge capacity of the metal and the voltage of the cell. Hence,

$$(Wh/cm^2)_{theo.} = V_{cell}(Ah/g)_-\rho_-\delta_-(1-\varepsilon_-) \quad (3)$$

and the usable energy is given by

$$(Wh/cm^2) = UL_{factor}(Wh/cm^2)_{theo}$$

where UL_{factor} is the fraction of the metal in the anode that is utilized. The theoretical capacity of the cell is the capacity if all the metal in the anode could be utilized in the discharge. However, only a fraction (50–80%) can be utilized due to pore blockage in the porous metal electrode.

An important parameter for any battery for vehicle applications is its power-to-energy ratio (N_{per}), which can be defined as

$$N_{per} = P/E = W/cm^2/(Wh/cm^2)$$

W/cm^2 can be expressed as

$$(W/cm^2)_{cell} = (A/cm^2)\ (V_- + V_+) = N_{per}(Wh/cm^2) \quad (4)$$

Since the A/cm^2 is equal for the negative and positive electrodes, we can express the A/cm^2 for the cell in terms of the value for the air cathode. Hence,

$$(A/cm^2) = (A/g)_+\rho_+\delta_+(1-\varepsilon_+) \quad (5)$$

Now using Equations (3) and (4),

$$(A/g)_+\rho_+\delta_+(1-\varepsilon_+) = N_{per}(Wh/cm^2)/(V_- + V_+)$$

Using Equation (2) and solving for the ratio of the electrode thicknesses,

$$\delta_+/\delta_- = (N_{per})\left(\frac{\rho_-(Ah/g)_-UL_{factor}}{[\rho_+(A/g)_+]}\right)\frac{(1-\varepsilon_-)}{(1-\varepsilon_+)} \quad (6)$$

$(A/g)_+$ is the current density at the air cathode and strongly influences the overpotential $(V_{ovp})_+$ at the cathode. The voltage at the cathode V_+ is given by

$$V_+ = (V_{equil})_+ + (V_{ovp})$$

The open-circuit voltage is given by

$$V_{oc} = (V_{eq})_- + (V_{eq.})_+$$

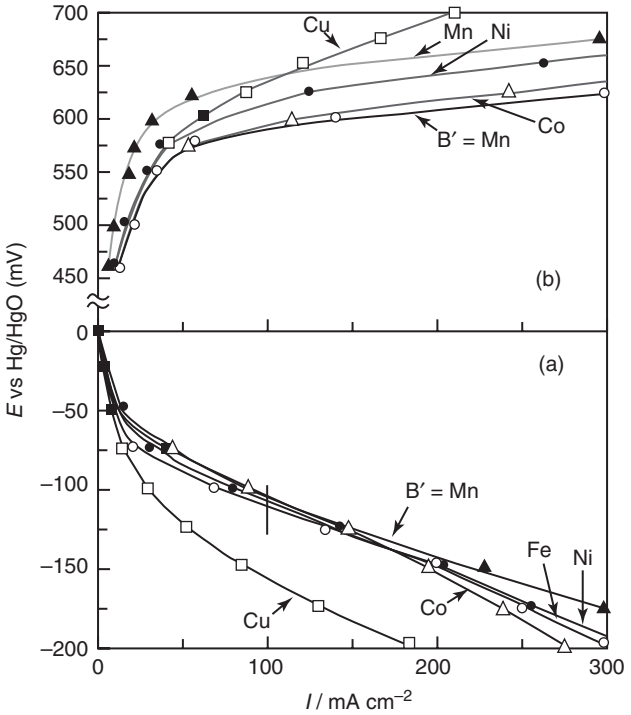


Figure 2. Overvoltage data for the air cathode. (Reproduced from Burke and Miller (2011). © Elsevier.)

The overpotential at the cathode is negative for discharge currents and positive for charging currents for the cell. The magnitude of the overpotential for the air electrode can be large (hundreds of millivolts) for both discharge and charge currents, but particularly large for charge currents. Research (Martin *et al*., 2009; Neburchilov *et al*., 2010; Yuasa *et al*., 2011; Wei *et al*., 2000) is presently being done to decrease the overvoltage using catalysts to permit higher values of A/g in the air cathode electrode (Figure 2).

The energy density of the cell requires calculation of the cell weight and volume. The weight of the cell is given by

$$(Wt/cm^2)_{cell} = (Wt/cm^2)_- + (Wt/cm^2)_+ + (Wt/cm^2)_{sep} + (Wt/cm^2)_{electrolyte}$$

The weights of the electrodes depend on their thickness and effective density including porosity. Hence,

$$(Wt/cm^2)_{cell} = \delta_-\rho_{metal}(1-\varepsilon_-) + \delta_+\rho_{+C\&cat}(1-\varepsilon_+) + \delta_{sep}\rho_{sep} + \delta_-\rho_{electrol.}\varepsilon_- + \delta_+\rho_{electrol.}\varepsilon_+ + \delta_{sep}\rho_{electrol.}\varepsilon_{sep} + \delta_{cc}\rho_{cc} \quad (7)$$

where ε is the porosity of the various cell components.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The thickness of the cell is simply

$$\delta_{\text{cell}} = \delta_{-} + \delta_{+} + \delta_{\text{sep}} + \delta_{\text{cc}} \tag{8}$$

The energy stored in the cell is given by

$$(\text{Wh/cm}^2) = V_{\text{cell}}(\text{Ah/g})_{-}\delta_{-}\rho_{\text{metal}}(1-\varepsilon_{-})\ \text{UL}_{\text{f canactor}} \tag{9}$$

The energy densities are given by

$$(\text{Wh/kg}) = (\text{Wh/cm}^2)/(\text{Wt/cm}^2)_{\text{cell}} \tag{10}$$

$$(\text{Wh/L}) = (\text{Wh/cm}^2)/\delta_{\text{cell}}$$

In calculating the cell energy densities for a specific metal–air chemistry, the key design parameters are $N_{\text{per.}}$, $\text{UL}_{\text{factor}}$, δ_{-}, component porosities ε, and the A/g, the current density at the air cathode. The discharge time of the cell at the maximum steady power is simply $1/N_{\text{per.}}$

The steady power capability of the cell is given by either

$$(\text{W/cm}^2) = N_{\text{per}}(\text{Wh/cm}^2)$$

or

$$(\text{W/cm}^2) = V_{\text{cell}}(\text{A/g})_{+}\delta_{+}\rho_{+}(1-\varepsilon_{+}) \tag{11}$$

The power densities are given by

$$(\text{W/kg}) = (\text{W/cm}^2)/\text{Wt/cm}^2$$

$$(\text{W/L}) = (\text{W/cm}^2)/\delta_{\text{cell}}$$

The steady power capability of the cell is proportional to A/g at the air cathode. The pulse power of the cell will be dependent on the resistance of the cell for short pulses of 5–10 s.

$$P = \text{EF}(1-\text{EF})\ V^2/R_{\text{cell}} \tag{12}$$

where EF is the efficiency of the pulse.

The voltage drop across the cell is the sum of the voltage drops across the anode and cathode electrodes and the separator. Hence,

$$V_{\text{oc}} - V = (\text{Ovp})_{\text{anode}} + \left[\left(\delta_{+}\varepsilon_{+}{}^{-1.5} + \delta_{-}\varepsilon_{-}{}^{-1.5}\right)/2 + \delta_{\text{sep}}\varepsilon_{\text{sep}}{}^{-1.5}\right]\kappa_{\text{electrolyte}}\left(\frac{\text{A}}{\text{cm}^2}\right) + (\text{Ovp})_{\text{aircathode}} \tag{13}$$

The effective resistance R_{cell} is given by

$$R_{\text{cell}} = V_{\text{oc}} - V/\text{A/cm}^2(\text{Ohm}-\text{cm}^2)$$

$$V_{\text{cell}} = V_{\text{oc}} - R_{\text{cell}}(\text{A/cm}^2) \tag{14}$$

Overpotential data for both discharge and charge currents are show in Figure 2 for various catalysts (Martin *et al*., 2009). The V_{cell} used in Equations (8) and (10) for the energy and power densities should include the cell resistance given in Equation (13).

This model can be used to project the performance characteristics of metal–air batteries. This will be done for Zn–air and Li–air batteries in the following sections. A summary of the input parameters for the metal–air cell calculations is given in Table 6.

3.4 Zn–air batteries

As indicated in Table 5, the potential energy density of the Zn–air battery is quite high (1300 Wh/kg), but the practical energy density is much lower. In this section, the previously discussed modeling equations are used to estimate the energy density and power capability that can be expected in practical designs. The electrochemistry of the electrodes in the Zn–air battery is quite complex, but it can be simplified as follows:

Zn anode:

$$\text{Zn} + 4\text{OH}^{-} \rightarrow \text{Zn(OH)}_4{}^{-2} + 2\text{e} \qquad V_{\text{eq.}} = 1.22 \tag{15}$$

$$\text{Zn(OH)}_4{}^{-2} \rightarrow \text{ZnO} + \text{H}_2\text{O} + 2\text{OH}^{-}$$

Air cathode:

$$1/2\text{O}_2 + \text{H}_2\text{O} \rightarrow 2\text{OH}^{-} \qquad V_{\text{eq.}} = 0.40 \tag{16}$$

Overall reaction:

$$\text{Zn} + {}^{1}/_{2}\,\text{O}_2 \rightarrow \text{ZnO} \qquad V_{\text{eq.}} = 1.62 \tag{17}$$

The electrolyte used in the Zn–air cell is KOH. The cell geometry is illustrated in Figure 1. The reactions shown are for the discharge of the cell. The reactions proceed in the reverse directions for charging the cell. Note that the oxygen reacted at the air electrode is taken from the ambient air. The construction and functioning of the air electrode (cathode) is complex (Reddy, 2011a; Schalkwijk and Scrosati, 2002; Thackeray *et al*., 2007; Marom *et al*., 2011; Zhang, Zhang, and Amine, 2011a; Horino *et al*., 2010; Smart *et al*., 2011; Hassoun *et al*., 2010; Nelson *et al*., 2002; Barnett, 2011; Burke and Miller, 2011; Burke and Miller, 2009a, b; Reddy, 2011b; Kowalczk, Read, and Salomom, 2007; Dahn; Lee *et al*., 2011; Girishkumar *et al*., 2010; Kraytsberg and Ein-Eli, 2011; Zhang *et al*., 2011b; Martin *et al*., 2009; Neburchilov *et al*., 2010; Yuasa *et al*., 2011; Wei *et al*., 2000). It is particularly difficult to develop the bidirectional air electrode needed for vehicle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 6. Summary of the input parameters for the metal–air cell calculations.

Metal-Air battery											
Cell input options	Electrode Material			Charge capacity Ah/g theoretical (Anode)	Density (g/cc)	Porosity*	Equilibrium voltage (V)	Overpotential (V)	Current density A/g (Cathode)	Resistivity (Ohm-cm)	Thickness (um)
	Cathode	Air cathode	Carbon + Catalyst	—	1.9	0.8	0.4	0.075	3	0.01	200
				—	1.9	0.8	0.4	0.11	6	0.01	150
				—	1.9	0.8	0.4	0.15	12	0.01	200
				—	1.9	0.8	0.4	0.2	18	0.01	200
		Current collector	Current collector 1	—	1.7	0.5	—	—	—	1.00E-02	0
	Anode	Metal anode	Zinc	0.82	7.1	0.5	1.22	0.1	—	1.00E-04	—
			Magnesium	2.2	1.74	0.5	2.7		—	1.00E-04	—
			Lithium	3.86	0.534	0	3	0.1	—	1.00E-04	—
		Protective layer*	No protective layer	—	0	0	—	—	—	0	0
			Protective layer	—	2.6	0.8	—	—	—	50	100
		Current collector	Current collector 1	—	8.9	0	—	—	—	1.00E-06	0
	Electrolyte		KOH/Zn	—	1.2	—	—	—	—	2.2	—
			NaCl/Mg	—		—	—	—	—		—
			LiCl(1M)/Li	—	1.1	—	—	—	—	16	—
	Separator		Separator 1	—	0.2	0.5	—	—	—	2	100

applications in which high power in charging is required for regenerative braking and battery recharging in an hour or less is desirable. High power air electrodes operating in the discharge direction have been developed for use in fuel cell applications.

Equations (1–14) can be used to prepare an EXCEL spreadsheet to analyze the cell energy density and power capability. The primary design inputs are the following:

Cell power-to-energy ratio N_{per}
Thickness of the air cathode δ_+
Current density of the air cathode A/g
Utilization factor of the anode UL_{factor}

Note that $1/N_{per}$ is the discharge time of the cell at the current density A/g. The voltages of the anode and cathode depend on the equilibrium voltage $V_{eq.}$ and the overpotentials η of electrodes. The voltage at an electrode is given by

$$V_{electrode} = V_{equil.} - \eta$$

The overpotentials depend on the current density and can be obtained from detailed modeling of the electrodes (White, 2005; Mao and White, 1992) or from test data (Martin *et al*., 2009; Neburchilov *et al*., 2010; Yuasa *et al*., 2011; Wei *et al*., 2000). There is considerable uncertainty concerning the characteristics of the air cathode because the overpotential depends critically on the catalyst being used. The electrode inputs used in the present analysis are given in Table 6.

For a given set of design inputs and electrode current A/g, the ratio of the electrode thicknesses and then the thickness of the anode (metal electrode) can be calculated from Equation (6). The current density (A/cm^2) and energy stored (Wh/cm^2) follow from Equations (5) and (9). The weight of the active materials in the cell and its thickness are given by Equations (7) and (8). The Wh/kg and Wh/L follow from Equation (10). The steady power density (W/kg and W/L) are calculated using Equation (10). The cell voltage, resistance, and the pulse power characteristics of the cell can be calculated using Equations (12–14) for various current densities (A/cm^2). These calculations will include charging currents (oxygen evolution) and will be an indication of the ability of the cell for regenerative braking and recharge. The charging efficiency is given by

$$EF_{charging} = (V - V_{oc})_{charging}/V_{oc}$$

The spreadsheet model was utilized to obtain the results shown in Table 7 for number of combinations of input parameters. The cell design parameters that varied in the calculations were the thickness and current density (A/g) of the air cathode and the discharge time ($1/N_{per}$) of the cell. Discharge times of 3, 5, and 10 h were considered. These are short discharge times for zinc–air batteries, but reasonable for batteries intended for vehicle applications. The results in Table 7 indicate that cells with energy densities in the range of 500–600 Wh/kg

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 7. Summary of calculated characteristics of the zinc–air cells.

Current Density, Air Cathode (A/g)	Discharge Time (h)	Cathode Thickness (mm)	Anode Thickness (mm)	V_{cell} (V)	A/cm^2	Energy Density (Wh/kg)[a]	Wh/L	Power Density, Steady (W/kg)	Power Density, Pulse 75% (W/kg)
18	10	0.2	6.7	1.02	0.137	495	1992	49	40
18	5	0.2	3.3	1.16	0.137	558	2178	126	103
18	3	0.2	2.0	1.22	0.137	576	2167	188	191
6	10	0.2	2.2	1.37	0.046	650	2468	94	65
6	5	0.2	1.2	1.39	0.046	636	2233	195	126
6	3	0.2	0.7	1.40	0.046	612	1972	317	199
6	10	0.15	1.7	1.40	0.034	656	2464	65	100
6	5	0.15	0.84	1.40	0.034	637	2195	127	202
6	3	0.15	0.50	1.40	0.034	610	1915	198	323

[a] $UL_{factor} = 0.65$.

and 2100–2400 Wh/L (not including packaging) could be possible for zinc–air batteries. The power capability of the cells is relatively low being in the range of 200–300 W/kg for cells with a discharge time of 3–5 h. The calculations indicate that an air cathode thickness of about 0.15 mm seems to result in reasonable anode thicknesses. A zinc utilization of 65–70% was assumed in the calculations based on the detailed modeling results given in White (2005) and Mao and White (1992).

3.5 Lithium–air batteries

As indicated in Table 5, the theoretical energy density of lithium–air batteries is the highest of the various metal–air chemistries. For that reason, there is considerable research underway (Girishkumar *et al*., 2010; Kraytsberg and Ein-Eli, 2011; Zhang *et al*., 2011b) to understand the science needed to develop lithium–air batteries for vehicle applications. The lithium–air cell utilizes lithium metal at the anode and a bifunctional air cathode. Uniform deposition of the lithium at the anode during recharge is critical to the safe and long cycle life operation of the cell. Cells can incorporate either an aqueous or an organic electrolyte.

The electrochemistry of the battery is relatively simple. In the case of cells using an organic (nonaqueous) electrolyte, the reactions for cell discharge are

Anode:

$$Li \rightarrow Li^+ + e$$

Cathode:

$$O_2 + 2Li^+ + 2e \rightarrow Li_2O_2 \qquad V_{eq.} = 3.1$$

$$O_2 + 4Li^+ + 4e \rightarrow 2Li_2O \qquad V_{eq.} = 2.91$$

The dominate reaction yields Li_2O_2, which is not soluble in the organic electrolyte. The reaction to form Li_2O_2 takes place in a porous carbon electrode with a catalyst and can be reversible at low rates. The open-circuit voltage of this cell is about 3 V. A thin protective layer is applied to the anode to prevent the formation of a barrier layer that impedes diffusion of the lithium ions. This and other challenges seem to have led to minimal work to develop rechargeable lithium–air cells using organic electrolytes. The reactions in charging are the reverse of those shown for discharge. It is important that the reaction rates are reasonably fast in both directions.

In the case of lithium–air cells using aqueous electrolytes, the reactions are

Anode:

$$Li \rightarrow Li^+ + e \qquad V_{eq.} = 3.04 \tag{15}$$

Cathode:

$$O_2 + 2H_2O + 4e \rightarrow 4OH^- \qquad V_{eq.} = 0.40 \tag{16}$$

$$4Li^+ + 4OH^- \rightarrow LiOH$$

Overall reaction:

$$4Li + O_2 + 2\,H_2O \rightarrow 4\,LiOH \qquad V_{eq.} = 3.44 \tag{17}$$

The LiOH is soluble in the aqueous electrolyte and thus does not clog the pores of the air cathode. The cathode reaction is reversible, so it is possible to develop electrically rechargeable cells.

Since lithium reacts vigorously in water, it is necessary to apply a protective film on the lithium metal anode. In addition, a buffer layer is needed between the lithium foil and the protective film to minimize the formation of dendrites during recharging of the cell. The protective film is a lithium-conducting glass ceramic and the buffer layer can be a solid material or a polymer electrolyte. With these protective layers, the lithium metal anode can be stable in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 8. Summary of calculated characteristics of the water-stable lithium–air cells.

Current Density, Air Cathode (A/g).	Protect Layer (Ohm-cm^2)	Discharge Time (h)	Cathode Thickness (mm)	Anode Thickness (mm)	V_{cell} (V)	A/cm^2	Energy Density (Wh/kg)[a]	Wh/L	Power Density Steady (W/kg)	Power Density Pulse 75% (W/kg)	Ohm-cm^2 Cell
6	20	3	0.2	1.1	2.48	0.034	2021	1758	600	574	23
6	6	3	0.2	1.1	2.96	0.034	2411	2098	716	1194	12
6	1	3	0.2	1.1	3.05	0.034	3066	2315	926	1908	10
6	1	5	0.2	1.6	3.06	0.034	3650	2540	738	1515	10
6	6	5	0.2	1.6	2.96	0.034	2986	2341	602	986	12
6	20	5	0.2	1.6	2.48	0.034	2502	1962	498	474	27
12	6	3	0.2	2.9	2.5	0.091	3134	2189	940	919	10
12	6	5	0.2	4.4	2.5	0.091	3558	2310	695	710	10
18	6	3	0.2	4.4	2.2	0.137	3041	1976	919	737	9
18	6	5	0.2	6.6	2.2	0.137	3340	2052	668	539	9

[a] ULfactor = 0.5, $\delta_{sep} = 150\,\mu$

the aqueous electrolyte. However, the protective layers do have high ionic resistance, which must be greatly reduced in order for the lithium–air cells to have power capability suitable for vehicle applications. Hence, it can be expected that lithium–air batteries would have high energy density (a theoretical value of 2450 Wh/kg including oxygen), but relatively low power capability. The open-circuit voltage of the cell using an aqueous electrolyte is 3.84 V. The cell is stable at that high voltage because about 3 V occurs at the buffer layer leaving only about 0.8 V at the air cathode with the aqueous electrolyte.

Most lithium–air cells developed to date have not been electrically rechargeable (i.e., primary cells). However, it is possible that rechargeable cells can be developed. As discussed in connection with the zinc–air cells, bifunctional air electrodes with aqueous electrolytes have been the subject of active research for many years. The recharge rates (mA/cm^2) at the air cathode are dependent on the use of catalysts. Progress in the development of catalysts that will permit increased recharge rates is being made. Owing to the voltage drops at the anode protective layers and the air cathode, it seems likely that the charge efficiency will be low even for recharge times of many hours.

The projected performance of lithium–air cells using an aqueous electrolyte can be made using Equations (1–14) that were applied previously to develop a spreadsheet model for the Zn–air cells. A major difference between the two chemistries is the need for an anode protective layer in the case of the lithium–air cell. Input parameters for the lithium–air calculations are included in Table 6. In the case of the protective layer, it is assumed that further development of that technology will result in lower resistance (reduced Ohm-cm) and thinner layers than is presently possible. Otherwise, the power capability of the cells will not be suitable for vehicle applications. A summary of calculated characteristics of the lithium–air cells is given in Table 8.

Although the calculated cell performance does not include cell packaging, the results shown in Table 8 indicate that the water-stable lithium–air cells show good promise for vehicle applications with energy densities >2500 Wh/kg and 1500 Wh/L for packaged cells. The power capabilities seem reasonable and are sufficiently high that the lithium–air cells combined with high power lithium-ion batteries or ultracapacitors should meet the power requirements for vehicles.

4 TEST DATA AND THE STATUS OF ADVANCED BATTERIES OF VARIOUS BATTERY CHEMISTRIES

The design and performance of advanced batteries have been discussed in the previous section. The energy densities of the batteries have been projected based on available information on the electrode materials and electrolytes. In this section, test data are presented for lithium and metal–air cells. Unfortunately, the data available are very limited especially for the most advanced chemistries.

4.1 Lithium batteries

Test data are available for cells using the various lithium chemistries that have been commercialized. Test results (Burke and Miller, 2009a, b) are summarized in Table 9. The energy densities of the cells vary significantly between the various chemistries, the sizes (Ah), and power capabilities. The maximum energy density of the commercially available cells using graphite in the anode is 150–170 Wh/kg, which is somewhat lower than the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 9. Summary of the performance characteristics of lithium-ion batteries of various chemistries.

Battery Developer/ Cell type	Electrode Chemistry	Voltage Range	Ah	Resist. mOhm	Wh/kg	W/kg 90% Efficiency[a]	W/kg Match Imped.	Weight (kg)	Density gm/cm^3
Enerdel HEV	Graphite/Ni MnO_2	4.1–2.5	15	1.4	115	2010	6420	0.445	—
Enerdel EV/PHEV	Graphite/Ni MnO_2	4.1–2.5	15	2.7	127	1076	3494	0.424	—
Kokam prismatic	Graphite/$NiCoMnO_2$	4.1–3.2	30	1.5	140	1220	3388	0.787	2.4
Saft Cylind.	Graphite/NiCoAl	4.0–2.5	6.5	3.2	63	1225	3571	0.35	2.1
GAIA	Graphite/$NiCoMnO_2$	4.1–2.5	40	0.48	96	2063	5446	1.53	3.22
Cylind.			7	3.6	78	1313	3472	0.32	—
A123 Cylind.	Graphite/iron phosphate	3.6–2.0	2.2	12	90	1393	3857	0.07	2.2
Altairnano prismatic	LiTiO/$NiMnO_2$	2.8–1.5	11	2.2	70	990	2620	0.34	1.83
Altairnano prismatic	LiTiO/$NiMnO_2$	2.8–1.5	3.8	1.15	35	2460	6555	0.26	1.91
Quallion Cylind.	Graphite/NiCo	4.2–2.7	1.8	60	144	577	1550	0.043	2.6
Quallion Cylind.	Graphite/NiCo	4.2–2.7	2.3	72	170	445	1182	0.047	2.8
EIG prismatic	Graphite/$NiCoMnO_2$	4.2–3.0	20	3.1	165	1278	3147	0.41	—
EIG prismatic	Graphite/iron phosphate.	3.65–2.0	15	2.5	113	1100	3085	0.42	—
Panasonic EV prismatic	Ni metal hydride	7.2–5.4	6.5	11.4	46	395	1093	1.04	1.8

[a] Power density $P = \text{Eff.}(1-\text{Eff.})\ V_{oc}{}^2/R$, $P_{\text{match. imped.}} = V^2/4R$.

Table 10. Fast charging characteristics of lithium batteries using various chemistries.

Charge rate	Percent Ah to Max. Charge Voltage		
	Nickel cobalt manganese (NCM)	Iron phosphate	Lithium titanate
3C (20 min.)	81%	92%	99%
4C (15 min.)	76%	90%	98%
5C (12 min.)	72%	85%	96%
6C (10 min.)	—	78%	94%

projections given in Table 4, because the projection values do not include packaging of the cells. In addition, most of the available cells have been designed to provide high power. The power values (W/kg) given in Table 8 can be achieved in lithium batteries for both discharge and charge pulses. As indicated in Table 10, lithium batteries can be fast charged in 10–15 min. Hence, in all respects, available lithium batteries have power capabilities suitable for vehicle applications.

There is essentially no test data available for packaged lithium cells using the advanced layered–layered nano-metal clusters media (NMC) composites in the cathode and silicone–carbon anodes. A summary (Barnett, 2011) of data for some laboratory cells using the advanced materials is shown in Figure 3. The data indicate the wide range of specific capacity (mAh/g) and voltages for the various chemistries.

4.2 Metal–air batteries

Modeling results for the metal–air batteries are given in Tables 7 and 8. The cell projections were done assuming electrically rechargeable cells even when the bifunctional air cathode technology for reasonable recharge times has not yet been developed. For this reason, most of the commercial metal–air cells are not rechargeable (i.e., primary cells). Nevertheless, several examples of those cells will be given to indicate the present status of the metal–air battery technology.

4.3 Zinc–air

Zinc–air coin cells for hearing aids are manufactured in very large quantity around the world. These small cells (150–300 mAh) have an energy density of about 400 Wh/kg and maximum power rating of about 5–8 mW/g. They are not rechargeable and have very low power and thus the technology is not suitable for vehicle applications.

There has been considerable R & D to develop larger cells for use in cell phones and vehicles of various sizes from passenger cars to buses. Some of the development has been recent (Revolt portable battery-Technology Brief, n.d.) and others in the more distant past (Goldstein and Koretz, 1994; Schimpf, 1995; Cheiky, Danczyk, Wehrey, 1990) as an effort to develop advanced batteries for electric vehicles. However, these efforts have not yet resulted in commercial products.

Revolt Technology in Switzerland is developing a rechargeable cell for cell phones (Revolt portable battery-Technology Brief, n.d.). A photograph of a 10-Ah, 1.2-V cell is shown in Figure 4. The cell shown has energy densities of 450 Wh/kg and 1040 Wh/L. The power capability of the cell is about 200 W/kg, which is much lower (about a factor of 10) than required for vehicle applications.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

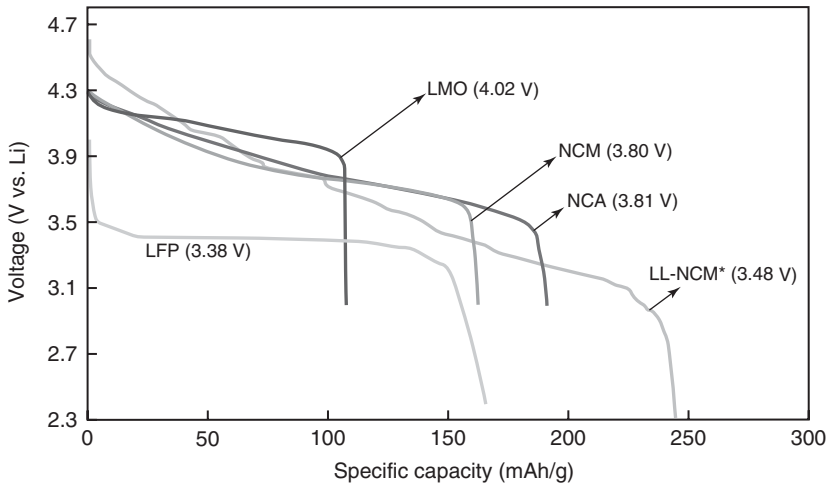


Figure 3. Discharge characteristics of lithium cells using various advanced cathode chemistries). (Reproduced from Barnett, 2011 © the Knowledge Foundation.)



Figure 4. A 10-Ah rechargeable, Zn–air cell Revolt technology.

As discussed in (Goldstein and Koretz, 1994; Schimpf, 1995; Cheiky, Danczyk, Wehrey, 1990), large Zn–air batteries have been assembled for testing in electric passenger cars, vans, and buses. Both mechanically rechargeable and electrically rechargeable systems were developed. The mechanically recharge units (battery switching) by *Electric Fuel* in Israel were large (110 kWh) and had an energy density of 230 Wh/kg and 230 Wh/L. The batteries were relatively low power with a power density of 100 W/kg. There was considerable in-vehicle testing of *Electric Fuel* batteries in vans and buses (Goldstein and Koretz, 1994). Electrically rechargeable cells were developed by AER Energy Resources Energy Resources (Schimpf, 1995) and Dreisbach Electromotive (DEMI) (Cheiky, Danczyk, Wehrey, 1990) in the 1990s. These batteries had energy densities of 180 Wh/kg and 150 Wh/L and pulse power of about 200 W/kg. Hence, the rechargeable Zn–air batteries of the 1990s had much lower energy density and power than the present lithium-ion batteries. It is uncertain (Table 7) whether the more recent developments in Zn–air batteries will result in batteries of high enough energy density and power capability for vehicle applications.

4.4 Lithium–air

In recent years, there have been many claims that lithium–air batteries can be the answer to the marketing of electric cars with ranges comparable to gasoline-fueled vehicles. In a recent news release (IBM, 2012), IBM claims a breakthrough that will permit the development of lithium–air batteries for a 500-mile range electric vehicle by 2020. This will require batteries with an energy density of about 2000 Wh/kg and adequate power capability to maintain the vehicle at 70–75 mph. In addition, the battery will require a recharge capability of several hours and a low cost and long calendar and cycle life of 10–15 years. As discussed earlier in this section, this will require technical breakthroughs in the areas of a low resistance protective film for the lithium metal anode and large advances in the performance of the bifunctional air cathode. Present technology in those areas is very much inadequate for the development of lithium–air batteries for vehicle applications.

The present status of lithium–air battery technology is given in Reddy (2011b), Girishkumar *et al*. (2010),

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

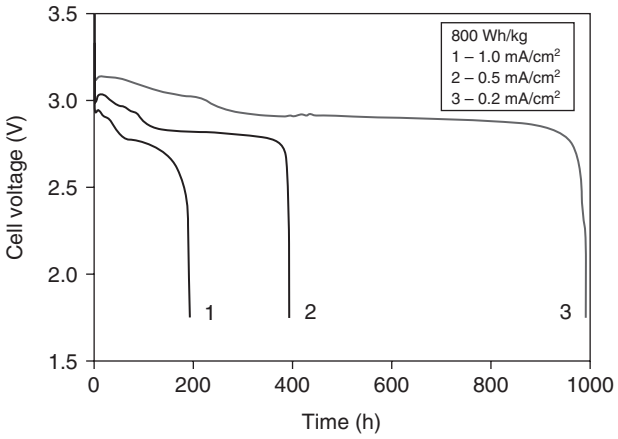


Figure 5. Lithium–air cell and discharge curves. (Reproduced with permission from T.B. Reddy, Linden's Handbook of Batteries (Reddy, 2011b) © The McGraw-Hill Companies, Inc.)

Kraytsberg and Ein-Eli (2011), Zhang *et al*. (2011b). An example of the present technology shown in Figure 5 is taken from Reddy (2011b). The energy density of the cell can be very high, but the discharge times are very long. Practical batteries for oceanic applications are being tested by *Polyplus* (Polyplus, 2009), but those batteries are not rechargeable and have discharge times of many months.

Projections of the performance of lithium–air batteries that could be suitable for vehicle applications are given in Table 8. These projections require technologies that do not presently exist, but much R & D is being done on lithium–air cell technologies that could result in advances making possible lithium–air batteries for vehicles.

5 VEHICLE CALCULATIONS USING ADVANCED BATTERIES

In this section, the application of the advanced batteries in PEVs will be considered.

5.1 Electric vehicles

The design parameter of primary interest for EVs is the range (miles). Using the high energy density advanced batteries, it is reasonable to consider ranges up to 300–500 miles. The battery for the EV is sized by the energy storage requirement (kWh), which depends on the energy use (Wh/mi) of the vehicle and the fraction of the energy stored that can be used on a regular basis. The battery pack weight and volume are design inputs and the required cell performance parameters (Wh/kg and Wh/L) depend on the packaging factors relating to cell and pack weight and volume. As discussed in Section 2, the vehicle energy use depends on vehicle weight and road load parameters and is in the range of 200–250 Wh/mi for most EV designs. The cell/battery packaging factors are highly dependent on battery pack design and are expected to improve markedly in future years from the present low values in the Leaf and Volt. It seems reasonable to assume values of 0.75 and 0.5 for the weight and volume packaging factors, respectively. Calculations of the resultant cell energy and power density requirements are shown in Table 11.

Comparison of these cell requirements with the cell performance projections for advanced lithium chemistries in Table 4 indicate that for well-designed battery packs and vehicles with reduced weight and road load, the advanced lithium chemistries may satisfy the cell requirements for up to 300-mile range, but not 400–500-mile range. The power density requirements are well within the capability of the advanced lithium batteries.

It appears that the metal–air batteries would be needed to extend the range of the electric vehicles beyond about 300 miles. The projected energy density of the zinc–air and Li–air batteries are given in Tables 7 and 8. The energy density of the metal–air batteries seems to be adequate for the long-range EVs, but there are questions concerning the adequacy of their power capability. It is likely necessary to combine a high power battery or ultracapacitor with the metal–air battery to meet the power requirements of the vehicles.

Table 11. Cell performance requirements for various range EVs.

Range (miles)	Energy Stored (kWh)	Pack Weight (kg)	Pack Volume (L)	Cell Weight (kg)	Cell Volume L	Cell Energy Density (Wh/kg)	Cell Energy Density (Wh/L)	Power (kW)	Cell Power Density (W/kg)
100	25	250	200	188	100	133	250	80	320
200	50	250	200	188	100	266	500	80	320
300	75	300	250	225	125	333	600	75	250
400	100	300	250	225	125	444	800	75	250
500	125	300	250	225	125	555	1000	75	250

Vehicle energy use 200 Wh/mi, usable energy fraction 0.8, weight packaging factor 0.75, volume packaging factor 0.5.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 12. Simulation results for the PHEV using zinc–air batteries.

Battery Type 1	Cycle	Range (miles)	Maximum Capacitor (kW)	Maximum Bat (kW)	Eff. Bat.	Maximum Capacity (kW)	Eff. Cap.	Wh/mi Bat.	Operation Mode	mpg 20 miles	mpg 40 miles	Mpg Charge Sustaining HEV
Batteries alone												
Rech. Zn–air	FUDS	66	30	30	0.84			139	Blended	139	137	39.4
32 kg bat	HW	63	20	20	0.83			156	Blended	169	169	41.1
	US06	93	36	36	0.72			101	Blended	48.5	48.5	30.1
With ultracapacitors												
Rech. Zn–air	FUDS	40	45	19	0.87	45	0.97	228	AE	None	None	54.5
32 kg bat	HW	38	45	19	0.81	45	0.97	242	AE	None	None	57.7
20 kg cap	US06	66	68	21	0.82	68	0.94	149	Blended	62.4	60	38.8

Weight of cells only.

5.2 Plug-in hybrid vehicles

At the present time, PHEVs such as the Volt are limited to about a 40-mile electric range in order to fit the battery into the vehicle without compromising its functionality. With the higher energy density of the advanced batteries, PHEVs will be possible with electric ranges of 50–75 miles. For 75-mile range, the battery pack would need to store about 19 kWh. If the energy densities of the cells are 300 Wh/kg and 650 Wh/L with the same packaging factors used in Table 11, the cell weight and volume for a 75-mile PHEV would be 63 kg and 29 L resulting in a battery pack weight and volume of 84 kg and 58 L. Hence, the battery pack in the PHEV would be much smaller than that in an electric vehicle. Assuming the same power rated electric drive in the PHEV as in the EV, the power requirements of the cells would be 1190 Wkg and 2586 W/L. These are high power requirements that may not be able to be met with the advanced lithium cells even using thin electrodes. If they cannot be met, the advanced battery can be combined with an ultracapacitor to provide the maximum power for acceleration and regenerative braking.

5.3 Combinations of advanced batteries and ultracapacitors

Computer simulations of PHEVs using the advanced batteries are presented in Burke and Miller (2010) and King *et al*. (2003). Simulation results are given for vehicle operation using the batteries alone and in combination with ultracapacitors. In a PHEV, if the batteries are power limited, the additional power is provided by the engine even when the vehicle is operating in the battery charge depletion mode. This type of operation is called the *blended* mode of operation. In the case of batteries combined with ultracapacitors, the ultracapacitors provide the peak power and the batteries provide the average power over the driving cycle. In addition, the ultracapacitors accept all the charge during regenerative braking. Results are given in Table 12 for a PHEV using a zinc–air battery and carbon/carbon ultracapacitors on various driving cycles.

All the simulations were performed using the following vehicle inputs:

$C_D = 0.27$, $A_F = 2.2\ m^2$, $f_r = 0.008$, test weight = 1650 kg (approx.)

Engine: Honda 1.3L, iVTEC engine map, scaled to 90 kW

Electric motor: Honda hybrid Civic AC PM 2006 efficiency map, scaled to 70 kW

DC/DC inverter: constant efficiency 0.96

Transmission: 5-speed manual (3.11, 2.11, 1.55, 1.0, 0.71, FD = 3.95), automatically shifted.

The characteristics of the battery and ultracapacitors are as follows:

Zinc–air batteries:

Zn–air 32 kg 60 Ah cells 180 in series 9.5 kWh useable
385 Wh/kg 156 W/kg, 95% effic., 616 W/kg, 75% effic.

Carbon/carbon ultracapacitors:

Symmetric C/C 20 kg 1350 F cells 110 in series
100 Wh usable
5.5 Wh/kg 2320 W/kg, 95% effic., 11.6 kW/kg, 75% effic.

The nominal energy storage unit voltage was 240 V (approximately) with the maximum currents limited to about 300 A even in the cases of the batteries alone. In all cases, the batteries were depleted to 30% state of charge (SOC) from 100% SOC in the charge depleting mode of operation.

Note that the zinc–air battery is power limited with a peak power density for efficient pulses of only 156 W/kg.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

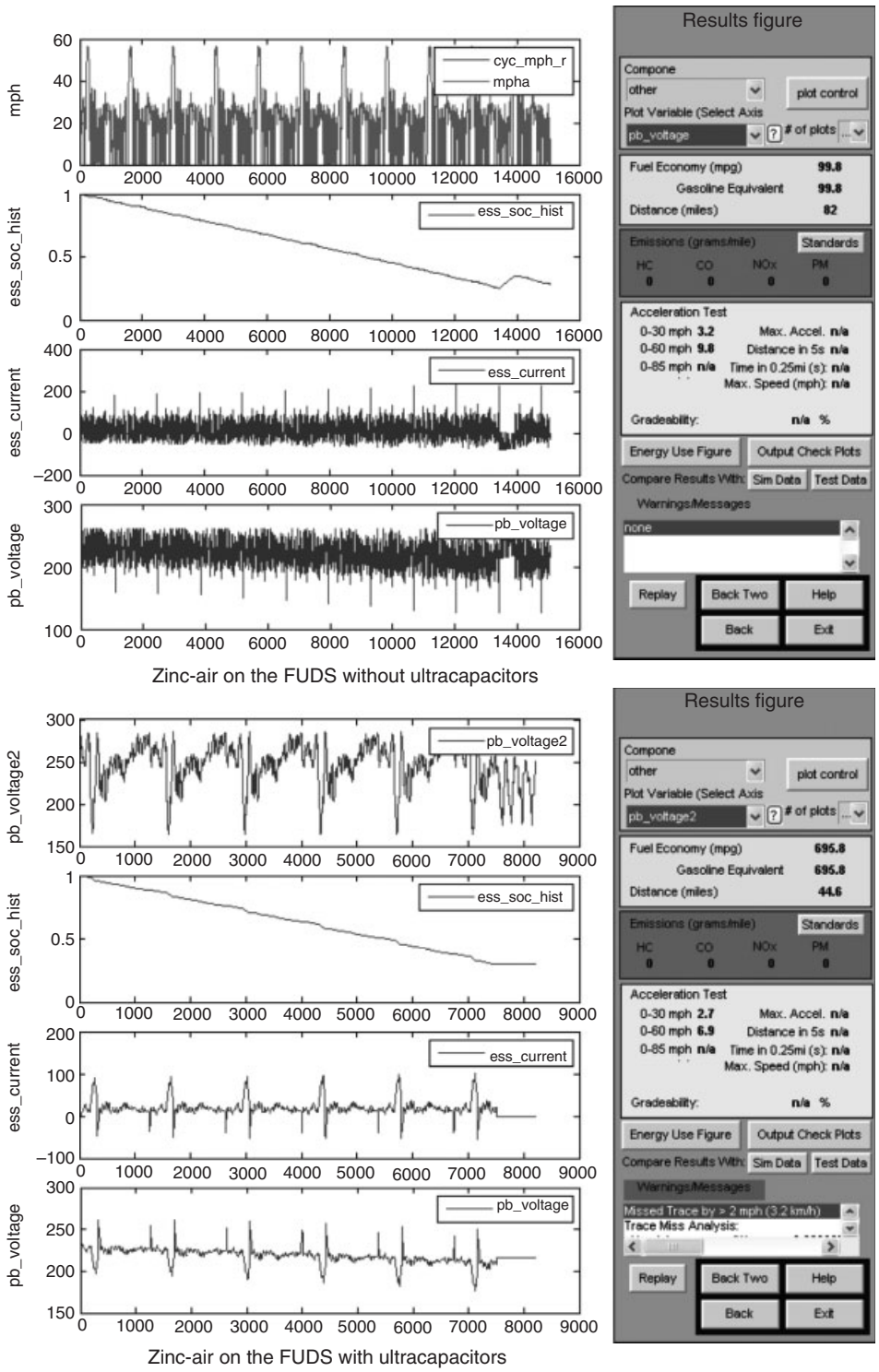


Figure 6. The zinc–air battery on the FUDS with and without ultracapacitors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The carbon/carbon ultracapacitor has a high power density of over 2 kW/kg for both discharge and charge pulses.

Simulations were made for the FUDS, Federal Highway, and ES06 driving cycles. For each of the driving cycles, runs were made for selected numbers of cycles to represent driving in the charge depleting and charge sustaining modes of operation. The all-electric range and energy use (Wh/mi) were determined for the charge depleting mode, and the fuel economy (mpg) was determined for all cases/modes in which the engine was operating.

The simulation results in Table 12 show the advantages of using ultracapacitors in combination with batteries especially ones with limited power capability. With the batteries in combination with the ultracapacitors, the PHEV was able to operate in the all-electric mode until the battery SOC was equal to 30% on the FUDS and Federal Highway driving cycles. In all cases for the ES06 driving cycle, the vehicle had blended operation (engine and electric drive both needed) in the charge depleting mode. The use of the ultracapacitors with the batteries permits all-electric operation of the vehicle over a wide range of driving conditions and more use of electrical energy (higher Wh/mi) for all the driving cycles. Even when the engine is needed in the charge depleting mode, the fuel economy (mpg) is higher by 50–100% using the ultracapacitors. The fuel economy in the charge sustaining mode is also higher for all the driving cycles using the ultracapacitors by about 30%. The acceleration times of the vehicle were lower using the ultracapacitors than for the batteries alone. With the ultracapacitors, the acceleration times were 2.7 s for 0–30 mph and 6.9 s for 0–60 mph. For the batteries alone, the acceleration times were 3.1 s for 0–30 mph and 8.6 s for 0–60 mph. Hence, in all respects, vehicle performance was improved using the ultracapacitors with the zinc–air battery.

The current and voltage responses of the battery with and without the ultracapacitors are shown in Figure 6 for the FUDS driving cycle. The effects of the load leveling of the power demand from the battery using the ultracapacitors are evident in the figure. Both the average currents and the peak currents from the batteries are lower by a factor of 2–3 using the ultracapacitors. The minimum voltages of the battery are significantly higher using the capacitors, and the voltage dynamics (fluctuations) are dramatically reduced. Hence, the stress on the battery and resultant heating are much reduced. The simulation results in Figure 6 also show that the ultracapacitors are utilized over a wide voltage range, indicating that a large fraction of their usable energy storage (100 Wh) is being used to load level the batteries. This is only possible using a DC/DC converter between the battery and the DC-bus. The use of the ultracapacitors will permit the use of the advanced batteries in PHEVs even when their power capability is limited.

6 SUMMARY AND CONCLUSIONS

This chapter was concerned with the research and development on advanced batteries for PEVs in the next 10–20 years. The primary motivation for the development of the advanced batteries is to permit the design of electric vehicles having ranges comparable to conventional ICE vehicles—300–500 miles. By advanced batteries are meant cell chemistries having energy densities significantly higher than those presently available. These chemistries include (i) lithium-ion cells using layered–layered composites of metal oxides in the cathode, mixtures of carbon and silicone in the anode, and advanced electrolytes permitting cell voltages of 4.5–5 V and (ii) metal–air cells utilizing zinc or lithium in the anode and an air cathode. Research is underway and laboratory cells are being prepared for most of these chemistries, but in nearly all cases, it is anticipated that many years of R & D will be needed before the most advanced batteries could be marketed for use in vehicles. It is expected that some of the new lithium chemistries will be the first of the advanced battery technologies to be marketed in the next 5–10 years. The metal–air batteries are much further in the future and likely will be required to approach the 300–500-mile range goal.

Spreadsheet models of cells using the various chemistries were developed to calculate their energy densities and power capability. In the case of the advanced lithium-ion chemistries, it was found that the largest improvements in cell performance will result from significant increases in cathode charge capacity mAh/g and increases in the cathode voltage. Using silicone–carbon anodes is not projected to have a significant improvement in cell performance compared to using graphite in the anode. For example, for a cell having a 50-μ thick cathode, the energy densities using graphite and a layered-composite cathode are projected to be 334 Wh/kg and 710 Wh/L at a cell voltage of 4.5 V compared to 289 Wh/kg and 620 Wh/L at the lower voltage of 4 V. The power potential using the higher cathode voltage is also significantly higher. Large differences in cell performance and power capability result from changes in the electrode thicknesses, and comparisons in cell performance for different electrode chemistries should be done at the same electrode thickness. The model results for the advanced lithium-ion cells indicate that energy densities of 350–400 Wh/kg and 750–800 Wh/kg (unpackaged) are possible at power capabilities suitable for vehicle applications. With improvements in plug-in vehicle design (lower weight and energy consumption-Wh/mi), the advanced lithium batteries could be used in 250–300-mile range vehicles.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Zinc–air and lithium–air cells were also modeled to project their performance. Both metal–air systems were modeled using an aqueous electrolyte (KOH) and a carbon/catalyst air cathode. In the case of the lithium–air cell, a ceramic protective film was used on the anode to make the cell water-stable. In the case of the zinc–air cells, the projected energy densities were 500–600 Wh/kg and 2000–2400 Wh/L (unpackaged) depending on the discharge time and air cathode current density assumed. The power capability of the Zn–air cells was relatively low in the range of 150–300 Wh/kg. The projected energy densities of the lithium–air cells were 2000–3000 Wh/kg and 2000–2500 Wh/L (unpackaged) depending on the resistance (Ohm-cm^2) assumed for the anode protective film and the current density of the air cathode. In all cases for the lithium–air calculations, it was assumed that large reductions in the present high resistance of the anode protective film would occur in the future. The power capability of the lithium–air cells with the improved protective film was 1000–1500 W/kg, which was much higher than that for the Zn–air cells. The higher power capability of the Li–air cells is primarily due to their higher cell voltage.

It appears that the metal–air batteries would be needed to extend the range of the electric vehicles beyond about 300 miles. The energy density of the metal–air batteries seems to be adequate for the long-range EVs, but there are questions concerning the adequacy of their power capability. It is likely necessary to combine a high power battery or ultracapacitor with the metal–air battery to meet the power requirements of the vehicles.

NOMENCLATURE AND ABBREVIATIONS

AER	all-electric range
ICE	Conventional vehicle with engine and transmission
EV	Electric vehicle (battery-powered)
HEV	Hybrid-electric vehicle (charge sustaining with electric motor and engine)
LMO	Lithium Manganese Oxide
NMC	Nickel Manganese Cobalt
PHEV	Plug-in hybrid electric vehicle with a rechargeable battery
PHEV-XX	PHEV vehicle with an all-electric range of XX miles
C_d	drag coefficient of the vehicle
A_f	frontal area of the vehicle (m^2)
$f_{rolling}$	rolling resistance of the tires (kg/kg)
W_V	inertia weight of the vehicle (kg)
L/100 km	fuel economy in terms of liters per 100 km
mpg	fuel economy in terms of miles per gallon
Wh/mi	electrical energy from the battery
$(kWh)_{useable}$	useable energy (kWh) from the battery
Wh/kg	energy density of the battery
W/kg	power density of the battery
Ah	Amp-hour capacity of a cell in the battery

RELATED ARTICLES

Overview of Electric, Hybrid and Fuel Cell Vehicles
Parallel Hybrid Electric Vehicles (Parallel HEVs)
Rechargeable Battery Basics
Power and Energy Requirements for Electric and Hybrid Vehicles
Battery Safety for Lithium Batteries in Vehicle Applications
Ultracapacitors in Hybrid and Plug-in Electric Vehicles
Regenerative Braking Systems
Energy Management Systems of EVs

REFERENCES

Barnett, B.M. (2011) Translating High Capacity Materials into High Energy Density, High Performance Cells, Presentation at the 7th Annual international Conference, Lithium Battery Power 2011, Las Vegas, Nevada, November.

Burke, A.F. and Miller, M. (2009a) Performance Characteristics of Lithium-ion Batteries of Various Chemistries for Plug-in Hybrid Vehicles, EVS-24, Stavanger, Norway, May (paper on the CD of the meeting)

Burke, A.F. and Miller, M. (2009b) The UC Davis Emerging Lithium Battery Test Project. Report UCD-ITS-RR-09-18, July.

Burke, A. and Miller, M. (2010) Lithium batteries and ultracapacitors alone and in combination in hybrid vehicles: Fuel economy and battery stress reduction advantages. Paper presented at the Electric Vehicle Symposium 25, Shenzhen, China, November.

Burke, A.F. and Miller, M. (2011) The power capability of ultracapacitors and lithium batteries for electric and hybrid vehicle applications *Journal of the Power Sources*, **196**(1, January), 514–522.

Cheiky, M.C., Danczyk, L., Wehrey, M. (1990) Rechargeable zinc-air batteries in electric vehicle applications. SAE Paper 901516, August.

Dahn, J. (2009), Electrically Rechargeable Metal-Air Batteries compared to Advanced Lithium-ion Batteries. Presentation slides, Dalhousie University, Canada.

Girishkumar, G., McCloskey, B., Luntz, A.C., *et al*. (2010) Lithium-air battery: promise and challenges. *Journal of Physical Chemistry Letters*, **1**, 2193–2203., July

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Goldstein, J.R. and Koretz, B. (1994) On-going Tests of the Electric Fuel Zinc-air Battery for Electric Vehicles. *Proceedings of the 11th Primary and Secondary Battery Technology and Applications*, Deerfield Beach, Florida.

Hassoun, J., Fernicola, A., Navarra, M.A., *et al*. (2010) An advanced lithium-ion battery based on a nanostructured Sn-C anode and an electrochemically stable LiTFSi-Py_{24}TFsi ionic liquid electrolyte *Journal of Power Sources*, **195**, 774–579.

Horino, T., Tamada, H., Kishimoto, A., *et al*. (2010) High voltage stability of interfacial reaction at the $LiMn_2$ O_4 thin film electrodes/liquid electrolytes with boroxine compounds *Journal of the Electrochemical Society*, **157**(6), A677–A681.

IBM Develops a lithium-Air Battery with a 500-Mile Range for Electric Cars (2012) PCWorld News release, January 13.

King, R.D., Song, D., Gikakis, C. *et al*., (2003) Ultracapacitor Enhanced Zero Emission Zinc Air Electric Transit Bus-Performance Test Results, EVS-20, Long Beach, California.

Kowalczk, I., Read, J., and Salomom, M. (2007) Li-air batteries: a classic example of limitations owing to solubilities *Journal of Pure Applied Chemistry*, **79**(5), 851–860.

Kraytsberg, A. and Ein-Eli, Y. (2011) Review of Li-air batteries: opportunities, limitations, and perspective *Journal of Power Sources*, **196**, 886–893.

Lee, J.S., Tai Kim, S., Cao, R., *et al*. (2011) Metal–air batteries with high energy density: Li-air vs. Zn-air *Journal of Advanced Energy Materials*, **1**, 34–50.

Mao, Z. and White, R.E. (1992) Mathematical modeling of a primary zinc/air battery *Journal of the Electrochemical Society*, **139**, 1105–1114.

Marom, R., Amalraj, S.F., Leifer, D., *et al*. (2011) A review of advanced and practical lithium battery materials *Journal of Material Chemistry*, **21**, 9938–9954., February

Martin, J.J., Neburchilov, V., Wang, H. *et al*., Air Cathodes for Metal-Air Batteries and Fuel Cells, 2009 IEEE Electrical Power & Energy Conference, paper 978-1-4244-4509-7/09

Neburchilov, V., Wang, H.J., Martin, J.J., *et al*. (2010) A review on air cathodes for zinc-air fuel cells *Journal of the Power Sources*, **195**, 1271–1291.

Nelson, P., Bloom, I., Amine, K., *et al*. (2002) Design modeling of lithium-ion battery performance *Journal of Power Sources*, **110**, 437–444.

Polyplus Develops Lithium metal Batteries for Underwater Use (2009) News release, July.

Reddy, T.B. (2011a) *Linden's Handbook of Batteries*, 4th, Chapter 26, edn, McGraw-Hill Publishers.

Reddy, T.B. (2011b) *Linden's Handbook of Batteries*, 4th, Chapters 13 and 33, edn, McGraw-Hill Publishers.

Revolt portable battery-Technology Brief, white paper taken from the Revolt Technology website, www.Revolttechnology.com

Schalkwijk, W.A. and Scrosati, B. (2002) *Advances in Lithium-ion Batteries*, Kluwer Academic/Plenum Publishers.

Schimpf, M. (1995) Rechargeable Zinc-Air Batteries-Market and Technology Overview, IEEE Xplorer, paper March.

Smart, M.C., Ratnakumar, B.V., West, W.C. *et al*., (2011) Electrolytes for Use in High Energy Lithium-ion Batteries with wide Operating Temperature Range, DOE Battery Review, Washington, D.C., May.

Thackeray, M.M., Kang, S.-H., Johnson, C.S., *et al*. (2007) Li_2MnO_3-stablized $LiMO_2$ (M = Mn, Ni, Co) electrodes for lithium-ion batteries. *Journal of Material Chemistry*, **17**(30), 3112–3125.

Wei, Z., Huang, W., Zhang, S. *et al*., (2000) Carbon-based air electrodes carrying MnO2 in Zinc-air batteries, *Journal of the Power Sources*, **91**, 83–85, 2000

White, L.J. (2005) An approximate analytical model for the discharge performance of a primary zinc/air cell.Master's thesis, Worcester Polytechnic Institute, January.

Yoshio, M., Brodd, R., and Kozawa, A. (eds) (2009) *Lithium-ion Batteries-Science and Technologies*, Springer Publishing.

Yuasa, M., Nishida, M., Kida, T., *et al*. (2011) Bi-functional oxygen electrodes using $LaMnO_3$/$LaNiO_3$ for rechargeable metal-air batteries *Journal of the Electrochemical Society*, **158**(5), A605–A610.

Zhang, Z., Zhang, L. and Amine, K. (2011a), Advanced Electrolyte Additives for PHEV/EV Lithium-ion Battery, DOE Vehicle Technologies Program, Annual Merit Review, Washington, D.C., June.

Zhang, T., Imanishi, N., Takeda, Y., *et al*. (2011b) Aqueous lithium/air rechargeable batteries, The Chemical Society of Japan, highlight Review, *Chemistry Letters*, **40**, 668–673.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto064
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Battery Safety for Lithium Batteries in Vehicle Applications

Andrew Burke
University of California, Davis, CA, USA

1 Introduction 1
2 Safety Concerns for Lithium Batteries 1
3 Abuse Testing of Batteries 2
4 Sudden Failure Modes of Lithium Batteries 3
5 Analysis and Testing for Thermal Runaway 4
6 Analysis of "Soft Short" Battery Safety Issues 8
7 Battery Management Systems to Mitigate Safety Issues 11
8 Summary and Conclusions 13
Related Articles 15
References 15

1 INTRODUCTION

There are safety concerns with all of the lithium battery chemistries. These concerns stem primarily from failures experienced with cylindrical 18650 $LiCoO_2$ cells in laptop computers. For this reason, lithium batteries have been used with caution in vehicle applications. The primary concern is the possibility of thermal runaway due to an internal short in the battery caused by a manufacturing flaw or abnormally high currents or temperatures due to a failure of the battery control unit or in the case of a vehicle accident. This concern has been addressed in two ways. First, extensive abuse testing of lithium batteries has been done to show that thermal runaway does not occur even if the cells or modules are subject to sudden compression, intrusion of sharp objects, dropping, fire, and so on. The second approach to battery safety is to provide a battery management system (BMS) that monitors the cell/module voltages and temperatures and alerts the vehicle control system/computer if any of the cell voltages or temperatures is outside a specified normal range. Development of a BMS unit for use with their battery is common practice for battery manufacturers. All of these aspects of lithium battery safety are discussed in this article.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 SAFETY CONCERNS FOR LITHIUM BATTERIES

The concern for safety is especially significant for lithium batteries because of their high energy density and capability to store large quantities of energy and the fact that they use organic electrolytes that are highly flammable and toxic. Battery failures can result from operation of the batteries outside of normal operating conditions due to a failure of the battery control and monitoring systems or due to a vehicle accident that could subject the battery to abnormal physical conditions such as crushing or metal penetration. These types of battery failures are addressed in abuse testing of the cells/batteries as discussed in Sections 3 and 5 of the chapter. A second type of battery failure, often referred to as *field failures*, can occur suddenly with little warning during what appears to be normal operation of the battery. These failures usually result from internal "shorts" in a cell due to flaws in the manufacture of the cell. This type of failure is considered in Sections 4 and 6. Primarily because of safety concerns, all lithium batteries are equipped with

Table 1. Relative thermal stability and safety of various lithium battery chemistries.

Lithium Cell Chemistry	Temperature of the Onset of Self-heating (°C)[a]	Heat Release Tendencies	Relative Thermal Stability[b]	Relative Safety Concerns[b]
$LiCoO_2$	185	High	6	6
$LiNi_xCo_yAl_zO_2$	195	High	5	5
$LiNi_xCoyMn_zO_2$	210	Medium	4	4
$LiMn_2O_4$	200	Low	3	3
$LiFePO_4$	225	Low	2	2
$Li_4Ti_5O_{12}/LiMn_2O_4$	200	Low	1	1

[a] Significant self-heating resulting in temperature rise of 1°C/min.
[b] 1 (best)–6 (worst).

Table 2. Summary of abuse testing categories and tests.

Mechanical abuse
- Mechanical shock
- Vibration
- Drop test
- Penetration (nail)
- Immersion (water)
- Crush

Thermal abuse
- Radiant heat (simulated fire)
- Thermal stability (maximum temp.)
- Accelerating rate calorimetry (ramp)
- Overheat/thermal runaway

Electrical abuse
- Short circuit
- Overcharge
- Overdischarge

Table 3. Example of core abuse tests performed on most cells.

Overcharge
- Low rate at 1C, high rate at 3–4C
- Overvoltage of 0.5–1.0 V
- Test for heat and gas generation and possible thermal runaway

Short circuit
- Hard short (1 mΩ)
- Intermediate short (equal to unit impedance)

Thermal stability
- Controlled ramp rate up to 200–250°C
- Test of onset of thermal runaway reactions
- Test flammability of vent gases

Controlled crush and penetration
- 50% max. deformation (load max < 1000x battery weight)
- Nail penetration test at several rates

an extensive battery management and monitoring system, which is intended to mitigate safety issues and, if possible, issue a warning to the vehicle operator of an impending failure. In most cases, the battery monitoring is done on a cell basis. BMSs are discussed in Section 7.

The safety concerns for various lithium battery chemistries are discussed in detail in the next sections. Detailed consideration is given to the analysis of the relative thermal stability of the different chemistries in Section 5. The conclusions of these considerations are summarized in Table 1. Since there are wide variations in the tendencies of the different chemistries to experience thermal events, including runaway, it is necessary to understand the differences and the reasons for them. Approaches for investigating these differences are discussed in the following sections.

3 ABUSE TESTING OF BATTERIES

Abuse testing of lithium batteries is well documented in the literature and the procedures used in the testing are also specified in detail (Battery Safety, 2011; Mikolajczak *et al*., 2011; Roth, 2009; Roth, Wunsch, and Orendorff). Abuse testing is done in several categories: (i) mechanical, (ii) thermal, and (iii) electrical. As shown in Table 2, there are a number of tests done in each category. In most cases, the test conditions would not be encountered in normal operation of the battery, but might be encountered due to a malfunction of the battery control system or in the event of a vehicle accident on the road. One of the objectives of the abuse testing is to determine the extent of the abuse that a particular battery can tolerate. Therefore, the abuse testing is often done to induce battery failure and destruction with the intent of comparing the tolerance of various lithium battery chemistries and mechanical designs.

An example of a core set of abuse tests is given in Table 3. The testing is usually done with the cell/battery at full charge. These tests are customarily performed by the developer to demonstrate the safety of their cells. Testing at extreme conditions to determine the safety limits of specific cell chemistries and designs is usually done at DOE

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Summary of hazard level designations for abuse testing.

Hazard Level	Description	Classification Criteria, Effect
0	No effect	No effect, no loss of functionality
1	Passive protection activated	No defect, no leakage, no venting, no fire or flame, no rupture, no explosion, no exothermic reaction or thermal runaway. Cell reversibly damaged. Repair of protection device needed
2	Defect/damage	No leakage, no venting, no fire or flame, no rupture, no explosion, no exothermic reaction or thermal runaway. Cell irreversibly damaged, repair needed.
3	Leakage $\Delta m < 50\%$	No venting, no fire or flame, no rupture, no explosion, weight loss <50% of electrolyte weight. (electrolyte = solvent + salt)
4	Venting $\Delta m \geq 50\%$	No fire or flame, no rupture, no explosion, weight loss $\geq 50\%$ of electrolyte weight.
5	Fire or flame	No rupture, no explosion, that is, no flying parts
6	Rupture	No explosion, but flying parts, ejection of parts of the active mass
7	Explosion	Explosion, that is, disintegration of the cell

Descriptions adapted from EUCAR and SAND2005-3123.



Figure 1. Battery failure testing at Sandia National Laboratory. (Reproduced from Roth, Winsch, and Orendirff, Sandia National Laboratory.)

National Laboratories such as Sandia National Laboratory (Roth, 2009; Roth, Wunsch, and Orendorff).

The response of the battery/cells to the abuse testing is often rated on a qualitative basis as indicated in Table 4. A response of "hazard level" of 1–4 is preferred, but levels up to 5 are probably acceptable. As shown in Figure 1, abuse testing can result in catastrophic disassembly of the cells without a fire or explosion (Level 6 hazard).

Most abuse testing reported in the literature at the present time (2012) has been done on relatively new cells, but in the future, abuse testing is needed on aged cells and full battery packs. It seems likely that full battery pack abuse testing has been done by battery and/or vehicle developers, but those results are not available in the literature.

4 SUDDEN FAILURE MODES OF LITHIUM BATTERIES

In vehicle applications, safety concerns are related to sudden failure of the batteries that can endanger the vehicle operator and/or lead to destruction of the vehicle. These failures are usually caused by a thermal runaway condition (a rapid increase in temperature and pressure) inside the battery, which can result in disassembly of the battery or a fire. Under some conditions, the thermal runaway can occur in a few seconds with little prior warning that something was wrong. Under more normal conditions, the BMS would alert the vehicle controller that the battery temperature was increasing out of the normal, allowable range, and action

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

would be taken to limit the temperature increase before thermal runaway occurred.

Present thinking is that sudden battery failures could be the result of internal shorts in a cell in the battery pack. The likely scenario (Barnett and Siriramulu, 2011; Barnett *et al*., 2009) is that there are "soft shorts" in the cells which normally result only in localized nonuniform distributions of current in the cells, but not large enough variations to result in localized high temperatures that can lead to thermal runaway conditions. In rare occasions, the "soft shorts" can lead to "hard shorts" and localized high currents and resultant heating that can lead to thermal runaway conditions. This occurrence is a very rare event (occurring in only one in several million cells from high quality battery manufacturers), but nevertheless, cells should be designed to avoid the possibility of thermal runaway. The analysis and testing for thermal runaway and how the possibility of it can be minimized or even eliminated are discussed in Sections 5 and 6.

A "soft short" in a cell is simply a small area (a few square millimeter) for which the resistance (Ω-cm^2) is significantly less than the average for the cell. The result will be that the current (A/cm^2) will be higher through that region of the cell and I^2R heating will result in the temperature being higher than elsewhere in the cell, Over time, this can lead to further degradation of the cell in that region and even a total failure of the separator, leading to direct contact of the positive and negative electrodes. This results in a "hard short", very high currents, very high local heating and temperature rise, and possible total failure of the cell and battery. The initial "soft short" could be the result of a significant nonuniformity in the cell electrode coating during manufacturing or a defect in the separator material. The most likely direct cause of a "hard short" is the presence of a very small metal particle that inadvertently got into the cell during manufacturing or assembly. Such a particle could cause a direct contact between the two electrodes as the cell ages in the vehicle. Overdischarge of the cell could lead to the deposition of copper from the anode current collector, which would also lead to a "hard short". Once a "hard short" is encountered, the cell temperature will increase very rapidly (in seconds) and there is little that can be done to prevent total failure of the battery.

5 ANALYSIS AND TESTING FOR THERMAL RUNAWAY

5.1 Thermodynamic and cell heating considerations

Determination of the change in temperature of a cell under conditions of rapid heating and resultant chemical change inside the cell requires complex modeling (Fang, Kwon, and Wang, 2010; Chen, Wang, and Wan, 2006; Kim, Pesaran, Spotnitz, 2007), but the essence of what is occurring can be expressed in the rather simple equation below:

$$\sum \mathrm{i}\,(m_{\mathrm{i}}C_{\mathrm{Pi}})\mathrm{d}T = (\,I^2R + \sum \mathrm{i}\,(\mathrm{mdot}_{\mathrm{i}}(H_{\mathrm{Ri}}))\mathrm{d}t - IT\,\mathrm{d}S + Q_{\mathrm{heating}} - Q_{\mathrm{loss}} \qquad (1)$$

Subscript "i" refers to the constituent components of the cell (Figure 2). Each of the components has a mass and specific heat ($m_{\mathrm{i}}C_{\mathrm{Pi}}$) associated with it as well as a heat release due to chemical reaction ($\mathrm{mdot}_{\mathrm{i}}$ (H_{Ri})). The heat release term depends on the kinetics of the reaction (mdot), which is highly temperature-dependent. It is this dependency and the magnitude of the heat of reaction that strongly affects the tendency of particular lithium battery chemistries to

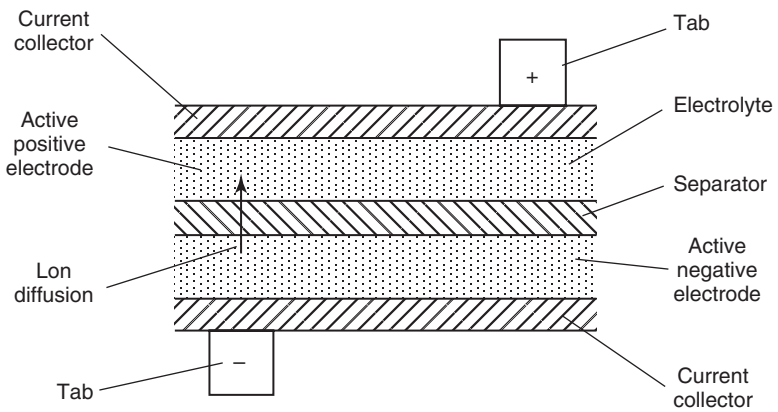


Figure 2. Schematic of the cell components.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

experience thermal runaway. $Q_{heating}$ and Q_{loss} are the heating and cooling rates provided to the cell from external sources.

It is of interest to evaluate Equation 1 for the simple case of a hard short in which only the resistant heating is considered. The maximum power discharge condition (matched impedance) corresponds to 50% efficiency for which one-half of the electrochemical energy in the cell is dissipated in heat. If the cell is insulated, an estimate of the temperature increase due to a hard short (very high current) will be the following:

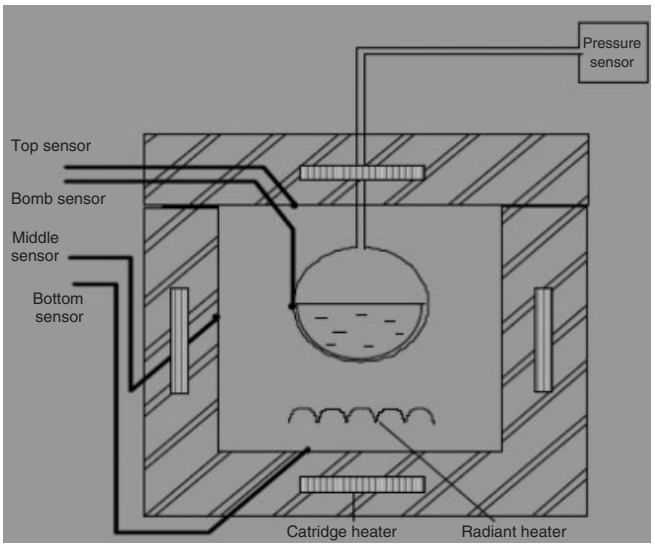
$$\Delta T = {}^1\!/_2\,(3.6)\ (\text{Wh/kg})/C_p, C_p = 1.1\ \text{kJ/kg}\ ^\circ\text{C}$$

Energy Density (Wh/kg)	ΔT (°C)
100	164
150	245
175	278
200	328

The temperature increase is strongly dependent on the energy density (Wh/kg) of the cell, indicating that the potential for thermal runaway is greater for the higher performance lithium chemistries. The actual increase will be much greater than shown in the table due to the heat release from the chemical reaction (H_{Ri}) at the high temperatures.

Equation 1 considers the total volume and mass of the cell and assumes uniform heating. Thus, the equation determines an average temperature and not the distribution within the cell or the effects of nonuniform heating as would be the case for internal shorts or the penetration of foreign objects. Uniform heating would be appropriate to evaluate the cell response to overcharging or an external short. There has been considerable research (Barnett, 2011; Smith *et al*., 2010; Kim, Pesaran, and Smith, 2008; Keyser, Kim, and Pesaran, 2010) concerned with modeling the response of lithium cells to localized heating. These analyses are complex and require the adaptation of multidimensional, time-dependent finite-element numerical methods to the cell. In addition, they require detailed descriptions of the cell construction and material properties and extensive thermochemical data to model the high temperature chemistry that occurs in the cell. These calculations indicate that localized heating (internal shorts) can lead to thermal runaway conditions in a few seconds under certain conditions of cell design, discharge current, and cooling. These computer results will be discussed later in Section 6.

Much of the available information relative to lithium battery safety is based on testing of cells in a special test apparatus known as an *accelerating rate calorimeter* (*ARC*),



Test chamber with cell



Figure 3. An accelerating rate calorimeter. (Reproduced by permission of Thermal Hazard Technology (2012).)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

which permits the heating of cells in a systematic way in a controlled, adiabatic environment until the cell undergoes exothermic reactions (often leading to thermal runaway and cell rupture). ARC testing and the information/conclusions gained from that testing are discussed in the next several sections.

5.2 Accelerating rate calorimetry (ARC) testing

Under normal battery operating conditions, the anode, cathode, and electrolyte materials in the battery are stable and the heat generated is due primarily to resistance heating (I^2R). The magnitude of this heating is significant, but the battery temperature can be controlled to 50–60°C by cooling without great difficulty. However, if the battery experiences much higher heating rates than I^2R and the battery temperature increases to much higher values (>100°C), the anode and cathode materials become unstable and begin to react with the electrolyte which itself begins to decompose. These reactions can be exothermic, leading to large increases in internal heating and resultant increases in the cell temperature. The ARC is designed to systematically increase the temperature of the cells and contains the cell fragments and gases if/when it ruptures (Figure 3). The ARC chamber (Thermal Hazard Technology) is constructed of steel and has the capability of controlling its surface temperatures to track the surface temperature of the test cell as it is heated and experiences exothermic reactions. In this way, the cell test is done under adiabatic conditions. The temperature and internal pressure of the cell is recorded during the test and the rate of temperature rise (°C/minute) is determined as a function of test time.

The standard procedure for most of the ARC testing is to gradually increase the cell temperature in steps using the heat-wait-seek approach. After the heat step, the temperature is maintained constant (wait) for 30–35 min, seeking whether the cell will enter into the exothermic mode in which the temperature of cell continues to increase without external heating. This test (Thermal Hazard Technology, 2012) is illustrated in Figure 4 for a 18650 $LiCoO_2$ cell. Note that in this test, thermal runaway started at about 105°C and took place over a period of about 20 min to a maximum temperature greater than 300°C.

Test results (Thermal Hazard Technology, 2012) for an overcharge test of the 18650 cell are shown in Figure 5. The overcharge was done to a voltage of 5 V at a constant current of about 1C. The cell became self-heating in 40 min at a temperature of about 50°C after the start of the overcharge test. The thermal runaway started at about 65 min at a temperature of 100°C. Onset of thermal runaway from the overcharge occurred relatively slowly and seemed to be controllable by cooling and termination of the overcharge.

External short (Figure 6) and nail penetration (Figure 7) tests (Thermal Hazard Technology, 2012) can also be performed in the ARC. In the case of the external short, thermal runaway started about 15 min after the "short" event at about 150°C and took place over about 3 min to a maximum temperature of 450°C. In the case of the nail penetration test, thermal runaway started about 180 min after the "nail" event at about 150°C and took place over about 5 min to a maximum temperature of 450°C. Especially, in the case of the nail penetration, the thermal runaway took place long after the heating event and an initial finding might be that neither test resulted in thermal runaway. However, the thermal runaway event was similar in both cases. All the test results shown (Figures 4–7) are for the 18650 cell. Unfortunately, corresponding data

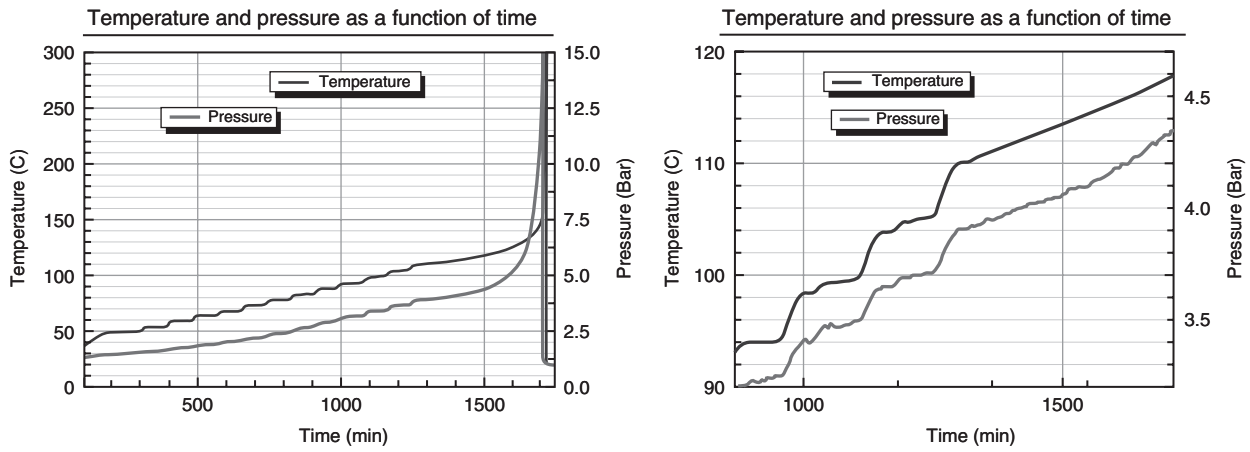


Figure 4. Heat-wait-seek test for a 18650 cell to thermal runaway. (Reproduced by permission of Thermal Hazard Technology (2012).)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

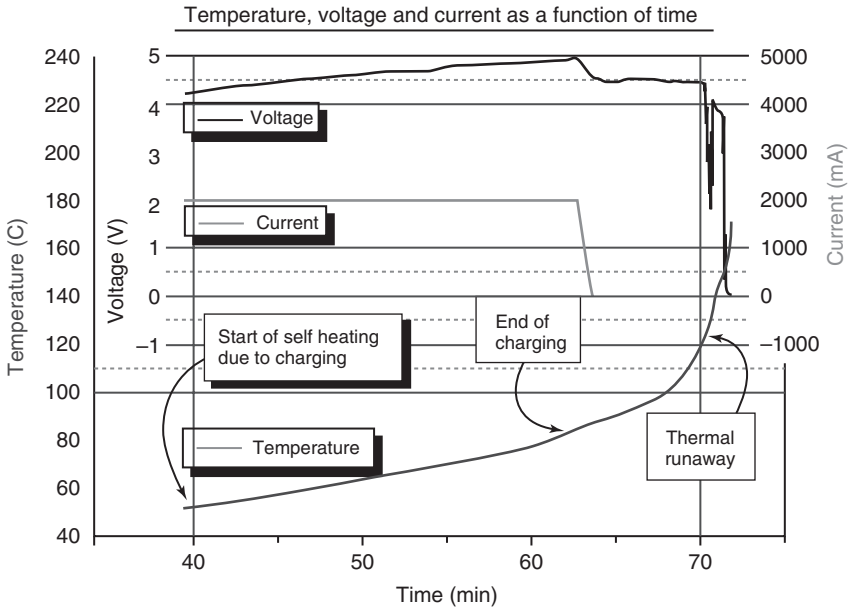


Figure 5. Overcharge test for a 18650 cell to thermal runaway. (Reproduced by permission of Thermal Hazard Technology (2012).)

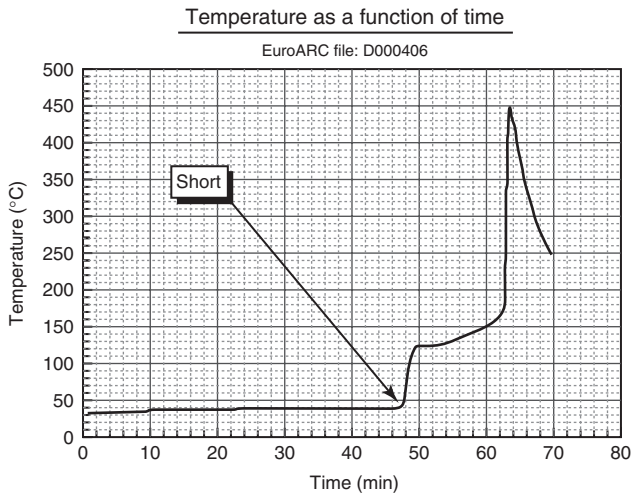


Figure 6. External "short" test for a 18650 cell to thermal runaway. (Reproduced by permission of Thermal Hazard Technology (2012).)

for other lithium battery chemistries do not seem to be available.

5.3 Thermal runaway sequences for lithium batteries

Test data for several thermal runaway events for 18650 $LiCoO_2$ cells are shown in Figures 4–7. In the case of the events with local heating (external short and nail penetration), the time in which the temperature rises rapidly to a maximum of 450–500°C is 3–5 min. This period is preceded by a longer time in which the temperature rises more slowly, but nevertheless, self-heating (exothermic reactions) is taking place in the cell. It is of interest to understand what is happening chemically in the cell during the complete thermal runaway event. In general terms, as shown in Figure 8, this is presently reasonably well understood for the lithium batteries. The initial step in the self-heating of the battery is the decomposition of the SEI layer and reaction of the anode material (graphite) with the electrolyte. This occurs in the temperature range of 80–120°C. The next steps involve reactions of the cathode materials (composite metal oxides) with the electrolyte and the evolution of oxygen from the cathode. This occurs at temperatures near 150 °C and signals the onset of the rapid rise in the temperature and in most cases ends in a rupture of the cell and venting of gases. Figure 8 shows how the rate of temperature rise (°C/min) changes rapidly from the initial to final phases of the self-heating of the cell. To avoid thermal runaway, it appears necessary to limit the cell temperature to <100°C even when self-heating becomes evident by adequate cooling of the cell. The details of the chemistry and associated temperatures at which the different steps occur vary with the lithium battery chemistry of interest as discussed in the next section.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

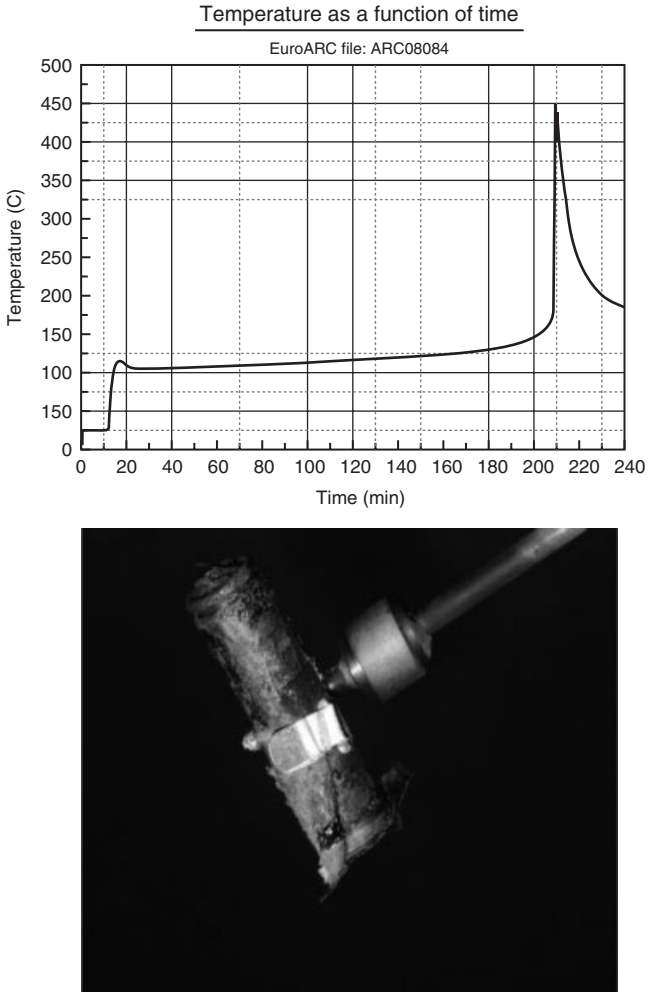


Figure 7. Nail penetration test for a 18650 cell to thermal runaway. (Reproduced by permission of Thermal Hazard Technology (2012).)

5.4 Comparisons of the thermal runaway tendencies of lithium batteries of various chemistries

Most of the thermal stability and abuse testing of lithium batteries has been done with 18650 cells of the $LiCoO_2$ chemistry. These cells are used in battery packs for laptop computers and other consumer electronics and are available in very large quantities. Some testing has been done of other lithium battery chemistries. ARC data (Roth, 2009; Roth, Wunsch, and Orendorff; Doughty, 2012) are shown in Figure 9 for several battery chemistries. The data show that there are large differences in the tendencies of the various chemistries to experience thermal runaway. The $LiCoO_2$ and Li NiCoAl composite cathode chemistries exhibit much higher tendencies for thermal runaway than the other chemistries containing Mn and phosphate. Not shown is the lithium titanate oxide cell which is the most thermally stable of all the lithium cell chemistries (Takami *et al*., 2009; House, 2007).

Thermal stability and abuse studies of the $LiFePO_4$ cells are given in Deveney *et al*. (2007). Temperature data for 10-Ah and 25-Ah cells are shown in Figures 10 and 11 for several types of tests. In Figure 10, test results are shown for short circuit and overcharge events for 10-Ah cells. In both cases, the maximum temperature reached in the cell was 100°C and no venting was experienced. In the case of the short circuit test (0.63 mΩ), the initial current was very high, the time to peak temperature was about 60 s, and no thermal runaway was exhibited. In the case of the overcharge test, the overcharge current was 2C and the voltage was limited to about 4.5 V. The internal cell temperature reached 100°C in about 250 s and then decreased after venting (no smoke, no fire).

Abuse test results for a larger 25-Ah $LiFePO_4$ cell are shown in Figure 11 for overcharge to 5 V and a nail penetration. In both tests, the cells vented. There was smoke, but no flame. It is interesting to note that 1C overcharge to 5 V resulted in thermal runaway to 310°C, but 2C overcharge to 4.5 V resulted in a temperature of only 100°C and no venting. The nail penetration test resulted in heating over several minutes and maximum temperatures of only 100–120°C with venting, but no thermal runaway. The results shown in Figures 9–11 indicate that $LiFePO_4$ batteries seem to be resistant to thermal runaway except in cases of marked abuse.

6 ANALYSIS OF "SOFT SHORT" BATTERY SAFETY ISSUES

Most of the thermal runaway events discussed in the previous sections concerned abuse conditions that could result from battery mismanagement or vehicle accidents. These events in most cases were preceded by periods of exothermic heating of the battery lasting many minutes. It seems reasonable to assume that these events could be avoided with the appropriate BMS and cooling strategy. In addition, these events would be the result of a vehicle accident or human error in the management of the battery.

This section is concerned with battery failures resulting from the large localized heating caused by the development of hard shots interior to a cell. These shorts would be unknown to the battery user and occur very infrequently. Hence, totally avoiding the occurrence of these shorts is likely not to be possible. What is then needed is a cell design and chemistry that will not result in thermal runaway even should a hard short occur. Much of the research

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

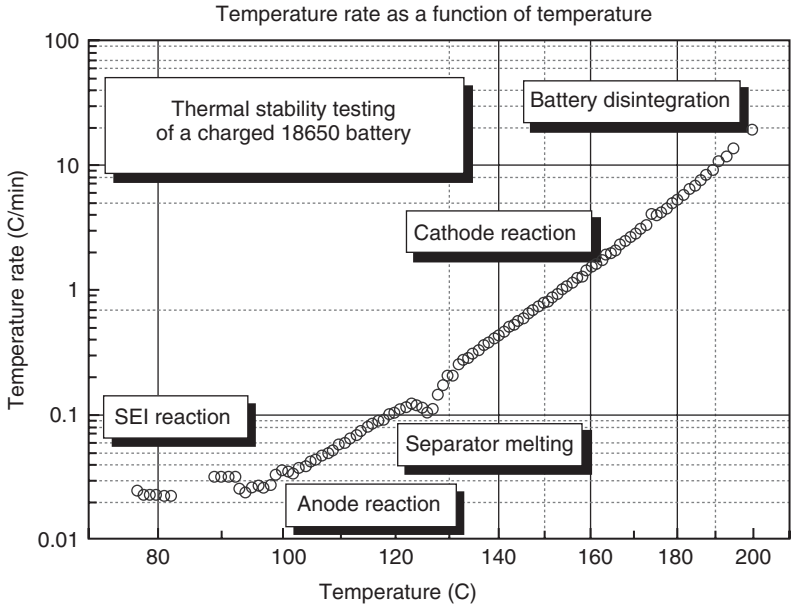


Figure 8. Temperature rate as a function of cell temperature and the chemistry occurring during self-heating and thermal runaway. (Reproduced by permission of Thermal Hazard Technology (2012).)

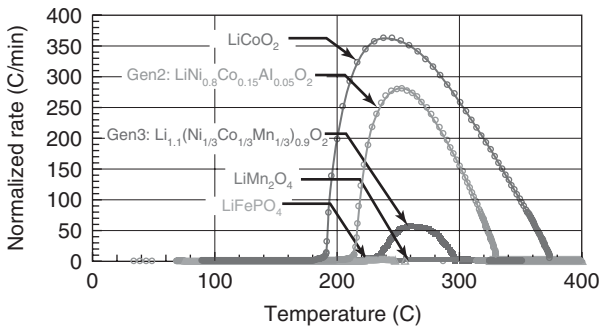


Figure 9. Comparisons of the thermal stability of lithium batteries of various chemistries. (Reproduced from Doughty (2012).)

concerning the recognition and consequences of "shorts" in batteries is thus directed toward understanding how the thermal runaway develops and how to design runaway proof cells. This research is considered in the next section.

6.1 Modeling of soft/hard short transitions and failures

Most of the research involving shorts and cell safety has involved computer modeling of the cells (Barnett and Siriramulu, 2011; Barnett *et al*., 2009; Barnett, 2011; Smith *et al*., 2010; Kim, Pesaran, and Smith, 2008; Keyser, Kim, and Pesaran, 2010). The models involve the solution of the three-dimensional, time-dependent partial differential equations (Fang, Kwon, and Wang, 2010; Chen, Wang, and Wan, 2006; Kim, Pesaran, Spotnitz, 2007) using the finite-element approach. It is critical to include the high temperature thermodynamics and kinetics of the chemical reactions between materials in the anode, cathode, and electrolyte in the cell. Details of the model are illustrated in Figure 12. A key element of the model is how to specify the characteristics of the "short" in terms of its size (mm), energy release (% of cell energy), and power (W). Some combinations of these parameters lead to thermal runaway and others do not. A typical model result leading to thermal runaway in a cylindrical $LiCoO_2$ cell is shown in Figure 13. Note that the average internal temperature of the cell is higher than on the external surfaces and the temperature increase occurs at an earlier time.

In this case, the time to thermal runaway to 500°C is about 10 s although internally the exothermic heating and average temperature increase was more gradual. Hence, it appears that thermal runaway due to internal shorts can occur over a very short time.

As noted previously, there are combinations of parameters for which thermal runaway does not result from the presence of a "short" in the cell. This is shown in Figure 14 for various values of energy and power when the short is initiated (Barnett, 2011). For small values of the relative energy, the power of the short can be quite large (>100 W) and there will not be thermal runaway. Hence, there is a wide region for all cells that should be thermal-runaway-free.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

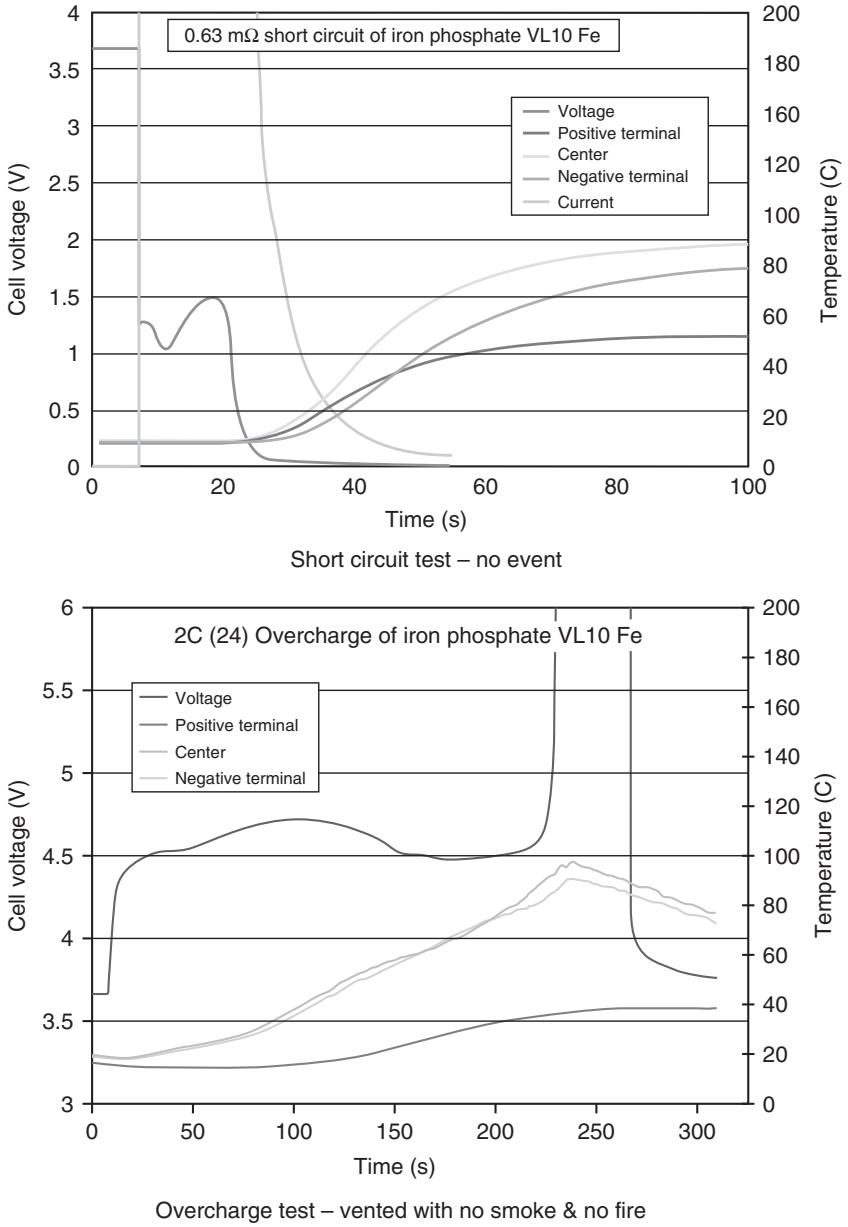


Figure 10. Short circuit and overcharge tests of the Saft 10-Ah $LiFePO_4$ cell. (Reproduced from Deveney *et al*. (2007). © Saft.)

Calculations have also been made (Barnett, 2011) comparing the response of cylindrical and prismatic cells to the occurrence of "shorts". As shown in Figure 15, it was found that for the same relative energy, the threshold power for thermal runaway was much higher for a prismatic cell than for a cylindrical cell of the same size (Ah). This is due to the higher heat loss from the prismatic cell. It is also evident from Figure 15 that for both types of cells, there is a sharp sensitivity to the initiating power—only a very small difference in power can have a large influence on whether thermal runaway occurs. In both cases, there is an extended period of exothermic heating and then a sudden temperature rise to runaway.

The model results presently available (2012) are incomplete, especially concerning the effects of lithium battery chemistry on the response of cells to soft/hard shorts. The results are also not conclusive regarding how to design thermal-runaway-free cells, but they seem to indicate that providing adequate cooling of the interior of the cells is important.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

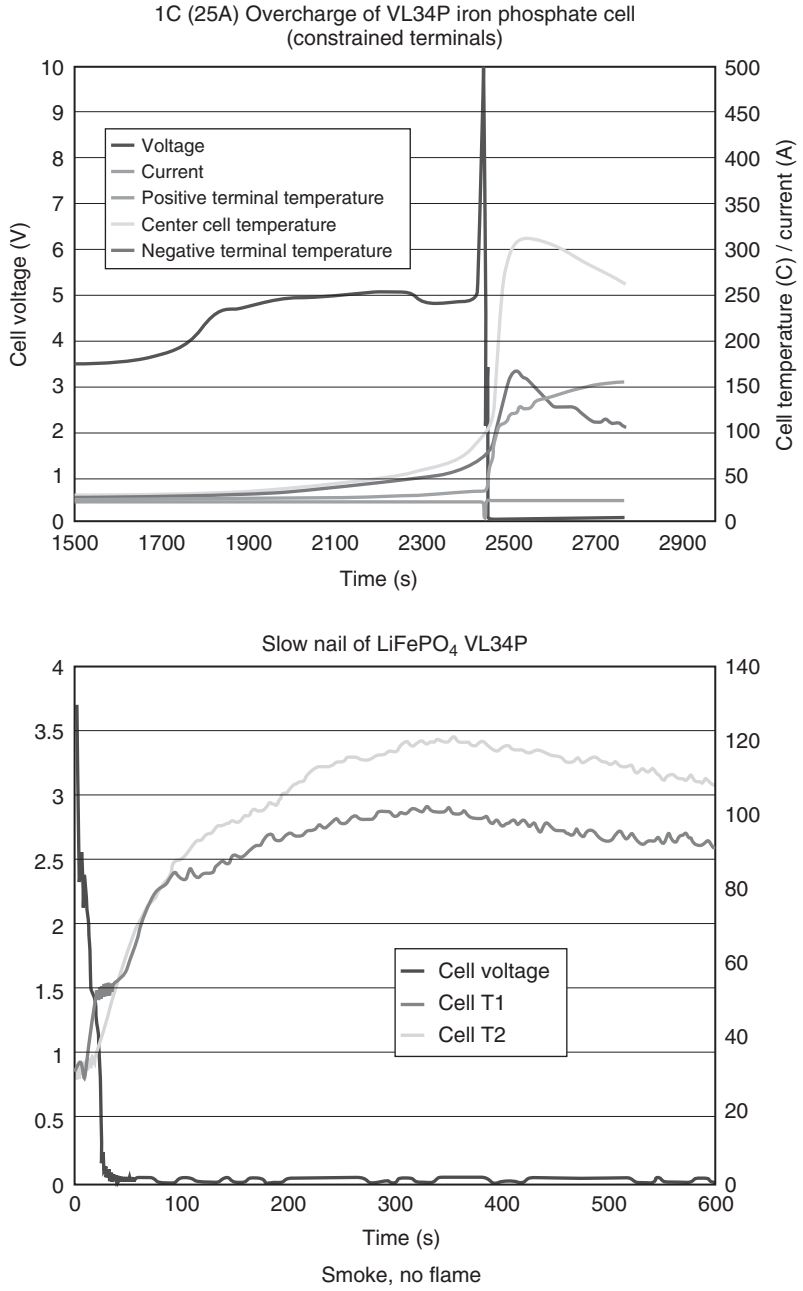


Figure 11. Abuse tests of the Saft 25-Ah $LiFePO_4$ cell. (Reproduced from Deveney *et al*. (2007). © Saft.)

7 BATTERY MANAGEMENT SYSTEMS TO MITIGATE SAFETY ISSUES

It is well recognized that providing an appropriate BMS is key to safe operation of lithium batteries. All BMSs monitor the voltages and temperatures of the cells and warn the vehicle controller when a cell condition is outside safe ranges. This monitoring function of the BMS should permit the battery to avoid abusive conditions during normal operation of the vehicle. Operation of the battery disconnect contactors by the BMS should also avoid abusive conditions in the event of a vehicle accident.

Utilizing the BMS to warn of the presence of soft/hard shorts or conditions that indicate "shorts" may have occurred is much less developed than the straightforward monitoring function. There has been considerable work

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

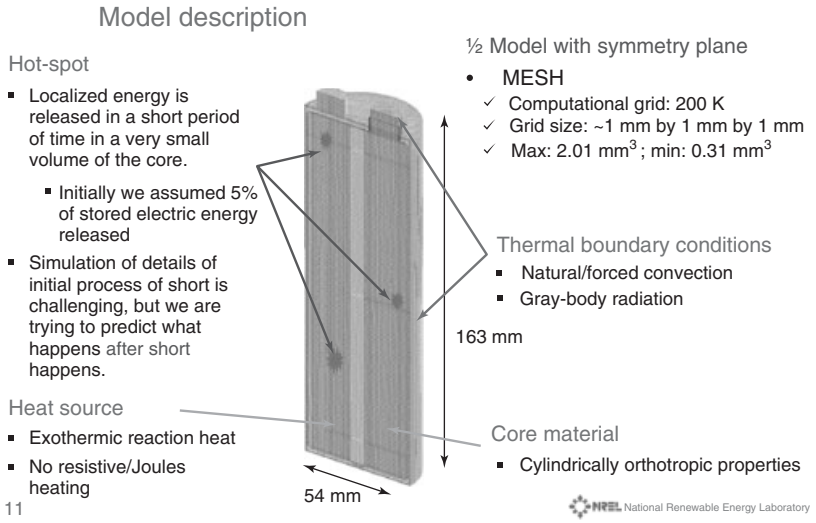


Figure 12. Description of the model for analyzing "shorts" in cells. (Reproduced from Kim, Pesaran, and Smith (2008).)

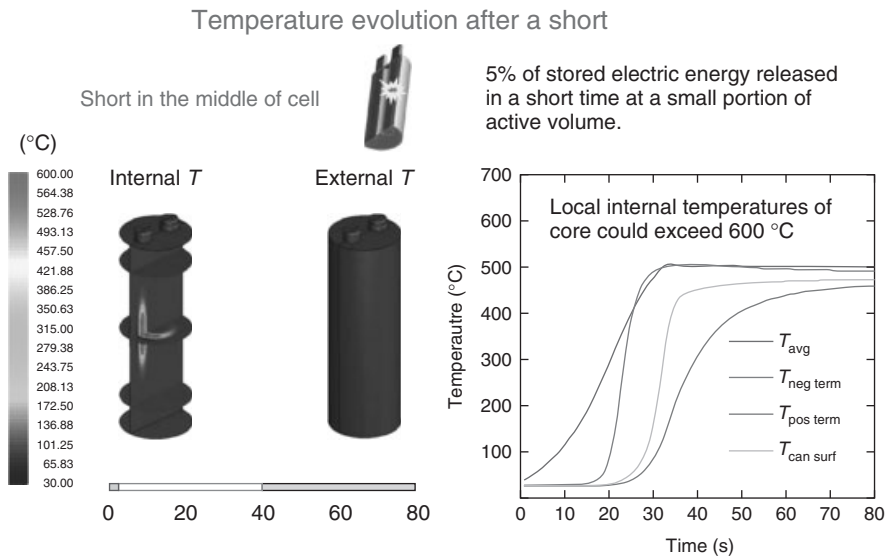


Figure 13. Calculated cell response to a "short" leading to thermal runaway. (Reproduced from Kim, Pesaran, and Smith (2008).)

(Mikolajczak *et al*., 2010; Schaffhauser, 2012; Darcy and Smith, 2010; Sanyo; Texas Instruments) concerned with (i) identifying indicators (Mikolajczak *et al*., 2010; Schaffhauser, 2012; Darcy and Smith, 2010) pointing to the presence or impending occurrence of "shorts" and (ii) development (Sanyo; Texas Instruments) of chips and printed circuit boards (PCBs) that implement in hardware and software the various approaches to possible battery failure detection. Most of this work has been done in connection with lithium-ion cells and units used in laptop computers. However, this work should be valuable for lithium battery applications in vehicles.

One approach to detecting the presence of "micro-shorts" in cells or battery packs is to analysis their voltage–current behavior during charging and rest periods. For example, it has been found (Mikolajczak *et al*., 2010) that cell behavior that indicated incipient faults include elevated self-discharge during rest periods, extended taper-current times, and charge capacity significantly higher than discharge capacity. In these cases, the batteries should be inspected and tested carefully to identify the source of the problem before continuing to use them.

Most of the thinking regarding detection of "shorts" during the normal use of the battery is concerned with

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

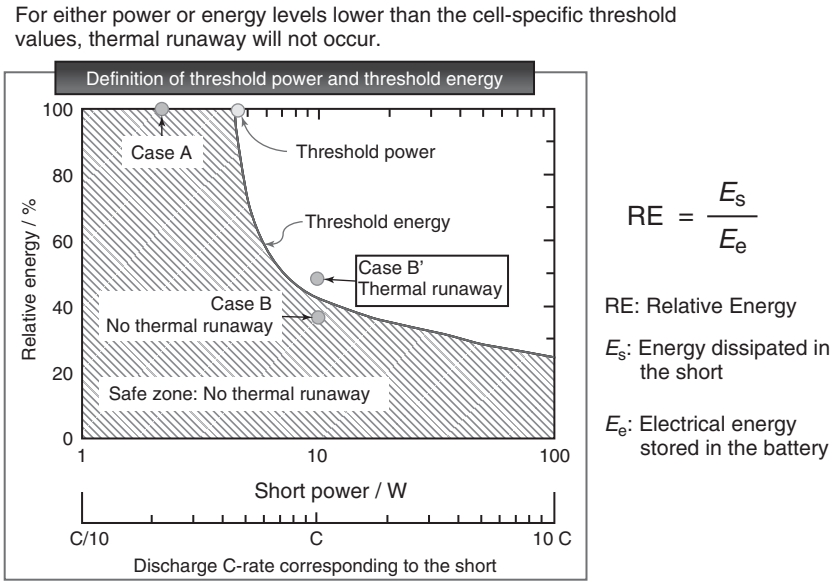


Figure 14. Combinations of relative energy and power (Barnett and Siriramulu, 2011) for which the cell will be free of thermal runaway. (Reproduced from Barnett (2011). © the Knowledge Foundation.)

detecting sudden changes in cell resistance or impedance and/or changes in temperature either at the surface of the cells or at the negative or positive terminals (Figure 13). These approaches can be combined with the monitoring function of the BMS. One approach (Schaffhauser, 2012) for detecting these changes is to apply a small current at specific frequencies to the cells which can indicate changes in the temperature of the SEI layer on the anode (Figure 8). If the onset of the exothermic heating of the cell can be detected and then mitigated, thermal runaway can be avoided even if soft/hard shorts are being developed. Some of the mitigation steps could be isolating the section of the pack affected from the other parts of the pack, increased cooling to the affected cells, and reduction of the peak currents/powers demanded of the pack or if it can be done safely as in a hybrid, open all contactors to disconnect the battery from the powertrain with continued cooling.

Unfortunately, there is little information/data in the literature showing that thermal runaway has been avoided in particular cases using any of the approaches and detectors discussed. All the electric vehicles being sold with lithium-ion batteries (Leaf, Volt, EVFocus, and EVFit) have BMS units to protect the batteries, but these units seem to only monitor the cells and report any abnormal voltages and temperatures to the vehicle controller, which alerts the driver and reduces the maximum power demanded from the battery. In addition, in nearly all cases, the energy used from the batteries is limited to 60–70% of the total stored in the batteries. This is done by not fully charging the battery and terminating the discharge before the battery is fully discharged. Most of the damage and stress to the cells occurs near full charge and full discharge, and these conditions are avoided by limiting the useable energy from the battery. In the future, when better instrumentation for diagnosing potential serious battery problems becomes available, it is likely that vehicle manufacturers will utilize a greater fraction of the total energy stored in the battery.

8 SUMMARY AND CONCLUSIONS

In this article, various aspects of the safety of lithium-ion batteries of different chemistries have been reviewed based on information available in the literature as of early 2013. The primary safety concern is the possibility of encountering a thermal runaway condition unexpectedly during otherwise normal operation of the batteries. This concern has been evaluated both through study of the abuse testing of cells and computer modeling of cell response to soft/hard shorts.

The abuse testing has shown that thermal runaway can be induced for all cells and chemistries under extreme conditions, but there are significant differences in the tendencies of the different chemistries to experience thermal runaway. The most informative testing has been done using ARC in which the cell temperature can be tracked under known conditions (heat input and heat loss). These tests indicate that thermal runaway in almost all cases

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

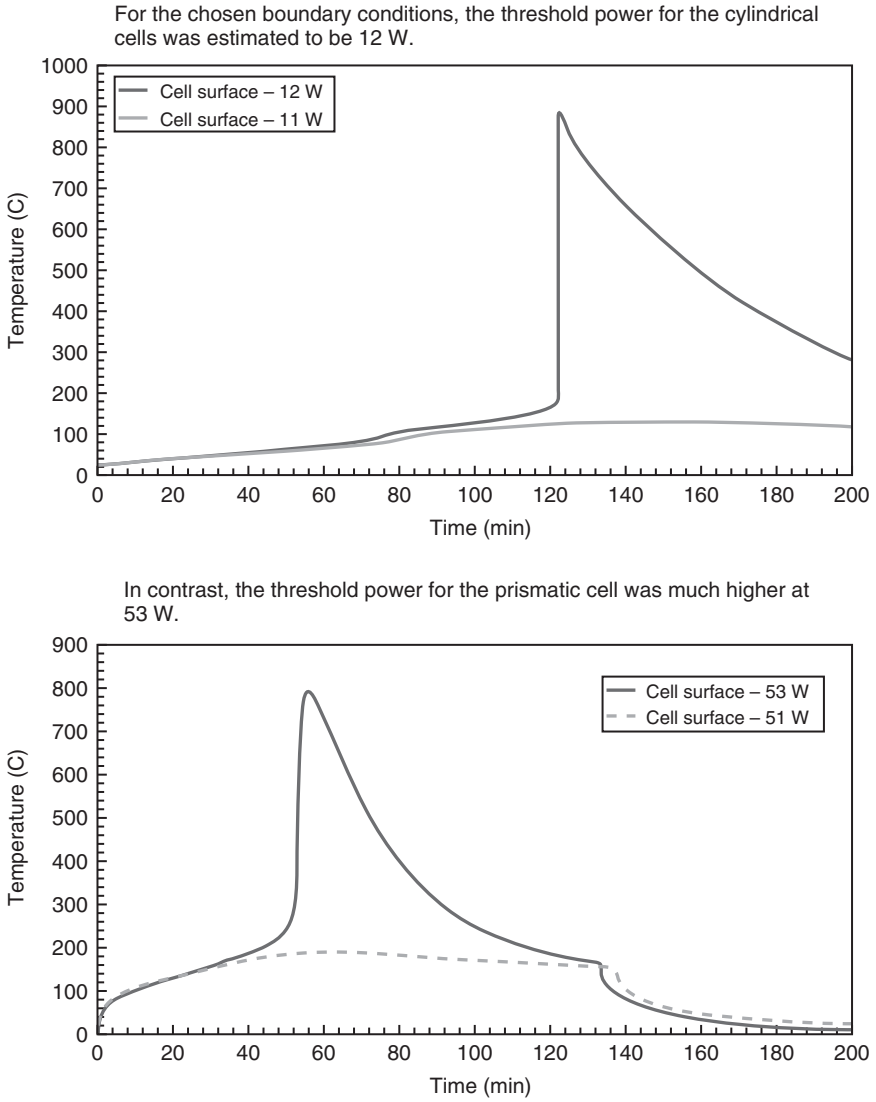


Figure 15. Thermal runaway characteristics of cylindrical and prismatic cells. (Reproduced from Barnett (2011). © the Knowledge Foundation.)

is preceded by a relatively long period (many minutes) of exothermic heating of cell. The temperature increase during the thermal runaway event is very rapid, lasting only seconds reaching temperatures of over 500°C in some cases. Under many conditions, the period of exothermic heating does not lead to thermal runaway. In these cases, the heat loss during the exothermic heating period results in a maximum temperature of <150°C and an uneventful failure of the cell with venting but no fire or explosion. Thermal runaway is most likely with $LiCoO_2$ cells and least likely with cells using Fe PO_4 and Mn oxide cathodes and/or lithium titanate oxide anodes. Thermal runaway is due primarily to the reaction of the electrolyte with the electrode materials at elevated temperatures and subsequent combustion of the electrolyte with the resultant products of the reactions. Any steps to reduce heat generation from these reactions and to increase the heat loss from the cell will minimize the possibility of thermal runaway.

Most of the studies of the response of cells to internal shorts have involved computer simulations. These studies have shown that for all chemistries, there are combinations of the geometry and energy release through the "short" and heat loss from the cell, which will lead to thermal runaway. Even in the case of shorts, the simulations indicate a relatively long period of exothermic heating (gradual temperature increase) before a sudden thermal runaway event. Adequate heat loss (cooling) during the exothermic heating time is critical for mitigating thermal

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

runaway. Hence, designing the cells for good heat loss and/or cooling and monitoring the temperature of the cells via a BMS so it becomes known quickly if/when the temperature (probably of a terminal) exceeds a specified value (probably 85–90°C) will greatly reduce the probability of thermal runaway due to internal shorts for lithium battery chemistries.

RELATED ARTICLES

EV Powertrain Parameters
Basic Consideration
Range Extender EV
Rechargeable Battery Basics
Advanced Batteries for Vehicle Applications
Batteries Indication and Management

REFERENCES

Barnett, B. (2011) Safety of Lithium-ion PHEV Cells: Cylindrical Versus Prismatic. *Proceedings of the 2nd Annual International Conference on Battery Safety 2011: Advancements in System Design, Integration, and Testing for Safety and Reliability*, Knowledge Foundation Conference, November 9–10, Las Vegas, NV.

Barnett, B. and Siriramulu, S. (2011) New Safety Technologies for Lithium-ion Batteries. Presentation at the 28th International Battery Seminar, March 14, Fort Lauderdale, Florida.

Barnett, B., Siriramulu, S., Stringfellow, R. *et al*. (2009) How to Mitigate/Prevent Safety Incidents in Lithium-ion Cells and Batteries. Presentation at the 26th International Battery Seminar, March, Fort Lauderdale, Florida.

Chen, S.C., Wang, Y.Y. and Wan, C.C. (2006) Thermal analysis of spirally wound lithium batteries. *Journal of the Electrochemical Society*, **153** (4), A637–A648.

Darcy, E. and Smith, K. (2010) Advanced mitigating measures for cell internal short risk. Electric Aircraft Symposium, April 23, Rohnert, California.

Deveney, B., Nechev, K., Guseynov, T. *et al*. (2007), *Large size lithium ion cells based on $LiFePO_4$ cathode material*. Saft presentation at 10th Electrochemical Power Sources R&D Symposium Program, August 20–23, Williamsburg, VA, USA, http://www.11ecpss.betterbtr.com/pdf%20folder%20tuesday/T13-Deveney 10th%20Electrochemical%20Power%20Sources%20R&D%20 Symposium%20BD.ppt.pdf (accessed 17 January 2014).

Doughty, D.H. (2012) Vehicle battery safety roadmap guide. NREL Report NREL/SR-5400-54404, October, http://www.nrel.gov/docs/fy13osti/54404.pdf (accessed 17 January 2014).

Fang, W., Kwon, O.J. and Wang, C.Y. (2010) Electrochemical-thermal modeling of automotive Li-ion batteries and experimental validation using a three-electrode cell. *International Journal of Energy Research*, **34**, 107–115.

House, V.E. (2007) Nano-based Lithium-ion (Titanate Oxide) Batteries for Electric Vehicles, EPA Pollution Prevention Through Nanotechnology, September 25–26, Arlington, Virgina.

Keyser, M., Kim, G.H. and Pesaran, A. (2011) Numerical and Experimental Investigation of Internal Short Circuits in a Li-ion C. Presentation at the 2011 DOE Vehicle Technologies Program Review.

Kim, G.H., Pesaran, A. and Spotnitz, R. (2007) A three-dimensional thermal abuse model for lithium-ion cells. *Journal of the Power Sources*, **170**, 476–489.

Kim, G.H., Pesaran, A. and Smith, K. (2008) *Thermal abuse modeling of Li-ion cells and propagation in modules*. 4th International Symposium on Large Lithium-ion battery Technology and Application, Tampa, florida, May 13–18, http://www.nrel.gov/vehiclesandfuels/energystorage/pdfs/43186.pdf (accessed 17 January 2014).

Mikolajczak, C., Harmon, J., White, K., *et al*. (2010) Detecting Lithium-ion Cell Internal Faults in Real Time. *News Release* (Mar 1),

Mikolajczak, C., Kahn, M., White, K. *et al*. (2011) Lithium-ion Batteries Hazard and Use Assessment, prepared by the Exponent Failure Analysis Associates, Inc. for the Fire Protection Research Foundation, July.

Proceedings of Battery Safety 2011-Advancements in System Design, Integration, and Testing for Safety and Reliability, Knowledge Foundation Conference (2011), November 9–10, Las Vegas, NV.

Roth, P. (2009) Abuse Testing of High Power Batteries, Sandia Laboratory presentation, May 19.

Roth, P., Wunsch, T. and Orendorff, C. (2009) Sandia Battery Abuse Testing Laboratory (BATLab), Sandia Laboratory presentation.

Sanyo (2012) Data Sheet, LV51138T, 2-cell Lithium-ion Secondary Battery Protection IC, specification and circuit diagrams.

Schaffhauser, D. (2012) John Hopkins Sensor Detects Overheating in Li-ion Batteries, Campus Technology, John Hopkins University.

Smith, K., Kim, G.H., Darcy, E., *et al*. (2010) Thermal/electrical modeling for abuse-tolerant design of lithium ion batteries, *International Journal of Energy Research*, **34**, 204–215.

Takami, N., Inagaki, H., Kishi, T., *et al*. (2009) Electrochemical kinetics and safety of 2-volt class Li-ion battery system using lithium titanium oxide anode, *Journal of the Electrochemical Society*, **156**, A128–A132.

Texas Instruments, bq6400 (2012) Single Chip or 4 cell Li-ion Battery management Controller with Power Pump Cell Balancing Technology, specification and circuit diagrams.

Thermal Hazard Technology (2012) www.thermalhazard technology.com (accessed 17 January 2014).

Thermal Hazard Technology (2012), Technical Information Sheets, No. 66, Application of the Accelerating Rate Calorimeter for Safety Testing of lithium-ion Batteries.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto065
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fuel Cell Powered Vehicles

Hengbing Zhao and Andrew Burke

University of California, Davis, CA, USA

1 Introduction 1
2 The Fuel Cell Vehicle Driveline and its Operation 1
3 The fuel cell system and optimal operation 6
4 Fuel Cell Vehicle Control Strategy and Simulation 10
References 17

1 INTRODUCTION

Fuel cell vehicles (FCVs) utilize a fuel cell with hydrogen as the fuel to produce electricity that is used by an electric motor to power the vehicle. The fuel cell system consists of a hydrogen storage tank, fuel and air supply circuits, the fuel cell stack, and water and thermal management subsystems. Owing to the highly transient operating conditions in automotive applications, hybridization using batteries or ultracapacitors is often adopted to reduce electrical stresses on the fuel cell and improve its life.

Section 2 of this chapter deals with FCV powertrain arrangements and several other system aspects such as onboard hydrogen storage and refueling. In Section 3, the fuel cell system architecture and operating principles are introduced, and various operating modes and optimum performances of a fuel cell system are discussed based on an optimum operating approach. Section 4 explores different power splitting strategies between the fuel cell and electrical energy storage (batteries or ultracapacitors) and the effect on the fuel economy of FCVs over various driving cycles. The fuel economy of FCVs is then compared with that of vehicles utilizing other advanced driveline technologies.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 THE FUEL CELL VEHICLE DRIVELINE AND ITS OPERATION

2.1 Powertrain arrangements

In automotive applications, fuel cell systems must be able to adapt to challenging operating conditions such as frequent start-up and shutdown and rapidly varying power demand. These conditions are easier to cope with if the fuel cell system is hybridized using batteries or ultracapacitors. In addition to mitigating the stress on the fuel cell via load leveling or power assist, the energy storage permits the capture of regenerative braking energy, which benefits vehicle fuel economy and can potentially permit downsizing the fuel cell system. Designers have a number of choices (Gao, 2005; Zhao and Burke, 2010a, 2010b; Thounthonga, Raël, and Davat, 2009; Uzunoglu and Alam, 2007; Lin *et al*., 2006; Bauman and Kazerani, 2008; Garcia-Arregui, Turpin, and Astier, 2007; Jeon, 2005; Schaltz, Khaligh, and Rasmussen, 2008; An, Lee, and Kim, 2008; Zhao, Burke, and Miller, 2011; Emadi *et al*., 2005; Lai and Nelson, 2007; Jeong and Oh, 2002; Ohkawa, 2004; Zolot, Markel, and Pesaran, 2004) to hybridize the fuel cell. These alternatives include the physical arrangement of the power sources, selection of the energy storage technology and devices, and the control strategy for splitting power between two power sources. Hence, there are several practical arrangements of the power sources (Figure 1). Each arrangement has its advantages and disadvantages relative to operating conditions, control complexity, development cost, vehicle performance, and fuel economy potential.

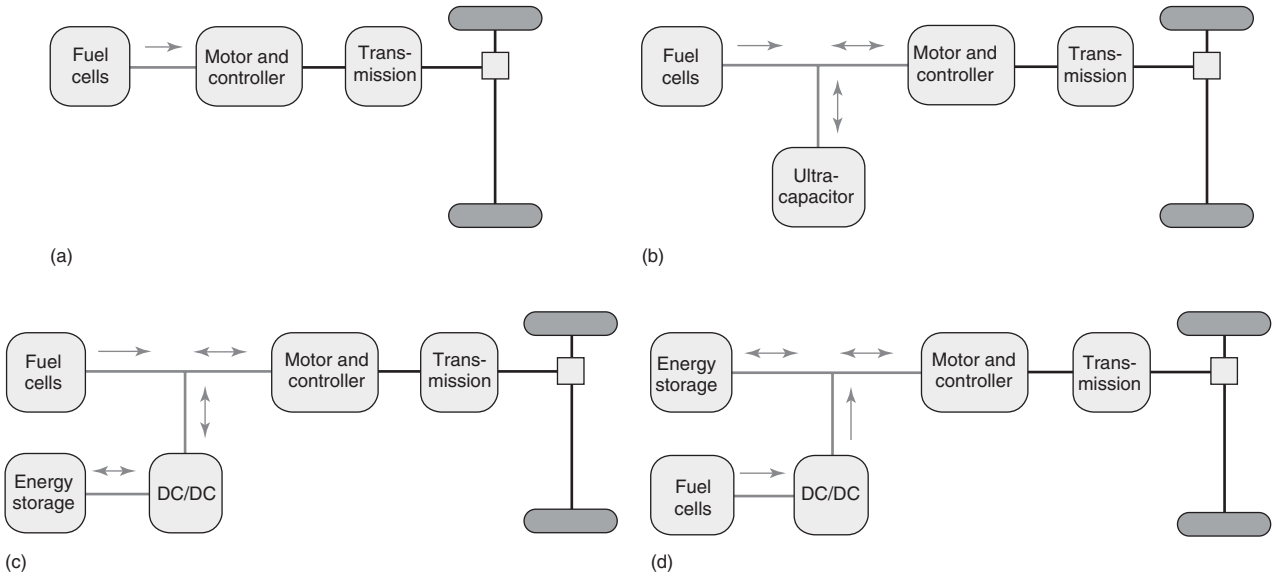


Figure 1. Powertrain configurations for fuel cell vehicles.

(a) *A direct hydrogen FCV without electrical energy storage*. This configuration is the simplest. No DC/DC converter is employed to control the DC-link voltage resulting in the fuel cell stack voltage being equal to the DC-link voltage. Because the dependency of fuel cell output current on the hydrogen and airflow rates limits its response to load transients, this configuration requires a higher power fuel cell stack and fast hydrogen and air supply systems to satisfy the large variations in load power. The DC-link voltage can experience large swings because of the slow response of the air supply system. This configuration will be used as the baseline for comparing different powertrain configurations (Zolot, Markel, and Pesaran, 2004; Zhao and Burke, 2009a).

(b) *FCVs with supercapacitors directly connected to fuel cells*. The supercapacitors are directly connected in parallel with the DC-link (fuel cell stack). In this case, the voltage of the ultracapacitor unit and fuel cell is equal. The relatively soft voltage–current characteristics of fuel cell allow supercapacitors to operate over a fairly wide range of voltages and to self-regulate the DC-link voltage fluctuation. The supercapacitors will absorb the excess power from the stack and the regenerative braking energy and provide a fraction of transient power for vehicle acceleration. A diode is utilized between the fuel cell and DC-link to prevent current from flowing into the fuel cell during regenerative braking of the vehicle. This configuration is the simplest of the hybridized powertrain arrangements (Zhao and Burke, 2010a; Garcia-Arregui, Turpin, and Astier, 2007; Lai and Nelson, 2007).

(c) *FCVs with electrical energy storage (supercapacitors or batteries) coupled in parallel with fuel cell stack through a DC/DC converter*. The fuel cell voltage is the DC-link voltage. The transient power provided by the energy storage is regulated by the DC/DC converter. The introduction of the DC/DC converter will maximize the utilization of supercapacitors or batteries during acceleration and cruise and regenerative braking. This configuration permits controlling the transient power from the fuel cell by applying different power split strategies such as power assist or load leveling to mitigate the stress on the fuel cell stack (Gao, 2005; Zhao and Burke, 2010b; Emadi *et al*., 2005; Lai and Nelson, 2007; Ohkawa, 2004). The state of charge (SOC) of supercapacitors or batteries can also be controlled within appropriate ranges.

(d) *FCVs with the fuel cell coupled with energy storage unit such as supercapacitors or batteries through a DC/DC converter*. The energy storage voltage is the DC-link voltage. The power provided by the fuel cell passes through the DC/DC converter (Zhao and Burke, 2010b; Emadi *et al*., 2005; Jeong and Oh, 2002; Zolot, Markel, and Pesaran, 2004). The converter regulates the fuel cell power to avoid large fluctuation of the DC-link voltage. The SOC of the battery or supercapacitor is also a factor that is determining the fuel cell output power.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Over the years, one of the auto companies developing FCVs has adopted most of these arrangements. The present Honda FCX Clarity, Toyota FCHV, and GM Fuel Cell Equinox Vehicles use the powertrain arrangement shown in Figure 1c. The energy storage is connected to the fuel cell via a DC/DC converter and acts as a power assist or load leveling unit. Ford developed the first research prototype pure FCV without hybridization (Figure 1a) in 1999. Ford's latest HySeries Drive uses powertrain configuration shown in Figure 1d, with a hydrogen fuel cell that operates as an onboard charger. Fiat developed in 2001 an APU (the first generation, Seicento Phase I) using the powertrain architecture shown in Figure 1d. The fuel cell acts as a battery charger, in order to extend the range of the vehicle. In 2003, Fiat presented a Seicento Elettra (the second generation) with a hybrid configuration shown in Figure 1c where the battery acts as a load leveling unit. The third prototype Panda Hydrogen uses the pure fuel cell powertrain architecture without the assistance of a battery or an ultracapacitor because of the improvement in fuel cell performance and cost.

2.2 Onboard hydrogen storage

Most FCVs operate on gaseous hydrogen from a hydrogen tank and oxygen from ambient air. The lower heating value of hydrogen is 121 MJ/kg and that of gasoline is 121.7 MJ/gal. Therefore, 1 kg of hydrogen has the same energy content as 1 gallon of gasoline. Hence, it is easy to relate the mpkgH of hydrogen in an FCV to the equivalent miles per gallon of gasoline. For a specified range RH_2 of the FCV, the weight of hydrogen to be stored is simply

$$WH_2 = RH_2/\text{mpkgH}$$

Hence, for a desired range of 300 miles in an FCV having a fuel economy of 80 mpkgH,

$$WH_2 = 3.75\ \text{kgH}_2$$

Hydrogen has good energy density by weight, but poor energy density by volume (MJ/L). As a result, it is not easy to store 3–5 kg of hydrogen onboard a mid-size passenger car. Hydrogen can be stored using different approaches to reduce storage volume in automotive applications (Broom, 2011; Burke and Gardiner, 2005). These approaches can be classified into two groups: physical storage (gas, liquid, or absorption) and chemical storage (hydrides). Hydrogen can be stored at high pressure in cylindrical tanks wound from carbon filaments. Compressed hydrogen at 350 bar (5000 psi) and 700 bar (10,000 psi) is a well-developed approach being used for by most car manufacturers in their FCVs—Honda, Toyota, Nissan, Daimler, and GM. The present technology for compressed gas hydrogen storage is 66 L/kg at 350 bar and 40 L/kg at 700 bar (Gardiner and Cunningham, 2001). Hence, storing 3.75 kgH_2 would require 250 L at 350 bar and 150 L at 700 bar. Hydrogen can also be stored as a liquid at cryogenic temperatures—at about 20°K. The volume required to store hydrogen as a liquid is 20–25 L/kgH_2 resulting in 75–95 L to store 3.75 kgH_2 (Gardiner and Cunningham, 2001). Storing hydrogen as a cryogenic liquid is presently the most volume efficient method, but the cryogenic systems are much less convenient than the compressed gas systems for vehicle applications. Only BMW has adopted liquid storage of hydrogen in their prototype vehicles.

Hydrogen can also be stored using gas absorption on materials with a large specific surface area such as activated carbon (Gardiner, 2004; Sevilla, Foulston, and Mokaya, 2010). Chemical storage in the form of metal hydrides in which the hydrogen is chemically bonded in covalent and ionic compounds at controlled temperature and pressure are also being studied (Young *et al*., 2004; Mori *et al*., 2005). In these cases, the hydrogen is released by heating the host material. The hydrides have good volume storage characteristics (15–20 L/kgH), but can be heavy (50–60 kg/kgH in the case of low temperature hydrides). High temperature hydrides (>300°C) can have reasonable good weight characteristics (15 kg/kgH).

A comparison of various hydrogen storage approaches for FCVs is shown in Table 1. Liquid hydrogen shows the highest system gravimetric and volumetric capacities. However, current liquid hydrogen systems use more than

Table 1. Comparisons of system metrics for various hydrogen storage technologies with the DOE goals (2015 and ultimate).

Hydrogen Storage Approaches	System Gravimetric Capacity (% kg H_2/kg System)	System Volumetric Capacity (g H_2/L System)
	2015 Goal: 5.5 Ultimate Goal: 7.5	2015 Goal: 40 Ultimate Goal: 70
Compressed gas		
5000 psi	3–4	15
10,000 psi	4–5	25
Liquid (LH_2) (20°K)	5–10	40–50
Activated carbon (77°K)	5	25
Hydrides		
Low temperature (<100°C)	1.8	70
High temperature (>300°C)	5.5	55

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Comparisons of the system energy densities of battery and hydrogen storage technologies.

Storage technologies	Watt-Hour per Kilogram	Watt-Hour per Liter
Batteries		
Lead acid	30	70
NiMt hydride	70	175
Lithium-ion	100	200
Compressed H_2		
5000 psi	1500	500
10,000 psi	1166	850
Liquid H_2	1600–3200	1500
Metal hydrides		
100°C	600	2000
300°C	1850	1600
Activated carbon	1665	850
Gasoline	11,660	8750
DOE goals for hydrogen storage (2015/ultimate)	1800/2500	1300/2300

30% of the energy in the hydrogen for liquefaction and cannot meet the DOE long-term targets for volumetric capacity. Also shown in the table are the DOE system development goals (Dillich, 2009; DOE report, 2009) for weight and volume. If the ultimate targets are met, the system volume to store 3.75 kgH would be 53.6 L with a weight of 50 kg. Comparisons of energy storage in hydrogen with various battery chemistries and liquid gasoline are shown in Table 2. Note that energy densities for storage in hydrogen are at least one order of magnitude higher than for lithium batteries. However, compared to gasoline, energy storage in hydrogen is only 0.25–0.30 as good as gasoline even if the DOE ultimate goals are met. As the fuel economy of FCVs is expected to be about twice that of conventional ICE vehicles, the short fall in energy storage would only be about a factor of two in hydrogen-fueled FCVs.

2.3 Fuel cell assembly and operation

2.3.1 Introduction

A fuel cell is a device that converts chemical energy from a fuel directly into electricity via controlled electrochemical reactions at two separate electrodes (Mench, 2008; Srinivasan, 2006; Spiegel, 2008). Fuel cells are usually named based on the electrolyte material between the two electrodes. Two types of fuel cells of particular interest for vehicles are the following: polymer electrolyte membrane fuel cell (PEMFC) and solid oxide fuel cell (SOFC). The SOFC utilizes a ceramic, solid oxide electrolyte and operates at a high temperature—around 1000°C. High temperature operation removes the requirement for a precious metal catalyst in the electrodes and enables the use of a variety of fuels, but results in a slow start-up and critical thermal insulation and sealing issues. Efforts are being made for developing lower temperature SOFC to deal with high temperature-related issues. The PEMFC utilizes a polymer electrolyte membrane and operates at relatively low temperatures—around 80°C. The low temperature operation of PEMFCs requires precious metal catalyst—platinum and also pure hydrogen as a fuel. However, the low temperature operation allows PEMFCs to start quickly and makes them suitable for use in vehicles, especially light-duty vehicles. Research is being done to develop non-noble catalysts to deal with cost and CO poisoning issues and also higher temperature membranes that function over 100°C to improve electrochemical kinetics and simplify water and thermal management. In this chapter, only PEMFCs are considered for the FCV application.

2.3.2 The assembly and operation of PEMFCs

All types of fuel cells have essentially the same structure and arrangement of internal components. Figure 2 shows a schematic of a generic PEMFC. An anode with fuel flow channels, anode gas diffusion layer (GDL), electrolyte, cathode GDL, and a cathode with airflow channels and coolant flow channels. Anode and cathode plates are made of thermal and electrical conductive materials such as graphite, composites, or metals with machined or stamped flow channels. They are designed to accomplish many functions, such as supplying reactants, removing heat and by-product—water, and collecting current and carrying current from cell to cell. The porous GDLs are made of electrically conductive carbon paper or carbon cloth to ensure that reactants are uniformly diffused over the active

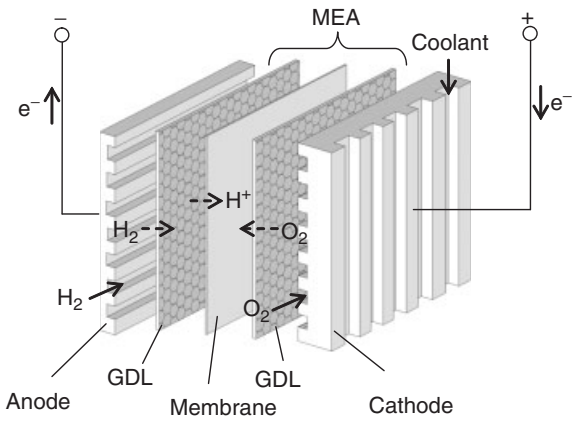


Figure 2. Schematic of a generic fuel cell.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

area of fuel cell electrodes. The GDLs are typically treated with the hydrophobic material—PTFE (Teflon) to ensure that the diffusion pores are not clogged with condensed by-product—water. The membrane of the PEMFC is a type of chemically resistant and durable polymer. The most commonly used membrane is Nafion. It performs as a separator and selectively transports cations—protons (H^+)—across the cell from the anode to the cathode. A thin catalyst layer—a mixture of platinum or platinum alloys, carbon black, and PTFE particles—is applied to the membrane on the anode and cathode sides according to the assembly process (Litster and McLean, 2004). This is the active electrode layer where the electrochemical reaction occurs. The anode GDL, membrane, and the cathode GDL are assembled through hot pressing to form the membrane electrode assembly (MEA).

Hydrogen is delivered to the anode catalyst layer through the anode flow field plate and the anode GDL and split into H^+ and electrons through the electrochemical oxidation reaction. The electrons flow through the anode GDL and the anode plates to the external circuit, whereas the H^+ ions diffuse (conduct) through the polymer electrolyte membrane to the cathode electrode. At the cathode electrode catalyst layer, the oxygen through the cathode flow field plate and the cathode GDL reacts with the H^+ ions that have permeated through the membrane and the electrons from the external circuit (via the adjacent anode plate) through the electrochemical reduction reaction. The products of the fuel cell are by-product water, waste heat, and electricity. The cells in the fuel cell are arranged in a bipolar manner with the electron current flowing from anode to cathode and finally into the external circuit.

A single cell produces less than 1.0 V. Multiple cells are stacked in series into a fuel cell stack to achieve a higher voltage for automotive applications. The output current of a fuel cell stack is proportional to the active area of the individual cells, and the stack voltage depends on the number of the cells in series. The electrical power that can be generated by a fuel cell stack is limited by its electrical resistance, which is dependent on the reaction kinetics and the rates of diffusion of the reactants in the membrane and the electrodes. For example, high pressure and high air stoichiometry improve the kinetics and reactant diffusion in the catalyst layers and enhance removing condensed water from the active area of the cathode, but high pressure and high airflow result in high parasitic losses and reduced system efficiency. Therefore, optimal operation of the fuel cell system is needed to achieve high vehicle fuel economy.

2.3.3 Start-up from subfreezing temperature

The presence of water in the fuel cell stack is one of the major challenges in the operation of FCVs in cold climates. Water is formed at the cathode of the fuel cell as a by-product and needs to be removed to avoid flooding in the cathode GDL. At the same time, the PEM needs to be well hydrated to maintain its high performance. The inherent presence of water creates a problem when FCVs are parked in a subfreezing weather environment. Rapid start-up of a fuel cell in a subfreezing environment is critical to the marketing of FCVs. FCVs require that the stack reaches 90% rated power in less than 30 s at temperatures as low as -20°C. Properly managing the start-up at subfreezing temperatures is necessary for avoiding fuel cell damage.

Different start-up and shutdown approaches have been proposed or adopted to address the rapid start-up in a subfreezing environment (Pesaran, Kim, and Gonder, 2005). The proposed solutions can be categorized into three main strategies: dry gas purge, keep-warm, and thaw and heat at start-up. The dry gas purging process before fuel cell shutdown can minimize residual water in the membrane, porous electrodes, and GDLs. The keep-warm uses thermal insulation to delay the fuel cell cool-down and freezing following system turn-off. Using insulation can delay freezing up to several days, but cannot eliminate freezing. Periodically heating the fuel cell via embedded electric heaters or heated air supply may be used to maintain fuel cell temperature above 0°C for long-term storage in a subfreezing environment. The thaw and heat method uses electric heaters to warm up the fuel cell directly or via the air supply circuit before the vehicle is driven. However, freeze/thaw cycles create a durability issue for the fuel cell system such as increased MEA internal resistance because of internal ice expansion and require a long warm-up period. Appropriate combinations of the dry gas purging and the thermal management in the fuel cell are likely the best approach to solve the rapid start-up issue.

2.4 Hydrogen refueling

Currently, prototype FCVs store compressed hydrogen in carbon-fiber-reinforced composite tanks or as liquid hydrogen in a cryogenic tank. These vehicles can achieve ranges up to about 400 miles. They require hydrogen refueling stations that provide pressurized hydrogen gas at high pressure (up to 5000–10,000 psig) and/or liquefied hydrogen at low temperature (-253°C). Hence, a new infrastructure (facilities and systems) must be constructed for producing and delivering hydrogen. During the early

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Comparison of various hydrogen refueling alternatives.

Hydrogen Refueling Methods	Overall Energy Conversion Efficiency[a]	Cost of Hydrogen[b]
Bulk liquid hydrogen is produced from natural gas from an existing central reformer, transported to a refueling station by truck, stored as a cryogenic liquid, and dispensed to the vehicle as a liquid or converted to a compressed gas before dispensing to the vehicle	—	\$2.3–2.4/kg
Bulk gaseous hydrogen is produced from natural gas from an existing central reformer, transported to the refueling station by existing pipeline, compressed on-site and stored as a compressed gas at 500 psi, and dispensed to the vehicle as a gas	—	~\$3.2/kg
Natural gas is transported to the refueling station by existing pipeline, hydrogen is produced by steam methane reforming or partial oxidation process, compressed and stored at 5000 psi, and dispensed to the vehicle as a gas	67% for methane reforming 69% for partial oxidation process	~\$4.7/kg for methane reforming ~\$4.4/kg for partial oxidation process
Gaseous hydrogen generated at the refueling station by electrolysis using grid electricity, stored as a compressed gas at 5000 psi, and dispensed to the vehicle as a gas	80%	~\$3.7/kg

[a]Efficiency is based on the higher heating value.
[b]On the basis of the value of the US dollar in 2001.

adoption of the FCVs, it seems likely that hydrogen refueling facilities/stations will be provided in clusters around some pilot cities (Yang and Ogden, 2010; Ogden and Yang, 2005). FCVs can also be refueled through mobile refueler. Hydrogen can be produced on-site at the stations from water with an electrolyzer or from natural gas with a reformer, then compressed and stored for dispensing to the vehicle. Hydrogen can also be delivered to the station as a high pressure gas or a cryogenic liquid by a tank truck from a central production site for storage and later dispensing to the FCVs.

Schoenung and Weinert analyzed and compared all hydrogen refueling station alternatives in terms of overall energy conversion efficiency and cost (Schoenung, 2001; Weinert and Lipman, 2006). Various hydrogen refueling approaches and associated energy conversion efficiency assuming 100% utilization of the facility are summarized in Table 3. Schoenung calculated the cost of hydrogen dispensed from the annual operating costs divided by the total GJ of hydrogen delivered per year. The cost of hydrogen is then converted to the cost per kilogram using the lower heating value of hydrogen (120 MJ/kg).

3 THE FUEL CELL SYSTEM AND OPTIMAL OPERATION

3.1 Fuel cell system

The fuel cell stack is the key element of a fuel cell system. However, without auxiliary components such as the air compressor, humidifier, and pressure and flow regulators, the stack itself would not function properly. Fuel cell system configurations vary considerably in different applications. A direct hydrogen fuel cell system, as shown in Figure 3, typically involves the following four major auxiliary subsystems: air supply and control, fuel supply and control, water management, and thermal management subsystems. The air supply subsystem consists of several interacting components—namely, an air compressor and expander, supply and return manifolds, and back pressure control valve (Cunningham, 2001; Chen and Peng, 2005). The fuel supply subsystem consists of a high pressure fuel tank, pressure regulator, supply manifold, hydrogen recirculation pump, and purge control valve. The water management subsystem includes air/fuel humidifiers or vapor injector and vapor condenser. The thermal management consists of the cooling loop for the stack and temperature control for humidifiers and a radiator (Badrinarayanan *et al*., 2001).

Water and thermal management are two of the critical issues in the operation of PEMFCs. The proton (H^+) conductivity of the Nafion membrane (Xu, Kunz, and Fenton, 2006) requires its humidification. The humidity of inlet air stream should be maintained at a high level (>60% RH at 80°C). Temperatures above 80°C will dry the membrane and result in high membrane resistance. High pressure and high flow rate of air will benefit the disposal of the condensed water in the cathode GDL. These subsystems interact and careful attention to these interactions is required. Optimal air and fuel supply and water and thermal management will improve the performance and life of the fuel cell stack and increase efficiency of the fuel cell system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

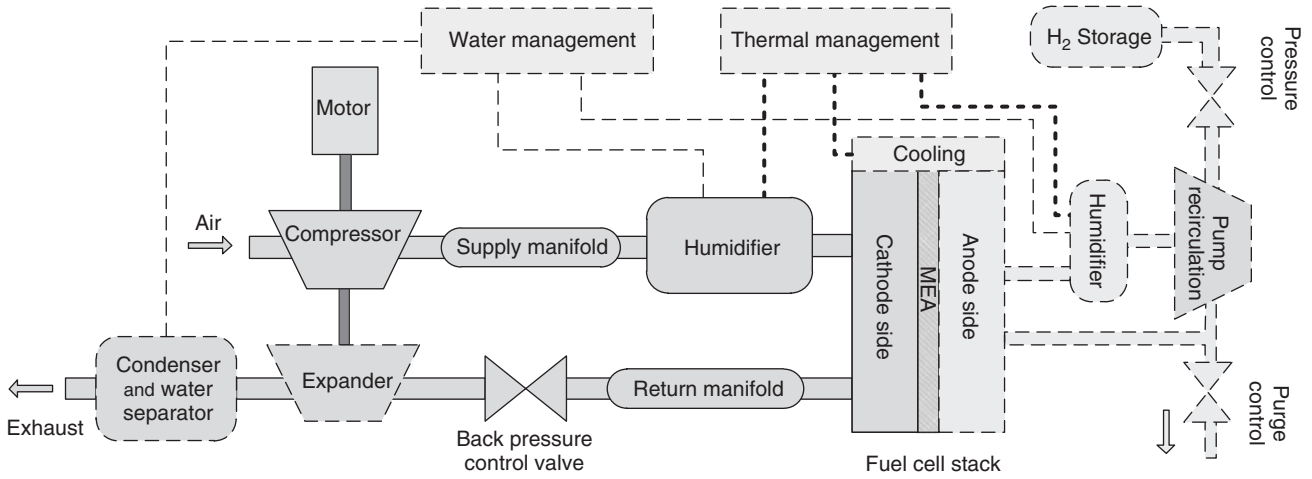


Figure 3. Schematic of a fuel cell system in a fuel cell vehicle.

3.2 Optimal operation

The fuel cell stack can deliver electricity at high efficiency (up to about 60%). However, the operation of the on-board auxiliaries significantly affects the performance and efficiency of the fuel cell system. PEMFC systems can consume up to 20% of the stack output to provide power to auxiliaries such as compressors, heaters, and pumps. The air supply system accounts for about 80% of the parasitic losses and has a dominant impact on the system efficiency and response time. The optimization of the fuel cell system operation is concerned with the analysis of various air supply configurations and their operation to maximize the net system power and system efficiency (Li and Liu, 2009; Sundström and Stefanopoulou, 2007).

There are many considerations for optimizing the operation of a fuel cell system. First, the fuel cell performance is sensitive to the mass flow of the reactants, which depends on the fuel cell stack design and operating conditions. The number of cells, the active area of the cell, and flow field design including the channel shape, dimensions and spacing, and the maximum allowable pressure drop are key stack design parameters. Secondly, in terms of operating conditions, temperature, relative humidity, operating pressure, and the air mass flow are the four key external variables that have a major impact on the performance of the stack. They determine the oxygen partial pressure at the cathode catalyst layer, which determines the resultant cathode overpotential as a function of stack current. Finally, the selection of the main auxiliary device—the compressor—accounts for most of the parasitic losses and determines the two key operating variables—pressure and mass flow.

The compressor that pressurizes and delivers air into the fuel cell has a direct effect on the system efficiency. High pressure operation of the fuel cell can give high power density and better water management, but it results high parasitic losses. There are a variety of compressors available for FCV applications. The turbo compressor and the twin-screw compressor are two well-suitable options for high pressure operation because of their low weight and small size. Kulp (2001) found that the turbocharger is more efficient than the twin-screw compressor, especially at low mass flows. However, a neutral water balance was more difficult to maintain with a turbocharger than in a twin-screw setup. In the present analysis, a twin-screw compressor is used. Pressure ratio (outlet pressure/inlet pressure) and mass flow rate are the two main parameters needed to match the fuel cell requirements.

The pressure drop across the stack depends on the humid air mass flow, stack back pressure, and channel flow field plate design following the Darcy–Weisbach Law. The optimum operating conditions can be determined as follows. The humid air flow rate is calculated from the back pressure, the dry air mass flow rate, and the interpolated pressure drop consistent with the maximum allowable pressure drop (0.4 atm in the present analysis). The optimum pressure drop across the stack can then be determined. For every triplet, the current density (J), the dry air mass flow ($\dot{m}$), the back pressure (P_r), and the net output power of $P_{net}(J, \dot{m}, P_r)$ can be calculated. All the values of $P_{net}(J, \dot{m}, P_r) > 0$ that fall within the safe operational region of the compressor are then scanned to find the one with max $P_{net}(J, \dot{m}, P_r)$.

$$P_{net}(J, \dot{m}_{optimal}, P_{r,optimal}) = \max[P_{net}(J, \dot{m}, P_r)] \quad (1)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

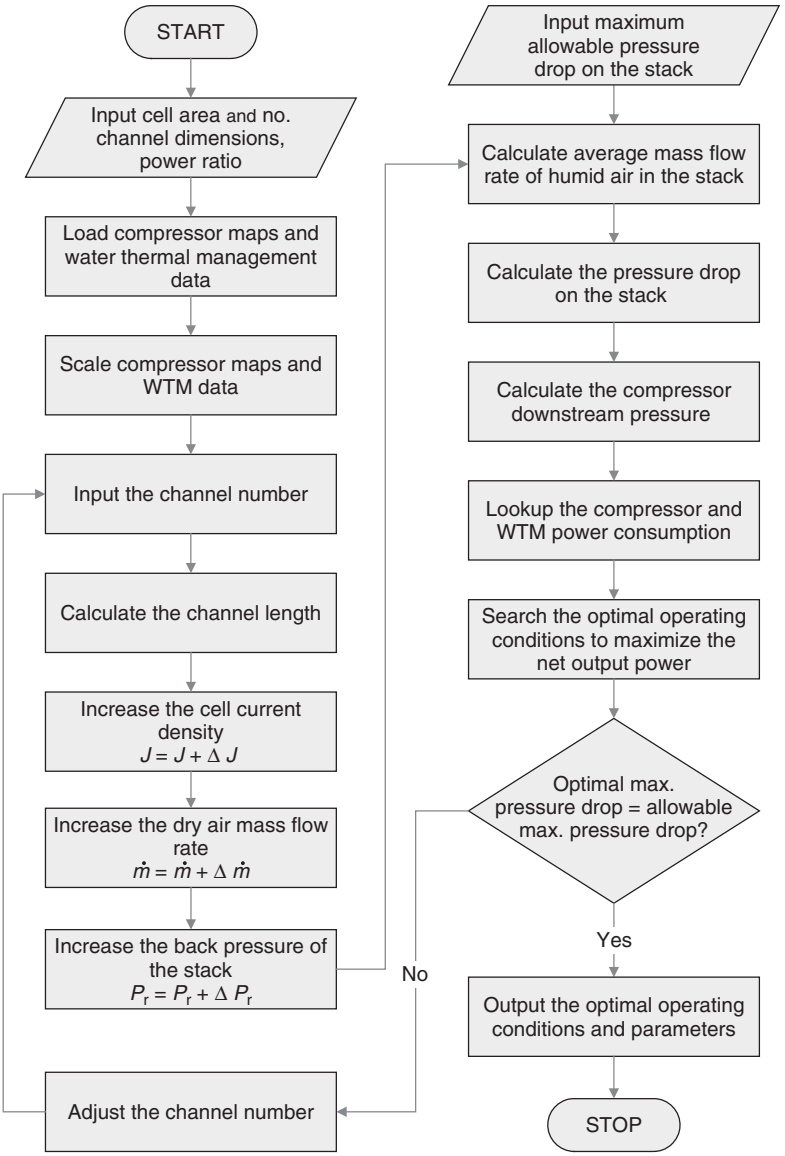


Figure 4. Flowchart of optimization of the fuel cell stack design and operating conditions.

In other words, the optimal mass flow $\dot{m}_{\text{optimal}}$ and back pressure $P_{\text{r,optimal}}$ will yield the maximum net power for each J value. The optimization process is shown in Figure 4 (Zhao and Burke, 2009b).

In the optimization process, the maximum pressure drop is obtained from the optimal results and compared with the input allowable pressure drop. If the maximum pressure drop in the stack matches the allowable pressure drop, the selected channel number is acceptable and the optimum operating conditions are also acceptable. Otherwise, the channel number is varied and the model is run until the maximum optimum pressure drop in the stack is acceptable (less than the allowable pressure drop). The above optimization of fuel cell system operation only gives an approximate calculation. For detailed analysis, commercial computational fluid dynamic (CFD) software programs such as CFD−ACE+, CFX, Flow-3D, and Fluent are usually employed to analyze the microfluidic flow in the diffusion layer of fuel cells and verify the cell and stack design.

3.3 Comparison of fuel cell operation modes

In a fuel cell system, a pressure control valve or regulator is used to control the back pressure. The back pressure can be kept constant at varied gas flow conditions (fixed back pressure operation) or can be varied with the change

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of the load (optimal varying back pressure operation). For both cases, the pressure difference across the membrane is minimized to reduce the stress on the membrane electrode assembly and the airflow or air stoichiometry ratio (SR) is optimized by avoiding operation of the compressor at low mass flow rate.

It is of interest to compare fixed and varying pressure operation of the system. Hence, the optimization approach was used to optimize the air SR for fixed back pressure operation and the SR and back pressure for optimal varying back pressure operation. For this analysis, a fuel cell system having 440 cells with the active area of 510 cm^2 and employing a twin-screw compressor is optimized. The fuel cell system was first optimized for varying SR and back pressure operation mode. Then the same system was optimized at the fixed back pressures of 2.0, 1.5, and 1.1 atm (Zhao and Burke, 2009b). A plot of system efficiency versus system net power is shown in Figure 5. The optimal polarization curves, the compressor responses, and the pressure drop across the stack for different operating modes are shown in Figures 6–8, respectively.

The comparisons of the results for the different operating modes indicate the following:

- The fuel cell system with optimal varying back pressure can achieve higher system efficiency over the full load range (Figure 5) and can produce more power than the fuel cell system operating at constant back pressures.
- For the same fuel cell system with different operating modes, lower constant back pressure operation has higher pressure drop across the stack than other operating modes because of higher ratio of water vapor partial pressure to dry air partial pressure (Figure 8).
- At low power demand, the fuel cell system operating at low pressure and at optimal back pressure has higher system efficiency than the fuel cell system operating at high pressure because of relatively low parasitic losses.
- At high power demand, the high pressure operating mode and the optimal varying back pressure operation mode can achieve higher system efficiency compared to the low pressure operation because of the high oxygen

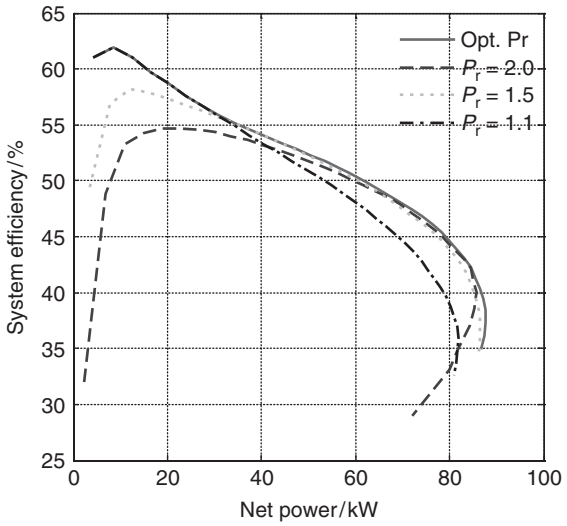


Figure 5. Comparison of the optimal system efficiency for different operating modes: optimal varying back pressure operation and fixed back pressure of 2.0, 1.5, and 1.1 atm.

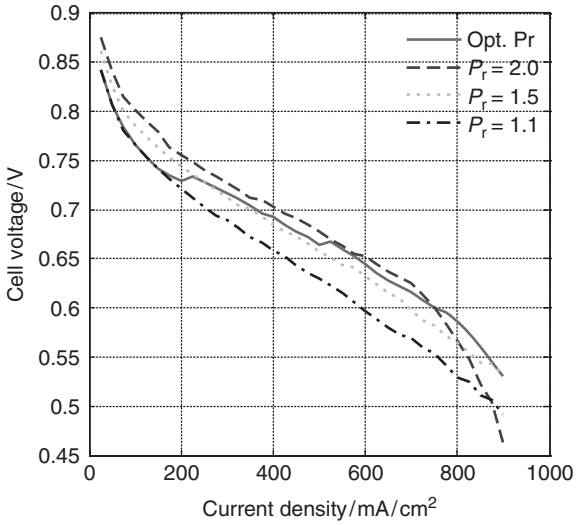


Figure 6. Optimal fuel cell polarization curves for different operating modes.

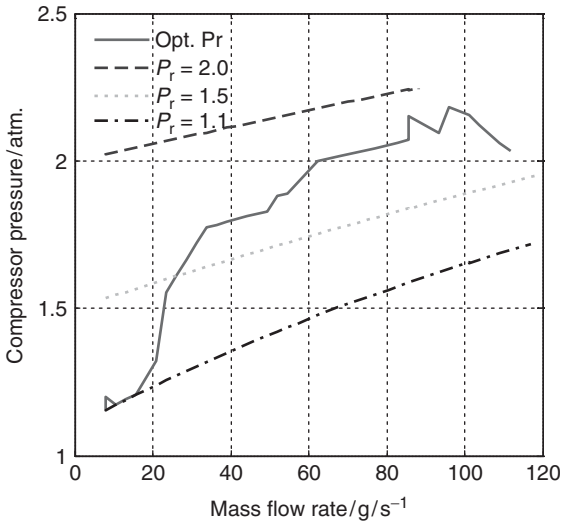


Figure 7. Optimal compressor quasi-steady responses for different operating modes.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

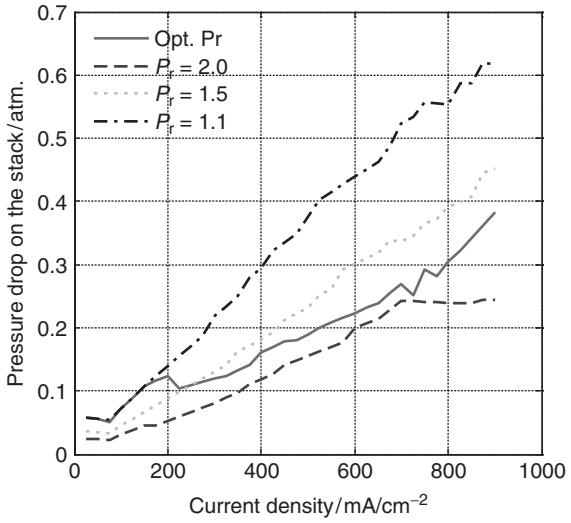


Figure 8. Pressure drop across the stack versus current density.

partial pressure at the catalyst layer and low pressure loss on the stack.

- At medium load demand, there is no apparent difference in the system efficiency for the different operation modes. However, low pressure operation requires a much larger humidifier than high pressure operation.
- The optimal operation can achieve higher efficiency over wide load change. However, coordinated control of the compressor and the back pressure valve can be complicated and is needed to avoid large transient voltage drops during rapid changes in power demand.

4 FUEL CELL VEHICLE CONTROL STRATEGY AND SIMULATION

4.1 Vehicle power splitting strategy

As discussed in the previous section, optimal operation of a fuel cell system varies the back pressure and air supply SR according to the change of the power demand. Coordinated control of the mass flow and pressure of the cathode and anode sides of the stack is required. This is the main drawback of the optimal operation of the fuel cell system for automotive applications. These rapid changes in the operating conditions of the fuel cell stack can have a major impact on the lifetime of the fuel cell stack because of the mechanical and thermal stresses on the MEA and the stack accessory components. Hybridization of the vehicle power train is an effective approach to mitigate the stress on the fuel cell stack by shifting most of the dynamic power demand to a second power source such as batteries and/or ultracapacitors. Another advantage of hybridization of an FCV is the ability to recover energy while decelerating through regenerative braking. In the hybrid configuration, the total power demand from the vehicle is satisfied by splitting the power between the fuel cell stack and the second power source, usually a battery pack. The power split strategy has a significant effect on the dynamics of the power demands of the fuel cell stack and the battery pack. The primary factors of interest for different power split strategies are impacts on the sizing of the power sources, durability of the fuel cell stack and battery, and vehicle fuel economy.

The fuel cell operation (power, voltage, current vs time) and hydrogen consumption (fuel economy) are closely related to the strategy utilized to split power between the fuel cell and the energy storage as the vehicle is operated over various driving cycles. The general objective of any control strategy is to operate the fuel cell system only in its high efficiency region, avoiding operation in the very low power and very high power regions. Power-assisted control and load leveling control strategies can be used in hybrid FCVs (Zhao and Burke, 2010a). Power-assisted control splits the power/current demand of the traction motor i_{motor} based on the fuel cell voltage V_{fc} and the energy storage SOC. The current command for the energy storage device i_{ess} is expressed in Equation 2 with the fuel cell providing the remaining current (Equation 3).

$$i_{\text{ess}} = f_{\text{fc}}(V_{\text{fc}}) \cdot f_{\text{ess}}(\text{SOC}) \cdot i_{\text{motor}} \tag{2}$$

$$i_{\text{fc}} = i_{\text{motor}} - i_{\text{ess}} \tag{3}$$

where f_{fc} and f_{ess} are factors related to fuel cell voltage and energy storage device SOC, respectively. If the fuel cell voltage remains relatively high, it will provide most of the current to the motor. When the fuel cell voltage becomes low, the energy storage device will provide a large fraction of the current demanded by the motor. Figure 9 shows an example of the splitting factors of f_{fc} and f_{ess} used in a fuel cell-battery hybrid vehicle with power-assisted control.

For load leveling control, the fuel cell provides relatively steady power and the energy storage device provides transient power. The fuel cell current command i_{fc} is calculated by averaging the traction motor current i_{motor} over a specified time period such as 60 s.

$$i_{\text{fc}} = i_{\text{av,60s}} \tag{4}$$

$$i_{\text{ess}} = i_{\text{motor}} - i_{\text{av,60s}} \tag{5}$$

The implementation of the control strategy for power split is shown in Figure 10. Both control strategies maintain

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

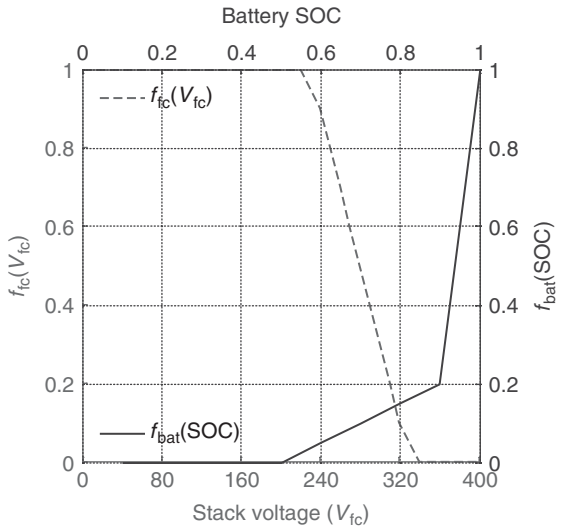


Figure 9. Power split factors for power-assisted control.

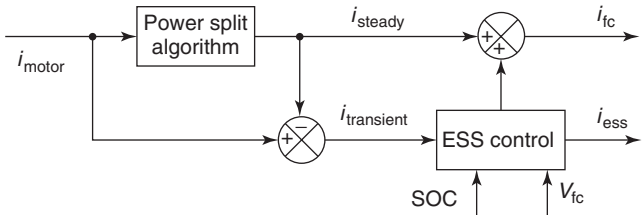


Figure 10. Schematic of power split control.

the SOC of the battery or supercapacitor within a specified range. Compared to the power-assisted control, load leveling control permits the fuel cell to operate within a relatively narrow high efficiency region. This mitigates the stress on the fuel cell and maximizes fuel cell life by utilizing the energy stored in the battery or supercapacitor to meet high power transients. However, a significant fraction of the transient power passes through the DC/DC converter for leveling the fuel cell current, which introduces significant losses in the power electronics.

For FCVs with supercapacitors coupled directly in parallel with the fuel cell, no DC/DC converter is employed. The voltages of the supercapacitor unit and fuel cell are equal. The current of the supercapacitors is governed by the differential Equation 6.

$$\frac{dV}{dt} = \frac{i_{cap}}{C} + R_{cap}\frac{di_{cap}}{dt}, V = V_{cap} = V_{fc} \qquad (6)$$

This arrangement avoids the losses in the DC/DC and takes maximum advantage of the high efficiency of the supercapacitors.

4.2 Fuel cell vehicle simulation

4.2.1 Vehicle simulation inputs

Simulations of the operation of fuel cell hybrid vehicles using various drive train arrangements and energy storage technologies (Li-ion batteries and supercapacitors) were performed using the UCD FCV simulation program (Zhao and Burke, 2010a, 2009a, 2009b). Simulations were performed using both the power-assisted and load leveling control strategies. In addition to the choice of drive train arrangements and energy storage technologies, the simulations have been run with different size energy storage units (kWh or Wh). In the simulations, the battery SOC is maintained between 0.6 and 0.8. In addition, the regenerative braking currents are limited to protect the batteries from over voltage. The minimum voltage of the supercapacitor is set at 50% of its rated voltage, which limits the maximum usable energy to 75% of the total energy stored in the capacitor. The supercapacitor SOC $= 1 - \left[\frac{(V_{rated} - V)}{(V_{rated}/2)}\right]$ is controlled to be between 0.95 and 0.2. Test data for the carbon/carbon supercapacitors (Burke and Miller, 2009a) and the lithium batteries (Burke and Miller, 2009b) are used to model the energy storage units.

Simulations were performed for mid-size passenger vehicles without energy storage, with supercapacitors directly connected in parallel with the fuel cell, with the fuel cell connected to supercapacitor DC-link via a DC/DC converter, and with supercapacitors and Li-ion batteries

Table 4. Vehicle simulation parameters.

Vehicle and system parameters	
Drag coefficient	0.3
Frontal area (m^2)	2.2
Rolling resistance	0.01
Vehicle hotel load (kW)	0.3
Vehicle mass without energy storage (kg)[a]	1500
Electric motor (kW)	75
Fuel cell stack and auxiliaries	
Maximum net power (kW)	87.6
Gross power (kW)	106
Number of cells	440
Cell area (cm^2)	510
Compressor (kW)	17.2
Energy storages	
Ultracapacitor capacity (Wh)	80–200
Ultracapacitor module number in series	160[b]
Battery energy capacity (kWh)	0.85–2.0
Battery capacity (Ah)	2.5–5.8
Battery cell number	144

[a] Vehicle mass recalculated based on the size and type of energy storage.
[b] 148 in the case with ultracapacitors connected without interface electronics.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

coupled with the fuel cell through a DC/DC converter. All the drive trains were simulated in the same vehicle having the road load characteristics shown in Table 4. The fuel cell system generated a net output power of 87.6 kW. The total vehicle mass was adjusted to reflect the type and capacity of the energy storage unit and was recalculated based on the specific energy of energy storage units. The rated traction motor power was 75 kW for all cases. An empirical efficiency map of a bidirectional DC/DC converter, indexed by the input/output voltage ratio and the output power, was employed in the simulations. The energy storage current is limited by the maximum charging/discharging current, which is calculated based on the open circuit voltage, the maximum/minimum voltage, and the internal resistance of the energy storage.

4.2.2 *Fuel cell vehicle simulation results*

Energy storage is utilized in the fuel cell driveline both to improve the fuel economy of the vehicle and to increase the life of the fuel stack by reducing the peak currents and current transients that it experiences. First, consider the improvements in fuel economy projected using the various driveline arrangements with lithium batteries and ultracapacitors. Detailed simulation results for the various drivelines for the FUDS and US06 cycles are given in Table 5. The fuel economy results are summarized in Table 6.

Table 5. Comparisons of energy consumptions and efficiencies of fuel cell vehicles using different drivelines with energy storage.

Powertrain Configuration		FCVs w/o ESS	FCVs with UCs Connected to Fuel Cells Directly	FCVs with UCs Connected to Fuel Cell DC-Link via a DC/DC Converter		FCVs with Batteries Connected to Fuel Cell DC-Link via a DC/DC Converter	
Power Splitting Strategy		—	—	Load Leveling	Power Assist	Load Leveling	Power Assist
ESS	Type/Capacity		APowerCap Ultracapacitor 100 Wh			Enerdell Li-ion Battery 1.5 kWh	
FUDS: 5 cycles							
Fuel economy	mpgge	68.0	85.0	78.8	79.2	72.8	78.6
Wheel energy (kWh)	Regenerative	0	−2.985	−2.887	−2.868	−2.216	−2.691
	Accelerating	7.141	7.122	7.179	7.175	7.143	7.148
	Mechanical brake	3.267	0.242	0.372	0.388	1.015	0.545
DCDC and ESS	Loss (kWh)	0	0.060	0.650	0.330	1.003	0.306
Motor efficiency	Regenerative	0	0.814	0.814	0.812	0.732	0.788
	Accelerating	0.784	0.770	0.770	0.770	0.770	0.770
Fuel cell stack	Efficiency	0.572	0.574	0.584	0.568	0.598	0.572
DCDC converter efficiency	Charge efficiency	—	—	0.941	0.948	0.936	0.940
	Discharge efficiency	—	—	0.950	0.874	0.951	0.864
	Round-trip efficiency	—	—	0.894	0.829	0.890	0.813
Energy storage efficiency	Charge efficiency	—	0.993	0.990	0.993	0.977	0.979
	Discharge efficiency	—	0.992	0.989	0.997	0.955	0.990
	Round-trip efficiency	—	0.985	0.980	0.990	0.933	0.969
US06: 5 cycles							
Fuel economy	mpgge	50.7	59.6	55.0	57.3	51.9	56.6
Wheel energy (kWh)	Regenerative	0	−2.193	−1.905	−2.356	−1.048	−2.221
	Accelerating	11.418	11.384	11.452	11.455	11.354	11.386
	Mechanical brake	3.462	1.211	1.540	1.092	2.307	1.167
DCDC and ESS	Loss (kWh)	0	0.14	0.59	0.280	0.989	0.368
Motor efficiency	Regenerative	0	0.859	0.850	0.881	0.636	0.829
	Accelerating	0.870	0.859	0.860	0.859	0.860	0.859
Fuel cell stack	Efficiency	0.542	0.555	0.561	0.551	0.574	0.555
DCDC converter efficiency	Charge efficiency	—	—	0.959	0.950	0.967	0.954
	Discharge efficiency	—	—	0.960	0.932	0.963	0.939
	Round-trip efficiency	—	—	0.920	0.886	0.930	0.895
Energy storage efficiency	Charge efficiency	—	0.984	0.977	0.983	0.960	0.968
	Discharge efficiency	—	0.983	0.975	0.991	0.913	0.963
	Round-trip efficiency	—	0.967	0.952	0.974	0.877	0.932

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 6. Comparison of fuel economy and improvement factor of different fuel cell vehicles.

Vehicle Topology	Drive Cycle	Fuel Economy/Improvement Factor	
		Power Assist	Load Leveling
FC-battery hybrid with 1500 Wh battery and power electronics	FUDS	78.6/1.16	72.8/1.07
	US06	56.6/1.12	51.9/1.02
FC-UC hybrid with 100 Wh UC and power electronics	FUDS	79.2/1.16	78.8/1.16
	US06	57.3/1.13	55.0/1.08
FC-UC hybrid with 100 Wh UC and without power electronics	FUDS	85.0/1.25	
	US06	59.6/1.18	
FCV without energy storages	FUDS	68.0/—	
	US06	50.7/—	

The effect of the control strategy on fuel economy improvement is of particular interest. In all cases, the improvement is calculated relative to the driveline without energy storage, in which case the fuel cell is load following and there is no energy recovery during braking of the vehicle. In general, the power-assisted strategy results in larger improvements in fuel economy than the load leveling strategy. This is due to the higher losses in the energy storage and the DC/DC in the case of the load leveling strategy that requires a greater fraction of the energy to pass through energy storage. However, the load leveling strategy mitigates to a larger extent current transients in the fuel cell stack. As shown in Table 6, the improvements in fuel economy using energy storage are modest in magnitude being in most cases less than 15%. The fuel economy improvements are slightly less for the US06 cycle than for the FUDS. The largest improvements projected are using ultracapacitors directly connected to the fuel cell bus without DC/DC electronics. In this case, the improvements are 18–25%. The simulation results indicate that the best approaches are to use ultracapacitors without DC/DC electronics or batteries with power-assisted control strategy. An ultracapacitor unit storing 100–120 Wh of usable energy or a lithium battery storing 1.5 kWh seems to be good energy storage solutions for FCVs. The influences of energy storage unit capacity on the fuel economy improvement for ultracapacitors and batteries are shown in Figures 11 and 12.

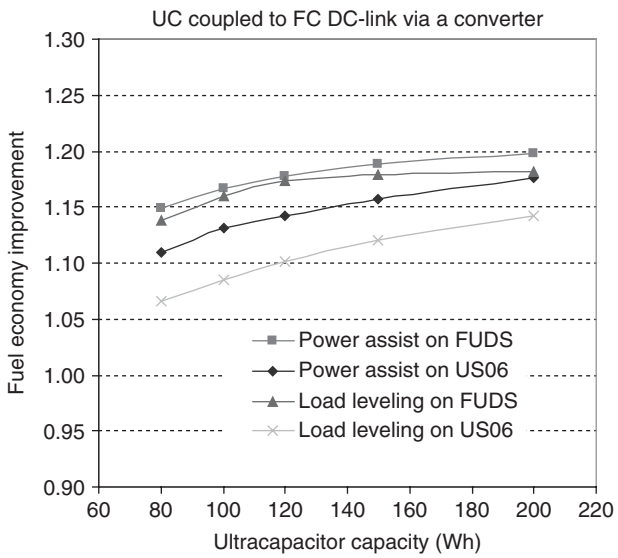


Figure 11. Fuel economy improvements of fuel cell vehicles utilizing ultracapacitors for different control strategies.

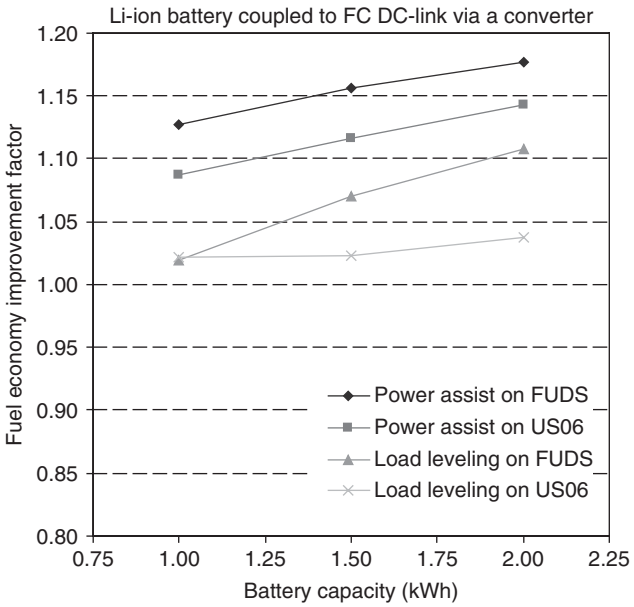


Figure 12. Fuel economy improvements of fuel cell vehicles utilizing lithium batteries for different control strategies.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

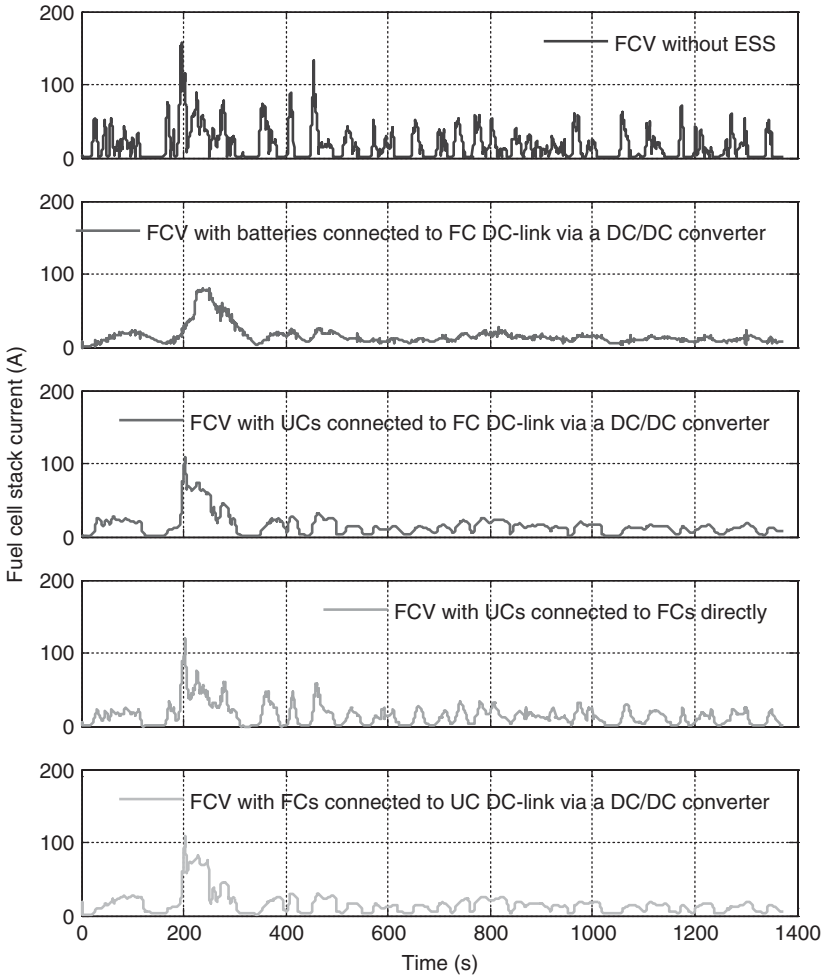


Figure 13. Comparison of the current transients for various driveline arrangements on the FUDS.

Comparisons of the simulation results for the fuel stack and energy storage unit currents using the different drive train configurations are shown in Figures 13 and 14. The results indicate that the use of energy storage in all cases significantly reduce the maximum current and the current transient dynamics and thus mitigates the stress on the fuel cell stack. However, the reduction of the current transients is significantly greater using the load leveling control strategy (Figure 14). Connecting the ultracapacitors directly across the fuel cell results in only slightly higher transients than with the load leveling strategy (Figure 13).

4.3 Comparison of the fuel economy of fuel cell vehicles with other advanced technology vehicles

As one of the future vehicle technologies, it is of interest to compare the fuel economy of FCVs with that of other advanced technology vehicles being developed to reduce fuel use and greenhouse gas emissions from passenger cars in years ahead. Computer simulations of the operation of mid-size passenger cars were performed for fuel cell-battery hybrid vehicles, hybrid electric vehicles (HEVs), and advanced conventional internal combustion engine (ICE) vehicles in future years—2015, 2030, and 2045—to project how much each technology would reduce energy consumption (Burke and Zhao, 2010; Ogden and Anderson, 2011).

The FCVs (FCHEV) used for comparison with the other vehicle technologies are fuel cell-battery hybrids with the lithium-ion battery connected to the fuel cell bus by a DC/DC converter. The converter controls the output power of the battery such that the output power of the fuel cell is load leveled. The engines used in the simulations of HEVs and conventional ICEs are spark-ignition engines. The engine efficiencies are increased in the simulations for

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

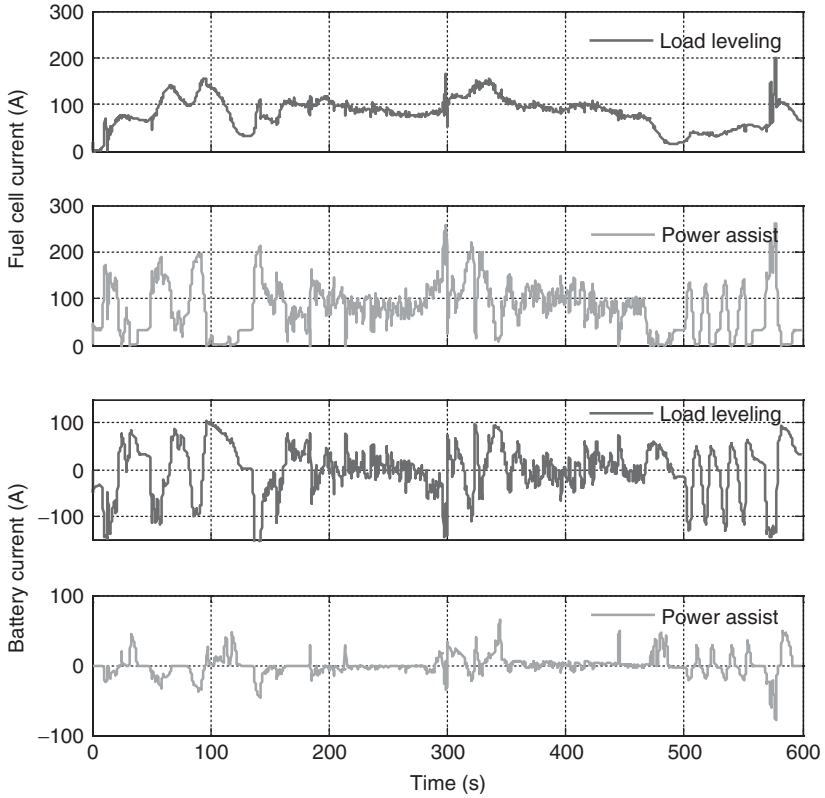


Figure 14. Comparison of the current transients for various driveline arrangements on the US06 cycle.

Table 7. Battery characteristics.

Year	Battery Type	Ampere-Hour	Watt-Hour per Kilogram	Resistance (mΩ)
2015	Lithium titanate oxide	4	35	1.1
2030/2045	Lithium titanate oxide	4	42	0.9

Table 8. Characteristics of the mid-size fuel cell passenger cars.

	2015	2030	2045
Vehicle configuration	2015	2030	2045
C_D	0.25	0.22	0.2
A_F (m^2)	2.2	2.2	2.2
Fr	0.007	0.006	0.006
FC (kW)	83.2	76.6	72.1
Fuel cell efficiency (%)	60	62	65
Motor (kW)	103	100	99
Battery (kWh)	0.93	0.85	0.85
Vehicle test weight (kg)	1516	1383	1366
Electronic account load (W)	220	240	260

future years based on expected significant improvements in engine technology (Kasseris and Heywood, 2009). The batteries used in the simulations of HEVs and FCHEVs are scaled from lithium titanate oxide batteries (Burke and Miller, 2009b) (Table 7). The vehicle characteristics—curb weight, drag coefficient, frontal area, and tire rolling resistance—are the same as assumed by DOE (Plotkin and Singh, 2005). A summary of the vehicle and powertrain characteristics used in the simulations is given in Table 8.

Simulations of the fuel cell-battery hybrid vehicles were performed for the FUDS, HWY, and US06 driving cycles for the years of 2015, 2030, and 2045, respectively. The simulation results for the fuel economy are given in Table 9. Also shown in the table are the FCV fuel economy projections of DOE (Plotkin and Singh, 2005).

The projected fuel economy and fuel savings for the different driving cycles are plotted in Figures 15 and 16. The simulation results indicate that large improvements in the fuel economy of mid-size fuel cell-hybrid passenger cars can be expected in 2015–2030. Further improvements are projected for 2045. Compared to a mid-size 2007 baseline passenger car, these improvements are 60% (2015) to 70% (2045) for fuel savings. The simulation results published

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 9. Fuel cell vehicle fuel economy 2015–2045.

	2015		2030		2045	
	UCD	DOE	UCD	DOE	UCD	DOE
FUDS	82.6	70	102.8	102	108.9	114
HWY	90.8	79	111.5	114	119.5	130
US06	61.3	—	76.2	—	82.3	—

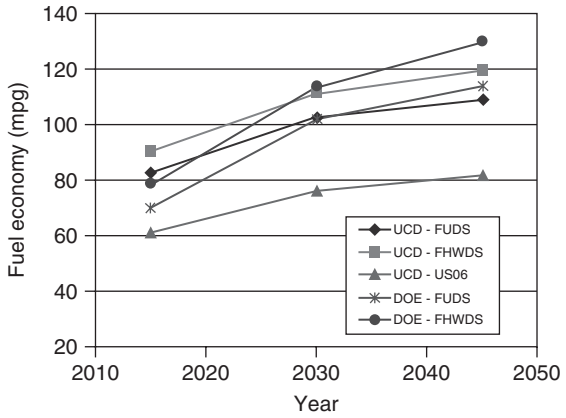


Figure 15. Fuel economy simulation results for future fuel cell hybrid vehicles.

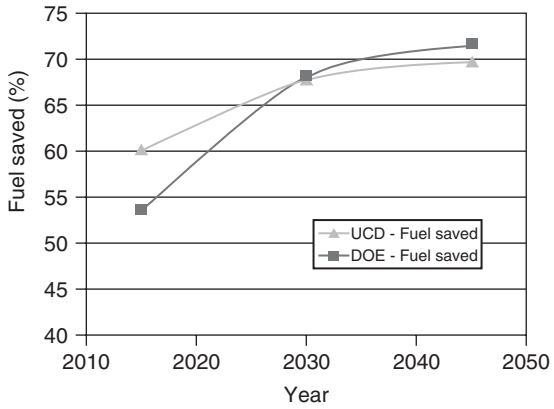


Figure 16. Fuel savings[a] for future fuel cell hybrids relative to 2007 passenger cars. ([a]% Fuel saved $= (1 - \text{mpg}_0/\text{mpg}) \times 100$, $\text{mpg}_0 = 34.5$, which is the average of the urban-highway dynamometer fuel economy of the 2007 baseline vehicle.)

by the DOE (Plotkin and Singh, 2005) are also plotted in Figures 15 and 16. The UCD and DOE projections for fuel cell-battery hybrids are in good agreement over the complete time period of the simulations with the agreement being closest in the 2030–2045 time periods.

Studies directed toward projecting the performance of advanced ICE vehicles and HEVs were also performed in

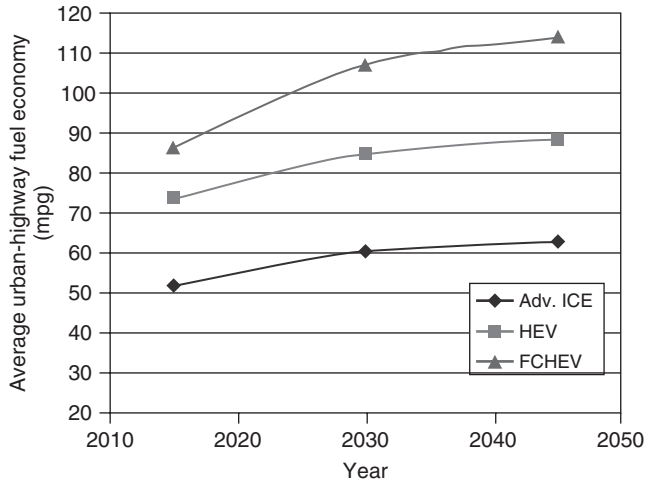


Figure 17. Comparisons of the fuel economies of future advanced vehicle technologies.

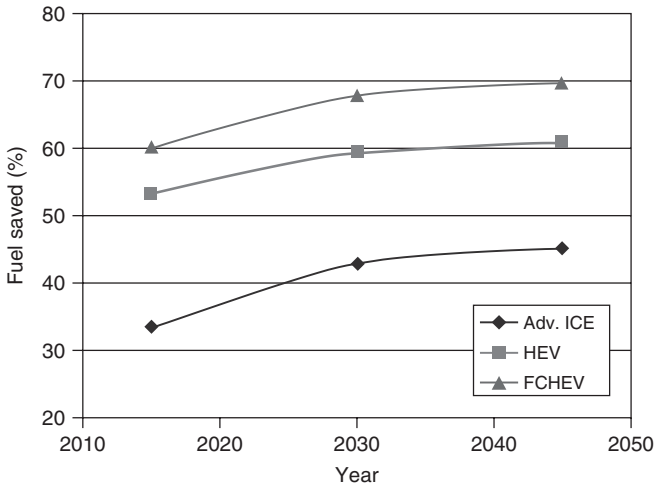


Figure 18. Comparison of fuel savings of future vehicle technologies.

(Burke and Zhao, 2010; Ogden and Anderson, 2011). The projected fuel economies and fuel savings for advanced ICEs, HEVs, and hybrid FCVs are plotted in Figures 17 and 18, respectively. The simulations show that FCVs achieve about twice the fuel economy of the improved conventional engine/transmission ICE vehicles, but only about 33% better fuel economy compared to the HEV vehicles. However, the resultant difference in the fuel savings of the FCVs compared to the hybrids is only 16%. This savings does not include consideration of the differences in the efficiencies of producing gasoline from petroleum and hydrogen from natural gas or coal.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

REFERENCES

An, S.J., Lee, K.I., and Kim, T.J. (2008) Performance analysis according to the combination of energy storage system for fuel cell hybrid vehicle. *International Journal of Automotive Technology*, **9** (1), 111–118.

Badrinarayanan, P., Ramaswamy, S., Eggert, A., and Moore, R., Fuel cell stack water and thermal management: impact of variable system power operation, ITS-UC Davis report UCD-ITS-RP-01-37 and SAE paper 2001-01-0537, March 2001.

Bauman, J. and Kazerani, M. (2008) A comparative study of fuel-cell–battery, fuel-cell–ultracapacitor, and fuel-cell–battery–ultracapacitor vehicles. *IEEE Transactions on Vehicular Technology*, **57** (2), 760–769.

Broom, D.P. (2011) *Hydrogen Storage Materials—The Characterization of Their Storage Properties*, Springer-Verlag London Limited, London. ISBN: 978-0-85729-220-9.

Burke, A.F. and Gardiner, M., Hydrogen Storage Options: Technologies and Comparisons for Light-duty Vehicle Applications, 2005, UCD-ITS-RR-05-01.

Burke, A.F. and Miller, M., Electrochemical Capacitors as Energy Storage in Hybrid-Electric Vehicles: Present Status and Future Prospects, EVS-24, Stavanger, Norway, May 2009a (paper on the CD of the meeting).

Burke, A.F. and Miller, M., Performance Characteristics of Lithium-ion Batteries of Various Chemistries for Plug-in Hybrid Vehicles, EVS-24, Stavanger, Norway, May 2009b (paper on the CD of the meeting).

Burke, A.F. and Zhao, H., Projected fuel consumption characteristics of hybrid and fuel cell vehicles for 2015-2045, EVS25, Shenzhen, China, Nov. 5–9, 2010.

Chen, D. and Peng, H. (2005) The thermodynamic model of membrane humidifiers for PEM fuel cell humidification control. *Journal of Dynamic Systems, Measurement, and Control*, **127**, 424–432. September

Cunningham, J.M., Air System Management for Fuel Cell Vehicle Applications, ITS UCDavis report UCD-ITS-RR-01-16, January 2001.

Dillich, S., Hydrogen Storage, 2009 DOE Hydrogen Program and Vehicle Technologies Program, Merit Review and Peer Evaluation Meeting, May 19, 2009

DOE report, Targets for Onboard Hydrogen Storage System for Light-Duty Vehicles, September 2009.

Emadi, A., Rajashekara, K., Williamson, S., and Lukic, S. (2005) Topological overview of hybrid electric and fuel cell vehicular power system architectures and configurations. *IEEE Transactions on Vehicular Technology*, **54** (3May), 763–770.

Gao, W. (2005) Performance comparison of a fuel cell-battery hybrid powertrain and a fuel cell-ultracapacitor hybrid powertrain. *IEEE Transactions on Vehicular Technology*, **54** (3) NO., 846–855.

Garcia-Arregui, M., Turpin, C., and Astier, S., Direct connection between a fuel cell and ultracapacitors, Clean Electrical Power Conference, Capri, May 2007, 474–479

Gardiner, M.R., Investigation of Cryogenic Hydrogen Storage with High Surface Area Carbon; Theory, Experiment, and System Design, Ph.D. thesis in Transportation Technology and Policy, University of California-Davis, September 2004.

Gardiner, M. and Cunningham, J., Compressed hydrogen storage for fuel cell vehicles. SAE paper 2001-01-2531, August 2001.

Jeon, S., Hyundai Supercapacitor Fuel Cell Hybrid Electric Vehicle, Proceeding of the 21st Annual Electric Vehicle Symposium, Monaco, April 2005.

Jeong, K.S. and Oh, B.S. (2002) Fuel economy and life-cycle cost analysis of a fuel cell hybrid vehicle. *Journal of Power Sources*, **105**, 58–65.

Kasseris, E.P. and Heywood, J.B. (2009) Comparative analysis of automotive powertrain choices for the next 25 years in *Road Vehicle Technology* (eds A. Schafer, J. Heywood, D. Jacoby, I.A. Waitz) Transportation in a Climate-Constrained World, Chapter 4, MIT Press, Cambridge, MA. ISBN: 978-0-26251-234-3.

Kulp, G.W. (2001) *A Comparison of Two Air Compressors for PEM Fuel Cell Systems*, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Lai, J.-S. and Nelson, D.J. (2007) Energy management power converters in hybrid electric and fuel cell vehicles. *Proceedings of the IEEE*, **95** (4April), 766–777.

Li, C.Y. and Liu, G.P. (2009) Optimal fuzzy power control and management of fuel cell/battery. hybrid vehicles. *Journal of Power Sources*, **192**, 525–533.

Lin, C.C., Kim, M.J., Peng, H., and Grizzle, J.W. (2006) System-level model and stochastic optimal control for a PEM fuel cell hybrid vehicle. *Transactions of the ASME*, **128**, 878–890.

Litster, S. and McLean, G. (2004) PEM fuel cell electrodes. *Journal of Power Sources*, **130**, 61–76.

Mench, M.M. (2008) *Fuel Cell Engines*, John Wiley & Sons, Hoboken, New Jerseya. ISBN: 978-0-47168-958-4.

Mori, D., and Haraikawa, N., High pressure metal hydride tank for fuel cell vehicles, Toyota presentation, IPHE International Hydrogen Storage Technology conference, June 19–22, 2005, Lucca, Italy.

Ogden, J. and Anderson, L. (2011) *Sustainable Transportation Energy Pathways-A Research Summary for Decision Makers (Chapter 4)*, University of California, Davis, Davis California, October.

Ogden, J.M. and Yang, C., Implementing a Hydrogen Energy Infrastructure: Storage Options and System Design, ITS Davis report UCD-ITS-RR-05-28, November 2005.

Ohkawa, A., Electric Power Control System for a Fuel Cell Vehicle Employing Electric Double-Layer Capacitor. SAE World Congress, Detroit, 2004, SAE 2004-01-1006.

Pesaran, A.A., Kim, G.H., and Gonder, J.D., PEM Fuel Cell Freeze and Rapid Startup Investigation, NREL Report, NREL/MP-540-38760, September 2005.

Plotkin, S. and Singh, M., Multi-Path Transportation Futures Study: Vehicle Characterization and Scenarios, Argonne Lab and DOE Report, 2005, ANL/ESD/09-5.

Schaltz, E., Khaligh, A., and Rasmussen, P.O., Investigation of Battery/Ultracapacitor Energy Storage Rating for a Fuel Cell Hybrid Electric Vehicle, IEEE Vehicle Power and Propulsion Conference (VPPC), Harbin, China, September 3–5, 2008

Schoenung, S. M., IEA Hydrogen Annex—Transportation Applications Analysis, Proceedings of the 2001 DOE Hydrogen Program Review, 2001, NREL/CP-570-30535.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Sevilla, M., Foulston, R., and Mokaya, R. (2010) Superactivated carbide-derived carbons with high hydrogen storage capacity. *Energy Environment Science*, **3**, 223–227.

Spiegel, C. (2008) *PEM Fuel Cell-Modeling and Simulation Using MATLAB*, Elsevier Inc. Burlington, MA San Diego, California London. ISBN: 978-0-12-374259-9.

Srinivasan, S. (2006) *Fuel Cells-From Fundamentals to Applications*, Springer Science+Business Media, LLC New York, NY.

Sundström, O. and Stefanopoulou, A. (2007) Optimum battery size for fuel cell hybrid electric vehicle—Part I. *Journal of Fuel Cell Science and Technology*, **4**, 167–175.

Thounthonga, P., Raël, S., and Davat, B. (2009) Energy management of fuel cell/battery/supercapacitor hybrid power source for vehicle applications. *Journal of Power Sources*, **193** (1), 376–385.

Uzunoglu, M. and Alam, M.S. (2007) Dynamic modeling, design and simulation of a PEM fuel cell/ultra-capacitor hybrid system for vehicular applications. *Energy Conversion and Management*, **48**, 1544–1553.

Weinert, J.X. and Lipman, T.E., An Assessment of the Near-term Costs of Hydrogen Refueling Stations and Station Components, ITS Davis Report UCD-ITS-RR-06-03, January 2006.

Xu, H., Kunz, H.R., and Fenton, J.M. (2006) Analysis of proton exchange membrane fuel cell polarization losses at elevated temperatures 120 deg C and reduced relative humidity. *Electrochimica Acta*, **52** (2007), 3525–3533. November

Yang, C. and Ogden, J.M., Build-up of a Hydrogen Infrastructure in the US, ITS-UC Davis report UCD-ITS-RP-10-05, February 2010.

Young, R., Chao, B., Fournier, G., and Pawlik, D.A., *et al.*, A hydrogen ICE vehicle powered by Ovonic metal hydride storage. SAE paper 04–606, 2004.

Zhao, H., and Burke, A.F., Optimum Performance of Direct Hydrogen Hybrid Fuel Cell Vehicles, EVS24, Stavanger, Norway, paper number: EVS-24-2070091, 13–16 May 2009a.

Zhao, H. and Burke, A.F. (2009b) Optimization of fuel cell system operating conditions for fuel cell vehicles. *Journal of Power Sources*, **186** (2), 408–416.

Zhao, H. and Burke, A.F. (2010a) Fuel cell powered vehicles using supercapacitors - device characteristics, control strategies, and simulation results. *Fuel Cells*, **10** (5), 879–896.

Zhao, H., and Burke, A.F., Effects of different powertrain configurations and control strategies on fuel economy of fuel cell vehicles, EVS25, Shenzhen, China, Nov. 5–9, 2010b.

Zhao, H., Burke, A.F., and Miller, M., Comparison of Hybrid Fuel Cell Vehicle Technology and Fuel Efficiency, International Conference of Hydrogen and Fuel Cell 2011, Vancouver, Canada, 15–18 May 2011.

Zolot, M., Markel, T., and Pesaran, A., Analysis of Fuel Cell Hybridization and Implications for Energy Storage Devices, The 4th International Advanced Automotive Battery conference, San Francisco, California, June 2–4, 2004.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto066
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

All-Wheel Drive Hybrid System

Eiji Sato

Toyota Motor Corporation, Toyota, Japan

1 Introduction 1
2 Background 1
3 System Outline 2
4 Control 6
References 9

1 INTRODUCTION

All-wheel drive (AWD) systems may be implements in hybrid vehicles with a number of alternative configurations that may extend to individual wheel control. Some aspects of these are considered in EV Powertrain Configurations and Basic Consideration.

One type of AWD hybrid electric vehicle (HEV) takes a combination approach in which a rear-wheel motor system is added to a front-wheel drive (FWD) hybrid system. As this combined hybrid system does not require a propeller shaft to mechanically connect the front and rear wheels, it is relatively easy to secure the space to install the rear-wheel motor system, and its installation results in only a small increase in weight. The rapid response of the rear-wheel motor in this system also allows refined front- and rear-wheel driving force control, in accordance with the driving conditions, and four-wheel regeneration. As a result, this improves fuel efficiency while enhancing dynamic performance on low friction road surfaces. In addition, installing drive motors for the front and rear wheels increases the overall driving force and allows the load to be evenly divided between the front and the rear wheels. Consequently, this system is particularly suitable for heavy vehicles such as minivans and sport utility vehicles (SUVs).

Using the current Lexus RX400h as a specific example, this chapter describes the configuration of the hybrid system used in an AWD HEV, its component technology, and its control principles.

2 BACKGROUND

Toyota Motor Corporation launched the Estima Hybrid as an AWD HEV that combines a rear-wheel motor system with an FWD hybrid system in Japan in 2001 (Figure 1). This vehicle is an AWD hybrid minivan without a propeller shaft. The front wheels are powered by a parallel hybrid system that combines a mechanical continuously variable transmission (CVT) with a single motor, and the rear wheels are powered by a motor drive system that is mechanically independent from the front wheels (Sasaki, 2005).

In 2004, Toyota launched the Lexus RX400h hybrid SUV (Figure 2) and Highlander Hybrid. These vehicles were equipped with a similar but higher power rear-wheel system than in the Estima Hybrid but combined with the Toyota Hybrid System II (THS II) at the front. The THS II is a power split two-motor type hybrid system that uses a planetary gear set. It should also be noted that the RX400h was also produced with an FWD version without the rear-wheel motor (Sadakata and Ikuta, 2005).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto067
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5



Figure 1. Estima hybrid. (Reproduced with permission from Sasaki (2005). © Toyota Motor Corporation.)



Figure 2. Lexus RX400h. (Reproduced by permission of Toyota Motor Corporation.)

3 SYSTEM OUTLINE

3.1 System configuration

Figure 3 shows the configuration of the AWD hybrid system. The front-wheel system is the THS II used in the Prius and other vehicles. Engine power is divided into two by the power split device, with a portion used to directly drive the front wheels and the remainder converted into electrical power by the generator. This power is then used to operate the motor in combination with power from the battery. In addition, the vehicle can be driven using the battery and motor even when the engine is stopped. A motor speed reduction device and a motor for driving the rear wheels were added to optimize this system for use in an SUV.

Table 1 lists the main specifications of each component in the AWD hybrid system and shows a comparison with the specifications of the components in the Prius.

3.2 Front-wheel unit

Figure 4 shows a cross-sectional view of the front-wheel hybrid system transmission for an SUV, and Table 2 shows its main specifications. This hybrid transmission includes a motor speed reduction device and a compound gear. The motor speed reduction device multiplies the output torque without increasing the length of the transmission.

The compound gear, which integrates two planetary ring gears, a counter (or transfer) gear and a parking gear, is laid out to the outside of the motor speed reduction gear and power split device. The use of this compound gear eliminates the chain in the previous hybrid system and

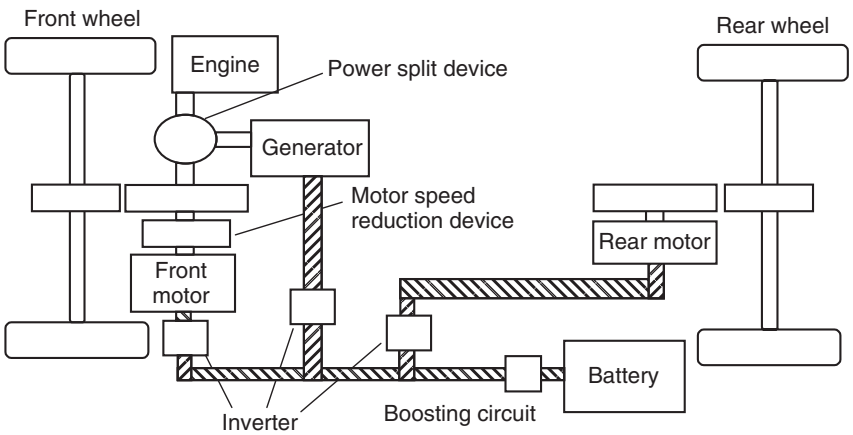


Figure 3. AWD HEV system configuration. (Reproduced by permission of Toyota Motor Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto067
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Specifications of THS II (comparison between SUV version and Prius).

Item		THS II (SUV)	THS II (Prius)
Engine	Type	3.3 L gasoline	1.5 L gasoline
	Maximum output (kW/rpm)	155/5600	57/5000
	Maximum torque (Nm/rpm)	288/4400	115/4200
Front drive motor	Type	Permanent magnet motor	Permanent magnet motor
	Maximum output (kW)	123	50
	Maximum torque (Nm)	333	400
	Maximum speed (rpm)	12,400	6400
Rear drive motor	Type	Permanent magnet motor	
	Maximum output (kW)	50	
	Maximum torque (Nm)	130	
	Maximum speed (rpm)	10,500	
Battery	Type	Nickel-metal hydride	Nickel-metal hydride
	Maximum output (kW)	45	25
	10 s rating output (kW)	36	21

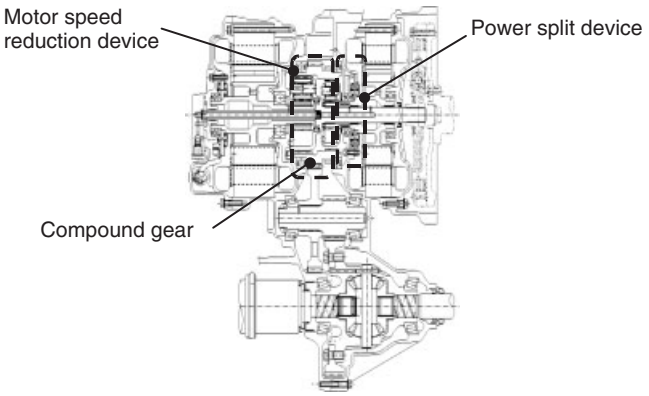


Figure 4. Cross-sectional view of front-wheel hybrid system transmission for SUV. (Reproduced with permission from Hata, Kamiya, and Nagamatsu (2005). © Toyota Motor Corporation.)

Table 2. Specifications of SUV hybrid transmission.

		For SUV	For 1.5 L HEV
Max. engine torque		288 Nm	115 Nm
Max. engine output		155 kW	57 kW
Motor	Type	Synchronous AC motor	←
	Max. output	123 kW	50 kW
	Max. torque	333 Nm	400 Nm
	Max. speed	12,400	6400 rpm
Motor reduction gear ratio		2.478	—
Differential gear ratio		3.542	4.113
Weight (including ATF)		125 kg	109 kg
Overall length		417 mm	430 mm

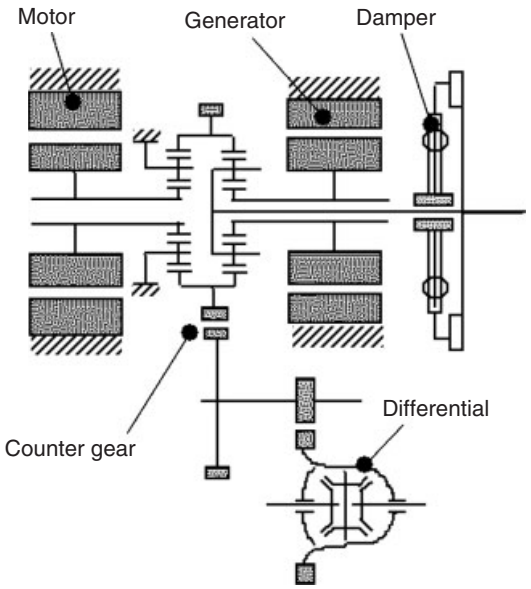


Figure 5. Schematic view of front-wheel hybrid system transmission for SUV. (Reproduced with permission from Hata, Kojima, and Wata (2006). © Society of Automotive Engineers Japan.)

makes the system more compact by reducing the number of axes from four to three.

Figure 5 shows a schematic view of the front-wheel system. From the right side (engine side) of the figure, the first axis includes a damper with a torque limiter, a generator, two planetary gears, and a motor. The planetary gear on the right acts as the power split device that divides the force from the engine into driving force for the vehicle and force to drive the generator. The planetary gear on the left acts as the motor speed reduction device. The compound gear, which integrates the two ring gears, counter gear, and parking gear, is laid out to the outside of these two planetary gears. Driving force is transmitted to the counter-driven gear and drive pinion gear on the second axis and the ring gear and differential on the third axis.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto067
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.3 Rear-wheel unit

In the rear transaxle, the distance between the input shaft from the motor and the output shaft to the side gear was shortened significantly. This results in a lightweight and compact three-shaft counter gear single-speed transaxle that can be integrated with the motor (Figure 6).

The rear-wheel unit in the Lexus RX400h uses an AC synchronous brushless motor based on that used in the Estima Hybrid. However, the operating voltage was increased to 650 V, which enabled a substantial increase in power by adjusting the number of coil turns in accordance with the voltage (Figure 7). As a result, coordinated control between the front-wheel and rear-wheel drive motors in accordance with the driving conditions creates rear-wheel driving force suitable for an SUV when the vehicle starts to move, while accelerating, and when climbing hills.



Figure 6. Rear-wheel unit. (Reproduced with permission from Sadakata and Ikuta (2005). © Toyota Motor Corporation.)

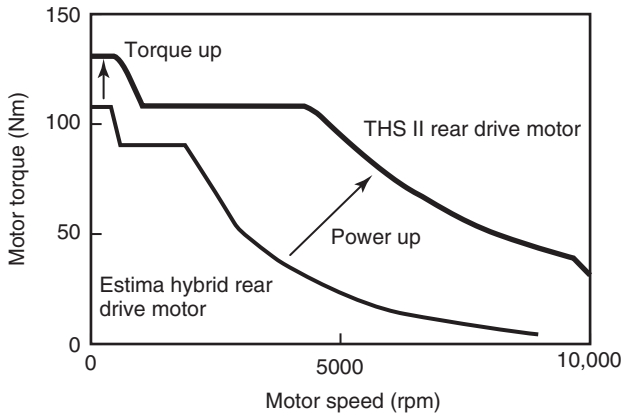


Figure 7. Rear drive motor torque profile. (Reproduced with permission from Kimura (2005). © Toyota Motor Corporation.)

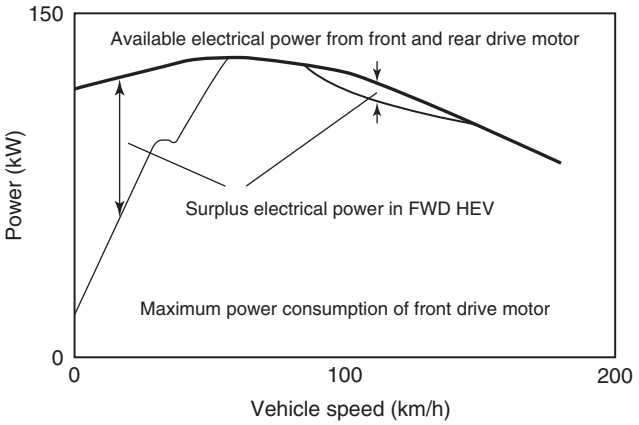


Figure 8. Surplus power in FWD HEV. (Reproduced with permission from Kimura (2005). © Toyota Motor Corporation.)

In an FWD HEV, surplus battery and generator power may occur in low speed regions. In the AWD HEV system, this surplus power can be converted to driving force by the rear-wheel motor (Figure 8). Although an FWD vehicle conventionally has superior dynamic performance than an AWD vehicle because of lower vehicle weight and other reasons, the conversion of surplus power by the rear-wheel motor gives the AWD HEV better dynamic performance than an FWD HEV, despite the weight increase caused by installing the rear-wheel unit.

3.4 Power control unit (PCU)

The THS II in the Prius contains a boost circuit that is capable of variable control up to a maximum of 500 V (Kikuchi and Matsubara, 2005). The boost circuit in the SUV hybrid system power control unit (PCU) is capable of variable control up to 650 V in accordance with the driving conditions. The increased motor size and the boosting voltage enabled a SUV hybrid system with a maximum motor output of 123 kW, which is approximately 2.4 times greater than the Prius THS II (maximum 50 kW). In addition, the PCU was designed to be compact enough for installation at the front of the vehicle, even while integrating the traction inverter for the rear-wheel motor, which has a maximum power of 50 kW (Figure 9).

Figure 10 shows the configuration of the primary circuit of the PCU. The inverters for the generator, front-wheel motor, and rear-wheel motor are voltage type inverters. These inverters drive the generator, front-wheel motor, and rear-wheel motor by the operation of switching devices using voltage boosted by the boost converter. The output from the generator and motor inverters is controlled by the three-phase voltage type pulse width modulation (PWM)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto067
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

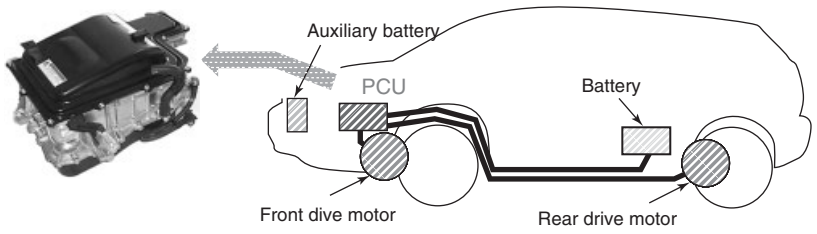


Figure 9. Layout and external view of PCU. (Reproduced by permission of Toyota Motor Corporation.)

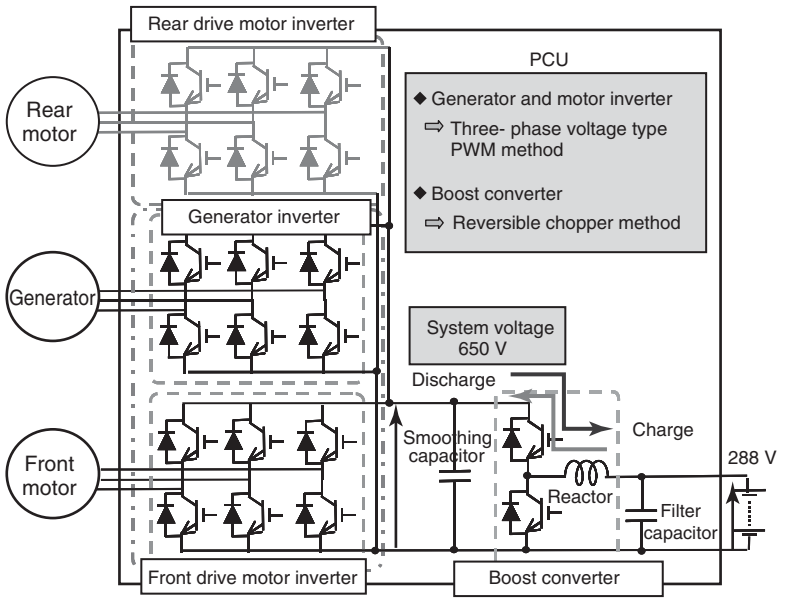


Figure 10. Configuration of primary circuit of PCU. (Reproduced with permission from Kikuchi and Matsubara (2005). © Toyota Motor Corporation.)

method. Voltage is applied to the generator and motor as sine wave or rectangular wave output depending on the rotational speed.

The PCU is water cooled. The generator and motor inverters are integrated at the top surface of the cooling circuit. The configuration can be made compatible with an FWD HEV by removing the rear-wheel motor inverter (Figure 11).

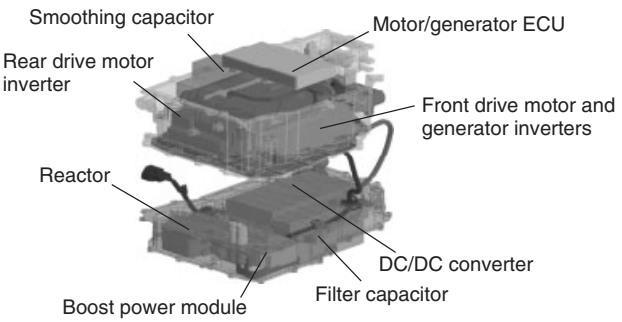


Figure 11. Configuration of PCU. (Reproduced with permission from Kikuchi and Matsubara (2005). © Toyota Motor Corporation.)

3.5 Battery

The AWD HEV system uses the same nickel-metal hybrid batteries as in the Prius. The number of cells was increased to 240 to adjust to the weight of an SUV. Instantaneous power is 45 kW with a 10 s rated value of 36 kW. The cells were also reduced in size to allow installation in a three-row SUV. Figure 12 shows how the batteries are installed under the second row of seats without affecting the internal space of the SUV. The main relay, the resistor to prevent inrush current, and the current sensor were integrated inside the battery pack, enabling a compact design.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto067
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

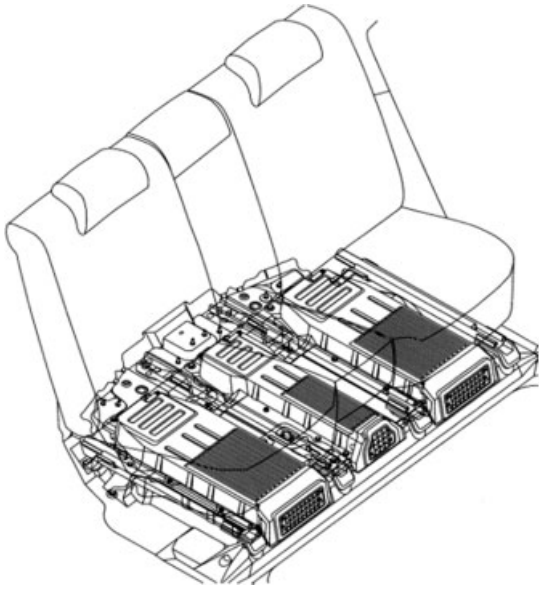


Figure 12. Battery pack. (Reproduced with permission from Kimura (2005). © Toyota Motor Corporation.)

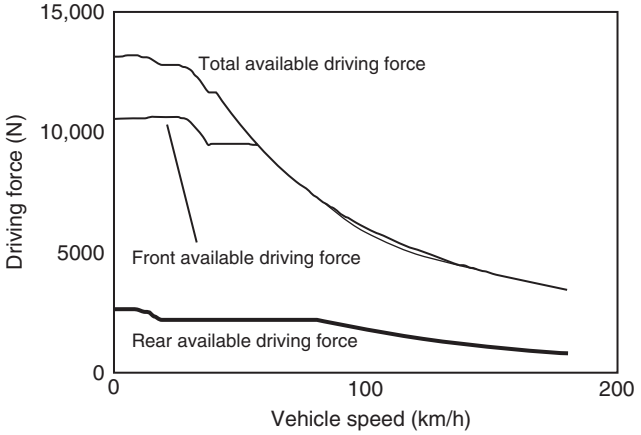


Figure 13. Capability of rear drive unit. (Reproduced with permission from Kimura (2005). © Toyota Motor Corporation.)

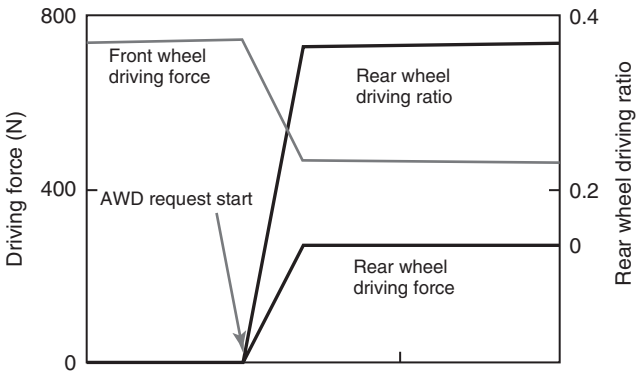


Figure 14. Rear driving force change speed. (Reproduced with permission from Kimura (2005). © Toyota Motor Corporation.)

4 CONTROL

4.1 Driving force distribution to front and rear wheels

The front-wheel unit is designed to generate sufficient performance for an FWD HEV (Kimura, 2005). In contrast, the power of the rear-wheel unit is restricted for packaging reasons. However, as the rear-wheel motor has power of 50 kW and 130 Nm, a 50 : 50 front-rear torque distribution is possible at an acceleration level of approximately 50% (Figure 13).

In this electrical AWD system, the front-wheel and rear-wheel units are installed with independent motors. Changes in the front-/rear-wheel driving force distribution ratio are implemented by changing the torque of the front- and rear-wheel motors. The distribution ratio can be changed rapidly as the driving force of the front and rear wheels is adjusted directly without intervention from hydraulics or other controls used in conventional AWD systems (Figure 14). This improves both fuel efficiency and vehicle stability.

4.2 Outline of AWD control

In consideration of driving force transfer efficiency, the AWD HEV system mostly operates the vehicle as an FWD HEV. AWD operation is only applied when necessary. As the rear-wheel unit is powered by electricity alone, constant output of driving force would require power generation by the front-wheel unit.

The reason for adopting FWD as the basic mode of operation is as follows: compared with FWD operation, AWD operation in this system reduces the torque of the front-wheel motor. This amount of torque is then added to the torque from the rear-wheel motor to enable AWD operation. This is preferable from the standpoints of overall driving force and the balance of the battery. If full-time AWD operation was adopted instead, the torque of the front-wheel motor would need to be increased to the negative side to drive the rear-wheel motor. The resulting power exchange between the front and rear-wheel motors would increase the losses of the electrical system and affect efficiency.

As described earlier, as the operation mode can be switched rapidly between FWD and AWD, vehicle stability can be guaranteed even when AWD is only engaged when necessary. AWD for vehicle stability is required when

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto067
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the vehicle starts to move, when the wheels begin to slip, or when turning left or right. AWD operation is also adopted for regeneration of kinetic energy during deceleration.

4.3 System operation

4.3.1 When vehicle starts to move

The engine is stopped when the vehicle is not moving. When the vehicle starts to move, AWD operation is selected and the front- and rear-wheel motors are driven by power from the battery (Figure 15).

4.3.2 Normal driving

During normal driving, the driving force from the engine is divided into two paths by the power split device (Figure 16). One path is used to directly drive the wheels. The other is used to drive the generator to generate power, which is then used to drive the motors.

4.3.3 Wide open throttle

When the driver accelerates the vehicle at wide open throttle, battery power is supplied to supplement the power from the engine (Figure 17). Maximum vehicle power is achieved by driving both the front- and rear-wheel motors.

4.3.4 Light load

When driving in conditions that may force the engine to operate in a low efficiency area, such as driving at low vehicle speeds or down a gradual hill, the engine is stopped and the vehicle is driven by the front-wheel motor (Figure 18). However, the engine is not stopped in

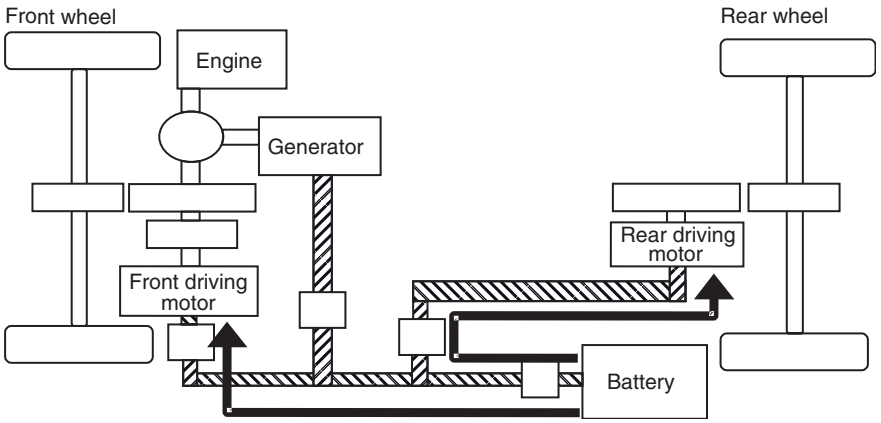


Figure 15. System operation (when vehicle starts to move). (Reproduced by permission of Toyota Motor Corporation.)

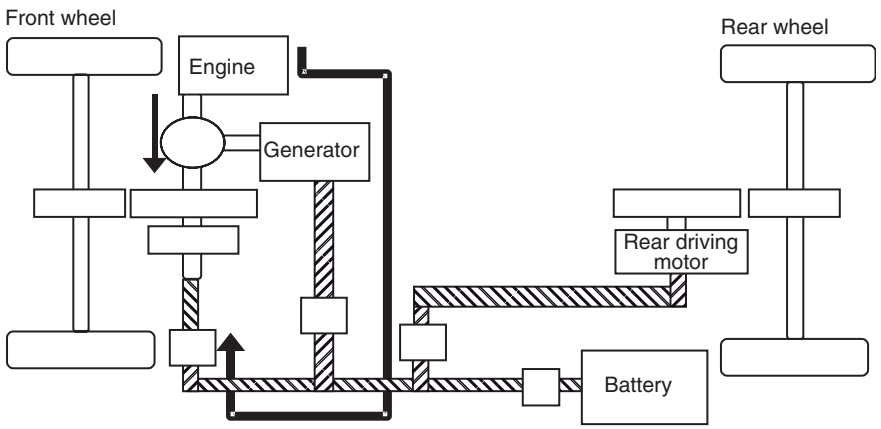


Figure 16. System operation (normal driving). (Reproduced by permission of Toyota Motor Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto067
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

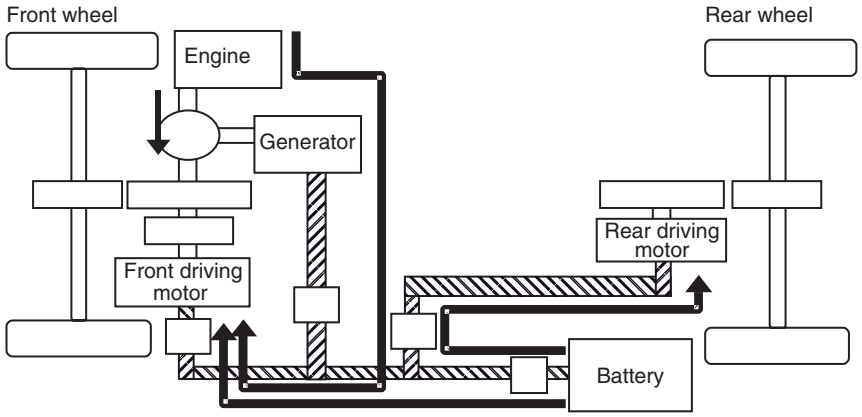


Figure 17. System operation (wide open throttle). (Reproduced by permission of Toyota Motor Corporation.)

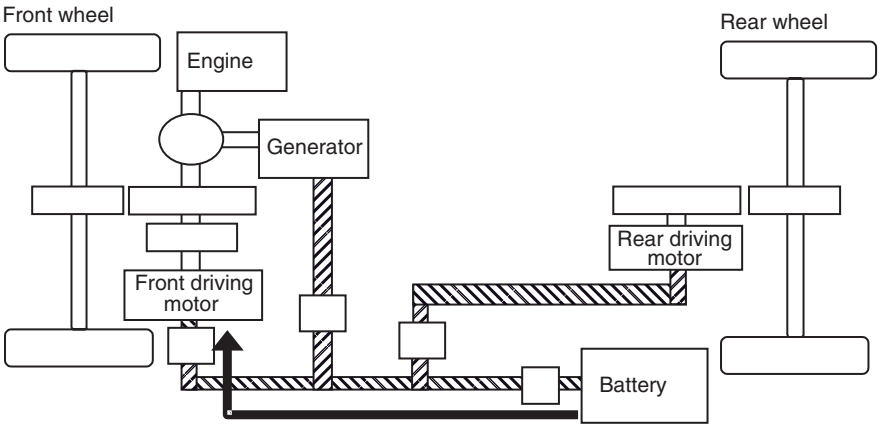


Figure 18. System operation (light load). (Reproduced by permission of Toyota Motor Corporation.)

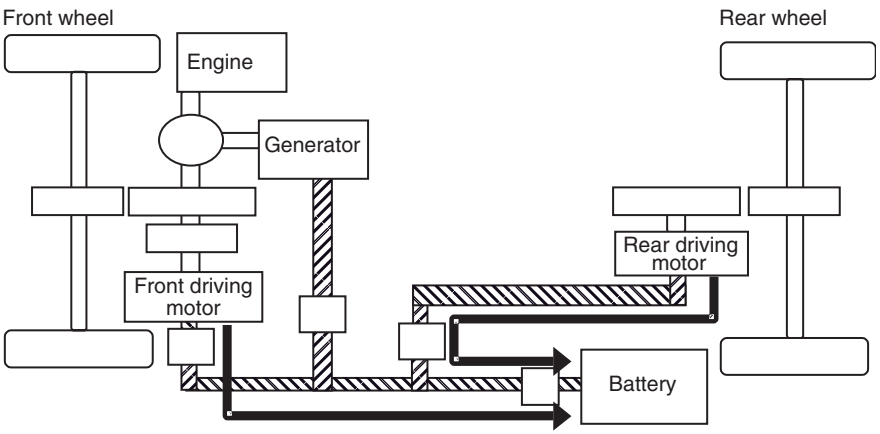


Figure 19. System operation (deceleration and braking). (Reproduced by permission of Toyota Motor Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto067
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

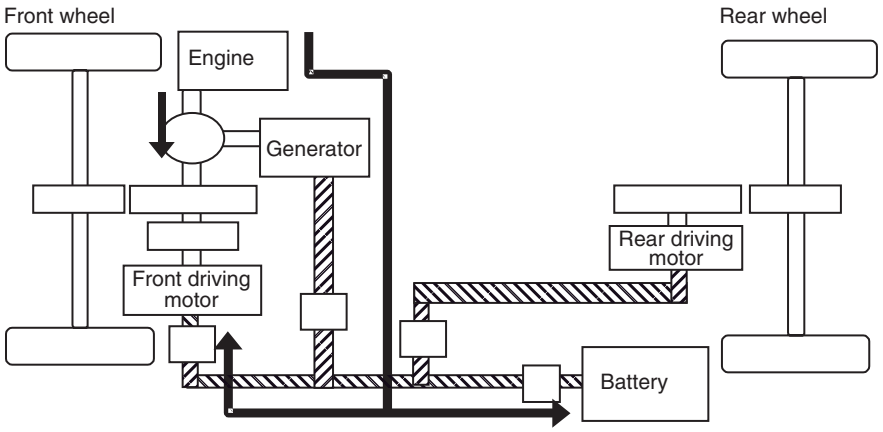


Figure 20. System operation (charging). (Reproduced by permission of Toyota Motor Corporation.)

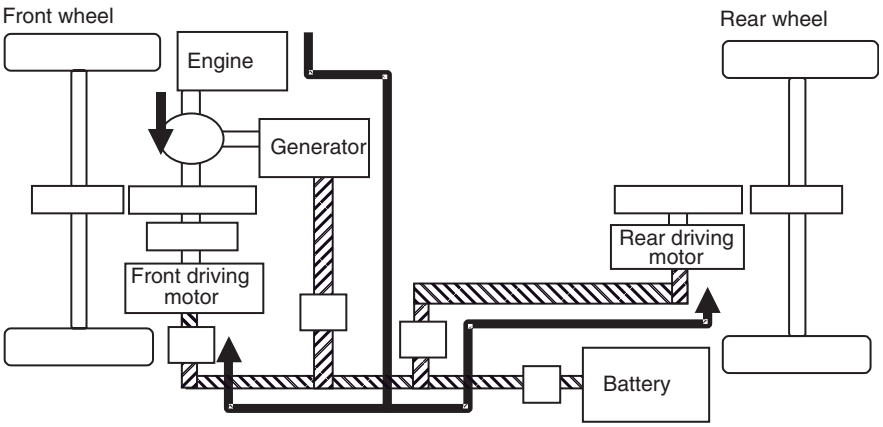


Figure 21. System operation (when front wheels slip). (Reproduced by permission of Toyota Motor Corporation.)

light-load regions if the battery charge is low or the engine coolant temperature is too low.

4.3.5 Deceleration and braking

When decelerating and braking, the wheels are driven by the front- and rear-wheel motors, which are operated as generators (Figure 19). The braking energy of the vehicle is regenerated into electrical power and used to charge the battery.

4.3.6 When charging

The system control maintains the battery charge at a constant level. If the amount of charge falls below a threshold level, the battery is charged using engine power through the generator (Figure 20).

4.3.7 When front wheels slip

When the system detects that the front wheels are slipping, a portion of the power from the generator is used to drive the rear-wheel motor (Figure 21). This reduces the driving force of the front wheels and generates driving force at the rear wheels. If the available power from the generator is not sufficient, the battery provides the necessary adjustment.

REFERENCES

Hata, H., Kamiya, M., and Nagamatsu, S. (2005) Development of a new hybrid transmission for FWD sports utility vehicles. *Toyota Technical Review*, **54** (1), 34–39.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto067
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Kikuchi, K. and Matsubara, K. (2005) Development of high output power control unit for hybrid SUVs. *Toyota Technical Review*, **54** (1), 50–55.

Kimura, A. (2005) Development of hybrid system for SUV. *Toyota Technical Review*, **54** (1), 23–28.

Sadakata, O. and Ikuta, Y. (2005) Introducing the Lexus RX400h. *Toyota Technical Review*, **54** (1), 16–22.

Sasaki, S. (2005) The history of hybrid technology in Toyota. *Toyota Technical Review*, **54** (1), 10–15.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto067
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Engine Management Systems

John Lahti
John Deere Power Systems, Waterloo, IA, USA

1 Introduction 1
2 Engine Management System Components 1
3 Engine Control Strategies 3
4 Individual Cylinder Models 13
5 Conclusion 15
Nomenclature 15
References 16
Further Reading 16

1 INTRODUCTION

This chapter provides an overview of the engine control strategies that are commonly used for diesel and spark ignition engines. Models are now routinely used within the electronic control unit (ECU) to predict parameters that are not measured. The models may also be used for calculating the required actuator positions. These models and their use in the control structure are described. Strategies are explained for modeling and controlling the airflow, exhaust gas recirculation (EGR), variable geometry turbocharger (VGT) vane position, fuel injection, and spark advance. Model fidelity is discussed and a new individual cylinder engine model is introduced.

Engine control strategies for diesel and spark ignition engines are slightly different because of the different combustion strategies, but for the most part the engine models that are used for controlling the engine are the same. The main control differences are in the way the fuel is delivered, how combustion is initiated, and the strategy for regulating the air to fuel ratio. With spark ignition engines, the torque is regulated primarily with the air throttle, while the fuel is normally delivered at a rate that results in a stoichiometric mixture in the cylinder for combustion. Diesel engines regulate torque by directly controlling the fuel injection mass, with the engine running lean most of the time. The fuel injection mass may be limited to prevent smoke when there is insufficient air for complete combustion. Engine models may be used in the controller to predict some of the control parameters. The models of engine flow, throttle flow, EGR, as well as the turbocharger models are the same for both engine types. In both applications, EGR is used to reduce emissions of nitrogen oxides (NO_x). The same models can be used with each engine type to predict the concentration of air in the manifolds and in the cylinder.

2 ENGINE MANAGEMENT SYSTEM COMPONENTS

Engine controls were originally implemented using mechanical devices such as the carburetor, mechanical diesel fuel injector, distributor with centrifugal or vacuum advance, and thermal bimetal actuators. Although these devices provided acceptable performance in many applications and were relatively inexpensive, they could not provide the level of control needed to meet the emission regulations of today. Many of the control functions performed by these devices are now done electronically using sensors and actuators. The sensors provide information about the operating condition of the engine while the actuators are used to regulate its operation. The ECU

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

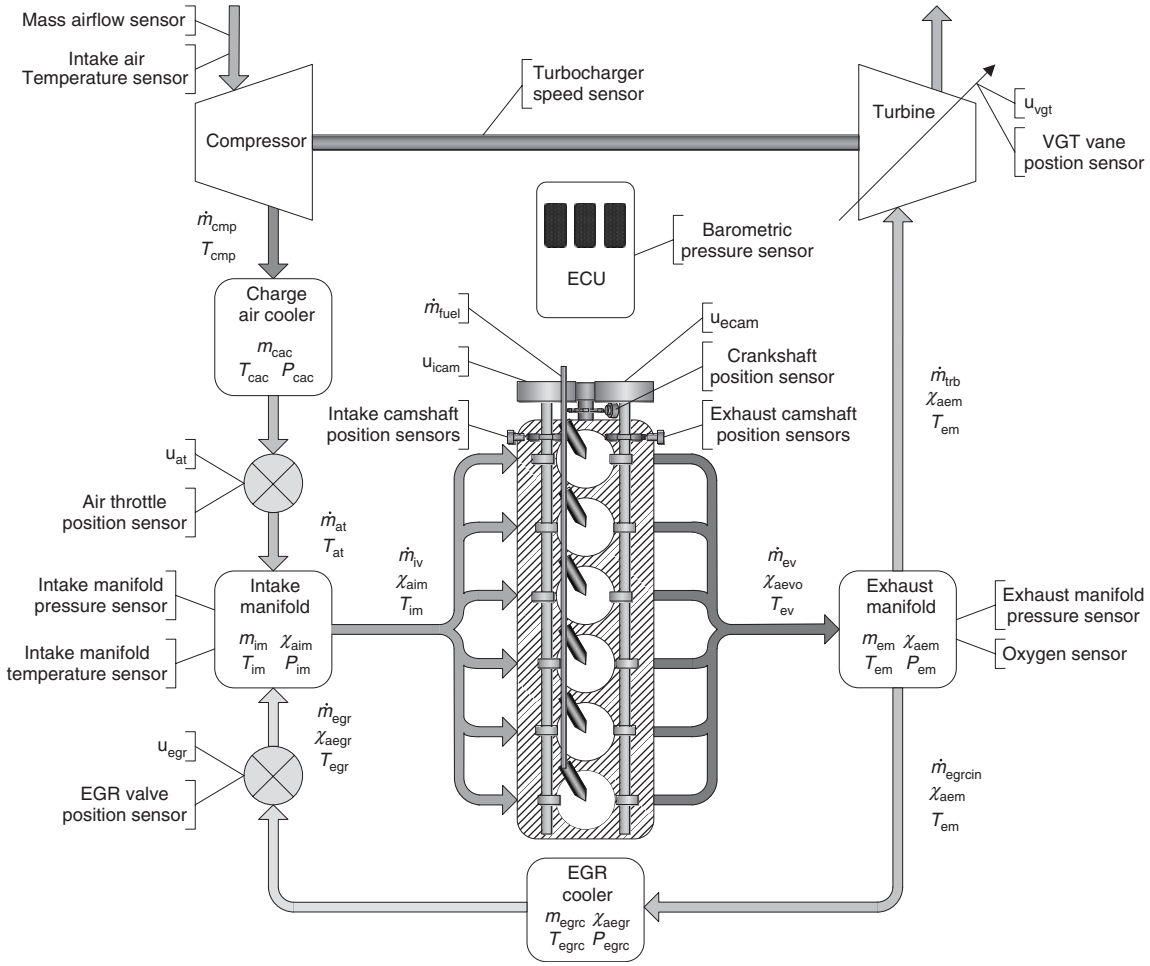


Figure 1. Engine components and model parameters.

processes information from the sensors and determines the desired position for each actuator.

Some of the components that make up the engine control system are shown in Figure 1. Also shown are model parameters described later.

2.1 Sensors

Some sensors interpret inputs from the driver of the vehicle. Examples of these include the accelerator pedal position, transmission range selector, and brake pedal switch.

Other sensors provide information about the operating condition of the engine. These include the coolant temperature sensor, intake air temperature sensor, and barometric pressure sensor. These signals change at a slow rate, allowing the sampling to occur at a slower rate than other sensors.

Some sensors provide information about the current state of the engine and may be used for feed forward and feedback control. These include the crankshaft position sensor, which is used for ignition and fuel injection timing as well as for calculating the engine speed. The camshaft position sensor along with the crankshaft position sensor determines where each cylinder is within the engine cycle. It may also be used to control camshaft phasing if the engine is equipped with variable valve actuation. The manifold air pressure sensor and mass airflow (MAF) sensor are used in the airflow calculations that determine the amount of fuel to inject and what spark advance is required. Oxygen sensors in the exhaust system provide feedback to the engine controller indicating whether the engine is running rich or lean.

2.2 Actuators

Actuators are devices that regulate operation of the engine. Examples of actuators include the fuel injector, air throttle,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

EGR valve, VGT turbine vanes, and ignition system. Actuators that have position control normally have a position sensor that is used with a feedback controller to maintain the desired position.

2.3 Controller

One of the factors contributing to widespread use of electronic engine controls has been emission regulations. Electronic controls make it possible to more accurately control the air to fuel ratio, spark advance, fuel injection timing, and EGR flow rate. Electronic controls can also improve performance, drivability, fuel economy, and integration with other vehicle systems.

Figure 1 shows some of the common sensors and actuators on an engine. The air throttle, EGR valve, and VGT vane are controlled using actuator commands: u_{at}, u_{egr}, and u_{vgt}, respectively. The intake and exhaust camshaft phasing are controlled using commands u_{icam} and u_{ecam}. Engines with electronic fuel injection regulate the fuel rate by controlling the duration of the fuel injector for each cylinder.

3 ENGINE CONTROL STRATEGIES

The block diagram for a typical engine control system is shown in Figure 2. The actuator controls are shown as just one block in this figure but the actuator control may have its own sensor and feedback controller. Proportional-integral-derivative (PID) controllers are commonly used for actuator position control. The actuator controller is within the control loop of the setpoint controller, which requires special consideration when selecting the controller gains. The actuator controls need to be fast enough during transient conditions to prevent the setpoint controller from making adjustments to the actuator setpoint because the actuator position and the corresponding engine response have not yet been achieved. To prevent dynamic interactions between these control loops the actuator control should ideally be more than 10 times faster than the setpoint control. A factor as low as 5 may be acceptable in certain cases.

The feed forward calculation shown in Figure 2 is a calculation of the required actuator position using the given setpoint command and known system parameters. Feed forward control allows the system to respond quicker under transient conditions because the required actuator position is calculated at each time step with essentially no lag. The feedback controller is different in that it is designed to remove control system error over a period with a certain time constant. Using the combination of feed forward and feedback control allows the system to respond quickly to changes in the setpoint command while still having the ability to correct for system changes or errors in the feed forward calculation. When the feed forward calculation is done correctly, the output of the feedback controller will be small.

The feed forward term can be obtained from tables, empirical models, or physics-based models. In some cases, the steady state position for the actuator is determined for each operating condition and those values are then placed in tables that are used to determine the feed forward term. This approach gets the actuator close to the required steady state position fast but it does not provide compensation for the system dynamic so the control under transient conditions is not as good as it could be. Empirical and physics-based models that more accurately account for the system dynamics can provide better transient response.

The main difference between empirical models and physics-based models is that empirical models generally require engine data for calibration whereas physics-based models are based mostly on first principles, allowing them to be calibrated with parameters such as component sizes and fluid properties.

An example of a physics-based model is the compressible gas flow equation for an orifice (Equation 1). This equation is commonly used for modeling airflow through the throttle. The parameter ψ is a function of the pressure ratio across

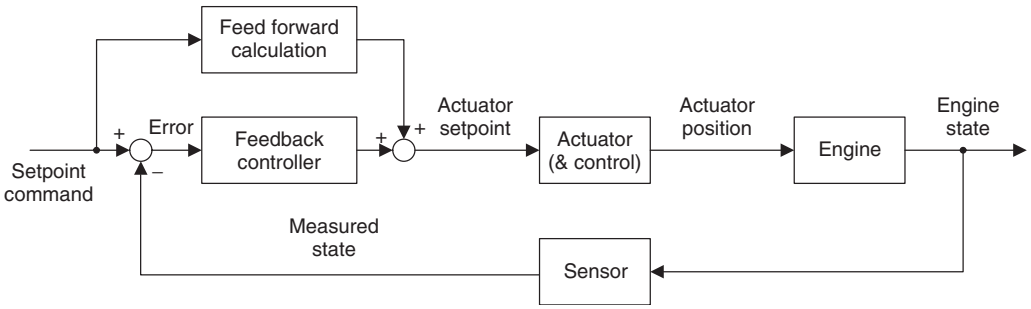


Figure 2. Engine control using sensor feedback.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

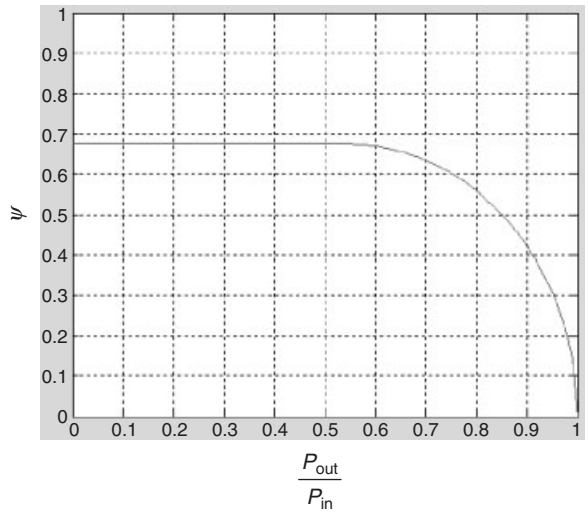


Figure 3. Parameter ψ for air throttle compressible gas flow equation.

the throttle and the ratio of specific heats for the fluid (Guzzella & Onder, 2010). Within the ECU, the parameter ψ is typically stored in a table covering a range of pressure ratios. The table lookup in this case is more efficient in terms of ECU throughput than the equation for ψ. A plot of ψ versus pressure ratio is shown for air in Figure 3. This model considers the throttle opening area, upstream air properties, and pressure ratio across the throttle. The effective area ($C_d \cdot A$) is typically calibrated using data from a flow bench and a table in the ECU with throttle position as the input.

$$\dot{m} = C_d \cdot A \left(\frac{P_{in}}{\sqrt{R_{in} \cdot T_{in}}} \right) \cdot \psi \tag{1}$$

$$\psi = \sqrt{k \cdot \left(\frac{2}{k+1} \right)^{\frac{k+1}{k-1}}}, \quad \text{for } P_{out} < P_{cr}$$

$$\psi = \left(\frac{P_{out}}{P_{in}} \right)^{\frac{1}{k}} \cdot \sqrt{\frac{2k}{k-1} \cdot \left(1 - \left(\frac{P_{out}}{P_{in}} \right)^{\frac{k-1}{k}} \right)}, \quad \text{for } P_{out} \geq P_{cr}$$

$$P_{cr} = \left(\frac{2}{k+1} \right)^{\frac{k}{k-1}} P_{in}$$

An example of an empirical model is the "speed–density" calculation for engine airflow using volumetric efficiency (VE) tables (Equation 2). This is a mean value model of the engine that does not account for the discrete events of each cylinder or the delays associated with the combustion cycle. The model does account for changes in manifold pressure, manifold temperature, and engine speed, making it a reasonably good method for predicting flow to the engine cylinders. This model is discussed in more detail in Section 3.1.1.

Models such as these can be used in a feed forward calculation to determine what manifold pressure is required to achieve the desired airflow to the cylinders or what throttle position is required to achieve a given airflow through the throttle.

In some cases, the desired control parameter does not have a sensor to provide a feedback signal. It may be impractical to have certain sensors because of cost or reliability. For example, it is not practical to measure the flow at the intake valve or to measure the mass fraction of air within the cylinder. If such parameters are important for controlling the engine, a model may be used to estimate these parameters so that the feedback controller can use them. Such a model is called an *observer* (Figure 4). The observer receives the same inputs as the real engine so dynamically it responds similar to the engine. The states of the model are compared to the measured states on the real engine allowing corrections to be made to the model, reducing the parameter estimation error.

In addition to providing state information to the feedback controller, information from the observer model and the model parameters may also be used in the feed forward calculation.

The following discussion describes how an observer model can be used to improve the engine airflow estimate that is used for controlling the air to fuel ratio and setting the spark advance.

3.1 Engine airflow

A mean value engine model is a model of the engine that does not consider the effects of individual cylinders. It assumes the flow through the engine is continuous as it would be in a gas turbine. When such a model is used for control purposes, the mass of air, fuel, or exhaust gas within each cylinder is calculated by evaluating the total mass going through the engine in one cycle and dividing by the number of cylinders.

3.1.1 Speed–density–flow

The speed–density model calculates flow to the engine cylinders using engine speed, the density of the fluid in the intake manifold, the displacement of the engine, and the VE (Equation 2). VE is the ratio of the actual flow to the theoretical flow that would be achieved if flow

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

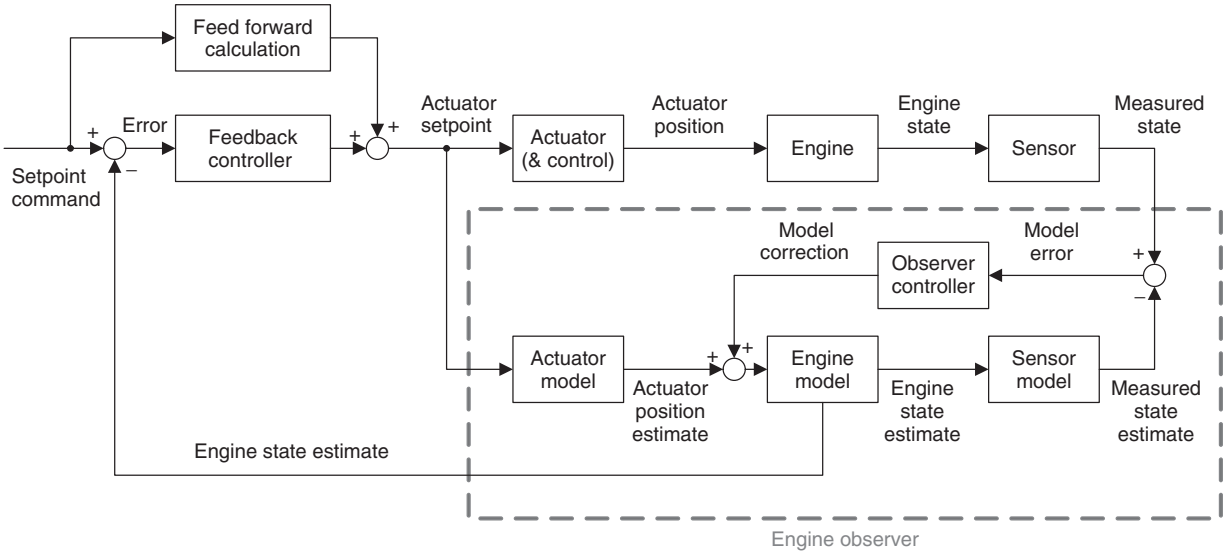


Figure 4. Observer-based engine control.

equivalent to the displaced volume of the engine at the intake manifold fluid density were achieved during each engine cycle. Sometimes, VE is specified with respect to the density of air at atmospheric conditions but this method is seldom used for control purposes. By setting Equation 2 equal to the measured mass flow from an engine test, it is possible to solve for the VE. The VE is sometimes calibrated using a table with axes of engine speed and intake manifold pressure. Turbocharged applications may require a more complex empirical model to accurately predict the VE.

$$\dot{m}_{\mathrm{iv}} = \left(\frac{N_{\mathrm{e}}}{120}\right) \cdot \rho_{\mathrm{im}} \cdot V_{\mathrm{disp}} \cdot \mathrm{VE} \tag{2}$$

This flow estimate includes both air and EGR entering the cylinder. During operation of the engine, the ECU calculates the mass entering each cylinder using the intake manifold fluid density, cylinder displacement, and VE as shown by Equation 3.

$$m_{\mathrm{cyl,in}} = \rho_{\mathrm{i}} \left(\frac{V_{\mathrm{disp}}}{n_{\mathrm{cyl}}}\right) \cdot \mathrm{VE} \tag{3}$$

The total mass in the cylinder at the time when the intake valve closes is equal to the mass that entered through the intake valve plus the residual mass that remained in the cylinder from the previous engine cycle (Equation 4).

$$m_{\mathrm{cyl,ivc}} = m_{\mathrm{cyl,in}} + m_{\mathrm{cyl,res}} \tag{4}$$

A separate model provides the estimate of $m_{\mathrm{cyc,res}}$. The residual mass will be affected by engine speed, manifold pressures, fuel rate, and valve timing.

The speed–density method provides reasonably good estimates of flow to the cylinder under steady state operation but there are several sources of error under transient conditions:

1. The mass of residual exhaust gas retained within the cylinder may change from one cycle to the next, making the VE under transient conditions different than the steady state value. When the intake manifold pressure is increasing the speed–density–flow estimate will be too low because the residual mass has not yet increased to the steady state value, which may cause the spark advance to be set too high resulting in engine knock.
2. The wave dynamics within the intake manifold will be changing under transient conditions and the pressure near the intake valve at the time when the valve closes may be different than it was during the steady state condition, resulting in a different VE.
3. On a port fuel injected engine, the fuel has to be injected before the intake stroke. The manifold pressure may change between the time at which the fuel was injected to the time when the intake valve closes, causing an error in the air to fuel ratio. The flow estimate will be too low when the intake manifold pressure is increasing and it will be too high when the intake manifold pressure is decreasing. The speed–density–flow estimate will cause the engine to run lean on tip-ins and rich on

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

tip-outs. This flow estimate tends to lag behind the true flow to the cylinders.

Some of the limitations listed here can be addressed by adding ad hoc features to the control software that make corrections under certain conditions. For example, when a tip-in is detected the spark advance can be reduced by several degrees to prevent knock. Alternately, to improve fuel control extra fuel can be added with a tip-in, or removed with a tip-out. Calibration of these corrections for all operating conditions can be very time consuming.

The speed–density–flow estimation can be improved by using predicted manifold pressure instead of a measured value that may have to be filtered. The method of using predicted manifold pressure is discussed more in Section 3.1.3.

3.1.2 Measured mass airflow

A MAF sensor measures flow in the air intake duct between the air cleaner and the throttle, or before the compressor on turbocharged engines. Under steady state conditions, this flow (plus the EGR flow rate) should match the speed–density–flow estimate. Under conditions where the intake manifold pressure is changing, the mass contained within the intake manifold will also be changing. For the case with no EGR flow, when the intake manifold pressure is increasing the measured MAF will be higher than the actual airflow into the cylinders. Likewise, when the intake manifold pressure is decreasing the measured MAF will be lower than the actual airflow into the cylinders. These characteristics are opposite to those of the speed–density calculation. The measured MAF tends to lead the true cylinder airflow.

The measured MAF would not provide very good fuel control under transient conditions if used directly. The real benefit to using a MAF sensor is that corrections can be made to the VE, improving the airflow estimate under steady state conditions. The speed–density model provides an estimate of flow to the cylinder while the VE is corrected with the MAF reading. This model-based approach to cylinder air charge estimation is described more in the following section.

3.1.3 Model-based cylinder air charge estimation

One way to improve the estimate of air entering the cylinder is to use a model-based approach. By constructing an observer model of the engine, it is possible to predict the rate of change in intake manifold pressure. The rate of change in manifold pressure can then be used to predict the pressure in the manifold at the time when the intake valve closes. The modeled version of the intake manifold pressure will have less variation than the measured value and will not require filtering. This approach can overcome the lag associated with the speed–density–flow estimate and provide a much more accurate estimate of cylinder air charge for the fuel injection calculations.

The observer model can also estimate the concentration of air and exhaust gas within the cylinder, allowing spark advance and other control parameters to be set for the expected state of the cylinder. This approach improves control under transient conditions.

3.2 Exhaust gas recirculation

EGR significantly increases the complexity of the models needed to predict the mass of air entering the cylinder and the composition of the mixture within the cylinder. Modeling airflow, EGR flow, and residual exhaust gas within the cylinder allows the air per cylinder and exhaust gas concentration to be calculated. This information can then be used to deliver the correct amount of fuel and to set the spark advance.

Diesel engines regulate the exhaust gas concentration for controlling NO_x. Since diesel engines commonly run lean, the calculation of residual and recirculated exhaust gas must consider the concentration of excess air retained in the exhaust gas.

In order to control the mass of air per cylinder and exhaust gas concentration, an estimate of these parameters is needed. The next section describes an engine model that can be used as an observer within the ECU to provide this information.

3.2.1 Exhaust gas recirculation model

The model described in this section includes airflow, exhaust flow, and EGR flow. This is a mean value engine model with lumped parameter manifold models. Lumped parameter means that the concentration of air and exhaust gas is assumed to be evenly distributed within each manifold. The pressure and temperature within the manifold is also assumed to be uniform. Manifold models are based on the principal of conservation of mass. Manifold pressure can be calculated using the ideal gas law with manifold fluid mass and estimated or measured manifold temperature as inputs.

The equations that follow are for an engine with direct fuel injection. The equations will be slightly different for an engine with port fuel injection.

The rate of change in intake manifold mass is equal to the MAF through the throttle, plus the EGR mass flow,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

minus the mass flow entering the engine through the intake valves (Equation 5).

$$\frac{dm_{im}}{dt} = \dot{m}_{at} + \dot{m}_{egr} - \dot{m}_{iv} \tag{5}$$

The rate of change in air mass within the intake manifold is equal to the MAF through the throttle, plus the EGR mass flow times the mass fraction of air contained in the EGR, minus the mass flow entering the engine through the intake valves times the mass fraction of air in the intake manifold (Equation 6).

$$\frac{dm_{aim}}{dt} = \dot{m}_{at} + \dot{m}_{egr}\chi_{aegr} - \dot{m}_{iv} \cdot \chi_{aim} \tag{6}$$

The mass fraction of air contained in the intake manifold is equal to the mass of air in the intake manifold divided by the total mass of air and exhaust gas in the intake manifold (Equation 7).

$$\chi_{aim} = \frac{m_{aim}}{m_{im}} \tag{7}$$

The rate of change in exhaust manifold mass is equal to the mass flow leaving the engine through the exhaust valves, minus the EGR mass flow, minus the mass flow through the turbine (Equation 8).

$$\frac{dm_{em}}{dt} = \dot{m}_{ev} - \dot{m}_{egr} - \dot{m}_{trb} \tag{8}$$

The rate of change in air mass within the exhaust manifold is equal to the mass flow leaving the engine through the exhaust valves times the mass fraction of air in the exhaust gas when the exhaust valve opens, minus the mass flow to the EGR cooler times the mass fraction of air in exhaust manifold, minus the mass flow through the turbine times the mass fraction of air in the exhaust manifold (Equation 9).

$$\frac{dm_{aem}}{dt} = \dot{m}_{ev} \cdot \chi_{aevo} - \dot{m}_{egrcin} \cdot \chi_{aem} - \dot{m}_{trb} \cdot \chi_{aem} \tag{9}$$

The mass fraction of air contained in the exhaust manifold is equal to the mass of air in the exhaust manifold divided by the total mass of air and exhaust gas in the exhaust manifold (Equation 7).

$$\chi_{aem} = \frac{m_{aem}}{m_{em}} \tag{10}$$

The mass flow in the cylinder at the time when the intake valve closes is equal to the mass flow that enters the cylinder through the intake valve plus the residual mass (Equation 11). Mass flow rates are used here because this is a mean value model. In a real engine, these would be discrete masses for each cylinder event.

$$\dot{m}_{ivc} = \dot{m}_{iv} + \dot{m}_{res} \tag{11}$$

The exhaust gas residual mass fraction is defined as the mass of residual exhaust gas divided by the total mass in the cylinder at the time when the intake valve closes (Equation 12). Engine speed, manifold pressures, fuel rate, and valve timing all affect the residual mass fraction.

$$x_r = \frac{m_{res}}{m_{iv} + m_{res}} \tag{12}$$

Equation 12 can be rearranged and the residual mass can be expressed as a flow rate as shown in Equation 13. Expressing the residual mass as a mass flow allows it to be used in the mean value model.

$$\dot{m}_{res} = \dot{m}_{iv}\left(\frac{x_r}{1 - x_r}\right) \tag{13}$$

The mass flow in the cylinder at the time when the exhaust valve opens is equal to the mass flow in the cylinder at the time when the intake valve closes plus the fuel mass flow rate (Equation 14). Mass flow rates are used here because this is a mean value model. In a real engine, these would be discrete masses.

$$\dot{m}_{evo} = \dot{m}_{ivc} + \dot{m}_{fuel} \tag{14}$$

The mass flow through the exhaust valve is equal to the mass flow in the cylinder at the time when the exhaust valve opens minus the residual flow (Equation 15).

$$\dot{m}_{ev} = \dot{m}_{evo} - \dot{m}_{res} \tag{15}$$

The MAF in the cylinder at the time when the intake valve closes is equal to the mass flow through the intake valve times the mass fraction of air within the intake manifold, plus the residual mass flow rate times the mass fraction of air in the cylinder at the time when the exhaust valve opens (Equation 16).

$$\dot{m}_{aivc} = \dot{m}_{iv} \cdot \chi_{aim} + \dot{m}_{res} \cdot \chi_{aevo} \tag{16}$$

The mass fraction of air in the cylinder at the time when the intake valve closes is equal to the mass of air in the cylinder at the time when the intake valve closes divided by the total mass in the cylinder when the intake valve closes (Equation 17).

$$\chi_{aivc} = \frac{\dot{m}_{aivc}}{\dot{m}_{ivc}} \tag{17}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The MAF in the cylinder at the time when the exhaust valve opens is equal to the MAF in the cylinder at the time when the intake valve closes minus the fuel mass flow times the stoichiometric air to fuel ratio (Equation 18). For modeling purposes, the fuel is assumed to react with a stoichiometric amount of air in the cylinder. This assumption only applies when the engine is running lean. If the engine is running rich, there will be no remaining air in the exhaust gas.

$$\dot{m}_{\text{aevo}} = \dot{m}_{\text{aivc}} - \dot{m}_{\text{fuel}} \cdot \text{AFR}_{\text{stoich}} \tag{18}$$

The mass fraction of air in the cylinder at the time when the exhaust valve opens is equal to the mass of air in the cylinder at the time when the exhaust valve opens divided by the total mass in the cylinder when the exhaust valve opens (Equation 19).

$$\chi_{\text{aevo}} = \frac{\dot{m}_{\text{aevo}}}{\dot{m}_{\text{evo}}} \tag{19}$$

The equations presented up to this point can be shown in block diagram form as illustrated in Figures 5 and 6. Figure 5 is a model of the total mass flow through the engine. Figure 6 is a model of airflow and air concentration. The orifice equation for throttle airflow was shown in Equation 1. The speed–density calculation was shown in Equation 2.

The EGR valve and cooler could be modeled as an orifice and a volume, or with an empirical model in a different form. The mass fraction of air in the recirculated exhaust gas could be modeled using a transport delay instead of a volume with uniform mixture distribution. The hardware configuration and desired level of model fidelity will determine which model is the best for the application.

The mass of air and exhaust gas entering the cylinder from the intake manifold depends on the pressure in the manifold and the composition of the mixture within the intake manifold. Cross coupling exists between the pressure and concentration terms. To achieve the same mass of

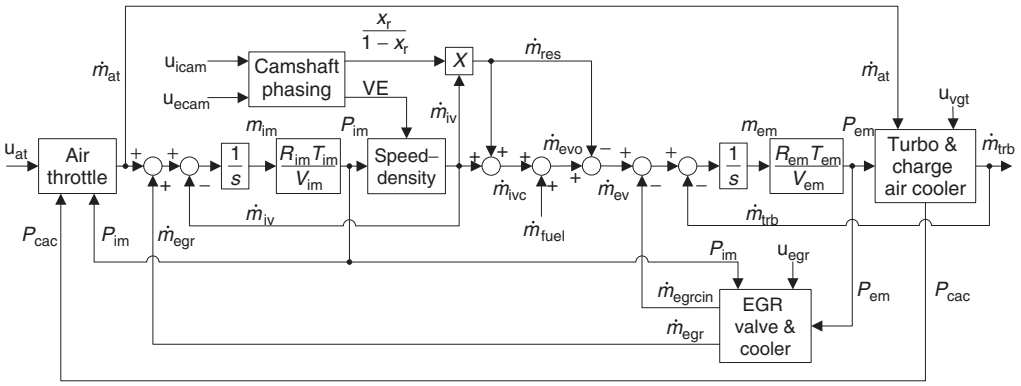


Figure 5. Engine mass flow model.

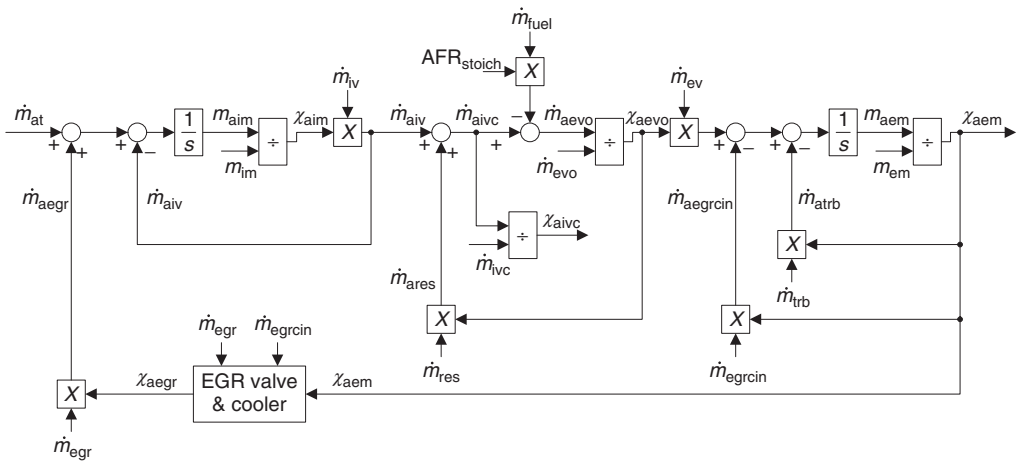


Figure 6. Engine air mass concentration model.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

air in the cylinder with a higher concentration of exhaust gas requires higher intake manifold pressure. In general, opening the EGR valve, closing the turbocharger vanes (higher back pressure), or closing the air throttle (lower intake manifold pressure) will provide more recirculated exhaust gas to the intake manifold, whereas closing the EGR valve or opening the air throttle will provide more air to the intake manifold. Depending on the operating range of the turbocharger, closing the vanes may increase both fresh airflow and EGR flow.

Certain actuators are better at providing fast response under transient conditions while others may be used to guide the system to an efficient operating point. For example, the air throttle provides the fastest air response but in many cases it is desirable to have the throttle opened all the way to minimize pumping losses. Likewise, the EGR valve can provide fast response for controlling the EGR rate but adjusting the VGT vanes also affects the EGR rate and there will be an optimum setpoint for each actuator under steady state conditions.

Different manufacturers have different strategies for controlling these actuators and the details are mostly proprietary. Some research papers have been published on this topic while additional work is still ongoing to determine the best approach. PID controls may provide acceptable performance in certain applications where the control response does not have to be very fast. The controls for these actuators are cross-coupled and very nonlinear, which can limit the gains used in a PID controller. Some researchers have proposed methods for decoupling the system, and others have proposed the use of sliding mode control to better handle the nonlinear characteristic of the system.

Multivariable control is an option that could provide very good control but is more difficult to implement. It uses optimization cost functions to provide a response that uses all the actuators in a way that can be calibrated to provide a response that is considered ideal or "optimal." The cost functions penalize factors such as excessive actuator movement, slow response, or control overshoot. Multivariable control works best with linear systems and the engine system is very nonlinear. One of the steps to implementing multivariable control is to create a model in state variable form, which may require the creation of several linear operating point models of the system. Another challenge is the proper handling of system constraints such as actuator limits, limits on manifold pressure, or turbocharger speed.

3.3 Fuel injection

On diesel engines, the driver demand torque (or governor torque) is primarily controlled with fuel. The other engine actuators respond as needed to provide the correct amount of air and EGR for the mass of fuel that is to be injected into the cylinder. Diesel engines typically run lean so the mass of air in the cylinder is not of much concern until operating at high loads where there may not be sufficient air in the cylinder to prevent smoke. Under a transient smoke-limited operating condition, the fuel injection mass may be limited until the other actuators can make adjustments to provide sufficient air to the cylinder.

With spark ignition engines, the driver demand torque is primarily controlled with the air throttle. The fuel injection quantity depends on the mass of air that is expected to be in the cylinder. A stoichiometric air to fuel ratio is normally maintained so that low levels of both hydrocarbons and NO*x* can be achieved. In addition, a stoichiometric air to fuel ratio allows the three-way catalytic converter to be most efficient at reducing emissions.

The rest of this section focuses on fuel injection controls for spark ignition engines.

3.3.1 Port fuel injection

Switching from carburetors to throttle body fuel injection offered the ability to accurately control the amount of fuel delivered to the engine. The real challenge was calculating how much fuel was required under transient conditions. A speed–density type model could predict the airflow to the engine cylinders with reasonably good accuracy but the changing intake manifold pressure affected the evaporation rate of the fuel and the resulting air to fuel ratio for the mixture entering the cylinders. One solution was to inject extra fuel with a throttle "tip-in" similar to the accelerator pump on a carburetor. Likewise, less than the normal amount of fuel could be injected when a throttle "tip-out" was detected. This ad hoc solution was difficult to calibrate for all operating conditions and did not provide very accurate control.

Aquino (Aquino, 1981) proposed a "wall-wetting" model of the intake manifold that accounted for the mass of liquid fuel within the manifold and the rate at which that fuel evaporated (Figure 7). By inverting this model, the fuel injection quantity can be adjusted to compensate for the liquid fuel accumulation and evaporation that occurs within the manifold.

The wall-wetting model was initially developed for throttle body fuel injection systems and then later applied to engines with port fuel injection. Although each port had independent fuel films that could be modeled separately it was common to continue using a single mean value wall-wetting model.

The following equations are used to model the wall-wetting process. The mass flow of injected fuel entering the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

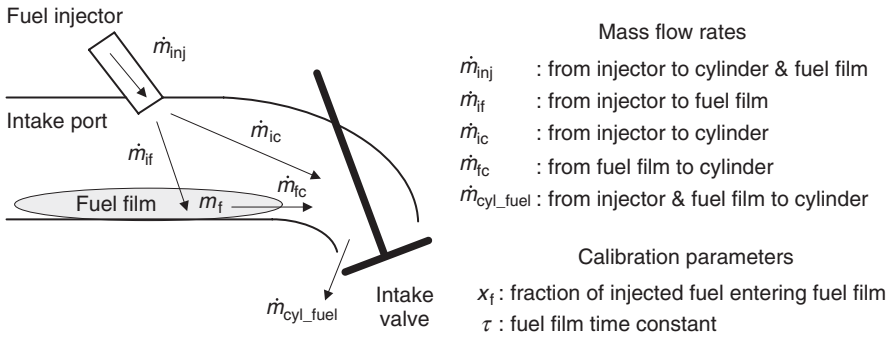


Figure 7. Port fuel injection wall-wetting model.

fuel film is equal to the mass flow of injected fuel times a factor x_f called the *impact factor* (Equation 20). The impact factor is the fraction of injected fuel entering the fuel film.

$$\dot{m}_{if} = x_f \cdot \dot{m}_{inj} \tag{20}$$

The mass flow of injected fuel not entering the fuel film is equal to the mass flow of injected fuel times one minus the impact factor (Equation 21).

$$\dot{m}_{ic} = (1 - x_f) \cdot \dot{m}_{inj} \tag{21}$$

The fuel film is assumed to evaporate with a time constant of τ (Equation 22). Fuel evaporating from the fuel film is assumed to enter the cylinder.

$$\dot{m}_{fc} = \frac{m_f}{\tau} \tag{22}$$

The change in mass of the fuel film is equal to the mass flow entering the film from the injector minus the mass flow that is evaporating and entering the cylinder (Equation 23).

$$\frac{dm_f}{dt} = \dot{m}_{if} - \dot{m}_{fc} \tag{23}$$

Figure 8 shows these equations in block diagram form. This model has the fuel injection mass flow as the input and the fuel mass flow entering the cylinder as the output.

This model can be inverted as shown in Figure 9 to have the mass flow entering the cylinder as the input and the mass of injected fuel as the output. The model in this form

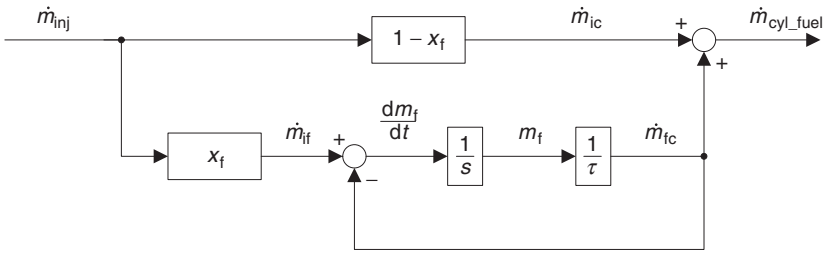


Figure 8. Wall-wetting model block diagram.

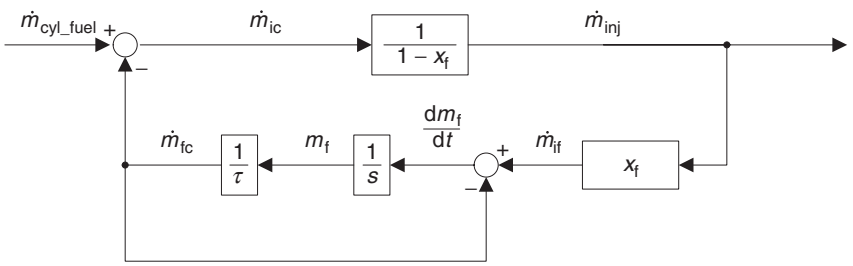


Figure 9. Inverted form of wall-wetting model.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

can be used in the feed forward calculation to determine the amount of fuel to inject. This model will cause more fuel to be injected during a tip-in, when less fuel is evaporating from the fuel film, and less fuel to be injected during a tip-out, when more fuel is evaporating from the film.

3.3.2 Gasoline direct injection

The wall-wetting model is not required for engines with direct fuel injection because all the fuel stays in the cylinder until combustion. Gasoline direct injection normally occurs early in the engine cycle before the intake valve closes. This provides more time for fuel atomization before combustion, reducing hydrocarbon emissions. In addition, the cooling effect of the injected fuel allows more air to enter the cylinder, which allows higher peak torque and power to be achieved.

3.3.3 Closed loop air to fuel ratio control

To achieve a high level of conversion efficiency with the catalytic converter the air to fuel ratio has to be controlled very close to the stoichiometric ratio. Engine to engine variation, variation in fuels, and purging of the evaporative emission canister can cause the engine to run rich or lean of stoichiometry. Closed loop control can be used to correct for these fueling errors. The closed loop fuel controller uses an oxygen sensor for feedback (Figure 10). The oxygen sensor is installed in the exhaust manifold or exhaust pipe and provides a signal that is related to the oxygen concentration in the exhaust.

The oxygen sensor is constructed from a ceramic material called *zirconium oxide*. The sensor is in the shape of a thimble that protrudes into the exhaust stream. The inner and outer surfaces of the sensor are coated with porous layers of platinum, which act as the electrodes. When there is a difference in oxygen concentration between the inner and outer surfaces of the sensor, oxygen ions pass through the ceramic material with reactions occurring at the platinum electrodes generating an electric potential that can be measured as a voltage.

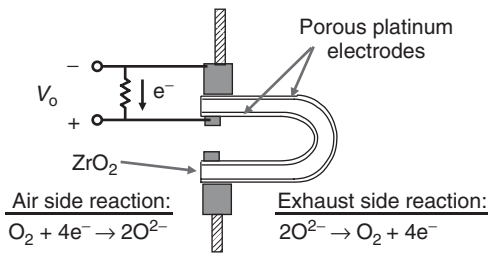


Figure 10. Oxygen sensor.

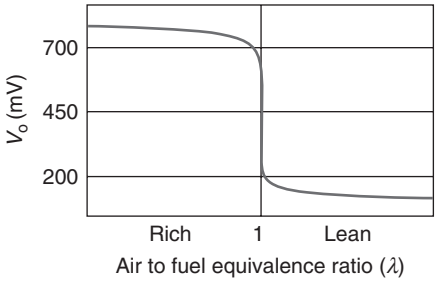


Figure 11. Oxygen sensor response curve.

The sensor voltage is related to the oxygen partial pressure at each electrode. The voltage can be approximated using the Nernst equation:

$$V_o = \frac{R \cdot F}{4F} \cdot \ln\left(\frac{P_{O_2\,\mathrm{ref}}}{P_{O_2\,\mathrm{exh}}}\right) \tag{24}$$

where R is the gas constant, T is the temperature, P is the partial pressure, and F is the Faraday constant. A typical response curve is shown in Figure 11. As the engine switches from running lean to rich, the oxygen sensor voltage increases significantly. The sensor operates as a switch, indicating whether the engine is currently operating rich or lean.

The sensor temperature has to be above a certain value for the reactions to occur at the platinum electrodes, generating the sensor output voltage. The minimum operating temperature for the sensor is about 300°C. Some sensors use an electric heating element to warm up the sensor so that it can be used for control sooner after a cold start.

The sensor is normally used with a proportion integral (PI) controller to adjust the mass of fuel that is injected. Figure 12 shows two examples: control with just integral control and control with a PI controller. With integral control, the fuel injection mass is adjusted up or down at each fuel injection event depending on the output of the sensor. When the oxygen sensor indicates that the engine is running lean, the integrator will increase the fuel injection mass until the oxygen sensor indicates that the engine is running rich; then, the fuel injection mass will start decreasing.

There is a delay between the time when fuel is injected and the time its effect is detected at the sensor. Part of this delay is associated with the engine cycle time and part is due to the transport delay for the exhaust gas to travel from the exhaust valve to the location of the sensor. The transport delay causes the air to fuel ratio at the engine to overshoot the setpoint and affects the period of the rich–lean cycling. The cycle time can be reduced by using a PI controller. Since a certain amount of overshoot is expected because

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

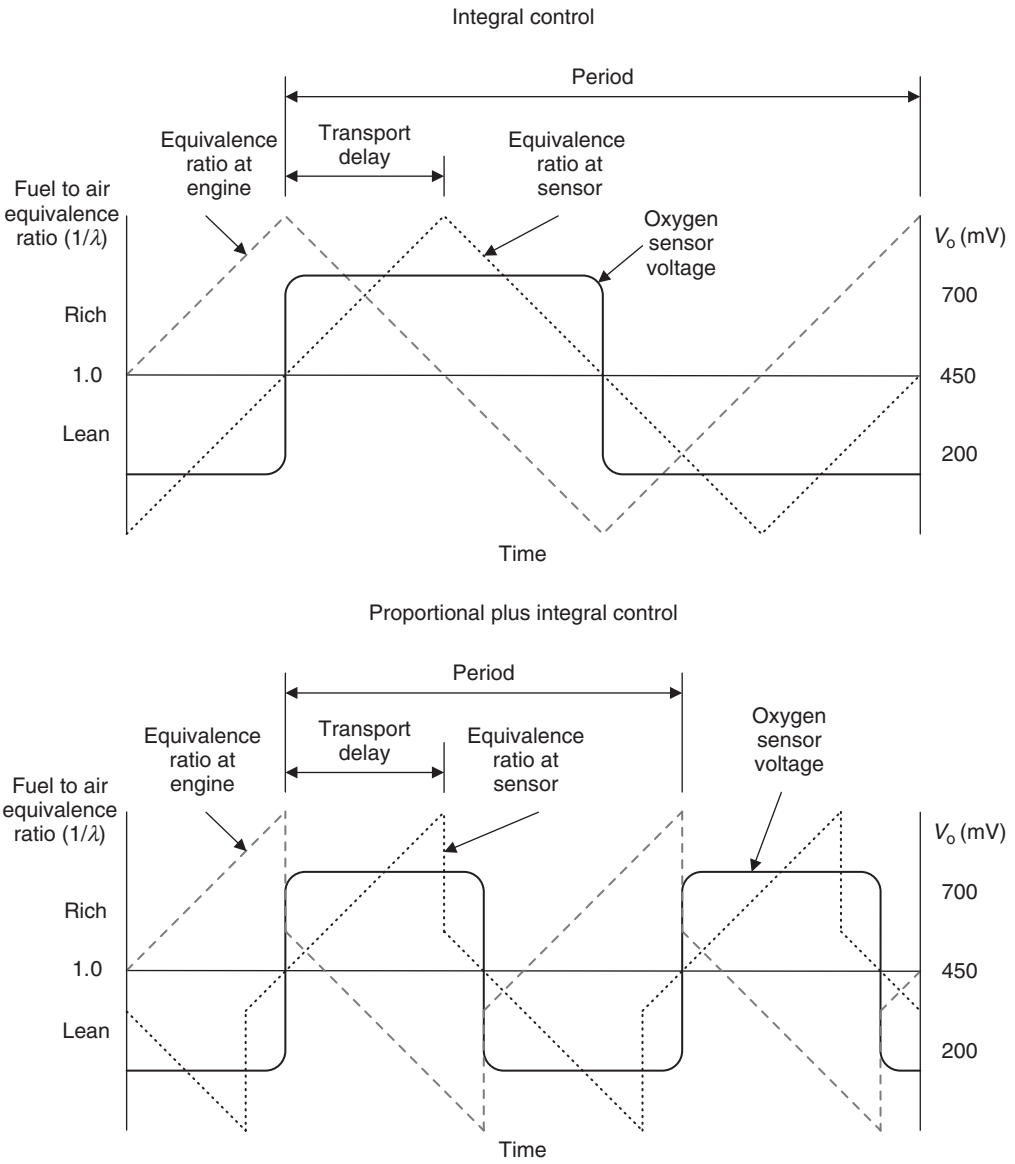


Figure 12. Air to fuel ratio control.

of the transport delay, the proportional term can be used to quickly change the fuel injection mass when a rich-to-lean or lean-to-rich transition occurs at the sensor. Figure 12 shows a reduction in the period of the rich–lean cycling with the PI controller.

3.4 Spark advance

Spark advance is normally set to a value that provides the maximum torque with the minimum amount of spark advance. This is called the *minimum spark advance for best torque* (MBT). Under some conditions with regular grade fuel the spark advance may have to be set lower to prevent engine knock. Knock occurs when the combustion mixture auto-ignites, making a knocking sound rather than burning as a flame front that propagates from the spark plug to the edge of the piston. Combustion may start off as a flame front and then auto-ignite once a certain temperature and pressure is achieved. Reducing the spark advance lowers the pressure during combustion, reducing the tendency to knock. The spark advance for the engine is normally calibrated using tables that are sometimes called maps. The base spark advance is normally set to the minimum of the MBT spark advance and the knock-limited spark advance.

The base spark advance table normally has engine speed as one axis and some indication of engine load on the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

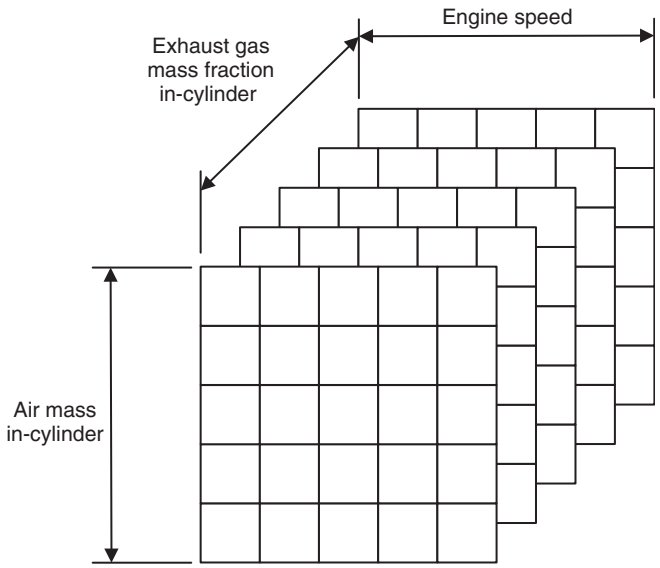


Figure 13. In-cylinder parameter-based spark advance table.

other axis. If only measured parameters are available the throttle position or intake manifold pressure are sometimes used as the second axis. If an engine airflow observer is implemented in the ECU, then the mass of air per cylinder could be used for the second axis.

The combustion burn rate is affected by the mass fraction of exhaust gas contained in the cylinder. If an estimate of exhaust gas mass fraction is available it could be used as a third table axis as shown in Figure 13.

The use of in-cylinder parameters for setting the spark advance provides good control under transient conditions because the advance is set using factors that ultimately affect the combustion process.

The spark advance may also be adjusted for non-standard operating conditions such as when the engine is colder or hotter than normal. These adjustments can be made using additional tables that adjust the spark advance based on coolant temperature or the estimated in-cylinder temperature.

Using only actuator positions or sensor measurements as inputs to spark advance tables would require many tables to cover all operating conditions and would not provide the level of transient control that is possible with in-cylinder estimated parameters.

4 INDIVIDUAL CYLINDER MODELS

As can be seen from what has been described up to this point, the mean value engine model can provide a lot of useful information for controlling the engine when implemented as an observer in the ECU. There are still some sources of error that may need to be addressed if more accurate control is required for advanced combustion strategies such as homogeneous charge compression ignition (HCCI). These limitations are as follows:

1. The real engine operates in discrete cylinder events with delays associated with the engine cycle that are not properly captured by the mean value model. This can cause errors in the residual cylinder mass and composition calculations.
2. The model does not account for the wave dynamics in the intake manifold that can significantly change the flow entering the cylinder under transient conditions from that of the flow predicted by the speed–density method using VE tables.
3. Engines with variable valvetrain systems offer a wide range of valve-opening strategies, making it difficult to accurately model all possible operating conditions with VE tables.

These limitations can be overcome by using a higher fidelity model that includes individual cylinders and by modeling the wave dynamics of the intake and exhaust manifolds. Such a model provides more information but requires more processing capacity from the ECU. This approach eliminates the need for VE tables.

A project at the University of Wisconsin–Madison created a real-time combustion and compressible gas flow model that could be used in an ECU as an observer (Lahti, 2004). The following discussion provides an overview of that model.

The wave dynamics were modeled using a process called the *method of characteristics*. The governing equations are the continuity equation and the momentum equation. Through a process or variable transformations the state of the fluid in the manifold runners can be defined using parameters called *Riemann variables*. One Riemann variable defines the right moving characteristic and the other defines the left moving characteristic. For isentropic flow, the Riemann variables remain constant as they propagate through the manifold runner. This modeling technique makes it possible to predict the state of the fluid at the valve and to predict the flow through the valve when the cylinder pressure is known. The reader is referred to Benson (Benson, 1982) for more details on the wave modeling techniques.

The method of characteristics was originally developed for solving wave problems on a drafting board. Later it was implemented as a computer program. Other wave analysis methods may provide more accuracy but the software code may not run fast enough to be used in a real-time

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

application. This method was found to work well when the isentropic flow assumption was not violated. If the valve timing was such that hot cylinder gases entered the intake runner during part of the intake event, the model would not accurately represent the wave dynamics at that point. If such a valve-opening strategy were required, a slightly more complex model could be implemented to more accurately model those effects.

An individual cylinder model was developed to calculate the temperature and pressure in the cylinder throughout the engine cycle. This information was used with the wave dynamics model to determine the flow through the valves.

The states of the cylinder were modeled using the first law of thermodynamics, conservation of mass, and the ideal gas law. The first law equation for the cylinder is

$$\dot{Q}_{\mathrm{cv}} - \dot{W}_{\mathrm{cv}} + \dot{m}_{\mathrm{iv}} \cdot h_{\mathrm{Si}} - \dot{m}_{\mathrm{ev}} \cdot h_{\mathrm{Se}} = \frac{\mathrm{d}(m u_{\mathrm{S}})\mathrm{cv}}{\mathrm{d}t} \tag{25}$$

This right side of the equation can be rewritten as

$$\frac{\mathrm{d}(m u_{\mathrm{S}})\mathrm{cv}}{\mathrm{d}t} = (\dot{m}_{\mathrm{iv}} - \dot{m}_{\mathrm{ev}}) \cdot u_{\mathrm{S}} + m_{\mathrm{cyl}} \cdot C_{\mathrm{v}} \cdot \frac{\mathrm{d}T_{\mathrm{cyl}}}{\mathrm{d}t} \tag{26}$$

It is now possible to solve for the rate of temperature change:

$$\frac{\mathrm{d}T_{\mathrm{cyl}}}{\mathrm{d}t} = \frac{\begin{array}{c}\dot{Q}_{\mathrm{comb}} + \dot{Q}_{\mathrm{wall}} - \dot{W}_{\mathrm{cv}} + \dot{m}_{\mathrm{iv}} \cdot h_{\mathrm{Si}} \\ -\dot{m}_{\mathrm{ev}} \cdot h_{\mathrm{Se}} - (\dot{m}_{\mathrm{iv}} - \dot{m}_{\mathrm{ev}}) \cdot u_{\mathrm{S}}\end{array}}{m_{\mathrm{cyl}} \cdot C_{\mathrm{v}}} \tag{27}$$

The temperature at each time step is calculated as

$$T_{\mathrm{cyl}}(t + \Delta t) = T_{\mathrm{cyl}}(t) + \frac{\mathrm{d}T_{\mathrm{cyl}}(t)}{\mathrm{d}t}\Delta t \tag{28}$$

The cylinder mass is updated using the mass flow rates at the valves

$$m_{\mathrm{cyl}}(t + \Delta t) = m_{\mathrm{cyl}}(t) + (\dot{m}_{\mathrm{iv}} - \dot{m}_{\mathrm{ev}}) \cdot \Delta t \tag{29}$$

The pressure at the new time step is calculated using the ideal gas law:

$$P_{\mathrm{cyl}}(t + \Delta t) = \frac{m_{\mathrm{cyl}}(t + \Delta t) \cdot R_{\mathrm{cyl}}(t + \Delta t) \cdot T_{\mathrm{cyl}}(t + \Delta t)}{V_{\mathrm{cyl}}(t + \Delta t)} \tag{30}$$

This model was evaluated using a single cylinder spark ignition research engine. Figure 14 shows the measured and estimated cylinder pressures through one engine cycle. During this test, the model was running in real time using a dSPACE rapid prototype control system. The measured and estimated cylinder pressures were nearly the same throughout the cycle. Similar results were obtained for different combinations of valve timing, valve lift, airflow, and engine speed.

The combustion heat release was modeled using a mathematical function called a *Weibe function*. Weibe functions are commonly used to model the mass fraction of fuel that has burned as a function of crank angle. The calibration parameters for the Weibe function were stored in tables

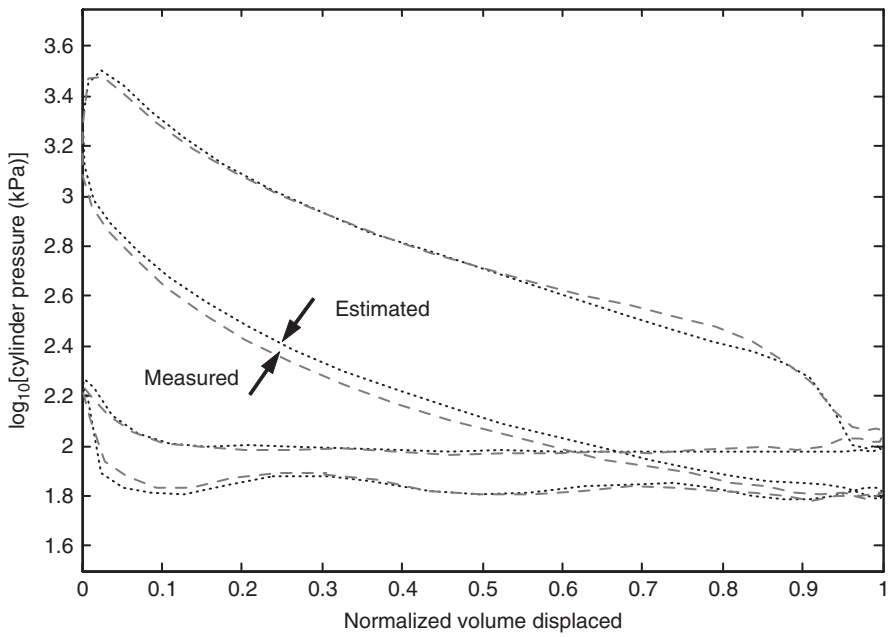


Figure 14. Real-time individual cylinder model data.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of the form shown in Figure 13 to be consistent with the strategy used for the spark timing. More complex models could be implemented to define the heat release for applications using premixed, partially premixed, or multiple direct inject combustion strategies.

With the individual cylinder models, each cylinder event is a transient event. The cylinder pressures and temperatures are continually changing. There is intermittent flow through the valves, and the mass within the cylinder keeps changing. This is much different than the mean value model where the flow is continuous and the engine cycles are assumed to occur with no delay.

One of the challenges to implementing HCCI in a production application is controlling the process under transient conditions. Current engine control strategies using mean value models are not able to provide information about the state of the engine with sufficient accuracy to control the process. The individual cylinder models offer an alternative that accurately represents many of the factors that affect when combustion begins. Information from the models could be used for controlling the trapped residual mass and EGR flow as a way of regulating the HCCI process.

A patent for the engine control technology described in this section is held by the University of Wisconsin Alumni Research Foundation (WARF) at the University of Wisconsin–Madison (Lahti & Moskwa, 2006).

5 CONCLUSION

In many cases, the engine parameters that need to be controlled are difficult or impractical to measure. It may be possible to model the important parameters using the actuator and sensor information that is available. Such a model is called *an observer*. There are several advantages to using observers: they provide the desired state feedback information without adding sensors, the modeled response does not require filtering like a sensed parameter, the model information can be used for feed forward calculations to provide better control response, and the observer can be used for diagnostics to detect changes in engine operation.

The mean value engine model may provide sufficient information for controlling the engine in some applications. As with any model, assumptions are made and those assumptions can be a source of error under certain conditions. The mean value model assumes continuous flow through the engine with no delays. This assumption may be acceptable for some applications but not for others.

The level of model fidelity for an application may vary depending on the requirements. Increasing model fidelity has many control advantages but the model must be capable of running in real time within the ECU, which has limited processing capacity. As processor capacity increases, in the future it may be possible to implement higher fidelity models such as the individual cylinder model, and models of the manifold wave dynamics. These higher fidelity models allow better control under transient conditions, which may resolve some of the implementation problems associated with alternative combustion strategies.

NOMENCLATURE

The variables listed below are used in the equations that follow.

Δt	controller time step
ρ_{im}	intake manifold fluid density
τ	wall wetting time constant
χ_{aegr}	mass fraction of air in EGR
χ_{aem}	mass fraction of air in exhaust manifold
χ_{aevo}	mass fraction of air when exhaust valve opens
χ_{aim}	mass fraction of air in intake manifold
χ_{aivc}	mass fraction of air when intake valve closes
ψ	compressible gas flow parameter
A	orifice area
$\text{AFR}_{\text{stoich}}$	stoichiometric air to fuel ratio
C_{d}	discharge coefficient
C_{p}	specific heat at constant pressure
C_{v}	specific heat at constant volume
ECU	engine control unit
EGR	exhaust gas recirculation
H	enthalpy
H_{s}	sensible enthalpy
k	ratio of specific heats
m_{aem}	mass of air in exhaust manifold
m_{aim}	mass of air in intake manifold
m_{cac}	mass in charge air cooler
m_{cyc}	mass in cylinder
$m_{\text{cyl,in}}$	mass that entered the cylinder from intake valve
$m_{\text{cyl,ivc}}$	mass in cylinder when intake valve closes
$m_{\text{cyl,res}}$	residual mass in cylinder
m_{egrc}	mass in EGR cooler
m_{em}	mass in exhaust manifold
m_{f}	wall-wetting model fuel film mass
m_{im}	mass in intake manifold
$\dot{m}$	mass flow through orifice
$\dot{m}_{\text{aevo}}$	mass flow of air (mean value model) when exhaust valve opens
$\dot{m}_{\text{aivc}}$	mass flow of air (mean value model) when intake valve closes

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

$\dot{m}_{at}$	mass airflow through throttle
$\dot{m}_{egr}$	mass flow of EGR
$\dot{m}_{egrcin}$	mass flow into the EGR cooler
$\dot{m}_{ev}$	mass flow through exhaust valve
$\dot{m}_{evo}$	mass flow when the exhaust valve opens
$\dot{m}_{fc}$	mass flow of fuel from film to cylinder
$\dot{m}_{fuel}$	mass flow of fuel to engine
$\dot{m}_{ic}$	mass flow of injected fuel entering cylinder
$\dot{m}_{if}$	mass flow of injected fuel entering fuel film
$\dot{m}_{inj}$	mass flow of injected fuel
$\dot{m}_{iv}$	mass flow through intake valve
$\dot{m}_{ivc}$	mass flow when intake valve closes
$\dot{m}_{res}$	mass flow of residual mass in cylinder
$\dot{m}_{trb}$	mass flow to turbine or exhaust pipe
MAF	mass airflow
n_{cyl}	number of cylinders
N_e	engine speed (rpm)
NO_x	nitrogen oxides
P_{cac}	pressure in charge air cooler
P_{cyl}	pressure in cylinder
P_{cr}	critical pressure
P_{egrc}	pressure in EGR cooler
P_{em}	pressure in exhaust manifold
PI	proportional-integral controller
PID	proportional-integral-derivative controller
P_{im}	pressure in intake manifold
P_{in}	orifice inlet pressure
P_{out}	orifice outlet pressure
$\dot{Q}_{comb}$	combustion heat release rate
$\dot{Q}_{cv}$	control volume heat transfer rate
$\dot{Q}_{wall}$	rate of heat transfer to the cylinder walls
R	gas constant
R_{in}	gas constant of flow entering orifice
T	time
T_{cac}	temperature in charge air cooler
T_{cyl}	temperature in cylinder
T_{egrc}	temperature in EGR cooler
T_{em}	temperature in exhaust manifold
T_{im}	temperature in intake manifold
T_{in}	temperature of flow entering orifice
U	internal energy
u_{at}	actuator command for air throttle position
u_{egr}	actuator command for EGR valve position
u_{ecam}	actuator command for exhaust camshaft position
u_{icam}	actuator command for intake camshaft position
u_{vgt}	actuator command for VGT vane position
U_S	sensible energy
V_{cyl}	volume of cylinder
V_{disp}	engine displacement
VE	volumetric efficiency
VGT	variable geometry turbocharger
$\dot{W}_{cv}$	rate of work done by the control volume
X_f	mass fraction of injected fuel entering fuel film
X_r	exhaust gas residual mass fraction

REFERENCES

Aquino, C.F. (1981) Transient A/F control characteristics of the 5 liter central fuel injection engine. SAE 810494.

Benson, R.S. (1982) *The Thermodynamics and Gas Dynamics of Internal-Combustion Engines*, vol. 1, Clarendon Press, Oxford.

Guzzella, L. and Onder, C.H. (2010) *Introduction to Modeling and Control of Internal Combustion Engine Systems*, 2nd edn, Springer, Heidelberg.

Lahti, J.L. (2004) Engine control using real time combustion and compressible gas flow models. PhD dissertation. University of Wisconsin–Madison.

Lahti, J.L. and Moskwa, J.J. (2006) Internal Combustion Engine Control System. United States Patent and Trademark Office, Patent No. 7,275,426 B2, March 31.

FURTHER READING

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*, McGraw-Hill, Inc., New York.

Heywood, J.B., Higgins, J.M., Watts, P.A., and Tabaczynski, R.J. (1979) Development and use of a cycle simulation to predict SI engine efficiency and NO*x* emissions. SAE 790291.

Iwadare, M., Ueno, M., and Adachi, S. (2009) Multi-variable air-path management for a clean diesel engine using model predictive control. SAE 2009-01-0733.

Jante, A. (1960) The Wiebe Combustion Law (Das Wiebe-Brenngesetz, ein Forschritt in der Thermodynamik der Kreisprozess von Verbrennungsmotoren), Kraftfahrzeugtchnik, vol. 9, pp. 340–346.

Lahti, J.L. and Moskwa, J.J. (2005) Engine Control Using Estimated Parameters from a Real Time Model of an Engine with Variable Valve Actuation. *Proceedings of the ASME International Mechanical Engineering Congress and Exposition*, IMECE2005-81362, Orlando.

Lahti, J.L., Snyder, M.W., and Moskwa, J.J. (2005) A Transient Single Cylinder Test System for Engine Research and Control Development. *Proceedings of the ASME International Mechanical Engineering Congress and Exposition*, IMECE2005-81323, Orlando.

Luenberger, D.G. (1971) An introduction to observers. *IEEE Transactions on Automatic Control*, AC-16(6), 596–602.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.

This article is © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

McBride, B.J., Sanford, G., and Reno, M.A. (1993) Coefficients for Calculating Thermodynamic and Transport Properties of Individual Species. NASA Technical Memorandum 4513, US Department of Commerce, National Technical Information Service, October 1993.

Oppenheim, A.K. and Kuhl, A.L. (1998) Life of fuel in engine cylinder. SAE 980780.

Vibe, I.I. (1956) Semi-Empirical Expression for Combustion Rate in Engines. *Proceedings of Conference on Piston Engines*, USSR Academy of Sciences, Moscow, pp. 185–191.

Wahlstrom, J. and Eriksson, L. (2010) Nonlinear input transformation for EGR and VGT control in diesel engines. SAE 2010-01-2203.

Woschni, G. (1967) A universally applicable equation for the instantaneous heat transfer coefficient in the internal combustion engine. SAE 670931.

Young, C.T. (1981) Experimental analysis of ZrO_2 oxygen sensor transient switching behavior. SAE 810380.

Young, C.T. and Bode, J.D. (1979) Characteristics of ZrO_2-type oxygen sensors for automotive applications. SAE 790142.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto068
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Turbocharging

Syed Shahed

3E Technology LLC, Rancho Palos Verdes, CA, USA

1 Introduction: Turbocharging History and Perspective 1
2 Benefits of Turbocharging—Diesel and Gasoline Engines 2
3 Turbochargers—Basic Structure and Functionality 3
4 Turbocharger Performance 6
5 Engine/Turbocharger Matching Basics 9
6 Advanced Engine Requirements and Turbo Technologies 10
7 Summary 16
Endnotes 17
Acknowledgment 17
References 17

1 INTRODUCTION: TURBOCHARGING HISTORY AND PERSPECTIVE

The first turbocharger powered by engine exhaust was developed by Alfred J. Buchi, Chief Engineer, Sulzer Brothers Research Department, and was patented[1] in 1905. He did propose a turbocharged diesel engine in 1915 but early application was to aircraft gasoline engines. The primary purpose of boosting the intake pressure to an internal combustion engine is to increase power. The power developed by an engine is a function of the amount of fuel burnt, which is a function of the amount of air available and effectively utilized (near 100% in premixed gasoline engines, about 60–70% in diesel engines). The volume of air required is 9000 (for gasoline) to 20,000 (for diesel) times the volume of fuel. Hence, the power developed by an engine is limited by the amount of air available. Turbocharging greatly increases the amount of air an engine is able to breathe, and hence is greatly effective in increasing the power density (power/displacement). The impact of boost pressure on power density at high altitude is obvious. Hence, the first application of turbocharging was to gasoline engines used in aircraft. In 1920, the LaPere biplane achieved a record altitude for the time—33,113 ft.

Turbocharging was applied to diesel engines by Elliott Company for large marine, industrial, and locomotive applications. The Cummins turbocharged diesel engine[2] achieved the pole position at the Indianapolis 500 in 1952. Simultaneously, the 1950s saw the application of turbocharging to off-road diesel engines in construction equipment (Caterpillar) largely for power density and smoke control. Application to on-road truck engines (Cummins, Volvo) followed soon after.

The 1970s saw a great acceleration of turbocharging and charge-cooling for all heavy-duty diesel engines, not only for power density but also for emissions control and improved efficiency. The 1990s saw a great acceleration of turbocharged diesel engines for passenger car applications in Europe driven largely by the need for fuel economy. Today, all heavy-duty equipment and vehicles use turbocharged diesel engines. Turbo-diesel passenger cars are around 50% of all cars in Europe and rapidly increasing in the United States. Turbocharging is such an essential part of diesel engines that today there are no serious diesel engine applications without turbocharging.

Application of turbocharging to gasoline engines for passenger car applications is attracting increasing attention.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.

This article is © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

The reason (as in early application) is increased power density, but this time not for high altitude but for engine downsizing and "de-throttling" at part load. This effect is discussed in greater detail in Section 2.

2 BENEFITS OF TURBOCHARGING—DIESEL AND GASOLINE ENGINES

As has already been noted, the benefits of turbocharging diesel engines are so overwhelming that there are no significant naturally aspirated diesel engines. Therefore, in discussing benefits of turbocharging in diesel engines, a comparison between turbocharged and naturally aspirated diesel engines is not important. Also, the benefits of turbocharged diesel engines in heavy-duty trucks and equipment is so well established that a discussion of turbocharged versus naturally aspirated diesel engines or gasoline engines is not relevant. However, turbo-diesel application in passenger vehicles, while well established in Europe, is still in its starting phase in the United States. Therefore, such a comparison is of much interest. Figure 1 shows this comparison for 2011 model year US vehicles. All vehicles are certified to meet the US Environmental Protection Agency (EPA) Tier II Bin 5 emission regulations. The data are taken from EPA fuel economy Web site and manufacturer Web sites. This is not a back-to-back comparison in the same vehicle and there may be other technology differences that contribute to the difference, but the overall picture is very effective, showing a 30–50% advantage in fuel consumption for turbo-diesel vehicles.

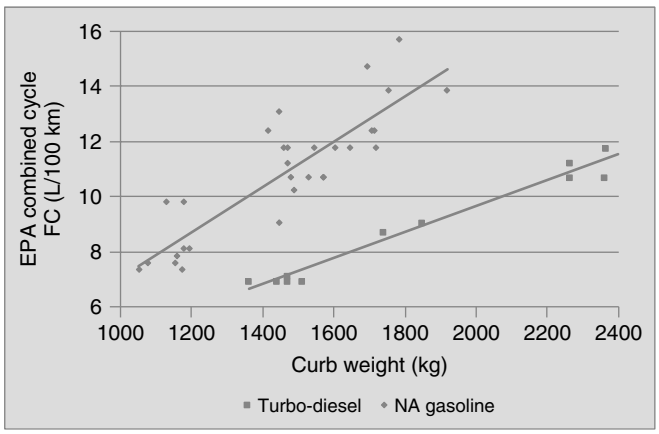


Figure 1. Fuel economy comparison between turbo-diesel and naturally aspirated gasoline US 2011 model year passenger vehicles.

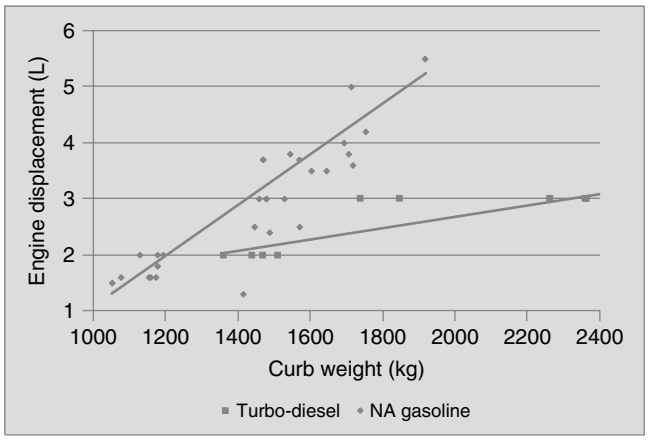


Figure 2. US 2011 model year turbo-diesel and naturally aspirated gasoline engine displacement in comparable vehicles.

One other significant factor needs to be pointed out. Figure 2 shows the engine displacement for exactly the same vehicles as in Figure 1. It is seen that turbo diesel engines are 30–50% smaller in displacement. This should have an impact on cost and somewhat mitigate the added cost of emissions control equipment. The same observation can be made about vehicle curb weight.

The reasons for fuel economy benefits are related to the high throttling losses in gasoline engines compared to turbo-diesel engines and to the differences in compression ratio (as well as energy content of a liter of each fuel). Smaller displacement diesel engines can be used partly because turbocharging increases power density but mostly because turbo-diesel engines have higher low speed torque and a small engine gives good drivability compared to a large, naturally aspirated gasoline engine.

Similar benefits can be shown in vehicles certified for European emissions regulations, and, as has already been observed, turbo-diesel penetration in European passenger vehicles is very high.

Turbocharging and downsizing is on the increase in gasoline engines also. Figures 3 and 4 show, respectively, the fuel consumption and displacement comparison of model year 2011 US passenger vehicles with naturally aspirated and turbocharged engine options. It is seen that downsized turbo gasoline engines give about 20% benefit in fuel consumption and enable about 50% reduction in engine displacement.

The fuel consumption benefit shown in Figure 3 is the result of using smaller engines, as shown in Figure 4. At light loads, gasoline engines are throttled. Throttling losses are a major factor in the poor fuel economy of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

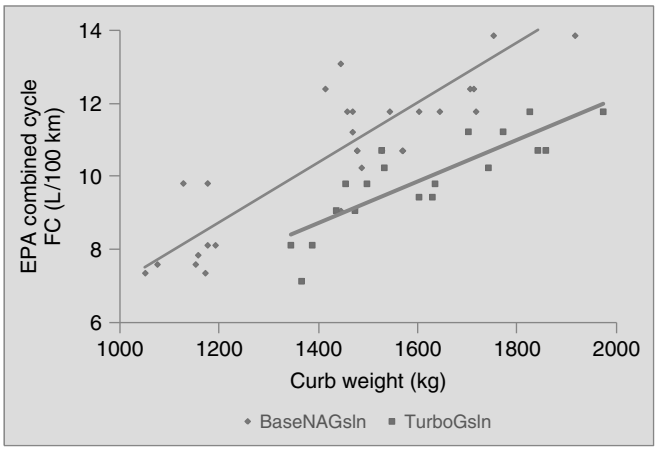


Figure 3. Fuel consumption in US model year 2011 vehicles—naturally aspirated and turbocharged gasoline engines.

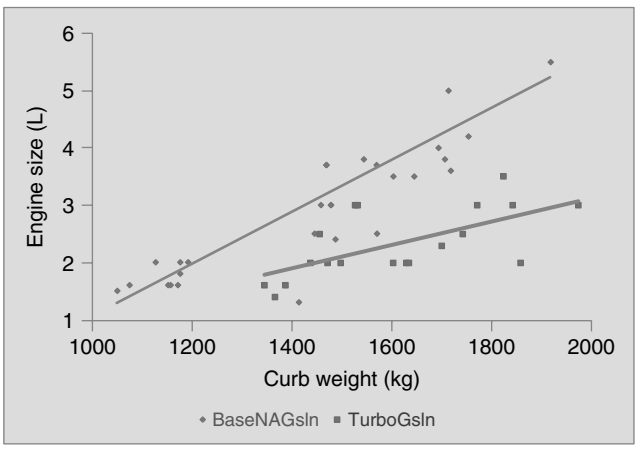


Figure 4. Engine displacement in US model year 2011 vehicles—naturally aspirated and turbocharged gasoline engines.

gasoline engines. If the engine displacement is large, greater throttling and greater throttling losses are involved. For smaller engines, less throttling is required for the same load, hence throttling losses are correspondingly lower. A detailed discussion of the mechanisms of these effects can be found in Walzer (2001). The benefits shown in Figure 3 are a major factor driving the growth of turbocharged, downsized gasoline engines in passenger vehicles.

Requirements placed on turbochargers for diesel engine and gasoline engine applications are different, and differences in design and materials result from this. However, the basic fluid mechanics and thermodynamics are the same. These aspects are discussed in the following sections.

3 TURBOCHARGERS—BASIC STRUCTURE AND FUNCTIONALITY

In this and the following sections, turbocharger designs and design methodology are discussed. The discussion is primarily from an engine design and development point of view, not from a detailed turbocharger design point of view. Turbocharger design is a highly specialized field involving a deep understanding of fluid mechanics and stresses in high speed rotating machinery. Requirements placed on turbochargers, turbo–engine interactions, design responses, design methodology, and advanced concepts are discussed. This is by no means a stand-alone turbocharger design manual.

The basic functionality of a turbocharger is illustrated in Figure 5.[3] Exhaust gas from the engine is made to pass through a turbine wheel, causing the wheel to rotate at a high speed. A compressor wheel is connected to the turbine on the same shaft and rotates at the same high speed inducting fresh air, imparting momentum and high velocity to this air. High velocity discharge air is slowed down aerodynamically in a diffuser/housing, and kinetic energy converted to high pressure/density. The inefficiencies of the compression and recovery processes increase the temperature of air, which is cooled in a charge cooler before being supplied to the engine.

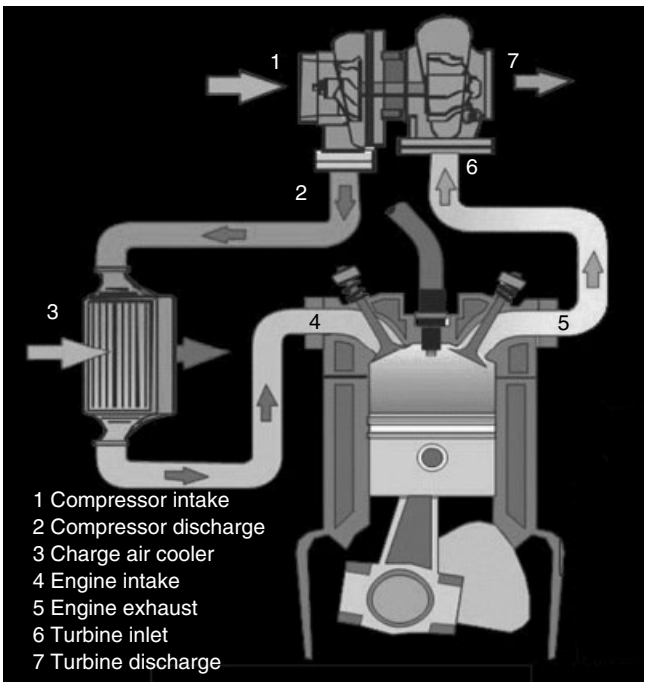


Figure 5. Pictorial representation of a turbocharged internal combustion engine. (Reproduced by permission of Honeywell Turbo Technologies.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

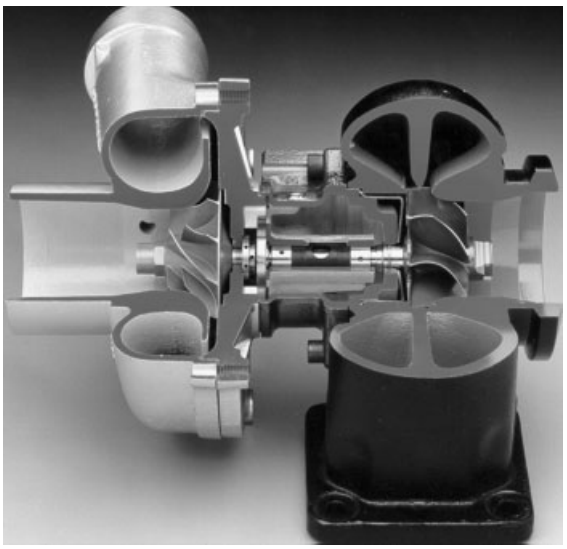


Figure 6. Cut-away illustration of a basic turbocharger. (Reproduced by permission of Honeywell Turbo Technologies.)

Figure 6 shows a cut-away view of a basic turbocharger to illustrate its design and componentry.

Figure 6 shows a turbine housing on the right-hand side (to be attached to the engine exhaust manifold) guiding the exhaust flow to the turbine wheel radially (design variations with axial flow and mixed flow are also used but are not discussed here). The center housing carries bearings (journal and thrust) for the shaft attached to the back of the turbine wheel and the compressor wheel on the left-hand side. The intake to the compressor is attached to the air filter and the discharge to the engine intake manifold usually with a charge cooler in between, as shown in Figure 5.

Power is developed by the turbine as a result of momentum transfer from the exhaust gases to the turbine wheel. Under steady-state conditions, this power is exactly equal to the power consumed by the compressor plus losses in the center housing. During acceleration, the turbine is required to develop greater power to accelerate the rotating mass to the new higher speed (and power) condition. The turbocharger has to "pick itself up by its bootstraps" during this process. A slight increase in fueling results in higher exhaust temperature, which slightly accelerates the turbocharger, resulting in increased airflow, which enables more fuel to be burnt resulting in higher speeds, and so on. The ability of the turbocharger to respond quickly to changes in demand is an important design consideration.

3.1 Compressor wheel and housing design

A primary requirement of the compressor wheel is to deliver the full range of flow that the engine requires. This implies not only mass flow rate but also the intake manifold pressure (density) required by the engine at all operating conditions. The efficiency with which the compressor is able to do this has an impact on the power demanded by it. This power is supplied by the turbine, which imposes a backpressure on the engine depending on its efficiency.

Compressor aero performance design objectives are met by the detailed design of blade shape at the inducer, in the body (including the hub), and at the exducer, as well as the aero design of the diffuser and housing. Centrifugal stresses are the most usual limiting factor, although air exit temperatures can impose a limit if pressures are high and the material softening temperature (as for aluminum) is reached. Conventional compressor wheel designs have a through hole at the center for the shaft to pass through. A nut at the free end secures the compressor wheel to the shaft.

Fatigue life is an important design consideration for both turbine and compressor wheels. High cycle fatigue failure is most usually associated with the high frequency vibrations of the blades. Design details associated with pressure fluctuations around the rotational path, blade geometry, and material properties are independent design parameters generally used to get satisfactory high cycle fatigue life. Low cycle fatigue failure is most usually associated with the centrifugal stress level and the frequency of changes in stress as in start/stop or slowing down/speeding up. Blade and hub geometry (e.g., blending radii at the blade/hub interface), material properties, and manufacturing processes all play a significant role in obtaining acceptable low cycle fatigue life.

Blade/hub geometry details can yield only small refinements in low cycle fatigue life because of the maturity of designs. Greater improvements are possible with major design changes[3] such as "boreless" compressor wheels (Figure 7), forged and machined wheels, counter gravity "pouring" of molten aluminum to minimize oxide inclusion (which can be crack initiation sites), and use of high strength material (e.g., titanium).

Prediction of fatigue life is a highly empirical process, with each manufacturer using a historic (and confidential) database of failure rate versus duty cycle. Design, material, and process changes are evaluated against this database to obtain fatigue life predictions such as shown in Figure 8 (Arnold, 2004).

3.2 Turbine wheel and housing design considerations

Design considerations include size/flow capacity, inertia/response, thermomechanical durability, aerodynamic

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

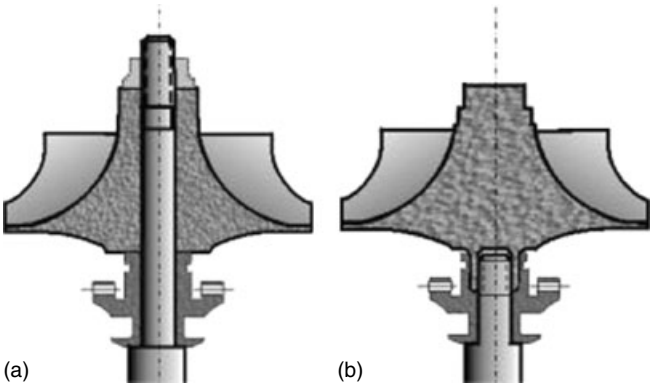


Figure 7. "Boreless" compressor wheel design (b) compared to conventional through-bore compressor wheel (a). (Reproduced by permission of Honeywell Turbo Technologies.)

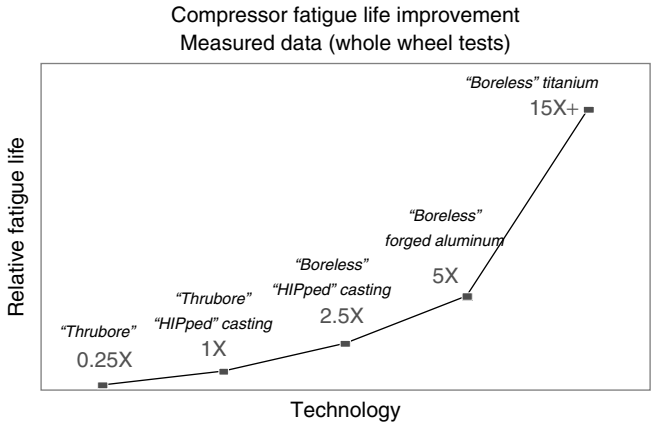


Figure 8. Improvements in fatigue life of a compressor wheel resulting from design, material, and process changes.

performance, and a host of other parameters, including cost. The turbine wheel has to be large enough (high inertia) to have enough flow capacity to handle the engine exhaust flow rate without choking and without imposing an unduly high backpressure on the engine. At the same time, it has to withstand high exhaust temperatures. Therefore, it is made of high nickel alloys, which tend to have high density, so much so that turbine wheels generally comprise more than 70% of the rotating inertia of a turbocharger. This sets up a design trade-off between the flow capacity and the response. The material, size (diameter), number of blades, blade thickness, back disk profile, and hub profile all require detailed design consideration of the trade-off between aerodynamic performance, inertia, thermomechanical integrity, and noise/vibration. This is done using highly specialized 3-D rotating flow and stress analysis computer codes in an iterative process for trade-off and design validation.

A primary requirement of turbine wheel design is that it should deliver the power required by the compressor without imposing high backpressure on the engine, that is, at a high efficiency. This is achieved by the geometry of the blades and details of the flow path. 3-D rotating flow computer codes are used to achieve aero design objectives.

The turbine housing has to withstand similar high temperatures as the turbine wheel. In addition, it has high temperature gradients because of direct exposure to the ambient on the outside. In general, the turbine housing has a center divider (shown in Figure 6) to preserve exhaust pulse energy of each cylinder. This sets up a design requirement to handle high thermal gradients/stresses in the center divider area. Sophisticated finite element analyses are done to choose the right material and the right detailed geometry to prevent turbine housing cracking.

Turbochargers rotate at very high speeds, with compressor wheel tip speeds reaching 550 m/s (somewhat lower for turbine tip speeds because turbine wheels are typically slightly smaller in diameter). A wheel burst and escape of high velocity pieces is a serious concern. Therefore, turbine (and compressor) housings have to be designed to contain wheel bursts, that is, not allow any broken parts to escape.

Centrifugal stresses are an important consideration in turbocharger wheel design because of its high rotational speed. A hole drilled through the center of a rotating disk doubles the centrifugal stresses. Turbine wheels are generally designed without a hole through the center. The shaft is welded to the back of the wheel (current practice is electron-beam welding versus earlier practice of friction welding). High speed balancing of the shaft–wheel assembly (as well as of the complete rotating assembly) is a highly specialized manufacturing process design and control technology. It has a healthy component of experience and empiricism and is usually treated as a "trade secret."

3.3 Center housing design considerations

The center housing carries the rotating assembly (turbine and compressor wheels and the shaft). Major design requirements for the center housing are to provide bearing capacities (rotational and thrust), damping to help reduce vibrations induced by high speeds and imbalance, cooling, and sealing. Each of these is considered in the paragraphs that follow.

Balancing of the rotating assembly is a critical design and manufacturing process requirement because of the high rotational speeds of the turbocharger and of the need to control shaft motion (displacement from the nominal

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

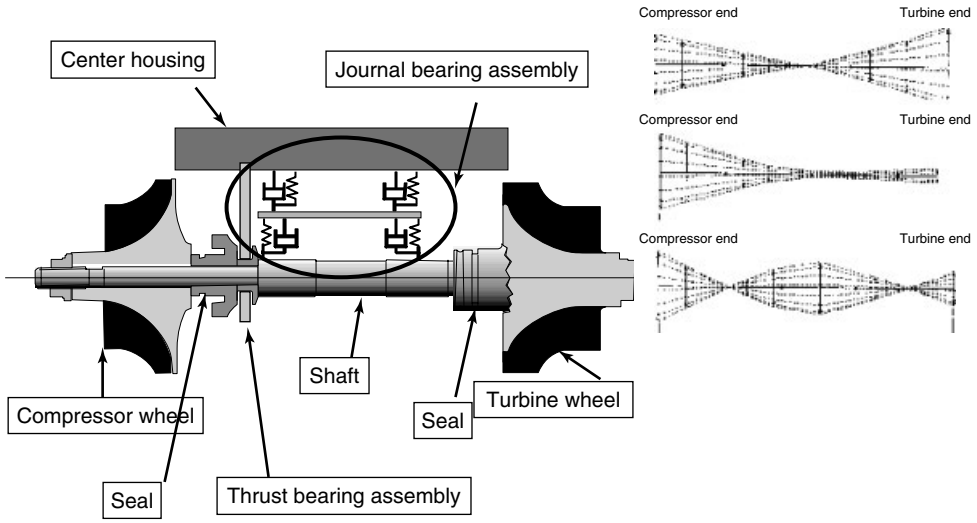


Figure 9. Modes of bending of the rotating assembly and the role of oil film damping in controlling shaft motion. (Reproduced by permission of Honeywell Turbo Technologies.)

geometric center of rotation). Shaft motion can set up local pressure cycles in the rotational path (causing high frequency vibrations in the blades) and cause noise and vibration and possible interference between the wheel and housing. The clearance between the wheel and housing represents an aerodynamic loss. Minimizing this loss requires that shaft motion be controlled at all rotational speeds. Residual imbalance (due to manufacturing tolerances, e.g., shaft–turbine wheel joining, compressor wheel assembly), flexibility of the shaft, and overhung loads of the compressor and turbine wheels all play a significant role in the satisfactory performance of the rotating assembly. Damping provided by bearings can mitigate shaft motion. High speed bearings have two hydrodynamic oil films: one between the shaft and the inner diameter (ID) of the bearing, and the other between the outer diameter (OD) of the bearing and the housing, as shown in Figure 9. Bearings can be free to rotate in the housing (fully floating) or pegged to not rotate but nevertheless gyrate at high speed to create a film between the OD and the housing. These oil films play a critical role in providing the necessary damping.

The net thrust load along the axis can be in either direction depending on the engine/turbo operating conditions. A thrust bearing and collar on the shaft and housing are designed to pick up oil, create a dynamic oil film, and provide enough load capacity to prevent metal-to-metal contact. The thrust bearing is placed near the compressor side because it is cooler. A careful look at Figures 6 and 9 show the placement of the thrust bearing and collar.

Proper sealing at the points where the shaft penetrates through the center housing is a significant design consideration to prevent oil from leaking out into the exhaust or intake flow path as well as to prevent gases from leaking into the center housing. A steel ring is usually used on the turbine side and the dynamic centrifugal separator (with or without deflectors) on the compressor side to provide an effective oil seal. The center housing runs at crank-case pressure (oil drains into the crank case), and aerodynamic design at the back of the wheels can be used to always maintain a slightly higher pressure on the air and exhaust sides to prevent oil from leaking into the gas path.

The oil drain from the center housing to the crank case is a sometimes neglected part of the design and can cause serious issues if oil does not drain freely.

Oil flow rate plays a critical role in carrying heat away from the center housing. Most often, a critical condition for cooling is not when the engine/turbo is running at high speeds and loads but during “hot shut down” when heat is still getting transferred from the hot turbine housing to the center housing and oil is not flowing. Sometimes, this can lead to coking of the oil in the bearings. Often, a heat shield is designed to present a resistance to the heat flow path from the turbine housing to the bearings. A careful look at Figure 2 on the right-hand side shows the placement of the heat shield.

Sometimes the center housing is cooled by providing coolant flow.

4 TURBOCHARGER PERFORMANCE

The parameters most usually used to measure the performance of a turbocharger are the following:

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- The ability of the compressor to supply air (flow rate and intake density/pressure) needed by the engine;
- Useful flow range of the compressor covering the full range of engine requirements;
- Compressor efficiency over the operating range;
- The capacity of the turbine to flow all exhaust without imposing a high backpressure on the engine;
- The efficiency of the turbine over the operating range (the losses in the center housing due to bearing and seal friction are conventionally included in turbine efficiency calculation);
- A host of thermomechanical attributes such as low noise and vibration, acceptable life (low cycle and high cycle fatigue included), no leaks, ability to withstand thermal stresses, fast response, and so on, in addition to cost targets.

Thermomechanical attributes have been discussed earlier. Performance attributes are discussed using example performance maps of compressors and turbines with engine maps superimposed on them. It is worth recalling the definition of some of the terms used in performance maps.

Total versus static temperature and pressure: high fluid velocities are involved in both compressors and turbines. Total and static properties are defined as

$$P_{\text{tot}} = P_{\text{stat}} + \rho \frac{V^2}{2}$$

$$T_{\text{tot}} = T_{\text{stat}} + \rho \frac{V^2}{2} C_{\text{p}}$$

Corrected flow: actual flow corrected to standard temperature and pressure conditions. The correction is derived using dimensional analysis but it is important to remember that corrected flow is *not* a dimensionless parameter.

$$\dot{\text{m}}_{\text{corrected}} = \frac{\dot{\text{m}}\sqrt{T_0/T_{\text{standard}}}}{p_0/p_{\text{standard}}}$$

It should also be noted that inlet temperature and pressure are relatively close to standard temperature and pressure for a compressor. Hence, the corrected flow is relatively close to the actual flow. However, for a turbine both inlet temperature (~800–1000°C) and pressure (2–4 bar) are considerably different from standard temperature and pressure. Hence, the corrected flow can be a factor of 2 or more different from actual flow.

The use of corrected flow enables comparison of different turbochargers.

Corrected speed: actual speed corrected to standard temperature and pressure conditions. The correction is derived using dimensional analysis but it is important to remember that corrected speed is *not* a dimensionless parameter.

$$N_{\text{corrected}} = \frac{N_{\text{actual}}}{(\sqrt{T_0/T_{\text{standard}}})}$$

Efficiency: on the compressor and turbine side takes into account aerodynamic losses in the wheel itself as well as in the entry and exit sections and housings. Mechanical losses are usually included with turbine efficiency calculation.

For the compressor side

$$\frac{\Delta E_{\text{ideal}}}{\Delta E_{\text{actual}}} = \frac{C_p \times (T_{06,\,\text{ideal}} - T_{00})}{C_p \times (T_{06} - T_{00})} = \frac{T_{00}(PR_{T-T}^{\gamma-1/\gamma} - 1)}{(T_{06} - T_{00})}$$

and for the turbine side

$$\frac{\Delta E\,(\text{Actual energy extracted})^*}{\Delta E_{\text{ideal}}(\text{Ideal energy extracted})} = \frac{C_{\text{p}} \times (T_{00} - T_{04,\text{actual}})}{C_{\text{p}} \times (T_{00} - T_{4,\text{ideal}})} = \frac{C_{\text{p}} \times (T_{00} - T_{04,\text{actual}})}{C_{\text{p}} \times T_{00} \times (1 - 1/PR_{T-S}^{\gamma-1/\gamma})}$$

Surge: the engine represents a restriction to compressor discharge. Under certain conditions, the compressor flow rate can be higher than the engine is able to accept. Under such conditions, flow separation can occur between the blades of the compressor wheel, resulting in reverse flow, noise, and possible damage to the wheel. When compressor discharge pressure is "relieved" by reverse flow, normal flow is reestablished, pressure builds up, and reverse flow can begin again. This repeated flow reversal phenomenon is termed *surge*.

Choke: can occur both on the compressor and turbine sides. Under both conditions, the maximum aerodynamic flow capacity has been reached and wheel is unable to handle (or supply) more flow regardless of the discharge condition.

4.1 Compressor performance

Figure 10 shows a typical heavy-duty diesel engine compressor map (Arnold, 2004). The compressor corrected flow is plotted on the *x*-axis and the compressor pressure ratio on the *y*-axis. Compressor characteristics are shown in dark gray, and engine operating lines superimposed on the same map are shown in light gray.

First, consider compressor speed lines. At a given (constant) compressor speed, as the flow increases (exit restriction is reduced), the pressure ratio drops. This continues as the exit flow restriction is reduced to a point where choking occurs and the flow can no longer increase

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

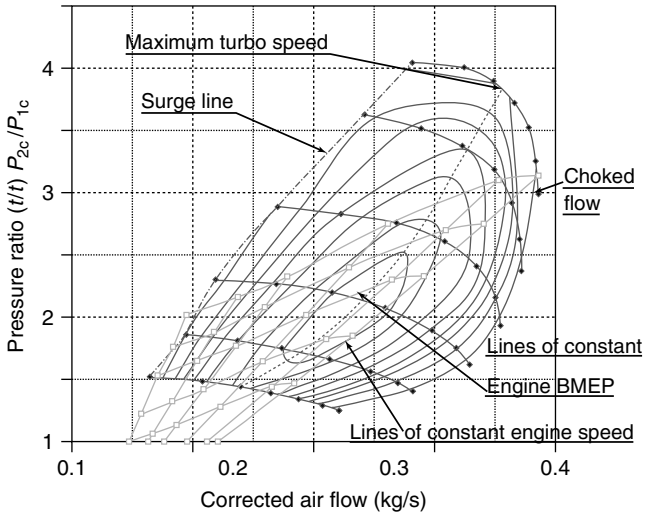


Figure 10. Representative heavy-duty diesel engine compressor map with engine operating conditions superimposed in light gray.

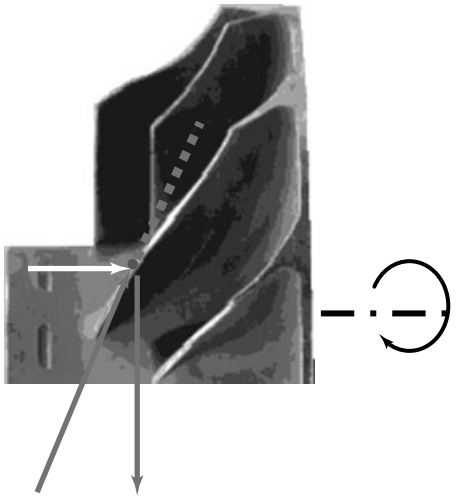


Figure 11. Illustration of inlet axial velocity, blade tip velocity, resultant velocity, angle of incidence, and blade angle at the compressor inlet.

(is no longer a function of discharge conditions). This is seen toward the right of the compressor map, where the pressure ratio drops nearly vertically without any further increase in the flow rate. Choke flow line is shown as the rightmost efficiency contour in Figure 10. At this line, the compressor efficiency is generally considered to be at its lowest acceptable value.

Next, consider the process as the exit restriction is increased, that is, the flow rate reduces and pressure ratio increases (along the same constant speed line). This can continue to a point where flow separation occurs in the compressor blades, resulting in reverse flow and surge, as described in the previous section. Surge line is shown as the leftmost efficiency contour in Figure 10. Operating the engine at conditions to the left of the surge line can result in damage to the turbocharger. Blade geometry is designed to give a wide usable flow range from surge to choke so that the engine operating map can be fitted within the usable flow range. The useful flow range can also be extended by providing a recirculation path or a "ported shroud" around the compressor intake.

Consider compressor efficiency. Compressor blade tip velocity is determined by the local radius and rotational speed of the compressor wheel. The incoming flow comes axially at compressor inlet (Figure 2). The angle of incidence is the resultant of the axial velocity and the blade tip velocity at that point. If the angle of incidence matches the blade angle, flow glides smoothly into the wheel and the efficiency is high. Figure 11 is an illustration of the velocity vectors at the inlet of the compressor wheel. Efficiency drops off on either side (flow rate, inlet velocity) because of the mismatch between the angle of incidence and the (fixed) blade angle. The high efficiency ridge, shown as the dotted line in Figure 10, represents the best match between the angle of incidence and the blade angle. On either side of the ridge, as the flow falls or increases, the resultant incidence angle is not exactly equal to the blade angle, resulting in flow vortices and aerodynamic losses. Therefore, efficiency drops off on both sides of the high efficiency ridge. Compressor wheel blade angle is designed to give the best compromise of efficiency over the engine operating map and the expected duty cycle.

Consider the peak pressure ratio capability of the compressor. This is determined largely by the speed at which the wheel can rotate without getting overstressed. Usually, the peak pressure ratio of the compressor is considerably higher than the operating pressure at full load so that the operating points of interest (depending on the duty cycle) may fall in the best efficiency region of the compressor.

In Figure 10, the engine operating lines are shown in light gray. The top light gray line is the full load line and is determined by the air–fuel ratio the engine requires (a requirement dictated by the combustion system and engine efficiency and emissions targets). It is seen from the figure that the full load–full speed point is at choke and the full load–idle speed point is at surge, using up the full useful flow range of the compressor without any margin of safety. The compressor designer has some flexibility in increasing the useful range at the expense of peak pressure ratio capability by changing the geometry of compressor

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

blades (the backward curvature). Below the full load line, operating lines at part loads are shown. Constant engine speed lines are also shown: rated speed line to the right, and idle speed line to the left.

Thus, compressor performance characteristics, that is, the flow capacity, flow range, efficiency, pressure ratio, and wheel speed, are all shown on the compressor map. The process of matching (discussed in Section 5) is used to evaluate the performance of the compressor on the engine.

4.2 Turbine performance

Figure 12 shows a turbine map suitable for the same heavy-duty diesel engine as used for the compressor map (Arnold, 2004).

Consider turbine flow characteristics. The top set of curves show the flow rate versus expansion ratio at various constant turbine speeds. It is seen that all speed lines collapse into essentially a single flow versus expansion ratio curve. It is also seen that at an expansion ratio of ~2.8 turbine flow capacity is reached as a result of choking. For higher flow rates, a larger diameter turbine (or a turbine with radically different flow characteristics) would be required.

The bottom set of curves show the efficiency versus expansion ratio for each speed. Only a small range of expansion ratios is shown. The efficiency curves as shown are generally referred to as the *eye brow curves*. A pictorial representation (dashed line) of efficiency over a wider range of expansion ratio (at a constant speed) is shown. In most datasets, only a narrow range (pictorially an "eye brow") is shown because usually the engine flow and turbine speed are self-adjusting so that the operating point falls near peak efficiency for that speed. Efficiency drops off on either side of the peak because the incidence angle changes, resulting in aerodynamic losses.

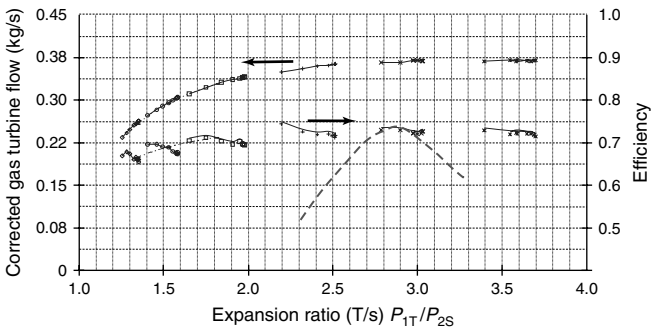


Figure 12. Illustrative turbine operating characteristics of a heavy-duty diesel engine turbocharger.

The calculation of turbine power is discussed in Section 5.

Turbine peak efficiency is seen to be around 72%. It is usually less for smaller turbines because the clearance area is a larger proportion of the total flow area.

5 ENGINE/TURBOCHARGER MATCHING BASICS

The purpose of "matching" is to select compressor and turbine size and performance characteristics that best match the needs and performance targets of the engine/vehicle/equipment. This is necessarily an iterative process, with the number of iterations considerably reduced by appropriate simulations. Further, there is some leeway in tweaking some design parameters of the turbocharger (such as *A*/*R* ratio, trim) to fine-tune the match.

Engine simulation, design targets, and candidate compressor and turbine maps are used as the starting point. It is best to start with engine airflow requirements.

Engine airflow requirement is calculated using power rating, brake specific fuel consumption (BSFC) target, and target air–fuel ratio (the effect of exhaust gas recirculation (EGR) is discussed later) set by the expected combustion system performance and emissions targets. Thus,

$$\text{Fuel rate (g/h)} = \text{BSFC (g/kWh)} \times \text{Power (kW)}$$

$$\text{Air mass flow rate (g/h)} = \text{Fuel rate (g/h)} \times \text{AFR}$$

Volume flow rate is set by engine displacement, speed, and target volumetric efficiency. Thus,

$$\text{Air volume flow rate (L/h)} = \eta_{\text{vol}} \times V_{\text{displ}}\ \text{(L)} \times \frac{n}{2}\ \text{(L/min)} \times 60\ \text{(min/h)}$$

The mass flow rate and the volume flow rate readily give the intake air density and, together with the intake manifold temperature (dependent on charge cooler, EGR rate, EGR cooler), the required intake manifold pressure can be calculated. Thus,

$$\text{Air density } (\rho) = \frac{\text{Mass flow rate}}{\text{Volume flow rate}}$$

$$\text{Intake manifold pressure (kPa)} = \rho RT$$

Knowing the engine airflow requirement and the compressor pressure ratio required to enable this airflow, compressor power requirement can be calculated. Thus,

$$\text{Power (kW)} = \frac{[m_{\text{a}} \times C_{\text{p}} \times T_{int} \times (P_{\text{out}}/P_{\text{in}})^{(\gamma-1)/\gamma}]}{\eta_{\text{c}}}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Note that the compressor efficiency is needed to calculate the power. The efficiency can be obtained by plotting the operating point on the compressor map and reading the corresponding efficiency. This process also yields the compressor speed.

The turbine has to supply the power required by the compressor. The power developed by the turbine is given by

$$\text{Power (kW)} = \eta_{\text{tm}} \times m_{(\text{a+f})} \times C_{\text{p}} \times T_{\text{exh}} \times \left[1 - \left(\frac{P_{\text{out}}}{P_{\text{in}}}\right)^{(\gamma-1)/\gamma}\right]$$

In the above equation, T_{exh} is the engine exhaust temperature, which is obtained by simulation or estimated using similarity with other engines. P_{out} is turbine outlet pressure and set by the exhaust system. There are two unknowns in this equation: η_{tm} and P_{in}, that is, turbine efficiency and engine exhaust pressure. An iterative solution using turbine performance map and turbine speed (equal to compressor speed because they are on the same shaft) readily yields these two values.

The engine operating point can be plotted on the compressor map and on the turbine map. It is not enough to plot just one operating point. This calculation has to be repeated preferably over a range of load and speed conditions and checked to see if the operating points fall within the operating range of the compressor, with efficiency requirements in mind. Figure 10 is such a representative plot. Note that the best compressor efficiency is around 50% engine speed and 50% load. Perhaps in this match, a case can be made that the engine duty cycle is such that the engine seldom runs at full load–full speed or at full load–idle speed. This match has no "surge margin" or "altitude capability." At higher altitudes, in order to maintain the air–fuel ratio, the pressure ratio requirement increases (because of decreasing intake air density), pushing the engine operating points into surge beyond the usable flow range of the compressor. If that is a major design requirement, a different compressor will have to be selected for this engine. Turbocharger manufacturers usually have a wide range of turbochargers with different performance characteristics available to choose from (and match).

Engine flow is plotted on the turbine map in a similar manner and shown in Figure 13 (Arnold, 2004). In this figure, the turbine flow curves for all turbine speeds are collapsed into a single curve and turbine efficiency lines are not shown. The engine flow operating line at full load is also shown.

A significant feature to note is the difference between the turbine wheel flow characteristic and the engine flow. The turbine wheel has to be sized to handle the flow at full load–rated speed. This implies that the turbine is too large for lower speed conditions and turbocharger inertia has an impact on turbo response. Wastegate turbines, variable geometry turbines, and multistage turbocharging to mitigate this problem are discussed in the following sections.

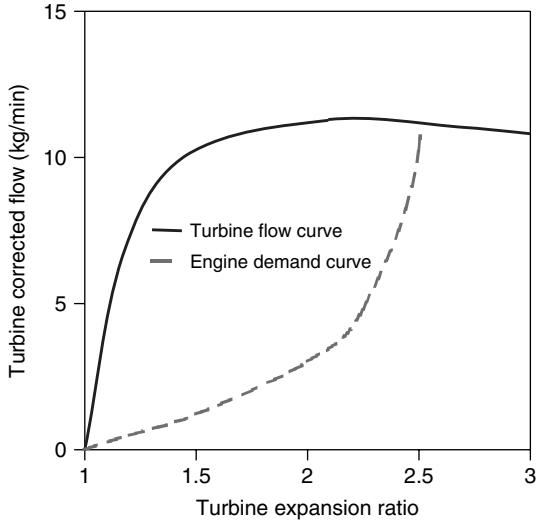


Figure 13. Illustrative simplified flow map of a heavy-duty diesel engine turbine. Engine full-load operating line is also shown.

6 ADVANCED ENGINE REQUIREMENTS AND TURBO TECHNOLOGIES

Some of the requirements and technologies described below can no longer be considered "advanced" but are grouped together for convenience.

6.1 Wastegate turbocharger

In the previous sections, we have considered the basic free-floating turbocharger. As discussed in Section 5, the basic match results in a turbine large enough for full flow conditions but too large at low engine flow conditions. This can result in a response issue.

To get around this issue, the turbine is intentionally designed to be too small for full flow conditions, and under such conditions a "wastegate" valve is opened to (partially) bypass the turbine. This prevents overspeeding the turbo and/or inducing too high a backpressure on the engine. It also "wastes" the exhaust energy. But this might be a net benefit under duty cycle conditions that require frequent accelerations and infrequent operation at full load–full speed (as in a passenger car). A picture of a cut-away turbine housing focusing on the wastegate valve is shown

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

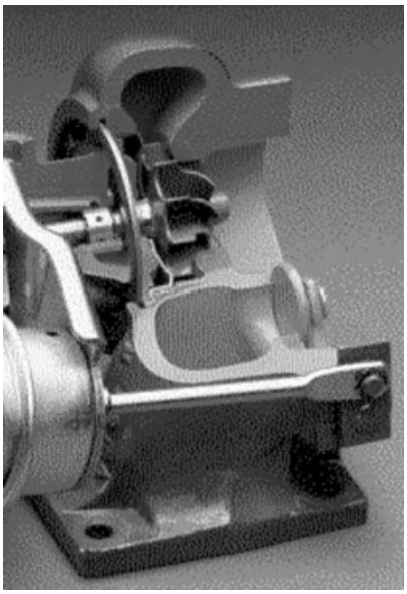


Figure 14. Illustration of a wastegate turbocharger. (Reproduced by permission of Honeywell Turbo Technologies.)

in Figure 14. In this rendering, the wastegate valve is actuated with a pneumatic actuator using compressor outlet pressure.

6.2 Variable-geometry turbocharging

Instead of using a small turbine and wastegate, a full-sized turbine wheel is retained and the issue of response is addressed by using a variable-geometry inlet in the turbine. Figure 15 is a collection of photographs of cutaways of a light-duty diesel engine turbocharger showing how one specific variable-geometry mechanism works (Petitjean *et al*., 2004).

Under low flow conditions, the vanes are closed (top right inset) to increase the velocity of exhaust gases coming at the turbine wheel. This keeps the turbo speed up and helps with low engine speed torque and response. Under high flow conditions, the vanes are open (bottom right inset). The turbine wheel size is adequate to handle full flow, thus the backpressure on the engine is not excessive (nor is the flow bypassed through a wastegate, saving fuel). The actuator shown in the main picture is a pneumatic actuator that can be intelligently controlled. An electric actuator shown in the inset enables greater flexibility.

Heavy-duty diesel engines and the forces involved, together with the possible use of the variable-geometry mechanism in synergy with engine braking, require a design capable of withstanding higher pressures and higher pressure fluctuations. Figure 16 shows a heavy-duty diesel engine variable-geometry turbo mechanism where vanes are supported on both sides rather than cantilevered as shown in Figure 15. Details of the moving vane mechanism are shown in the inset to the left. Also, the pneumatic actuator works using the vehicle's pneumatic pressure system because the force required to move the vanes is high. Hydraulic actuation using engine oil is an alternative.

Figure 17 shows the performance characteristics of a variable-geometry turbine. Turbine flow characteristic at vane fully open position is shown as the top dark gray curve. Each successive dark gray curve under it is the turbine flow characteristic with the vanes progressively

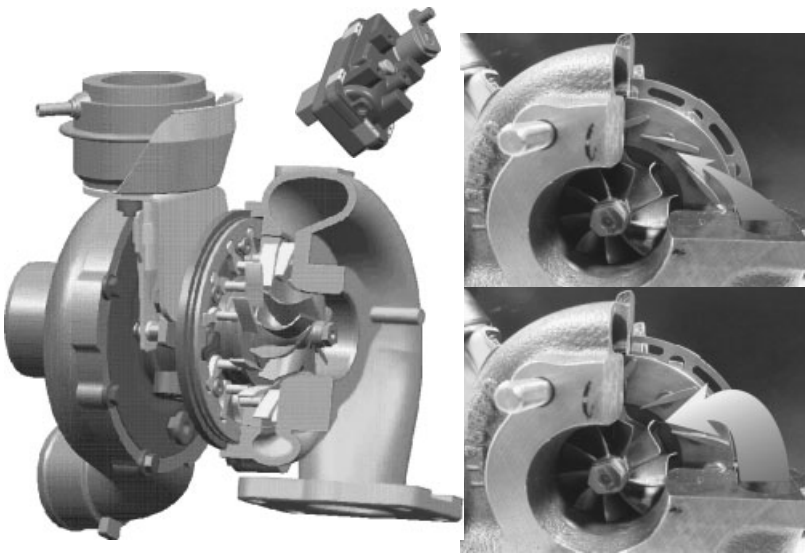


Figure 15. Illustrative variable-geometry turbocharging for light-duty diesel engine. (Reproduced by permission of Honeywell Turbo Technologies.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

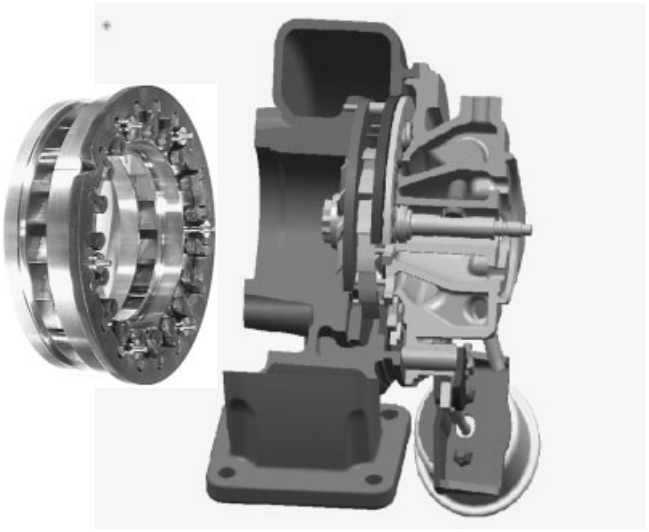


Figure 16. Illustrative view of a heavy-duty diesel engine variable-geometry turbocharger. (Reproduced by permission of Honeywell Turbo Technologies.)

closed. It is seen that, by appropriately choosing the vane position for the load (and speed) condition of the engine, a better match between the turbine flow characteristics and the engine flow requirements can be made.

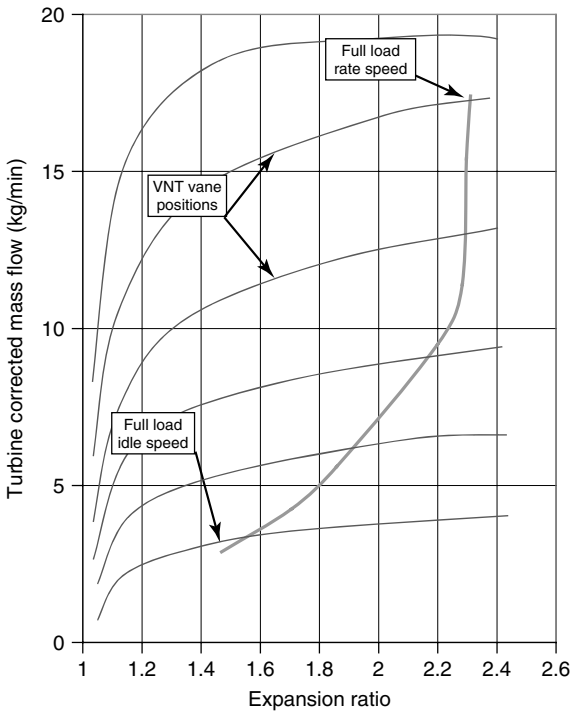


Figure 17. Turbine flow characteristics with various vane positions. Also shown is the full-load engine flow curve (light gray).

6.3 Exhaust gas recirculation (diesel engines)

EGR is necessary to control the engine nitric oxide (NO*x*) emissions. EGR presents problems both on the compressor and turbine sides. High pressure (short) loop (HPL)-EGR is discussed first. A schematic of a representative HPL-EGR arrangement is shown in Figure 18 (Arnold, 2004).

In a well-designed and well-matched turbocharger, the exhaust pressure is lower than the intake pressure over a large part of the engine load–speed range, giving a net positive pumping work contribution. Thus, driving the EGR involves increasing the backpressure on the engine and undoing the pumping work benefit of turbocharging. At the same time, in order to maintain the same combustion system performance (particulate emissions), the oxygen/fuel ratio has to be maintained at or near the pre-EGR levels. The inert constituents of the EGR (nitrogen, carbon dioxide, water vapor) proportionately use up space in the cylinder which fresh air would have occupied. Therefore, the intake manifold pressure (density) has to be increased to supply the needed air. In addition, because the EGR temperatures are high, the pressure has to be further increased to compensate for the lower density.

A variable-geometry turbine can be used to increase the backpressure to drive EGR and at the same time increase turbo speed and drive more fresh airflow. On the compressor side, the increase in intake manifold pressure (therefore compressor pressure ratio) tends to drive the engine operating curve more toward the surge line. The compressor has to be rematched to avoid surge. Another possible problem could be (if pressure ratio increases too high) that the compressor outlet temperatures may be high enough to soften aluminum. In some applications, titanium compressor wheels are being used, as well a precooler upstream of the aluminum charge air cooler. Figure 18 illustrates this.

A low pressure (long) loop (LPL)-EGR system presents a different set of challenges. EGR is usually drawn from downstream of the diesel particulate filter (DPF) and fed to the intake of the compressor. In addition to high pressure requirements, this increases the temperature even more because the compressor inlet temperature is increased, and then goes through a compression process further increasing its temperature. In spite of the efficiency of the DPF, there is enough particulate level in the EGR stream (and it accumulates on the walls over time), so that this presents a severe fouling challenge both to the compressor wheel and to the charge air cooler. Further, even with low sulfur fuel, acid condensation and corrosion remain serious issues to be considered. Compressor wheels coated with a corrosion-resistant material are usually considered.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

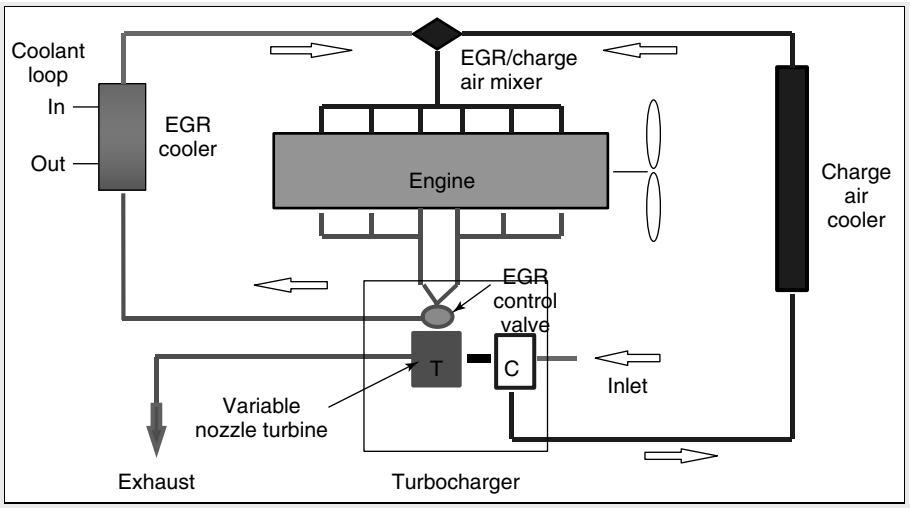


Figure 18. Illustration of a high pressure loop EGR system with variable geometry turbocharging.

6.4 Multistage and sequential turbocharging

High EGR rates increase the intake manifold pressure needed, as discussed earlier. Figure 19 shows the average peak compressor pressure ratios used between 1980 and 2004 in heavy-duty diesel engines in the United States (Arnold, 2004). The gradual, linear increase in compressor pressure ratio between 1980 and 2002 is a reflection of engine development and increase in power rating of the same basic engine (more air, more fuel, and more power). The step increase in 2002 reflects the implementation of EGR systems to meet the stringent (at the time) new NO*x* emissions regulations.

While the illustration in Figure 19 uses a heavy-duty diesel engine database, the same trend applies to light-duty diesel engines. Even as the need for compressor ratio is increasing, the sensitivity to response from the air handling system is also increasing. Multistage turbocharging is being implemented to meet these needs. As the name implies, two turbochargers are used. In the series and series-sequential configuration, the high pressure turbocharger is smaller in size (because it handles less volume flow because of higher pressures). This makes the system more responsive. Also, the multiplication effect of the compression ratio in the two stages gives the possibility of very high overall compressor pressure ratios without getting into surge.

Figure 20 shows a series turbocharging system (Mattes, 2007) with bypass valves on both turbines and on one

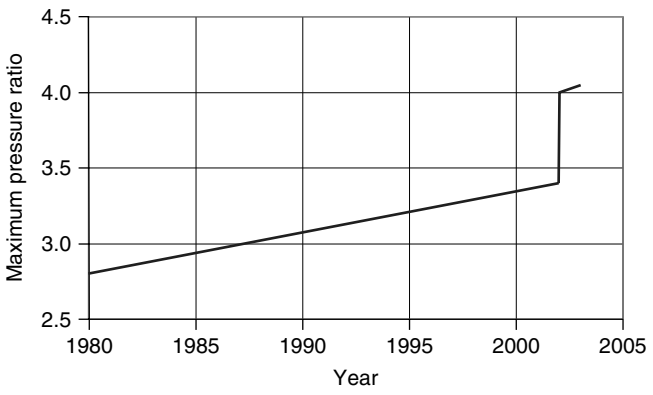


Figure 19. Peak compressor pressure ratios implemented in on-highway US heavy duty diesel engines.

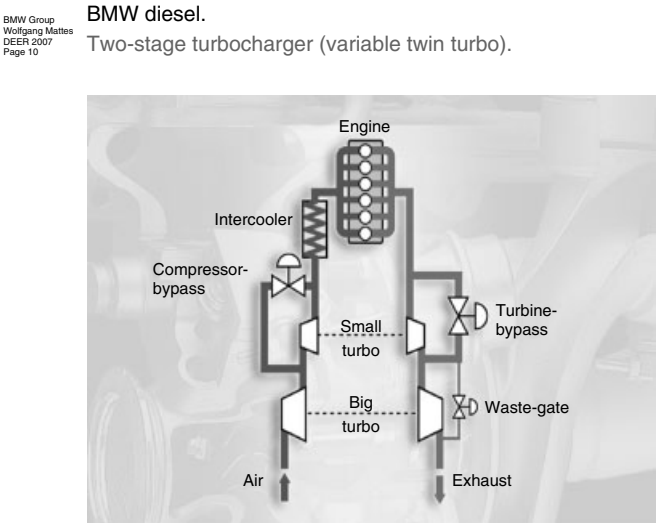


Figure 20. Multistage series turbocharging with bypass valves.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

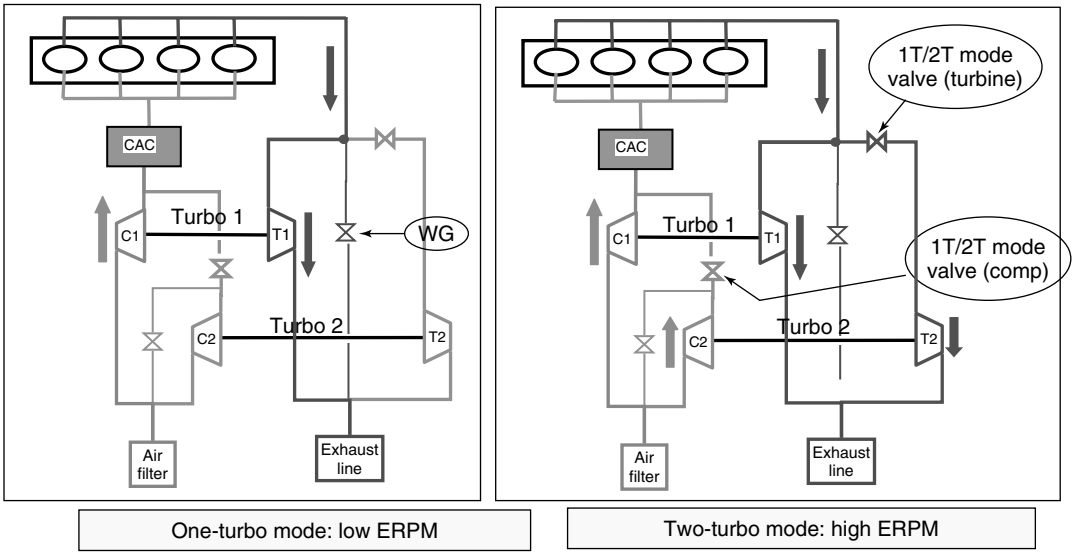


Figure 21. Parallel sequential turbocharging.

compressor, which enables the turbochargers to be used sequentially or together depending upon the load–speed conditions.

Figure 21 shows a parallel-sequential configuration (Portalier *et al*., 2006). In this version, both turbos (two-turbo mode) are used under high flow conditions and a single turbo (one-turbo mode) is used under low flow conditions. Both turbos may be (but do not have to be) of the same size. Figure 22 shows a picture of the production configuration.[4]

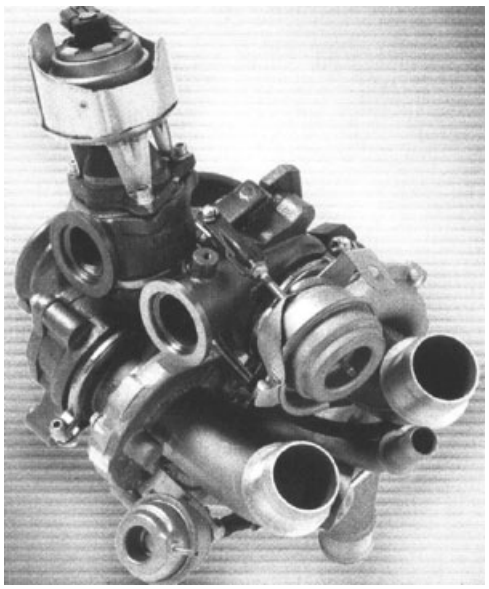


Figure 22. Production implementation of parallel sequential turbocharging.

6.5 Single-sequential turbocharging

Figures 23 illustrates a rather innovative implementation of series multistage compression on a compact single axis design (Arnold, 2007).

Back-to-back compressor wheels are designed, and the discharge from the front compressor wheel is supplied to the rear wheel, resulting in the second stage of compression. Figure 24 shows the impact of this design change on the compressor flow map. It is seen that, in a compact, single-axis machine, very high compressor ratios and a wide useful flow range is obtained.

A slightly different implementation of the same basic idea (back-to-back compressor wheels) is shown in Figure 25. In this implementation, the two compressor wheels are used in a parallel flow mode. This enables

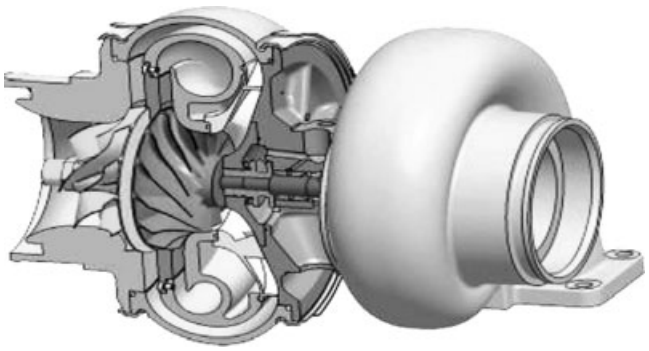


Figure 23. Series multistage compression designed on a single shaft.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

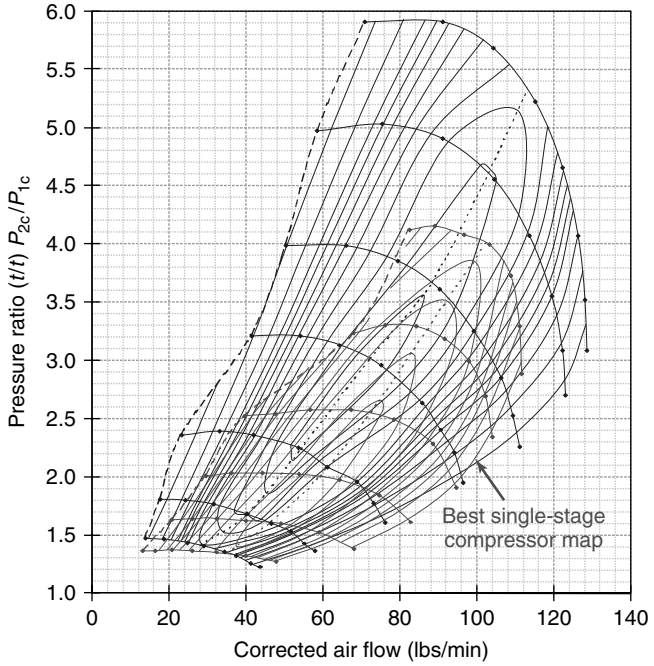


Figure 24. Compressor performance map with multistage compression designed on a single shaft compared to the best single-stage compressor map.

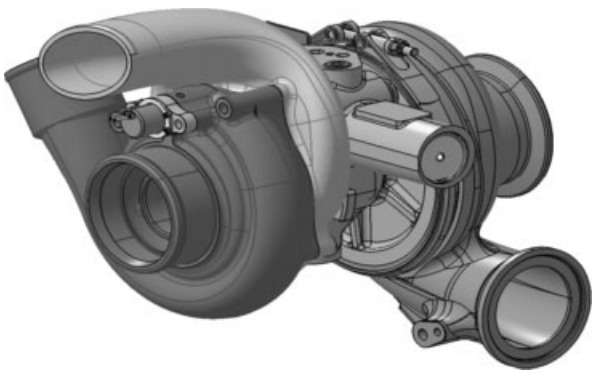


Figure 25. Parallel flow compressor packaged on a single shaft.

a reduction in the compressor wheel and turbine wheel diameters for the same airflow, thereby greatly reducing inertia and improving response.

Note that there are two fresh air inlets, one for the front compressor and one for the rear. Both compressors discharge into the same outlet. By suitably matching the aero characteristics of the two compressors, both high pressure ratio and wide flow range can be obtained, as shown in Figure 26.

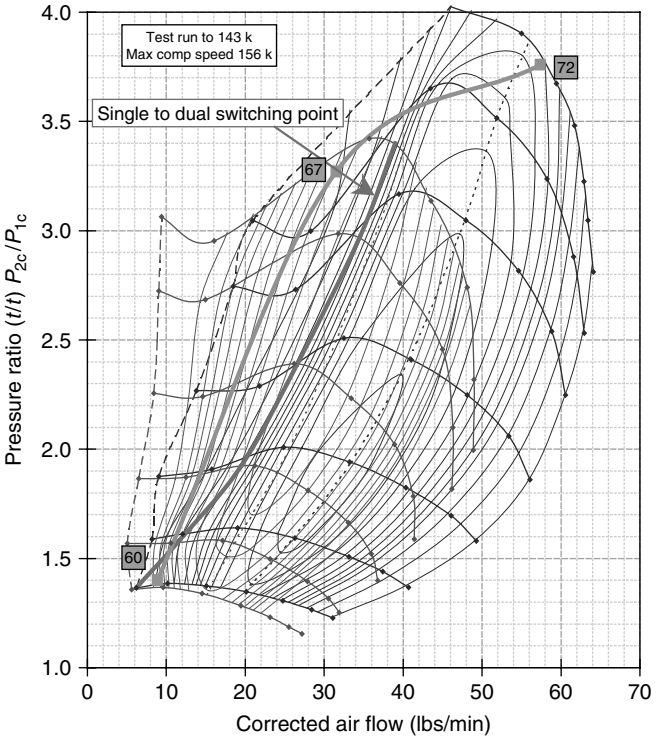


Figure 26. Performance map of parallel flow compressor designed on a single shaft.

6.6 Gasoline engine turbocharging

The benefits of gasoline engine downsizing and turbocharging (as also of turbocharged diesel engines) have already been discussed in Section 2. The discussion here is limited to the turbocharger itself. There are several major issues with gasoline turbocharging.

- The need for a wide flow range because of the typically much wider speed range of gasoline engines;
- Much higher (than diesel engines) exhaust temperatures;
- Heightened sensitivity to response due to passenger car applications.

Fortunately, there are mitigating factors also. First, the pressure ratios required are not very high, usually limited by engine knock considerations and, second, the airflow rate required for the same amount of fuel is not very high because of stoichiometric (rather than lean) mixture ratios.

In the 1980s, considerable effort was put into developing ceramic turbine wheels, which addressed both temperature and response issues. These were put into production for a limited time but ultimately did not achieve major applications because of "foreign object damage." The flow range issue was addressed by using wastegate turbochargers.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

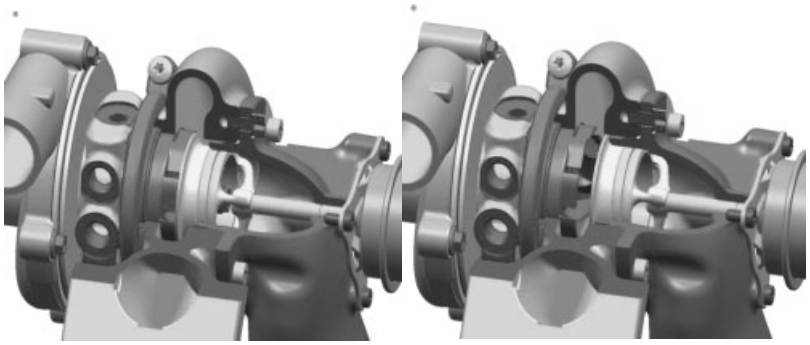


Figure 27. Illustrative variable-geometry implementation for gasoline engine turbocharging. (Reproduced by permission of Honeywell Turbo Technologies.)

More recent developments include the design of variable-geometry turbocharging with sliding vanes rather than moving vanes to handle higher thermal distortions associated with higher exhaust temperatures (Petitjean *et al*., 2004).

Figure 27 shows one such design, with the two sides illustrating low flow and high flow positions. Note that all the vanes are fixed on the disk behind the turbine wheel. The slider slides in and out, controlling the flow area into the turbine (and hence the velocity of the incoming exhaust).

Figure 28 shows a moving-vane design adapted especially for high temperature gasoline engine applications with the use of high temperature austenitic steel. Note the use of additional disks on the pivoted and free sides of the moving vanes, all mounted on the same package, to keep the effects of thermal distortion under control.[5]

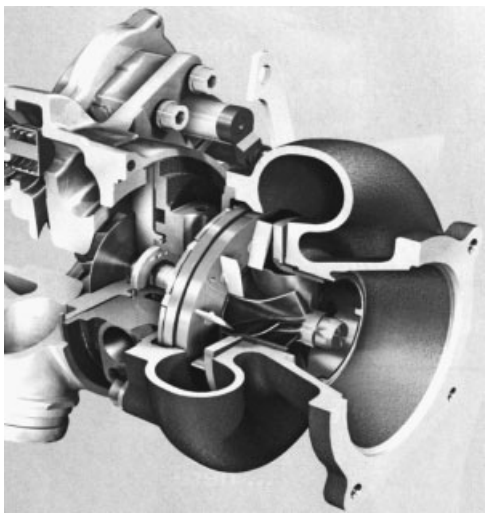


Figure 28. Moving-vane variable-geometry mechanism implemented on a production gasoline engine.

7 SUMMARY

The first patent for a turbocharger design was issued in 1905, and turbocharging has seen significant and continuous improvements since then. The first application of turbocharging was for reciprocating gasoline aircraft engines and has proceeded to on highway and off-highway heavy-duty diesel engines, light-duty passenger car diesel engines, and now to passenger car gasoline engines, thus covering all reciprocating engines. Turbocharging is increasingly recognized as a critical part of engine/vehicle performance, emissions, and fuel economy.

In heavy-duty diesel engines, turbocharging has increased power density, enabled emissions reduction through charge-cooling and EGR, and improved fuel economy. In light-duty diesel engines, turbocharging has given similar improvements and 30–50% better fuel economy compared to equivalent gasoline engines. Turbocharging is so significant to diesel engines that there are no important diesel engines without turbocharging.

In light-duty passenger-vehicle gasoline engines, turbocharging has enabled engine downsizing and reduced throttling over the drive cycle. This has improved fuel economy by about 20% while enabling 40–50% downsizing of engines.

Fatigue life, size/weight/inertia, high temperature capabilities, noise, and oil leakage are important mechanical design considerations. Performance/aerodynamic design considerations include efficiency, wide flow range, and flow and pressure capability. Material, manufacturing process, and detailed blade and hub design all play a significant role in determining both the performance and mechanical design capabilities of a turbocharger. Detailed geometric design features of turbine and compressor wheels are rapidly reaching a highly mature stage. Further improvements are expected at a system level with variable-geometry and multistage turbocharging.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Turbochargers have to be carefully matched to meet specific engine requirements such that the engine operating lines lie well within the useful flow range of the compressor and the turbine is able to supply the power needed by the turbocharger as well as handle engine exhaust flow without imposing a high backpressure on the engine. EGR and DPFs impose additional requirements on turbocharger matching and design.

Variable-geometry turbochargers are available for light- and heavy-duty diesel engines and gasoline engines. Variable-geometry turbocharging improves response and fuel economy.

Increasingly, the stringent requirements imposed by the need for high rates of EGR, faster response, and high lowspeed torque drive the need for multi-stage turbocharging. Parallel, series, and sequential turbocharging, as well as multistage turbocharging packaged on a single shaft giving a compact design, have been designed, developed, and brought to production.

Turbocharging has become an increasingly significant and integral part of engine design for all classes of diesel and gasoline engines, and also for alternate fuel engines.

ENDNOTES

1. Automotive Engineer, July 15, 2010.
2. Cummins Turbo Technologies, http://cumminsengines.com/cummins-indy-racing-heritage.
3. Honeywell Turbo Technologies—training material.
4. SAE Automotive Engineering International, November 2006.
5. SAE Automotive Engineering International, October 2006.

ACKNOWLEDGMENT

The author wishes to acknowledge the encouragement and help offered by Honeywell Turbo Technologies and their permission in using several key figures.

REFERENCES

Arnold S. (2004) Turbocharging technologies to meet critical performance demands of ultra-low emissions diesel engines. SAE Paper 2004-01-1359, 19th Cliff Garrett Turbomachinery Award Lecture.

Arnold S. (2007) Single sequential turbocharger: a new boosting concept for ultra-low emission diesel engines. SAE 2008-01-0298.

Mattes W. (2007) *DEER Conference 2007*.

Petitjean D., Bernardini L., Middlemass C., and Shahed S.M. (2004) Advanced gasoline engine turbocharging technology for fuel economy improvements. SAE 2004-01-0988.

Portalier, J., Blanc, J.C., Garnier, F., *et al*. (2006) Twin Turbo Boosting System Design for the New Generation of PSA 2,2 Liter HDI Diesel Engines. *Proceedings of Thiesel Conference 2006*.

Walzer P. (2001) Future powerplants for cars. SAE 2001-01-3192.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto069
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Bearings and Bearing Design for Transmissions

Reinhart Malik, Ernst Masur, and Andreas Schick
Schaeffler AG, Herzogenaurach, Germany

1 Introduction 1
2 From the Component to the System: Bearing Applications in Transmissions 2
3 Bearing Design for Transmissions 9
4 Bearing Damage and Preventing Damage 12
5 Summary 15
References 15
Further Reading 15

1 INTRODUCTION

The primary aim in modern transmission manufacturing is to achieve greater efficiency. This applies both to the considerable influence of the transmission on the overall energy efficiency of a vehicle and to the costs. The transmission is one of the most valuable systems in a vehicle. Over the last few years, vehicle and transmission manufacturers have responded to the need for efficiency with, among other things, strong diversification in the design variants described in detail in the previous chapters. The torque density of vehicle transmissions—that is, the maximum torque transferred in relation to the overall installed size—has increased significantly as modern engines offer more torque at low speeds (thanks also to the trend for turbocharging), whereas the mounting space available for the transmission is shrinking.

However, what all transmissions have in common, despite the different design variants and the individual models, is that the engine speed is transmitted to the output speed via rotating shafts. Engine development gives rise to ever-increasing torques that lead to high loads on the moving parts. Therefore, the bearings installed in the transmission have a decisive influence on the service life of the transmission.

Owing to the conditions present in the transmission, bearings are to some extent subject to specific loads that do not occur in this form in other applications. For example, frequently changing axial and radial loads occur simultaneously in the transmission, and dynamic loads prevail. This applies even for normal operation due to the forces that arise on the gear wheel mesh. Shaft deflection also occurs, as well as impact loads due to uneven roads.

The trend for light construction means that transmission housings are often made from light metal (aluminum and, increasingly, magnesium). These materials demonstrate relatively high coefficients of thermal expansion giving thermal distortion that has to be compensated. The temperature range within which a vehicle transmission has to operate without any problems ranges from -40 to $+60\,^{\circ}\mathrm{C}$ external temperature, whereby the temperature of the transmission oil can be considerably higher.

The transmission oil itself is usually used to lubricate the bearings. As the transmission oil is no longer changed over the entire life of passenger cars today, oil contamination and oil ageing represent significant challenges for bearing developers. The fact that transmission oils with low viscosity are increasingly being used only compounds the issue. Despite this, the development of a load-bearing lubricating film must be guaranteed at all usage temperatures.

Due to the friction that arises, the bearing points also contribute directly to the overall efficiency of the transmission and thus to the fuel consumption of the vehicle. A modern rolling bearing may reach an efficiency level of over 99%, but the multiplication of many small

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

losses at individual bearing points results in a measurable influence on the CO_2 emissions of the overall vehicle.

2 FROM THE COMPONENT TO THE SYSTEM: BEARING APPLICATIONS IN TRANSMISSIONS

There is not and cannot be one typical transmission bearing. In industrial practice, every bearing planned for transmissions produced in high quantities is designed specifically for the individual use case. The best design solution is selected for each bearing point, meaning that one individual transmission can contain more than two dozen different types of bearing construction (Figure 1).

The bearings installed in a transmission are primarily rolling bearings in which a rolling element runs between an inner and an outer ring. The main components of a rolling bearing (Figure 2) are:

- The *rolling elements*, generally designed as balls or rollers.

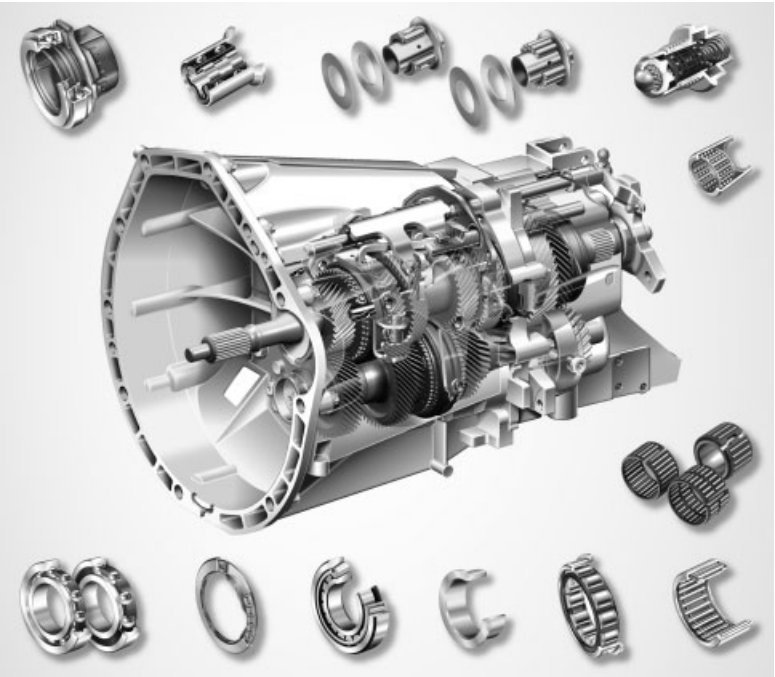


Figure 1. Types of bearing construction in a typical manual transmission. (Reproduced with permission from Schaeffler Group.)

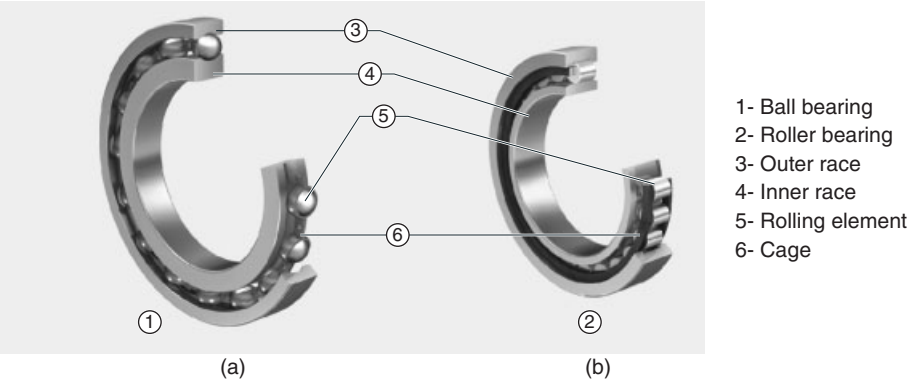


Figure 2. Main components of rolling bearings as ball bearings (a) or roller bearings (b). (Reproduced with permission from Schaeffler Group.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- The *tracks* on which the rolling element runs. Generally, the tracks consist of an inner and an outer ring. However, in some cases, the component or shaft to be supported is itself used as the track.
- The *cage*, used to protect the rolling elements against mutual contact and slipping.
- The *seal*, if the rolling bearing is to be protected against penetrating media.
- The *lubricant*, if the rolling bearing is lubricated with grease. In vehicle transmissions, lubrication is provided almost exclusively by the transmission oil.

The diversity of design variants for rolling bearings reflects the multitude of applications in transmissions. These applications differ particularly with regard to the possible occurrence of axial and radial forces. As it is the design of the rolling element that varies most strongly, it makes sense to look at the bearings based on the respective design variants and, in this context, to consider the respective transmission applications. Figure 3 shows an overview of the types of rolling elements used most frequently in transmissions.

2.1 Ball bearings

As the name indicates, ball bearings use balls as rolling elements. Ball bearings were used in automotive engineering long before the car was invented, for example, in bicycles. Industrial mass production began with the invention of the ball mill by Friedrich Fischer in 1883. Fischer, who founded the company FAG Kugelfischer, designed a grinding machine that, for the first time, enabled a large number of steel balls to be produced with identical dimensions. Transmission manufacturing today uses primarily deep groove ball bearings and angular contact ball bearings; they differ in particular in the arrangement of the running surfaces.

2.1.1 Deep groove ball bearings

The deep groove ball bearing is the classic ball bearing *per se*. Here, the balls run in a groove in the outer ring and one in the inner ring. Deep groove ball bearings are suitable for applications in which primarily radial forces occur. The absorption of axial forces is limited to approximately 10% of the maximum radial force absorption (Figure 4). If the axial forces are too high, this reduces the life of the bearing considerably.

Deep groove ball bearings are suitable as locating bearings for guiding the shaft, for example. Due to their small contact surfaces, they produce relatively low friction, but this very fact also limits their absolute load capacity. This means that deep groove ball bearings are primarily used as the main bearing in smaller transmissions with input torques of up to 250 Nm. One factor that determines the average dynamic load capacity and thus the service life of a bearing is the material used. The standard steel used is 100Cr6, but higher quality heat-treated materials are also widely used.

For transmission applications, deep groove ball bearings are being increasingly designed with an integrated elastomer seal for the cage in order to protect the running surfaces against contamination, for example, small particles in the transmission oil. The grease enclosed in the bearing provides the lubrication. These bearings are known as *clean bearings.* The synthetic rubber NBR (nitrile butadiene rubber) has proven its value as a sealing material. In individual cases, a similar effect can be achieved by providing a closed bearing housing with small oil bores on the inner ring.

Rolling elements		Bearing type
	Steel ball	Ball bearings
	Cylindrical rollers	Cylindrical roller bearings
	Needle rollers	Needle roller bearings, needle roller and cage assemblies
	Tapered rollers	Tapered roller bearings

Figure 3. Design variants of rolling elements. (Reproduced with permission from Schaeffler Group.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

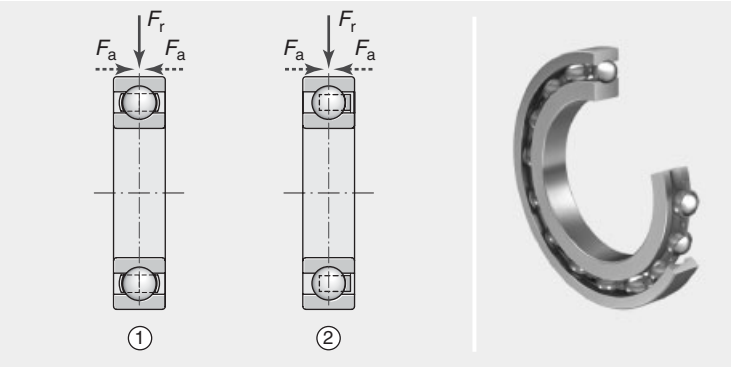


Figure 4. Structure of a deep groove ball bearing. (Reproduced with permission from Schaeffler Group.)

When a ball bearing is designed as a locating bearing, the environment and the intended assembly sequence must be considered so as to ensure that the bearing housing and assembly flange are coordinated optimally. This can be achieved, for example, by including the retainer as part of the assembly in the delivery scope of the bearing manufacturer and thus reducing the assembly time at the transmission manufacturer.

2.1.2 Angular contact ball bearings

The angular contact ball bearing is a type of ball bearing in which the running surfaces of the inner and outer ring are diagonally offset to the bearing axis. This means that greater axial forces (but only in one direction) can be absorbed in addition to the radial forces (Figure 5). With identical dimensions, more balls can be used, which leads to an increase in the load capacity.

A radial load can only be applied to an individual angular contact ball bearing if a minimum axial load is present at the same time. This means that angular contact ball bearings are often installed in pairs and mirror-inverted. The load capacity of the bearing can also be increased by a double row design, provided there is sufficient space for this. Here, we must differentiate between a tandem ball bearing, where both ball sets are of identical size (Figure 6), and a tandem bearing that consists of one large, higher load capacity ball set and one smaller, supporting ball set (Figure 7).

In transmissions, angular contact ball bearings are generally used for medium sized load situations in which a deep groove ball bearing is no longer sufficient, but the use of a tapered roller bearing is not yet necessary. The level of friction that occurs is between that for a deep groove ball bearing and that for a tapered roller bearing.

The combination of an angular contact ball bearing on one side of the shaft in conjunction with a tapered roller bearing on the other side is possible in principle if the forces occurring are known precisely. This type of design would be implemented in a rear-axle differential, for example.

If the lubrication fails, angular contact ball bearings have relatively good emergency properties. If there is insufficient lubricant, they do not allow the shaft to seize up quickly—a situation that could occur above all due to a lack of sliding

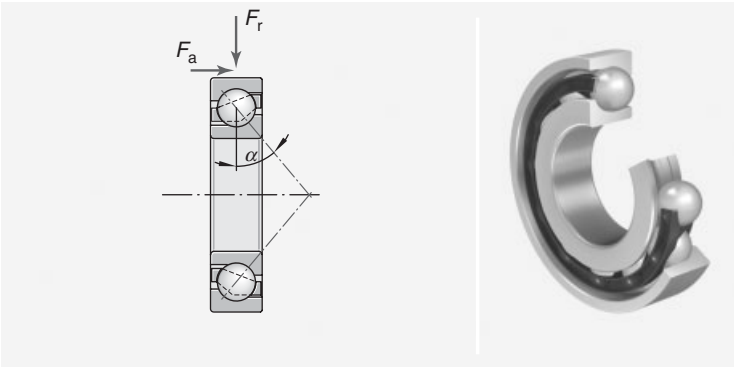


Figure 5. Structure of an angular contact ball bearing. (Reproduced with permission from Schaeffler Group.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

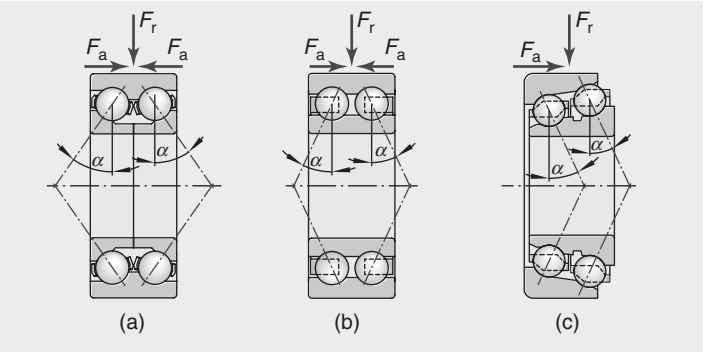


Figure 6. Double row angular contact ball bearing with (a) divided and (b) undivided inner ring. (c) Tandem ball bearing. (Reproduced with permission from Schaeffler Group.)



Figure 7. Tandem angular contact ball bearing for applications in rear-axle transmissions. (Reproduced with permission from Schaeffler Group.)

friction. One disadvantage of angular contact ball bearings compared with roller bearings is the greater sensitivity to external dirt effects, which is the result of highly loaded point-type contact.

2.2 Roller bearings

In a ball bearing, the rolling elements have a point-type contact to the running surfaces, whereas in a roller bearing, the contact is linear. This increases the static load capacity considerably, and the axial load capacity of a radial bearing can also be increased specifically by the design. The greatest disadvantage of almost all roller bearings is the inherently higher friction. In many cases, the ball bearing is also the more cost-effective design. The design variants needle roller and cage assembly and drawn-cup roller bearings are an exception.

2.2.1 Cylindrical roller bearings

Cylindrical roller bearings can be designed as radial or axial bearings in accordance with their main loading direction. To a limited extent, axial forces can also be transferred by a radial bearing via the design of the end retainer in the outer and inner ring. Both design variants can be found in vehicle transmissions. Radial cylindrical roller bearings (Figure 8) are typically used in the non-locating bearing arrangement of the pinion shaft in transmissions—that is, shafts that have an output drive shaft. As cylindrical roller

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

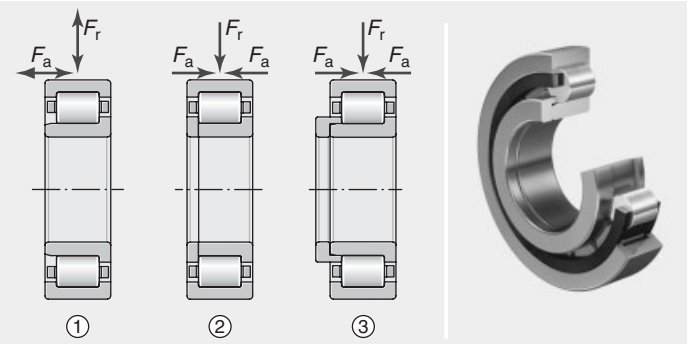


Figure 8. Radial cylindrical roller bearing. (Reproduced with permission from Schaeffler Group.)

bearings are relatively rigid, they are often used in light metal housings.

2.2.1.1 Radial and axial needle roller bearings. Needle roller bearings are used where low section height is advantageous. They are a type of cylindrical roller bearing in which the ratio of length to diameter of the rolling element is greater than 2.5 : 1 (Figure 9). In practice, the design variants are clearly differentiated and irrelevant for individual design cases.

In automatic transmissions, most bearings are axial needle roller bearings (Figure 10). This is because the forces that occur in planetary gear trains cancel each other out to a large extent and only the axial forces that occur due to the helical gearing of the sun gear/ring wheel pairing have to be supported by the bearing. In the radial direction, the bearing

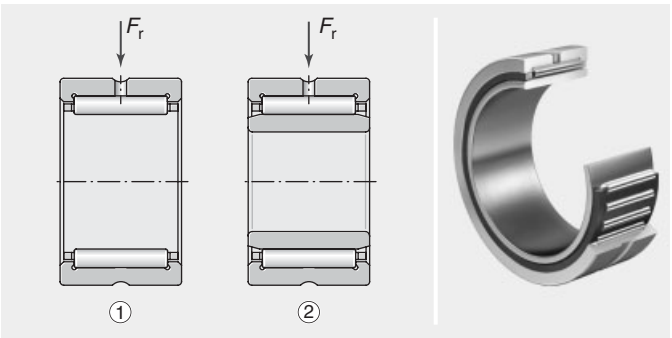


Figure 9. Structure of a radial needle roller bearing. (Reproduced with permission from Schaeffler Group.)

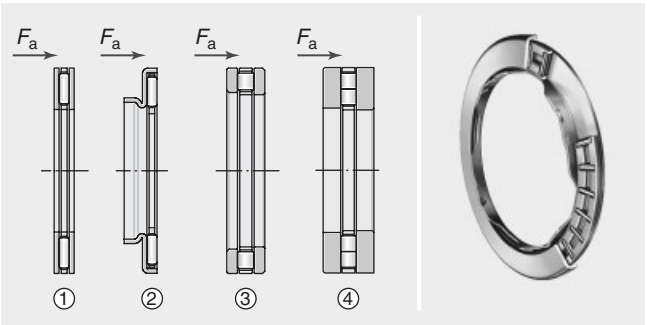


Figure 10. Different design variants of axial needle roller bearings and cylindrical roller thrust bearings. (Reproduced with permission from Schaeffler Group.)

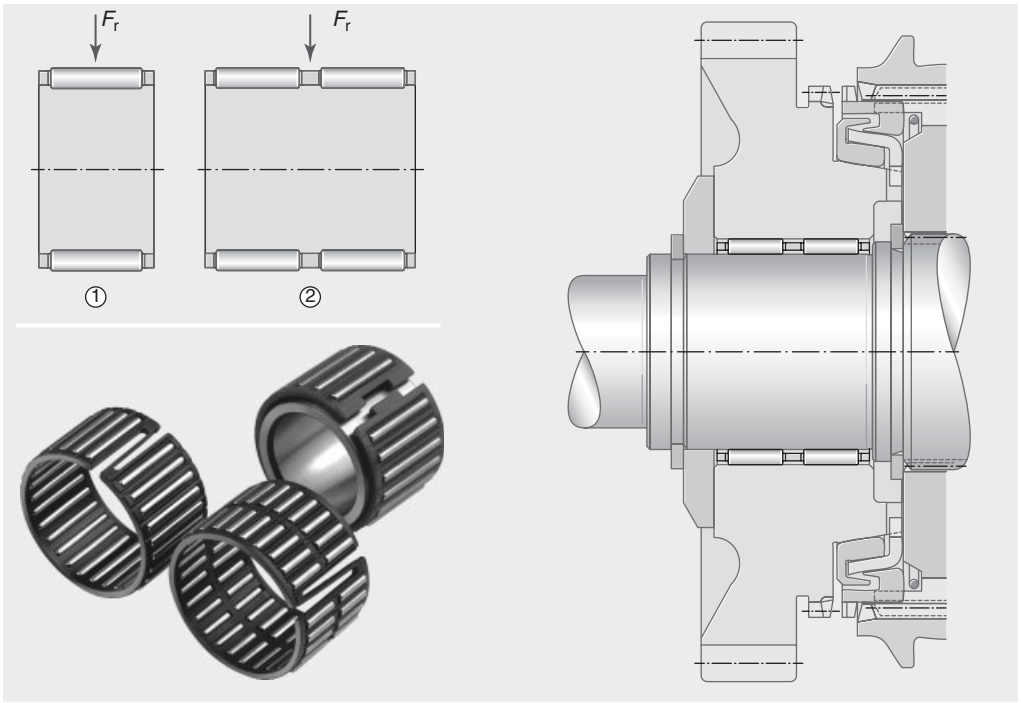
Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

in automatic transmissions primarily takes on management functions. Additional functions can be assigned to the thin-walled discs of the bearing, such as oil throttle points or fixtures that counteract faulty assembly.

2.2.1.2 Needle roller and cage assemblies. Needle roller and cage assemblies have no outer and inner ring—the rolling elements run directly on the surrounding components. The rolling elements can be very small in relation to the bearing diameter. If needle-shaped rolling elements are run in a cage, they take up very little space, which is why they were used increasingly in transmission manufacturing in the 1950s. They have almost completely eliminated the plain bearings formerly used.

Needle roller and cage assemblies are used for the bearing arrangement in constant mesh gears (Figure 11). In the design, it is very important to consider the ambient conditions: to cool the synchronizer, sufficient oil supply must be ensured in all circumstances by designing the cage correspondingly (e.g., with grooves). When speed gears are not engaged, the bearing load is very uneven due to the vibrations that occur. This leads in particular to high loads on the cage. Suitable design measures must be used to protect the cage against fractures. When speed gears are engaged, the static loads are very high as there is no relative movement between the speed gear and the shaft. With the right design of the clearance between the cage and the shaft, it is possible that the rollers do undergo a slight relative movement, thus preventing damage.

The planet gears of automatic transmissions also have needle roller and cage assemblies as bearings. Here, "crank pin cages" that can absorb very high acceleration forces are used. The name comes from the manufacturing of two-wheeled vehicles where these bearings were first used. They are therefore optimized for minimal friction. In comparison with earlier bearing designs, the friction reduction is up to 2 kW for the entire automatic transmission. When designing the crank pin cages, it is important to note that centrifugal forces of up to 7500 g are at work in the planet gears (Figure 12). In combination with the gear wheel meshes—in particular for torque steps—these centrifugal forces bring strong fluctuating loads into the bearing. The life is therefore influenced more by the ambient conditions, in particular the safeguarding of good lubrication, than by static loads (Hertzian contact pressure).



F_r = Radial force
1- Single-row needle roller cage assembly
2- Double-row needle roller cage assembly

Figure 11. Different types of needle roller and cage assemblies as used in speed gears. (Reproduced with permission from Schaeffler Group.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

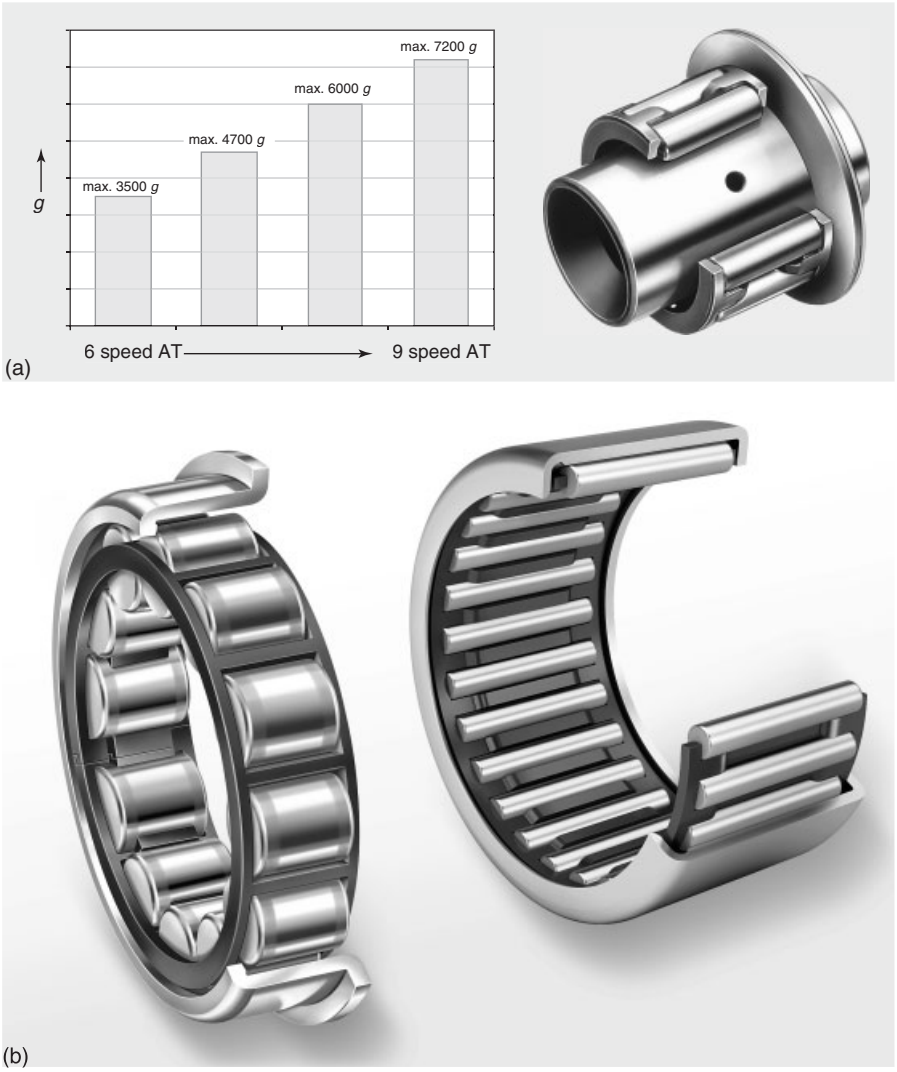


Figure 12. (a) Centrifugal forces on planetary bearings in automatic transmissions. (Reproduced with permission from Schaeffler Group.). (b) Drawn-cup roller bearings. (Reproduced with permission from Schaeffler Group.)

To achieve an optimum service life, the trend is toward developing bearings and planet gear sets together and supplying them directly to transmission manufacturers as an assembled component. The advantage is not only in the shorter assembly time but also in the optimized design with regard to oil supply.

2.2.1.3 Drawn-cup roller bearings. Drawn-cup roller bearings are not manufactured by mechanical processing (turning, hardening, grinding, and honing). Instead, they are formed by deep drawing and therefore offer considerable cost advantages, provided large quantities that justify the higher tooling costs can be achieved. However, in high volume production, maximum precision and process stability must be ensured, as mechanical post-processing is not feasible. The manufacturing process is important even at the material selection stage. Low carbon steels that achieve the required strength through case hardening after processing are generally used.

In particular, drawn-cup roller bearings are used for the non-locating bearing arrangement in a locating and non-locating bearing system. What makes them so useful is the excellent ratio of required space to load capacity. Thus, in comparison with a cylindrical roller bearing, either a considerable amount of space can be saved or the load capacity can be considerably increased while keeping the dimensions the same. The fact that, as a drawn part, the outer ring of drawn-cup roller bearings is relatively thin is generally an advantage but must be considered with regard to maximum pressure. This is particularly true if

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

aluminum or magnesium is used for the drawn-cup rather than steel.

2.2.2 *Tapered roller bearings*

In a tapered roller bearing, tapered rolling elements run on an inner ring that has a conical track and an inner ring shoulder (Figure 13). The higher level absorbs the axial forces and the lower level is the actual track.

Tapered roller bearings can also only absorb axial forces in one direction. Therefore, in transmissions, they are installed almost without exception in pairs and mirror-inverted. The important thing in the design of tapered roller bearings is that axial and radial forces must not be considered independently of one another. If a radial load is applied to this type of bearing at a certain contact angle, backpressure forces with an axial load component arise within the bearing. This vector must be absorbed by the counter bearing in addition to the axial forces working externally.

Tapered roller bearings are used in automotive engineering primarily where very high loads occur and high bearing arrangement rigidity is required, for example, in trucks but also in passenger cars with high engine power. Typical uses cases are differential transmissions (Figure 14) or bearing arrangements in the output shaft. The friction of tapered roller bearings is inherently higher due to the larger contact surfaces. This is why ball bearings are preferred as long as they can control the forces that occur.

As tapered roller bearings require more lubrication than ball bearings, particular importance must be attributed to this aspect during component design. Specific oil bores are often implemented in the housing and the shaft. Oil feed devices such as distributor plates or drip trays are optional.

A high degree of integration of components is also becoming increasingly important in tapered roller bearings. For example, double row design variants in which the rolling elements run in a common outer ring are used successfully.

2.3 Other types of bearings in transmissions

For manual transmissions, a shift rod bearing arrangement is required. This can be implemented via a ball cage that can execute a longitudinal movement. The bearings in the transmission are often assigned further functions, for example, a contact that switches on the rear light when reverse gear is engaged. In automated manual and dual clutch transmissions, the contact signal is also an input parameter for the transmission control system. The trend for gearshift mechanisms is also toward pre-assembled complete units that include the shift rod, the shift forks, the detents, the bearings, and the assembly aid.

As described, rolling bearings have largely eliminated the plain bearings previously used in transmissions. Bushes are still used for radial loads and thrust washers for axial loads in some cases, in particular in automatic transmissions (Figure 15).

3 BEARING DESIGN FOR TRANSMISSIONS

Transmission bearings are designed specifically for the respective transmission and the installation location; "catalogue bearings" are virtually unknown. It is therefore beneficial if the transmission manufacturer—in some cases, the vehicle manufacturer themselves—works together with the bearing supplier from an early design stage. If the two parties define the space required for the bearing arrangement together, technically optimum solutions that also offer high efficiency are possible, for example, by planning the assembly flange as part of the bearing housing.

The process of bearing development should always begin with a precise analysis of the entire transmission. In addition to the design conditions and the forces and torques at work, the oil supply in the transmission must also be considered. Only then should a specific

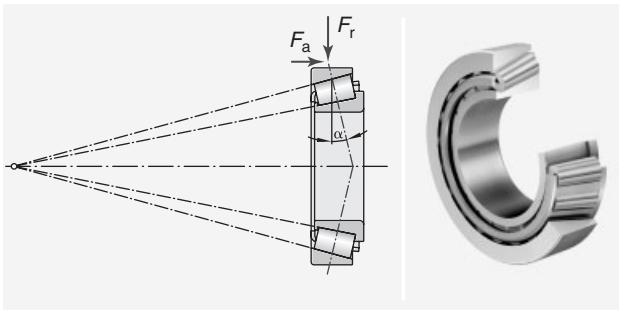


Figure 13. Structure of a tapered roller bearing. (Reproduced with permission from Schaeffler Group.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

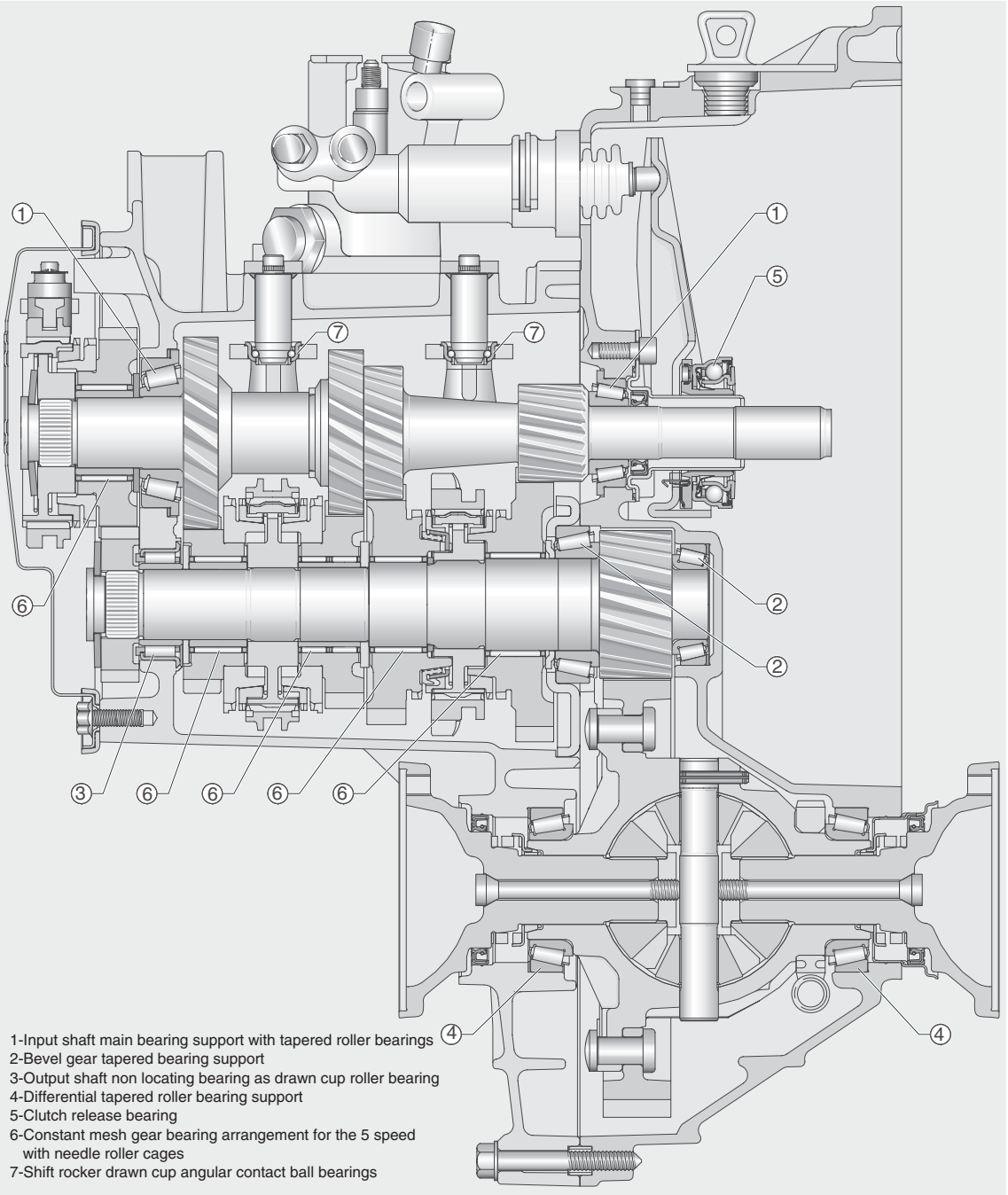


Figure 14. Installation positions of different bearings in a front-drive manual transmission. (Reproduced with permission from Schaeffler Group.)

bearing concept and its design implementation be selected. Therefore, it is not possible to assign specific bearing construction types to applications generally. Figure 16 illustrates this using the example of a bearing arrangement for a main shaft. The first variant, a locating and non-locating bearing arrangement system, has the lowest friction. In contrast, the adjusted bearing arrangement is extremely rigid and requires the least space. Alternatively, a floating bearing arrangement shows no definite axial fixture and causes the least difficulties with regard to axial positioning of, for example, shifting devices.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

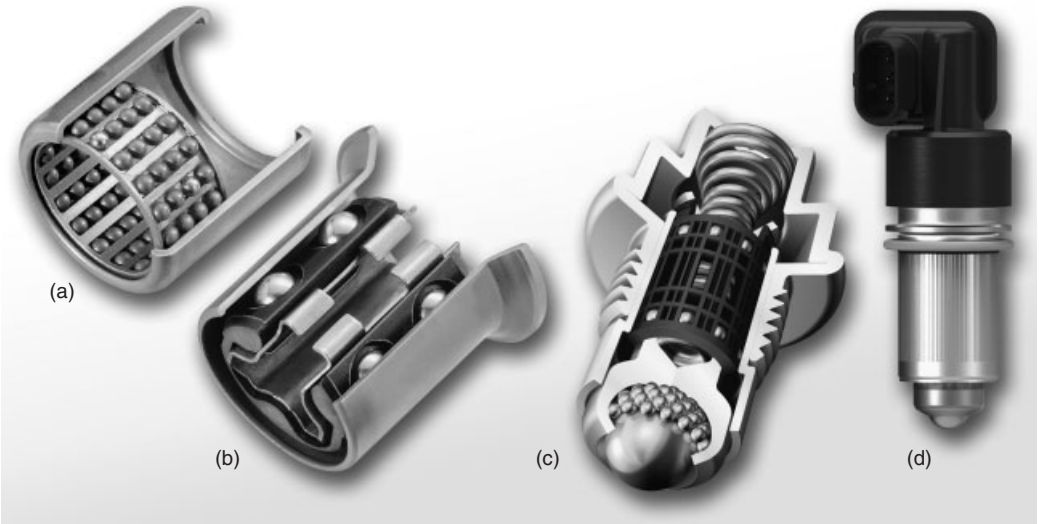


Figure 15. Linear bearing installation. (a) Shift rod bearing for rotated and limited linear motion. (b) Shift rail bearing for rotated and limited linear motion. (c) Detent pin. (d) Detent pin with integrated sensor or switch contact. (Reproduced with permission from Schaeffler Group.)

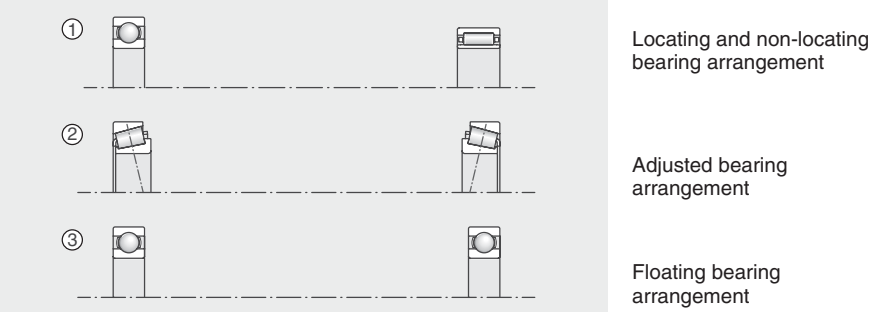


Figure 16. Main shaft transmission bearing support variants. (Reproduced with permission from Schaeffler Group.)

3.1 Service life

The most important design criterion of every bearing is its service life. The service life is the life during which no component failure occurs under typical load conditions and defined cases of misuse. In automotive manufacturing, the service life is generally 300,000 km for passenger cars, which corresponds to a usage time of approximately 5000 h. As the transmission components are not constantly subject to full load during the service life, this results in typical lives of between 500 h for passenger cars and 6000 h for heavy trucks (Naunheimer, Bertsche, and Lechner 2007).

The basis for the durability calculation is the fatigue theory developed by Lundberg and Palmgren (1947). According to this theory, the dynamic load rating C is decisive for bearing fatigue; C is defined as the load of unchanging size and direction at which a sufficiently high quantity of identical bearings achieves a nominal life of one million revolutions.

This calculation has since been considerably extended, and the following parameters are included in the extended calculation method:

- The extent of the bearing load
- Material fatigue limits
- Degree of surface separation by the lubricant
- Cleanliness in the lubricating gap
- Lubricant properties
- Inner load distribution and friction in the bearing

Beyond these parameters, modern simulation tools such as the BEARINX program from Schaeffler also consider the influence of heat treatments and surface coatings as

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

well as the influence of the overall system—for example, the bending of transmission shafts under load. In reality, loads occur that are not reflected by DIN ISO 281 for the durability calculation. Therefore, the power flow in the entire transmission, including the gearings, must be modeled. The load spectrum used for the calculation differs considerably depending on the vehicle manufacturer.

Nowadays, the calculation also enables the property of the lubricating film to be calculated for every load point: for example, the friction state (mixed friction, hydrodynamic friction, etc.) of the bearing is always known. The direction of flow of the oil and the temperature of the oil must also be analyzed. This makes it possible to adjust the bearing geometry such that a load-bearing lubricating film is built up under all operating conditions.

Although the calculation was previously used to verify a bearing arrangement created by a design engineer, the procedure today is almost reversed; initially, based on calculations, a set of rolling elements is optimized in a simulated transmission environment to the extent that it fulfills the service life demands (without being excessively large); then the entire bearing and, if applicable, additional peripherals such as the assembly aids mentioned, are designed. The bearing developer generally selects the rolling element set used for the calculation from an existing library of bearings already implemented and then optimizes it step-by-step.

Important options for optimizing the life include changing the geometry—provided this is possible within the mounting space—and adjusting the material, including the heat treatment. Other methods for increasing the strength are case hardening and strength peening. The standard basic material for transmission rolling bearings is still steel type 100Cr6 (carbon content 0.9%, chrome content 1.35%). Steels with higher chrome and magnesium content are also used. The life of the bearing can also be increased up to ten times by selecting the ideal heat treatment (i.e., aligned to use).

Maximum surface hardness is not, however, the key to the longest service life. This is because under some circumstances, the rolling elements can roll over dirt particles transported with the transmission oil more easily in a softer surface compared to a very hard surface that, in an identical operation, could cause permanent damage.

To achieve the required service life in practice, knowledge of the typical damage patterns is also necessary (Section 4).

3.2 Friction

In light of the increasingly stringent CO_2 limits, minimal frictional losses are becoming even more important for the entire powertrain. After the frictional losses in the gearing, the losses through the rolling bearing are important. Various studies have demonstrated that simply by optimizing the bearings, the overall friction in a manual transmission can be reduced by up to 2% and in a rear-axle differential by up to 4%.

Using modern simulation tools, the friction of a rolling bearing can be predicted to a precision of 10–15%. In turn, the accuracy of the calculation for the entire transmission depends on how well the environmental influences, such as oil churning losses, are considered.

In this encyclopedia, the friction optimization of transmission bearings is addressed in a separate chapter (see Chapter "Rolling Bearings and Bearing Design for Transmissions" in this encyclopedia). The chapter also discusses the surface coating as an option for low friction rolling bearings.

3.3 Acoustics

The use of guide bearings can reduce the relative movement of a shaft to the transmission housing, thereby improving the acoustic properties of the powertrain. One example is the use of polygon bearings that pretension a shaft against the housing, that is, they eliminate the shaft clearance.

The definition of the bearing points has a considerable influence on the vibration behavior of the shaft under dynamically changing loads. Therefore, the bearing points should not be defined considering only the aspects strength and the mounting space.

4 BEARING DAMAGE AND PREVENTING DAMAGE

In principle, bearing damage must be avoided by making the bearings of the correct size and using a suitable design construction. However, due to the objectives of light construction and downsizing, as well as the economic pressure in the automotive industry, the classic engineer virtue "if in doubt, a little bit more" is no longer an option. If, however, bearings are to be adapted precisely to the loads present, it is important to be aware of any potential damage that may occur and to plan suitable countermeasures during development.

Just like any other machine element, a transmission rolling bearing changes over the course of its use. Therefore, there must be a differentiation between normal signs of use and damage that can lead to restricted function or even failure of the component. Some typical types of damage for transmission bearings are discussed below.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.1 Fatigue damage

"Natural" bearing damage is the material fatigue that occurs at some point for every bearing. However, if the design is correct, it occurs only after the end of the service life. Fatigue is caused by alternating shearing stresses near to the surface of the structure of a bearing material. An inner ring raceway is more affected than the rolling element or the outer ring raceway. Fine cracks appear and gradually multiply.

As this type of damage during the service life can be eliminated at the outset by choosing an appropriate design, fatigue-related wear on the surface is particularly important in practice. It arises when the maximum shearing stress shifts as a result of fluctuating load at the surface. In some circumstances, very low expansion plastic deformations can occur there (<0.1 mm). If this damage occurs frequently, it is visible to the naked eye and is called *gray staining* due to its appearance (Figure 17).

If gray staining occurs, the first thing to check is whether the optimum type of bearing construction was selected for the loads occurring in the axial and radial direction. Switching to a different type of bearing construction may resolve the situation. Triggers for this type of surface damage can also be individual particles in the lubricant, which are pressed into the material by the rolling element, causing a plastic deformation. Therefore, the formation of the lubricating film must also be analyzed.

4.2 Wear damage

Wear damage comprises all damage patterns where the surface gradually changes. There is a differentiation between abrasive and adhesive wear. Adhesive wear arises through the direct atomic contact between the sliding contact surfaces, whereas abrasive wear is caused by frictional forces that arise on the running surface.

A frequent form of wear damage for transmissions is riffling in constant mesh gears (Figure 18). It arises because the bearing that is not moved relative to the shaft is subject to high loads. Due to the lack of speed, sufficient lubricating film cannot be built up, meaning that there is direct metallic contact. As the needles of the speed gear bearing always come to rest at the same point on the outer ring, longitudinal grooves occur—known as riffling. A design idea can provide help here; the very small rolling movements of

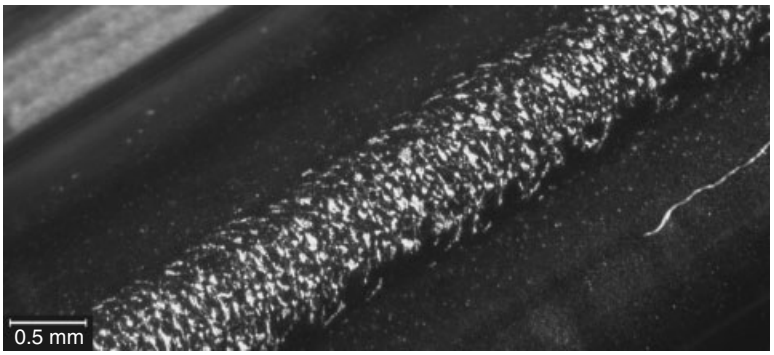


Figure 17. Macroscopic image of the gray staining on a roller sleeve. (Reproduced with permission from Schaeffler Group.)

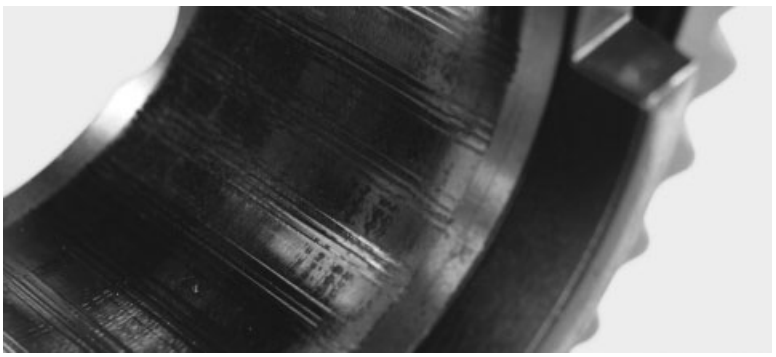


Figure 18. Riffling in a constant mesh gear. (Reproduced with permission from Schaeffler Group.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the needles are used to move the entire cage slightly on the shaft, creating even, uniform wear over the entire surface.

4.3 Corrosion damage

Even though every transmission is in principle protected against oxidation by the lubricant, specifically through the additives used in the lubricant, corrosion damage can still occur from time to time. The cause is usually contact corrosion, which occurs when a material pairing is unevenly coated with lubricant over a constant period. This type of damage is particularly relevant in the transmission area of vehicles that are stationary for long periods, because they are only used on a seasonal basis (combine harvesters, ski run snowplow, etc.). Special surface coatings can be used to resolve this problem.

4.4 Bearing overload

Extreme load spectrums that were not considered in the bearing design can lead to premature fatigue and then to bearing fracture. One example of this is the constant rally use of a road vehicle. Repeated load impacts through extreme contraction of the shafts in the transmission can lead to a fracture in this type of situation; thin-walled bearings such as roller sleeves are particularly at risk here. Bearing fractures can also be caused by primary damage at the transmission shafts or the planet gear sets of an automatic transmission. As a result, the load on the bearing is so uneven that the bearing deforms and eventually breaks (Figure 19).

In modern automotive development, individual misuse tests are performed as standard, for example, driving over a curb at high speed. These tests are reflected in the load spectrums used for the bearing design. Any further damage occurring must be analyzed in each individual case and the relevance for the actual operation of the vehicle must be considered.

4.5 Temperature failures

Overall, temperature failures in transmission bearings are very rare. However, if they do occur, for example, during extreme load tests (e.g., repeated journeys over high mountain passes), they usually indicate insufficient lubrication. A self-reinforcing effect can occur because an increased oil temperature leads to a decrease in the oil viscosity. However, the thinner the oil, the more difficult it is for the load-bearing lubricating film to form—and therefore, the greater the waste heat arising from the mixed friction.

Bearing cages that are, in the case of transmission rolling bearings, often made from colored metals or plastics suffer particularly from excessive temperatures. If the local temperature exceeds the permitted value for a long period, the cage can become permanently deformed or even melt completely in extreme cases.

4.6 Special features in hybrid vehicles

In hybrid vehicles, the electric motor is often installed directly in the bell housing of the automatic transmission or in the direct vicinity of the differential. This led to the

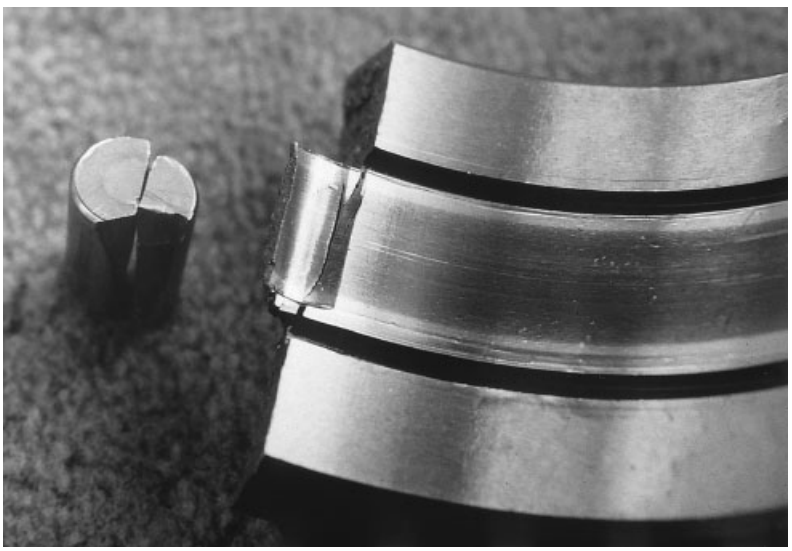


Figure 19. Forced fracture on the bearing of a planet gear. (Reproduced with permission from Schaeffler Group.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

fear that the magnetic fields occurring could influence the bearing function—for example, metallic particles in the transmission oil could move in the direction of the field and, under some circumstances, could also be deflected to the bearing surfaces. In existing hybrid vehicles, this issue has been considered to some extent in the oil conduit in the transmission. No increased damaged has been observed in the field so far.

5 SUMMARY

The requirements concerning the design of transmission bearings are constantly increasing. The classic criterion of a sufficient service life is supplemented by lower frictional losses and—thanks to the competitive pressure in the automotive industry—a cost-optimized design. This leads increasingly to modules that perform further functions in addition to pure bearing arrangement.

Using modern calculation tools, it is possible to optimize the entire shaft/bearing system of a transmission numerically first. This approach enables different types of bearing construction as well as materials and after-treatment methods for the respective installation location to be varied very quickly and conflicting goals to be resolved. It is therefore impossible to provide hard and fast rules for individual applications.

With proper design and by using modern methods, it is not difficult to reflect the required service life. However, engineers must still have knowledge of typical damage patterns, particularly when new technologies are being tested.

REFERENCES

Lundberg, G. and Palmgren, A. (1947) Dynamic capacity of rolling bearings. *Acta Polytechnica Mechanical Engineering Series*, **1** (3).

Naunheimer, H., Bertsche, B., and Lechner, G. (2007) *Automotive Transmissions: Fundamentals, Selection, Design, and Application*, 2nd edn, Springer, Heidelberg, Germany.

Schaeffler (2006) Failure analysis: INA bearing failure mode archive. Technical product information 109, Herzogenaurach, Germany.

FURTHER READING

Brändlein, J., Eschmann, P., Hasbargen, L., *et al*. (2009) *Ball and Roller Bearings*, 3rd edn, Vereinigte Fachverlage, Mainz, Germany.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto073
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Mechanics of Contacting Surfaces

Homer Rahnejat[1] and Patricia Margaret Johns-Rahnejat[2]

[1] *Loughborough University, Loughborough, UK*
[2] *Previously at Imperial College, London, UK*

1 Introduction 1
2 Contacting Solids Under Normal Load 1
3 Elastohydrodynamics 5
4 Friction 7
Notations 8
References 9

1 INTRODUCTION

There is hardly any system; mechanism, device, or machine that does not incorporate load bearing surfaces. These are the interfaces formed by contact of a pair of solids, often limiting the performance of the system as a whole. The same is true of the natural world itself. For example, the load that a species can carry is often a measure of its skeletal structure and the limiting joint stresses. The working efficiency of species is not only a function of their respiratory and circulatory system but also their joints' friction. Therefore, similar to the natural world, machines have performance constraints in terms of load and speed. These are two key parameters that need to be considered in the operational performance of all machines. The geometrical, mechanical, and physical properties of contacting surfaces determine their performance, thus the machine which they are a part. The subject that deals with the fundamentals of load bearing surfaces is called *contact mechanics*. This chapter intends to familiarize the reader with the fundamental aspects of the subject. Two aspects are of interest; firstly the behavior of the contacting bodies under normal load and secondly their interactions in relative motion.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto075
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 CONTACTING SOLIDS UNDER NORMAL LOAD

The subject of contact mechanics was initiated with the works of Pascal (1653), where the pressure generated between a pair of solid smooth surfaces with an interfacial area A under a normal applied load W can be determined as $p_{\mathrm{m}} = \frac{W}{A}$. This is the mean (as opposed to the maximum) pressure that is generated between a pair of solids in contact. The contact area A is the footprint area (impression) that one of the solids, considered as a rigid body, would make on the other having an effective stiffness of the pair. The footprint shape and area are functions of the geometry of the solid surfaces.

The deformation behavior of contacting solids varies according to the applied load and their mechanical properties. Broadly, this can be classified into two categories. One category is concerned with the global deformation of solids, where significant changes in the shape of solids occur. These various shapes of the solids under load are known as *mode shapes*. Depending on loading and compliance of the solids, there may be many such mode shapes. In all cases, the extent of deformation can be comparable to the geometrical dimensions of the solids themselves. The other category is concerned with the localized deformation of the solids in the region of their contact, being far smaller than the overall geometrical dimensions of the solids themselves. The difference between these forms of elastic deformation, global versus local, is the extent of elastic deformation

strain (Rahnejat, 2010). Therefore, with localized contact deformation, there is no appreciable change in the shape of the loaded contacting solids themselves (i.e., very small strains). This chapter is devoted to the contact mechanics of solids with localized deformation.

There are certain salient features for localized deformation of solids in contact or impact. These were noted by Hertz (1881). In Pascal's equation, the contact of solids is considered to be over a flat region. This implies that surfaces with infinite radii of curvature are idealized. In practice, the contact occurs over a finite region. In other words, all surfaces have finite radii of curvature, however large these may be. Hertz (1881) noted that localized deformation occurs, when elastic deflection of solid surfaces in contact, δ, is far smaller than the footprint contact dimensions, which are in turn far smaller than the radii of curvature of the surfaces at the point of contact. Thus, for an assumed circular footprint contact of radius a, made by an equivalent rigid spherical indenter of radius R, pressed onto an elastic plane of large dimensions $\delta \ll a \ll R$. For example, a sphere of several millimeter radius loaded onto a compliant surface yields a circular contact footprint radius of a few tenths of a millimeter, causing a deflection of several tenths to a few micrometers.

As in dynamics problems where a simple model is a mass–spring system, in contact mechanics, one strives to make a basic model, comprising an ellipsoidal rigid solid loaded against an elastic surface of assumed infinite dimensions termed an *elastic half-space* or a *semi-infinite elastic solid*. This is assumed in order to reduce the effects often associated with defined boundaries. The radius of the rigid solid should consider the radii of contact of the actual physical surfaces and their elastic properties should incorporate those of the same surfaces as well.

Figure 1 shows the different contact configurations. Figure 1a is that of a pair of closely conforming solids (i.e., radii R_1 and R_2 are very similar in size). This type of contact is referred to as *conforming*, good examples of which are journal bearings, where $\frac{R_1}{\Delta R} > 100$, $c = \Delta R = R_2 - R_1$ is the nominal clearance and R_1 is the journal radius. R_2 is the radius of the bearing bushing or shell. Figure 1b shows the partially conforming (or counterformal) contact of two solid surfaces. A good example is that of a ball bearing or a roller bearing (end view) in its raceway groove. The degree of conformity R_1/R_2 is no longer near unity (as in Figure 1a). Typically, for the contact of rolling elements to raceway grooves, this ratio is between 0.8 and 0.9, whereas for journal bearings, it is usually in excess of 0.99. This small difference in conformity makes a significant change in the mechanics of contact. Figure 1c shows the extreme case of a counterformal contact with zero degree of conformity (R_2 being very large). The area of contact becomes infinitesimally small. This is an idealized case, used for modeling purposes. In practice, counterforming contacts are showed by Figure 1d.

The contact area is determined by the degree of conformity. The angle φ in Figure 1 shows the extent of contact. In the case of journal bearings this could extend to the value of π. Therefore, a large contact area would result, whose dimensions are comparable with the radii of the contacting solids. Thus, the use of Hertzian theory is inadmissible. For the partially conforming contacts and those with a counterforming nature, one may consider the Hertzian theory. For example, for a roller in a raceway groove, the small value of φ yields a thin rectangular strip footprint, whose width is a function of the applied load and whose length is approximately that of a right cylindrical roller, except for its relieved axial end extremities. These forms of contact, usually with high loads and small contact area, are termed *concentrated contacts*. The footprint shape is a function of the applied load, elastic properties of contacting solids, and their radii of curvature. These are known as the *principal radii of contact*, which are usually given in the planes zx and zy in Figure 2 for an equivalent ellipsoidal solid representing the contact of a ball–inner race in a ball bearing.

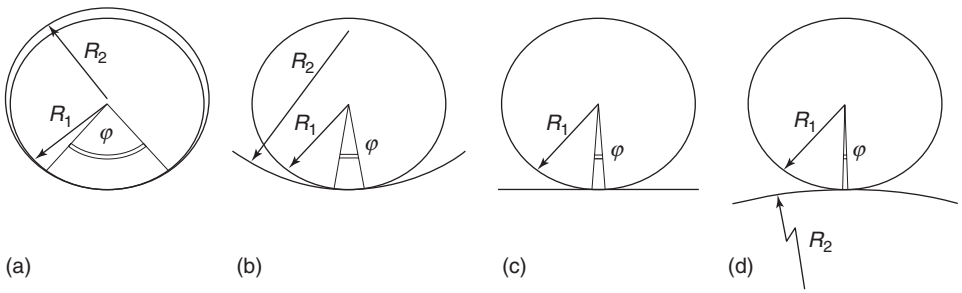


Figure 1. (a–d) Various contact configurations.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto075
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

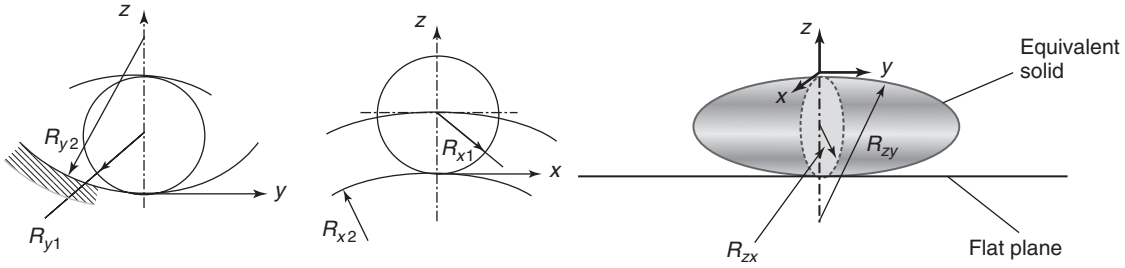


Figure 2. Actual geometry and the equivalent ellipsoidal solid.

The radii for the equivalent ellipsoidal solid, contacting a semi-infinite elastic plane, are obtained as:

$$\frac{1}{R_{zx}} = \frac{1}{R_{x1}} + \frac{1}{R_{x2}} \quad \text{and} \quad \frac{1}{R_{zy}} = \frac{1}{R_{y1}} - \frac{1}{R_{y2}} \tag{1}$$

Note that a concave surface is represented by a negative radius of curvature, in this case the groove radius R_{y2}. For the counterforming contact, the equivalent radius, $R_{zx} < R_{x1} < R_{x2}$, thus it is referred to as the *reduced radius*. For the partially conforming surfaces, $R_{zy} > R_{y2} > R_{y1}$, hence the equivalent radius is referred to as the *increased radius*. It is clear that for case (c) in Figure 1, the radii of curvature $R_{x2} = R_{y2} \to \infty$ (very large) and the ellipsoidal solid becomes a sphere. As the ellipsoidal solid is considered to be rigid, the elastic modulus of the semi-infinite elastic plane is adjusted accordingly and is referred to as the *equivalent* or *effective elastic modulus*:

$$\frac{1}{E^*} = \frac{1-\upsilon_1^2}{E_1} + \frac{1-\upsilon_2^2}{E_2} \tag{2}$$

If both the solids are made of the same material, Equation 2 simplifies to $E^* = \frac{E}{1-\upsilon^2}$. This is known as the *plane strain elastic modulus*. The contact stresses are considered to be compressive $\sigma_z(x,y) = -p(x,y)$, where x, y is the small flattened contact region (for a circular footprint x and $y \le a$). Outside the contacting region (x and $y > a$), $p = 0$. The contact of an ellipsoidal solid with a semi-infinite elastic half-space is an elliptical footprint, with an aspect ratio (Figure 3):

$$e_p^* = \frac{b}{a} \approx \left(\frac{R_{zy}}{R_{zx}}\right)^{\frac{2}{3}} \tag{3}$$

In ball bearings, the direction of ball rolling is along the minor axis of this contact ellipse. Also note that for a circular point contact $R_{zx} = R_{zy} = R_{x1} = R_{y1} = R$ (radius of a sphere) and thus $e_p^* = 1$ (a circular contact footprint of radius a). The contact of a crowned roller with its inner

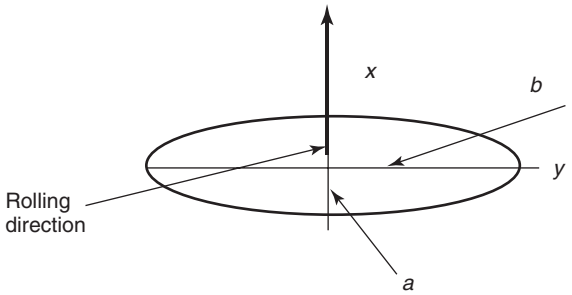


Figure 3. An elliptical footprint.

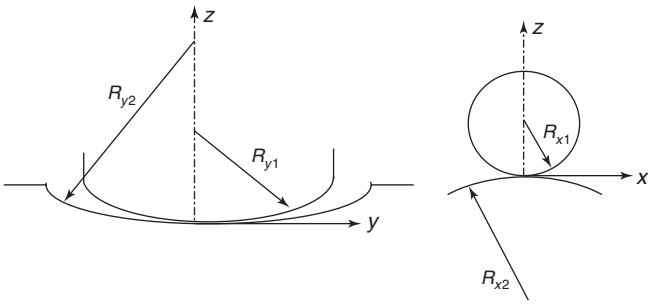


Figure 4. Roller–race contact geometry.

raceway groove (Figure 4) can also be represented by an ellipsoidal solid against a semi-infinite elastic flat plane as in Figure 1. The ellipsoidal solid would usually be long and slender (i.e., $R_{zy} \gg R_{zx}$). Therefore, the elliptical aspect ratio would be quite large, normally with a value in excess of 10. In fact, the contact footprint of a roller of any axial profile (crown barreled as in Figure 4, crowned at the edges or with small dub-off radii at its extremities) with a flat would usually be a dog bone or dumbbell shape because of sharp generated pressures at its rather abrupt axial edge profile (Johns and Gohar, 1981).

Disregarding the spread in the footprint shape at its extremities, the contact of a pair of loaded rollers (such

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto075
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

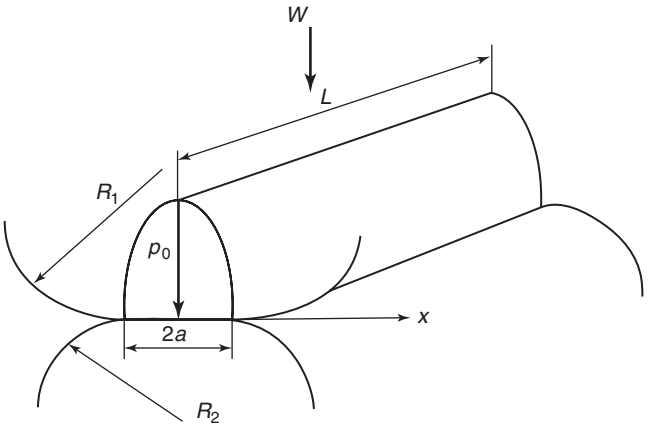


Figure 5. Elastic line contact of cylinders.

as a roller against the inner race, ignoring any roller axial profile) yields a thin rectangular footprint of width $2a$ and length L (nearly the length of the roller itself) as shown in Figure 5. As $a \ll L$, the contact footprint is often referred to as a *line contact*, where an analytic solution for generated pressures and contact deflection was found by Hertz (1881) along the contact width $2a$, considering the pressure distribution to be uniform along the length of contact (Figure 5). For a two-dimensional solution, considering the actual length of the contact, a numerical finite difference or finite element technique is required. The solution is referred to as *finite line contact* (Johns and Gohar, 1981; Rahnejat *et al*., 2009).

A narrow band pressure distribution is generated because of the applied load W, locally flattening the surfaces and forming a thin rectangular contact footprint. Using purely dimensional considerations, the contact half-width may be obtained in the form:

$$a = k\left(\frac{WR}{LE^*}\right)^n \tag{4}$$

where the term in the bracket has the dimension $\frac{FL}{L^F/L^2} = L^2$, with $F \equiv$ force, $L \equiv$ length.

Hence, in order to make the right-hand side of Equation 4 dimensionally compatible with its left-hand side, $n = 1/2$. Hertz (1881) showed that dimensionally, an elliptical pressure distribution is required over the footprint band for long line contacts, and an ellipsoid is needed over a circular footprint for contacting spherical surfaces. In the general case of a point contact with differing principal radii of curvature along orthogonal axes, such as a ball in an annular groove (a ball-race contact is an example), one requires an ellipsoidal pressure distribution over an elliptical footprint. The localized small deflection δ is as the result of the generated pressures. Table 1 provides a summary of classical Hertzian contact mechanics for cases of elastic line, general elliptical point, and the special case of circular point contacts. Note that the mean pressure p_m is the Pascal pressure and for a circular point contact $R_{zx} = R_{zy} = R$.

Hertzian theory applies to an assortment of load bearing surfaces encountered in the various automobile systems, particularly in the engine and drivetrain systems. This is because many of the conjunctions (a word used to describe contact of surfaces with a lubricant film usually present) are subject to high loads applied to infinitesimal contact areas of counterforming or partially conforming solids of revolution. In other words, they are regarded as concentrated contacts. Typical concentrated contacts include cam–tappet pair (Kushwaha, Rahnejat, and Jin, 2000), teeth of an engaged helical gear pair of the transmission (De la Cruz, Theodossiades, and Rahnejat, 2010), and those of hypoid gears of the differential (Litvin *et al*., 2002). The instantaneous contact of the cam against a flat tappet or a roller may be considered as an equivalent cylinder with the instantaneous radius of cam at the point of contact (Gohar and Rahnejat, 2008). This varies according to the cam lift profile. The length of the contact is that of the cam width. The profile of the cam width and that of the tappet are considered to be flat. Therefore, at any instant of time, the

Table 1. Relationships for Hertzian contacts.

Variable	Elastic line contact	Circular contact	Elliptical contact
Contact half-width or radius	$a = \left(\frac{4WR}{\pi LE^*}\right)^{1/2}$	$a = \left(\frac{3WR}{4E^*}\right)^{1/3}$	$\sqrt{ab} = \left(\frac{3W\sqrt{R_{zx}R_{zy}}}{4E^*}\right)^{1/3}$
Maximum and mean contact pressures	$p_0 = \frac{4}{\pi}p_m = \left(\frac{WE^*}{\pi RL}\right)^{1/2}$	$p_0 = \frac{3}{2}p_m = \left(\frac{6WE^{*2}}{\pi^3R^2}\right)^{1/3}$	$p_0 = \frac{3p_m}{2} = \left(\frac{6WE^{*2}}{\pi^3R_{zx}R_{zy}}\right)^{1/3}$
Contact load	$W = 2aLp_m$	$W = \pi a^2 p_m$	$W = \pi abp_m$
Contact center deflection	$\delta = \frac{W}{\pi LE^*}\left[\ln\left(\frac{L^3\pi E^*}{2RW}\right) + 1\right]$	$\delta = \frac{\pi p_0 a}{2E^*} = \left(\frac{9W^2}{16E^{*2}R}\right)^{1/3}$	$\delta = \frac{1}{2}\left(\frac{9W^2}{2E^{*2}\sqrt{R_{zx}R_{zy}}}\right)^{1/3}$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto075
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

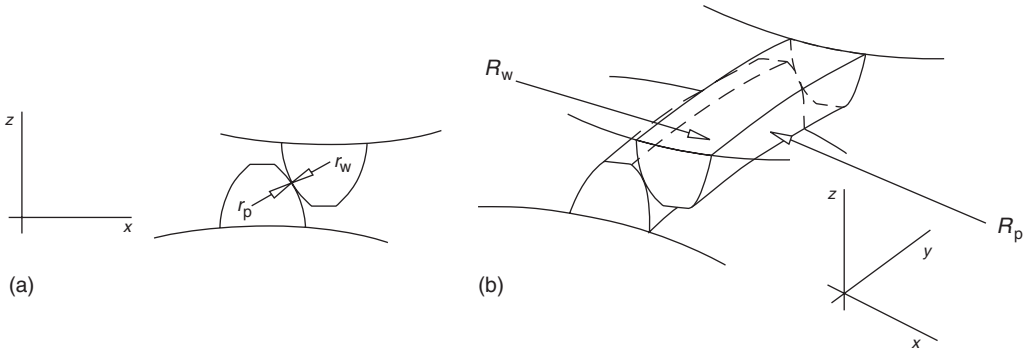


Figure 6. (a, b) Teeth pair contact.

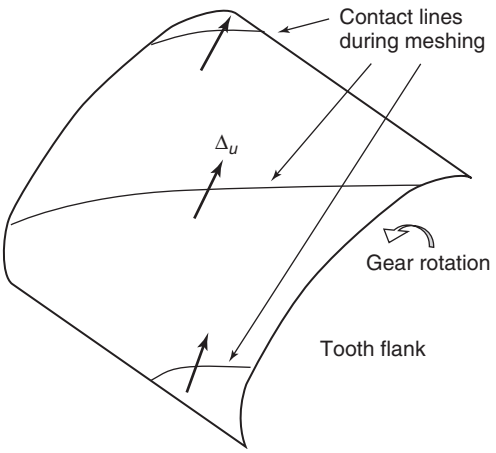


Figure 7. Procession of contact line during a meshing cycle.

contact footprint is a thin rectangular strip similar to the example in Figure 5.

Figure 6 shows typical contact of a gearing teeth pair. Clearly, as in the previous example in Figure 2, at any instant of time, an equivalent ellipsoidal solid can be determined according to the principal radii of teeth pair in the region of contact. For spur and helical gears, the transverse principal radii of contact in the *zy*-plane are very large, so can be ignored. This yields a slanted thin rectangular contact strip across the tooth flank as the instantaneous contact footprint. The contact line is shown in Figure 7, where the sliding motion (Δu is the relative sliding velocity between a pair of contacting teeth flanks) occurs along the interface and at an angle to the thin strip (De la Cruz *et al*., 2012).

In general, all forms of bevel gears have curved surfaces in the *zy*-plane of contact and in the *zx*-plane. Therefore, the hypoid gear pair of the differential can be represented by an instantaneous equivalent ellipsoidal solid; R_{zx}, R_{zy} contacting a semi-infinite elastic half-space of effective elastic counterface modulus E^*. The footprint shape is clearly a slanted ellipse with lubricant entrainment into it at an angle. If one fixes the direction of sliding (lubricant entrainment), the ellipse appears to precess about the contact normal as shown in Figure 8 for a pair of differential hypoid gear teeth through a meshing cycle (Mohammadpour, Theodossiades, and Rahnejat, 2012).

The Hertzian pressure distribution is for dry contact of solids of revolution in a concentrated contact. Figure 9a shows a typical ellipsoidal pressure distribution for an elliptical point contact footprint. For the line contact condition, a typical pressure distribution is shown in Figure 9b. For a finite line contact of a roller against a flat plane, there are pressure peaks at the sharp edges of the contact as shown in Figure 9c (Johns and Gohar, 1981). These pressure spikes are reduced where edges of the roller bearings are provided with a small relief radius, referred to as a *dub-off*. The contact line in Figure 7 is in fact similar to the footprint in Figure 9c.

3 ELASTOHYDRODYNAMICS

The concentrated contacts described in engine and drivetrains are usually lubricated conjunctions. Grubin (1949) proposed the mechanism of elastohydrodynamic (EHD) lubrication, based on his work with Ertel (Figure 10). He postulated that as a film of lubricant is entrained into the contact of a pair of elastic solid surfaces in concentrated contact and in relative motion, its viscosity is increased significantly because of the generated pressures that closely follow the pressure profile predicted by the Hertzian theory. An almost parallel thin lubricant film h_0 is formed in the flattened Hertzian contact footprint. The lubricant becomes almost incompressible, acting similar to an amorphous solid. As the lubricant film flows toward the contact exit, its viscosity reduces to comply with the principle of continuity of flow. The localized reduction in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto075
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

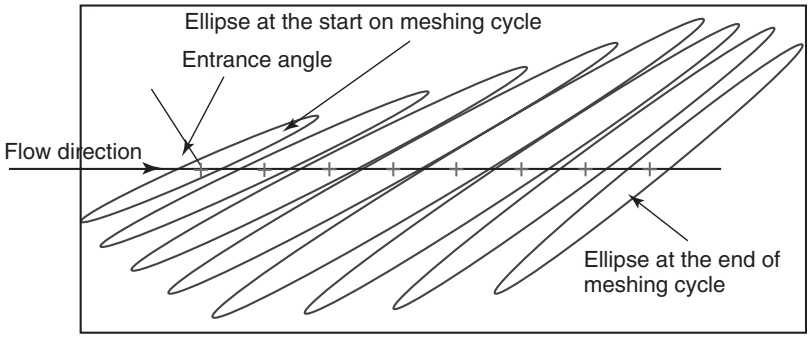


Figure 8. Precession of elliptical contact footprint of a hypoid gear teeth pair.

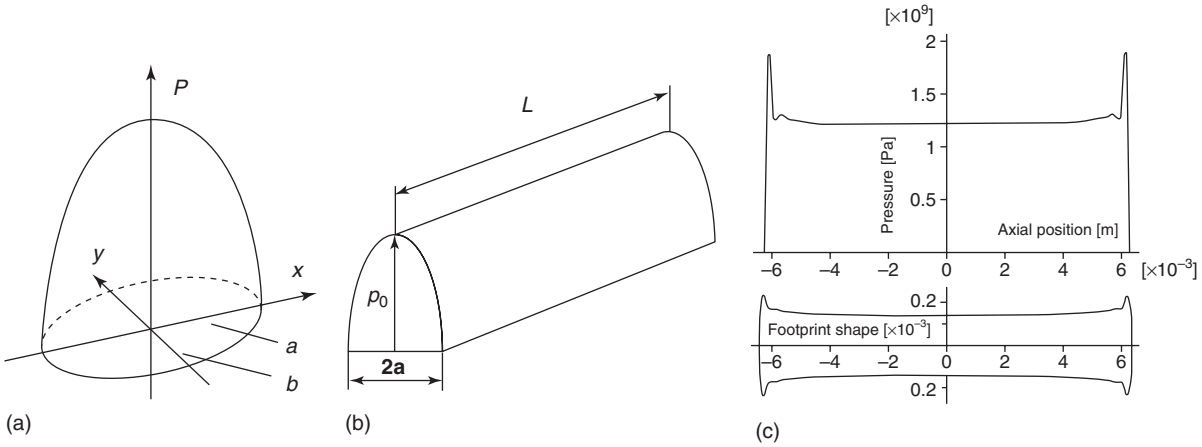


Figure 9. Pressure distribution for different contact footprints. (a) Elliptical point contact, (b) long line contact, and (c) finite line contact.

viscosity reduces the lubricant load carrying capacity, thus diminishes the gap (minimum film thickness, $h_{\min}$) and increases the pressure, which is exhibited by a pressure spike (pip) near the contact exit.

The main interests in prediction of film thickness and the corresponding pressure distribution are to ascertain friction, wear, and fatigue behavior of the contact. These are the limiting performance parameters for any conjunction. Unlike Hertzian theory, prediction of pressure distribution and film thickness (Figure 10) requires numerical analysis. This is usually carried out by finite difference methods (Jalali-Vahid *et al*., 2001; Dowson, Ehret, and Taylor, 1999). Some film thickness equations have been obtained from the results of many simulation studies. The following are a representative sample:

For line contact (Dowson and Higginson, 1966):

$$h^*_{\min} = 1.6G^{*0.6}U^{*0.7}W^{*-0.13} \tag{5}$$

For finite line contact (Mostofi and Gohar, 1983):

$$h^*_0 = 1.67G^{*0.421}U^{*0.541}W^{*0.059} \tag{6}$$

For elliptical point contact (Hamrock and Dowson, 1977):

$$h^*_{\min} = 3.63U^{*0.68}G^{*0.49}W^{*-0.073}(1 - e^{-0.68e^*_p})$$

where:

$$U^* = \frac{u\eta}{E^*R}, \quad W^* = \frac{W}{E^*RL} \text{ (line contact)},$$

$$W^* = \frac{W}{E^*R^2_{zx}} \text{ (point contact)},$$

$$G^* = E^*\alpha, \quad e^*_p = \frac{a}{b}, \quad \text{and} \quad h^* = \frac{h}{R_{zx}}$$

Wear and fatigue are adequately covered by Gohar and Rahnejat (2008). The remainder of this chapter is devoted to friction.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto075
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

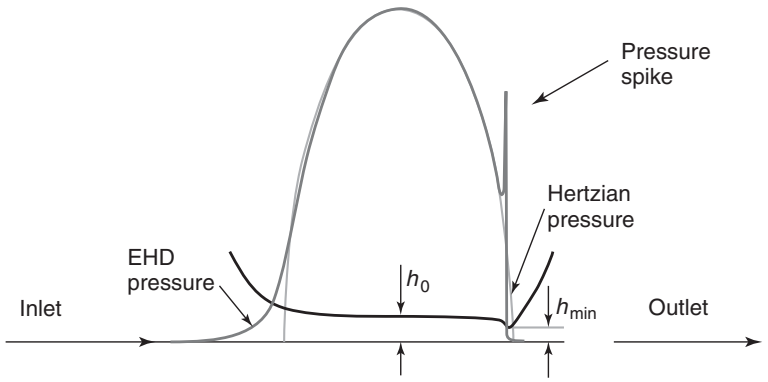


Figure 10. Elastohydrodynamic conjunction.

4 FRICTION

Study of friction originates from the work of Amontons (1699). Friction is defined as the resistance to the tangential relative movement of a pair of surfaces in contact. As real surfaces are rough, then Amontons noted that friction is as the result of interaction of rough topography of the surfaces. Therefore, it is clear that for dry surfaces, friction is independent of the nominal (apparent) area of contact, as the interacting roughness peaks only account for a small proportion of the apparent contact area. Under the application of load, a pair of opposing asperities in contact cold weld together. These junctions need to be broken for any ensuing relative motion, the resistance to which is termed *adhesive friction*. In addition, asperities on the harder surface can elastically deform (flatten) or plastically plow through their counterparts on the softer counterface. This mechanism of friction is termed *plow* or *deformation friction*. The described mechanisms lead to dry or boundary friction, F, which is directly proportional to the applied load W. This was shown by Amontons (1699): $F = \mu W$, where the coefficient of proportionality is known as the *coefficient of friction* μ, which is a function of the material of the counterfaces, typical asperity geometry and distribution on the surfaces (Greenwood and Tripp, 1971).

Friction can be viewed as an energy sink. It is a major source of inefficiency in almost all machines and mechanisms. Therefore, except for the few cases where friction is an essential feature of a system, such as for locomotion, traction, and braking, its reduction is a primary design aim and is becoming progressively more important because of the diminishing traditional sources of energy.

The main principle of lubrication is to separate the contacting surfaces with a film of fluid of low shear strength and avoid shearing and deformation of asperities on solid surfaces with much higher shear strengths. Unlike for the rough surfaces, viscous friction arising from shear of a lubricant film is directly proportional to the area covered by the lubricant. In practice, as in nature itself, almost all wetted conjunctions are subject to a mixed regime of lubrication. This means that thinness of a lubricant film enables interaction of some surface asperities. Therefore, friction occurs as the result of viscous shear of a fluid film as well as some boundary interactions.

Depending on the geometry of the conjunction, film thickness can be predicted by an equation such as Equation 5 or 6. This is compared with the root mean square (RMS) of composite surface roughness (average roughness of the two counterface surfaces: ψ_1, ψ_2) to ascertain the likely degree of boundary interactions. If the RMS surface roughness is $\psi = \sqrt{\psi_1^2 + \psi_2^2}$, then Stribeck (1907) defined an oil film parameter as $\lambda = {}^{h}/_{\psi}$. The regime of lubrication and an idea of the likely coefficient of friction are indicated by the Stribeck graph (Figure 11). Note that $\lambda > 3$ indicates a fluid film regime of lubrication, where insignificant boundary interactions would be expected. The coefficient of friction, μ, would normally be in the range 0.005–0.01, meaning that less than 1% of the input energy would be lost because of friction. At the other extreme, the lubricant film would be so thin as to allow significant direct contact of surfaces (boundary lubrication). The coefficient of friction would depend on the surface roughness and the shear strength of the counterface materials. For steel-on-steel contacts, the coefficient of friction would be in the range 0.2–0.3. It is clear that significant frictional power loss would ensue.

Viscous shear of a film of lubricant of dynamic viscosity η subjected to relative sliding motion of contact surfaces at velocity Δu and conjunctional pressure gradient ${}^{dp}/_{dx}$ is:

$$\tau = \pm \frac{h}{2}\frac{dp}{dx} + \frac{\eta \Delta u}{h} \tag{7}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto075
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

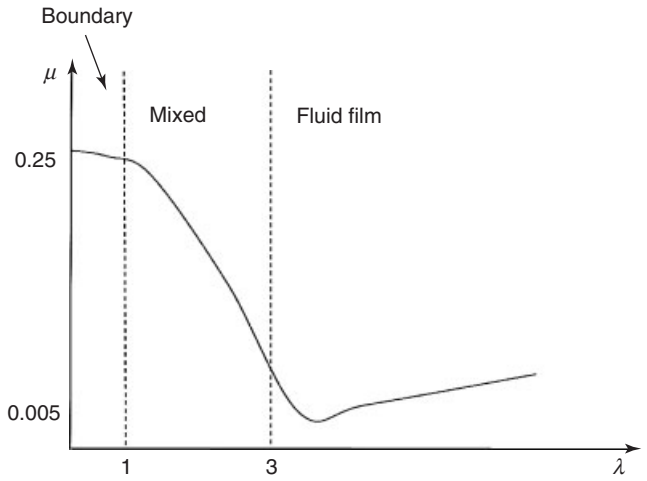


Figure 11. Stribeck graph for regimes of lubrication.

The first term in Equation 7 is often quite small compared with the latter in the flat region of the EHD film, as h is diminutive (a few tenths of micrometers) and the pressure gradient is only significant in the inlet wedge of the contact. Therefore, viscous friction is obtained as:

$$F_{\mathrm{v}} = \tau(A - A_{\mathrm{a}}) \tag{8}$$

where $A_{\mathrm{a}} = \pi^2(\zeta\kappa\psi)^2 A F_2(\lambda)$ is the area of asperity summits in contact in mixed regime of lubrication. $\zeta\kappa\psi$ is the roughness parameter (see Notations); a measure of surface topography assumed as a Gaussian distribution described by $F_2(\lambda)$ as a function of the oil film parameter, λ (Greenwood and Tripp, 1971).

Boundary friction occurs over the summit of opposing asperity pairs, where high load intensity (over a diminutive area) is assumed to induce non-Newtonian shear behavior of the lubricant at the limiting Eyring shear stress, τ_0. For most engine oils: $\tau_0 = 2{-}3$ MPa, for transmission fluids: $\tau_0 = 4{-}5$ MPa, and for differential oils: $\tau_0 = 5{-}7$ MPa. Boundary friction is obtained as:

$$F_{\mathrm{b}} = \tau_{\mathrm{o}} A_{\mathrm{a}} + \xi W_{\mathrm{a}} \tag{9}$$

where ξ is the pressure-induced coefficient of boundary shear strength of the softer of the counterfaces. For steel-on-steel contact: $\xi = 0.17$. The share of load carried by the asperities is

$$W_{\mathrm{a}} = \frac{8\sqrt{2}}{15}\pi(\zeta\kappa\psi)^2\sqrt{\frac{\psi}{\kappa}}E^* A F_{5/2}(\lambda) \tag{10}$$

Therefore, the total friction can be obtained as:

$$F_{\mathrm{T}} = F_{\mathrm{b}} + F_{\mathrm{v}} \tag{11}$$

The coefficient of friction is found from the same relationship originally noted by Amontons (1699). Referring back to Figure 11, various analyses have shown that typically: $\mu = 0.005$ (Fluid film lubrication) $- 0.09$ (mixed lubrication) for ball and rolling element bearings as well as cam–follower pairs and $\mu = 0.005$ (Fluid film lubrication) $- 0.15$ (boundary lubrication) for piston rings.

This chapter has provided an introduction to the important subject of contact mechanics, with emphasis put on concentrated counterformal or partially conforming contacts. Interested readers should refer to more in-depth texts such as Rahnejat (2010), Gohar and Rahnejat (2008), and Johnson (1985).

NOTATIONS

A	Apparent (nominal) contact area
A_{a}	Asperity contact area
a, b	Hertzian contact radii
E	Young's modulus of elasticity
E^*	Equivalent plane strain modulus of elasticity
e_p^*	Ellipticity ratio
F	Friction
F_{b}	Boundary friction
F_{T}	Total (mixed regime of lubrication) friction
F_{v}	Viscous friction
h	Film thickness
h_0	Central contact film thickness
h_{min}	Minimum (exit) film thickness
L	Contact length
p_{m}	Mean contact pressure
p_0	Maximum contact pressure
$R_{zx,zy}$	Equivalent radii of curvature in zx and zy planes of contact
$R_{x1,2}$	Principal radii of contacting solids in the zx-plane
$R_{y1,2}$	Principal radii of contacting solids in the zy-plane
u	Speed of lubricant entraining motion
W	Contact load
W_{a}	Load share of contacting asperities
α	Lubricant pressure viscosity coefficient
Δu	Sliding speed
δ	Contact deflection
η	Lubricant dynamic viscosity
κ	Average asperity summit (tip) radius
λ	Stribeck's oil film parameter
μ	Coefficient of friction
φ	Extent of contact
σ	Normal direct stress
τ	Shear stress
τ_0	Eyring shear stress

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto075
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

υ Poisson's ratio
ξ Pressure-induced coefficient of boundary shear strength
ψ Average surface roughness of contiguous solids
ζ Asperity density per unit contact area

REFERENCES

Amontons, G. (1699) De la résistance causée dans les machines. *Memoirs of the Royal Academy of Sciences*, 206–226.

De la Cruz, M., Chong, W.W.F., Teodorescu, M., *et al*. (2012) Transient mixed thermo-elastohydrodynamic lubrication in multi-speed transmissions. *Tribology International*, **48**, 17–29.

De la Cruz, M., Theodossiades, S., and Rahnejat, H. (2010) An investigation of manual transmission drive rattle. *Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics*, **224**, 167–181.

Dowson, D., Ehret, P., and Taylor, C.M. (1999) Past, present and future studies in elastohydrodynamics. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, **213**, 317–333.

Dowson, D. and Higginson, G.R. (1966) *Elastohydrodynamic Lubrication*, Pergamon Press, Oxford.

Gohar, R. and Rahnejat, H. (2008) *Fundamentals of Tribology*, Imperial College Press, London.

Greenwood, J.A. and Tripp, J. (1971) The contact of two nominally flat rough surfaces. *Proceedings of the Institution of Mechanical Engineers*, **185**, 625–633.

Grubin, A.N. (1949) Contact stresses in toothed gears and worm gears in *Book 30 CSRI for technology and Mechanical Engineering*, DSRI Trans., Central Scientific Research Institute, Moscow.

Hamrock, B.J. and Dowson, D. (1977) Isothermal elastohydrodynamic lubrication of point contact. Part III: fully flooded results. *Transactions of the American Society of Mechanical Engineers, Journal of Lubrication Technology*, **99** (2), 264–276.

Hertz, H. (1881) Uber den kontakt elastischer körper. *Journal für die Reine und Angewandte Mathematik*, **92**, 156.

Jalali-Vahid, D., Rahnejat, H., Jin, Z.M., and Dowson, D. (2001) Transient analysis of isothermal elastohydrodynamic circular point contacts. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, **215**, 1159–1173.

Johns, P.M. and Gohar, R. (1981) Roller bearings under radial and eccentric loads. *Tribology International*, **14** (3), 131–136.

Johnson, K.L. (1985) *Contact Mechanics*, Cambridge University Press, Cambridge, UK.

Kushwaha, M., Rahnejat, H., and Jin, Z.M. (2000) Valve-train dynamics: a simplified tribo-elasto-multi-body analysis. *Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics*, **214**, 95–108.

Litvin, F.L., Fuentes, A., Fan, Q., and Handschuh, R.F. (2002) Computerized design, simulation of meshing, and contact and stress analysis of face-milled formate generated spiral bevel gears. *Mechanism and Machine Theory*, **37**, 441–459.

Mohammadpour, M., Theodossiades, S., and Rahnejat, H. (2012) Elastohydrodynamic lubrication of hypoid gear pairs at high load. *Proceedings of Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, **226** (3), 183–198.

Mostofi, A. and Gohar, R. (1983) Elastohydrodynamic lubrication of finite line contacts. *Transactions of the American Society of Mechanical Engineers, Journal of Lubrication Technology*, **106** (4), 598–604.

Pascal, B. (1653) *Treatise on the Equilibrium of Liquids*, *L'Académie des Sciences*, Paris.

Rahnejat, H. (ed.) (2010) *Tribology and Dynamics of Engine and Powertrain*, Woodhead Publishing, Cambridge, UK. ISBN 978-1-84569-361-9

Rahnejat, H., Johns-Rahnejat, P.M., Teodorescu, M., *et al*. (2009) A review of some tribo-dynamics phenomena from micro- to nano-scale conjunctions. *Tribology International*, **42** (11–12), 1531–1541.

Stribeck, R. (1907) Die Wesentliechen ichen Eigenschaften Gleit und Rollen Lager or: ball bearings for various loads. *Transactions of the American Society of Mechanical Engineers*, **29**, 420–463.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto075
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Synchronizers—Gear Change Process, Loads, Timing, Shift Effort, Thermal Loads, Materials and Tolerances

Syed T. Razzacki

Retired, Chrysler LLC, Auburn Hills, MI, USA

1 Introduction 1
2 Gear Change Process 1
3 Synchronizer Components: Design Features and Functionality 2
4 Significant Parameters and Derivations 3
5 Design Process 10
6 Friction and Ring Material Compatibility with Lubricant Oil 16
7 Performance Failure Conditions 16
Nomenclature 17
Related Articles 18
References 18
Further Reading 19

1 INTRODUCTION

In a manual, synchromesh transmission, friction clutches called *synchronizers* are used to synchronize the rotational speed of the transmission output shaft and the gear to be engaged to actualize and secure smooth and noiseless gear transition. The size and location of synchronizers in transmissions varies for passenger cars and trucks. Increasing trend toward higher engine power and higher engine speeds due to multivalves per cylinder in pass cars and larger engines in trucks necessitate higher shift efforts. Regardless, the driver still demands smooth shiftability. These conflicting expectations require greater efficiency from the synchronizer design. In order to meet these conflicting expectations larger size synchronizers as well as multicone synchronizers are selectively inserted between the tall ratio gears to efficaciously sustain greater loads and provide slick gear change.

In sizing and locating the synchronizers, it is important to ensure minimizing the effect of system inertia and relative speeds of the rotating components. It must be recognized that synchronizer endures incessant punishment, more so than any other transmission components, and is expected to continue to work flawlessly for the life of the vehicle. Additionally, characteristically different driving habits ranging from silky smooth to sportive neck-snapping shifts affect the synchronizer design parameters for performance.

2 GEAR CHANGE PROCESS

Engine develops maximum power at high speed and it is desirable to have greater power available for excellent acceleration. A device called *transmission* was developed carrying a series of gears on two parallel shafts with ratios varying in descending order. It is attached to the engine that lets the engine run at the speed of maximum power at any vehicle road speed.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2.1 Early concept transmission

In 1895, M. Emile Levassar designed a three-speed transmission for use in Panhard et Levassar automobile. It had no mechanism for smoother gear change. Considerable skill was required to change gears without damaging the gear teeth as the selected gear was moved axially along a splined main shaft to engage with a countershaft (or lay shaft) gear, which was already turning resulting in clash. Hence, these transmissions were called *clash* (or *crash*) *boxes* and were very inefficient from performance standpoint.

2.2 Constant mesh

The concept of constant mesh gears was introduced later with sliding dog teeth to engage the gears. The constant mesh gears are always rotating as a driving and driven set. Sliding dog teeth improved the gear change process as a skillful driver with coordinated hand and foot movement could achieve clash-free shift.

2.3 Synchromesh

Again the gears are in constant mesh; rotating as a driving and driven pair except here the synchronizers are inserted between the speed gears, hence synchromesh. The gears rotate freely all the time, an intended gear, engaged through the synchronizer, can only transmit the torque. The synchronizer acts as a friction clutch to bring the relative speed of the transmission gear, engine, clutch disk, and the output shaft instantaneously to zero and aligns teeth for smooth gear change. Hence, synchronizer provides essential mechanism for ease of gear change.

3 SYNCHRONIZER COMPONENTS: DESIGN FEATURES AND FUNCTIONALITY

The synchronizer assembly is selected and designed to meet the load requirements of a specific application. Single cone, double cone, and triple cone synchronizers are used in the modern transmissions with medium to high torque engines. Typically, a synchronizer assembly, inserted between two speed gear wheels, slidingly engages either one intended. Synchronizer assembly comprises the following components, and respective design features are briefly described.

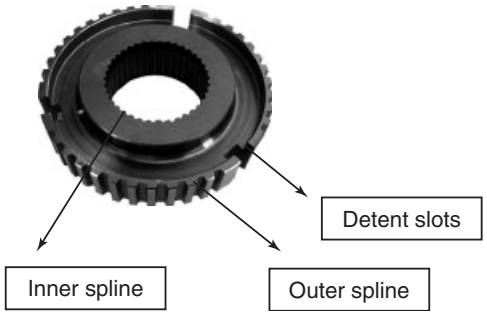


Figure 1. Synchronizer hub.

3.1 Hub

Hub has inner and outer splines; it is rotationally connected to the output shaft, spline fitted with internal splines. Three equally spaced slots are designed on the OD for housing the detent assemblies (Figure 1).

3.2 Sleeve

Sleeve has inner splines, slidingly fitted to the external splines of the hub, and travels on synchronization to lock with the clutching teeth of the intended gear. It has three annular grooves on the internal splines, equally spaced, for housing the three detents. An annular recess around the outer surface is provided for shift fork legs that instrumentally slide the sleeve to left or right when activated by the external applied force at the gearshift lever for consequent gear engagement. Traditionally, three splines at each of the three strut pockets are back tapered for a suck in engagement feel as well as anti-jump-out mechanism (Figure 2).

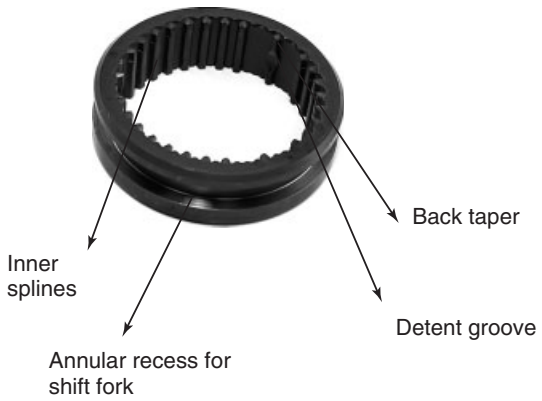


Figure 2. Synchronizer sleeve.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.3 Strut detents

Three strut detent assemblies per synchronizer are designed; the detent strut bumps or balls are nested in the three annular grooves of the sleeve in neutral position. These are primary energizing elements and provide initial indexing load for the friction (blocker) ring and set it ready for oncoming sleeve.

3.4 Friction (blocker, or baulk) ring

Bronze friction rings with oil wiping threads, used in the earlier design and still in vogue, having three equally spaced slots for the strut detent to push on. Various design and manufacturing developments have occurred to tackle the higher loads. Powder metal rings are being used coated with different robust friction materials for adequate and stable coefficient of friction at specified shift force. Friction material durability and efficiency are also significant considerations (Figure 3).

3.5 Friction cone

Friction cone could be integral part of the gear or separate slottedly locked in the gear. It generates cone torque required for synchronization when the friction ring surrounds it in full contact. The surface finish, roundness, and straightness are essential design requirements to ensure achieving the cone torque intended by design.

3.6 Gear locking (clutching) teeth

The sleeve traverses from initial stage, pushing the friction ring, which generates the friction torque and subsequently cone torque when synchronization is complete, and sleeve passes freely through the blocker ring and is sucked in by the gear locking teeth to engage the gear. The chamfer angle of the clutching teeth is designed 1° or 2° less than that of the sleeve pointing angle for easy engagement.

3.7 Spline chamfer angle

Sleeve spline and friction ring teeth pointing chamfer angles are designed the same. It is desirable to design clutching teeth chamfer 1° or 2° less than the sleeve chamfer angle to urge the sleeve in for gear lockup. The sleeve and blocker ring chamfers contact to generate index torque, whereas sleeve and clutching teeth chamfers contact to index the gear for engagement.

4 SIGNIFICANT PARAMETERS AND DERIVATIONS

Smooth, noiseless, and flawless gear change are the design objectives of the synchronizer. In order to achieve these objectives, synchronizer torque analysis will be conducted by addressing certain significant physical parameters.

4.1 Breakthrough load (BTL) and proximity

Also known as *push through load*, breakthrough load (BTL) effectively sets the blocker ring into block position. The BTL starts to build as soon as the force applied at the shift lever initiates sleeve movement via shift fork and stays on until the sleeve tooth chamfer contacts the blocker ring tooth chamfer. Looking at the graph in Figure 4, the following observations can be made:

- axial distance from sleeve tooth pointing to the blocker ring tooth pointing contact is called *proximity;*
- BTL drop-off short of proximity will unload blocker ring too soon and inhibit oil wiping process resulting in gear clash;

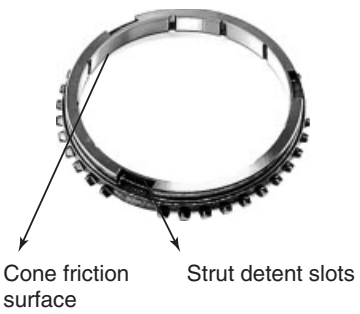


Figure 3. Synchronizer friction ring.

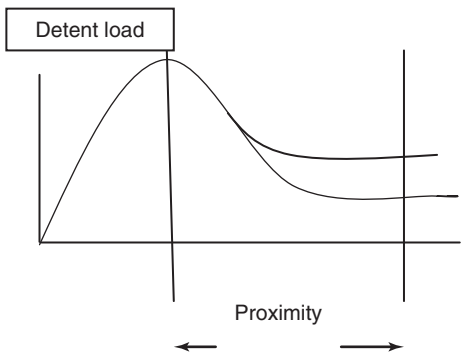


Figure 4. Detent load versus proximity.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- BTL prolonging beyond proximity will cause ring sticking; and

BTL is a function of detent spring rate, strut bump or ball height, coefficient of friction between detent ball and the sleeve, and the ramp angle of the annular groove in the sleeve.

4.2 Ball detent analysis

Ball detent used extensively in modern synchronizers is the subject of calculations later. Mathematically analyzing the forces on one of the three ball detents, BTL can be calculated from the following derivations.

As shown in Figure 5, taking sum of forces on the sleeve in x- and y-direction

$$\begin{aligned} F_A &= N_S \sin\theta + f_S \cos\theta \\ f_S &= \mu N_S \\ F_A &= N_S(\sin\theta + \mu\cos\theta) \end{aligned} \tag{1}$$

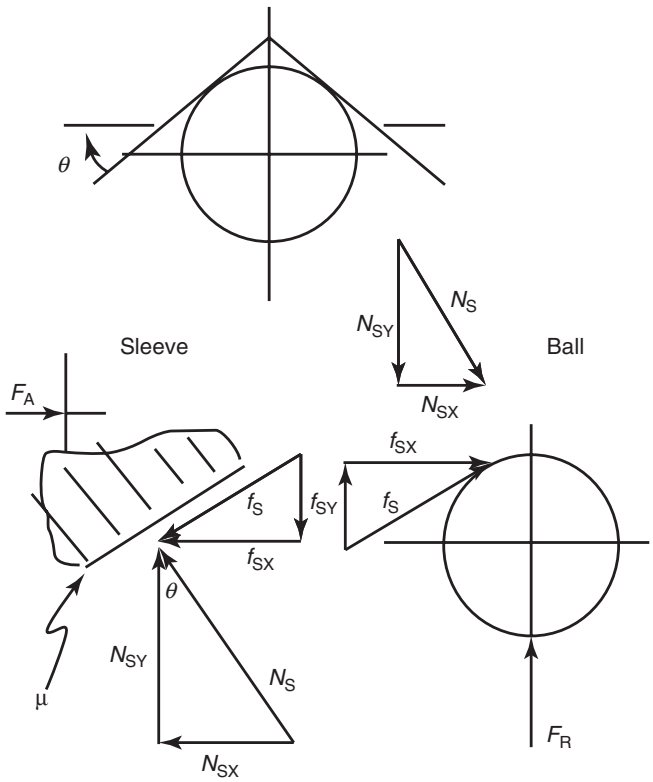


Figure 5. BTL—free body diagram. (Reproduced from Razzacki, 2004. Copyright © SAE International. Reprinted with permission.)

taking sum of forces on ball

$$\begin{aligned} F_A &= N_S \sin\theta + f_S \cos\theta \\ &= N_S \sin\theta + \mu N_S \cos\theta \\ F_R + f_S \sin\theta &= N_S \cos\theta \\ F_R &= N_S(\cos\theta - \mu\sin\theta) \end{aligned} \tag{2}$$

Substituting for N_s from Equation 1 in Equation 2,

$$\begin{aligned} F_R &= F_A \frac{\cos\theta - \mu\sin\theta}{\sin\theta + \mu\cos\theta} \\ F_A &= F_R \frac{\sin\theta + \mu\cos\theta}{\cos\theta - \mu\sin\theta} \\ &= F_R \frac{\mu + \tan\theta}{1 - \mu\tan\theta} \\ \text{BTL} &= 3 \times F_A \\ &= 3 \times F_R \frac{\mu + \tan\theta}{1 - \mu\tan\theta} \end{aligned} \tag{3}$$

By magnitude, the BTL should be smaller than the axial force applied at the sleeve groove, and too low could create clash condition. From experience 9–11 lbf (around 45N), BTL is sufficient to start ring indexing.

4.3 Cone torque

The chamfer of the oncoming sleeve tooth contacts with the blocker ring tooth chamfer and in the process pushes the inner conical surface of the blocker ring axially on to the external conical surface of the gear. Consequently, the lubricant oil is wiped out and friction force is developed in the direction of the cone angle generating cone torque for synchronization. The cone torque is primarily a function of the axial force applied to the sleeve, the cone angle, the surface coefficient of friction, and active cone diameter. Cone torque is calculated from the following equation, refer Figure 6:

$$T_C = \frac{F \times \mu_C \times R}{\sin\alpha} \tag{4}$$

The cone torque is countered by the index torque. The cone torque must be greater in magnitude to overcome the index torque to successfully complete synchronization,

$$T_C \geq T_I \tag{5}$$

4.4 Index torque

When the synchronizer ring is indexed and the sleeve has traversed the proximity distance, the sleeve pointing

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

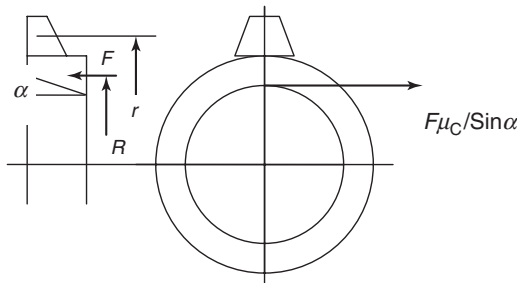


Figure 6. Cone torque—free body diagram. (Reproduced from Razzacki (2004). Copyright © SAE International. Reprinted with permission.)

contacts the blocker ring pointing and a friction force is generated between the two chamfer surfaces. This friction force is in the direction of pointing angle resulting in what is known as *index torque*. The index torque is a function of axial force applied to the sleeve, the tooth pointing chamfer angle, the pitch diameter of the blocking teeth, and surface coefficient of friction between the tooth chamfer surfaces of sleeve and blocker ring. The index torque is calculated from the following derivations (refer Figure 7):

$$T_I = F_I \times r$$

Summation of forces in *x*-direction on sleeve:

$$F_I = N_S \cos\beta - f_S \sin\beta = N_S(\cos\beta - \mu_p \sin\beta)$$

Summation of forces in *y*-direction on sleeve:

$$F = N_S(\sin\beta + \mu_p \cos\beta)$$

$$N_S = \frac{F}{\sin\beta + \mu_p \cos\beta}$$

$$F_I = F\frac{\cos\beta - \mu_p \sin\beta}{\sin + \mu_p \cos\beta}$$

$$T_I = F \times r\frac{\cos\beta - \mu_p \sin\beta}{\sin\beta + \mu_p \cos\beta}$$

$$= F \times r\frac{1 - \mu_p \tan\beta}{\mu_p + \tan\beta} \quad (6)$$

In the inequality 5, substituting for T_C from Equation 4 and for T_I from Equation 6 gives

$$\frac{F \times \mu_C \times R}{\sin\alpha} \geq F \times r \times \frac{1 - \mu_p \tan\beta}{\mu_p + \tan\beta} \quad (7)$$

Inequality 7 can be simplified to

$$\tan\beta \geq \frac{\frac{r}{R} - \mu_p \frac{\mu_C}{\sin\alpha}}{\frac{\mu_C}{\sin\alpha} + \mu_p \frac{r}{R}} \quad (8)$$

It can be observed that the inequality above has four interdependent significant synchronizer parameters. Treating this as equation nomograms have been developed to help size, select, and verify the parameters of synchronizer for a given application. The nomograms are shown in Figures 8 and 9a and b and explained in Section 4.6.

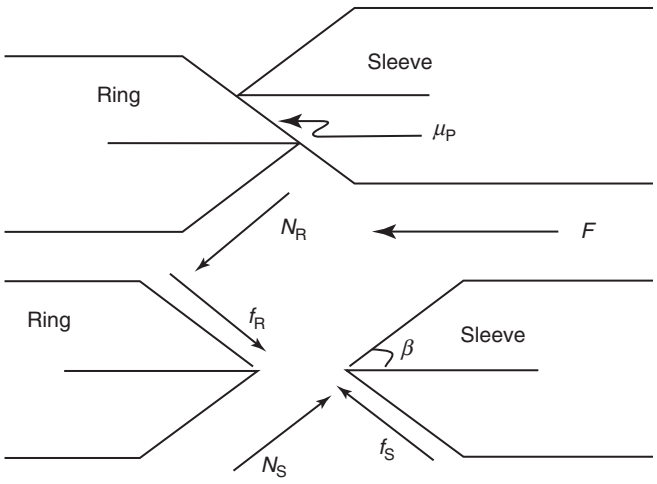


Figure 7. Index torque—free body diagram. (Reproduced from Razzacki, 2004. Copyright © SAE International. Reprinted with permission.)

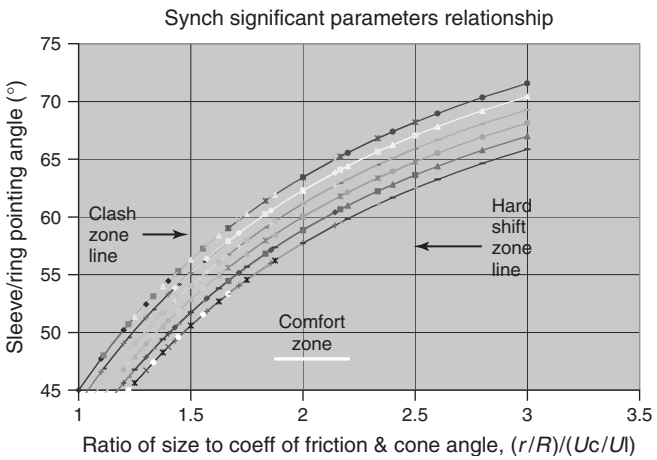


Figure 8. Significant parameter relationship. (Reproduced from Razzacki, 2004. Copyright © SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

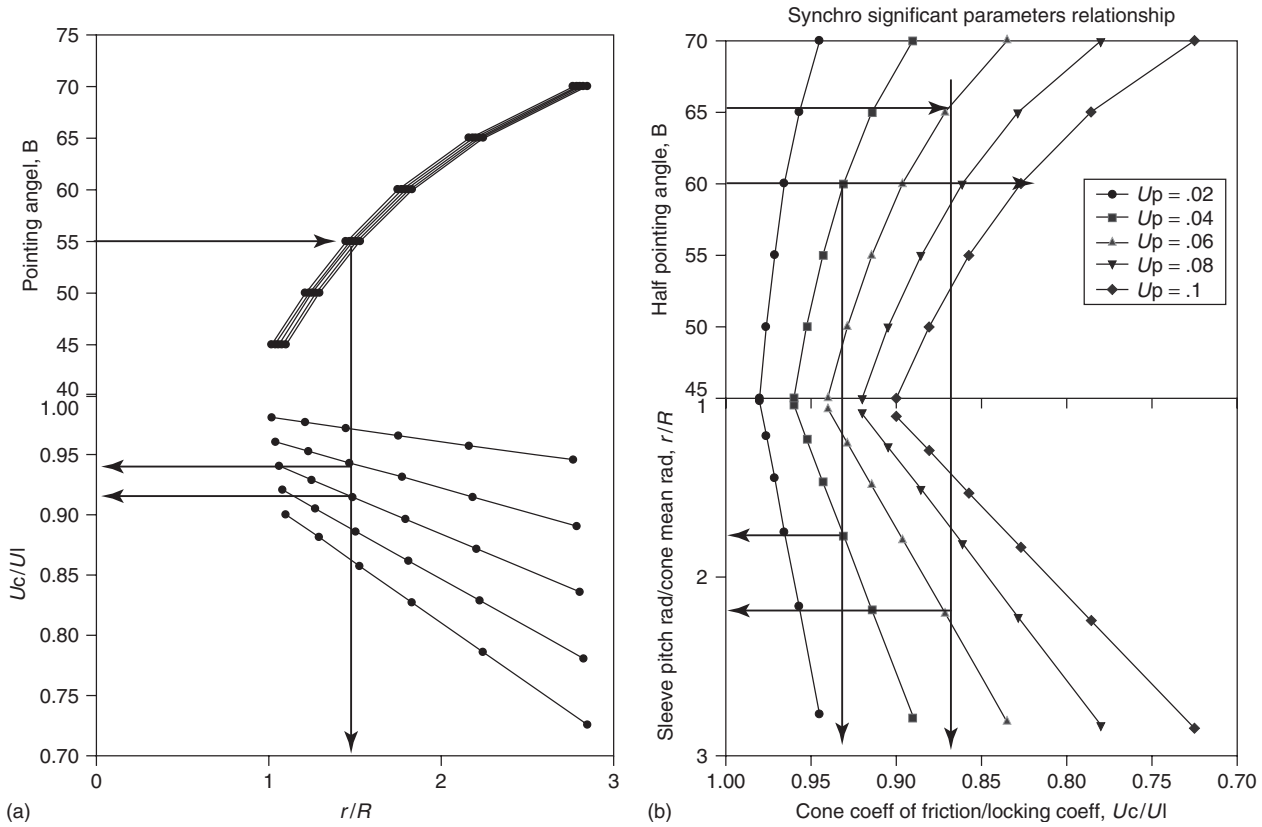


Figure 9. (a) Pointing angle and coefficient of friction versus size. (b): Pointing angle and size versus coefficient of friction. (Reproduced from Razzacki, 2004. Copyright © SAE International. Reprinted with permission.)

4.5 (Hat type) strut detent analysis

The strut detent arrangement in the initial stage of synchronizers design consisted of a hat type strut actuated with two circular springs, one on each side. This configuration is still conveniently and suitably useful and design analysis for this scheme of strut design is explored below.

4.5.1 Free body diagrams

The three figures in this section show detail of various forces acting to develop the analysis. Figure 10 shows the interface between sleeve pointing and ring pointing in a radial sense. Figure 11 shows a circumferential view of the strut contact with the ring, and Figure 12 shows the

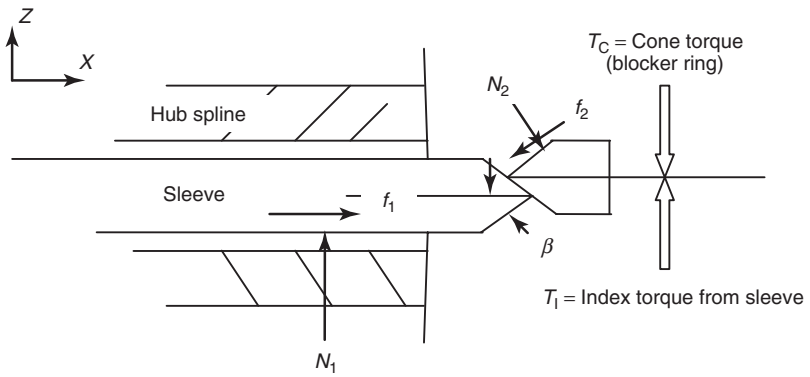


Figure 10. Sleeve chamfer meets the ring chamfer.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

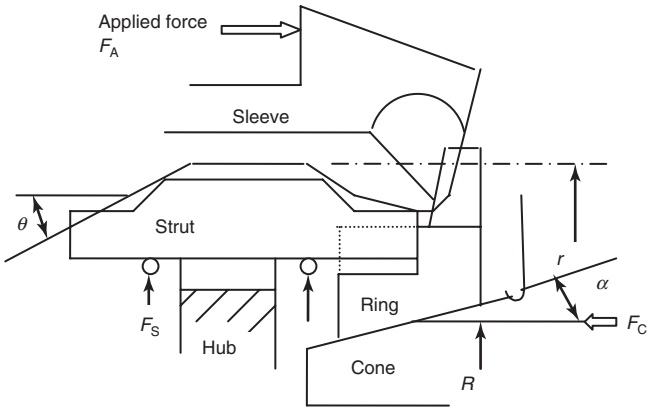


Figure 11. Sleeve loading the strut to index the blocker ring.

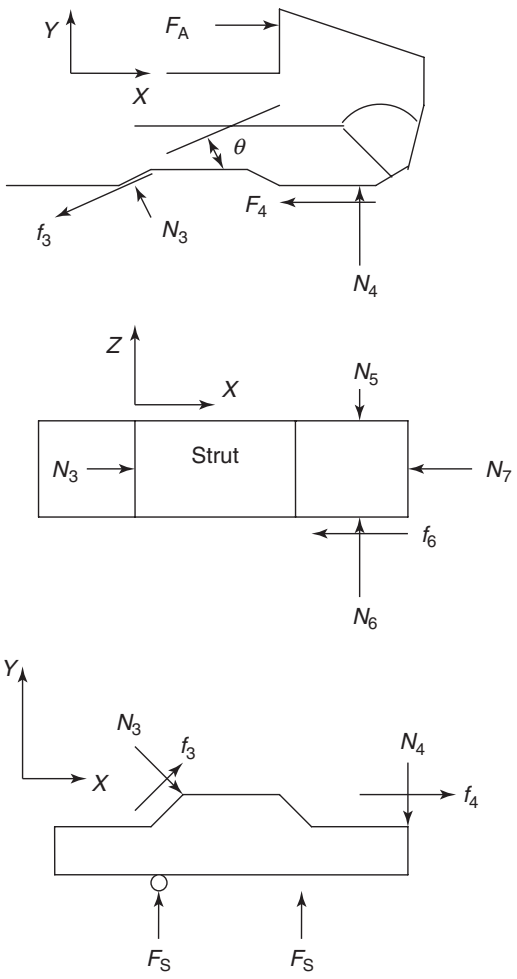


Figure 12. Sleeve and strut free body diagrams.

appropriate free body diagrams

$$T_C = \frac{F_C \times \mu_C \times R}{\sin\alpha}$$

$T_S = T_C - T_I$ where T_S is torque reaction on strut.

4.5.2 *Free body diagram equilibrium equations*

$$f_1 = \mu_{SH} N_1$$
$$f_2 = \mu_2 N_2$$
$$f_3 = \mu_3 N_3$$
$$f_4 = \mu_4 N_4$$
$$f_5 = 0$$

(No sliding after BTL developed)

$$f_6 = \mu_6 N_6$$

4.5.3 *Summation of forces on sleeve*

Summation in x-direction:

$$\begin{aligned} F_A &= f_1 + f_2\cos\beta + N_2\sin\beta \\ &\quad + f_3\cos\theta + N_3\sin\theta + f_4 \\ F_A &= \mu_{SH}N_1 + N_2(\sin\beta + \mu_2\cos\beta) \\ &\quad + N_3(\sin\theta + \mu_3\cos\theta) + \mu_4 N_4 \end{aligned} \tag{9}$$

Summation in z-direction:

$$\begin{aligned} N_2\cos\beta &= N_1 + f_2\sin\beta = N_1 + \mu_2 N_2\sin\beta \\ N_1 &= N_2(\cos\beta - \mu_2\sin\beta) \end{aligned} \tag{10}$$

4.5.4 *Summation of forces on strut*

Summation in x-direction:

$$\begin{aligned} N_3\sin\theta + f_3\cos\theta + f_4 &= f_6 + f_7 \\ N_3(\sin\theta + \mu_3\cos\theta) + \mu_4 N_4 &= \mu_6 N_6 + N_7 \end{aligned} \tag{11}$$

where $N_7 \geq 0$

Summation in z-direction:

$$N_5 = N_6$$

Summation in y-direction:

Let F_{ST} = total radial strut load

$K = \frac{N_4}{F_{ST}}$ = Percent of strut load, strut not in detent

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

$$F_{ST} + f_3 \sin\theta = N_3 \cos\theta + N_4$$
$$N_3(\cos\theta - \mu_3 \sin\theta) = F_{ST}(1 - K)$$
$$N_3 = \frac{F_{ST}(1 - K)}{\cos\theta - \mu_3 \sin\theta} \quad (12)$$

4.5.5 *Summation of forces and torques on ring*

Summation of forces in *x*-direction:

$$F_C = N_2 \sin\beta + f_2 \cos\beta + N_7$$
$$F_C = N_2(\sin\beta + \mu_2 \cos\beta) + N_7$$
$$= \text{resultant axial load on cone} \quad (13)$$

Summation of torques:

$$T_C = T_S + T_I = \frac{\mu_C F_C R}{\sin\alpha}$$
$$F_I = N_2 \cos\beta - f_2 \sin\beta$$
$$F_I = N_2(\cos\beta - \mu_2 \sin\beta)$$
$$T_I = R_P N_2(\cos\beta - \mu_2 \sin\beta) \quad (14)$$

where T_I is the index torque.

4.5.6 *Breakthrough load (BTL)*

$$\text{BTL} = N_3(\sin\theta + \mu_3 \cos\theta) + \mu_4 N_4 \quad (15)$$

Substituting Equation 12 in Equation 15:

$$\text{BTL} = \frac{F_{ST}(1 - K)}{\cos\theta - \mu_3 \sin\theta}(\sin\theta + \mu_3 \cos\theta) + \mu_4 K F_{ST}$$
$$\text{BTL} = F_{ST}\left[\mu_4 K + \frac{(1 - K)(\mu_3 + \tan\theta)}{1 - \mu_3 \tan\theta}\right] \quad (16)$$

Solving for normal forces: substituting Equations 10 and 12 in Equation 9

$$F_A = N_2[(\mu_{SH} + \mu_2)\cos\beta + (1 - \mu_{SH1}\mu_2)\sin\beta] + F_{ST}\left[\mu_4 K + \frac{(1 - K)(\sin\theta + \mu_3 \cos\theta)}{\cos\theta - \mu_3 \sin\theta}\right]$$

Substituting for N_2 from above equation in Equation 14 for index torque:

$$T_I = R_P\left[\frac{1 - \mu_2 \tan\beta}{\mu_{SH} + \mu_2 + (1 - \mu_1\mu_2)\tan\beta}\right]\left[F_A - F_{ST}\left(\mu_4 K + \frac{(1 - K)(\mu_3 + \tan\theta)}{1 - \mu_3 \tan\theta}\right)\right] \quad (17)$$

4.6 The nomograms

The nomogram in Figure 8 depicts the relationship of pointing chamfer angle with the size of the synchronizer, cone coefficient of friction, and cone angle for a given μ_p. This relationship resulted from the necessary condition in (inequality) 5 and the algorithms derived from it in (inequality) 8, and it can be observed that smaller the size to cone friction ratio smaller the pointing chamfer angle for a given μ_p and together would result in clash. On the other hand, bigger the ratio bigger the pointing chamfer angle and the combination would result in hard shift. Again, from computations based on inequalities 5 and 8, plotted in the nomogram in Figure 8, it is clear that for a given μ_p, size to coefficient of friction ratio above 2.5 could result in hard shift and below 1.5 in clash and comfortable shiftability between the two.

Figure 11a and b shows the graphical representation of the same relationship as in Figure 8 except here all four parameters are separately charted. These charts elicit that, for a given μ_p, greater the *r*/*R* ratio greater the pointing angle resulting in hard shift, whereas smaller the *r*/*R* ratio smaller the pointing angle resulting in clash. Any of the three nomograms can be used conveniently as desired.

4.7 System drag consideration

Excessive losses in the transmission including clutch drag, frictional, and fluid churning losses adversely affect the synchronization process. The system drag assists up shift by slowing the gear train and resists speeding up the gear train during downshift. It is hence understood that the index torque should be designed to compensate for the total drag. Applying the basic principle of synchronizer design, the synchronization torque must be greater than the index torque at every instant during the synchronization event. Therefore, synchronizer design, be it for conventional manual transmission, or the modern dual clutch (DCT) transmission, specially the wet clutch DCT, must satisfy the following two criteria.

1. Cone torque (T_C) must be greater than index torque (T_I) at every point during synchronization event to prevent clash or grating noise, that is, $T_C > T_I$.
2. Index torque must be high and able to shift in cold conditions to overcome the total system drag (T_D) to insure gear shiftability, that is, $T_I > T_D$.

The above-mentioned two requirements establish the boundary limits for index torque as follows. The index torque derivation shows that it is inversely proportional

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to sleeve/ring pointing angle, and higher index torque can be obtained by designing steeper pointing chamfer angle. Imperatively at the same time attention must be paid to maintain synchronization torque always greater than index torque by tweaking the cone coefficient of friction up as needed, which might warrant larger cone angle as well.

Total drag torque (T_D) is sum of clutch drag, fluid churning, and system friction losses:

$$T_D = T_d + T_C + T_f \tag{18}$$

Clutch drag torque is an uncertain quantity and its directional values should be calculated from existing empirical formulae, such as

$$T_d = 6.6 \times 10^{-13} \frac{\eta \times \Delta n \times F_N \times R_F^3 \times b}{l_F} \tag{19}$$

Subsequently, laboratory tests should be conducted to measure clutch drag torque at various speeds and temperatures for correlation with calculated values and most importantly to assist in selecting adequate pointing angle and cone coefficient of friction in the initial stage of design.

Torque due to fluid churning (T_{Ch}) is a function of drive pinion speed and a constant k dependent on viscosity, quantity, and temperature of oil:

$$T_{Ch} = k \times N_i \tag{20}$$

Torque due to frictional losses (T_f) consists of bearing friction, shift and clutch actuation mechanism bending and deflection, and all other system friction in the transmission.

The objectives of fuel economy and associated smooth shifting are largely dependent on clutch drag and transmission system efficiency: lower the drag higher the efficiency and higher the fuel economy and smoother the gear transition.

4.8 Synchronizer location and sizing

In a synchromesh, all gears are turning hence the inertia of all gears must be overcome in order to make a gear selection. It is advantageous to locate the 1/2 synchronizers on the intermediate shaft, whereas the other synchronizers on the input shaft as a best compromise from the stand point of reflected inertias and relative speeds of the rotating components. Vehicle size, engine power output, and number of transmission ratios and steps are required for synchronizer analysis. Synchronization torque and synchronization time relationship is crucial in sizing the synchronizer. Given synchronization time, which is dependent upon initial velocity of the output synchronizer and gear ratio step, its relationship to synchronization torque can be derived by solving differential equation of motion for both input side and output side. In a dual clutch transmission, ratio steps for skip shift would be higher and should be considered for upshifts and downshifts. The solution to the differential equation at the completion of synchronization is

$$t = \frac{\omega I_C (n-1)}{T_C \pm T_D} \tag{21}$$

Synchronization torque requirements can be calculated for a given synchronization time with known gear ratios and steps. The total drag that slows the input must be factored in the calculations as it assists the upshift and resists the downshift.

Next important factor in sizing the synchronizer is the total energy dissipation and rate of energy dissipation and can be calculated as follows: Total energy

$$E = \pm \frac{T_C \omega (n-1)}{2} \times t \tag{22}$$

Rate of energy dissipation

$$\frac{E}{t} = \pm \frac{T_C \omega (n-1)}{2} \tag{23}$$

Rate of energy dissipation per unit area of synchronizer cone is used to size the cone width and cone mean radius as indicated in Equation 24:

$$E/t/A = \frac{T_C \omega (n-1)}{4\pi R_r W_C} \tag{24}$$

4.9 Synchronization time (*t*)

200 ms is typically used in layshaft transmissions. Following observations can be made from the calculations above:

- Synchronization time is directly proportional to speed differential and inertia and inversely proportional to synchronization torque, hence higher the axial apply force shorter the synchronization time.
- Energy dissipation is directly proportional to synchronization torque, which is proportional to axial force, hence higher the axial force higher the energy dissipation. Assuming synchronization torque constant through the shift, the speed differential becomes overriding factor. As an example, a family of curves was generated for specific application graphically representing energy dissipation per unit area for miscellaneous shifts and axial apply forces (Figure 13).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

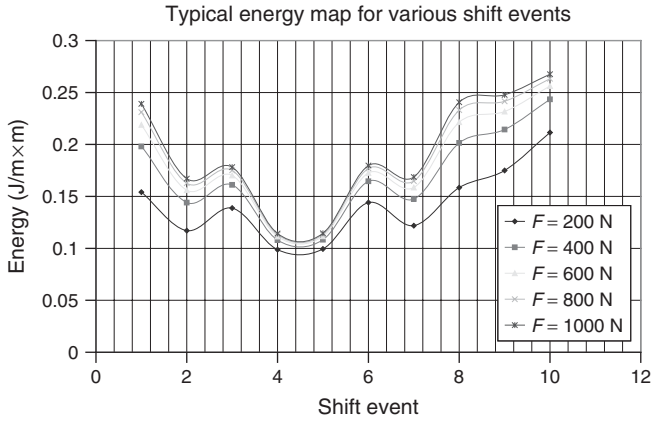


Figure 13. Rate of energy dissipation per unit area versus misc shift events. (Reproduced from Razzacki and Hottenstein, 2007. Copyright © SAE International. Reprinted with permission.)

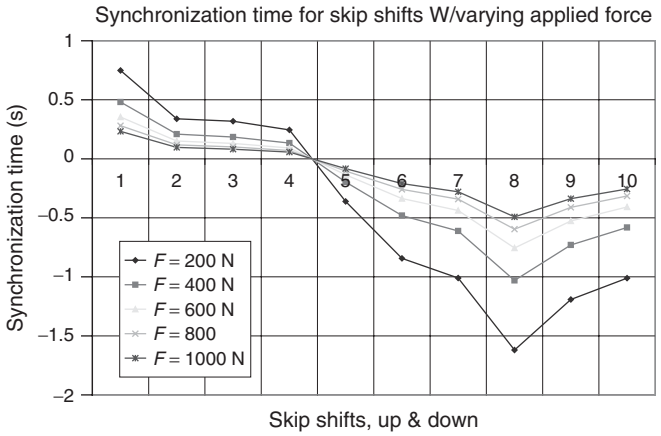


Figure 14. Synchronization time for misc shifts at various axial forces. (Reproduced from Razzacki and Hottenstein, 2007. Copyright © SAE International. Reprinted with permission.)

- For a known drag torque value, the axial force can be determined from synchronization torque/time relationship for upshifts, downshifts, and skip shifts. As an example, a family of curves was generated for a specific application graphically representing synchronization time for miscellaneous shifts for various values of axial apply force (Figure 14).
- Based on calculated synchronization torque, the synchronizer diameter and width can be established to package axially and radially for a specific application.
- Multiple friction surfaces may have to be incorporated especially in positions with higher numerical gear ratios to suitably adjust to higher synchronization torque requirements and packaging constraints.

Figures 13 and 14 assist in recommending appropriate axial apply force based on compliance of synchronization time and energy dissipation criterions.

5 DESIGN PROCESS

An important feature of the synchronizer is the aspect ratio, which is the relationship between its diameter and the width. A wide synchronizer sleeve slides smoothly to meet the blocker ring and clutching teeth squarely. A favorable aspect ratio allows maintaining tighter clearances for ease of operation.

Heavier vehicles and high power engines effectively require greater shift time and effort. Increased synchronizing moment is required to mitigate the consequent events. Aspect ratio adequately selected in a synchronizer having single cone friction surface, with relatively high coefficient of friction, would sufficiently meet the demand. Theoretically, a multicone synchronizer with multiple friction surfaces would boost the synchronizing moment by a factor of root mean square of the sum of inner and outer cone radii as shown in the equation below. Hence, the effort required to synchronize the angular velocities of input and outputs as well as the engagement time are significantly reduced.

$$T_C = \frac{F \times \mu_C}{\sin\alpha}\sqrt{(R_I^2 + R_O^2)}$$

It is generally assumed that the cone torque multiplies by a factor of number of friction surfaces. However, the measure of each friction surface radius, intermediate and inner, is smaller from the one above it. Plus the clearance between each friction surface. These factors do not support the above assumption.

A typical dual cone synchronizer assembly is shown in Figure 15.

After selecting any of the arrangements described above, and the physical parameters of the synchronizer namely sleeve and blocker ring pointing chamfer angle, cone angle, cone coefficient of friction, and the size, the most critical step is to design, dimension, and tolerance the synchronizer components. The intended objective of the design process should be to dimension and tolerance the individual components in a manner such that along with the selected parameters the functional objectives are achieved satisfactorily. The process consists of charting the synchronization events and iteratively dimensioning, stacking, and tolerancing for the best results. Hat type strut is used through the process of dimensioning and tolerancing. The synchronization episode has been broken up into following six distinct events as shown in Figure 16. Each event is descriptively explained in the ensuing sections.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

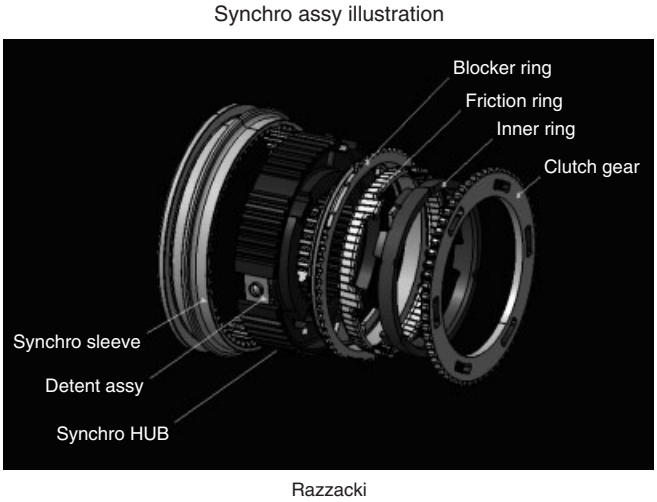


Figure 15. Dual cone synchronizer assembly.

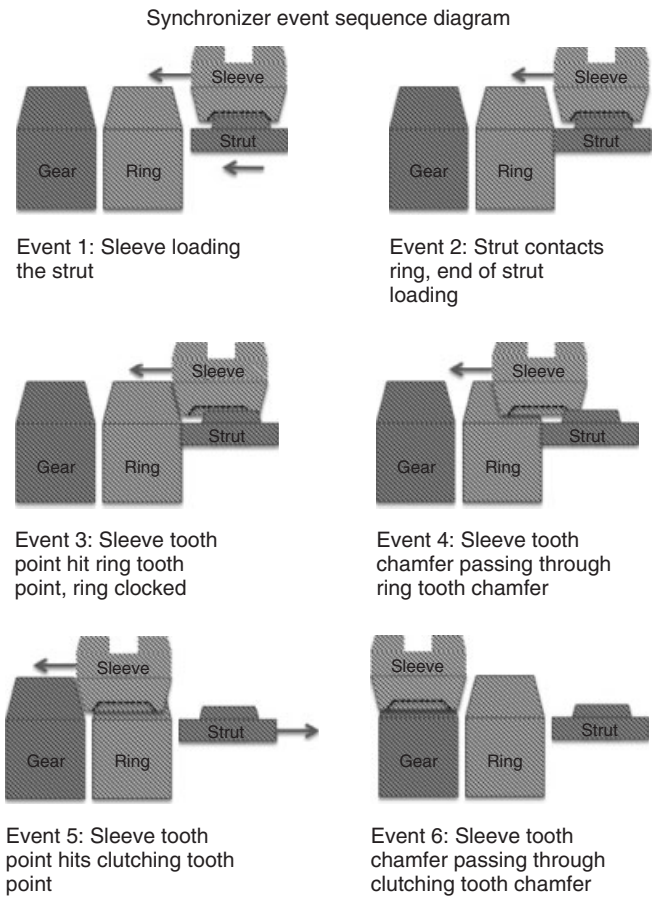


Figure 16. Synchronizer event sequence diagram.

5.1 Event I: strut contacts blocker ring

This is the starting point of the blocker ring energizing system, and the mechanism moves the blocker ring up to the gear cone prior to contact between the indexing teeth on the sleeve and the teeth on the blocker ring. The strut loading starts to generate BTL. What is also called *zero (0) point* will be the first contact of the strut on the ring. The zero point implies maximum strut length, maximum ring lug thickness, and maximum gage point offset. On the other extreme, the last contact point implies minimum strut length, minimum ring lug thickness, and minimum gage point offset. Hence, the total differences are as follows:

$$\text{Max} - \text{min strut length} = (L_{\text{ST max}} - L_{\text{ST min}})$$
$$\text{Max} - \text{min ring thickness} = (L_{\text{Rmax}} - L_{\text{Rmin}})$$
$$\text{Max} - \text{min gage point} = (G_{\text{max}} - G_{\text{min}})$$

Taking first contact point as zero point, the last contact will occur at a distance

$$(L_{\text{ST max}} - L_{\text{ST min}}) + (L_{\text{Rmax}} - L_{\text{Rmin}}) + (G_{\text{max}} - G_{\text{min}}) \quad (25)$$

The event is pictorially illustrated in Figure 17. The strut loading starts at the point of first contact, and earlier the loading begins the better.

The components involved in this event are sleeve, detent strut/ball, and detent spring.

5.2 Event II: end of strut loading, strut out of detent

This is the end of strut loading. The strut snaps back and the sleeve moves on toward the blocking ring. It is to be noted here that the detent load is a function of the ramp angle of the sleeve annulus groove and significantly influences the magnitude of BTL.

Detent profile is critical in achieving desirable BTL. As the desirable BTL has been calculated in Section 4.1, the

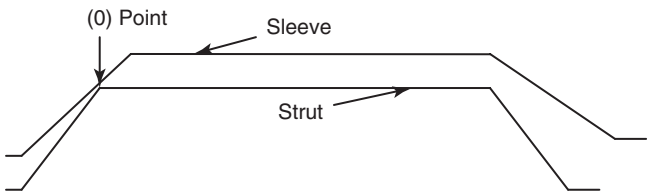


Figure 17. First contact, sleeve to strut. (Reproduced from Razzacki, 2004. Copyright © SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

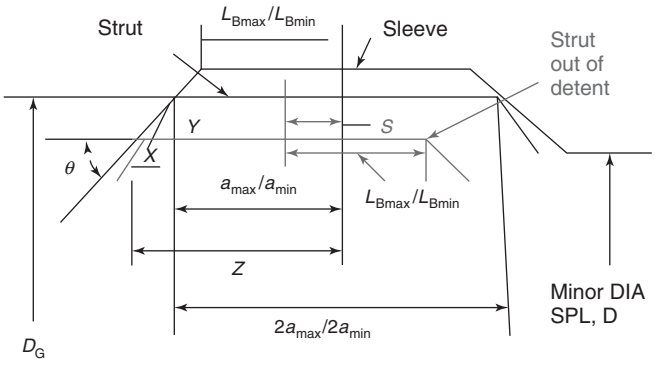


Figure 18. Strut detent profile. (Reproduced from Razzacki, 2004. Copyright © SAE International. Reprinted with permission.)

task at hand is to dimension the detent profile accordingly as shown in Figure 18.

Computing X & Y:

$$X = \frac{Y}{\tan\theta}$$

$$Y_{max} = \frac{D_G - D_{min}}{2}$$

$$Y_{min} = \frac{D_G - D_{max}}{2}$$

For X to be minimum ramp angle and minor diameter should be maximum, or Y minimum, hence

$$X_{min} = \frac{Y_{min}}{\tan\theta_{max}} = \frac{D_G - D_{max}}{2\tan\theta_{max}} \tag{26}$$

For X to be maximum ramp angle and minor diameter should be minimum, or Y maximum, hence

$$X_{max} = \frac{Y_{max}}{\tan\theta_{min}} = \frac{D_G - D_{min}}{2\tan\theta_{min}} \tag{27}$$

Computing groove width Z: for Z to be minimum gage dimension a and X should be minimum,

$$Z_{min} = a_{min} + X_{min} \tag{28}$$

For Z to be maximum gage dimension a and X should be maximum,

$$Z_{max} = a_{max} + X_{max} \tag{29}$$

The distance strut out of detent S can be found as follows:

$$S_{max} = Z_{max} - L_{B\,min} \tag{30}$$

$$S_{min} = Z_{min} - L_{B\,max} \tag{31}$$

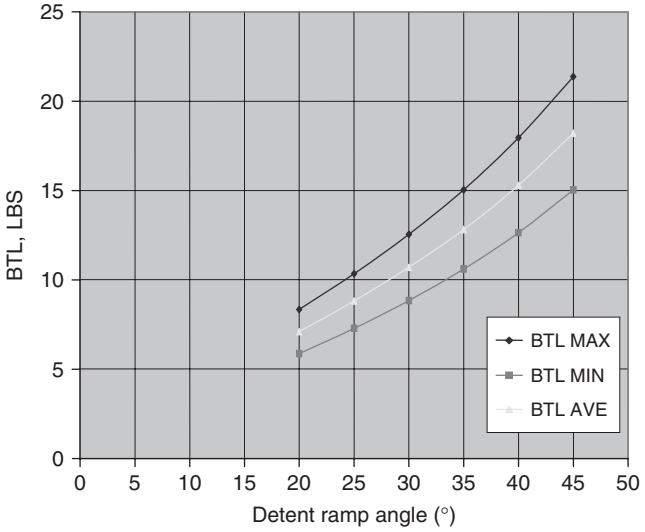


Figure 19. BTL versus sleeve groove ramp angle. (Reproduced from Razzacki, 2004. Copyright © SAE International. Reprinted with permission.)

Equations 24 through 30 can be used to design the detent profile for reasonable detent load to achieve the desired BTL. The sleeve groove ramp angle contributes significantly to the detent load and the BTL as such and it is illustrated in Figure 19. It can be observed that a ramp angle of 30° for a given application will yield the desired BTL.

The components involved in this event include sleeve, strut/ball, and detent spring.

5.3 Event III: sleeve point hits ring point, ring clocked

In this event for the sleeve point to hit the ring point the soonest, the gap between them should be the minimum. The minimum gap is obtained with maximum ring width, maximum strut, and maximum gage point offset at "0" point condition. Similarly, the maximum gap can be obtained with minimum ring width, minimum strut, and minimum gage point offset. Referring to Figure 20, the gap between sleeve tooth pointing and ring tooth pointing is called *proximity* =

$$(L_R + L_{ST} + L_B) - L_{SL} \tag{32}$$

The gap between the sleeve tooth pointing and ring tooth pointing is shown in Figure 20, Section A-A.

If sleeve and ring teeth have rake angle, then using trigonometry L_R will increase by a fraction and L_{SL} will diminish by a fraction affecting the proximity by a fraction too.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

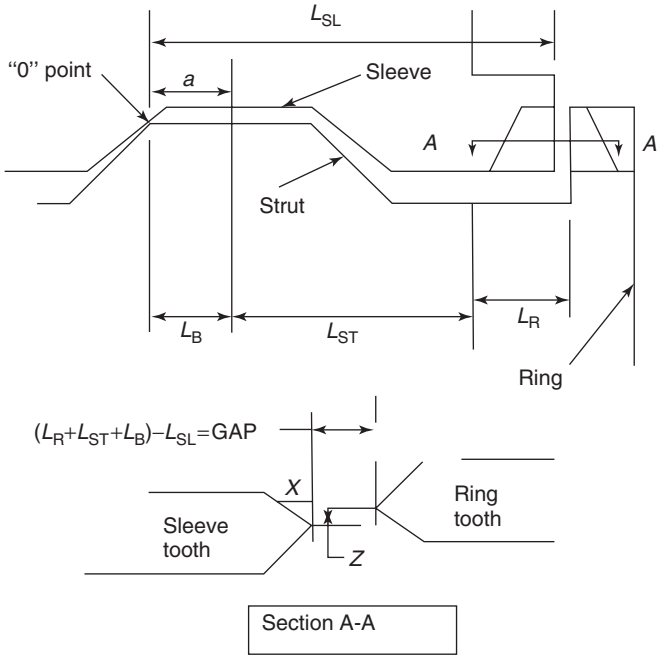


Figure 20. Proximity dimensioning. (Reproduced from Razzacki, 2004. Copyright © SAE International. Reprinted with permission.)

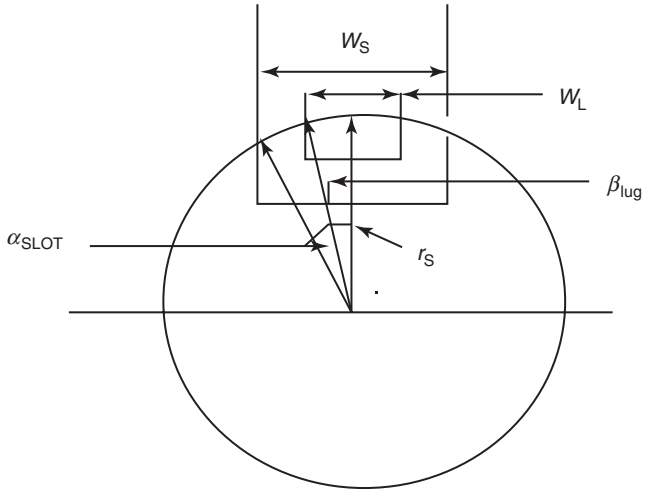


Figure 21. Clocking angle. (Reproduced from Razzacki, 2004. Copyright © SAE International. Reprinted with permission.)

During this event, as soon as the sleeve pointing contacts the ring pointing the blocker ring starts to clock and indexes with the oncoming sleeve. The clocking angle is a function of the widths of the lug integral to the ring and slot width in the hub. The lug and the slot widths should be dimensioned adequately and minimum maximum clocking angles should be calculated to insure that it is not too low allowing not enough time for BTL to develop and not too high when the ring would take too much time to set for the oncoming sleeve.

The clocking angle is calculated by applying trigonometry using the lug and slot widths and the radius at the lug as shown in Figure 21. Maximum clocking is obtained from maximum slot width, minimum lug width, and minimum radius.

$$\sin\alpha_{\mathrm{SLOT}} = \frac{W_{\mathrm{S}}}{2\times r_{\mathrm{S}}},$$

$$\alpha_{\mathrm{SLOT}} = \sin^{-1}\frac{W_{\mathrm{S}}}{2\times r_{\mathrm{S}}} \tag{33}$$

$$\sin\beta_{\mathrm{LUG}} = \frac{W_{\mathrm{L}}}{2\times r_{\mathrm{S}}}$$

$$\beta_{\mathrm{LUG}} = \sin^{-1}\frac{W_{\mathrm{L}}}{2\times r_{\mathrm{S}}} \tag{34}$$

Clocking angle

$$\psi = \alpha_{\mathrm{SLOT}} - \beta_{\mathrm{LUG}} \tag{35}$$

From experience, the clocking angle should be less than 4 but greater than 3° ($4>\psi>3$), and the lug and slot width should be dimensioned accordingly.

In Figure 20, Section A-A, the dimension Z between the tooth points is an arc. Angle between the center of a sleeve tooth and the center of a space is

$$\frac{360}{N+N} = \frac{180}{N}$$

$$\mathrm{AL}(180/N) = \frac{180}{N}\times r_{\mathrm{R}} \tag{36}$$

$$\mathrm{AL}(\psi) = \psi \times r_{\mathrm{R}} \tag{37}$$

In Equations 36 and 37, the angles $(180/N)$ and ψ are in radians.

For x to be small, the pointing angle should be maximum,

$$\tan\beta = \frac{z}{x}$$

$$z = \mathrm{AL}\left(\frac{180}{N}\right) - \mathrm{AL}(\psi) \tag{38}$$

Components involved in this event are sleeve, blocking ring, reaction cone, and friction ring.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5.4 Event IV: sleeve chamfer through ring chamfer

Given the pressure angle, the ring tooth thickness and the circular space width of sleeve spline can be calculated. Obtaining four values, maximum and minimum for each, they can be compared to determine the combination of tolerances at which the ring teeth thickness to sleeve space widths would have positive or negative clearance. The combination of tolerances can be selected that provide desirable fit and feasible manufacturability.

5.4.1 Ring tooth width calculation

Using the ring outer diameter (D_t) and tooth width at the pitch diameter (t_{DP}), two minimum and two maximum values of ring tooth width can be calculated (refer Figure 22).

The ring tooth width can be calculated by applying Equation 39 available in any gear design book:

$$\frac{t_t}{D_t} = \frac{t_{DP}}{D_P} + \text{INV}\varphi - \text{INV}\varphi_t,$$

or

$$t_t = D_t\left(\frac{t_{DP}}{D_P} + \text{INV}\varphi - \text{INV}\varphi_t\right) \tag{39}$$

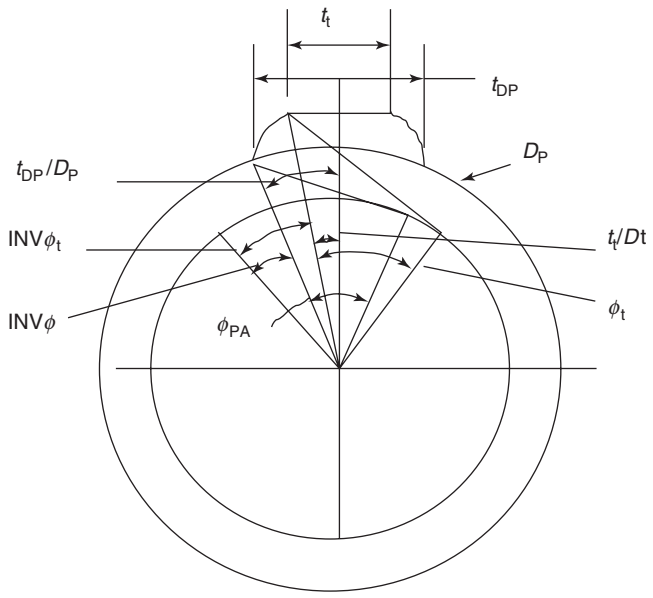


Figure 22. Ring tooth width illustration. (Reproduced from Razzacki (2004). Copyright © SAE International. Reprinted with permission.)

5.4.2 Sleeve spline circular space width calculation

Again, using Equation 39, the sleeve spline space width that would yield four values, two maximums and two minimums can be calculated. Comparing the values of ring tooth width with the sleeve spline space width, the combination of tolerances can be assessed that yield positive running clearance.

5.4.3 Sleeve tooth width calculation

Having calculated sleeve tooth space width, the tooth width can be calculated as follows:

$$t_{(\text{tooth+space})} = 1\ \text{tooth} + 1\ \text{space} = \frac{\pi D_t}{N_T} \tag{40}$$

The sleeve space width is determined in Section 5.4.2. Hence, the sleeve tooth width can be computed as follows:

$$t_S = t_{(\text{tooth+space})} - t_{\text{space}} \tag{41}$$

5.4.4 Total sleeve travel from "0" point through ring chamfer

Finally, it is possible to determine the distance traveled by the sleeve pointing chamfer from zero point to the ring pointing chamfer (Figure 23).

From Figure 23,

$$\begin{aligned}\tan\beta &= \frac{t_R}{2 \times a_R} \\ &= \frac{t_S}{2 \times a_S} \\ &= \frac{t_R + t_S}{2(a_R + a_S)}\end{aligned} \tag{42}$$

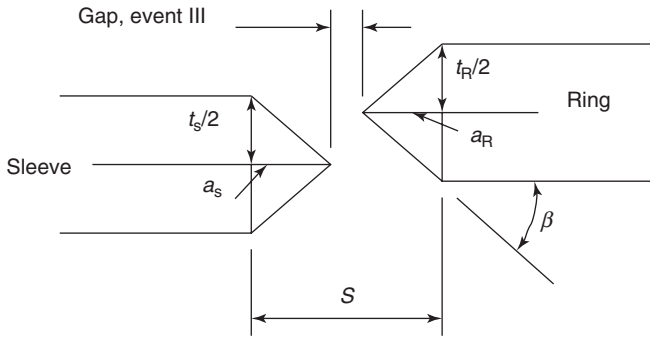


Figure 23. Sleeve chamfer through ring chamfer. (Reproduced from Razzacki (2004). Copyright © SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

or

$$a_R + a_S = \frac{t_R + t_S}{2 \times \tan\beta}$$

Total travel

$$S = \text{GAP} + a_S + a_R$$
$$= \text{GAP} + \frac{t_R + t_S}{2 \times \tan\beta} \quad (43)$$

Hence, for minimum distance traveled by sleeve in event IV,

$$S_{\min} = \text{GAP}_{\min} + a_{R\min} + a_{S\min} \quad (44)$$

For $S_{\min}$ use from event III the minimum GAP and from event IV, the values of t_R and t_S for conditions assigned for minimum values. Similarly, for $S_{\max}$ use from event III the maximum GAP and from event IV, the values t_R and t_S for conditions assigned for maximum values.

5.5 Event V: sleeve tooth point contacts clutching tooth point

Here again, it is possible to stack up dimensions to calculate the distance traveled by sleeve from zero point to meet the clutching tooth point (Figure 24), see the illustration in Figure 20.

The distance sleeve pointing has to travel from zero point to meet the clutch tooth pointing can be computed by stacking up the GAP in event III along with the ring and clutching tooth dimensions as follows:

$$P = \delta + G + L_{RW} + L_{ST} + L_B - L_{SL}$$
$$P = \delta + G + W_{RT} + \text{GAP} \quad (45)$$

For $P_{\min}$, the dimension L_{SL} maximum and all other dimensions minimum will be required. For $P_{\max}$, the dimension L_{SL} minimum and all other dimensions maximum will be required.

5.6 Event VI: sleeve tooth chamfer through gear clutching tooth chamfer

Clutching teeth chamfers traditionally are cut, whereas the modern gear forging technology is producing as-forged clutching teeth with rounded chamfers for smooth operation. Such forgings are called *monoblocks*.

This is the final event when the sleeve travels from the zero position, and its chamfer passes the gear clutching tooth chamfer to complete the gear engagement. During this event, the blocker ring is completely unloaded and freely gets back to zero position marking the end of cone

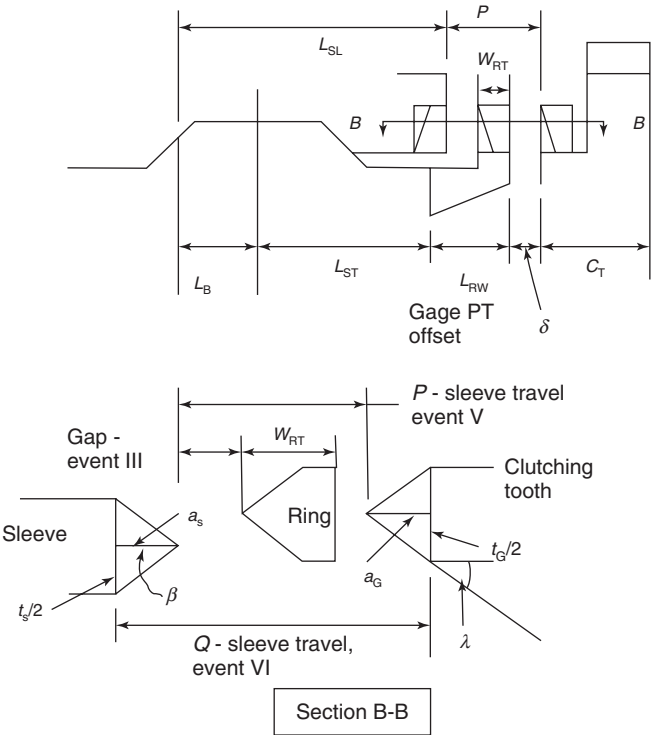


Figure 24. Distance P, sleeve tooth pointing to clutching tooth pointing. (Reproduced from Razzacki (2004). Copyright © SAE International. Reprinted with permission.)

torque. The total distance traveled by sleeve pointing from zero position to go past the gear clutching tooth chamfer can be computed as follows: refer to Figure 20, Section B-B,

$$Q = P + a_S + a_G \quad (46)$$
$$\tan\beta = \frac{t_S}{2 \times a_S},$$

or

$$a_S = \frac{t_S}{2 \times \tan\beta} \quad (47)$$
$$\tan\lambda = \frac{t_G}{2 \times a_G},$$

or

$$a_G = \frac{t_G}{2 \times \tan\lambda} \quad (48)$$

Substituting values from Equations 47 and 48 in Equation 46

$$Q = P + \frac{t_S}{2 \times \tan\beta} + \frac{t_G}{2 \times \tan\lambda} \quad (49)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6 FRICTION AND RING MATERIAL COMPATIBILITY WITH LUBRICANT OIL

Engines with greater torque and power are being used in modern vehicles including passenger cars, trucks, and tractors. And, yet the comfort of lower shift effort and smoother gear transition is imperatively demanded. High dynamic coefficient of friction is required to meet this demand, which can be secured by boundary lubrication, that is, the cone and blocking ring surfaces are separated with a boundary layer. The friction elements are designed with mechanisms to wipe out the oil film from cone surfaces to prevent hydrodynamic pressure from developing. Furthermore, selection of optimized friction material and lubricant for wear is essential factor as the comportment of friction and ring material is contingent upon the lubricant oil. It is necessary to engineer the lubricant with adequate additives such as extreme pressure, friction modifiers, and detergents/dispersants to enhance thermal stability. Hence, it is earnestly recommended that tests be conducted to establish oil compatibility for the friction and ring materials.

6.1 Friction material

The selection of friction materials is application specific and is made with keen consideration for wear and friction. Some of the friction materials being currently utilized are listed in the following:

- *Copper forging alloys*, such as aluminum bronze, are selected for synchronizer ring based on composition, strength, and stability as required. Friction rings are forged or cast. Fine pitch internal threads are cut and finished along with fine flats on top of the threads as oil wiping mechanism to generate synchronization torque. Grooves are machined along the width of the internal threaded cone of the ring to allow rapid removal of the wiped oil and subsequent quick development of cone torque.
- *Paper lining* used as friction element that provides adequate coefficient of friction and is capable of dissipating kinetic energy to the amount of 3150 kJ/m^2. However, its wear rate is high and resistance to thermal degradation is low.
- *Sintered Friction Lining*, which is a complex mixture of metallic and nonmetallic powders, provides high thermal capacity. It is capable of high load level and maintains low surface temperature.
- *Woven carbon fiber* has high unit loading capacity and is used in automotive and truck applications. Its high energy capability allows smaller synchro system. It provides stable coefficient of friction in that the static and dynamic coefficients of friction are relatively same.

6.2 Friction ring material

Copper forgings were largely used in the early transmissions and are still in vogue in some low power applications. However, the modern casting technology advances that allow maintaining tighter tolerances have made cast ring prevalent in current transmissions.

7 PERFORMANCE FAILURE CONDITIONS

Accurate design and meticulous manufacture of synchronizer components are key factors in accomplishing the objectives of flawless gearshift. Some of the failure conditions will be enumerated here along with reasons thereof.

7.1 Clash

Clash is the most annoyingly common performance failure in a transmission. Insufficient cone torque due to partially energized synchro ring causes clash. There are a host of factors stemming from defective design and/or fallacious manufacturing and quality of the components that result in clash. Listed below are some of the reasons for clash.

7.1.1 Clash due to ring-related problems

- Worn or yielded ring
- Worn or too wide thread flats on bronze rings
- Poor ring to cone conformity: roundness, flatness, run out, etc
- Sticking rings
- Ring blocking teeth worn or broken
- Incorrect ring clocking

7.1.2 Clash due to mismachined parts

- Sleeve detent groove
- Fork groove in sleeve
- Hub to sleeve spline backlash and tilt excessive
- Fork pads mismachined or worn
- Fork pads deflection excessive or uneven

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

7.1.3 Clash due to inadequate strut snap back

- Excessive shift linkage friction
- Excessive strut clearance due to ring wear
- Mismachined sleeve strut groove
- BTL too low

7.1.4 Clash due to incorrect torque ratio

- Low or poor cone microfinish specs
- High oil viscosity causing cold clash
- Improper lubricant additives
- Incorrect ring or sleeve pointing angles
- Ring or sleeve pointing off center
- Incorrect proximity

7.1.5 Clash due to clutch drag and related problems

- Clutch housing misaligned
- Bushing lube inadequate

7.1.6 Clash due to high index torque

Index torque higher than the cone torque will always cause clash. Since index torque is a function of sleeve pointing angle, it is determinatively critical to design so that the index torque remains lower than the cone torque throughout the synchronization event. Hence, sleeve pointing angle and cone angle must be selected to maintain $T_C > T_I$.

7.2 Double bump

During the synchronization process, immediately after the cone torque has peaked and prior to completion of the event, a sensation of another smaller peak is felt, which is defined as double bump (Figure 25).

Following are the reasons for double bump:

- BTL too high
- Coefficient of friction too high

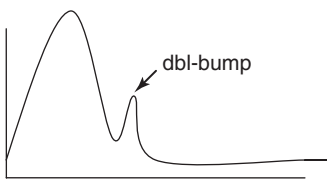


Figure 25. Cone torque versus sleeve travel.

- Cone angle too low
- Sleeve pointing angle too high

7.3 Gear jump out

The sliding sleeve splines engage with the gear clutching teeth that transmit the power. Transmission lugging could distress the hard parts to the extent to cause jump out. Listed below are some of the reasons for jump out:

- Insufficient or worn out back taper
- Broken or worn out fork pads
- Worn out fork sleeve groove
- Worn gear clutching teeth due to lugging

7.4 Hard shift and block out

Hard shift and block out occurs either due to intemperate harsh shifting or due to improper design of synchronizer components.

Breakneck power shift in a flash forcing the sleeve to crash through and not letting the blocker ring to index or the synchronizer torque to develop.

Hard shift and clash consideration in the design phase would reduce the possibility of block out. Pointing angle, cone angle, shift time, and shift effort relationship should be maintained as depicted in Figures 9a and b. Furthermore, blocker ring chamfer angle damage or gear clutching tooth chamfer angle damage would also result in block out.

Excessive clutch drag, windage, and oil churning could cause block out as well.

NOMENCLATURE

A	Unit area
AL	Arc length
a	½ detent groove width at gage point
a_R	X-coordinate of ring pointing angle
a_S	X-coordinate of sleeve pointing angle
a_G	X-component of gear clutching tooth pointing
BTL	Break through load
b	Width friction lining
C_T	Gear back face to clutching tooth front face
D	Minor diameter sleeve splines
D_G	Gage diameter sleeve detent ramp
D_P	Pitch diameter ring/sleeve tooth
D_t	Outer diameter ring tooth
E	Total energy
F	Axial force sleeve groove

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

F_A	Axial load to overcome detent spring reaction
F_N	Number of friction surfaces
F_R	Reaction force, detent spring
F_S	Spring actuating force
F_{ST}	Total radial strut load
F_I	Indexing force
f_1	Frictional force, hub spline to sleeve spline
f_2	Frictional force, sleeve pointing to ring pointing
f_3	Frictional force, strut ramp to sleeve pocket ramp
f_4	Frictional force at sleeve contact at strut edge
f_5	=0 (no sliding after N_7 developed)
f_6	Frictional force, strut contacts one side of ring pocket
f_R	Friction force, ring
f_S	Friction force, sleeve
G	Gage point offset
K	Percent detent load, strut not in detent
L_B	½ strut bump length
L_R	Ring lug width, axial
L_{RW}	Total ring width
L_{SL}	Length, sleeve front face to rear gage point
L_{ST}	½ strut length
l_f	Clearance per friction surface
N	Number of ring teeth/sleeve spaces
N_R	Normal force, ring
N_S	Normal force, sleeve
N_T	Number of sleeve teeth
N_1	Normal force, hub spline to sleeve spline
N_2	Normal force, sleeve pointing to ring pointing
N_3	Normal force, strut ramp to sleeve pocket ramp
N_4	Normal force, sleeve contact at strut edge
N_5	Normal force, strut contact at one side of ring pocket
N_6	Normal force, strut contact one side of ring pocket
N_7	Normal force, strut contact at other side of strut pocket
ΔN	Rubbing speed, rpm
n	Gear step
P	Sleeve tooth point to clutching tooth point from zero point
Q	Sleeve tooth chamfer to clutching tooth chamfer from zero point
R	½ cone gage dia
R_F	Radius, friction surfaces
R_r	Overall ratio reduction
r	½ pitch diameter sleeve/ring tooth
r_R	Ring radius
r_S	Radius at slot for ring lug in sleeve
S	Sleeve tooth point to ring tooth point from "0"
T_C	Cone torque
T_D	Total system drag
T_d	Clutch drag torque
T_f	Torque due to frictional losses
T_I	Index torque
T_S	Torque reaction on strut
T_{Ch}	Torque due to fluid churning
t	Synchronization time
t_G	Gear clutching tooth width
t_R	Ring tooth width
t_S	Sleeve tooth width
t_t	ring tooth width at ring OD
t_{DP}	Ring tooth width at pitch dia
W_C	Combined vehicle weight
W_L	Ring lug width
W_S	Hub slot width
W_{RT}	Ring tooth thickness, axial
X	X-component of sleeve ramp angle
Y	Y-component of sleeve ramp angle
Z	½ sleeve detent groove width
α	Cone angle
α_{SLOT}	Angle at any side of hub slot
β	Pointing angle, sleeve/ring
γ	Angle at any side of lug
δ	Distance, cone gage point to front of clutching tooth
θ	Sleeve detent ramp angle
λ	Clutching tooth pointing angle
η	Dynamic viscosity
μ	Coefficient of friction, detent strut/ball to sleeve
μ_C	Coefficient of friction, cone surface
μ_l	Locking Coefficient, $\sin\alpha$
μ_p	Coefficient of friction, sleeve to ring pointing
μ_{SH}	Coefficient of friction, sleeve spline to hub spline
ψ	Ring clocking angle
φ_{PA}	Pressure angle, ring/sleeve
φ_t	Transverse pressure angle
ω	Initial velocity of output side of synchronizer

RELATED ARTICLES

Possibilities of Coil Springs and Fibre Reinforced Suspension Parts
Designing Twist Beam Axles

REFERENCES

Razzacki, S.T. (2004) Synchronizer design: a mathematical and dimensional treatise. SAE Paper No. 2004-01-1230. Society of Automotive Engineers: PA, USA. 10.4271/2004-01-1230.

Razzacki, S.T. and Hottenstein, J.E. (2007) Synchronizer design and development for dual clutch transmission (DCT).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

SAE Paper No. 2007-01-0114. Society of Automotive Engineers: PA, USA. 10.4271/2007-01-0114.

FURTHER READING

McEwen, E. (1949) The Theory of Gear Changing. *Proceedings of the Institution of Mechanical Engineers: Automobile Division*, vol. 3 no. 1 30-40, 10.1243/PIME_AUTO_1949_000_009_02 January. Institution of Mechanical Engineers.

SAE International (1990) *AE-15 Gear Design, Manufacturing and Inspection Manual*, Society of Automotive Engineers, PA, USA.

Socin, R.J. and Walters, L.K (1968) Manual transmission synchronizers. SAE technical paper #680008, Society of Automotive Engineers: PA, USA. 10.4271/680008.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Syed Razzacki.
DOI: 10.1002/9781118354179.auto077
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Automatic Transmissions—Gear Train Combinations, Components, Design Considerations, Hydraulic System, Packaging, Manufacturing, Assembly

Takashi Shibayama
JATCO Ltd., Fuji City, Japan

1 Introduction 1
2 Basic Principle of Shifting by using a Planetary Gear Set 2
3 Gear Train Combination for Multiple Step Shifting 6
4 Components for Providing the Shifting Function 6
5 Hydraulic System 6
6 Centrifugal Balance System 10
7 Cooling and Lubrication 10
8 Packaging and Assembling 13
9 Manufacturing 16
Acknowledgments 16
References 18
Further Reading 18

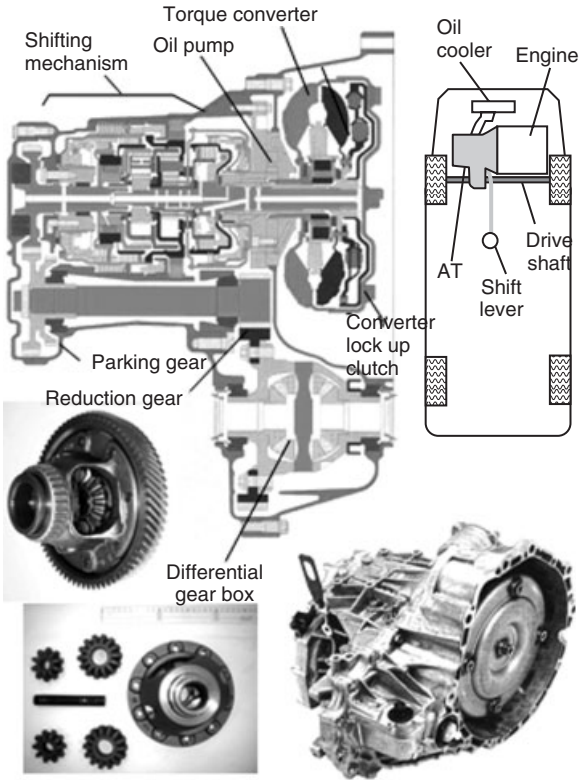
1 INTRODUCTION

In order to understand the basic structure and the basic operating principle of the automatic transmissions (ATs), the following topics are described.

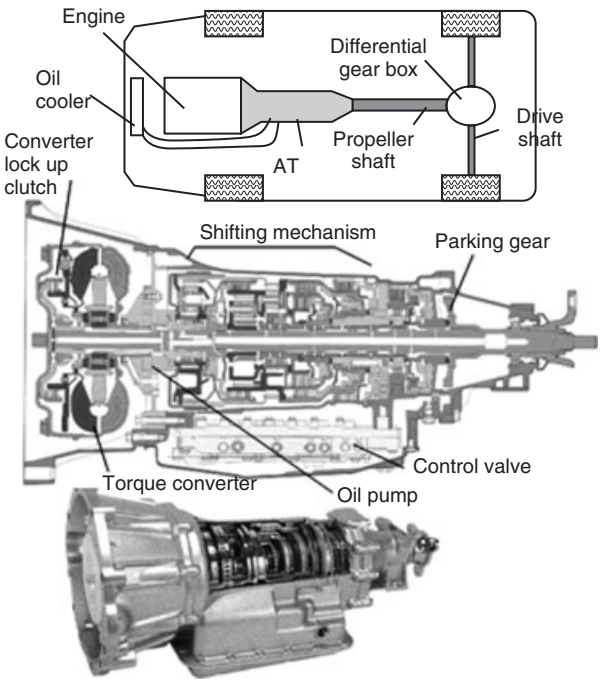
Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

1.1 Typical layout and structure of a step AT

Figure 1a shows a typical step AT installation layout for a front wheel drive vehicle and its typical structure. Engine output torque is transmitted to the torque converter. The shifting mechanism is built with several clutches, brakes, and planetary gear sets. Through the reduction gears and differential gear box, transmission output torque is transmitted to the drive shaft. The torque converter also contains a lockup clutch system by which converter slipping losses can be eliminated during cruising conditions. In general, the automatic transmissions have the mechanical parking mechanism, which serves to mechanically lock the transmission output shaft. Hence, a parking gear is added on the output shaft as is shown in the figure. An oil pump is driven by the hub of the torque converter outer cover. Figure 1b shows a typical step AT installation layout for a rear wheel drive vehicle and its typical structure. The basic layout and structure is similar to that of a step AT for a front wheel drive vehicle, except it does not have a differential gear box drive as with front drive designs. In addition, the rotating axis is straight as shown in the figure. Its output torque is transmitted to the in-line propeller shaft. In step AT development, the parking system is important. Designers have to pay attention to obtain good park performance and good shift lever operation feeling. Figure 2 shows a typical parking system. Shift operation force and feeling is adjusted by shaping of manual plate and detent spring characteristics. The details of an AT are



(a) Front wheel drive vehicle



(b) Rear wheel drive vehicle

Figure 1. (a,b) A typical step AT installation layout. (Reproduced by permission of Jatco, Ltd.)

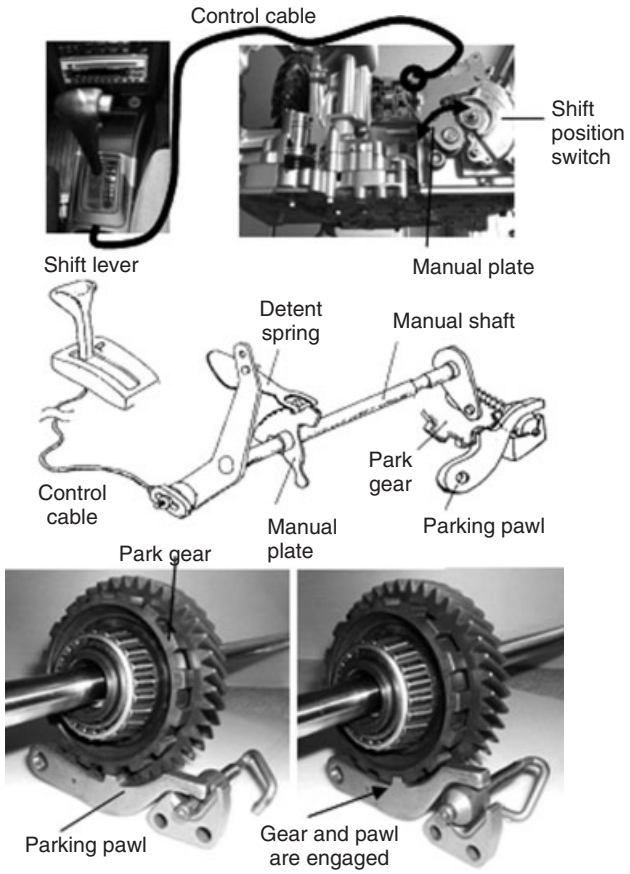


Figure 2. A typical parking system. (Reproduced by permission of Jatco, Ltd.)

explained in this chapter, using a typical 5-speed AT for a rear wheel drive vehicle as an example.

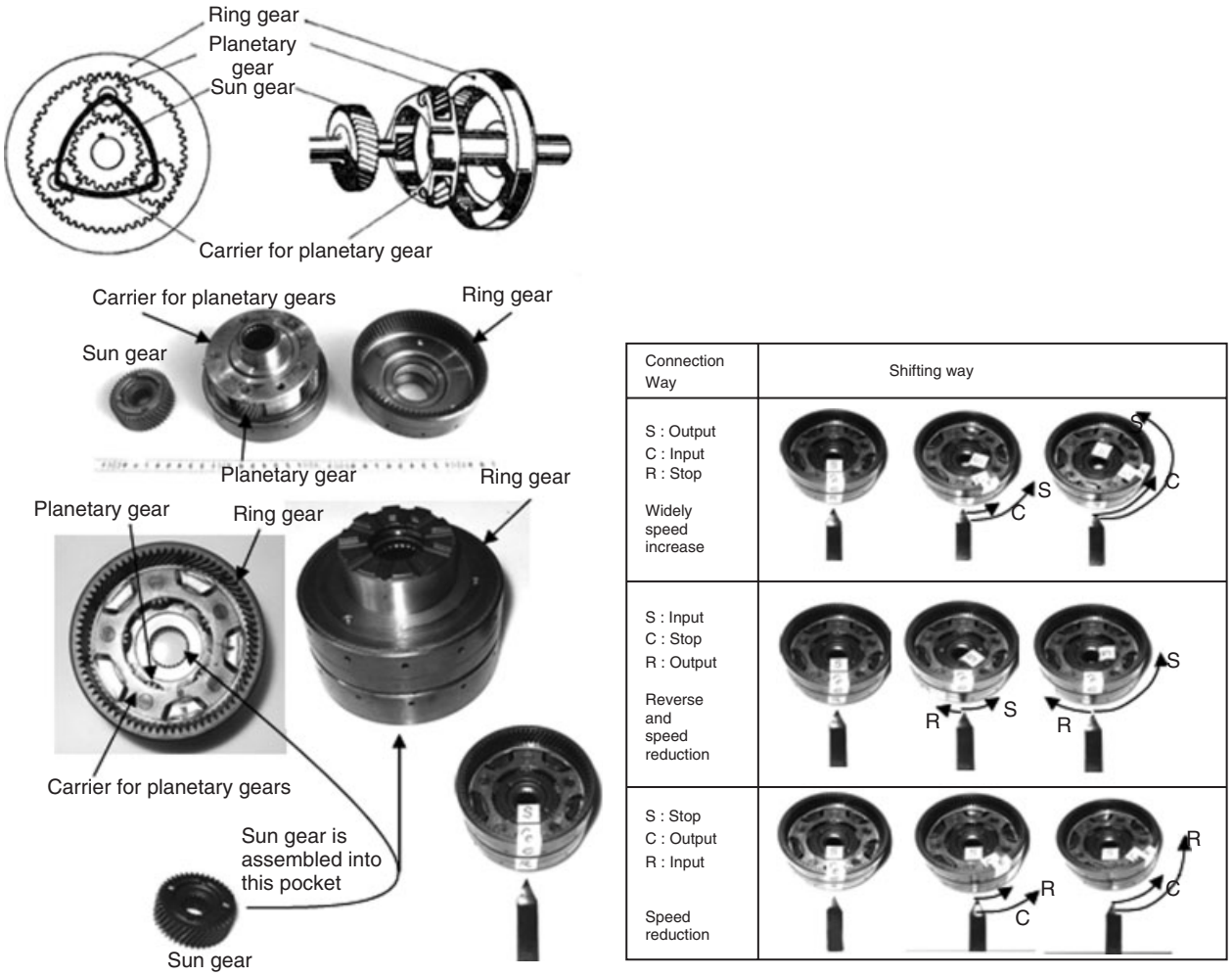
2 BASIC PRINCIPLE OF SHIFTING BY USING A PLANETARY GEAR SET

The first step is to study the shifting principle of a planetary gear set. It is very fundamental but very important to understand the shifting mechanism of a step AT. In Figure 3, a typical planetary gear set and its actual parts are shown (Society of Automotive Engineers, 1962).

The important elements are the sun gear, the carrier, and the ring (or annulus) gear and the symbols S, C, and R are used for those elements.

S, sun gear
C, carrier
R, ring gear

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Element \ Case		0	1	2	3	4	5	6	7
Sun gear	S	No connection	Input	Input	Stop	Output	Output	Stop	Any two element of the three element are engaged each other
Carrier	C	No connection	Output	Stop	Input	Input	Stop	Output	
Ring gear	R	No connection	Stop	Output	Output	Stop	Input	Input	
Gear set function		Neutral	Widely speed reduction	Reverse and speed reduction	Speed increase	Widely speed increase	Reverse and speed increase	Speed reduction	Direct connection

Figure 3. A typical planetary gear set and the shifting principle. (Reproduced by permission of Jatco, Ltd.)

In Figure 3, there is an important table that shows the possible shifting combinations obtained by one planetary gear set.

A planetary gear set can make several kinds of shifts. As in Figure 3, speed reduction, speed increase, or reverse function can be obtained by connecting its elements to input, output, or stationary elements. The obtained function depends on which element is connected to input, output, or stationary elements. Here, the gear ratio can be calculated from following fundamental formula (Society of Automotive Engineers, 1973).

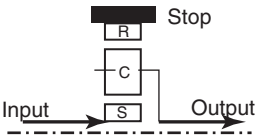
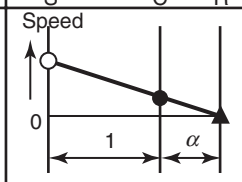
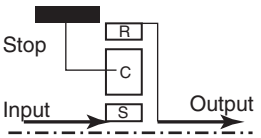
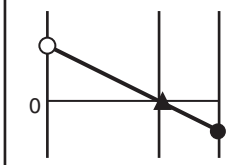
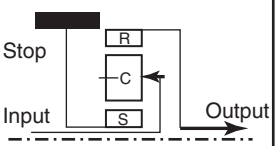
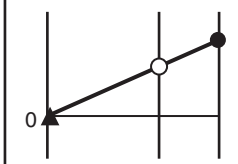
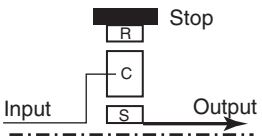
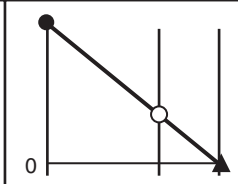
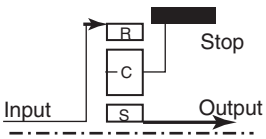
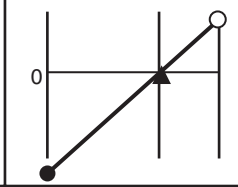
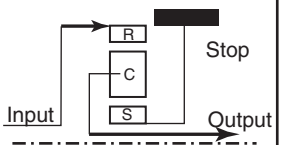
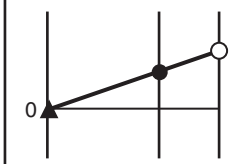
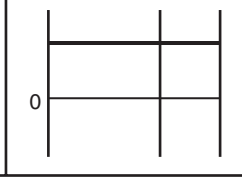
$$(N_R - N_C) = -\alpha(N_S - N_C)$$

where

N_R = rotational speed of ring gear
N_C = rotational speed of carrier
N_S = rotational speed of sun gear

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Ex. $\alpha = 0.5$

	Gear element: Sun (S)	Gear element: Carrier (C)	Gear element: Ring (R)	Toruque flow	Function of gear set	Gear ratio output speed/ input speed	Lever expression (○ Input ● Output ▲ Stop)
①	Input	Output	Stop		Widely speed reduction	$\frac{\alpha}{1+\alpha}$ = 0.33 When $\alpha = 0.5$	
②	Input	Stop	Output		Reverse and speed reduction	$-\alpha$ = –0.5 When $\alpha = 0.5$	
③	Stop	Input	Output		Speed increase	$1+\alpha$ = 1.5 When $\alpha = 0.5$	
④	Output	Input	Stop		Widely speed increase	$\frac{1+\alpha}{\alpha}$ = 3.0 When $\alpha = 0.5$	
⑤	Output	Stop	Input		Reverse and speed increase	$-\frac{1}{\alpha}$ = –2.0 When $\alpha = 0.5$	
⑥	Stop	Output	Input		Speed reduction	$\frac{1}{1+\alpha}$ = 0.67 When $\alpha = 0.5$	
⑦	Any two element of the three element are engaged each other				Direct connection	1	

α : Z_S / Z_R Z_S : Tooth number of sun gear

Z_R : Tooth number of ring gear

Figure 4. Possible shifting function obtained by one planetary gear set. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

$$\alpha = Z_S/Z_R$$

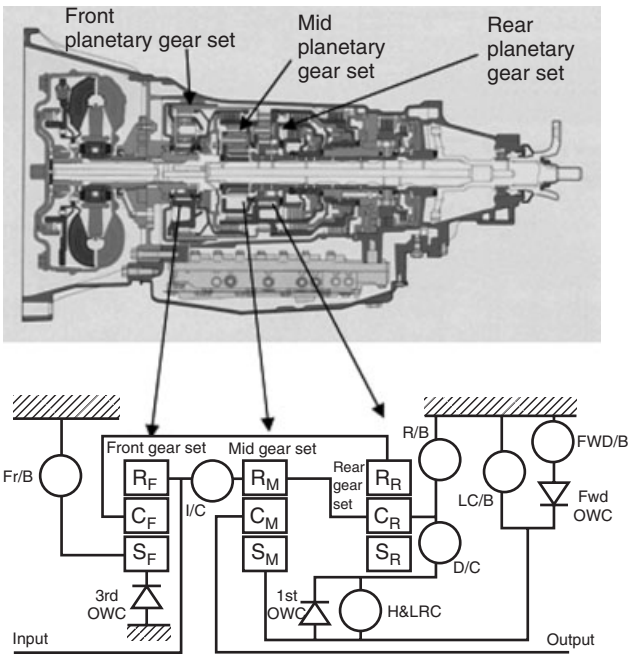
Z_S = tooth number of sun gear
Z_R = tooth number of ring gear

For example, if carrier is stopped, $N_C = 0$. Then,

$$N_R = -\alpha N_S$$

In this case, the reverse function is realized. In the same way, several functions can be provided. In the figure, detailed function and possible speed ratios are described.

In Figure 4, the lever expression is introduced. The expression is very popular in the step AT industry to determine ratios. The vertical axis indicates the rotating speed of each element and horizontal axis indicates speed ratio obtained by tooth ratio of ring gear and sun gear.



Shift position \ Clutch brake			Fr/B	I/C	H&LR/C	D/C	R/B	LC/B	Fwd/B	1st OWC	Fwd OWC	3rd OWC
While driving	D	1	∗		∗			∗	●	▼	▼	▼
		2	∗			●		∗	●		▼	▼
		3	∗		●	●				∗		▼
		4		●	●	●				∗		
		5	●	●	●					∗		∗
	R		∗		∗		●			▼		▼
While engine braking	D	1	●		●			●	∗	∗	∗	∗
		2	●			●		●	∗		∗	∗
		3	●		∗	●				▼		∗
		4		●	∗	●				▼		
		5	∗	●	∗					▼		▼
	R		●		●		●			∗		∗
			Front brake	Input clutch	High and low, reverse clutch	Direct clutch	Reverse brake	Low coast brake	Forward brake	1st one way clutch	Forward one way clutch	3rd one way clutch

●: Engaging and transmitting torque
∗: Clutch or brake slipping speed is zero
Blank: Disengaging and slipping

▼: Slipping speed is zero and transmitting torque
∗: Slipping speed is zero
Blank: Slipping

Figure 5. A stick diagram and connection of the rotating members of a typical conventional 5-step AT. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

By utilizing this lever expression, it is easier to understand how speed reduction or speed increase can be obtained by visualizing mechanical leverage. Input, output, and stationary elements can be visualized by this expression.

3 GEAR TRAIN COMBINATION FOR MULTIPLE STEP SHIFTING

In actual step ATs, several planetary gear sets are combined to obtain multiple steps shifting. Here, in this encyclopedia, a typical conventional 5-step AT with three planetary gear sets will be described to comprehend such a Gear Train Combination Shifting System (Yamaguchi and Sugihara, 2002, pp. 16–22).

Figure 5 shows a stick diagram and connection of the rotating members. To facilitate shifting function, some ring gear or sun gear or carrier should be braked or stopped. This is accomplished by brakes as indicated in the figure, for example, Fr/B (Front Brake) and R/B (Reverse Brake).This gear train combination is actually adopted in the JATCO RF507E model, which is in mass production. Rectangles indicate elements of the planetary gear, sun gear, carrier, and ring gear. White circles indicate multiple wet plate clutches. Moreover, there are three one-way mechanical clutches in this system that can transmit torque in only one direction. Such one-way clutches are often adopted because they are very useful to provide a smooth shift feeling. Using a one-way clutch, connecting or disconnecting is automatically achieved during shifting.

Figure 5 also indicates the detail connection table to make a 5-speed AT with reverse range. Because the one-way clutches can only transmit torque in one direction, bypass clutches are connected to obtain the other direction torque transmitting capacity when desired. Bypass clutches are engaged to transmit engine braking torque when it is necessary, for example during desending mountains.

In Figure 6, a detailed connection and lever expression of this typical 5-speed AT with reverse range is described, and in Figure 7, torque flow of this 5-speed AT is described.

4 COMPONENTS FOR PROVIDING THE SHIFTING FUNCTION

Figure 8 shows clutches and brakes that are used for providing shifting function and other power transmitting components. Multiple wet plate clutches, band brake, drum, one-way clutch, torque converter, gear sets, and shafts. Detailed explanation of multiple plate clutches is described in previous Chapter 2.8 (see Clutch Wet). In addition, torque converter is described in previous Chapter 2.9 (see Automotive Torque Converters). One-way clutches can only transmit torque in one direction. A band brake is often used for step ATs, which has a unique characteristic of torque transmitting capacity as shown in the figure. The band brake has a different capacity according to the drum rotation direction.

Up to this chapter, basic principles and mechanisms have been explained using the example of a conventional 5-speed AT. Recently, 6-, 7-, and 8-speed ATs have become popular, and the number of gears and the clutches tend to increase to provide such multiple steps ATs; therefore, the cost, the size, and the weight also tend to increase. So, the study of gear combination has become very important.

For the development of a step AT, many requirements must be considered: for example, reducing the numbers of clutches/brakes or numbers of parts, easy handling in assembly, or simplify the shift control. Recently, ZF made big progress in the gear train combination by computer-aided development.

As is shown in Figure 9, ZF adapted a very smart gear train combination for its 8-speed AT (SAE paper 2009-01-0510). This gear train combination has very limited number of shift elements, three multi-disc clutches and two multi-disc brakes. There are only two open clutches or brakes causing slip losses so the losses become minimal. During shifting, the system only requires to open one clutch/brake and close another clutch/brake, which makes the control system simple. With those points and others, this newly developed ZF 8-speed AT gear train is a very advanced system. Because aspeed has already been developed.

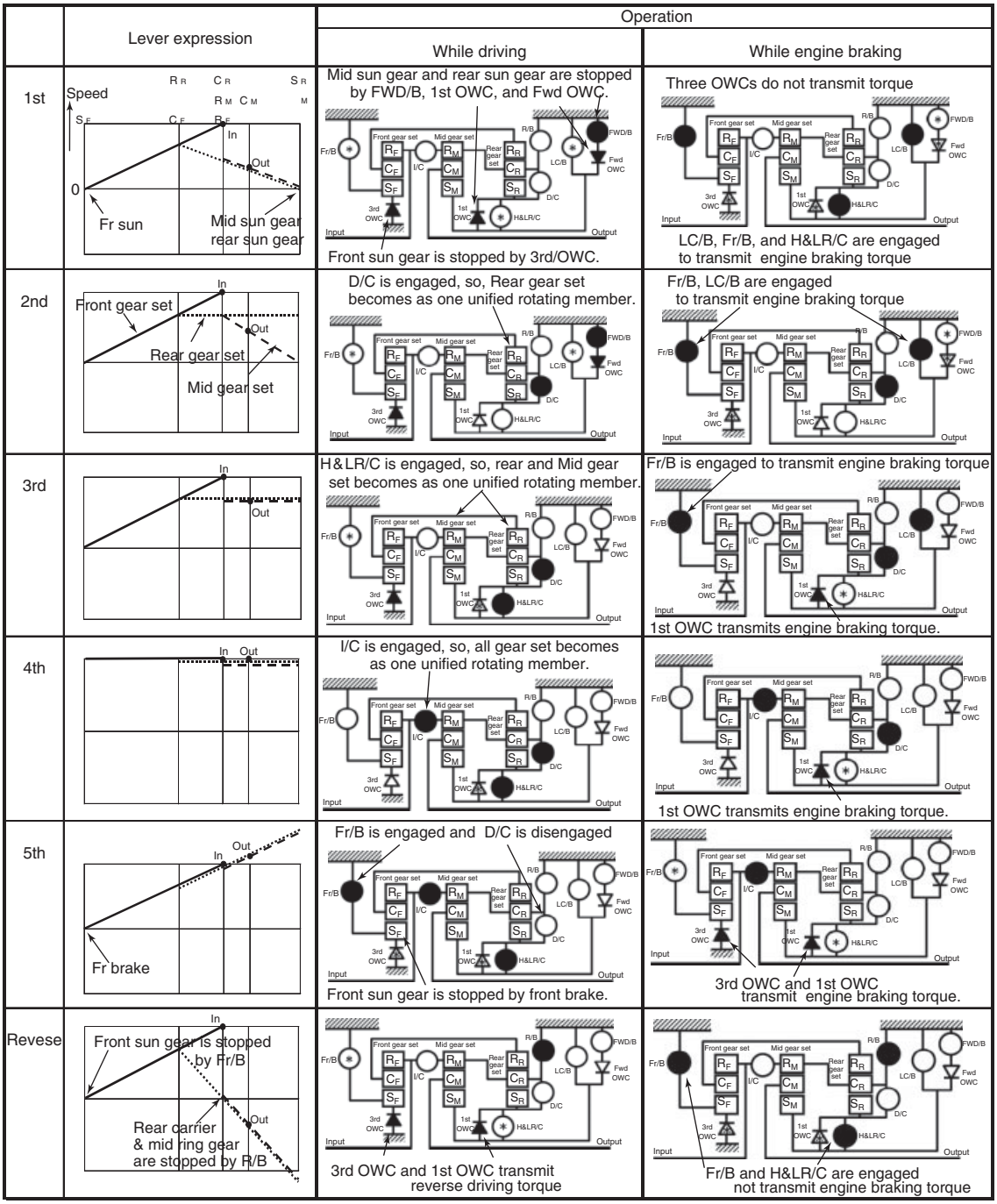
5 HYDRAULIC SYSTEM

As is described in Sections 2–4, shifting operation is done by hydraulic actuators of clutches and brakes. In addition, there are other important roles provided by the hydraulics. Figure 10a indicates the basic concept of utilizing hydraulic benefits. Indeed, hydraulics have many benefits: actuation, power supply, cooling, lubrication, shift feeling adjustment, cleaning, and so on. (Stebar *et al*., 1990, pp. 827–840).

Oil is delivered to each device through control valves after flow has been generated by the oil pump.

Usually, there are several of these devices. The whole oil circuit tends to become a little bit complex as is shown in Figure 10b, which is a typical oil circuit of a step AT. Using such an oil circuit, oil flow is switched or delivered

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Shift position		Clutch brake	Fr/B	I/C	H&LR/C	D/C	R/B	LC/B	Fwd/B	1st OWC	Fwd OWC	3rd OWC
While driving	D	1	⋆		⋆			⋆	●	▼	▼	▼
		2	⋆			●		⋆	●		▼	▼
		3	⋆		●	●				⋆		▼
		4		●	●	●				⋆		
		5	●	●	●					⋆		⋆
	R		⋆		⋆		●			▼		▼
While engine braking	D	1	●		●			●	⋆	⋆	⋆	⋆
		2	●			●		●	⋆		⋆	⋆
		3	●		⋆	●				▼		⋆
		4		●	⋆	●				▼		
		5	⋆	●	⋆					▼		▼
	R		●		●		●			⋆		⋆

● :Engaging and transmitting torque
⋆ :Clutch or brake slipping speed is zero
Blank : Disengaging and slipping

▼ :Slipping speed is zero and transmitting torque
⋆ :Slipping speed is zero
Blank : Slipping

Engaging and transmitting torque
Slipping speed is zero
Disengaging and slipping
OWC slipping speed is zero and transmitting torque
OWC slipping speed is zero
OWC slipping

Figure 6. A detailed connection and lever expression of a typical conventional 5-step AT. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

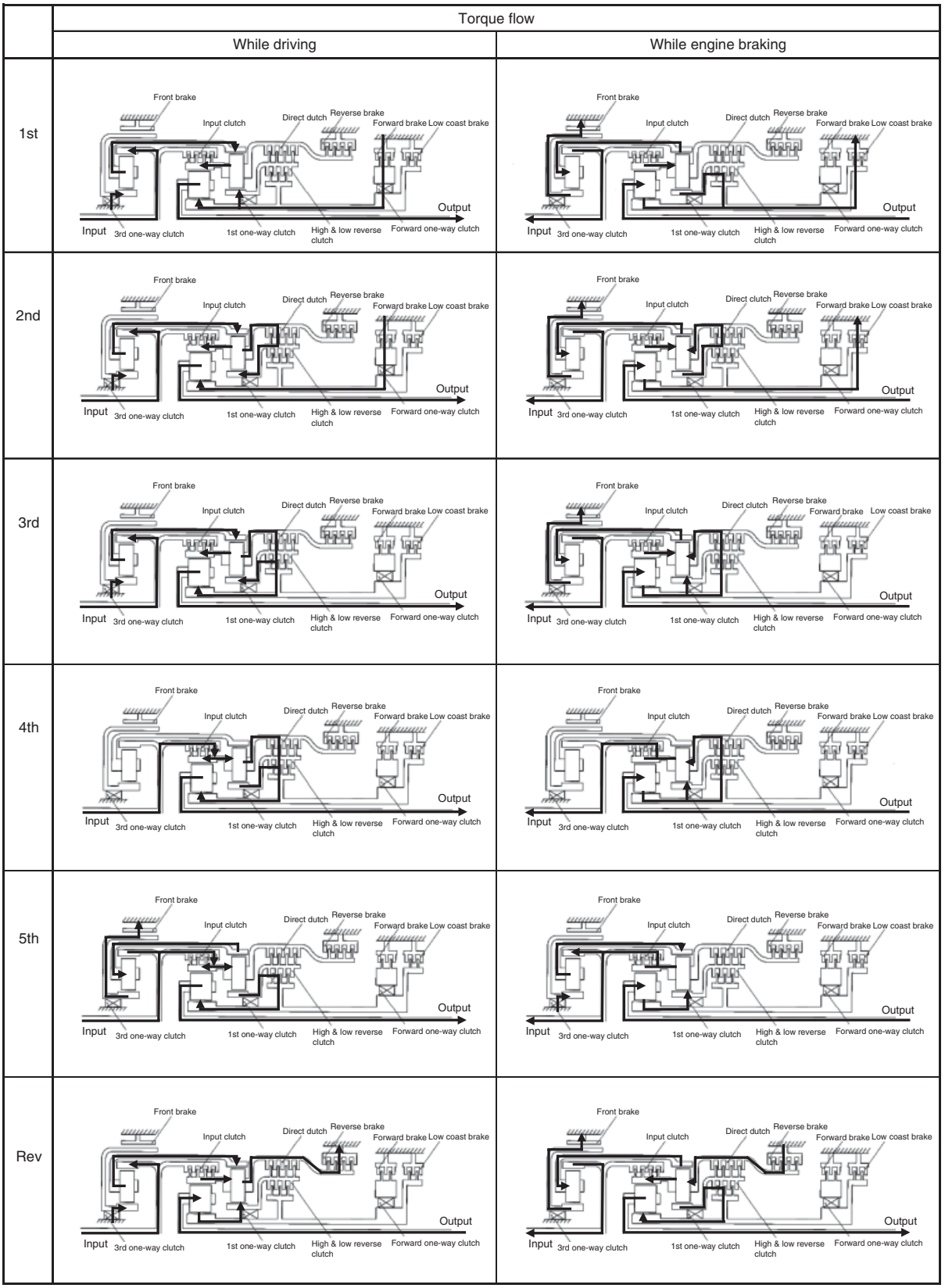


Figure 7. Torque flow of a typical conventional 5-step AT. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

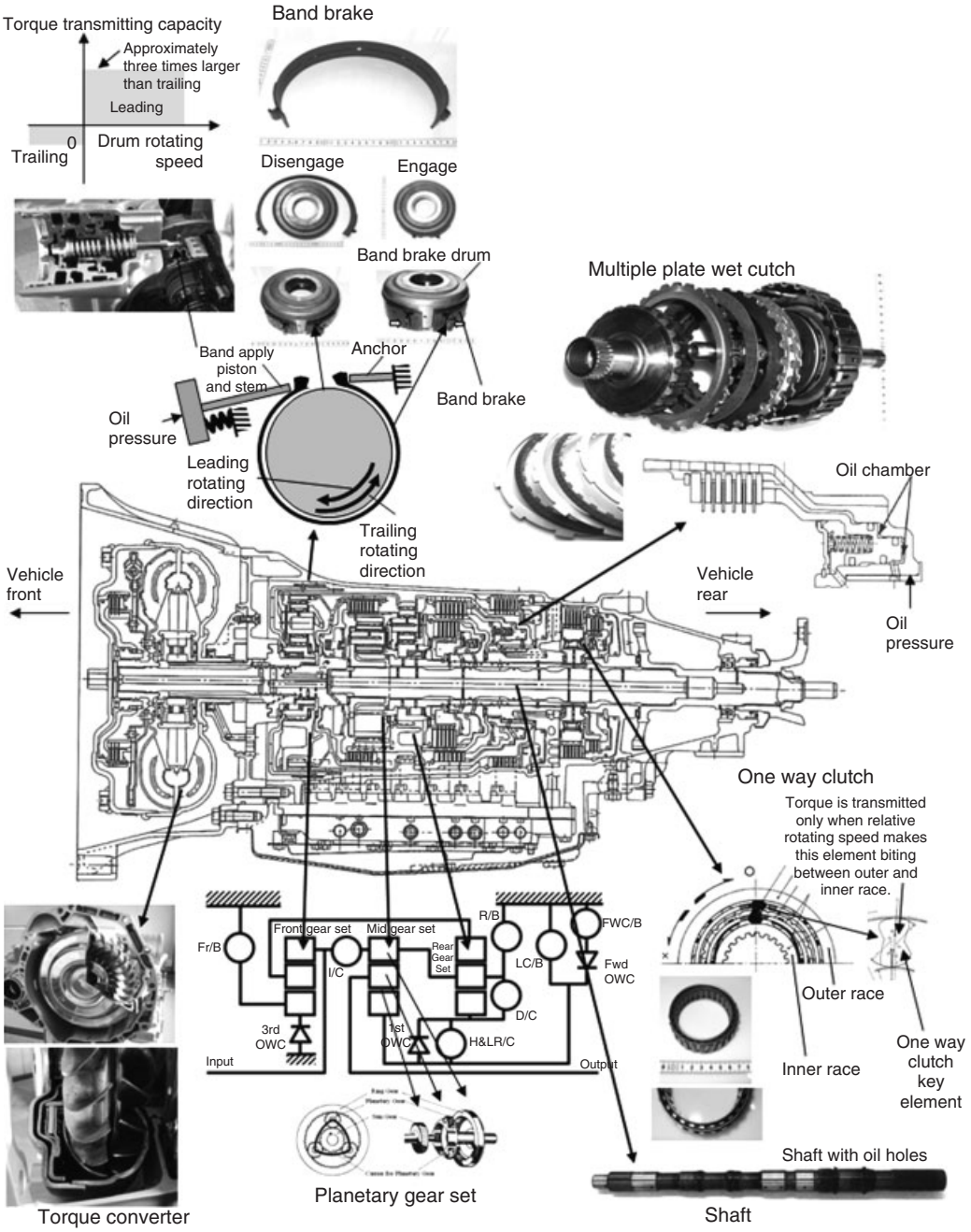


Figure 8. Components that are used for providing shifting function and other power transmitting components. (Reproduced by permission of Jatco, Ltd.)

to each of the devices. Figure 10c shows a typical oil circuit for forward clutch operation and for converter lockup clutch operation (Society of Automotive Engineers, 1994, pp. 523–539).

Pictures of actual components and parts are shown in Figure 11. Here, a typical oil pump and typical control valve assembly are shown. A control valve assembly has a similar structure to an electrical circuit board. Upper valve body has many ditches and the lower valve body has many ditches for establishing the oil circuits. A separator plate is sandwiched between the two valve body halves and has many holes to connect upper ditches and lower ditches oil passages. Oil flow will pass through several holes in the separator plate before reaching its

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

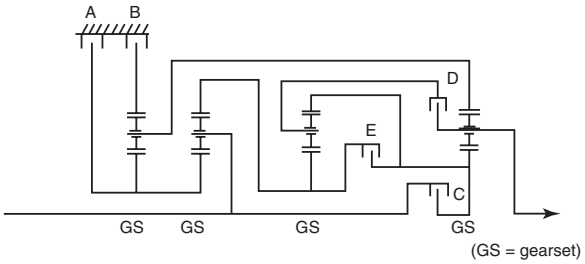


Figure 3 8-speed transmission diagram

Gear	Brake		Clutch			Ratio i	Gear step
	A	B	C	D	E		
1	●	●	●			4,696	1,50
2	●	●			●	3,130	1,49
3		●	●		●	2,104	1,26
4		●		●	●	1,667	1,30
5		●	●	●		1,285	1,29
6			●	●	●	1,000	1,19
7	●		●	●		0,839	1,25
8	●			●	●	0,667	Total 7,05
R	●	●		●		–3,297	

Figure 9. An advanced gear train combination for ZF's 8-step AT. (Reproduced by permission of Tatsuo Matsuda, ZF, Japan © ZF Friedrichshafen AG.)

objective devices. Sometimes, orifices are provided in the separator to restrict oil flow to adjust actuating timing.

6 CENTRIFUGAL BALANCE SYSTEM

In rotating wet clutch systems, avoiding influence of centrifugal pressure is very important. Figure 12a explains centrifugal pressure effect. Although clutch pressure is released, oil remains in the clutch apply chamber. This oil in the clutch apply chamber has a mass and therefore provides a force when the clutch is rotated. This centrifugal pressure can apply the clutch "unintentionally" at higher rotational speeds and therefore must be counteracted.

If the clutch rotating speed is low enough, the piston return spring force is large enough to overcome the centrifugal apply force due to rotation of the clutch. However as the clutch rotating speed becomes high, clutch apply force by centrifugal pressure exceeds the return spring force. Figure 12b shows such a case. The clutch piston is stroked against the return springs and provides an apply force to the friction material. Drag torque and excess heating are generated on the friction material surfaces. This condition must be avoided for reliability and durability.

To avoid this issue, a balancing or canceling system is usually adopted in the wet clutch assembly. There are two ways to make a balance chamber or to make an oil removal system.

Figure 12c indicates the preferred way to make a cancel chamber. In this system, as is indicated in the figure, the balance or cancel chamber is added to provide an equal and opposite centrifugal force to cancel the applying force by the centrifugal pressure in clutch apply chamber. This system is often adopted in recent newly developed ATs because the cancel performance is reliable, but this system requires another oil seal and another oil supplying circuit.

Another approach is to make an oil removal system. Figure 12d indicates one example, which is utilizing a check ball valve. Usually, the check ball is made of steel. This steel ball receives two forces: one is centrifugal force originated by its own weight and the other force is flow force originated by clutch oil pressure. As is shown in Figure 12d, if clutch rotating speed is low, centrifugal force is small, so the steel ball acts as an oil seal. Conversely, if clutch rotating speed becomes high, centrifugal force becomes large and the steel ball moves toward outer diameter along the sloped surface in the steel ball cylinder.

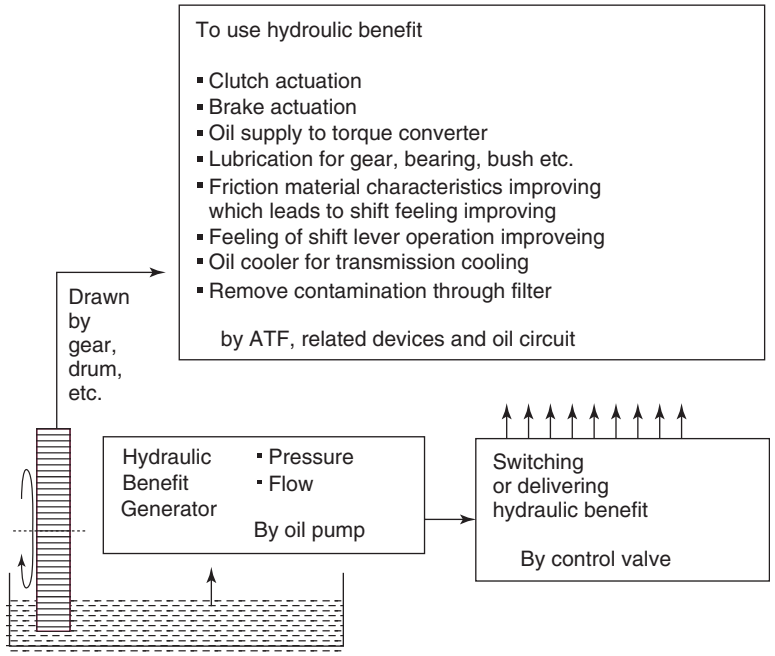
Actually, the steel ball position depends on clutch rotating speed and pressure at the clutch apply inlet hole. In this way, seal or drain is switched according to the solid line and dashed line in the figure. Flow force changes according to flow rate around the steel ball, so, there is some hysteresis between the solid line and dashed line. This ball system was very popular in the past but time lag during oil filling and other sensitivities brought pretty much difficulty into development, so, this system tends to be replaced by the balance system.

7 COOLING AND LUBRICATION

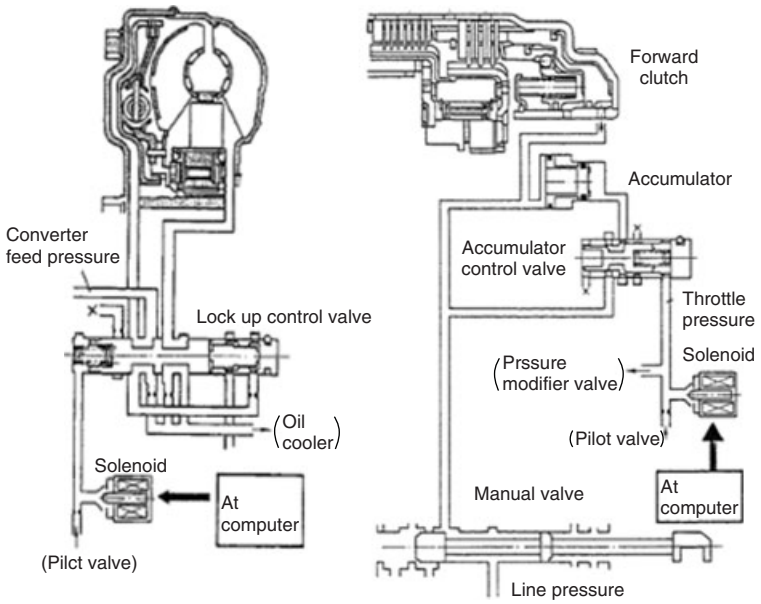
Cooling and lubrication are very important in AT development. Figure 13a explains a typical circuit for cooling and lubrication.

The largest heat generator is the torque converter. Therefore, usually output flow from the torque converter is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) Utilization of hydraulic benefit



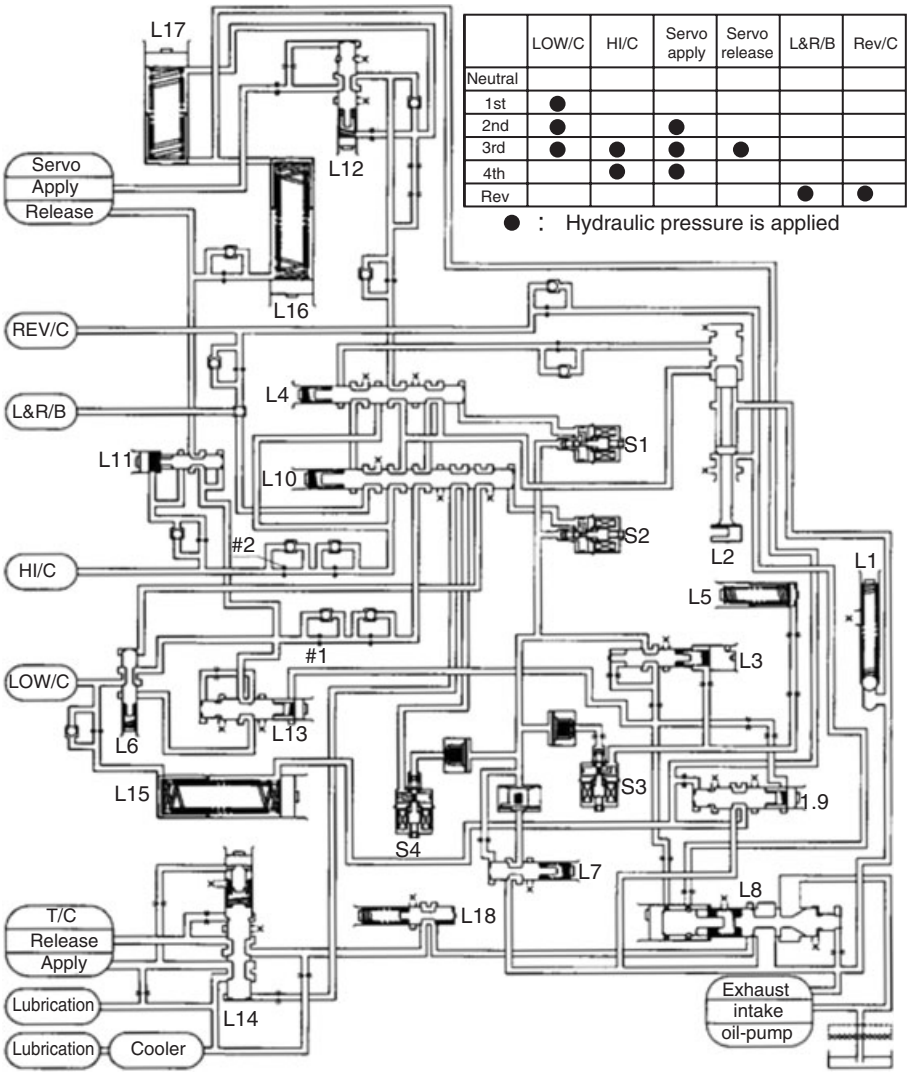
(c) Torque converter clutch and forward clutch control circuits

Figure 10. (a,b) Typical oil circuits of a step AT. (Reproduced by permission of Jatco, Ltd.)

usually sent directly to the oil cooler, which provides the required ATF (automatic transmission fluid) cooling. From the cooler, this cooled oil is sent back to the transmission where it enters the lubrication circuit, where it is used to lubricate and cool various parts such as bearings, bushings, gears, and clutch friction material. Usually, the shafts have holes for lubrication oil flow, and sometimes, additional pipes for routing lubrication flow are used.

Oil flow rate is adjusted by oil hole diameter or number of oil holes on the shafts.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



	LOW/C	HI/C	Servo apply	Servo release	L&R/B	Rev/C
Neutral						
1st	●					
2nd	●		●			
3rd	●	●	●	●		
4th		●	●			
Rev					●	●

L2	Manual valve	L11	Timing valve	S1	Shift A solenoid valve
L4	Shift A valve	L13	Neutral control valve	S2	Shift B solenoid valve
L6	Neutral shift valve	L14	Lock-up control valve	S3	PL duty solenoid valve
L9	Accumulator control valve	L15	LOW/C accumulator	S4	LU duty solenoid valve
L10	Shift B valve	L16	Servo release accumulator	#1, #2	Orifice

(b) Typical oil circuit of step AT

Figure 10. (*Continued*)

After lubrication, the oil comes back into oil pan and is drawn up by the pump to begin its journey through the transmission circuits once again.

Figure 13b indicates a typical cooling system for a rear wheel drive AT. In the radiator lower tank, a heat exchanger is installed. This heat exchanger provides ATF cooling using engine coolant and cool air. As is shown in the figure, hot ATF comes from AT and cooled ATF returns back to transmission.

Cool air also cools the transmission case directly.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

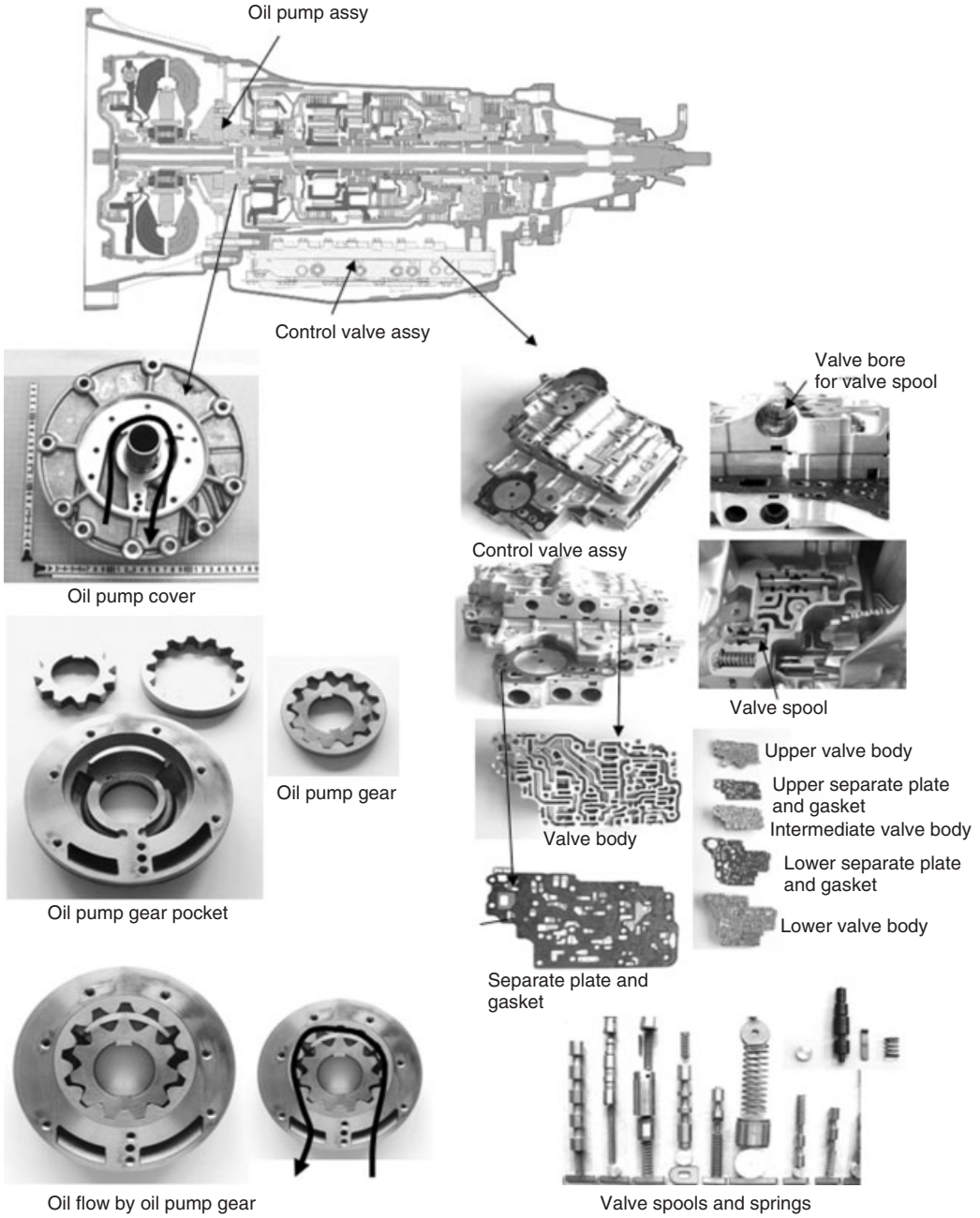


Figure 11. Actual components and parts for hydraulic system. (Reproduced by permission of Jatco, Ltd.)

Sometimes, for severe use vehicles such as heavy-duty trucks or taxis, an additional cooler (or radiator) that uses cool air for cooling is adopted.

Figure 13c indicates another cooling system. An independent oil cooler can be used that is attached on the transmission case by bolts.

This system has more flexibility for various vehicle installations but tends to have less cooling capacity than the radiator system described earlier. Therefore, sometimes, both a radiator in tank cooler and an independent attached cooler are adopted.

8 PACKAGING AND ASSEMBLING

For packaging and assembling, the designer and engineer investigate several issues.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

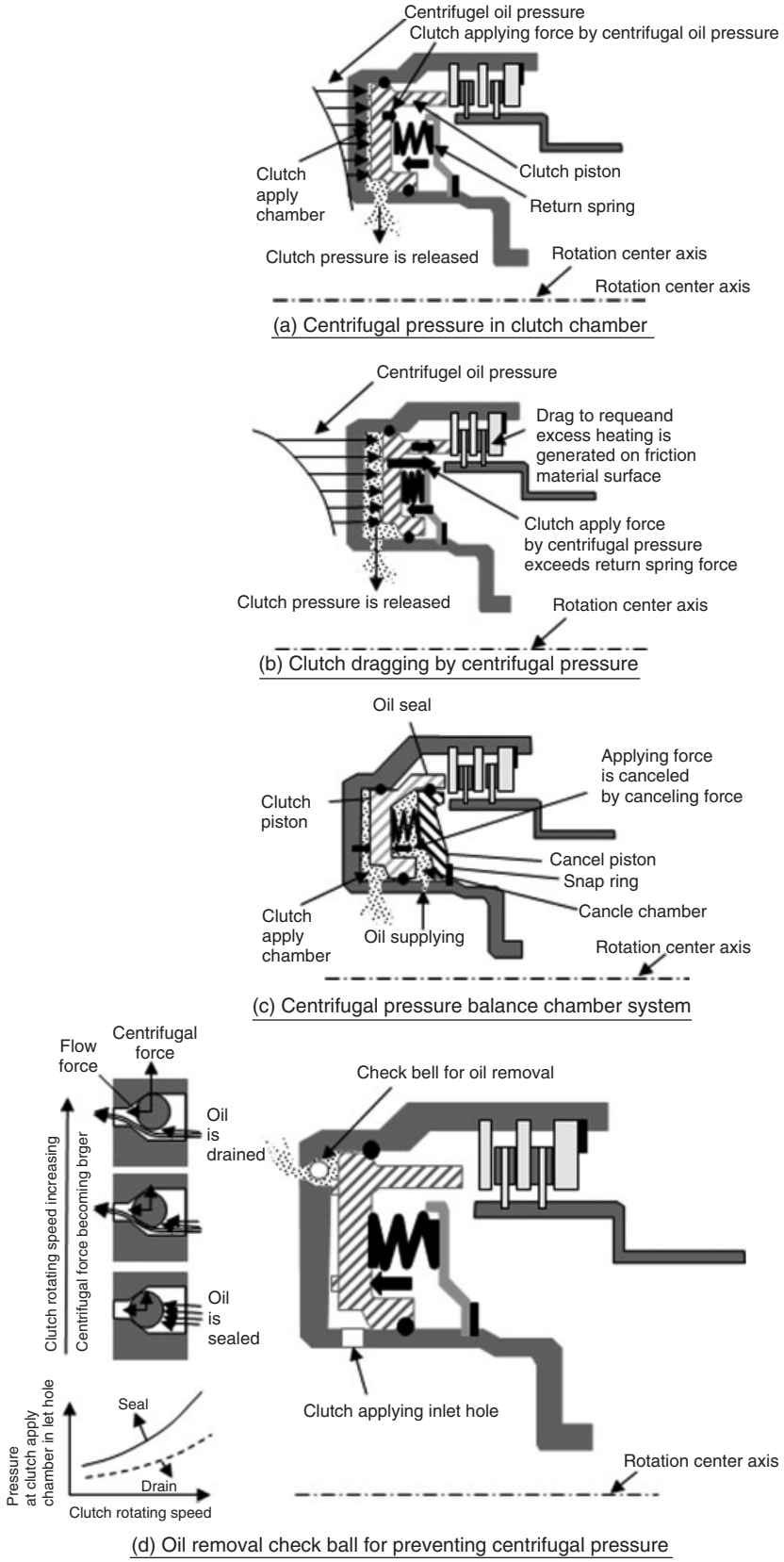


Figure 12. (a–d) Balance system for centrifugal pressure in the clutch chamber. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

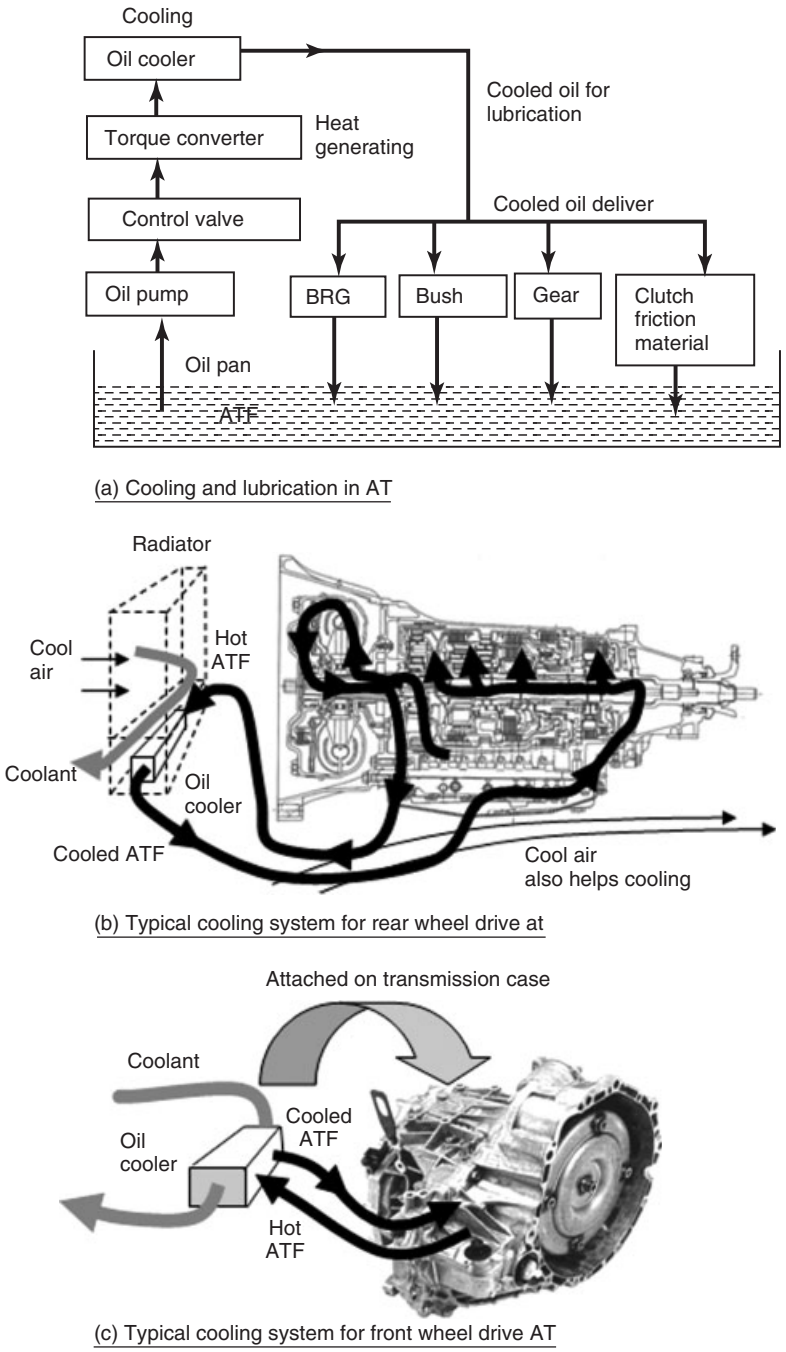


Figure 13. (a–c) A typical circuit for cooling and lubrication. (Reproduced by permission of Jatco, Ltd.)

Figure 14 indicates such issues. The figure shows an example of lineup and common parts strategy. Among several family models, some parts or components are shared. Using such a strategy, cost, and developing speed is improved.

In this example, the PTO (power take off) gear layout is also investigated for commercial vehicles. Such investigations are also important to make the lineup simpler (Shinohara *et al*., 1989).

Figure 15 shows other examples of assembly issues. Here, work ease and cycle time for assembly are investigated by computer simulation. Recently, such simulation has become very important to establish a good mass production factory line, because such

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

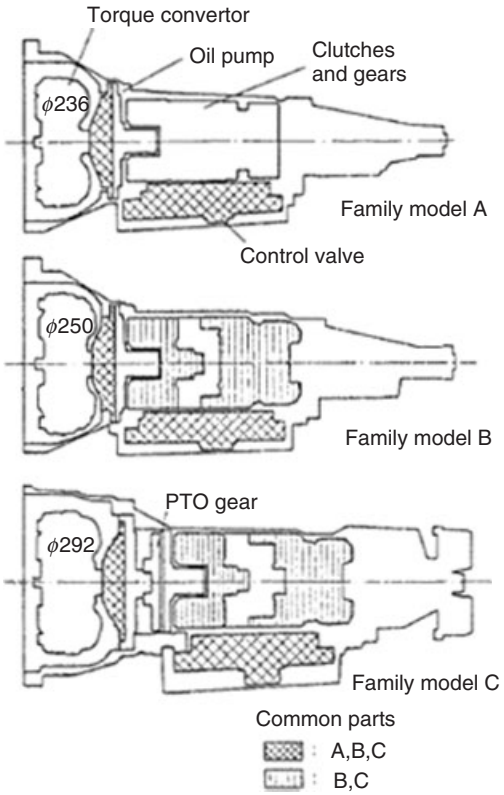


Figure 14. An example of lineup and common parts strategy. (Reproduced by permission of Jatco, Ltd.)

simulation can reduce or eliminate actual modification during production preparation (Takatori *et al*., 2009, pp. 25–42).

Three-dimensional simulation technology makes it possible to investigate such issues during the early preparation stages.

9 MANUFACTURING

For manufacturing, designers and engineers have to investigate various things and in this area; technology and innovation are also making great progress day by day (Martin and Redinger, 1993, pp. 579–602).

Figure 16a shows one example of a pressure die casting simulation. In pressure die casting, molten aluminum flow is very important. If flow is not adequate, as is shown in the figure, an unfilled area will create a void. Such void can cause oil leak or some weakness issue for the case.

To predict and avoid such unfilled areas in early development stages is very important to reduce large amounts of time and financial loss. Figure 16a includes another computer-aided engineering (CAE) system example for die design. Die design can be semiautomatically done aided by computer. Tooling design and lead time can also be improved using CAE. Figure 16b shows a tooling simulation example (Takagi, Ito, and Makino, 2008).

ACKNOWLEDGMENTS

The author wishes to thank Mr. Tatsuo Matsuda who works for ZF Japan, for his very helpful advice to improve the contents of this subscription, and also would like to thank all colleagues in JATCO and JATCO US.

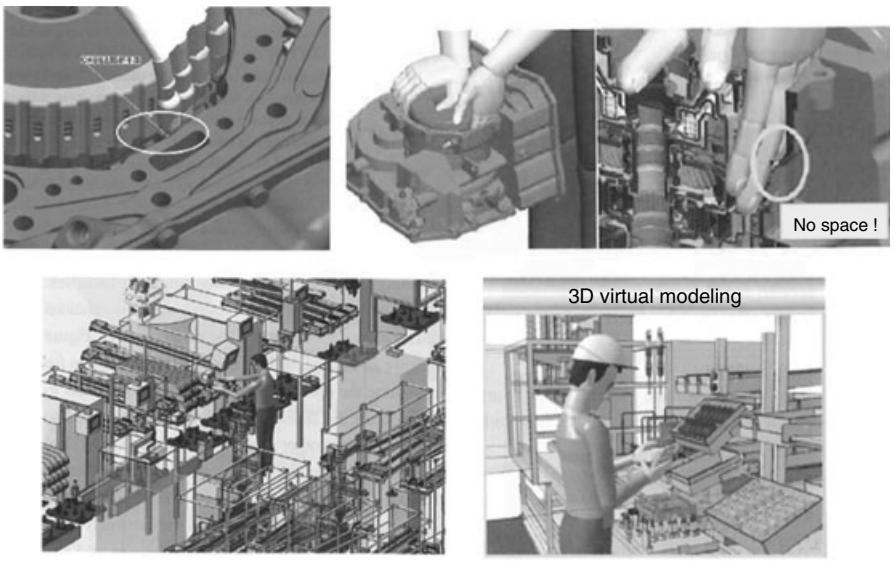


Figure 15. An example of investigation for working easiness and cycle time in assemble work. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

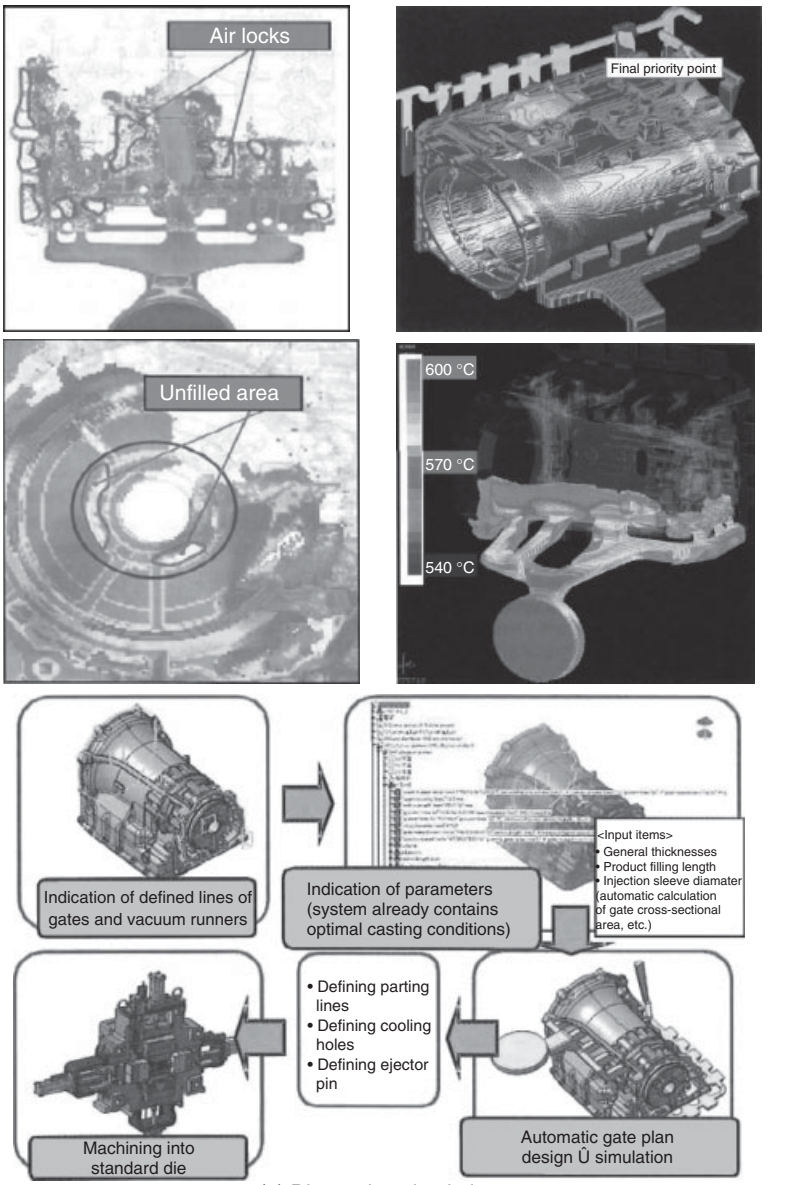


Figure 16. An example of simulation for manufacturing. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

REFERENCES

Berthold Martin and Charles J. Redinger (1993) Electronic four-speed automatic transaxle. SAE Technical Paper No. 930671. Society of Automotive Engineers (1994) *Design Practices, Passenger Car Automatic Transmissions*, 3rd, edn, AE-18 (Advances in engineering; vol. 18), pp. 579–602.

Minoru Shinohara, Takashi Shibayama, Ohtsuka, K. *et al*. (1989) Nissan electronically controlled four-speed automatic transmission. SAE Technical Paper 890530. Society of Automotive Engineers (1994) *Design Practices, Passenger Car Automatic Transmissions*, 3rd, edn, AE-18 (Advances in engineering; vol. 18), pp. 523–539.

Society of Automotive Engineers (1962) *Design Practices: Passenger Car Automatic Transmissions*, AE-1 & 2 (Advances in Engineering, Vol. 1, 2), Society of Automotive Engineers, Warrendale, PA.

Society of Automotive Engineers (1973) *Design Practices: Passenger Car Automatic Transmissions*, AE-05 (Advances in Engineering, Vol. 5),, Society of Automotive Engineers, Warrendale, PA.

Society of Automotive Engineers (1994) *Design Practices: Passenger Car Automatic Transmissions*, 3rd, edn, AE-18 (Advances in Engineering, Vol.18), pp. 523–539.

Russell F. Stebar, Ellard D. Davison, and Linden, J. (1990) Determining frictional performance of automatic transmission fluids in a band clutch. SAE Technical Paper 902146. Society of Automotive Engineers (1994) *Design Practices, Passenger Car Automatic Transmissions*, 3rd, edn, AE-18 -- (Advances in engineering; vol. 18), pp. 827–840.

Takagi, S., Ito, H., and Makino, T. (2008) Simultaneous design with related sections in *Jatco Technical Review* No. 7, pp. 50–54.

Takatori, K., Saito, Y., Kitagawa, H., *et al*. (2009) Introducing Jatco's new 7-speed AT for RWD cars 'the best products in the World by the smoothest operations' in *Jatco Technical Review* No. 8, pp. 25–42.

Yamaguchi, T., Sugihara, T., Inaba, T., and Shirato, K. (2002) Weight reduction technologies for powertrain parts of a new 5-speed AT in *Jatco Technical Review*, Jacto, Japan, pp. 16–22, No. 3.

SAE Paper (2009) SAE paper 2009-01-0510, *SAE. Intl. J. Engines*, **2** (1), 314–326.

FURTHER READING

Jatco (2000-2012) *Jatco Technical Review* No. 1, No. 11.

Society of Automotive Engineers (2012) *Design Practices, Passenger Car Automatic Transmissions*, 4th, edn, AE-29 (Advances in engineering; vol. 29), Society of Automotive Engineers, Warrendale, PA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto081
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The Variable Pulley CVT

J.G.L.M. van Spijk[1] and A. Englisch[2]

[1]*Bosch Transmission Technology B.V., Tilburg, The Netherlands*
[2]*LuK GmbH & Co. KG, Bühl, Germany*

1 Introduction 1
2 Construction and Operation 1
3 Variator 4
4 Difference in Principle Between Belt and Chain 7
5 Push Belt Design 7
6 Chain Design 10
7 Oil Pump and Hydraulics 11
8 CVT Assembly 16
References 16
Further Reading 17

1 INTRODUCTION

Leonardo da Vinci was already aware of the importance of continuously variable transmissions (CVTs) in making driving power available to the load with the right transformation. (Figure 1) Centuries later, the Dutch inventor and car manufacturer Mr. Hub van Doorne (DAF) took the challenge to design a CVT that was suitable to be produced in automotive mass production. The rubber V-belt pulley variator proved to be a success in the market, but attracted a specific segment of people who had problems with manual gear shifting and who were looking for a cost-effective automatic transmission. The urge to overcome the resulting image and also to respond to the need for ever higher power density motivated the development of an alternative for the rubber V-belt. This resulted, on the one hand, in the mass-produced steel push belt, directly from Mr. van Doorne's initiative; on the other hand, in the chain that now is offered in mass production as well. In this chapter, the pulley variator system is described, for both push belt and chain. The system description concentrates on the most specific items for CVT, which are the variator part and its controls.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 CONSTRUCTION AND OPERATION

Figure 2 shows a state-of-the-art front wheel drive CVT. The structure of this CVT consists of a torque converter (TC), an oil pump, a drive–neutral–reverse (DNR) device, a variator (consisting of a primary pulley set, a push belt or chain, and a secondary pulley set), a set of reduction gears, a differential, a parking mechanism, a hydraulic control unit, and a transmission control unit (TCU).

2.1 Most characteristic components of the CVT

The most characteristic components of the CVT include the following:

- Launch device.
- A torque convertor, which is the most commonly used device in a modern CVT. This device provides good performance and controllability during drive-off situations, in comparison to a wet-plate clutch, which is beneficial for its dimensions and cost.
- Forward, neutral, and reverse mechanism, usually realized with a planetary gear set, to be able to switch among Drive, Neutral, and Reverse.

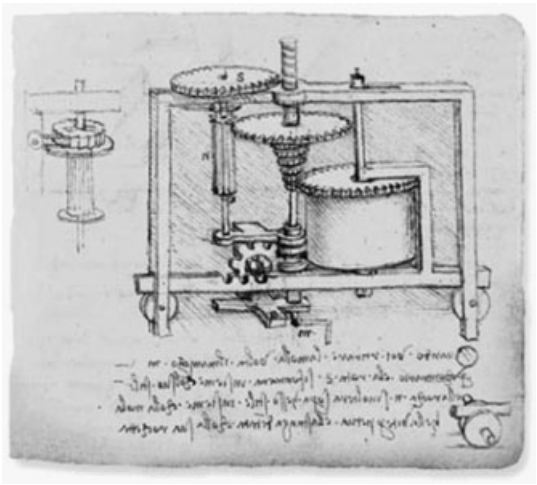


Figure 1. Da Vinci's CVT.

- A variator, which consists of a driving and a driven pulley. Each pulley comprises two conical sheaves that face each other. One of these sheaves is axially movable by a controlled pressure. A push belt/chain runs in the V-groove between the two sheaves.
- A final drive and differential. The final drive ratio reduces the variator output speed to vehicle wheel speed. The differential divides the torque over the driven wheels.
- Hydraulic/electric actuation, controlled by a TCU. This unit controls the transmission behavior.
- A pump. The pump generates hydraulic energy for actuation, lubrication, and cooling.

2.2 Transmission architecture

The main components of a CVT can be arranged in several layouts. The choice for a specific layout depends on the requirements of packaging, comfort, cost price, efficiency, and so on. Figures 3–7 show a number of possible layouts with their major advantages and disadvantages. These figures are simplified in order to emphasize the differences within component arrangement.

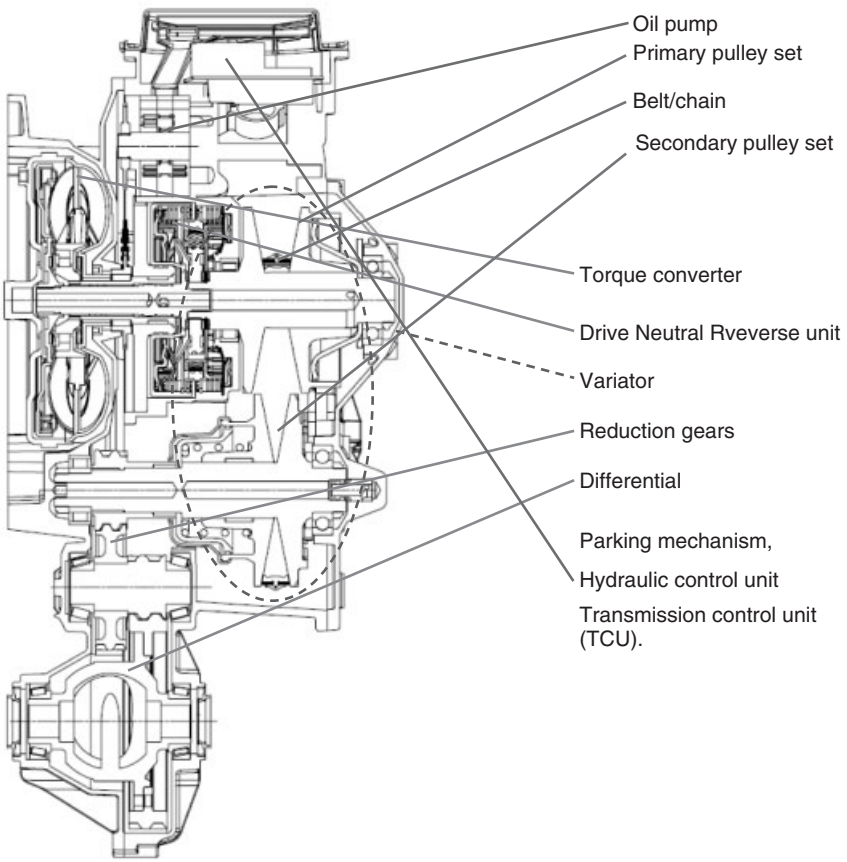
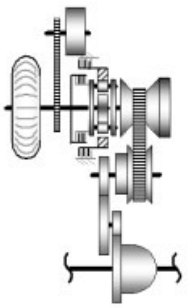


Figure 2. CVT cross section. (Reproduced by permission of Bosch Transmission Technology B.V.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



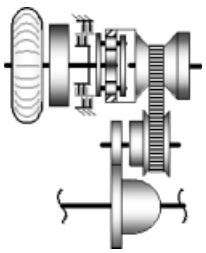
Features:
Torque converter, chain driven pump, DNR-set, variator, final drive, differential.
Remark: most common CVT-layout
Advantages:
* Compact pump; easy to change pump type/supplier
* Suits most FWD cars
Disadvantages:
* Extra pump drive component, cost
Production examples:
Jatco CVT0, CVT2, Honda 30mm-CVT, Hyundai CVT

Figure 3. Schematic architecture corresponding to Figure 2. (Reproduced by permission of Bosch Transmission Technology B.V.)



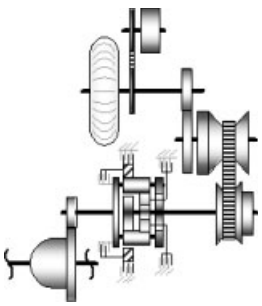
Features:
Clutches of DNR-set act as launch element, pump is mounted on end of primary side
Advantages:
∗ Cost, weight, & packaging (no torque converter and separate pump drive)
Disadvantages:
∗ Launch performance
∗ Comfort
Production example:
Punch VT2

Figure 4. CVT with wet-plate starting clutch. (Reproduced by permission of Bosch Transmission Technology B.V.)



Specific features:
DNR-set reduces speed & reverses direction of rotation in forward mode; no intermediate shaft
Advantages:
∗ Less components, from 4 to 3 shafts
∗ Packaging & weight
∗ Lower input speed for variator
Disadvantages:
∗ Increased input torque for variator
∗ Lower wheel torque in Reverse
∗ Efficiency loss in DNR
Production example:
Daihatsu D18C

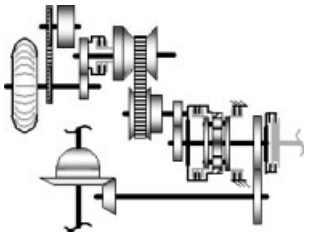
Figure 5. Three-shaft layout. (Reproduced by permission of Bosch Transmission Technology B.V.)



Specific features:
Input reduction, 2-stage DNR-set on secondary side
Advantages:
∗ Large transmission ratio coverage
∗ Primary length
∗ Efficiency
Disadvantages:
∗ Cost
Production examples:
Jatco CVT7

Figure 6. Two-step CVT. (Reproduced by permission of Bosch Transmission Technology B.V.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Specific features:
Input reduction, clutch, output reduction, DNR-set on secondary side, clutch rear wheel.
Advantages:
* This layout suits most rear-wheel-drive vehicles
* Easy start/stop option (because of primary clutch)
* Option to switch rear-wheel-drive modus on/off
Disadvantages:
* Transmission length (primary side)
* Transmission weight
* DNR-set has to transfer higher torque
* Efficiency at high speed
Production examples:
Subaru Lineartronic

Figure 7. Longitudinal FWD-CVT. (Reproduced by permission of Bosch Transmission Technology B.V.)

3 VARIATOR

3.1 Torque and clamping

Assuming the widely used architecture with a push belt as presented in Figure 2, positive driving torque comes from the engine, supplies the required torque to the pump, passes via the TC and DNR, and drives the variator from the primary pulley set. Here, half of the torque is transmitted to the moveable sheave and the other half to the fixed sheave on the shaft and both come together in the push belt (Figure 8).

The torque transfer between the push belt and firstly the primary and thereafter the secondary pulley sets is made possible by hydraulic actuation. The required clamping force is calculated according to the formula:

$$F_{\mathrm{ax,sec}} = \frac{T\,\mathrm{sec}^{*}\cos(\lambda)^{*}S_{\mathrm{f}}}{2^{*}\mu^{*}R_{\mathrm{sec}}} \tag{1}$$

where

μ = friction coefficient between push belt and pulley
R_{sec} = running radius on secondary pulley set
λ = pulley angle
S_{f} = safety factor on clamping force
T_{sec} = torque at secondary shaft.

The controlled oil pressure (p_{sec}) on the secondary cylinder area (A_{sec}) and the centrifugal effect of the oil on the cylinder ($f_{\mathrm{c,sec}}$) together with the spring force (F_{spring}) result in the clamping force secondary ($F_{\mathrm{ax,sec}}$):

$$F_{\mathrm{ax,sec}} = {p_{\mathrm{sec}}}^{*}A_{\mathrm{sec}} + {\omega_{\mathrm{sec}}}^{2*}f_{\mathrm{c,sec}} + F_{\mathrm{spring}} \tag{2}$$

The function of the compression spring is to pretension the push belt under the condition of no engine running and no oil pressure during towing or transport.

From the condition that the clamping force is secured by Equation 1, the variator will stay in a given ratio by control of the clamping force on the cylinder of the primary pulley ($F_{\mathrm{ax,prim}}$). For a stationary condition of ratio, the necessary $F_{\mathrm{ax,prim}}$ is determined by the KpKs value according to the formula:

$$\mathrm{KpKs} = \frac{F_{\mathrm{ax,prim}}}{F_{\mathrm{ax,sec}}} \tag{3}$$

Example with typical KpKs values as a function of safety and ratio (Figure 9).

To be able to shift from Low to OD ratio, the $F_{\mathrm{ax,prim}}$ has to be higher than $F_{\mathrm{ax,sec}}$ * KpKs and vice versa. The radial shaft load generated by clamping the push belt has to be supported by the variator system. The force is estimated by the formula:

$$F_{\mathrm{shaft}} = \frac{4^{*}F_{\mathrm{ax,sec}}{}^{*}\tan\lambda^{*}\sin\left(\frac{\alpha}{2}\right)}{\alpha} \tag{4}$$

where

α = wrapped angle.

This results in the variator loading conditions, which can serve as an input for the calculation of the bearing loads.

3.2 Stiffness

The variator accommodates the mechanical stops for the low and overdrive ratios. The stops prevent the outer ring of the push belt running out of the outer diameter of the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

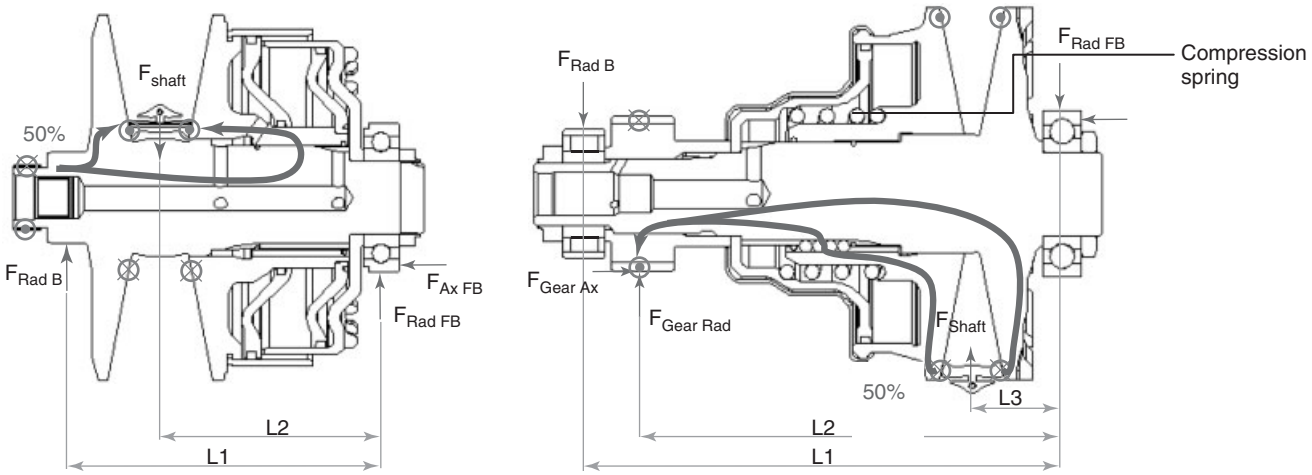


Figure 8. Primary and secondary pulley sets. (Reproduced by permission of Bosch Transmission Technology B.V.)

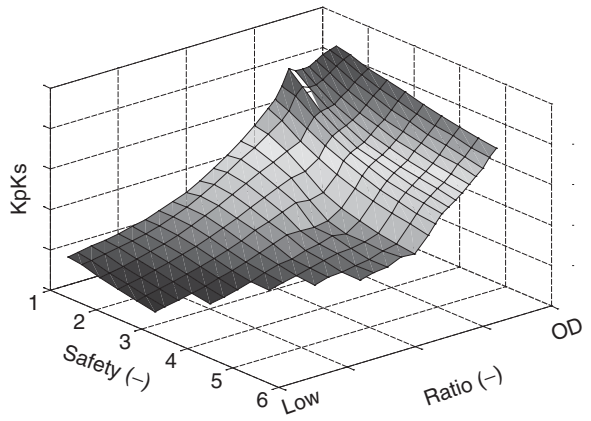


Figure 9. KpKs mapping. (Reproduced by permission of Bosch Transmission Technology B.V.)

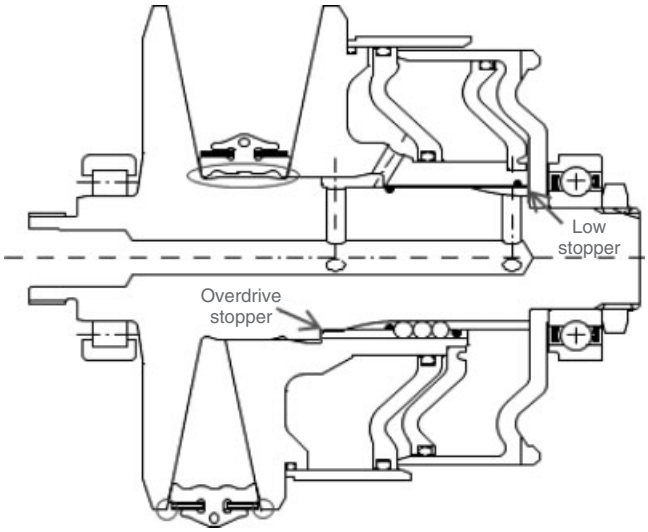


Figure 10. Primary pulley cross section. (Reproduced by permission of Bosch Transmission Technology B.V.)

pulleys, and toward the shaft center, it prevents the push belt element from running onto the shaft.

In practice, different designs are applied: both stops on one pulley set (primary), as shown in Figure 10, or a stop between movable pulley shaft and plunger on each pulley set (with a stepless pulley design).

A variator design, which is capable of transmitting the torque with high efficiency, ensures that the total axial deformation of the sheaves does not exceed a certain value. Axial deformation of the sheaves results in undesirable radial sag and consequently spiral running of the pushbelt/chain around the pulley. The radial component of this running (slip) does not contribute to the torque transmission and results in the loss of efficiency (Figure 11).

The torque is transmitted from the shaft with the fixed sheave to the moveable sheave by a design feature that gives low backlash, high tangential stiffness, and low sliding force. Typically, a so-called ball groove design is applied, in which at three positions around the shaft, half circular grooves are in the shaft and the inside of the moving sheave, in which balls are positioned with a tight tangential fit. Lower cost solutions as a single roller in just one groove are also applied (Figure 12).

3.3 Materials for pulley sets

The combined demands of high bending and torsional loads and tribological robustness of the contact require a material that combines ductility and a hard surface,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

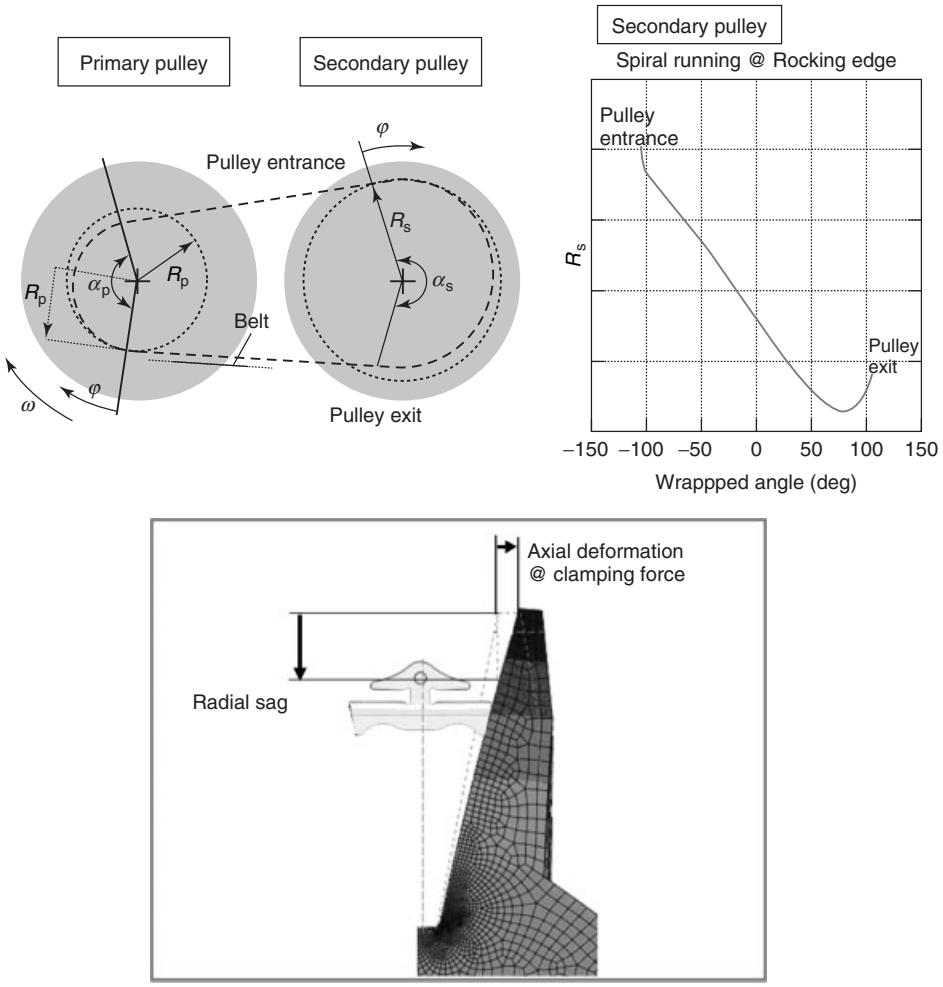


Figure 11. Radial sag—cause of spiral running. (Reproduced by permission of Bosch Transmission Technology B.V.)

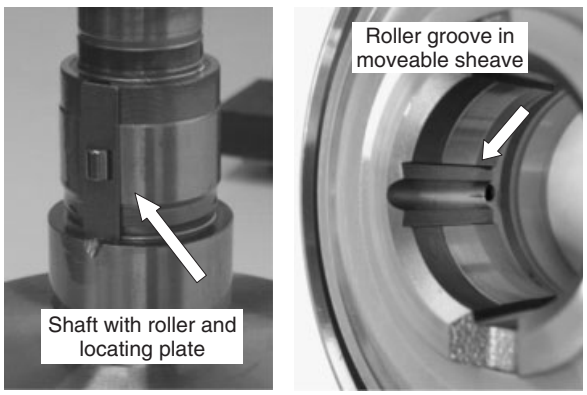


Figure 12. Sheave guiding by single roller. (Reproduced by permission of Bosch Transmission Technology B.V.)

which can be found in case hardening steels. The pulley material is a compromise among formability, machinability, heat treatability, strength, and cost. Typical materials are 20MnCrS5 and 20CrMoS5.

3.4 Tribology

The contact between push belt and pulley is lubricated and cooled by dedicated CVT oil. Together with the right setting of surface hardness and texture, this lubrication and cooling function limits wear and overheating of variator parts. Bosch performs oil validation testing, according to the so-called CVT fluid test (Brandsma *et al*., 2003). The fact that oil is also used in other contacts in the CVT, such as clutches and gears, as well as to perform hydraulic power transfer and control functions, makes dedicated CVT oil necessary.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

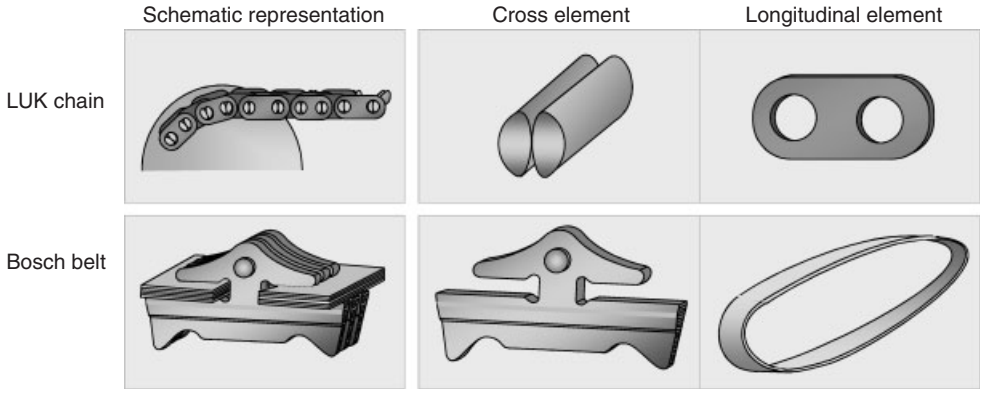


Figure 13. Components of chain and belt. (Reproduced by permission of Bosch Transmission Technology B.V.)

4 DIFFERENCE IN PRINCIPLE BETWEEN BELT AND CHAIN

The main function of the medium between the two pulleys, whether it is a chain or a belt, is to transmit power, by means of torque and speed, from one pulley to the other. The way in which this is realized, however, is different for the chain and the belt. For a better understanding, the transfer of torque is explained stepwise.

The first step is the transfer of torque from the driving pulley into the medium, which must be able to receive this torque from the driving pulley. The chain receives this torque via pins and the belt via elements (Figure 13, cross element). The transfer is, in both cases, enabled by a contact force between the pin or element and the pulley, which by friction creates a tangentional force. At stationary conditions, the sum of the tangentional forces of the pins or elements at their contact radius is in balance with the driving torque. So far, there is no distinguishing difference between the belt and the chain. The difference between the two products occurs in the second step, namely, the transfer of the received torque toward the driven pulley.

To get a picture of the second step, the layout of the two products must be examined a little closer. First is the chain. The sets of pins in the chain do not have direct contact with each other, but are attached to each other by the so-called links. It is the function of these links to transfer the received torque, by means of force, from one set of pins toward the next set, and so on, until the pins that are in contact with the driven pulley. Therefore, the function of these links is to transfer the torque from one set of pins to the other, by means of pulling force. Finally, the pins that make contact with the driven pulley transfer their pulling force into torque at the sheave similar to the pins on the driving side, namely by friction between the two contacts areas.

Next is the process for the belt. The elements in the driving pulley have absorbed the torque and feel a tangentional force. The elements transfer this force to their neighbors with whom they are in direct contact. In this way, they directly transfer the received torque by pushing the neighboring elements toward the driven pulley. The function of the loops in the push belt is just to keep the elements together under the radial forces inside the wrapped angle in the pulleys and guide them from one pulley to the other. Again, like the chain, the elements that are in contact with the driven pulley transfer their torque by friction between the two contact areas.

To be complete, it must be added that the loops do have a limited function in the transfer of torque. This can be understood by imagining the radial force of the elements in the wrapped angles, which is supported by the rings. The rings do have a relative speed to the elements, and this combination creates a positive or negative driving torque, depending on the speed ratio of the belt.

With this, the functions of the different parts of the belt and chain are described and it is understood how torque is transferred between the two pulleys. With this, it is made clear why the mediums transferring the torque have these typical different names, (pull) chain and (push) belt.

5 PUSH BELT DESIGN

5.1 Construction

The metal push belt consists of two sets of loops and a number of wedge-shaped steel elements (Figure 14). The compression type belt is designated as a push belt as the compressed elements act as a solid column to transfer torque from one pulley set to the other. The thin steel loops are the main structure of the belt. The loops are fitted

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

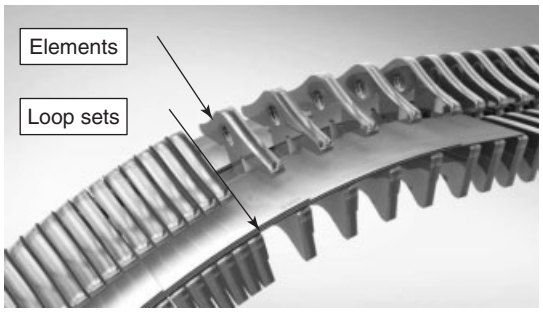


Figure 14. Push belt. (Reproduced by permission of Bosch Transmission Technology B.V.)

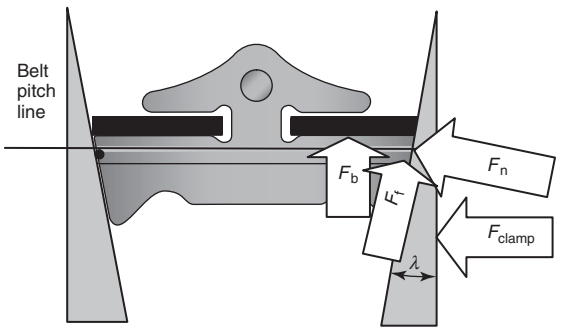


Figure 15. Loads on variator section. (Reproduced by permission of Bosch Transmission Technology B.V.)

closely together to form a nested set with no play. There are two loop sets, one on each side of the belt assembly. Each loop is approximately 0.2 mm thick, and a loop set typically consists of 6, 9, 10, or 12 loops. The loop sets give the push belt assembly high tensile strength with maximum flexibility. A special profile on the inner side of each loop optimizes the lubrication area and minimizes the friction losses between the sliding loops. This design, combined with the selected material, results in a compact system with a high power density and fatigue strength. The steel belt operates in a lubricated condition.

5.2 Belt length and pitch definition

The optimal length of a push belt is determined for each specific transmission layout. The design length of a push belt can vary continuously, as this belt length depends on the diameter of the loop sets and not on the thickness of the elements. Once the diameter of the loops has been determined, a quantity of elements with slightly different thickness is selected to "fill" the belt. The official product code (<element type/width>/<number of loops in each loop set>/<thickness of elements>/<inner diameter of the innermost loop>: e.g., 24/9/1.5/208.8) identifies the diameter of the innermost loop and not the diameter of the belt itself (which is defined by pitch line on the segments). Running radii are always defined on the pitch line height.

5.3 Maximum flank angle

In some applications, the belt needs to change ratio at standstill. In that case, it is not allowed to have a self-locking effect between the element flank angle and the pulley angle (λ). Therefore,

$$\lambda > \arctan(F_f{}^* F_n^{-1}) \tag{5}$$

And because $\mu = F_f{}^* F_n^{-1}$

$$\lambda > \arctan(\mu) \tag{6}$$

(Figure 15)

5.4 Maximal push force

The push force through the elements may never exceed the tensile force in the loop sets. If the push force would exceed the tensile force, the stack of elements would buckle. The tensile force is indirectly applied to the loop sets by the clamping force (F_{clamp}) on the pulleys and the pulley angle λ.

$$F_b = F_{clamp}{}^* \tan(\lambda) \tag{7}$$

5.5 Alignment sensitivity

The alignment is the axial distance of the primary (fixed) shaft running surface relative to the secondary (fixed) shaft running surface. In straight running condition, this distance is exactly the belt width at the pitch line. The fixed running surfaces of the two pulleys are diagonal to each other; with changing ratio, the alignment changes.

As shown in Figure 16, at two ratios the pushbelt runs in a straight line. For proper functioning, the alignment in the transmission should be set as indicated by the belt supplier.

5.6 Belt efficiency

The efficiency of a push belt variator depends on a number of parameters. It varies with ratio, torque, and speed. However, the applied clamping force also has a strong influence. With variation in the safety factor, as explained in formula 1, the clamping force varies.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

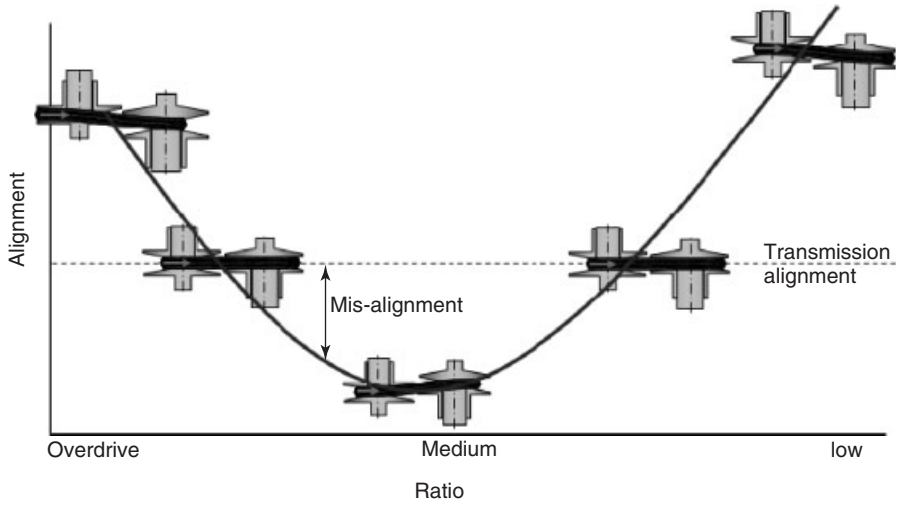


Figure 16. Alignment as function of ratio. (Reproduced by permission of Bosch Transmission Technology B.V.)

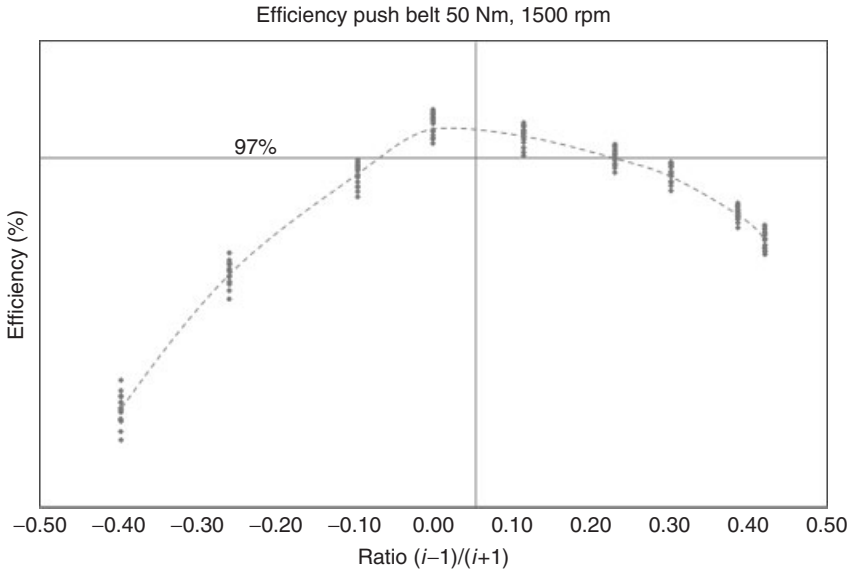


Figure 17. Belt efficiency diagram. (Reproduced by permission of Bosch Transmission Technology B.V.)

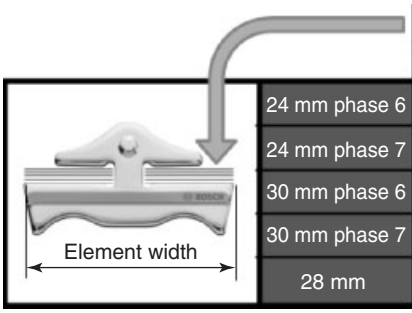
The example that is given for the push belt shows measured results at the given speed and load condition, with a safety factor of 1.3 (Figure 17).

5.7 Belt torque capacity

A push belt uses push force to transmit torque. Because steel can withstand more push force than tensile force, this enables a maximal power density. The contact surface with the pulley is also maximal and has a continuous character, which leads to low noise levels. The belt capacity is mainly determined by stress levels in the loops. These stress levels are a combination of tensile and bending stresses experienced by the belt. Stress is affected by rotational speed, ratio coverage, torque level, distance between primary and secondary pulleys, belt length, belt clamping safety factor, loop material, and belt dimensions. A variety of belt sizes, varying in width and number of loops within the loop set, is available for various applications. Typically, the wider the belt and the more rings in each set are, the higher will

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Torque capacity production belts.



	Number of loops in each loop set			
	6	9	10	12
24 mm phase 6		150 Nm		180 Nm
24 mm phase 7	150 Nm	215 Nm		
30 mm phase 6				350 Nm
30 mm phase 7			350 Nm	
28 mm				410 Nm

be the belt capacity. Table 1 shows some typical torque capacities for push belt configurations. Actual torque capacities depend on the application at hand.

The improvement from phase 6 to phase 7 is an improvement in the purity of band material.

6 CHAIN DESIGN

The chain consists of rocker pins and link plates. The rolling contact between the rocker pins causes a significant reduction in frictional losses. Beside the functional properties (efficiency and overall ratio), it benefits from its modularity. Owing to the construction of link plates and rocker pins, the chain can be matched to the loads and geometrical boundary conditions of the specific application.

The initial applications of the LuK CVT chain were mainly found in the moderate and upper torque range and were covered without exception by the chain type 08 (Figure 18). The number 08 reflects the average pitch distance in millimeters. In the course of discussions on increasing the power density of the subassemblies, the further development of "small" chains (07 and 06) was driven forward and a high performance capacity was identified.

The chain pattern, as shown in Figure 19, can be designed for optimized load distribution. Depending on the arrangement of the link plates, it is possible to reduce the bending of the rocker pins or the load on the outer links. The chain does not require specific alignment measures.

The acoustic behavior, which is impelled by the polygonic effect of the chain, can be improved on the one hand by the chain pattern and on the other using a guide rail. Chain pattern means, in this case, the arrangement of short and long links. If all the links would have the same pitch length, it causes a single tone, which might be easily perceived in a vehicle. By a "randomized" arrangement of the short and long links, this single tone can be eliminated.

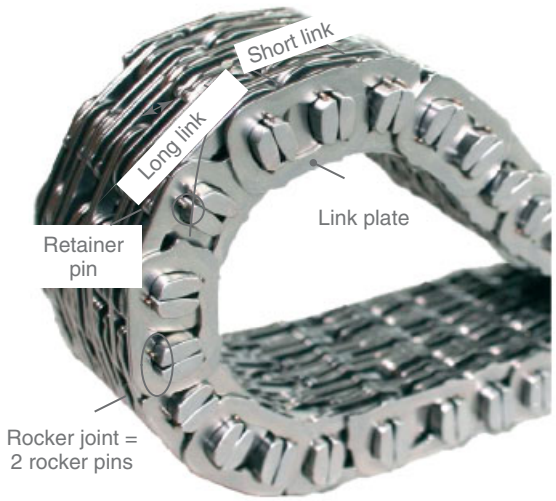


Figure 18. Chain design. (Reproduced by permission of LuK GmbH & Co. oHG.)

The function of the guide rail is to suppress oscillation effects of the free chain section and is in contact with the outside profile of the links during operation (Figure 20).

The specification of an absolute value for the chain's own efficiency is very difficult, as the exact configuration of the adjacent structure (pulley set, etc.) must be considered. However, a measurement of the efficiency profile of chain type 06 (Figure 21) shows the typical efficiency characteristic of a chain variator.

6.1 Chain types

The chain type 07 represents, in simplified terms, a scaling down of the type 08 with additional detailed optimizations. Figure 22 shows the geometrical and performance opportunities for the different chain types.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

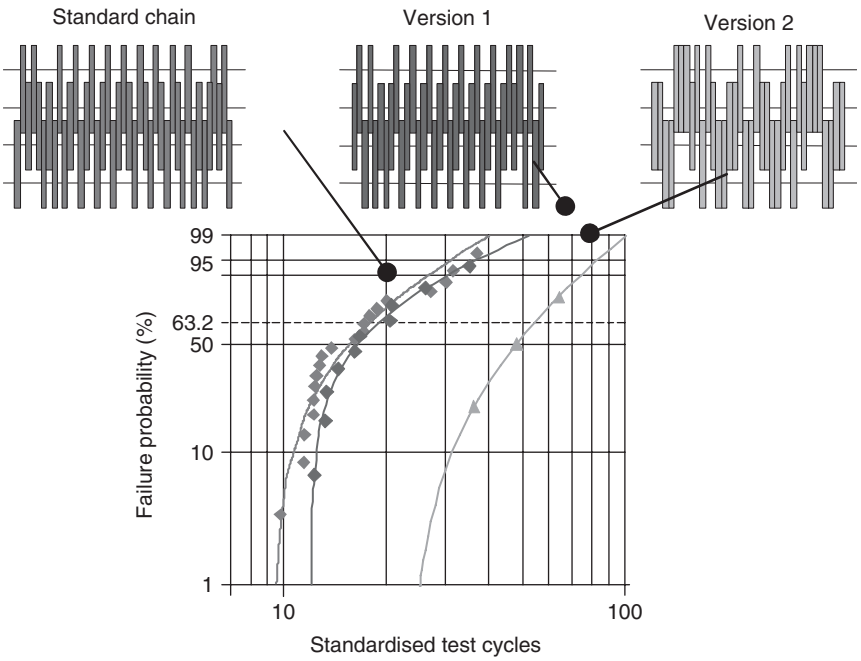


Figure 19. Chain pattern. (Reproduced by permission of LuK GmbH & Co. oHG.)



Figure 20. Guide rail. (Reproduced by permission of LuK GmbH & Co. oHG.)

As described in the preceding section, a reduction in the size of the chain type (transition from 08 to 07) brings a series of possible advantages. In order to answer the question as to what extent the size of the chain can be sensibly reduced further, an even smaller version of chain type 06 was designed and assessed using simulations and tests. As the individual parts become smaller, their strength decreases first. On the other hand, this leads to effects that can have a positive influence on the loading of the chain. Where a smaller chain pitch is present, for example, the frictional force between the chain and the pulley set is split between a larger number of rocker joints.

This leads to more uniform load distribution. Once all these conditions are considered, the chain type reduced in size by 10% (transition from 07 to 06) can be expected to show the performance capacity shown in Figures 22 and 23. In terms of chain width, it achieves approximately the performance capacity of chain type 07. This chain offers applications in passenger cars with low or moderate torque values and even in the two-wheel sector. A motivation in reducing the size of the chain is the possibility of influencing acoustic characteristics. The chain 06 offers a reduced weight and a lower chain pitch, which gives a reduction in the excitation level because of polygonal running. Furthermore, the smaller pitch changes the excitation frequencies.

7 OIL PUMP AND HYDRAULICS

7.1 General hydraulic control

Hydraulic control is normally used for CVT control because of the high power density and the flexibility in the actuation. The components in a CVT require a specific oil flow and oil pressure to function correctly. A pump generates the required hydraulic power, a flow at a certain pressure. The pump flow is pressurized and guided toward the different components by numerous valves, both passively and actively controlled.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

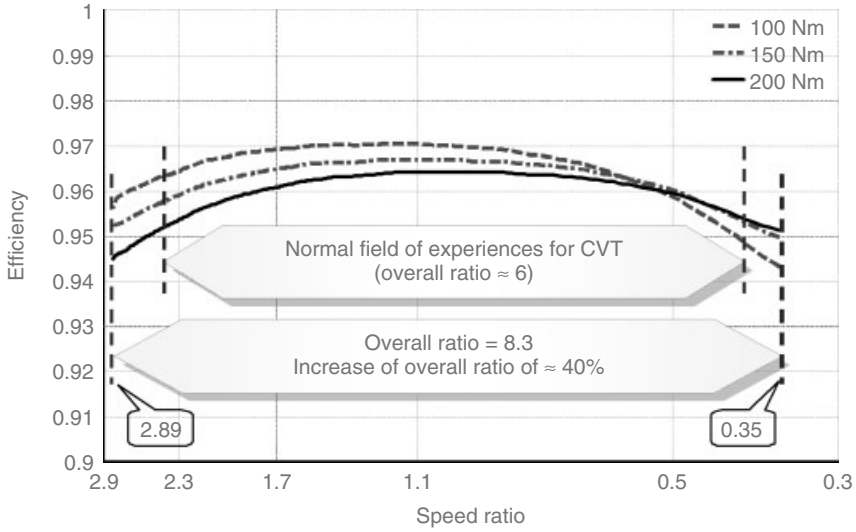


Figure 21. CVT chain efficiency data. (Reproduced by permission of LuK GmbH & Co. oHG.)

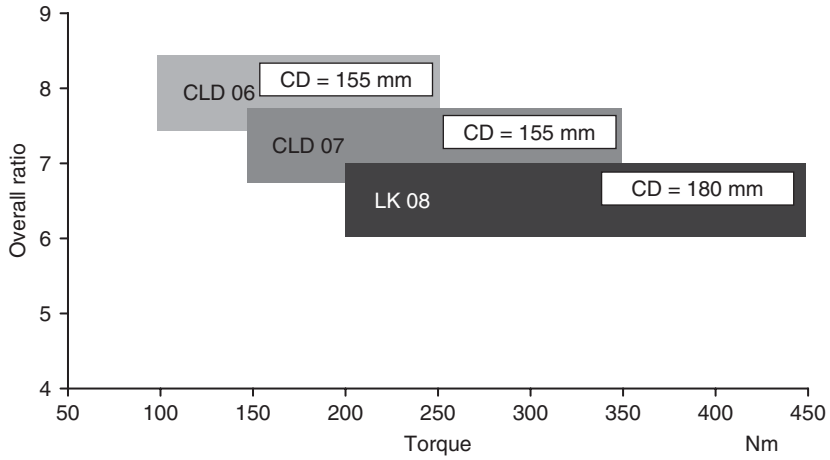


Figure 22. Overall ratio/torque capacity. (Reproduced by permission of LuK GmbH & Co. oHG.)

7.2 Pump requirements

Depending on application and calibration, one of the following conditions (Figure 24) determines the pump size (cc/rev):

- Dynamic variator shifting conditions (Figure 24):

– Emergency stop from overdrive toward underdrive at low engine speed
– Kick-down starting from a low engine speed
– Programmed step-mode shift at low engine speeds.

Coordinated power train control concepts based on drive torque control in some cases lead to more gradual engine speed and ratio transients, occasionally bringing some relief to these requirements.

- Clutch/brake engagement.

Usually, the shifting variator imposes the critical demands in terms of pressure and flow. However, with ever-decreasing idle speed, the idle/launch condition also becomes decisive for the sizing of the engine-driven pump. The pump is driven by the engine, which operates over a large speed range. This speed range directly results in a changing pump flow. To prevent unnecessary large flows into the hydraulic system, pump flow control can be applied. The reduced maximal flow makes it possible to use smaller valves and can reduce parasitic pumping losses.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

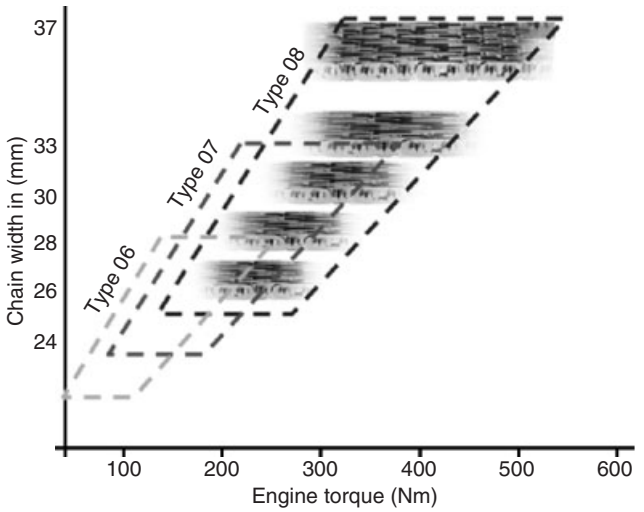


Figure 23. Chain types and capacity. (Reproduced by permission of LuK GmbH & Co. KG.)

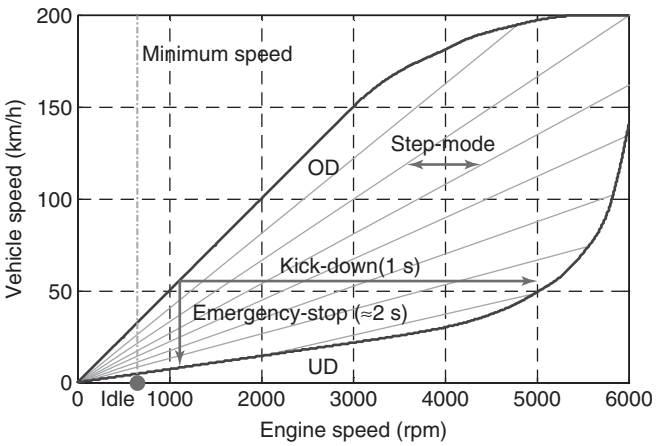


Figure 24. CVT variogram indicating shifts requiring large flow. (From Van Der Sluis, 2003. Copyright © 2003 SAE International. Reprinted with permission.)

7.3 Pump choice

Currently, most CVT pumps are gear or gerotor type pumps based on designs from stepped automatic transmissions. Recently, vane type pumps have been introduced into CVT from the field of power steering technology (Van Der Sluis, 2003).

Typically, leak-tightness of gear and gerotor type pumps is defined by the tolerances of parts that define the critical radial clearance inside the pump. Closer tolerances mean better efficiency but also more expensive parts for which the risk of degrading efficiency due to wear increases. In order to apply these pump types for higher pressure applications such as CVT, cost increasing measures such as radial pressure compensation, additional seals, or complex rotor profiles are required.

In vane pump designs, the discussed leakage is avoided. Centrifugal force on the vanes results in a sealing contact between vane and body, as shown in Figure 25. This increases volumetric efficiency very effectively.

7.4 Pump system

A further advantage of a vane type of pump as shown in Figure 25 is that it basically consists of two independent pumps gathered around a single shaft. This offers the possibility to create flow control by disconnecting one of the pumps from the hydraulic system.

7.5 Typical CVT-related hydraulics properties

The control strategy of a CVT influences transmission efficiency in two ways:

- By the power requirement for the actuation of the variator.
- By the internal losses, which are functions of the applied actuation forces.

Compared to other transmission types, the CVT variator actuation requires relatively high oil pressures and large oil flows, which results in a CVT-typical pump specification.

In Section 3.1, the calculation of the required clamping force (1) has been explained.

Typically, the maximal pump pressure in a CVT varies between 30 and 70 bar.

7.6 Hydraulic variator control layouts

The main and most specific hydraulic function in a CVT is the variator control function. Generally, secondary pulley control is used to control the desired clamping force and primary pulley control for the desired ratio. To have robust clamping force control, the secondary pulley is controlled using pressure control. Primary pulley control can be chosen between flow and pressure controls. There are two principal types of CVT variator control systems (Figure 26).

7.6.1 Dependent control system

The primary pressure is derived from the secondary pressure through a 3/3 directional valve that controls transmission ratio. This simple and low cost circuit has the following features.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

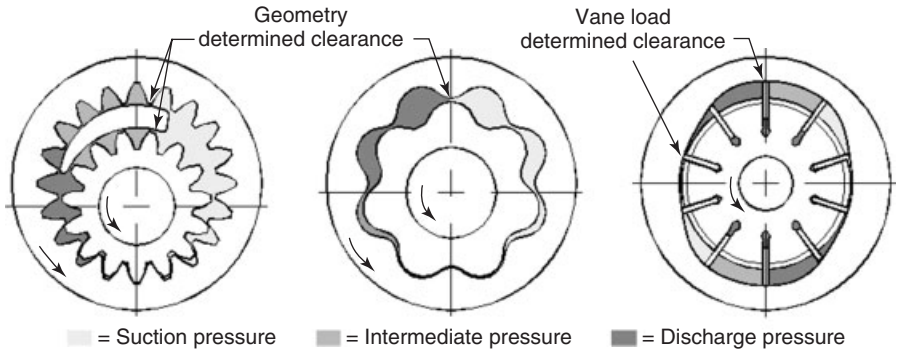


Figure 25. Radial clearances in gear, gerotor, and vane pumps. (From Van Der Sluis, 2003. Copyright © 2003 SAE International. Reprinted with permission.)

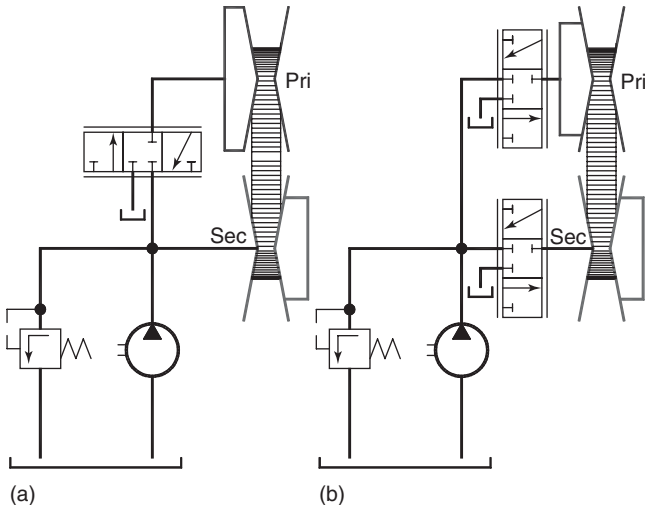


Figure 26. Variator control, dependent (a) versus independent (b) control. (Reproduced by permission of Bosch Transmission Technology B.V.)

- The variator pulley cylinder surface ratio A_{pri}/A_{sec} is set to the maximum KpKs that is required to keep the variator in overdrive plus an additional factor to enable dynamic variator shifting. Fast shifting toward underdrive, during which the secondary pulley cylinder is filled, is most critical in relation to variator slip.
- There is no distinction between the secondary pulley pressure and the pump pressure.
- The minimum secondary pressure is limited by other system pressures, and therefore no extreme low secondary pulley pressure (slip control) (Noll *et al*., 2009) is possible.

7.6.2 Independent control system

The circuit contains one 3/3 valve per cylinder fed by a line pressure control valve. The maximum cylinder pressure switches between the primary and the secondary cylinders. This circuit has the following features.

- The cylinder surface ratio A_{pri}/A_{sec} can be chosen to meet the most frequently occurring KpKs ratio. In this way, primary and secondary pressures lie close to each other most of the time. Moreover, for this circuit, the secondary cylinder surface area is an important parameter to determine pump size. For cost reasons, it can also be decided to use equal cylinder surface areas.
- The pump has to provide all shift flow to give ratio change.
- The secondary cylinder pressure can be reduced to ambient pressure and the pressure in the primary cylinder, when possible.
  As this setup can reduce pulley actuation force to a minimum, this offers the best potential for an "ideal" clamping strategy.
- The line pressure usually is controlled slightly above the maximum of secondary and primary pressures. As primary pressure can rise above secondary pressure, pump pressure will be high more often. Both effects negatively influence the efficiency of the hydraulic system.
  This negative effect can be solved when the hydraulic control is designed in such way that line pressure is directly connected to the highest of primary or secondary pressure. This system is called *SMART* (Van der Sluis *et al*., 2006).

7.7 Global hydraulic scheme layout

Current CVT designs contain several control functions that have to be supplied with pressurized oil. The hydraulic layout can be designed according to a specific priority level (for a typical complete diagram, see Figure 27):

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

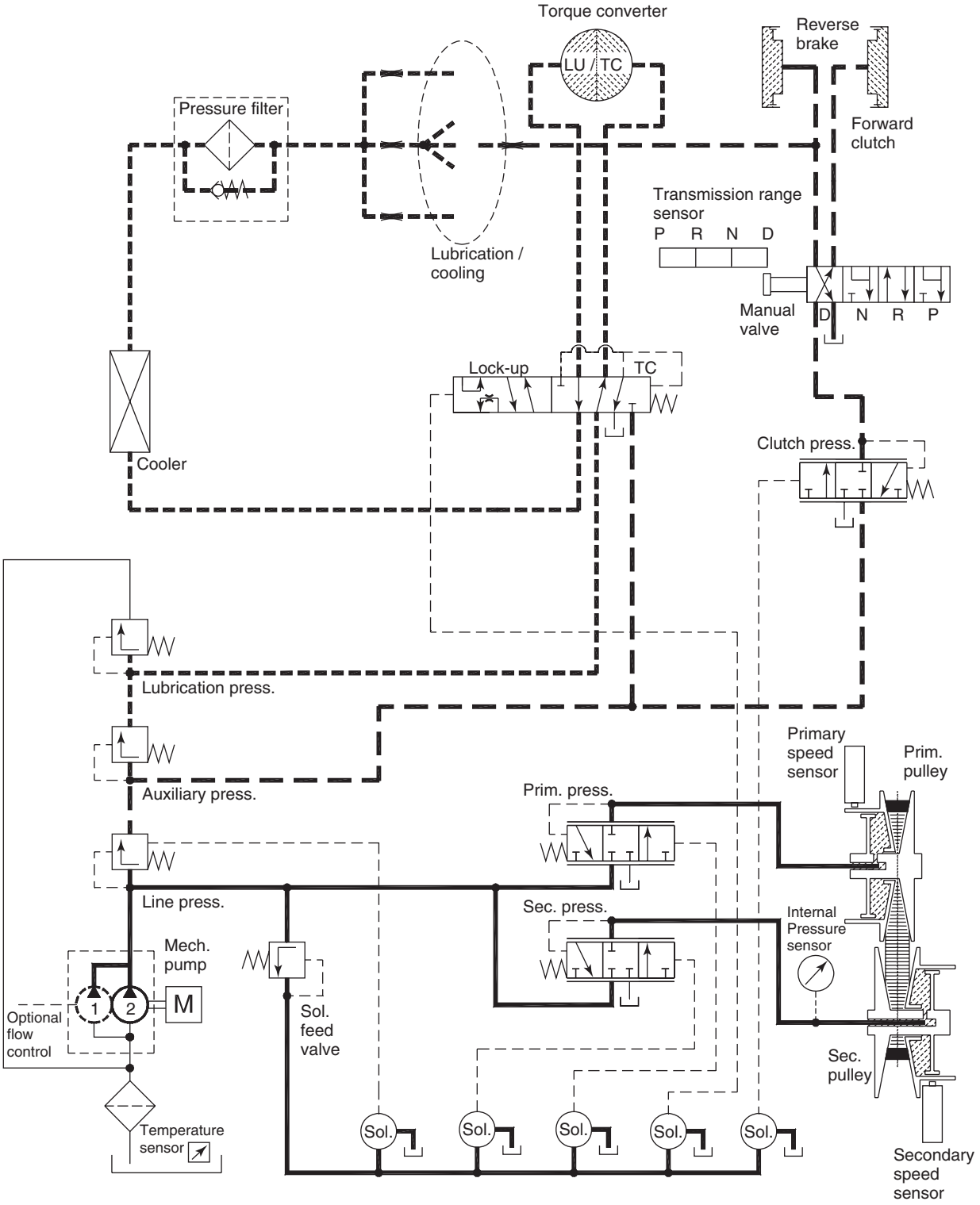


Figure 27. Hydraulic functions in a CVT, cascade system. (Reproduced by permission of Bosch Transmission Technology B.V.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Priority 1: ——— Variator control functions, including solenoid actuation

Priority 2: – – – Auxiliary control for clutch/brake and TC controls

Priority 3: ------ Lubrication control.

In critical flow conditions, pressures will break down in reversed priority level.

This way, the most vital functions, namely variator control and solenoid supply, remain operational as long as possible.

7.8 Control actuators

As shown in Figure 27, typically five main functions (four in case of dependent control) are active/pilot controlled. Different types of solenoids can be used. The solenoids types do differ, that is, oil usage/leakage, accuracy, cost, and noise/vibration/harshness.

7.8.1 Different types of pilot control actuators

- Bleed type solenoid (pressure controlled by bleeding oil)
- Spool type solenoid (3/3 pressure control valve)
- On/off solenoids, controlled by pulse width modulation.

7.8.2 Other types of control

- Direct pressure/flow control (without pilot pressure)
- Stepper motor control (e.g., primary control with mechanical ratio feedback)
- Electromagnets applying direct force to hydraulic valves.

8 CVT ASSEMBLY

In general, the assembly of components in a CVT transmission is comparable to other transmission types. For a CVT, the variator—the combination of pulleys and push belt/chain—is the specific transmission part. Each pulley subassembly comprises two conical sheaves that face each other. One of these sheaves is axially moveable by a controlled pressure. Each pulley has for this reason a pressure chamber.

The contact surfaces on which push belt/chain is clamped contribute to an efficient torque transfer and generate the friction force. Special requirements for the material properties and roughness requirements have to be met.

For proper functioning of the variator, the alignment in the transmission should be set to the value given by the supplier. To keep the tolerance of the alignment value, special attention has to be given to the positioning of both the pulley parts in the transmission during assembly. With normal production spread, shimming on one of the pulleys is often needed to guarantee the alignment value.

During the complete assembly of the transmission, special attention is required to leakage checking of (sub) components. As the variator is hydraulically controlled to generate the required clamping force, any leakage in the system will lead to a less efficient transmission. To control the involved hydraulic components, there is a functional check, generally performed in the production line (Jatco Technical Review No. 11):

1. Performance of the oil pump in delivering the required pressure.
2. Performance of the control valves in regulating the pressure.
3. Absence of any leakage in the oil passages in the transmission from the oil pump to the hydraulically actuated devices.
4. Reliable assembly of pressure-related parts.

A final complete functional transmission check is performed at the end of line testing.

REFERENCES

Brandsma, A., Drogen, M. van Ginkel, E., *et al*. (2003) Van Doorne CVT fluid test: a test method on belt-pulley level to select fluids for push belt CVT applications. SAE 2003-01-3253. Society of Automotive Engineers: Warrendale, USA.

Jatco Technical Review No.11 (2012), Development of Final Test Technologies for Embodying Added Value, Hideyuki Chigira, Japan.

Noll, E., van der Sluis, F., van der Dongen, T., van der Velde, A. (2009) Innovative self-optimizing clamping force strategy for the pushbelt CVT. SAE paper 2009-01-1537. Society of Automotive Engineers: Warrendale, USA.

Van Der Sluis, F. (2003) A new pump for CVT applications. SAE paper 2003-01-3207. Society of Automotive Engineers: Warrendale, USA.

Van der Sluis, F., Van Dongen, T., Van Spijk, G.-J., *et al*. (2006) Fuel Consumption Potential of the Pushbelt CVT. *Proceedings of FISITA 2006 World Automotive Congress no. F2006P218*, Yokohama, Japan, 2006, CD-ROM. http://www.bosch.nl/transmission_technology/en/downloads/FISITA_2006_Fuel_consumption_potential_of_the_pushbelt.pdf (accessed 16 December 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

FURTHER READING

Englisch, A., Fischer, E., Götz, A., *et al*. (2001) The compact high value CVT transmission. GIF 2011.

Society of Automotive Engineers (2000) CVT design guide march 2000. SAE J2525. Society of Automotive Engineers: Warrendale, USA.

Kruessmann, M. (2011) Driving CVT into a new era. VDI 2011.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto083
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Traction Drive CVT

Hirohisa Tanaka
Yokohama National University, Yokohama, Japan

1 Introduction 1
2 Film Thickness, Shear Model, and Spin 3
3 Full- and Half-Toroidal Variators 5
4 Endurance Test of Traction Drive Disks 7
5 Speed Ratio Control System 8
6 Actual Transmission Design 9
Nomenclature 10
References 10

1 INTRODUCTION

The history of the traction drive in detail was given by Heilich III and Schube (1983) in the book of Traction Drives. This chapter gives a brief overview of the history of traction drive continuously variable transmissions (CVTs) applied to automobile transmissions. The GM Research started researching toroidal CVTs in 1928. About the same time as GM developments, the Austin Company in the United Kingdom had launched a similar design under license, and this was produced in small number for several models in 1934. It demonstrated the considerable smoothness and drivability advantages of a CVT, but suffered from reliability problems and was withdrawn from production after 2 years (Gott, 1991).

Their scientific research works were opened in the GM's symposium on the Rolling Contact Phenomena in 1962. Figure 1 shows a research traction test machine for evaluating the performance and durability of fluid and steel materials by Hewko *et al*. in GM (1962).

In 1958, Perbury Engineering in the United Kingdom developed the same type of full-toroidal variator for a small passenger car with a hydromechanical torque control system (Fellows and Greenwood, 1991). Comparing to the full-toroidal CVT, a half-toroidal CVT has a feature of low spin of the elliptical contact, which reduces heat generation and gives a steep traction curve in the microslip region. In 1958, C.E. Kraus invented an equal torque transmission mechanism of multiple rolling elements of the half-toroidal variator by supporting tangential force through oil hydraulic pressure line (Kraus, 1972). This was applied to a small car, but was not commercially sold on the issues of durability of both traction steel materials and thrust ball bearings, and of another reason on the supporting mechanism of multiple rolling elements.

An important component is the traction fluid; Monsanto and Sun Oil provided traction fluid called *Santotrac* in 1968. This oil has high traction characteristics in the range of room temperature; however, it cannot bear high shear traction at the high temperatures for automobile use. Figure 2 shows temperature toughness of a newly developed traction fluid with alicyclic compound, Fluid 2, compared to a fluid with aromatic compound, Fluid 1, at a microslip of 4% test (Tsubouchi *et al*., 1990).

The former has a high stiffness of molecular bonds, as schematically shown in Figure 3.

On the tribological researches, the concept of elastohydrodynamic lubrication principle was proposed by Ertel and Grubin in 1939, and Dowson and Higginson succeeded to obtain spike pressure at the outlet of the contact numerically in 1959 (Hamrock and Dowson, 1981). Their analyses were useful for estimating the thin-film thickness in the elliptical contact. An innovative proposal on the traction force estimation using thermal elastic–plastic fluid model was made

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto084
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

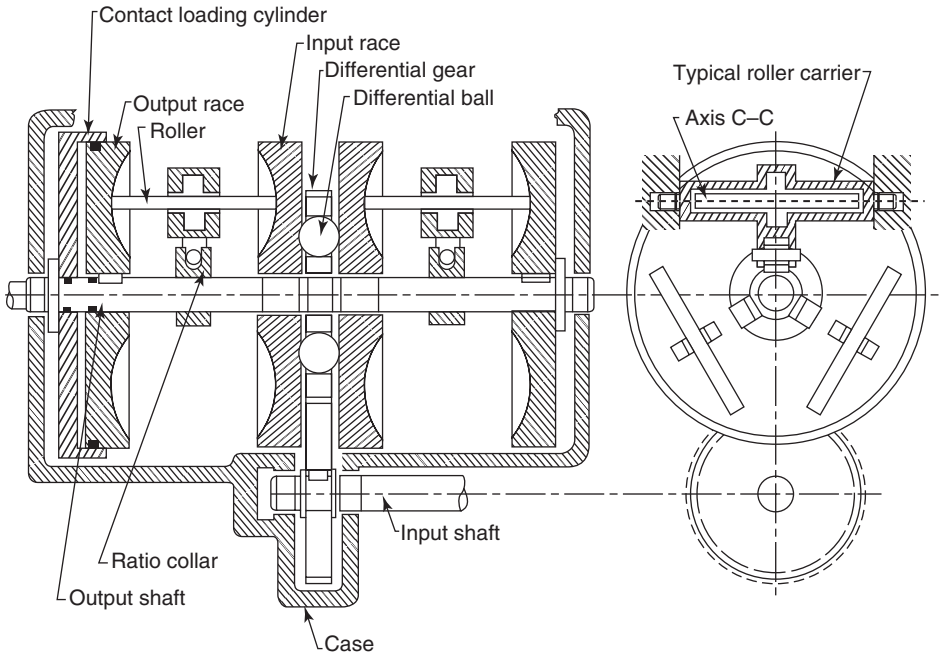


Figure 1. Research traction test machine by GM Research Lab. (Reproduced from Bidwell, 1962. © Elsevier.)

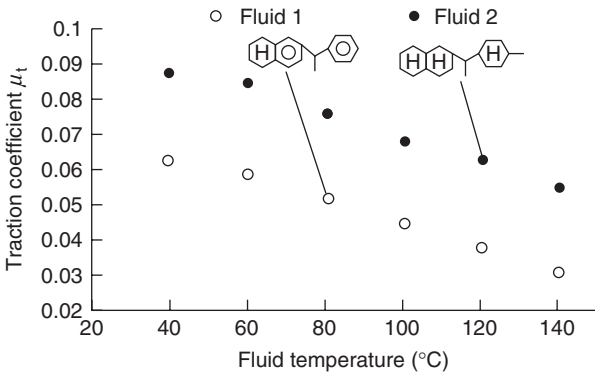


Figure 2. Traction coefficient measurement versus fluid temperature (°C) of two types of fluid with aromatic (Fluid 1) and alicyclic (Fluid 2) compounds. Pure roll test: slide roll ratio of 4%, $u = 4.1$ m/s, and $P_{max} = 1.1$ GPa. (Reproduced from Tsubouchi, 1990. © Elsevier.)

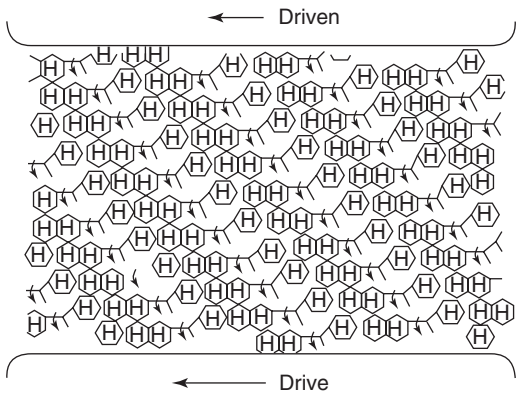


Figure 3. Molecular bonding model of traction fluid. (Reproduced from Tsubouchi, 1990. © Elsevier.)

by Tevaawerk (1980), which made it possible to calculate the efficiency of traction drives theoretically.

The traction drive CVT transmits tangential force F_t through glass-transitioned fluid by applying a contact force F_c between two rolling elements with a microslip as shown in Figure 4.

This microslip is usually called *creep*, $\mathrm{Cr} = (u_1 - u_2)/u_1$, and it is empirically known that the traction coefficient, $\mu_t = F_t/F_c$, increases with Cr in the microslip region and decreases after the saturation peak shown in Figure 5 because of temperature increase by gross slip. The working point is usually within the linear range of $\mathrm{Cr} = 1 \sim 2\,\%$ in Figure 5 and under a maximum traction coefficient of $\mu_t = 0.08$.

The rolling elements are pressurized by loading mechanism over the maximum Hertzian pressure of $P_{max} = 1.5$ GPa at a fluid temperature of $T_\theta = 75^\circ$C.

For designing a CVT, prediction of life is also important. In 1947, Lundberg and Palmgren proposed a stress-volume concept for subsurface-oriented failure of rolling bearings.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto084
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

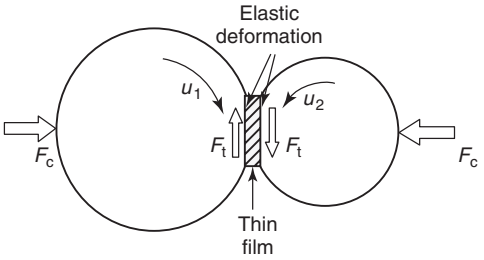


Figure 4. Traction drive image; F_c: contact force and F_t: traction force.

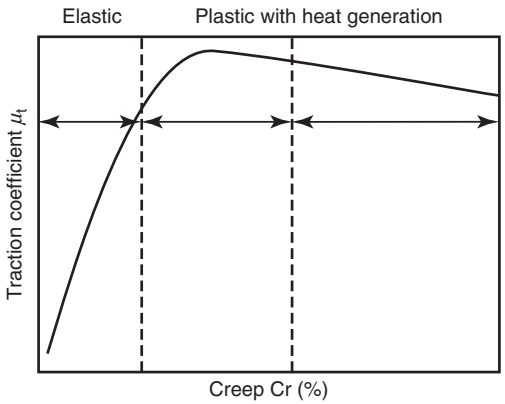


Figure 5. Typical traction curve: relation between creep Cr and traction coefficients μ_t.

Traction drive and ball bearing are definitely different on the tangential force action, but this formulation is simple and makes it possible to estimate the first-order rolling fatigue life (Coy, Loewenthal, and Zaretsky, 1976). Palmgren (1964) made also a database on the loss of several types of bearings. On the research of flaking, Suh (1973) gave many hints for heat treatment and purification of the steel materials.

In Japan, R&D of traction drive CVT for automobile use were started in 1979 and a 3L gasoline turbo passenger car came onto the market in 1999. A chronology of traction drive CVT history is shown in Table 1.

2 FILM THICKNESS, SHEAR MODEL, AND SPIN

2.1 Hertzian pressure in elliptical contact of toroidal variator

The toroidal variator transmits the traction force through the elastically deformed elliptical contact patches as shown in Figure 6 for the half-toroidal variator.

The size of this elliptical contact is calculated using the Hertzian theory (1881). Brewe and Hamrock (1977) proposed a simplified expression on the ellipticity parameter K and elliptic integral of the second kind ξ using equivalent radii R_x (rolling direction) and R_y:

$$\xi \simeq 1.0003 + 0.5968\left(\frac{R_x}{R_y}\right)$$

$$\text{and } K = \frac{a}{b} \simeq 1.0339\left(\frac{R_y}{R_x}\right)^{0.636} \quad (1)$$

Table 1. A chronology of traction drive CVT history.

1877	Patent of toroidal friction drive (Hunt)
1928–1933	Road test of the toric traction drive transmission (GM Research Lab.)
1934	The same time as GM developments the Austin Company in the United Kingdom had launched a similar design under license
1939	Concept of EHL theory (Ertel) and formulation (Grubin)
1947	Rolling fatigue analysis by proposing stress volume principle (Lundberg and Palmgren)
1949	Industrial application of CVT (Kopp Ball Variator)
1958	Road test of full-toroidal CVT mounted on the Hilman Minx (Perbury Engineering)
1959	Road test of half-toroidal CVT mounted on the American Motors Nash Ramber (Kraus)
Principle of EHL theory (Dowson and Higginson)	
1962–1968	Fundamental study on the traction drive characteristics (Hewko)
1968	Santotrac traction fluid onto the market (Monsanto and Sun Oil)
1976	Prediction of rolling fatigue life on the basis of stress volume principle (Coy)
1978	Formulation of the thermal traction force characteristics (Johnson and Tevaarwerk)
1980	Road test of a single cavity half-toroidal CVT mounted on a 1.6L passenger car (Machida)
Open innovation by showing S–N curves of traction drive steel with heat treatment (Machida)	
1985	Stability analysis of a half-toroidal CVT (Tanaka)
1990	Open innovation by showing property of Idemitsu synthetic traction fluid (Tsubouchi and Hata)
1991	Road test of a double-cavity half-toroidal CVT mounted on a 3L turbo passenger car (Nakano)
1999	390 Nm double-cavity half-toroidal CVT mounted on a 3L gasoline turbocharged passenger car came onto the market (NISSAN MOTOR CO, LTD)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto084
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

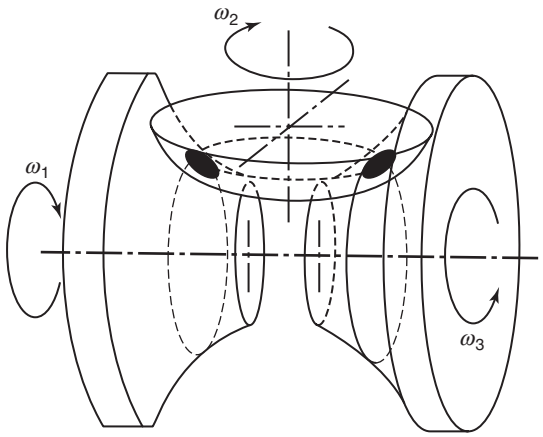


Figure 6. Elliptical contact patches of a half-toroidal variator. (From Tanaka, 2000. Reproduced by permission of Corona Publishing Co. Ltd.)

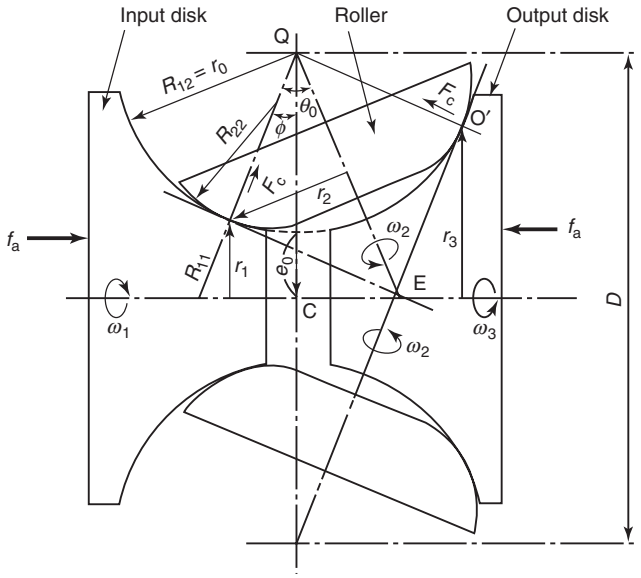


Figure 7. Geometrical parameters of a half-toroidal variator.

As an example of a half-toroidal variator in Figure 7, geometrical parameters are

$$r_{ax} = \frac{r_1}{\cos\varphi} = \frac{r_0(1 + k_0 - \cos\varphi)}{\cos\varphi},$$
$$r_{ay} = -r_0,\ r_{bx} = r_0,\ r_{by} = f_0 r_0 \tag{2}$$

where $k_0 = e_0/r_0$, $f_0 = R_{22}/r_0$, r_0 is the cavity radius and φ the tilting angle.

Then, the equivalent radii at inner (subscript i) and outer (subscript o) conjunctions are expressed by

$$R_{x\mathrm{i}} = \frac{r_0(1 + k_0 - \cos\varphi)}{(1 + k_0)},\ R_{y\mathrm{i}} = \frac{f_0 r_0}{(1 - f_0)},$$
$$R_{x\mathrm{o}} = \frac{r_0[1 + k_0 - \cos(2\theta_0 - \varphi)]}{(1 + k_0)},\ R_{y\mathrm{o}} = \frac{f_0 r_0}{(1 - f_0)} \tag{3}$$

where θ_0 is the half-cone angle of the power roller at the contact patch.

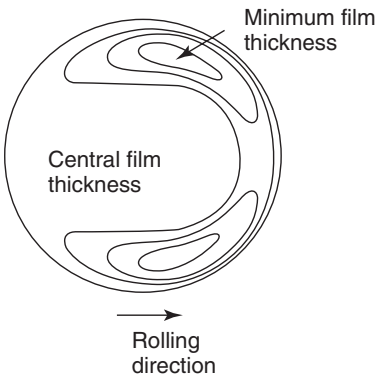


Figure 8. Schematic expression of film thickness. (Reproduced from Höglund, 1984. © Luleå University of Technology.)

2.2 Film thickness and traction image

Höglund (1984) succeeded in visualizing the film thickness between a rolling steel ball and a sapphire disk lubricated by traction fluid. The schematic contour line of film thickness is shown in Figure 8. It can be seen that the flat area of central film thickness in the center and minimum film thickness at both edges.

Hamrock and Dowson (1981) expressed this central film thickness H_c using nondimensional parameters of speed U, material G, ellipticity K, and load W.

Figure 9 is an image of traction force transmission through the thin film and Figure 10 is a measured example

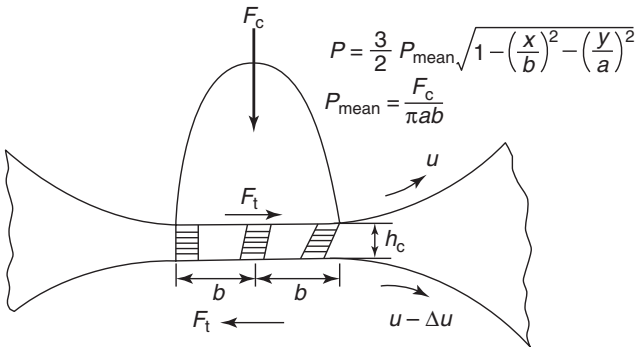


Figure 9. Image of traction force transmission through the thin film.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto084
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

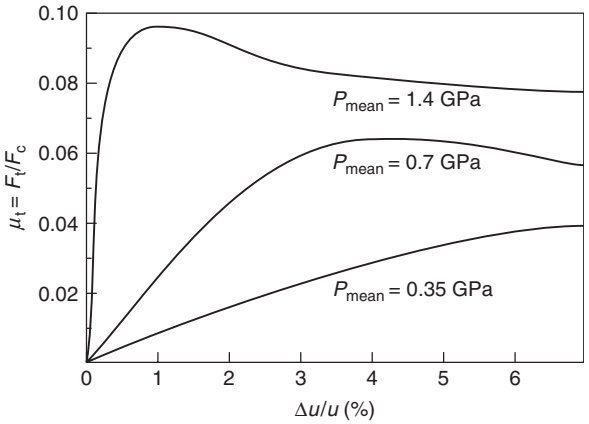


Figure 10. Example of traction curve measurement of the Santotrac 50 fluid, $u = 24$ m/s and $T_\theta = 75^\circ$C. (Reproduced with permission from Kraus, 1972. © C.E. Kraus.)

of a traction curve (Kraus, 1972). In these examples, it should be noted that the contact patch should be pressurized over 1.4 GPa for traction drive.

2.3 Shear model

The traction curve is important to predict a margin before gross slip. The curve is a relationship between microslip and traction coefficient. The traction coefficient is calculated theoretically by including microslip and spin on the elliptical contact. The shear stress is separated into an elastic region and a plastic region with heat generation (Tevaarwerk and Johnson, 1979). The shear stress of the thin film increases with creep linearly in the elastic region, but saturates at a lower value of limiting shear stress as shown in Figure 11.

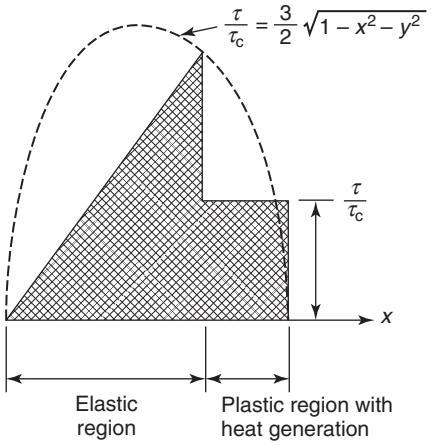


Figure 11. Elastoplastic shear stress model. (Adapted with permission from Tevaarwerk and Johnson 1979. © ASME.)

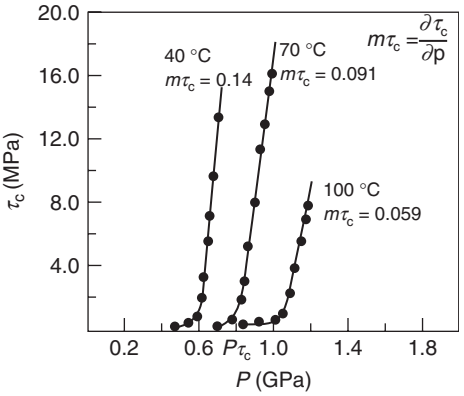


Figure 12. Measurement of limiting shear stress τ_c against pressure P for three temperatures of the Santotrac 50 fluid. (Reproduced from Höglund, 1984. © Luleå University of Technology.)

The grade of deterioration is calculated by iteration between the heat transfer mechanism of the rolling element and the limiting shear stress decrease of the traction fluid. An example of the glass-transition pressure and limiting shear stress of traction fluid as measured by Höglund is shown in Figure 12.

2.4 Spin

There are three major shearing motions in the elliptical contact. Creep in the rolling direction is the main component for traction force generation; sideslip perpendicular to the rolling direction plays an important role of generating the tilting force of the rolling element for speed ratio change; and spin normal to the elliptical surface decreases limiting shear stress of traction fluid because of heat generation. The spin occurs in the geometrical layout, as shown in Figure 13, by the pitch line of the contact not crossing the intersection of the rotating axes of the two rotating elements.

The film motion with spin is expressed schematically in Figure 14.

The equivalent traction force with spin is transmitted at an offset point of force pole from the elliptical contact center by f_p (Mägi, 1974).

3 FULL- AND HALF-TOROIDAL VARIATORS

3.1 Geometry

Full- and half-toroidal variators are characterized by the geometrical configuration of cone angle of the roller. In Figure 15, the cone angle of the full toroidal is $2\theta_0 = 180^\circ$,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto084
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

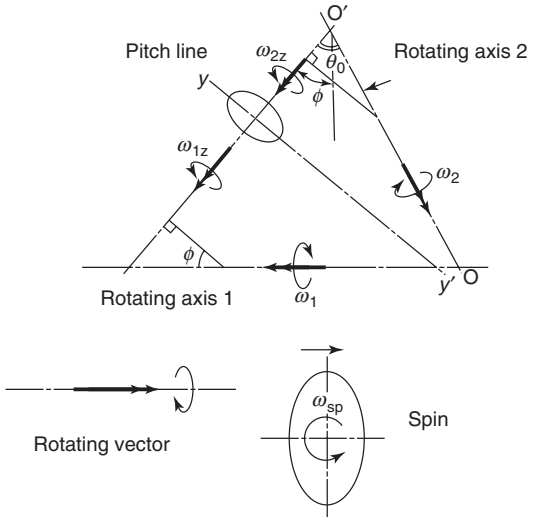


Figure 13. Spin: rotating motion of ellipsoid; the pitch line of the contact not crossing the intersection of the rotating axes of the two rotating elements.

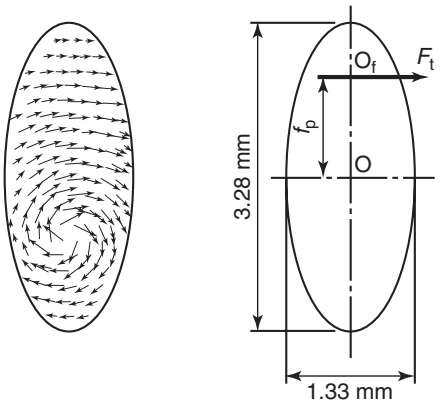


Figure 14. Schematic shearing model with spin and force pole O_f.

and, on the other hand, the half toroidal is smaller than 180°.

The choice of cone angle affects on the transmission efficiency because of spin and thrust forces on the rollers. The full toroidal has the advantage of no thrust force, but the disadvantage of high spin on the traction contact because the rolling axis of the roller is parallel to the pitch line of the elliptical contact at all tilting angles of the roller. On the other hand, the half toroidal has the advantage of small spin on the elliptical contact, but the disadvantage of supporting high speed and large thrust force of the roller. These mean that the full toroidal has issues on speed transmission efficiency, whereas the half toroidal has issues on torque transmission efficiency (Tanaka *et al*., 1995).

3.2 Comparison of axial loading force and traction curves

It is significant to compare both toroidal variators on the traction curve and loading force by changing the cone angle. As an example, by fixing the speed ratio range of $e_s = 0.5:1$ to $2:1$, disk radius of $r_0 = 40$ mm, transmitted torque of $T_1 = 340$ Nm, a working traction coefficient of $\mu_t = 0.05$, and the same contacting size of disk radii of r_1 and r_3 in Figure 16. The loading force is compared and shown in Figure 17, where it can be seen that the

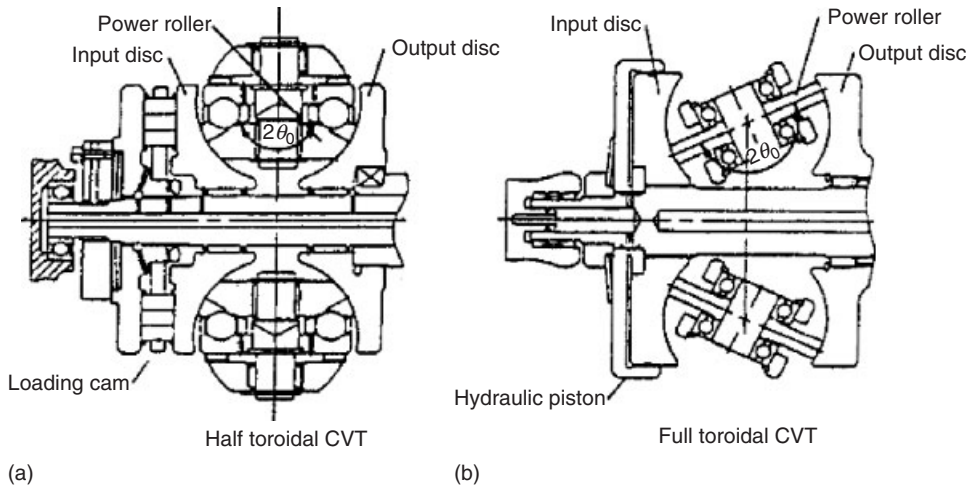


Figure 15. Configuration of (b) full- and (a) half-toroidal variators. (From Imanishi, Machida and Tanaka, 1996. Reproduced by permission of Takashi Imanishi.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto084
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

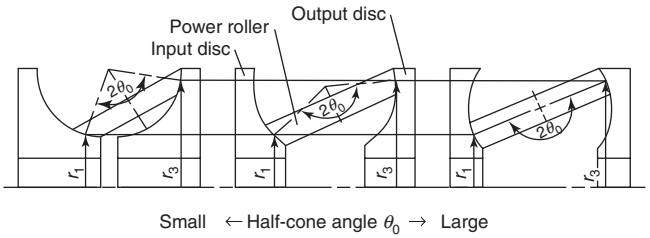


Figure 16. Schematic comparison of toroidal geometry by changing cone angles at a same disk size. (From Imanishi, Machida and Tanaka, 1996. Reproduced by permission of Takashi Imanishi.)

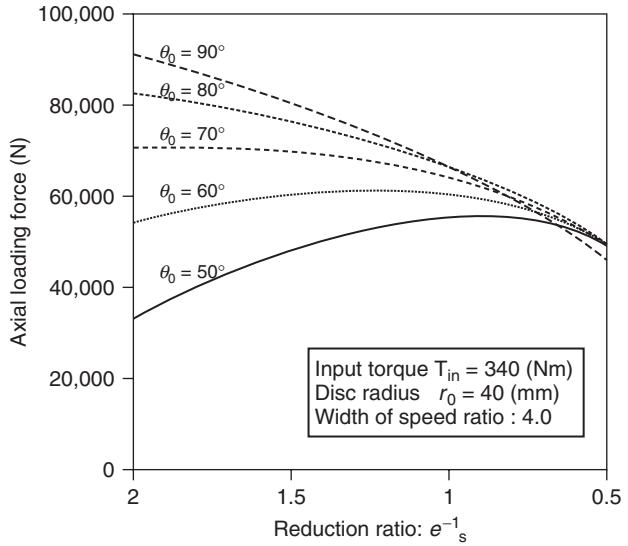


Figure 17. Axial loading force comparison by changing cone angles for transmitting torque of $T_1 = 340$ Nm at a traction coefficient of $\mu_t = 0.05$ in whole reduction speed ratio range; full toroidal ($\theta_0 = 90^\circ$) and half toroidal ($\theta_0 < 90^\circ$). (From Imanishi, Machida and Tanaka, 1996. Reproduced by permission of Takashi Imanishi.)

full-toroidal variator, $\theta_0 = 90^\circ$, needs double the loading force at maximum reduction ($1/e_s = 2$) against maximum speed increase ($1/e_s = 0.5$). The full toroidal cannot use mechanical loading, but only use a hydraulic loading system as shown in Figure 15b. On the other hand, the design of half-cone angle of $\theta_0 = 60^\circ$ keeps it to be almost constant in whole speed ratio range, which makes it possible to use a torque proportional cam loading system as shown in Figure 15a (Imanishi, Machida, and Tanaka, 1996).

The effect of cone angles on the traction curves are also shown in Figure 18. These are calculation examples at a working condition of speed ratio $e_s = 1$, input torque $T_1 = 340$ Nm, input rotational speed $N_1 = 4000$ rpm, cavity radius $r_0 = 40$ mm, and temperature $T_\theta = 75^\circ$C. The traction

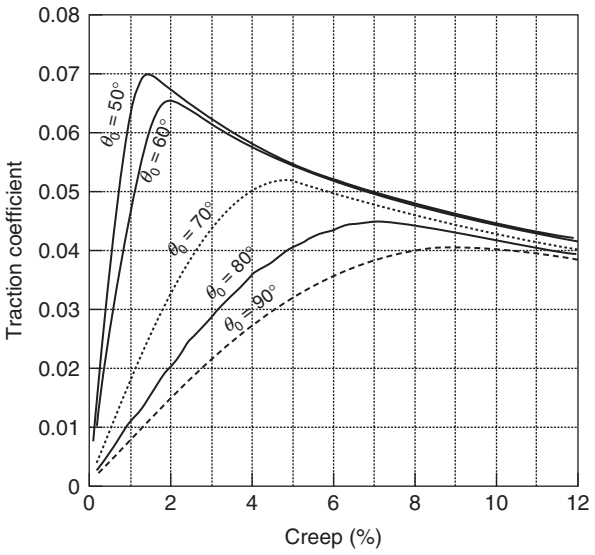


Figure 18. Effect of half-cone angles θ_0 on the traction curves; calculation at $e_s = 1$, input torque of $T_1 = 340$ Nm, input rotational speed of $N_1 = 4000$ rpm, cavity radius of $r_0 = 40$ mm, and temperature $T_\theta = 75^\circ$C. (From Imanishi, Machida and Tanaka, 1996. Reproduced by permission of Takashi Imanishi.)

curve of the half-toroidal variator is steep compared to the full toroidal because of low spin on the elliptical contact, which makes it possible to select a working point with higher traction coefficient.

4 ENDURANCE TEST OF TRACTION DRIVE DISKS

Rolling fatigue occurs usually on the driven elements because of the large tensile forces acting on the traction surface. Heat treatment giving compressive residual stress in the surface is one useful measure for elongating the life. It is well known that there are two types of flaking. One is subsurface flaking, due to reciprocal shear stress in the material, and the other is surface flaking, due to large tangential force transmission. In both cases, the original point of flaking is the place where harmful non-ferrous hard inclusions exist such as Al_2O_3 or $CaO \cdot Al_2O_3$. After removing these inclusions, surface flaking occurs because of the decrease of viscosity of traction fluid or the depletion of compressive residual stress of the steel by cyclic strong shearing. Figure 19 shows endurance test results of four half-size toroidal variators of 9–73.5 kW capacity (Machida, Aihara, and Tanaka, 1991).

In the figure, the black circles denote occurrence of flaking due to the decrease of viscosity of traction fluid by severe shearing in early developmental stage. Figure 20

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto084
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

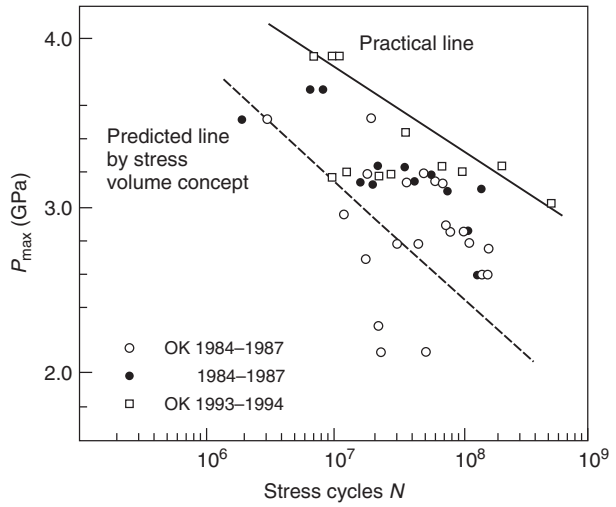


Figure 19. Endurance test of a half-toroidal variator (black circles are failed). (From Machida, Aihara and Tanaka, 1991. Reproduced by permission of Hisashi Machida.)

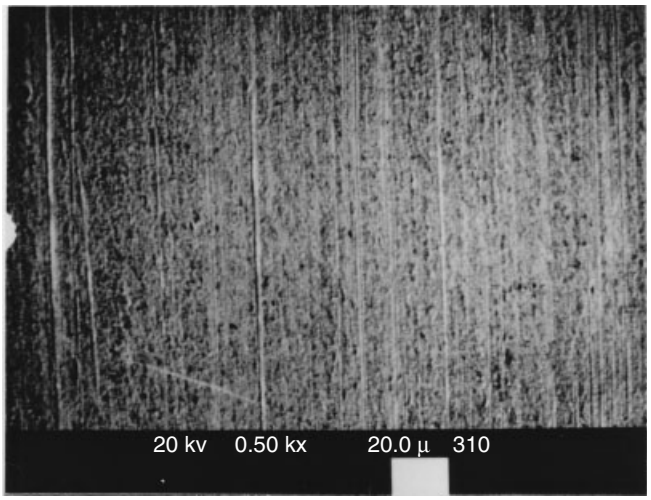


Figure 21. Center of the running track magnified 500 times (Machida, Aihara, and Tanaka, 1991); straight lines of grinding pattern of 0.2 μm in roughness still remains. (Reproduced by permission of Hisashi Machida.)

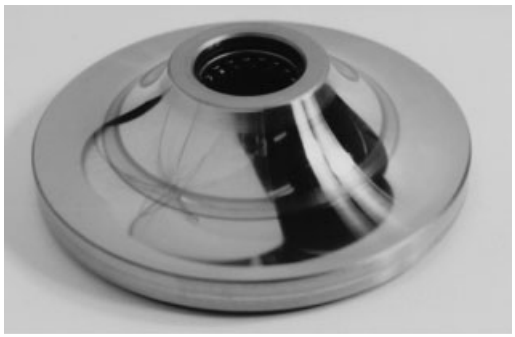


Figure 20. Sound disk after 10^8 running traction drive; transmission power of 73.5 kW at $P_{max} = 2.15$ GPa and 110°C. (Reproduced by permission of Hisashi Machida.)



Figure 22. Speed ratio changing diagram: tilting the rollers changes the ratio between the outer driving disks and the center-driven disks. (From Tanaka, 2000. Reproduced by permission of Corona Publishing Co. Ltd.)

is a sound disk photo after 10^8 running of 73 kW power transmission at $P_{max} = 2.15$ GPa.

Figure 21 is a photo of the center of running track magnified 500 times (Machida, Aihara, and Tanaka, 1991), in which straight lines of the grinding pattern of 0.2 μm in roughness still remain. This means that the traction film separates both rolling elements.

5 SPEED RATIO CONTROL SYSTEM

The variator changes its speed ratio by controlling the tilting angle of the rollers as shown in Figure 22.

The tilting force is induced by giving sideslip on the elliptical contact. The sideslip is another microslip perpendicular to the rolling direction. Figure 23 shows that the offset x of

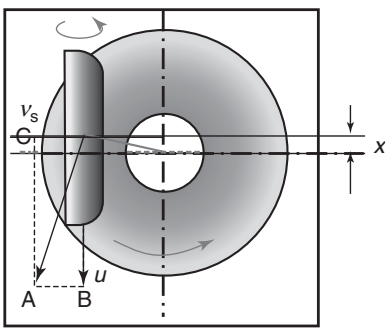


Figure 23. Sideslip (v_s) generation by giving offset (x) to the roller. (From Tanaka, 2000. Reproduced by permission of Corona Publishing Co. Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto084
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

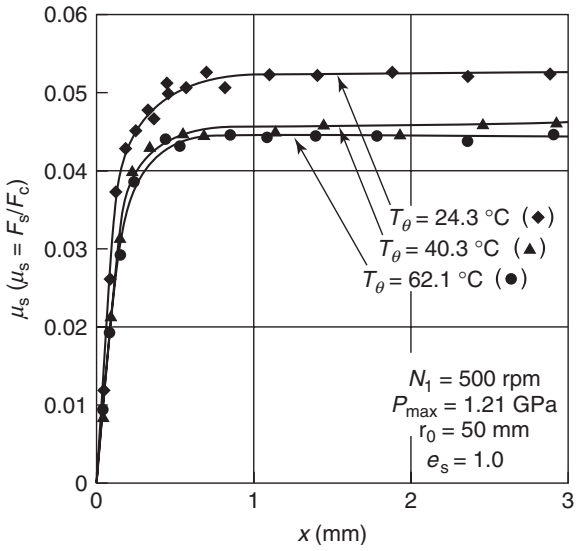


Figure 24. Measurement of nondimensional sideslip force at one contact. (From Tanaka, 2000. Reproduced by permission of Corona Publishing Co. Ltd.)

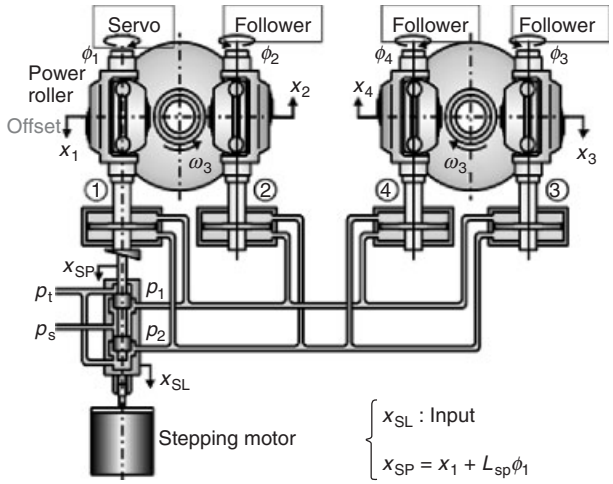


Figure 25. Speed ratio control system of a half-toroidal variator with four rollers; equal traction force transmission is achieved using hydraulically supporting pistons, and the servomechanism has two feedback loops of offset and tilting angle through a precise cam with gain L_{sp}. (From Tanaka, 2000. Reproduced by permission of Corona Publishing Co. Ltd.)

the rolling center gives rise to the sideslip. The sideslip v_s (direction C) is caused by the difference of rolling directions between the disk (direction A) and the roller (direction B).

Figure 24 is a measured example of the relation between the offset and the nondimensional sideslip force μ_s normalized by contact force. It can be seen that the offset x is adequate to be 0.5 mm at a maximum to prevent the decrease of traction force (Tanaka, 2000).

The control principle of half toroidal and full toroidal is different. The half toroidal controls the speed ratio of the variator, whereas the full toroidal controls torque in relation to the loading pressure. The half-toroidal rollers are supported by a hydraulic piston with offset and tilting angle control servomechanism as shown in Figure 25.

On the other hand, the full-toroidal rollers are supported by hydraulic pistons with caster angle without the need for a tilting angle control servomechanism as in Figure 26 (Fucks, Hasuda, and James, 2004) and hence, give speed ratio automatically.

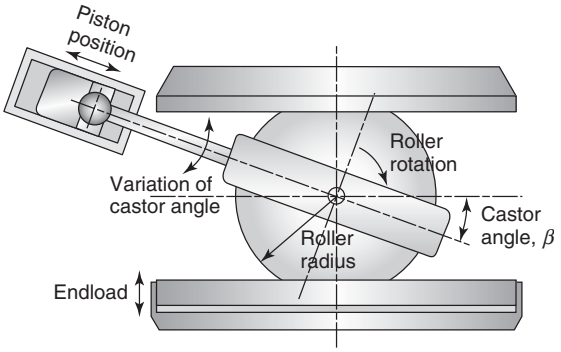


Figure 26. Torque control system of a full-toroidal variator; the roller supported by a hydraulic piston with caster angle. (Reproduced from Fucks, Hasuda and James, 2004. © R. Fucks, Y. Hasuda and I. James.)

6 ACTUAL TRANSMISSION DESIGN

A half-toroidal CVT was mounted in a 3L gasoline turbo passenger car in 1999 and remained in series production for several years. The transmission layout and the cross-sectional view are shown in Figures 27 and 28.

There are two toroidal cavities with two roller assemblies in each transferring drive from the end disks to central output disks. The input to the transmission is provided by

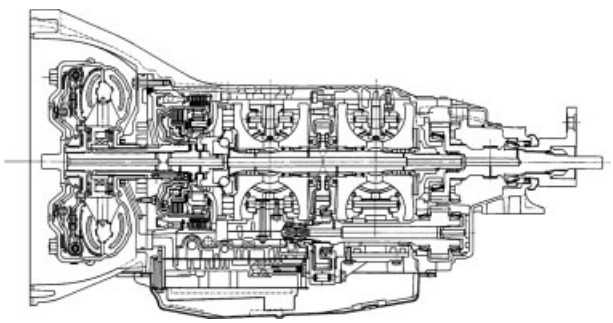


Figure 27. Cross-sectional view of half-toroidal CVT for 3L turbo-gasoline FR car with 387 Nm/4000 rpm, 280PS/6500 rpm, and speed ratio coverage 4.36. (Reproduced by permission of JATCO Ltd and Nissan Motor Co., Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto084
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

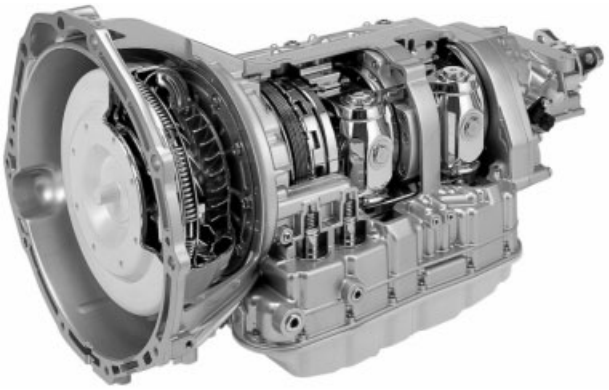


Figure 28. Photo of inner structure of half-toroidal CVT. (Reproduced by permission of JATCO Ltd and Nissan Motor Co., Ltd.)

a torque converter that acts as a starting device, as well as increasing the ratio range. The CVT has two hydromechanical speed ratio control mechanisms for forward and reverse drives. Official data on the fuel economy at a Japanese 10–15 mode measurement is 9.7 km/L, which is 8% higher than the same model with 4AT.

NOMENCLATURE

Cr	creep, $\mathrm{Cr} = (u_1 - u_2)/u_1$
e_s	speed increase ratio ($e^{-1}{}_s$: reduction ratio)
F_c	contact force
F_t	tangential force
r_0	cavity radius
v_s	sideslip (microslip perpendicular to the rolling direction)
x	offset of the rolling center of the roller for generating sideslip
θ_0	half-cone angle
μ_s	nondimensional sideslip force normalized by contact force
μ_t	traction coefficient, $\mu_t = F_t/F_c$
τ_c	limiting shear stress of traction fluid
φ	tilting angle of roller
ω_{sp}	spin (rotating motion of ellipsoid)

REFERENCES

Bidwell, J.B. (1962) Rolling Contact Phenomena. *Proceedings of a Symposium held at the GM Research Laboratories*, Elsevier, p. 160.

Brewe, D.E. and Hamrock, B.J. (1977) Simplified solution for eliptical contact deformation between two elastic solids. *Journal of Lubrication Technology*, **99**(4), 485–487.

Coy, J.J., Loewenthal, S.H., and Zaretsky, E.V. (1976) *Fatigue Life Analysis for Traction Drives with Application to a Toroidal Type Geometry*, NASA TN D-8362, Cleveland, OH.

Fellows, T.G. and Greenwood, C.J. (1991) The design and development of an experimental traction drive CVT for a 2.0L FWD passenger car. SAE Paper 910408.

Fucks, R., Hasuda, Y., and James, I.B. (2004) Dynamic performance analysis of a full toroidal IVT — a theoretical approach. CVT Congress 04CVT-30, Sacrament.

Gott, P.G. (1991) *Changing Gears: The Development of the Automotive Transmission*, SAE Inc., Warrendale. ISBN: 1-56091-099-2.

Hamrock, B.J. and Dowson, D. (1981) *Ball Bearing Lubrication: The Elestohydrodynamics of Elliptical Contacts*, John Wiley & Sons, Inc., Hoboken. ISBN: 471-03553-X.

Heilich, F.W., III, and Schube, E.E. (1983) *Traction Drives: Selection and Application*, Marcel Dekker Inc, New York.

Höglund, E. (1984) Elstohydrodynamic lubrication, Interferometic measurements, lubricant rheology and subsurface stress. Doctoral thesis. 32D, Luleo Univ. of Tech.

Imanishi, T., Machida, H., and Tanaka, H. (1996) *A Geometric Study of Toroidal CVT — A Comparison of Half Toroidals and Full Toroidals*, CVT'96, Yokohama, JSAE 9636411, Tokyo, p. 107.

Kraus, C.E. (1972) *Rolling Traction Analysis and Design*, Excelermatic, Inc., Austin, TX.

Machida, H., Aihara, S., and Tanaka, H. (1991) Oil film and surface damage in traction drive for automobiles. *JSME International Conference on Motion and Power Transmission*, Hiroshima.

Mägi, M. (1974) On efficiencies of mechanical coplaner shaft power transmissions. Dr. Dissertation. p. 38, Chalmers Univ. of Tech.

Palmgren, A. (1964) *Grundlagen der Wälzlager Technik*, Frank'sche Ver., Stuttgart.

Suh, N.P. (1973) The delamination theory of wear. *Wear*, **25**, 111–124.

Tanaka, H. (2000) *Toroidal CVT*, Corona Publishing Co. Ltd, Tokyo, p. 58 (in Japanese). ISBN: 4-339-04550-0.

Tanaka, H., Machida, H., Hata, H., and Nakano, M. (1995) Half-toroidal traction drive CVT for automobiles — traction drive materials, transmission design and efficiency. *JSME International Journal Series C*, **38**(4), 772–777.

Tevaawerk, J.L. and Johnson, K.L. (1979) The influence of fluid rheology on the performance of traction drive. *Journal of Lubrication Technology*, **101**, 266.

Tevaarwerk, J.L. (1980) Thermal Influence on the Traction Behavior of Elastic/Plastic Model. *Proceedings of the 7th Leeds-Lyon Symposium on Tribology*, 302–309.

Tsubouchi, T., Hata, H., Abe, K. *et al*. (1990) Development of New Traction Fluid for Automobile Use. *Proceedings of the 17th Leeds-Lyon Symposium on Tribology*, 439–443.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto084
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Torque Transfer with AWD Systems

John A. Barlage, Todd L. Perttola, Larry Pritchard, and Tom Foster
BorgWarner, Inc, Auburn Hills, MI, USA

1 Introduction 1
2 Transfer Cases for RWD-Based AWD 1
3 Power Transfer Units for FWD-Based AWD 7
4 Couplings for on-Demand AWD 9
5 Conclusion 16
Further Reading 16

1 INTRODUCTION

Historically, AWD (all-wheel drive) was offered almost exclusively in pickup trucks and sport utility vehicles (SUVs), which were purchased primarily for their cargo loads and all-terrain capabilities (such as traction and mobility). In recent years, the demand for AWD systems has expanded. Drivers of passenger cars and crossover utility vehicles (CUVs) are demanding the technology for both traction and enhanced vehicle stability and handling. As the applications of AWD become more widespread, the technology becomes more complex, as is demonstrated in the following sections.

1.1 RWD-based AWD

In order to adapt a rear-wheel drive (RWD) vehicle to AWD, a transfer case is required to transfer power from the transmission to the front secondary drive axle. The transfer case is mounted between the transmission and the rear axle. The function of the transfer case is to transfer and manage torque from the transmission to both the front and the rear axles. A chain or geartrain (usually helical gears) is used to drive the front axle. Some transfer cases have some type of torque management device to control the torque distribution to the front and rear axles.

1.2 FWD-based AWD

In front-wheel drive (FWD) vehicles, the engine is typically mounted transversely or what is more commonly known as an *east–west* powertrain layout. In order to adapt the vehicle to AWD, a power transfer unit (PTU) is mounted to the differential case of the transaxle. This unit takes rotational motion from the transaxle and turns it 90° to drive a longitudinal propshaft that is attached to a secondary axle through a torque-coupling device.

2 TRANSFER CASES FOR RWD-BASED AWD

This section provides an in-depth analysis of the functional characteristics and construction of various transfer case types, used for RWD-based AWD. The transfer case is usually mounted directly to the back of the transmission and manages the torque distribution to the front and rear axles (Figure 1).

2.1 Part-Time AWD

Part-time AWD usually has a number of different operating ranges (high, low, and neutral) and modes (two-wheel and four-wheel) and is typically used in truck/SUV applications. The driver may select two-high (2H), four-high (4H), four-low (4L), and neutral.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

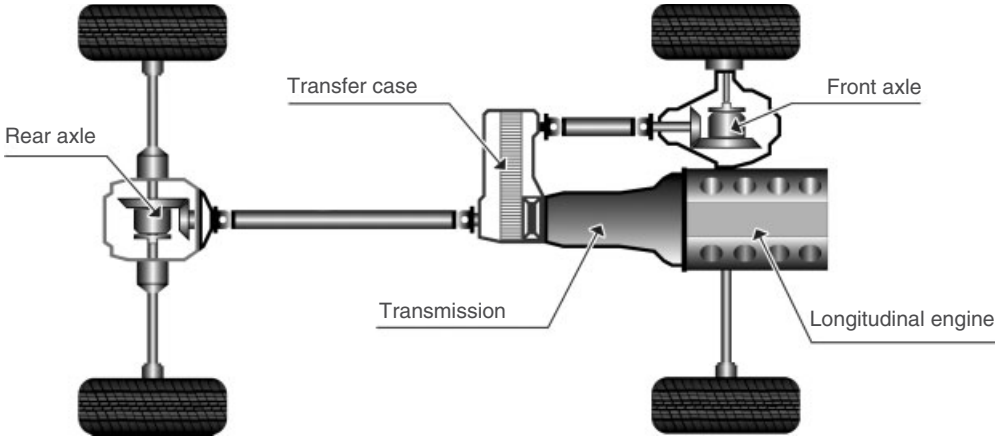


Figure 1. Typical RWD-based AWD architecture. (Reproduced by permission of BorgWarner, Inc.)

Two-high is normally used when driving on paved, dry roads. In 2H, only one axle (one set of wheels) is being driven (two-wheel drive). Four-high is used when additional traction is required, such as driving on ice or snow-covered roads or when operating off-road where the terrain is relatively flat without significant irregularities. In 4H, both axles (driving all four wheels) are mechanically connected and driven. The 4L mode operates similarly to 4H, but with a reduced gear ratio that multiplies the input torque.

Figure 2 shows the typical components of a part-time AWD transfer case. The operating modes of this type of transfer case are 2H, 4H, and 4L. Some systems will also have a neutral mode.

The 2H mode, used on paved, dry roads, provides power to only the rear axle and therefore the maximum efficiency to propel the vehicle. Auxiliary axle disconnect systems are often used when operating in 2H to stop all or part of the front axle from rotating to improve fuel economy and

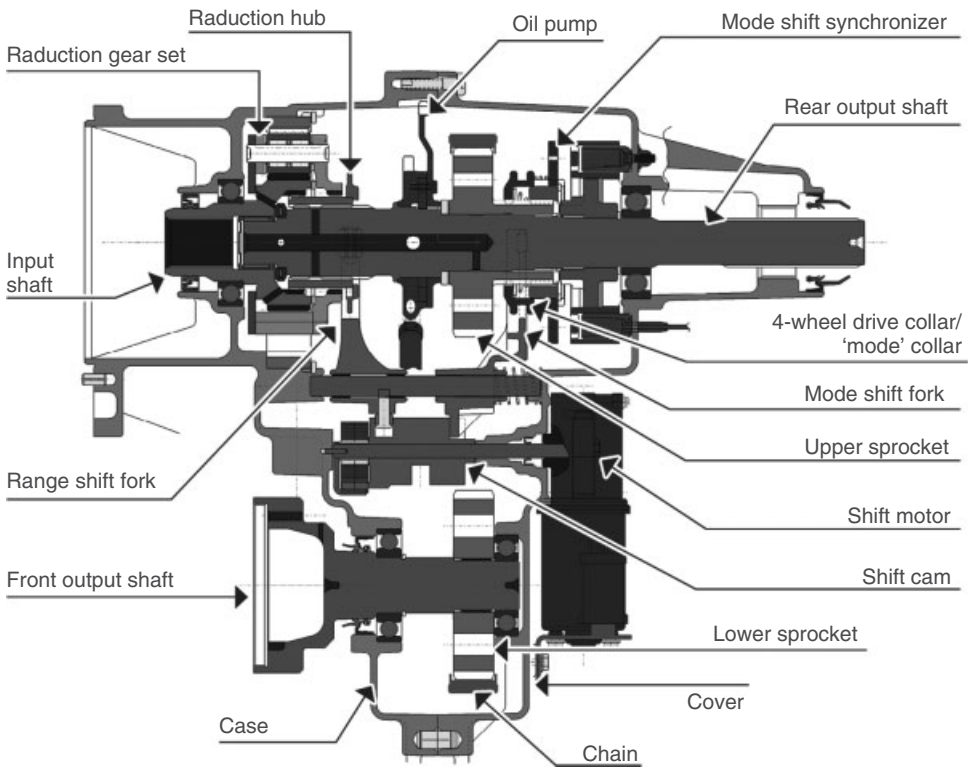


Figure 2. Typical transfer case components. (Reproduced by permission of BorgWarner, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

reduce noise, vibration, and harshness (NVH). In 2H, the front and rear wheels can rotate at different speeds because the front and rear axles are not locked together.

In 2H, power goes from the transmission straight through to the rear driveline (propshaft and axle). The input shaft and rear output shaft are connected together with a sliding dog clutch (reduction hub). Most components that drive the front wheels in 4L or 4H are disconnected to maximize fuel economy.

In 4H, the front and rear drivelines (propshafts and axles) are mechanically coupled together through some type of dog clutch arrangement and rotate at the same speed. There is no gear reduction in the transfer case in 4H. The torque at each propshaft is determined by the road or surface conditions, but the speed of front and rear propshafts remains equal. The input shaft and the rear output shaft are connected together with a reduction hub, whereas the rear output shaft and upper sprocket are connected together through a lock-up collar. Torque from the transmission drives the transfer case input shaft, which in turn drives the rear output shaft and upper sprocket. The front output shaft is driven by a chain that connects the upper and lower sprockets.

Four-low operation is similar to 4H in that the front and rear axles are locked together. However, the 4L mode also provides a gear reduction that multiplies the input torque before it is transmitted to the outputs. The additional torque is helpful in situations where maximum tractive effort is required, such as deep sand, deep mud, or when ascending steep grades. The additional gear reduction also increases the engine braking capability, allowing the driver to use the throttle instead of the brakes for ground speed control when descending steep grades.

The reduction hub is slid rearward connecting the rear output shaft to the planetary carrier. The rear output shaft and upper sprocket are connected together by the lock-up collar. Power from the transmission drives the transfer case input shaft and planetary sun gear (as they are splined together). Rotating the sun gear causes the planetary carrier to rotate in the same direction, but at a reduced speed as the planetary ring gear is fixed to the case. The rear output shaft and upper sprockets also rotate at this reduced speed because the output shaft is connected to the carrier by the reduction hub and the upper sprocket to the output shaft by the lock-up collar. The front output shaft is driven by a chain that connects the upper and lower sprockets.

In neutral, the power flow from the transmission to the transfer case outputs is interrupted. The vehicle can be towed in neutral without removing the propshafts. Care must be taken in the system design to ensure proper lubrication to the rotating transfer case components during towing, because the transfer case is equipped with a positive displacement lubrication system driven by the rotation of the mainshaft.

The reduction hub is positioned such that it is not connected to the input shaft or the planetary carrier. Power from the transmission is disconnected from the mainshaft. Rotational input from the wheels is isolated from the transmission.

2.1.1 Shifting from 2H to 4H while driving

Some part-time systems have a feature that allows the driver to shift from two-wheel to four-wheel drive while driving down the road at speed. This feature is commonly referred to as *shift-on-the-fly* (*SOF*). In SOF systems, a synchronizing device is employed to synchronize (match) the speeds of the front and rear output shafts, so the drive collar or dog clutch that connects the upper sprocket and mainshaft can be engaged smoothly. A variety of synchronization devices are used including mechanical cone-type synchronizers (similar to those used in manual transmissions) and electromagnetic clutches (such as those used in a number of BorgWarner's transfer cases). Additional details on synchronizers can be found in Synchronisers—Gear Change Process, Loads, Timing, Shift Effort, Thermal Loads, Materials and Tolerances.

Cone type is the most prevalent form of synchronizer used in transfer cases. Most manual transmission synchronizers are of the single cone design. However, in instances where there are particularly high synchronization torque demands, such as when shifting a transfer case from high to low range (high speed differential and transmission inertia) or mode shifting at high speeds, double or triple cone synchronizers—or unusually large-diameter single synchronizers—are required. Using intermediate rings increases the number of friction surfaces, resulting in increased frictional force, frictional torque, and heat dissipation.

Key components of a double-cone synchronizer are shown in Figure 3 and include the shift sleeve, synchronizer hub, struts, blocker rings, and gear cones.

Internal splines on the sleeve with angular-pointed ends mesh with the matching angular-pointed ends of the drive dogs on the drive gears at the completion of the shift. The sleeve has a circumferential groove that engages the shift fork, which then moves the sleeve axially. Notches in the internal teeth locate the pre-synchronization components (struts and springs).

The hub is attached to the output shaft (mainshaft) through internal splines with external splines that allow the sleeve to slide on it. Three notches on the circumference at each end of the hub engage corresponding lugs on the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

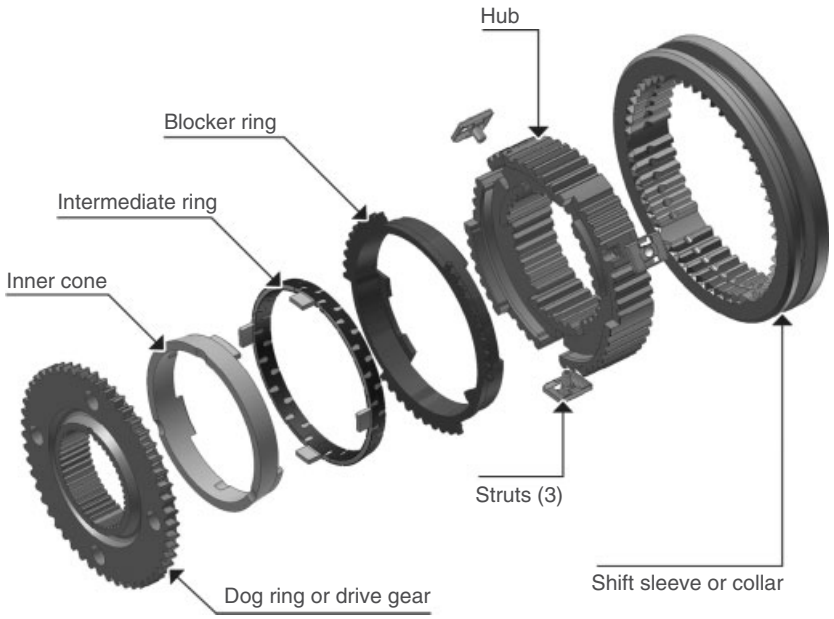


Figure 3. Components of a double-cone synchronizer. (Reproduced by permission of BorgWarner, Inc.)

blocker rings to prevent them from rotating. The hub has three axial slots for the pre-synchronization components.

The blocker rings were traditionally made from a special brass or bronze alloy. Modern production methods use organic or inorganic friction materials that are bonded to a steel or powdered metal core. The blocker rings have a cone-shaped surface on the inside diameter and "roof-shaped" teeth and lugs, which engage slots in the hub, on the outside diameter. The internal cone surface interfaces to an external cone surface on the drive gear. The blocker ring's cone surfaces have thread or groove patterns and axial grooves to allow faster displacement of the lubricant from the cone interface. The faster the lubricant can be displaced, the earlier the frictional torque increases, thus reducing the slip phase and resultant heat build-up.

Three radially movable struts are located on the circumference of the hub and spring loaded against three adjacent detent notches in the inside diameter of the shift sleeve. The struts can be ball bearings or roller bearings; or can be formed from sheet metal, powdered metal, or other processes.

The final element is the synchronizer cone, which is either machined onto the drive gear or is a separate steel piece welded onto the gear. The cone has "roof-shaped" clutching teeth machined at the outside diameter of the cone that mesh with the internal teeth in the sleeve when the shift is completed. In manual transmissions, the drive gear is supported on the mainshaft with roller or needle bearings. In transfer cases, the drive gear is integral with the upper sprocket.

During a shift, the shift fork slides the sleeve axially into position with the blocker ring. Owing to the chamfered sides on the detent notches inside the collar, the struts are compressed toward the center axis of the shift hub. The struts press the blocker ring axially against the cone on the drive gear. This action produces a frictional torque that causes the blocker ring to rotate slightly (limited by the hub groove width). The chamfered teeth on the sleeve contact the blocker ring teeth preventing premature axial movement of the sleeve. As the axial force from the shift fork increases, the resulting frictional torque brings the speed of the hub and drive gear (or sprocket) together, thus synchronizing them. When the speeds are equal, the frictional torque is eliminated. Because the shift force continues to act on the blocker ring teeth, the sleeve rotates the blocker ring and drive gear, allowing the teeth on the sleeve to slip into the blocker ring gaps and then mesh with the clutching teeth on the drive gear completing the shift. Operation of multiple-cone synchronizers is similar to the single cone design, except the intermediate cups and cones float between the main cones.

2.1.2 BorgWarner electromagnetic synchronizer

BorgWarner uses electromagnetic synchronizers in part-time transfer cases to facilitate the 2H to 4H shift while driving at road speeds. An electromagnetic synchronizer

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

consists of a stationary electromagnetic coil with a rotor or coil housing rotating around it and an armature placed in front of the rotor as shown in Figure 4.

The electromagnetic coil is an annular steel housing filled with insulated copper wire wrapped around the center axis. When voltage is passed through the wire, the resulting current generates an attracting magnet field.

The rotor is the component that rotates around the electromagnetic coil in close proximity to the outside diameter, inside diameter, and face of the coil. When the coil is energized, the rotor becomes part of the magnetic circuit. This part is usually attached to the mainshaft (rear output shaft). The armature is a flat plate usually attached to the upper sprocket via a collar.

When front axle synchronization is desired, the electronic control unit (ECU) energizes the magnetic coil with full available vehicle voltage (12 to 16 V). The shift system moves the armature, either touching it to or placing it in close axial proximity to the rotor. A magnetic flux circuit is created when the armature and rotor are touching each other. This is accomplished when the magnetic field generated by the coil flows through the rotor into the armature and back to the coil. This magnetic field causes an attractive force between the armature and the rotor.

The attractive force and the differential speed between the rotor and the armature causes a frictional torque to be imparted to the armature and thus to the upper sprocket. The upper sprocket translates the rotational force to the chain, then to the lower sprocket, and then to the front output shaft of the transfer case. The front propshaft and front axle assembly are attached to the front output shaft of the transfer case and are brought up to the same rotational speed as the rotor, thus the front driveline is rotating at the same speed as the rear driveline. The coil is shut off based on a prescribed time delay for synchronization or by input of front and rear driveline speed sensors.

2.2 Full-Time AWD

A typical full-time AWD system is always engaged and delivers power to all four wheels at all times. The power delivery is facilitated by an inter-axle or center differential in the transfer case. Full-time systems have a fixed torque split between the front and the rear axles. For example, the front to rear torque split for the BorgWarner 44–84 transfer case used in the Hummer H2 is 40 : 60, which means 40% of the input torque is distributed to the front axle and 60% to the rear axle.

The epicyclic (planetary) type is used as a center differential in the majority of full-time AWD applications, although the bevel gear type is still used in a few applications. Figure 5 depicts a typical full-time single speed transfer case (BorgWarner 44–79, used in the Cadillac STS) with an open-center differential. This center differential is an example of a transfer case using no ring gear and two sun gears. Power from the transmission flows to the planetary differential carrier. Power is then split between the front and the rear sun gears with the power going through one differential sun gear to the rear output shaft and the other sun gear to the front output drive sprocket, chain, then driven sprocket and front output shaft.

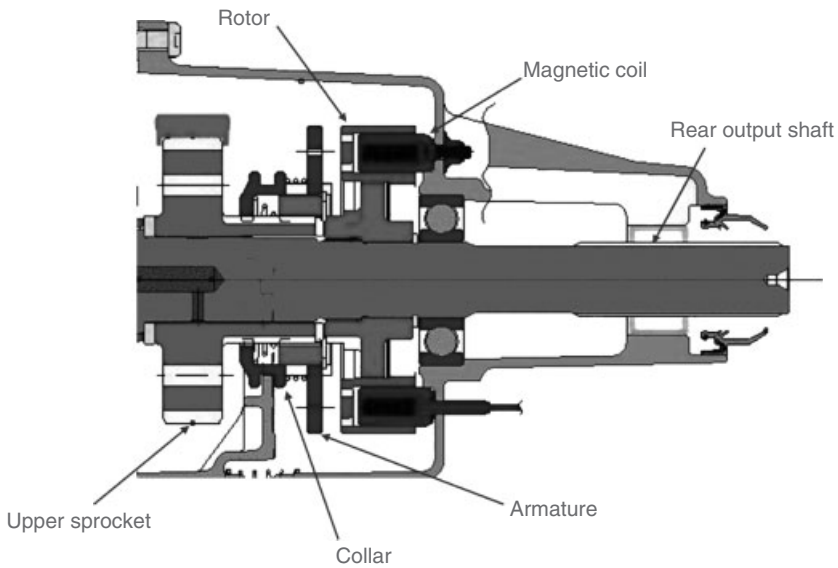


Figure 4. BorgWarner electromagnetic mode synchronizer components. (Reproduced by permission of BorgWarner, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

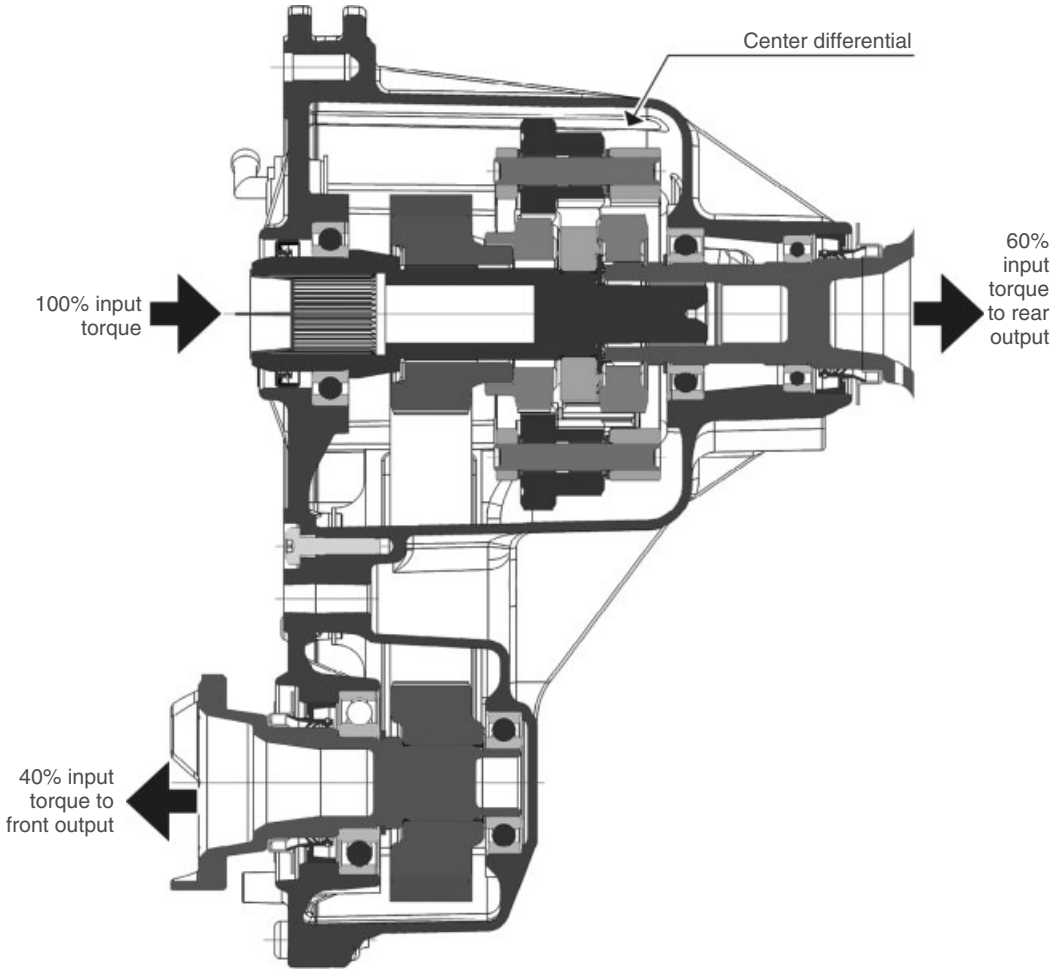


Figure 5. BorgWarner model 44–79 transfer case with open center differential. (Reproduced by permission of BorgWarner, Inc.)

For maximum traction, torque through the center differential must be controlled to ultimately distribute power to the tires with the greatest traction. Torque control, also known as *modulation*, can be done passively or actively with a torque biasing device or with brake-based traction control (TC) systems.

2.3 On-Demand AWD

An on-demand transfer case is directly connected to the rear (primary) axle and only distributes torque to the front (secondary) axle as required. Of the various AWD systems, on-demand systems are the most compatible with anti-lock braking systems (ABSs) and TC systems. In an on-demand AWD system, torque to the secondary axle is modulated by a torque management device; typically a wet clutch module within the transfer case.

Torque can also be directed to the secondary axle to improve vehicle handling on slippery or dry pavement. As the torque management device engages and disengages very quickly, a vehicle equipped with an on-demand transfer case can be driven on dry pavement without encountering the torque wind-up condition associated with a part-time system in the 4H or 4L mode.

Figure 6 shows a typical on-demand transfer case (BorgWarner 44–12, used in the Ford Explorer) in the on-demand (automatic) mode. When the automatic mode is selected, the electromagnetic clutch is engaged to synchronize the front and rear drivelines. The clutch is brought to a minimum current level that stages the components for any subsequent clutch increase and delivers approximately 35–70 Nm of torque to the front driveline. This torque level is too low to cause any torque wind-up problems during tight cornering maneuvers. The control system then monitors the speed difference between the front and the rear output shafts and increases the torque to the front output shaft if rear wheel slippage (speed difference between the front and the rear

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

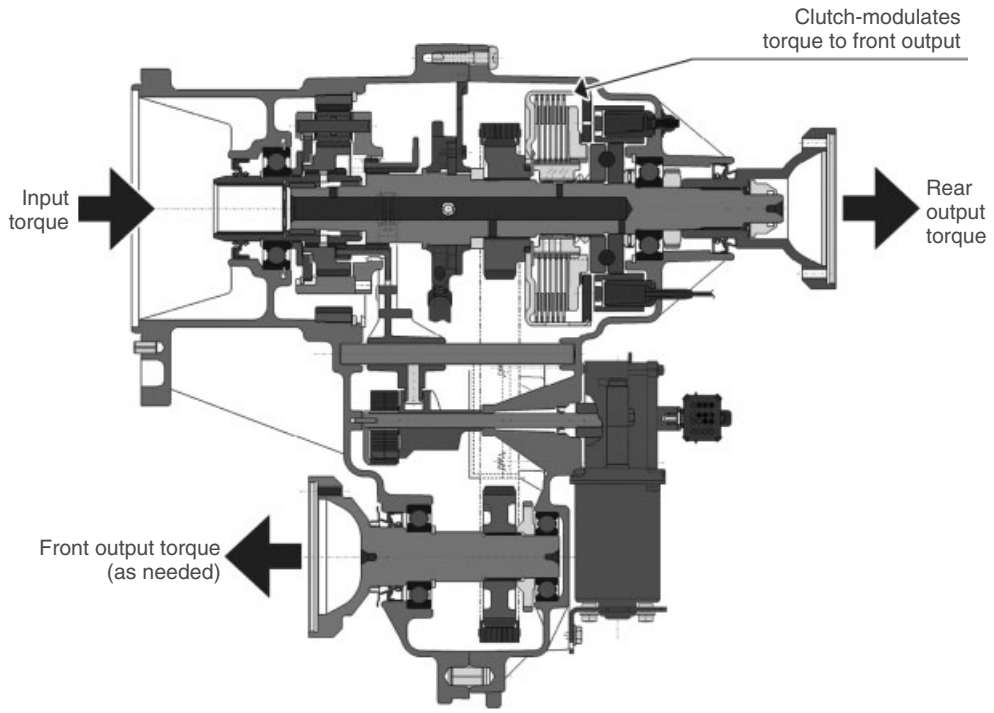


Figure 6. BorgWarner model 44–12 on-demand AWD transfer case. (Reproduced by permission of BorgWarner, Inc.)

output shafts) is sensed. In addition to output speed difference, the level of torque delivered can be determined from other vehicle parameters such as speed, throttle position, and accelerometer outputs. Once wheel slip is eliminated, torque to the front driveline is reduced to the staging level.

The BorgWarner 44–12 transfer case also has 4L and 4H operating modes. In these modes, maximum current is applied to the clutch resulting in a full lock condition similar to a part-time system. Power flow in 4L and 4H is the same as in a part-time transfer case. A variety of different torque management devices is used in on-demand transfer cases ranging from passive types to electronically controlled types.

3 POWER TRANSFER UNITS FOR FWD-BASED AWD

The term PTU is used to describe the automotive gearbox that is added to an FWD vehicle in order to facilitate the AWD architecture. The main difference between a PTU and a transfer case is that the PTU contains a right-angle gear system. This right-angle gear system is necessary because a PTU is mounted to a transverse (east–west) oriented transaxle, whereas a transfer case is mounted to a longitudinal (north–south) transmission.

Similarly to the transfer case, the PTU transfers power from the primary axle or wheels to the secondary (or auxiliary) axle. PTUs are most common on FWD-based AWD systems that typically incorporate the transverse (or east–west) powertrain layout. The PTU in FWD-based AWD vehicles is the focus of this section. There are, however, rear engine AWD vehicles that incorporate a PTU to deliver power to the front (auxiliary) axle.

There are two main types of AWD systems that utilize PTUs, which are full-time and on-demand. The main focus of this section is the PTU required for an on-demand system. We cover full-time PTUs later in the section. The primary shaft of the PTU is usually located on the output axis of the transaxle and is splined to the transaxle differential housing.

Figure 7 shows the typical packaging confines of a PTU. The halfshaft interconnecting linkshaft must pass through the primary PTU shaft. As a result, the primary shaft is hollow. This is one of the main sizing constraints of the PTU. The linkshaft must have a sufficient diameter to meet the torque requirements and duty cycle to drive the front wheel. The other typical sizing constraint for the PTU is the engine block and oil pan packaging envelope. With the size of the pass-through hole for the linkshaft and the engine and oil pan clearance defined, the inside and outside boundaries of the PTU are defined in the critical area of the primary shaft.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

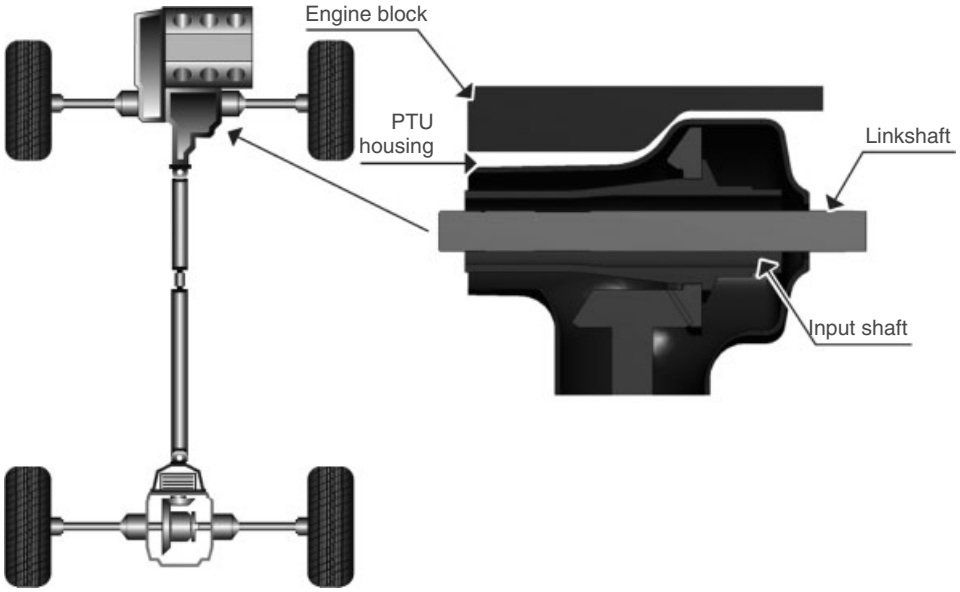


Figure 7. PTU integration to a transaxle. (Reproduced by permission of GKN Driveline.)

The two types of right-angle drive systems typically used are spiral bevel and hypoid gear systems.

The main difference between the spiral bevel and the hypoid gear systems is that the hypoid gear system has an offset. This is the difference between the centerline of the pinion and the centerline of the ring gear. Additional details on hypoid gears can be found in the chapter Basic Open Differentials.

The offset in the hypoid gear increases the contact ratio of the gear system, thus reducing NVH and providing more overall torque carrying capacity. The offset also creates more sliding at the ring/pinion interface. This sliding generates heat, pressure, and, without the correct lubrication, extreme wear. This resulting wear is the reason for extreme pressure additives in gear oil. This sliding also allows for lapping during the manufacturing process. Lapping refines the gear interface by rotating the gear members together under load with an abrasive fluid in the interface until an optimum contact pattern is established. The pinion offset is typically around 15–20% of ring gear outside diameter.

If there is not adequate space remaining to package a right-angle gearset in the PTU that meets the AWD system duty cycle, then one or more parallel axis gearsets are employed to facilitate locating the ring gear farther away from the confined area. These additional gearsets now change the single stage (hypoid set only) PTU to a two- or three-stage PTU.

In addition to creating packaging flexibility for the hypoid, adding the second or third stage of gearing may also reduce the torque requirement of the hypoid gearset. This is assuming that the parallel axis gear arrangement is a speed-increasing and torque-reducing arrangement. This gear arrangement may conversely be a speed-reducing and torque-increasing arrangement. Although not desirable, this arrangement may be required for some package environments. The consequence of adding multiple gear stages for improved packaging and torque optimization is that parasitic losses such as gear mesh losses, bearing and seal spin losses, oil churning, and increased mass will reduce the overall PTU efficiency and increase NVH.

The other main PTU packaging concerns are in the areas of the steering rack, exhaust, floor pan, and engine cradle cross members. All of these areas complicate the exact placement of the output shaft of the PTU. The other major issue with PTU placement is that the area behind the engine block intended for the PTU has minimal airflow. As a result of the gear sliding from the hypoids and other losses, heat is generated, which must be transferred to the lubricant and housing. In some extreme cases, liquid coolers have been added to help this situation. Most PTUs are not force cooled (primarily owing to cost and integration issues), so addressing the generated heat is yet another challenge during the design and development process.

Other noteworthy PTU design features that make it different from other gearboxes include the use of gear oil with extreme pressure modifiers similar to rear axle oil and a shaft-to-shaft seal between the hollow input shaft and the linkshaft to keep the transmission fluid separate from the PTU oil.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4 COUPLINGS FOR ON-DEMAND AWD

In this section, we cover couplings employed to manage the torque transfer between the front and the rear axles of AWD drivelines. On-demand AWD can be applied to either an FWD-based or an RWD-based AWD architecture. These are shown in Figure 8. In an FWD-based on-demand AWD driveline, the front axle is the primary driven axle and the function of the coupling is to transfer torque to the rear axle (secondary axle). In an RWD-based on-demand AWD driveline, the rear axle is the primary driven axle and the function of the coupling is to transfer torque to the front axle (secondary axle). The following coupling technologies can be applied in either of these architectures: dog clutch, passive, or active couplings.

4.1 Dog clutch for truck and SUVs

The simplest form of inter-axle control is the dog clutch arrangement used in part-time AWD systems. This type of clutch transmits power by engaging metal teeth or splines. It only allows direct mechanical engagement or disengagement without slippage or progressive torque transmission. Dog clutches are manually engaged using a lever or cable arrangement or electrically engaged by some type of electric motor/gear system.

Although the dog clutch is a very effective device for off-road traction, its use is fairly limited because of negative drivability issues when driving on-road. Because the system does not allow for speed differentiation between the front and the rear axles, driving on pavement can result in severe tire scrubbing and tight-corner binding. In addition, undesirable handling behaviors may occur during cornering. Dog clutch systems are only used on a part-time basis under extreme on- or off-road conditions, whereas on-road, these systems are typically disengaged. Dog clutch systems are typically offered on RWD-based pickup truck and SUV applications.

4.2 Passive and active on-demand AWD coupling control systems

Torque management technologies can be further categorized by the type of control system used (passive or active). Passive systems do not have an external control system to sense external variables. The operating characteristics of the passive systems are fixed and cannot be changed during operation to achieve a different set of operating characteristics. For example, the operating characteristics of a viscous coupling are determined by the number and diameter of the plates, viscosity of the fluid, percentage fill of fluid, and so on. The application of passive systems is sometimes limited because these systems can potentially interfere with the operation of other systems such as ABSs and electronic stability systems.

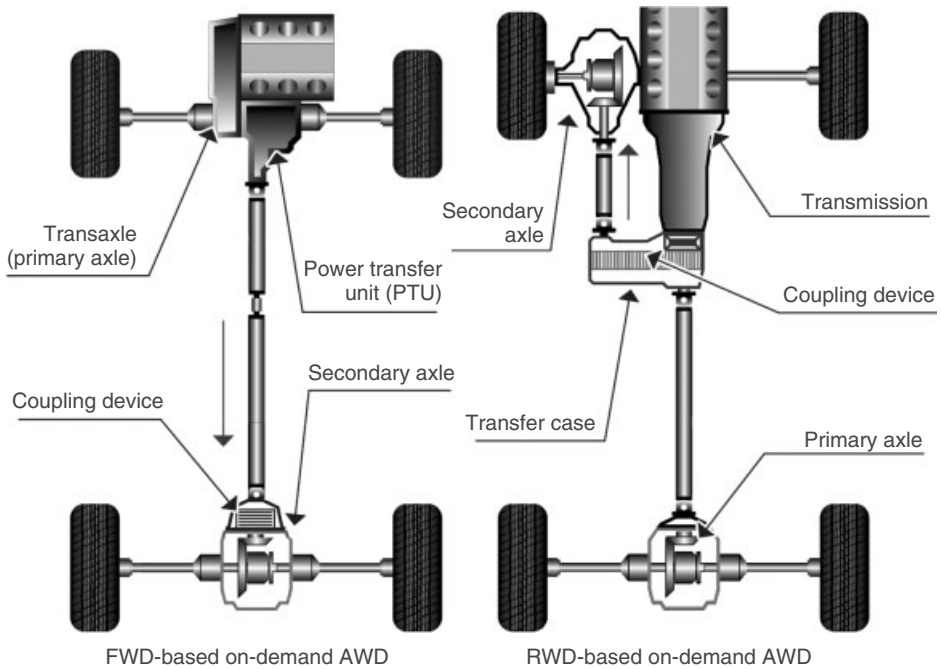


Figure 8. On-demand AWD drivelines. (Reproduced by permission of BorgWarner, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Active systems, on the other hand, use an external control system consisting of sensors to detect numerous variables (input speed, output speeds, and yaw velocity) and an ECU to process the signals that are sent to the actuation mechanism. Active systems are very versatile and can be integrated with other systems such as anti-lock braking, TC, stability control, and chassis control for improved vehicle-level performance. The system's capability is only limited by the number of sensors used and the control algorithm employed. Active systems can also be designed to be adaptive to compensate for hardware changes such as a degradation in the performance of the torque management device (clutch wear).

4.3 Speed-sensing passive couplers

On-demand passive couplers are self-actuated based on the speed difference across the device. They are also referred to as *speed-sensing*. Passive couplers generate a locking torque in response to the speed difference acting across the unit. There are typically three different methods to create a passive-type coupler:

1. Fluid friction
2. Hydraulic pump clutch actuation
3. Hydraulic resistance

4.3.1 Fluid friction type

The most common type of coupling using fluid friction is the viscous coupling. The coupling consists of a sealed housing with alternating inner and outer plates and is filled with a specific quantity of silicone fluid. One set of plates is connected to the housing of the coupling, which is then connected to the rear axle. Another set of plates is connected to the inner hub of the coupling, which is then connected to the front axle.

There are two operating modes in a viscous coupling, which are the viscous mode and the hump mode. The viscous mode is based on transmission of shearing forces in fluids. If the opposing surfaces of an inner plate and an outer plate move relative to one another in a fluid, a shear stress is produced in the fluid filling the gap. In a viscous coupling, this relative motion of the surfaces is achieved through a difference in the speed of rotation of the inner and the outer plates.

The silicone fluids used for viscous couplings are clear, nontoxic, and usually have a nominal viscosity between 5000 and 300,000 centistokes. The length of the molecule chains determines the flow properties—the longer the molecule is, the greater will be the viscosity of the fluid; the higher the nominal viscosity of the fluid is, the greater will be the torque that can be transmitted. The performance characteristics of the viscous couplings can be easily tuned to the vehicle in which it is to be used. The width of the gap between the inner and the outer plates also affects the torque transmission curve. The smaller the gap is, the greater will be the velocity gradient and the greater will be the transmitted torque at a given speed difference.

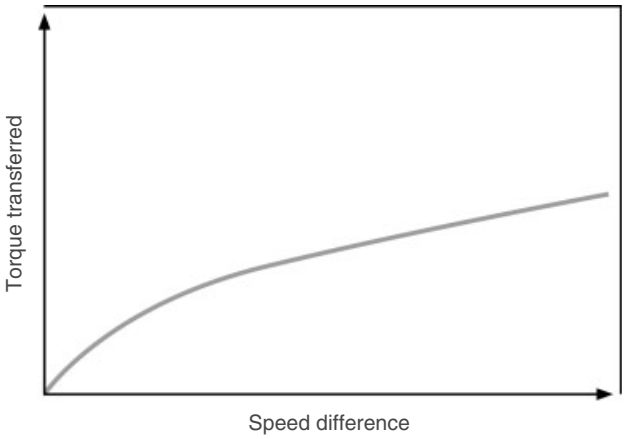


Figure 9. Generic viscous mode torque characteristic. (Reproduced by permission of BorgWarner, Inc.)

If a speed difference is present between the front and the rear axles, the outer housing rotates at a different speed than the inner shaft and the plates shear the viscous fluid. With increasing speed difference, there is an increase in torque transfer. The torque characteristic of this type of torque transfer device is said to be degressive. Figure 9 shows a typical speed-difference versus torque transferred. As the figure illustrates, the torque transferred at low speed differences is relatively high and increases quickly. As the speed difference increases, the rate of torque-transferred increase goes down. The physical properties of silicone fluid provide for significantly greater viscosity stable behavior across a wide range of operating temperatures as compared to using a mineral-based fluid.

The viscous coupling normally operates in the viscous mode, where torque transfer is accomplished by viscous shear (as described earlier). Under extreme driving conditions, the coupling will experience high speed difference and the fluid will begin to increase in temperature. The relatively large coefficient of thermal expansion of the silicone fluid causes the fluid inside the viscous coupling to expand considerably as the temperature increases. After several seconds of high speed difference, the fluid is expanded to such an extent that it fills all the available space inside the coupling, causing pressure to increase rapidly. This increase in pressure amplifies the throttling

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

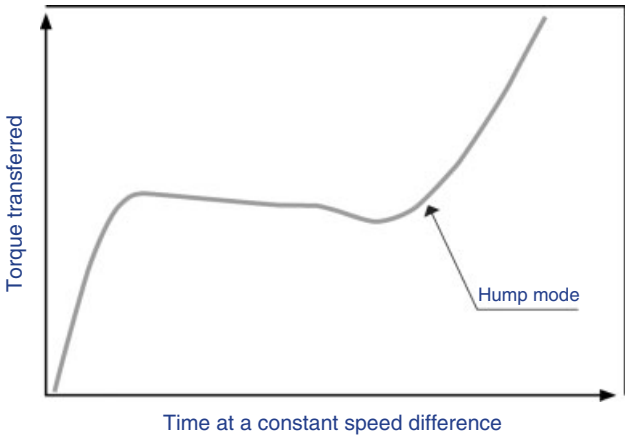


Figure 10. Generic hump mode torque characteristic. (Reproduced by permission of BorgWarner, Inc.)

effect between the plates, forcing the plates together and causing metal-to-metal friction to occur. The result is a substantial increase in torque transmission known as *self-induced torque amplification*, or *hump mode* (Figure 10), so named because of the hump-like shape of the torque curve. This torque amplification feature not only serves to protect the coupling against overheating but also provides additional traction in excess of axle skid torque.

The point where the hump mode is activated can be adjusted by varying the ratio of air and viscous fluid in the coupling. These adjustments allow the hump mode to be tuned so that it only engages under extreme driving conditions, or not at all (depending on the desired action).

Because this is a passive device, the final torque characteristic tuning selected must cover both winter and summer driving conditions as well as minimize tight corner binding. To ensure compatibility with certain ABSs, in the case of an FWD-based AWD vehicle, a free-wheel mechanism can be incorporated between the coupling and the rear axle to allow the rear wheel to overrun (allows the rear axle to be temporarily disconnected) during vehicle braking.

4.3.2 Hydraulic pump clutch pack actuation types

There are several different mechanizations using a hydraulic-type pump to develop hydraulic pressure in response to a speed difference. This resulting hydraulic pressure is then applied to engage a clutch pack. Two typical methods of generating hydraulic pressure are as follows:

1. Mechanical pump type
2. Rotary blade type

4.3.2.1 Mechanical pump type. A mechanical pump-type system can use a gerotor-, vane-, or piston-type pump, combined with an apply piston and a set of friction-based clutches. Typical applications use a gerotor-type pump. The housing can be connected to one of the axles (input) and the rotor can be connected to the other axle (output). When there is a speed difference between the axles, the pump begins turning and the fluid flow is initiated. This flow is then directed to an apply piston that will in turn press against a series of interleaved clutch plates. The greater the speed difference is, the greater will be the hydraulic pressure, and therefore the greater will be the amount of torque transferred.

Figure 11 shows a typical torque transfer versus speed difference curve for a gerotor-type pump. The curve can be shifted left or right in the design phase depending on the vehicle requirements. Examination of the graph in Figure 11 suggests that as the speed difference increases, the torque delivered is nearly proportional. The curve is actually progressive and the ease of engagement at lower speed differences helps to smooth out the engagement of the device. This correlation actually helps to protect the device from high speed partial engagement situations. Because the device relies on mineral-based oil for the hydraulic fluid, a temperature compensating bimetallic valve is typically used to adjust fluid flow based on the temperature-dependent viscosity properties of the fluid used. This valve helps ensure adequate performance over a wide range of fluid temperatures and viscosities.

4.3.2.2 Rotary blade type. Another type of self-actuating coupling that combines certain aspects of the viscous coupling and the pump/clutch-type coupling is the rotary blade coupling (RBC). This device consists of

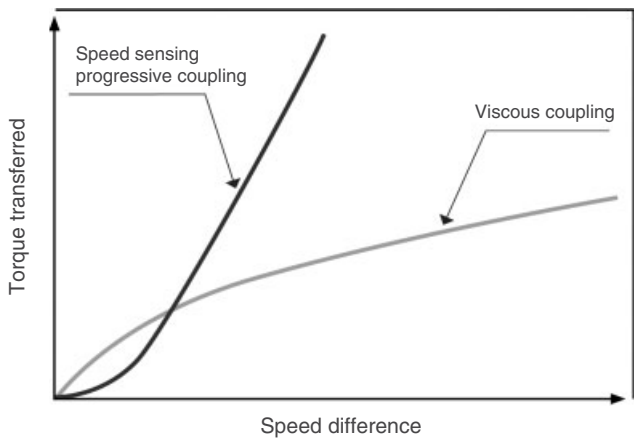


Figure 11. Generic gerotor-type passive progressive coupling torque characteristic. (Reproduced by permission of BorgWarner, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a sealed chamber filled with silicone fluid containing a tri-bladed disc, an apply piston, and a clutch pack. When there is a speed difference across the unit, the relative speed of the blades moving through the silicone fluid develops a pressure gradient that pushes on the apply piston, which in turn exerts axial force on the clutch pack.

With increasing speed difference, the apply pressure is generated and thus a higher torque transfer is achieved. The triblade design also offers the opportunity for a different geometry on the front and backside of the blade, which allows for an asymmetric torque characteristic. This feature is useful when tuning a vehicle's throttle-off handling behavior.

4.3.3 Hydraulic resistance type

Hydraulic resistance couplings typically use an axial piston pump in combination with a hydraulic valve body to generate a locking pressure proportional to the speed difference. Basically the device creates a hydraulic lock effect to resist the speed difference. The valve body typically has a temperature-compensating mechanism and the ability to reduce the locking torque characteristic based on the absolute speed of the device.

The input shaft of the coupling contains a cam profile, whereas the housing (output side) contains the individual piston-type pumps. The pistons ride along the cam profile generating the hydraulic pump effect.

4.4 Active couplers actuation methods

On-demand active couplers are externally controlled by an ECU and generally fall into three categories of actuation:

1. Electromagnetic
2. Electrohydraulic
3. Electromotor

As compared to the various passive coupler technologies covered in the previous section, active couplers offer significant vehicle-level benefits. Active systems are externally controlled with access to various sensors (input speed, output speeds, yaw velocity, etc.) and an ECU to process these signals, directly controlling the actuation mechanism. Active systems are very versatile and can be integrated with other systems such as anti-lock braking, TC, stability control, and chassis control for improved vehicle-level performance. The system's capability is only limited by the number of sensors used and the control algorithm employed.

4.4.1 Electromagnetic actuated coupling

The conventional design for an electromagnetic actuated torque coupling consists of an electromagnetic coil; an all-metallic primary clutch pack and a three-piece welded clutch housing assembly with a nonconductive metal ring for electromagnetic field management; an armature; and a secondary clutch pack. Energizing the coil creates a magnetic field flow around the coil through the clutch housing and through the primary clutch plates. The metal armature is attracted by the magnetic field, which in turn, compresses the steel plates of the primary clutch. This action generates the torque in the primary clutch, which is then amplified by the cam mechanism to engage the secondary clutch (Figure 12). With increasing current to the coil, there is a corresponding increase in torque transfer.

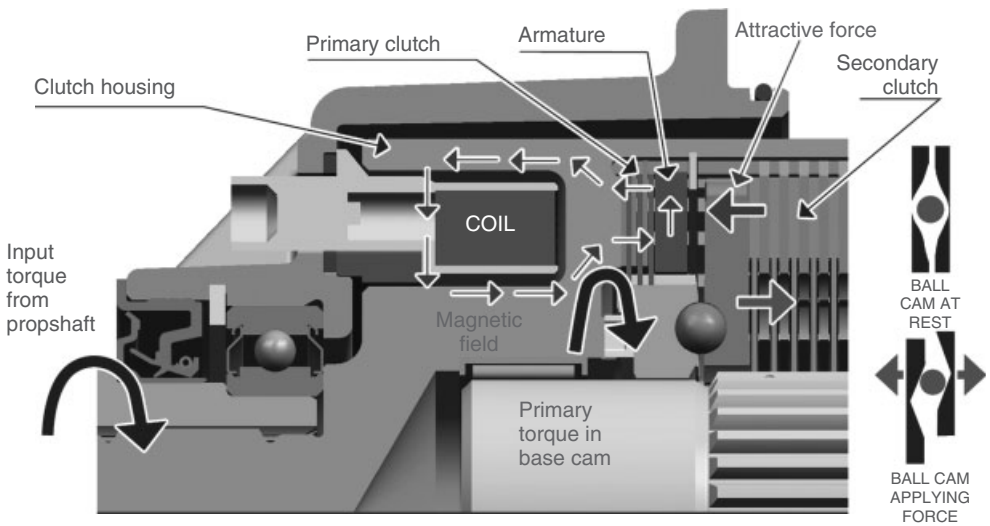


Figure 12. Conventional electromagnetic coupling function. (Reproduced by permission of BorgWarner, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The cam mechanism consists of a set of concentric balls reacting inside two opposing cams. The available torque in the primary clutch causes the input cam to rotate at a different speed than the output cam. The relative rotation of the cams allows the balls to move along the ramps of the cams to create axial force toward the base cam and the apply cam. The axial force that is put to the apply cam is directed toward the secondary clutch pack. This primary clutch torque represents about 10% of the output torque supplied by the system.

The secondary clutch pack consists of sets of friction and separator plates. The friction plates are splined to the output shaft. The separator plates are steel plates with specialized processing and finishing and are splined to the input shaft through the clutch housing. When the axial force from the ball-cam assembly (cam mechanism) is applied to this clutch pack, the compression of the clutch pack and the relative rotation of the input and the output components create the secondary clutch torque, which is about 90% of the output torque supplied by the system. The magnetic field path must travel through the primary clutch to attract the armature. This requires that the inner and the outer primary clutches be made of steel to allow the magnetic field to pass through.

In addition, a complex three-piece welded clutch housing assembly with a nonconductive metal ring is required for electromagnetic field management. Because the magnetic field must travel through the coil housing and across the metallic clutches, the field must bridge multiple air gaps. Because of these gaps, manufacturing and assembly processes must be highly controlled to minimize torque variations. This type of AWD coupling can be found in a wide range of applications.

Electromagnetic coupling technology can also be applied in such extreme vehicle applications as in the Porsche 911 Turbo (Type 997) in which the coupling transfers torque from the rear axle to the front axle.

Recently, there has been an evolutionary change to the primary clutch design and function. Instead of driving a magnetic circuit through the clutch, the primary clutch is now electromechanically actuated. An electromechanical force is generated by an electromagnet attracting the armature in a manner similar to a solenoid function. Three apply pins are connected to the armature and directly pull the primary apply plate which, in turn, compresses the primary clutch.

The compression action generates the torque in the primary clutch, which is then amplified by the cam mechanism to engage the secondary clutch. Because the electromagnetic field directly attracts the armature, there is only one air gap that the field must bridge. Because of this reduced gap, an all-metallic primary clutch design is no longer required nor is complex three-piece steel clutch housing necessary. Instead, an organic friction material-faced primary clutch and aluminum clutch housing can now be used.

In addition to the reduced complexity of this design, the use of an organic friction-faced primary clutch has an additional benefit. In the conventional design, the steel primary clutches tend to generate metallic debris over its lifetime: this wear can change the properties of the oil and deteriorate the frictional properties of the primary and secondary clutches. These changes can result in NVH issues. However, with the new actuation design, the source of metallic debris is eliminated because the primary clutches in this model are faced with organic-type friction material, thereby ensuring excellent NVH performance over the life of the application.

Electromagnetic couplings can also be applied to transfer case applications. The working principle is similar to that in the previously described conventional design. Energizing the coil creates a magnetic field flow around the coil, attracting the armature to the rotor. The armature is connected to one side of the cam mechanism, whereas the rotor is connected to the other side of the cam mechanism. The action of the armature generates the torque in the primary clutch, which is then amplified by the cam mechanism to engage the secondary clutch. This action is demonstrated back in Figure 12. With increasing current to the coil, there is a corresponding increase in torque transfer.

This type of RWD-based on-demand AWD transfer case application can be found in a wide range of vehicles from passenger cars to SUVs. This design also facilitates front axle disconnect systems because of its relatively low drag torque characteristic. This same technology can be used to provide both on-demand AWD to the rear axle of an FWD-based AWD vehicle and side-to-side torque transfer. Such a system is called a *twin on-demand layout*. The right and left wheels are individually coupled to the ring gear via the electromagnetic couplings.

Torque transfer to the rear wheels is achieved without a differential gearset. By actuating the couplings, torque can be transferred on demand to the rear wheels. This layout also provides significant side-to-side traction capability. In addition, if the couplings are controlled in conjunction with additional sensors (to detect the occurrence of lateral acceleration and yaw, for example), significant vehicle dynamics performance can be gained such as power-on torque vectoring in a turn.

4.4.2 Electrohydraulic actuated coupling

These active devices actively control hydraulic pressure to directly engage the clutch pack via an apply piston. The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

hydraulic pressure can be generated in the following three typical ways:

1. Mechanical pump in conjunction with an electric feeder pump
2. Electric pump in conjunction with a pressure accumulator
3. Direct hydraulic pressure control using an electric pump

4.4.2.1 Mechanical pump in conjunction with an electric feeder pump. A typical mechanization of this type of system consists of a piston-type pump, which generates a pressure when there is a speed difference across the unit. The hydraulic pressure from this mechanical pump is electronically controlled by a variable flow valve to direct the apply pressure to the clutch. There is also a mechanical overload valve to handle torque truncation for driveline torque overload protection and a supplemental electric pump for added system response. Figure 13 shows a schematic representation of how a typical piston-type pump functions.

The performance of the electrohydraulic actuated coupling device can be tuned based on the software algorithms that control the position of the valve. The ECU is integrated into the device and measures the fluid temperature directly so that the software control can compensate for changing viscosity of the fluid. Fluid temperature measurement is used to gain a better understanding of internal temperature of the device and contains a built-in thermal protection strategy. There are three separate pumping circuits for this type of system, which are the piston pump, the electric feeder pump, and the apply piston. Each of these circuits is arranged in such a manner that the total output of the circuit is a smooth, continuous flow rate.

4.4.2.2 Electric pump in conjunction with a pressure accumulator. This further advancement to this design is the so-called "power pack" approach, in which an electric motor drives a hydraulic pump, in conjunction with a pressure accumulator and a variable flow control solenoid valve. The oil pump is controlled to charge the pressure accumulator to a high supply pressure and the solenoid valve meters the necessary pressure on-demand to apply the clutch. Such systems can apply the clutch independent of speed difference. Figure 14 shows such a system applied as an FWD AWD coupling.

4.4.2.3 Direct hydraulic pressure control using an electric pump. This more recent advancement in electrohydraulic clutch actuation not only provides direct

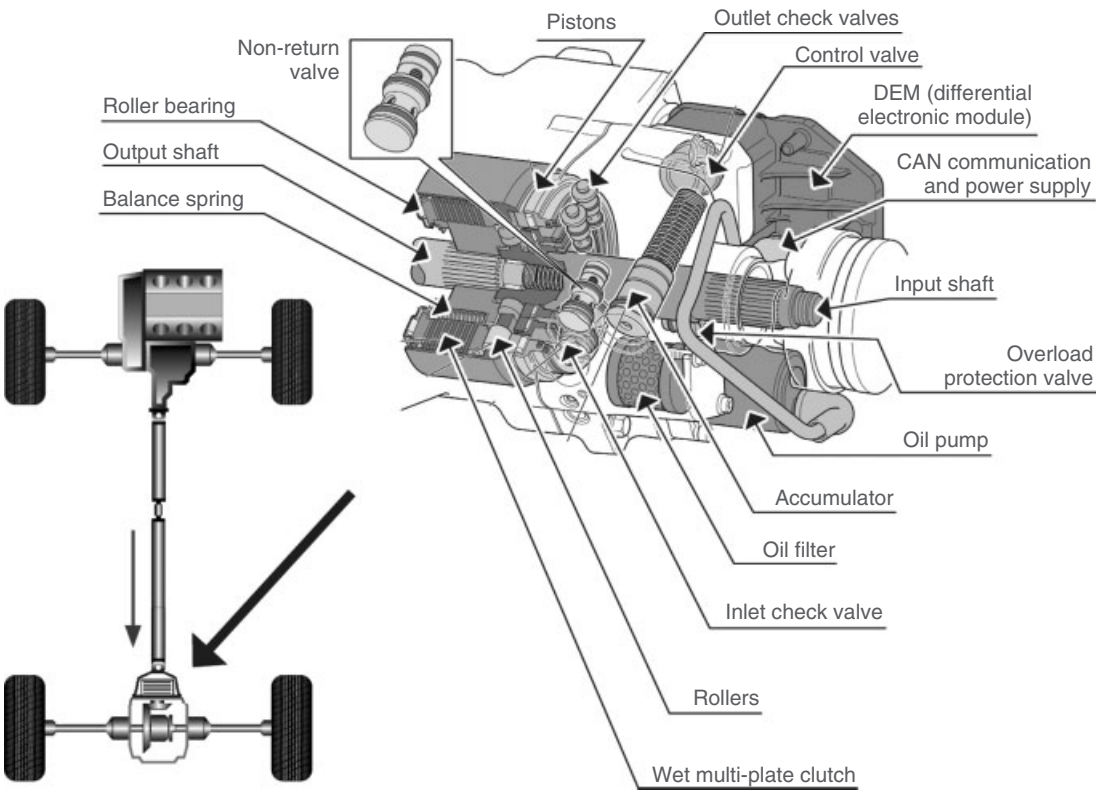


Figure 13. Electrohydraulic coupling application. (Reproduced by permission of BorgWarner, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

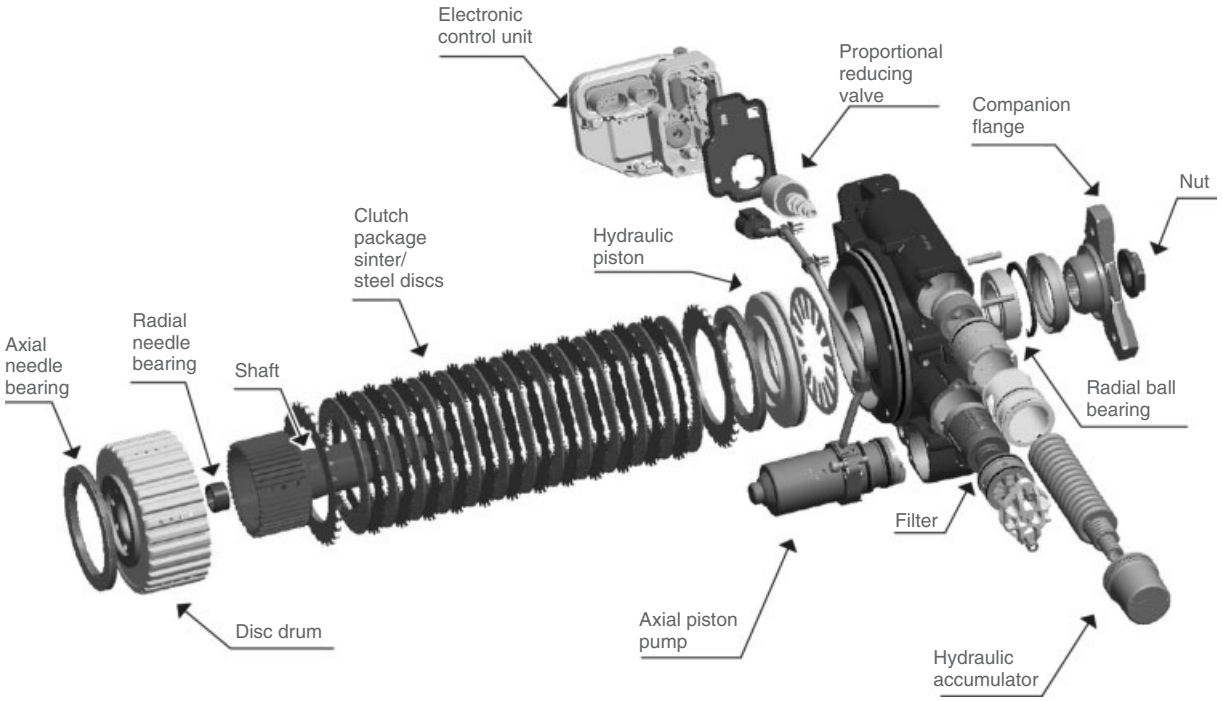


Figure 14. "Powerpack"-based electrohydraulic coupling. (Reproduced by permission of BorgWarner, Inc.)

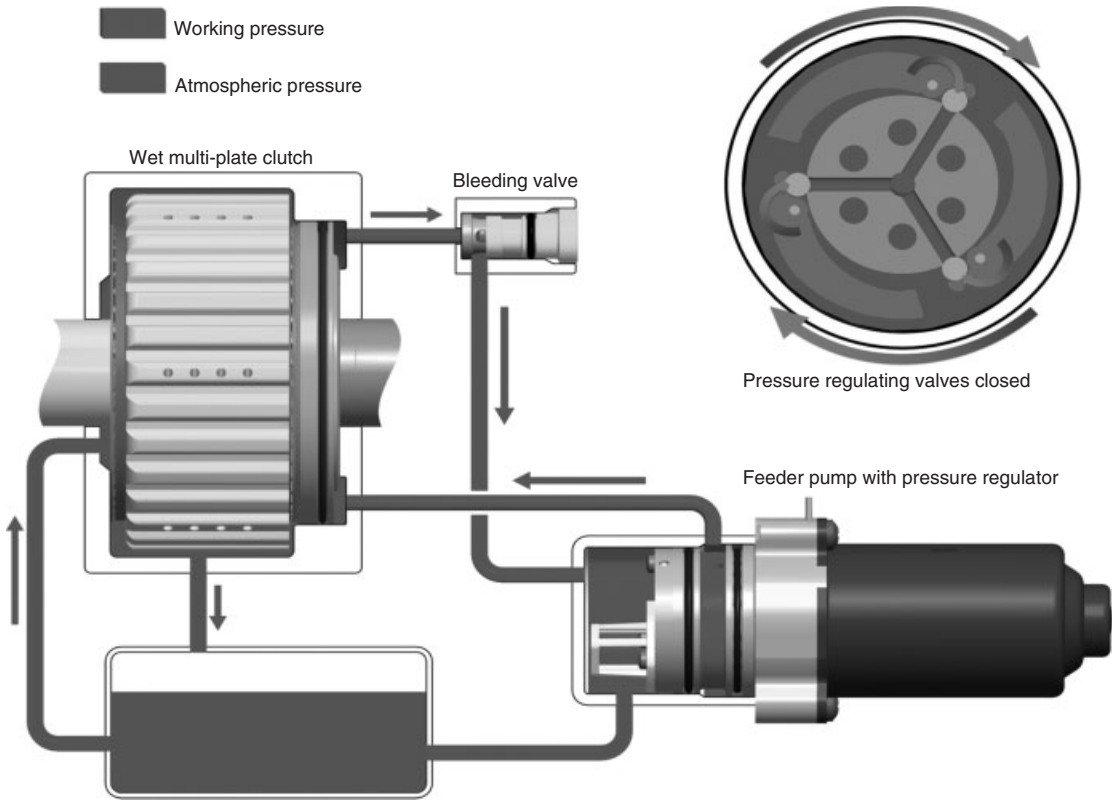


Figure 15. Direct hydraulic pressure control using an electric pump. (Reproduced by permission of BorgWarner, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

actuation of the clutch independent of speed difference, it does so without a pressure accumulator or pressure control solenoid or pressure sensor. An example of this type of system shown in Figure 15 uses an axial piston pump in combination with a centrifugal overflow valve mechanism, which allows accurate pressure control based on controlling the current to the pump motor. This type of system provides accurate and responsive performance with fewer components and with reduced complexity.

4.4.3 Electromotor actuated coupling

These active devices use a controlled electric motor to mechanically actuate the clutch. The most common approach is to use an electric motor driving a cam mechanism. Mechanical cam actuation consists of the electric motor driving a cam-type mechanism to translate rotational displacement into axial displacement. This can be achieved in several ways. An example is a system that uses an electric motor rotating a cam profiled output, which in turn causes the scissor-like mechanism to directly rotate a unidirectional ball cam mechanism, thereby inducing an axial force to compress the clutch pack.

5 CONCLUSION

FWD- and RWD-based AWD drivelines have unique architectures and also have some common torque transfer requirements. Actively controlled AWD torque transfer devices continue to grow in popularity as they enhance vehicle traction and handling performance while also contributing to higher system efficiency as compared to passive-type AWD systems. Actively controlled AWD torque transfer technologies will continue to evolve to deliver reduced mass and cost as well as increased response performance as part of fully integrated vehicle system together with brake-based TC and brake-based stability systems.

FURTHER READING

Society of Automotive Engineers (1989) 'Applications of viscous couplings for traction control in passenger cars'. Paper no. 890524. H. Taureg - Viscodrive GmbH, G. Herrmann - Viscodrive GmbH.

Society of Automotive Engineers (1990) 'Induced torque amplification in viscous couplings'. Paper no. 900557. Ing. H. Taureg - Viscodrive Gmbh, J. Horst - Viscodrive Gmbh.

Society of Automotive Engineers (1991) 'Development of a rotary tri-blade coupling for four-wheel drive cars'. Paper no. 910806. Satoshi Ashida - Toyota Motor Corp., Yukihiro Tanigawa - Toyota Motor Corp., Hiroaki Asano - Toyoda Machine Works, Ltd., Masaji Yamamoto - Toyoda Machine Works, Ltd., Yoshio Kojima - Toyota Central Res. & Develop. Labs., Inc., Kazunori Yoshida - Toyota Central Res. & Develop. Labs., Inc.

Society of Automotive Engineers (1995) 'The new venture gear NV249 on-demand transfer case for 1996 MY Grand Cherokee'. Paper no. 952646. Jim Brissenden - New Venture Gear, Sankar Mohan - New Venture Gear, Mark Dober - Chrysler Corp.

Society of Automotive Engineers (1995) 'A descriptive analysis of gerodisc type limited slip differentials and all wheel drive couplings'. Paper no. 952642. Murat N. Okcuoglu - Asha Corp.

Society of Automotive Engineers (1995) 'All-wheel and four-wheel-drive vehicle systems'. Paper no. 952600. Wesley M. Dick - Dana Corporation.

Society of Automotive Engineers (1999) 'New venture gear 261 transfer case - a robust manual shift transfer case'. Paper no. 1999-01-1261. Stephen M. Dolan - New Venture Gear Inc.

Society of Automotive Engineers (2004) 'P/M applications for heavy duty transfer case'. Paper no. 2004-01-0487. Marc Legault - BorgWarner Torque Transfer Systems, Jim Collins - BorgWarner Torque Transfer Systems.

Society of Automotive Engineers (2005) 'The NP244 transfer case chain noise reduction using a Gemini HyVo® chain system'. Paper no. 2005-01-2298. Steven Becker - Magna Drivetrain, New Process Gear Inc., Robert McAfee - Magna Drivetrain, New Process Gear Inc., Jeffrey Swanson - Magna Drivetrain, New Process Gear Inc.

Society of Automotive Engineers (2005) 'New technology for the management and distribution of torque in modern automotive drivetrains'. Paper no. 2005-01-0630. William R. Kelley - Borg Warner.

Society of Automotive Engineers (2006) 'Design of powder metal transfer case sprockets'. Paper no. 2006-01-0397. Marc Legault - BorgWarner Torque Transfer Systems.

Society of Automotive Engineers (2007) 'Development of an AWD coupling and controls for a high performance sports car'. Paper no. 2007-01-0661. Brian B. Ginther - Borg Warner TorqTransfer Systems, Chris Kowalsky - Borg Warner TorqTransfer Systems.

Society of Automotive Engineers (2007) 'Development of NexTrac™ electronic driveline coupling for front-wheel drive based all-wheel drive applications'. Paper no. 2007-01-0660. John Barlage - Borg-Warner TorqTransfer Systems, Joseph Mastie - BorgWarner Torq-Transfer Systems, Donn Niffenegger - BorgWarner TorqTransfer Systems.

Society of Automotive Engineers (2008) 'Approaches to achieving AWD torque accuracy'. Paper no. 2008-01-0303. David Haselton - BorgWarner TorqTransfer Systems.

Society of Automotive Engineers (2012) 'Development of a standard spin loss test procedure for 4WD transfer cases'. Paper no. 2012-01-0306. Michael Kirk - Chrysler Group LLC, Syed M. Ilyas - Chrysler Group LLC, Thomas D'Anna - FEV Inc, Andreas E. Perakes - Ford Motor Co., Craig S. Ross - General Motors Company.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto086
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Clutch Wet

Yuji Fujii, Nimrod Kapas, and Jau-Wen Tseng

Ford Motor Company, Dearborn, MI, USA

1 Introduction 1
2 Wet Clutch Structure 2
3 Operating Mechanisms 6
4 Clutch Design Process 10
5 Summary 14
Acknowledgments 14
Related Articles 14
References 14

1 INTRODUCTION

An oil-lubricated friction device has been utilized in automotive applications since the early twentieth century (Gott, 1991). It was first introduced as a means to simply couple or decouple rotating drivetrain components through a friction force (see Mechanics of Contacting Surfaces). A frictional interface of the wet clutch is lubricated with a transmission fluid (see Tribological Optimisation in the Powertrain) for cooling to protect the device from thermal damage. The fluid flow provides a significantly higher rate of heat rejection as compared to a dry friction device (see Dry Clutch). Largely speaking, there are two types of wet friction devices: wet plate clutch and wet band brake (Fujii *et al*., 2003). At present, the wet clutch is most commonly employed in a mass-production planetary-gear-based automatic transmission to alter torque paths for automated gear ratio changing (see Automatic Transmissions—Geartrain Combinations, Components, Design Considerations, Hydraulic System, Packaging, Manuf., Assembly). The wet clutch is also found in a hydrokinematic torque converter (see Automotive Torque Converters) to establish a mechanical connection, as desired, between an engine and an automatic transmission input shaft. Most recently, it is applied as a torque coupling device in a hybrid electric vehicle powertrain or HEV (see Micro, Mild and Full Hybrid, Power Split Configurations). It switches on and off a transmission of engine torque into a driveline on demand. The wet clutch has also evolved as a tool to control a torque transmission level from one component to another. Specifically, a clutch actuation force (see Clutch Actuation) is modulated to achieve a desired level of torque transmitted across the slipping frictional plates. For example, the wet clutch technology is utilized as a vehicle launch device in a lay shaft automatic transmission, which is often referred to as *dual-clutch transmission* or *DCT* (see Dual Clutch Transmissions (DCT)—Layouts, Clutch Selection, Packaging, Actuation, Manufacturing & Assembly). Its slip control is a key to a smooth vehicle launch without inducing undesirable driveline noise, vibration, and harshness (NVH) disturbance (see Drive Train Noise, Vibration and Harshness, Launch Control) in the absence of a torque converter. A limited slip differential system or a torque split device in a transfer case may also employ a wet clutch-based device (see Passive and Active Limited Slip Differentials).

Although the wet clutch technology has been around over 100 years, its design continues to evolve to meet ever-increasing challenges in fuel economy improvement, drivability enhancement, and shift quality refinement. For example, in a typical 6-speed planetary transmission system, two to three wet clutch packs are open at any time during vehicle operations. They collectively result in nonnegligible amount of viscous drag, measurably affecting overall driveline efficiency (see Tribological

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Optimisation in the Powertrain and Automatic Transmissions—Geartrain Combinations, Components, Design Considerations, Hydraulic System, Packaging, Manuf., Assembly). Automotive original equipment manufacturers (OEMs) and suppliers continue to look for clutch plate designs to lower viscous drag. During a typical shift event, one clutch is brought to an engagement, whereas another clutch is released in order to alter torque paths within gear sets. A mismatch in torque level and handshake timing between the on-coming and off-going clutches affects shift quality (Winchell and Route, 1961). It is important that the two clutches behave in a consistent and predictive manner in order to avoid an objectionable shift feel to a vehicle occupant. However, in practice, it remains a challenge to optimize a clutch deign to achieve low drag and consistent engagement and release controllability at the same time. New technologies, such as DCT, employ a wet clutch as a vehicle launch device to transmit torque from engine to driveline without the use of a hydrokinematic torque converter. The wet clutch slip control becomes very critical during vehicle launch to smoothly transmit a desired amount of drive torque from an engine side to a driveline without incurring an undesirable driveline NVH disturbance.

Regardless of its functional importance in drivetrain systems, wet clutch design analysis and optimization remain a challenge because of the complexity of physical processes in clutch operations. Open clutch analysis involves an interaction between multiphase fluid and rotating plates with grooves (Aphale, Schultz, and Ceccio, 2010; Takagi *et al*., 2011). Air bubbles are introduced at a high rotating speed, significantly affecting drag torque level. Engagement and release analysis requires a transient moving-boundary partial lubrication modeling (Cho *et al*., 2011). Fluid flow dynamically interacts with macro features such as plate grooves and microgeometry such as surface asperities. While the recent advancement in computational fluid dynamics methodologies may lead to a usable design support tool in the near future, clutch design improvement effort continues to heavily rely on a trial-and-error approach at time of writing.

This chapter presents an overview of the wet clutch technologies, including device construction, operating mechanisms, physical characteristics, and design practice. A typical wet clutch structure, components, construction materials, and key geometric features are described in Section 2. The wet clutch operating characteristics are explained in depth in Section 3. It includes the description of physical processes for open clutch, engagement and release behaviors in relation to fuel economy, automatic shift quality, and durability requirements. A clutch design practice based on the Systems Engineering is reviewed in Section 4. It also describes the recent trend in clutch design features, analytical tools, and test methodologies, followed by the concluding remarks in Section 5.

2 WET CLUTCH STRUCTURE

Wet clutches come in a variety of packaging configurations for different drivetrain applications. For example, a multiplate clutch for automatic shifting differs considerably from a single plate lockup clutch placed in a fluid-filled torque converter. Their design features also continue to evolve to meet ever-demanding controllability and durability requirements. However, the underlying operating principles remain the same to couple and decouple rotating components through wet friction. Accordingly, all the clutch designs include the common functional elements. This section describes the major components that are found in a typical wet clutch device. A clutch system for a planetary-gear-based automatic transmission is used for the illustration purpose as shown in Figure 1. Friction plates and separator plates are alternatively placed within the clutch housing. A set of friction and separator plates are called a *clutch pack*. The friction plates are mounted on a center hub, which is driven by a rotating drivetrain element. Separator plates are splined to the housing that is connected to another drivetrain element. The interface between the plates is lubricated with a transmission fluid. The plates are retained between the retainer ring on one side and the apply piston on the other end. When the clutch is open, friction plates and separators plate are allowed to rotate at different speeds while shearing fluid at the interface. When clutch engagement is commanded, an actuator applies a loading force onto the apply piston (see Clutch Actuation). A hydraulically driven actuator, depicted in Figure 1, is most commonly utilized for automatic transmission applications. The piston is stroked against a return spring while squeezing oil film at the frictional interface. The torque is initially transmitted through viscous shear between sliding plates. As the oil film is squeezed out, mechanical contact takes place between the friction and separator plates, partially transmitting torque through mechanical friction. As the engagement continues, the mechanical friction constitutes the entire engagement torque. The clutch engagement completes when the friction and separator plates are securely coupled. The construction and function of major components are described later for additional details.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

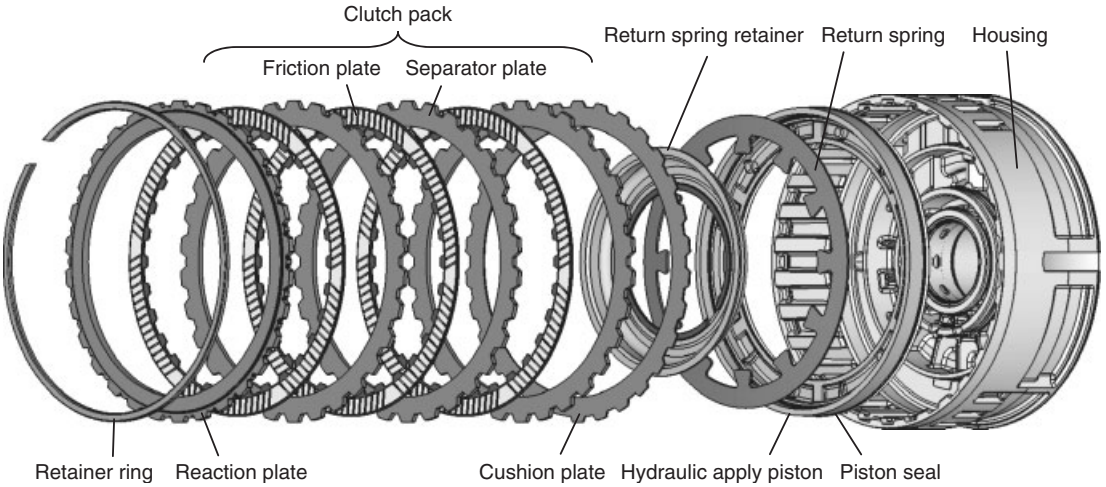


Figure 1. Major components of wet clutch system.

2.1 Major components

2.1.1 Friction plate

The primary function of the friction plates is to transmit torque from its spline-connected drivetrain element to the separator plates through wet friction (see Mechanics of Contacting Surfaces, Tribological Optimisation in the Powertrain). It is important that the friction plate provides stable friction characteristics under all operating conditions for smooth engagement, consistent release, and controlled slip. The friction plates must also possess sufficient static torque capacity when the clutch is functioning as a torque holding device. As illustrated in Figure 1, a typical friction plate, commonly referred to as the *double-sided plate*, is constructed with a friction material bonded to both sides of the steel core plate that has either internal or external spline. When the packaging space is limited, the friction material can be bonded on one side only, known as *single-sided friction plates*. In this design, the steel core plate also acts as a mating surface and heat sink for the adjacent friction plate. Thus, the steel core plate for the single-sided friction plate is significantly thicker than that for the doubled-sided plate. Shallow grooves are cut or pressed into the surface of the friction material lining in an application-specific pattern as a passive lubrication flow control device as later described in Sections 3 and 4. Alternatively, deep grooves can be formed between small friction material patches bonded to the core plate.

A friction material is made of organic fibers that are bonded together with resin as depicted in Figure 2. The structure is reinforced with synthetic fibers and certain chemical compounds for enhanced thermal and frictional

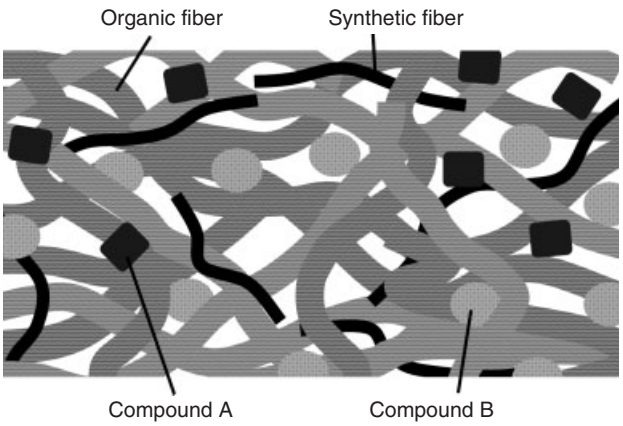


Figure 2. Friction material structure.

properties (Hirano *et al*., 2007). The compositions of fibers and the amount of pores within the structural matrix are considered important design parameters that affect oil flow within the material. In order to also adjust frictional characteristics, its surface layer may be embedded with friction modifier compounds. Both the structure and elements of friction materials are considered competitive information and not publically disclosed.

2.1.2 Separator plate

The function of separator plates is to transmit torque from its frictional interface to another gear element. The separator plate also acts as a heat sink to absorb heat generated during engagement and release. Thus, its thickness or thermal mass becomes an important design parameter to meet targeted slip conditions and duty cycles. The separator

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

plate is typically made from cold-rolled steel using either fine blanking or conventional stamping processes. A surface treatment for special micro-texturing may be employed to condition lubrication flow at the interface for fine-tuning frictional characteristics.

2.1.3 Reaction plate

A reaction plate, which is referred to as *pressure plate*, is the component located at end of the clutch pack as illustrated in Figure 1. It is placed directly against a retainer ring. It provides the reaction force against the clutch actuator, uniformly supporting friction and separator plates. Owing to the strength requirement, the pressure plate has a thick cross section and is typically made of powder metal. In some packaging deigns, a drivetrain component that is adjacent to the clutch pack may act as a pressure plate, providing the reaction force.

2.1.4 Retainer ring

A retainer ring is commonly used to retain clutch components such as a pressure plate and a return spring. It is made from spring steel and typically has a rectangular cross section. The retaining ring is formed by a coiling process. The ring has a circular shape in free state and is conformed into a noncircular shape in an installed position. This may present as an issue in applications that have small ring engagement depth. The ring may be pushed out under loaded conditions. For designs that require better conformance or higher retention force, a noncircular free-state shape retainer ring can be used.

2.1.5 Apply piston

In an automatic transmission system, a hydraulic actuator is most commonly utilized, although other types such as electromechanical or electromagnetic devices are also available (see Clutch Actuation). Its apply piston transforms hydraulic pressure into axial force exerted onto the clutch pack. Among the various types of piston designs, a stamped steel piston with a bonded seal and a machined aluminum piston with a loose seal are commonly employed in production applications. The reciprocating seal placed on the piston prevents the leakage of pressurized hydraulic fluid. Its design selection depends on seal drag requirements, sealing pressure, dynamic duty cycles, bore diameter, and seal groove tolerances. The most commonly used seals are O-rings, D-rings, and lip-type seals. A typical clutch piston is designed with a single hydraulic apply area. Alternatively, a clutch actuator may possess dual piston areas for enhanced torque controllability during engagement. Only the first piston area is initially pressurized, providing a low pressure-to-torque control gain for better slip management. Once the dynamic event is completed, the second area is pressurized to achieve high static torque capacity for securely holding the clutch pack. The dual-gain system can also be implemented through different approaches. For example, two hydraulic areas may act in opposing directions with differential cross sections. In this arrangement, the clutch has lower torque capacity when both areas are pressured while higher torque capacity is achieved by actuating a single area. In certain applications such as an automated engine start–stop system (see Micro, Mild and Full Hybrid), it is desired to lock the clutch piston in a pressurized position using an isolating valve without constantly delivering the pressurized fluid. However, maintaining a steady pressure level is difficult because of minute, yet continual fluid leakage across the piston seal. A hydraulic accumulator (see Automatic Transmissions—Geartrain Combinations, Components, Design Considerations, Hydraulic System, Packaging, Manuf., Assembly) may be utilized to hold the clutch piston in place, complementing the isolating valve.

2.1.6 Piston return spring

A return spring pushes the piston back to its released position when the hydraulic pressure is lifted. A coil spring pack and a disk-type spring are most commonly used in automatic transmission clutches. The coil spring pack consists of multiple coil springs that are held together by a steel spring retainer. The disk spring, which is illustrated in Figure 1, is stamped from spring steel and formed into a conical shape through heat setting and tempering processes. The load is obtained by compressing and flattening its conical geometry. While both types of spring provide the same function, they have distinct load characteristics. The coil spring pack produces a relatively linear load curve with up to 10% variability. The disk spring, on the other hand, has a nonlinear load behavior. It generally has the highest spring rate at the installed height. The nonlinear characteristics provide robustness against stroke distance change caused by friction material wear. Another benefit of the disk spring is packageability with its small axial length. However, the disk spring tends to have a higher load variation in the range of $\pm 20\%$.

2.1.7 Cushion plate

An optional cushion plate is inserted between the apply plate and the hydraulically operated clutch actuator, as depicted in Figure 1, to mitigate the initial engagement shock. It may take a shape of a waved plate or a coil spring.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

It is sometimes difficult to precisely control hydraulic pressure as the piston approaches its end of stroke. A sudden change in piston motion may result in an undesirable pressure overshoot under certain operating conditions. The cushion plate provides the cushioning effects, slowing down the piston motion for smoother transient pressure profiles. It is often found for wet clutches utilized for drive and reverse engagements in a planetary-based automatic transmission.

2.1.8 *Apply plate*

An apply plate (not shown in Figure 1) may be inserted between the clutch pack and the clutch actuator piston. The function of the apply plate is to uniformly distribute loading from the actuator onto the clutch pack. Accordingly, it generally has a larger thickness for higher structural integrity as compared to separator plates. Depending on design geometry and maximum load requirement, the apply plate can be made from powder metal, fine blanking, or conventional stamping.

2.1.9 *Clutch housing*

Clutch housing encases all the major clutch components. As illustrated in Figure 1, a hydraulic piston is directly placed at the end of the housing case, forming a hydraulic chamber. A hydraulic control system feeds a pressurized fluid into this chamber for clutch actuation. The housing has spline joints that directly engage separator plates and transmit torque to either a transmission case or a rotating gear member. Holes or slots are added to the housing to provide cooling lubrication flow into the frictional interfaces and to exhaust it to the outside of the housing. The clutch housing is typically made of a formed steel cylinder with a welded hub or constructed with machined aluminum. Steel is a common choice of material for many applications. Cylinders can be formed by different manufacturing processes, such as flow forming, transfer die forming, or a combination of stamping and Grob forming. The use of aluminum allows the inclusion of complex geometries, casted spline, and nonuniform cross sections. However, extensive machining may be required at higher cost to create hydraulic passage, snap ring grooves, and piston bores. The aluminum housing tends to have lower structural integrity because of its material strength and higher porosity.

2.1.10 *Balance dam*

Wet clutches are classified into two types based on their specific roles in drivetrain applications, namely the brake clutch and the rotating clutch. While they provide the same function to couple and decouple gear elements, the power flow arrangements are different. The brake clutch is utilized to ground a gear member to a nonrotating element. In case of a planetary gear set, this provides a means to alter torque path for automatic gear ratio changing. The rotating clutch, on the other hand, couples two gear members, rotating at different speeds. When the clutch is fully engaged, the gear elements become mechanically coupled and rotate in unison. The rotating clutch is widely employed in a planetary-gear-based automatic transmission and also in DCT as a launch device.

As illustrated in Figure 3, the brake clutch and the rotating clutch have a slightly different design feature. The housing of the rotating clutch rotates at all time, exerting centrifugal pressure onto the fluid trapped in the hydraulic chamber. A special design consideration must be made to avoid the so-called clutch drift-on where the centrifugal pressure overcomes return spring force, inadvertently stroking the apply piston for partial engagement. Two design approaches are available to mitigate this condition:

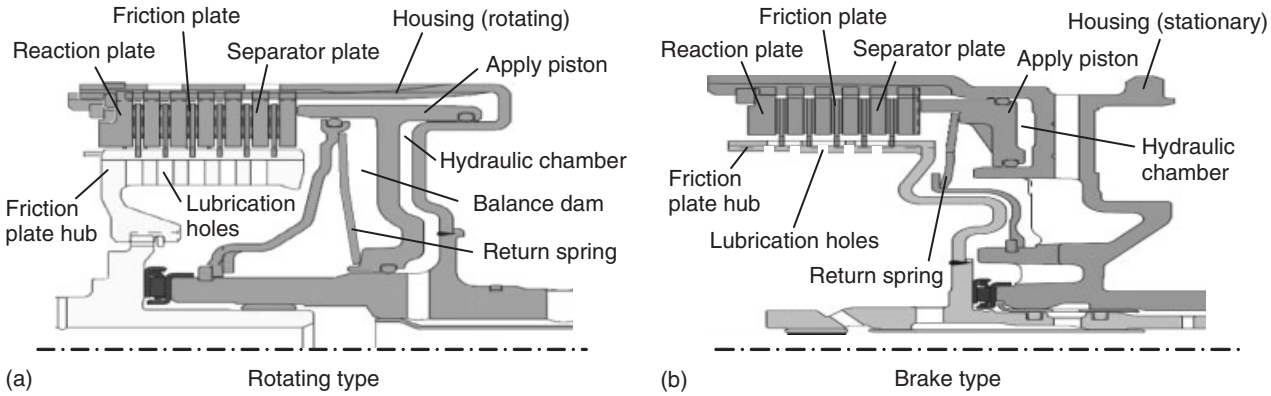


Figure 3. (a,b) Wet clutch design layout.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

one is to exhaust fluid in the hydraulic chamber by means of a ball check valve that is designed to open above a specific rotational speed. The second approach is to create a balance dam on the other side of the apply piston. A lubrication flow channel is designed to route transmission fluid into the balance dam. During the transmission operation, the centrifugal pressures on both sides of the apply piston cancel each other.

2.2 Lubrication design

The interface temperature rises significantly during repeated clutch engagements or extended slip control. In the construction of clutch assembly, a special attention must be paid to dissipate heat from the frictional interface to prevent the degradation of wet friction characteristics. That is, the temperature must not exceed a thermal limit of the friction material at any time. High temperature also accelerates the decomposition of transmission fluid additives, affecting wet frictional behaviors. A transmission lubrication system (see Tribological Optimisation in the Powertrain) is designed to deliver a targeted amount of cooling fluid into the clutch hub. The oil holes and slots are designed on the hub to uniformly distribute the lubrication flow across the clutch pack. The fluid flow rate must be at a sufficient level to provide adequate cooling effect. At the same time, it should not be excessive to avoid detrimental effects on open clutch drag and engagement behaviors. The drainage channels are designed into the clutch housing to efficiently exhaust fluid from the clutch pack.

3 OPERATING MECHANISMS

Clutch behaviors are highly sensitive to system configuration, design features, and operating conditions. Each application requires specific adjustments to achieve desired performance targets. This section describes open clutch behaviors, engagement and release mechanisms, followed by thermal characteristics. As previously mentioned, clutch drag reduction translates into fuel economy gain in an automatic transmission system. Robust clutch controls for superior drivability and shift quality rely on consistent and predictable clutch engagement and release processes. New drivetrain technologies such as DCT require demanding clutch duty cycles under challenging thermal conditions. The in-depth knowledge of wet friction mechanisms is a key to developing robust clutch systems. Clutch operating mechanisms and underlying physical processes are described later, based on multiplate clutch pack geometry for a planetary-gear-based automatic transmission.

3.1 Open clutch

When a multiplate clutch pack is open, friction and separator plates rotate independently at a relative slip speed, shearing the transmission fluid in-between. Each frictional interface is typically designed to have a clearance of 0.1–0.2 mm. The narrow clearance is critical to maintain engagement controllability, however, at the expense of open clutch efficiency. The clutch drag torque depends on many factors such as a number of friction interfaces, groove geometry, slip speed, fluid properties, flow rate, and temperature. Figure 4 illustrates typical drag characteristics as a function of slip speed for a clutch pack inside a warmed-up transmission system. Drag torque rises linearly in low slip range, peaks in the middle, and drops at high speeds. The physical explanation for this behavior is well established based on the fluid mechanics. The lubrication fluid flows into the friction interfaces from the inner hub as described in Section 2.2. The grooves on rotating friction plates pump and distribute fluid between the plates. At low speed, the friction interface is filled with lubrication fluid with fully developed laminar flow. As illustrated in Figure 4, the linear dependence of clutch drag on slip speed can be captured well by the Newton's law of viscosity (Kitabayashi, Li, and Hiraki, 2003):

$$T \approx rA\left(\eta \cdot \frac{v}{h}\right) \tag{1}$$

where T is drag torque, v is sliding speed, h is clearance, r is effective friction plate radius, η is dynamic viscosity, and A is friction surface area. At higher speed, the drag torque deviates from the linear behavior. This is because the centrifugal pumping effect exceeds fluid supply level, entraining air into the interface. The transition from single-phase to multiphase flow regime occurs around the peak torque location in Figure 4. Small air bubbles initially develop at the interface. The air volume grows larger at a

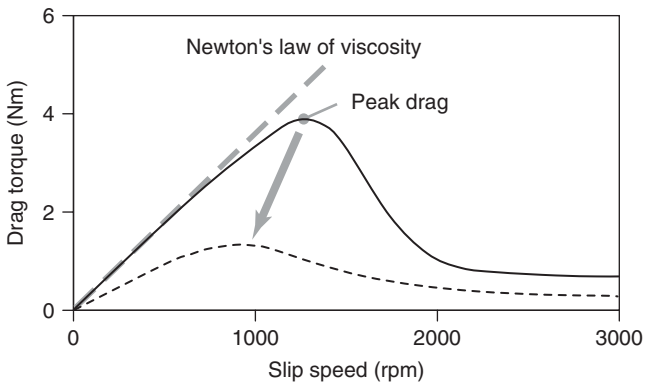


Figure 4. Clutch drag characteristics.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

higher slip speed, displacing the fluid and reducing viscous drag. One of the challenges in wet clutch design is to lower the overall drag curve, as depicted in the figure, without compromising engagement controllability. There are various design factors that can be adjusted to reduce the peak drag and shift its location (speed). Groove geometry, friction plate waviness, and fluid flow rate are among them. Although the principal mechanisms are understood for open clutch drag, it remains challenging to accurately model the effects of design variables because of complex flow geometry (Aphale, Schultz, and Ceccio, 2010; Takagi *et al*., 2011). Friction material and lubrication fluid also affect clutch drag characteristics. However, their changes are far more costly and require much longer lead time than geometry modifications.

3.2 Engagement

Clutch engagement is the process to couple friction plates and separator plates through wet friction. It begins with open clutch condition and ends when all the plates are securely grounded or coupled to rotate in unison. Clutch engagement torque, generated by wet friction, is directly transmitted to a vehicle drive shaft during automatic shifting. High sensitivity of clutch torque to operating conditions makes shift controls challenging. For example, Figure 5a illustrates the effects of temperature on transmission shift quality that can be actually observed in a vehicle (Fujii *et al*., 2003). The clutch actuator delivers nearly identical hydraulic pressure to the apply piston at hot and cold conditions. However, drive shaft torque profiles, which directly reflect clutch torque, differ significantly, creating inconsistent shift feel. Clutch engagement characteristics tend to be design-specific and vary widely, depending on applications. Thus, comprehensive knowledge on engagement mechanisms as well as underlying physical principles becomes a valuable asset to address unique challenges encountered during a clutch development process.

When a clutch engagement is commanded, an apply piston is hydraulically stroked to press all the friction and separator plates against the reaction plate. Lubrication fluid is squeezed out from the interfaces through complex flow geometry. During the initial phase, clutch torque is transmitted by wet friction that is primarily based on viscous shear between the rotating plates. As the fluid layer becomes thinner, a mechanical contact takes place between friction material asperities and separator plate, partially transferring torque through mechanical friction. Mechanical friction torque grows larger through the engagement process while viscous torque drops because of rising temperature and diminishing fluid volume at the frictional

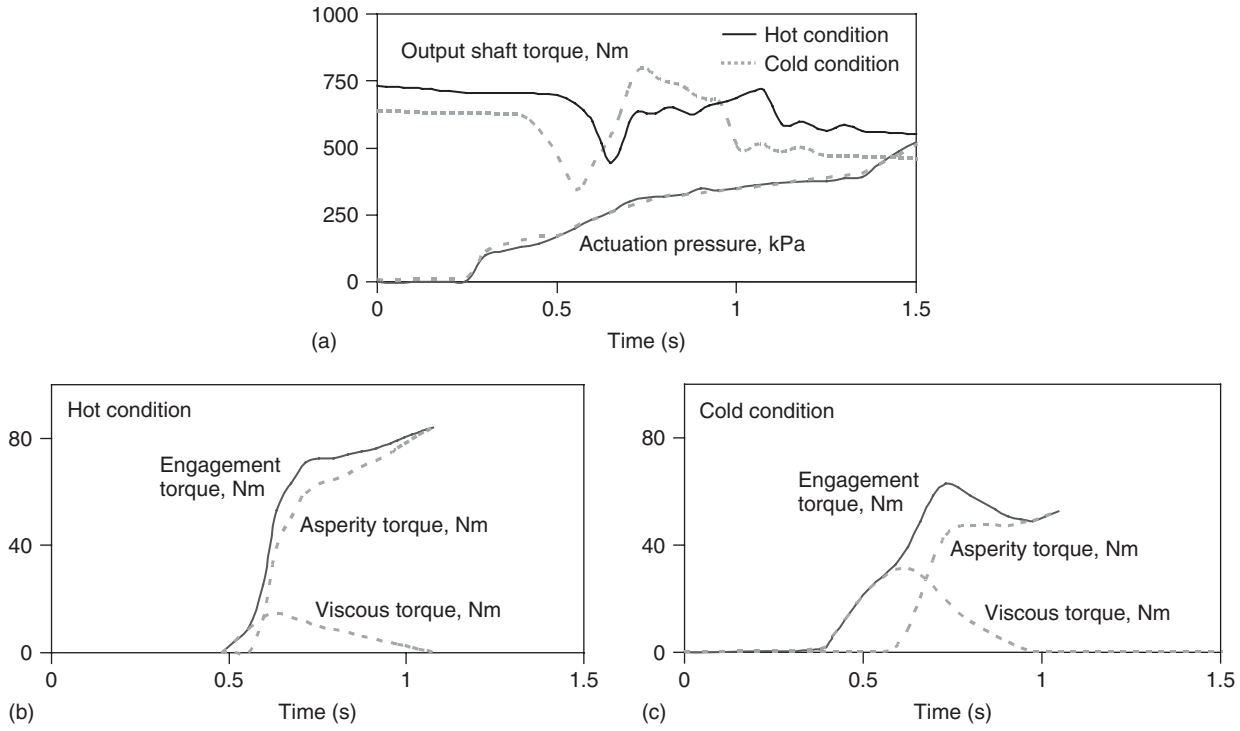


Figure 5. (a–c) Example of engagement behaviors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

interface. Mechanical friction eventually constitutes the entire wet friction, transferring torque through asperity contact. A real contact area of the asperities is surprisingly small and limited to a small percentage of a nominal friction surface area even under loaded conditions (Otani and Kimura, 1994).

Clutch engagement characteristics are largely affected by the squeeze film process that is highly sensitive to both geometry and fluid conditions. The classical fluid mechanics theory can be employed to describe the physical principles involved in the squeeze film process (Hamrock, 1994). As illustrated in Figure 6, a relative rotation between the friction and separator plates carries the Couette flow. Lateral and translational plate motion creates pressure gradient, driving the Poiseuille flow. The Couette flow and the Poiselle flow move fluid in and out of the groove. The groove acts as a centrifugal pump to push the fluid across its channel. The fluid is transported by pressure gradient into the permeable friction material that has a porous structure. This complex squeeze film process can be captured by the average Reynolds equation (Patir and Cheng, 1978, 1979) derived from the Navier–Stokes equation under the assumption of uniform pressure across film thickness (z-direction):

$$\frac{\partial}{\partial x}\left(\frac{\phi_x \cdot \rho \cdot h^3}{12 \cdot \eta} \cdot \frac{\partial p}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\phi_y \cdot \rho \cdot h^3}{12 \cdot \eta} \cdot \frac{\partial p}{\partial y}\right) = \frac{\partial}{\partial x}\left(\frac{\rho \cdot \overline{h_T} \cdot v}{2}\right) + \frac{\partial}{\partial x}\left(\frac{\rho \cdot \phi_s \cdot v}{2}\right) + \frac{\partial}{\partial t}(\rho \cdot \overline{h_T}) - u \tag{2}$$

where x and y are spatial coordinates; ϕ_x, ϕ_y, and ϕ_s are empirical flow factors that represent the effect of surface asperities on fluid flow; ρ is oil density; h is film thickness; p is fluid pressure; η is dynamic viscosity; $\overline{h}_t$ is average film thickness; v is linear sliding velocity, t is time; and u is exuding velocity of fluid into the porous friction material. The left-hand side of Equation 2 represents the Poiseuille flow driven by pressure gradient, and the first term on the right-hand side describes the Couette flow driven by sliding motion.

The average Reynolds equation forms a basis of comprehensive clutch engagement analysis. When coupled with heat transfer and asperity contact model (Greenwood and Williamson, 1966), its solution, in principle, provides fluid film thickness and asperity contact load for computing transient clutch engagement torque. However, a numerical method is yet to be established to solve the average Reynolds equation for complex three-dimensional clutch geometry. A practical application of the average Reynolds equation requires a considerable simplification in model geometry (Berger, Sadeghi, and Krousgrill, 1996; Fujii, Tobler, and Snyder, 2001). Nonetheless, it still provides valuable insight into wet friction characteristics as illustrated in Figure 5b and c (Fujii *et al*., 2002, 2003). Under hot fluid condition, clutch engagement torque largely consists of asperity contact torque. However, at cold condition, a combination of large viscous torque and delayed asperity contact results in a torque peak during the engagement process. This engagement behavior explains a large output shaft torque oscillation in Figure 5a observed in a vehicle at cold conditions. A recent trend in the high fidelity multi-physics clutch modeling is described in Section 4.2.

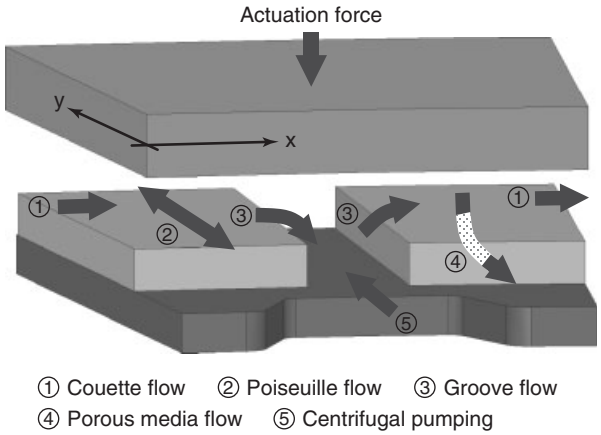


Figure 6. Fluid flow during engagement.

3.3 Release

The wet clutch release characteristics are particularly important for a clutch-to-clutch shift application (see AT Control—Actuation Methods & System Integration, Gear Choice, Gear Shift Strategy & Process, Adaptive Features). During the shift event, an off-going clutch must be released synchronously when another clutch reaches a certain torque level during its engagement process (Winchell and Route, 1961). A missed synchronization leads to inconsistent shift feel or unpleasant shift shock. However, a wet clutch often exhibits erratic breakaway in response to slowly dropping hydraulic control pressure. Figure 7 illustrates three distinct release behaviors often observed in an actual vehicle. Case 1 shows a nominal release behavior as the hydraulic pressure is gradually reduced. The clutch breaks away around 0.15 s when the holding torque capacity becomes smaller than the torque exerted onto the clutch pack. The torque transition is smooth before and after the breakaway. In Case 2, the clutch breaks away at the same timing. However, it goes through undesirable torque

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

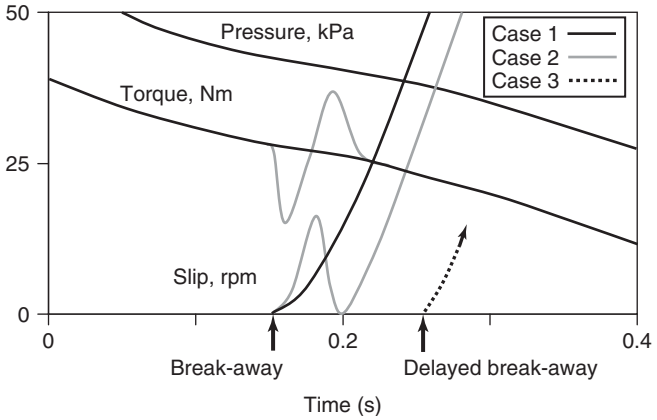


Figure 7. Example of release behaviors.

oscillation, momentarily re-engaging the clutch pack at around 0.2 s. In Case 3 where only the slip speed is shown in the figure, the clutch breakaway is significantly delayed, missing clutch-to-clutch synchronization timing, even through the hydraulic pressure follows the same profile as Case 1 and Case 2.

There are very few formal studies published in open literature for clutch release characteristics regardless of its importance. However, there is a significant amount of hands-on knowledge in industry obtained from product development experiences. This includes the physical steps involved in clutch release processes as illustrated in Figure 8. In Stage 1, the clutch is fully engaged. The static friction between the friction and separator plates provides a sufficient holding torque against the external load exerted on the clutch pack. In Stage 2, the external load overcomes the holding capacity as the hydraulic actuator pressure drops. The friction plate breaks away from the separator plate, whereas the transition from static to dynamic friction takes place. In Step 3, the sliding motion of the friction plate distributes the lubrication fluid from the groove into the frictional interface, establishing the partial lubrication condition (Hamrock, 1994). As the friction plate accelerates its motion, more fluid is entrained into the interface. The overall wet friction torque drops toward a drag level when the hydraulic actuator pressure is totally lifted.

The erratic release behavior depicted in Case 2 (Figure 7) can be explained by the release model in Figure 8. A certain plate design has a sizable fluid reservoir and drainage at the frictional interface. At the moment of breakaway, a large amount of fluids spreads into the interface, moving directly from Step 1 to Step 3. This jump creates a torque drop that is larger than static-to-dynamic transition. However, once the friction plate starts sliding, the fluid is quickly exhausted into the large drainage, reestablishing the asperity contact. The clutch torque momentarily rises, re-engaging the friction plate. The release behavior in Case 3 is often observed in actual transmission systems. It is thought to be caused by the variability of static friction under the presence of uncontrolled noise factors such as the interface temperature. The breakaway is delayed if the static friction becomes larger for a given level of hydraulic actuator pressure.

3.4 Thermal consideration

A wet clutch generates a large amount of heat during a coupling process. Thermal conditions must be adequately managed for robust slip controls, high durability, and application-specific duty cycles. The effects of temperature on clutch operating behaviors are very complex. However, the previous discussions on engagement and release mechanisms should prove useful to provide insight into specific thermal dependency for any applications. The following part focuses on the thermal considerations mainly related to durability and duty cycles.

Figure 9 illustrates transient thermal behaviors during consecutive clutch engagements for a planetary-gear-based automatic transmission. The friction material surface temperature rises during each engagement. The cooling effect is limited while it remains engaged. Once the clutch is released, the lubrication fluid flows into the clutch pack to cool down the frictional interfaces. However, the surface temperature does not drop to the pre-engagement level if a cooling duration is not sufficient. Accordingly, the peak temperature becomes higher during the subsequent engagements. A drive cycle thermal analysis is an important aspect of a clutch development process, especially for the applications that require demanding clutch duty cycles. It establishes cooling requirements and supports design optimization to adequately dissipate the thermal

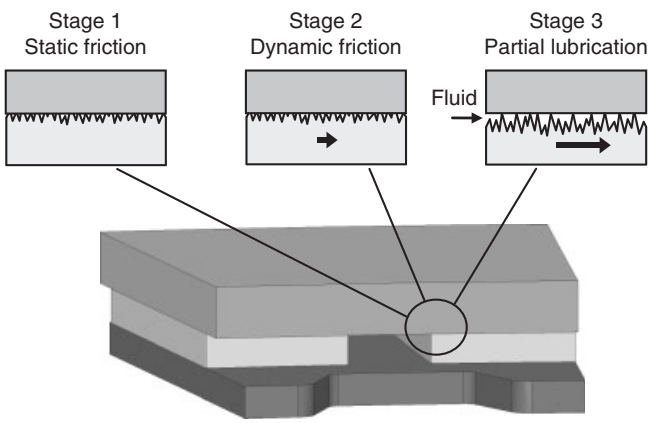


Figure 8. Clutch release mechanisms.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

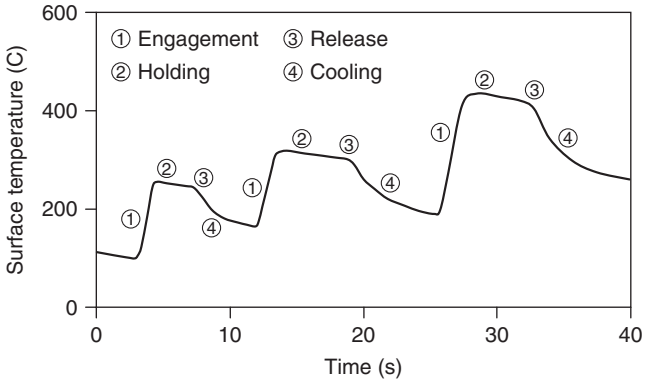


Figure 9. Example of clutch thermal cycles.

energy through convection, conduction, and radiation (see Tribological Optimisation in the Powertrain). A first-order clutch thermal cycle analysis is relatively straightforward. Assuming an average clutch torque T and a linear reduction of slip speed from ω_0 to zero over the duration of Δt, the total thermal energy for each engagement, E, can be readily estimated through the following equation:

$$E \approx T \cdot \omega_0 \cdot \left(\frac{\Delta t}{2}\right) \tag{3}$$

Equation 3 can be utilized to determine clutch cooling requirements, following specific engagement sequences in a target drive cycle. A typical cooling flow rate is 100–200 mL/min per each interface for a multiplate clutch pack in a planetary gear transmission. For demanding applications such as DCT launch clutch, the cooling requirement is substantially higher.

Generally speaking, a higher temperature accelerates the degradation of friction material and transmission fluid properties. It promotes a microstructural loss of fibers and resins, wearing off the friction lining. Fluid additives, such as a friction modifier agent, are decomposed at a higher rate, altering wet friction characteristics. In practice, a conventional step-ratio transmission system is designed to keep the bulk fluid temperature under 120°C range to protect additive properties. Its clutch duty cycle is monitored to prevent friction lining temperature from exceeding 300–400°C range. Excessive thermal conditions lead to serious failure modes that may result in a spontaneous loss of clutch functions. For example, a friction material surface can be glazed, temporarily or permanently losing its frictional capability (Yesnik, 2002). A friction lining may break off or delaminate from the core plate due to a severe structural damage. When the interface temperature exceeds the thermal limit of a separator plate, catastrophic thermoelastic instability may occur, resulting in a complete seizure of a clutch pack. More specifically, thermal deformation of a separator plate causes a local hot spot, which in turn promotes further plate deformation, resulting in a severely warped or coned separator plate (Audebert, Barber, and Zagrodzki, 1998; Zagrodzki and Farris, 1998).

4 CLUTCH DESIGN PROCESS

A challenge for higher fuel economy and superior drivability has transformed a once-mundane clutch design process into a sophisticated design optimization that requires the intimate knowledge of clutch operating principles. The new drivetrain systems demand a wet clutch with lower drag, enhanced torque controllability and higher durability. In order to balance overall clutch characteristics, a complex interaction between various design variables must be accounted, including component geometry, friction materials, and fluid properties. At present, a typical clutch design process follows a well-established practice from the Systems Engineering (INCOSE (International Council on Systems Engineering), 2010; US DOT (Dept. of Transportation), 2007) as a part of drivetrain system development. The Systems Engineering approach aims at reducing costly and lengthy engineering iterations in order to achieve drivetrain-level performance requirements. Among the clutch components, the friction plate assumes of particular importance because of its direct impact on both open clutch and transient torque characteristics. Figure 10 shows the so-called V-model from the Systems Engineering, highlighting its application to the friction plate design process as an example. In principle, all other components also follow the same developmental process in parallel. On the left side of V, the specifications and the requirements for a target driveline system are cascaded down to a clutch pack and to a friction plate level, including packaging constraints, performance requirements, operating conditions, and target costs. A thorough analysis of the requirements is a critical step toward finding a successful clutch design. The engineering iterations take place, as required, between the component, subsystem, and system levels on the left side of the system V process.

As one of the basic building blocks, friction plate design activities are defined at the bottom of the V-model. Geometric features are evaluated, accounting for primary interactions with other factors such as material characteristics, fluid properties, and operating conditions. The complexity of physical processes described in Section 3 makes the design optimization challenging, especially in the absence of predictive CAE (computer-aided engineering) tools for engagement and release analysis. Performance

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

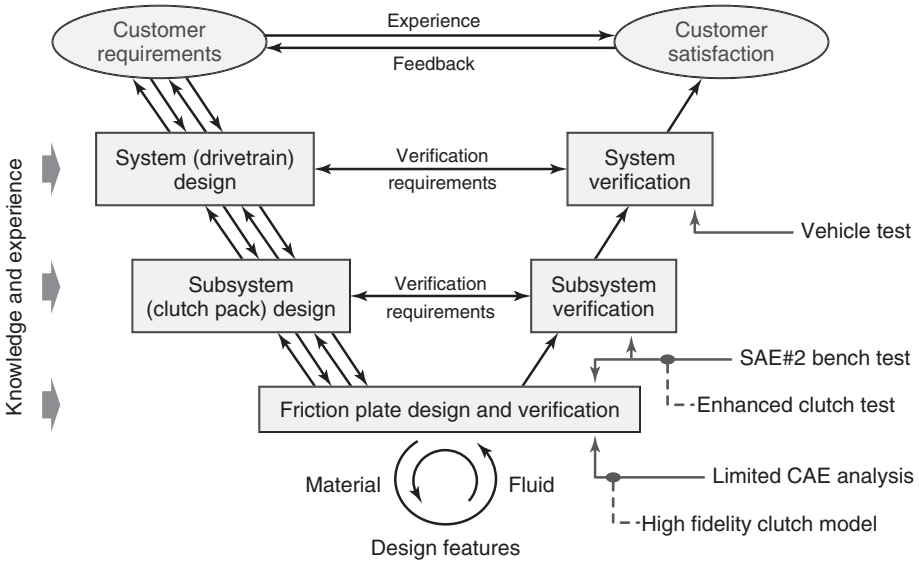


Figure 10. V-model for clutch design process.

evaluation tests are conducted on component benches, typically adapting the industry-standard SAE#2 test procedure (Fanella, 1994). The performance verification follows the right-hand side ladder of the V-model upward from the friction plate level. Thorough requirements analyses conducted on the left side of V form a basis for developing a successful performance verification plan.

The engineering iterations for performance verification on the right-hand side of the V-model are costly and time-consuming. For example, even if a clutch plate and a clutch pack pass verification tests as a component, a shift quality may still fail to meet the target metrics at a vehicle level. This can push the development work back to the subsystem or component levels after having invested significant resources for the prototypes and vehicle-level shift calibration work. A well-designed Systems Engineering approach is built on up-front planning, requirements analysis and robust engineering at all levels on the left side of V to reduce hardware changes and firefighting work downstream on the right side of V for a shorter development time.

The remainder of this section describes the recent trend in clutch performance optimization with a specific focus on friction plate design features. Then, two emerging engineering technologies, namely the high fidelity multi-physics clutch model and the enhanced clutch test method, are presented. They are expected to proliferate in the clutch design process to significantly enhance the up-front Systems Engineering practice.

4.1 Friction plate features

Friction plate designs are typically adjusted for every production application to optimize torque characteristics. Among the design variables, groove geometry is considered the most practical to fine-tune open clutch drag as well as engagement and release behaviors. The groove acts as a passive flow control device. The effects of groove changes are generally confined to torque characteristics without inadvertently affecting other requirements. Figure 11 illustrates three friction plates that highlight important groove design considerations, namely (i) overall pattern, (ii) groove shape, and (iii) groove volume. The plate A and the plate B have a radial groove pattern, whereas the plate C has a parallel pattern. There is no well-established engineering methodology to optimize overall groove layout. Accordingly, a number of distinct patterns are introduced to new production systems, largely based on empirical evidence and prior production experience. Although the pattern analysis remains challenging, analytical tools are routinely utilized to optimize individual groove geometry for open clutch conditions. For example, the plate A and the plate C have a conventional straight edge groove, whereas the plate B has a groove with complex asymmetric geometry. CFD (computational fluid dynamics) simulations are conducted to fine-tune flow geometry to enhance pumping capabilities for a given application. Coupled with groove shape optimization, there is a clear trend in production to widen or deepen the groove as illustrated in the plate B. A large groove increases average fluid film thickness between rotating plates, significantly reducing viscous

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

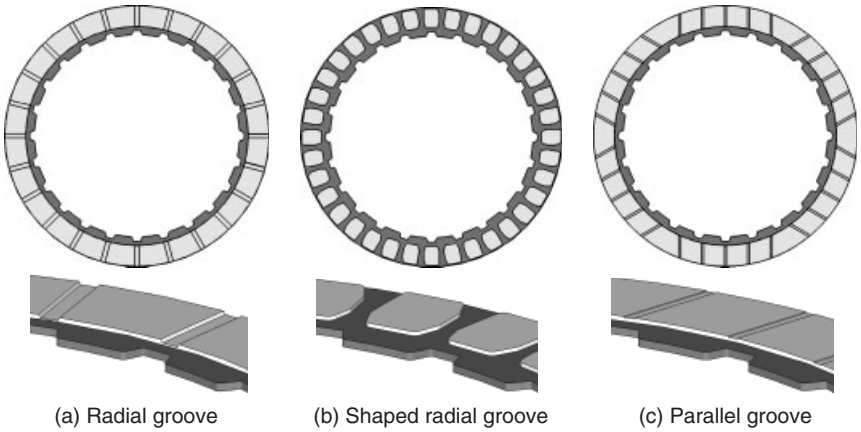


Figure 11. (a–c) Friction plate design features.

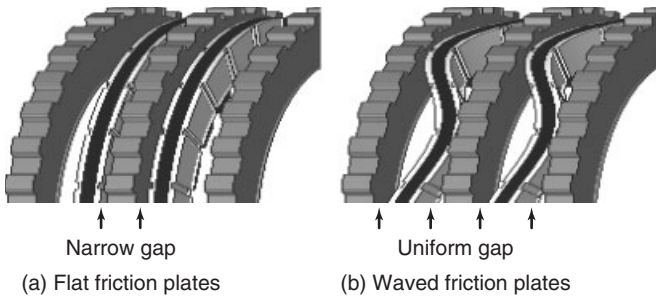


Figure 12. (a,b) Clutch plate clearance.

shear. However, the fuel economy benefit comes at the expense of clutch controllability. A large fluid volume at the interface tends to make the engagement and release behaviors more sensitive to operating conditions.

Another important design trend is a waved friction plate. It is known that both friction and separator plates slide back and forth along spline teeth when the clutch is open. This causes the clearance between some plates to be very small, resulting in large viscous drag, as illustrated in Figure 12a. To prevent the plates from sticking together, a friction plate is formed with several gentle waves in its circumferential direction as depicted in Figure 12. The waves in the figure are exaggerated for the purpose of illustration. The waves help to maintain uniform clearance on average between the plates for reduced viscous drag as in Figure 12b. The waves are designed to collapse at relatively small force levels during the clutch engagement.

4.2 High fidelity clutch model

The functionality of a wet friction device is built on complex physical processes that involve fluid–structure interactions. Its characteristics are highly sensitive to geometric features as well as operating conditions. There have been a number of attempts to analytically, computationally, and empirically capture wet friction behaviors since the 1970s, including Wu (1971, 1978), Patir and Cheng (1978, 1979), Yang, Lam, and Fujii (1998), Fujii, Tobler, and Snyder (2001), Fujii *et al*. (2002, 2003), Cao *et al*. (2004), Deur *et al*. (2005), Yuan, Liu, and Hill (2007), and Ivanovic *et al*. (2009). The earlier research activities often targeted certain physical processes to gain insight into specific clutch characteristics. Today's understanding on clutch operating mechanisms, described in Section 3, is built on those research accomplishments.

The inclusion of detailed geometric features in clutch simulations became possible in the 1990s with the advancement of FEM (finite-element method) and CFD technologies (Berger, Sadeghi, and Krousgrill, 1996; Davis, Sadeghi, and Krousgrill, 2000; Aphale, Schultz, and Ceccio, 2010; Takagi *et al*., 2011). In addition, the recent progress in multi-physics modeling methodologies offer a realistic possibility first time to enable up-front clutch performance analysis through high fidelity simulations. All the

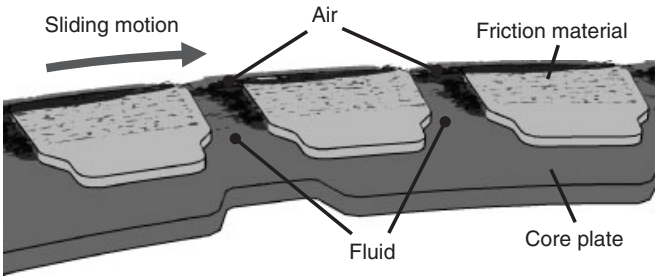


Figure 13. High fidelity clutch simulations.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

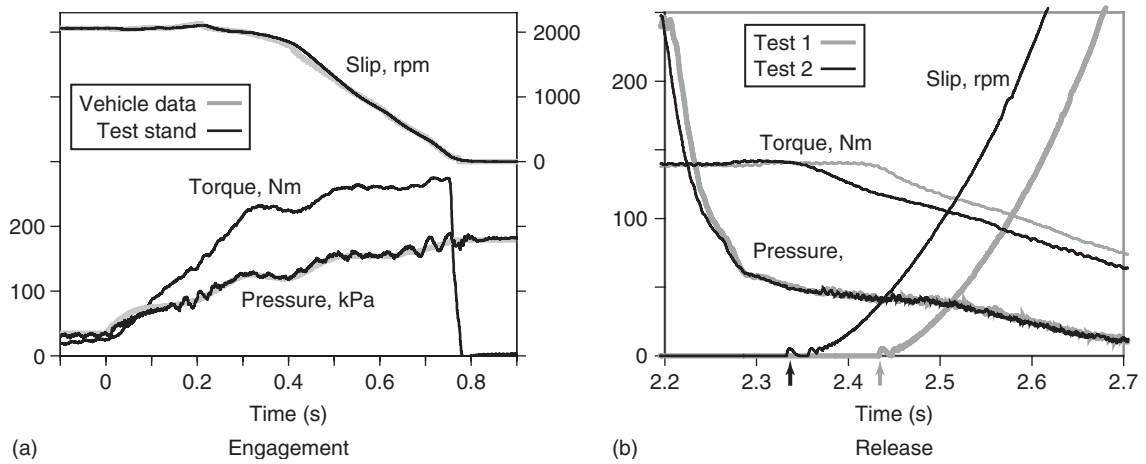


Figure 14. (a,b) Enhanced clutch test methodology.

key physical processes, such as multi-phase flow, squeeze film, porous diffusion, flow through asperities, and heat transfer, can be coupled together for open clutch and transient simulations while allowing rotational and translational plate motions (Cho *et al*., 2011; Cho, 2012). Figure 13 shows an example of three-dimensional high fidelity multi-phase flow simulations for open clutch drag analysis. A friction plate with complex groove geometry slides at 2000 rpm against a separator plate (not shown in the figure) that is 0.2 mm apart. The figure shows three-dimensional iso-surface where air volume fraction locally exceeds 0.9. Air penetrates into the interface from the outer edge of the plate and travels along the leeward side of the grooves. The presence of air is also observed over the segmented friction lining in the simulation. An analytical examination of flow behaviors provides valuable insight into the effects of groove depth, width, and wall shape for increased air entrainment and reduced drag. A new generation of high fidelity multi-physics modeling tools and methodologies appears promising for greatly enhancing up-front clutch design capabilities.

4.3 Enhanced clutch test method

A conventional clutch performance verification test follows the SAE#2 test standard that was developed in the 1970s. The SAE#2 standard defines the specifications for test rig, test procedures, and conditions to facilitate technical communications between OEM and suppliers (Fanella, 1994). The SAE#2 test stand is a type of an inertia-absorption component dynamometer. An electric motor spins friction plates at a target rotational speed, whereas separator plates are mounted on a stationary housing. When a clutch engagement is commanded, the motor becomes decoupled from the clutch plates. A pneumatic or hydraulic actuator applies a force onto a clutch piston, following a step-function profile. The engagement completes when the friction plates are securely grounded. The SAE#2 standard continues to provide a valuable common ground to cross-examine relative performance of various plate designs, friction materials, and the effects of fluid properties. However, SAE#2 procedures do not adequately represent clutch controls in contemporary drivetrain systems. Accordingly, SAE#2 test data do not correlate very well with clutch behaviors observed in an actual vehicle. At present, an enhanced SAE#2 clutch test stand is becoming increasingly available, equipped with programmable actuator and slip controller (Fujii *et al*., 2006). Figure 14 shows the examples of advanced clutch tests conducted on such an enhanced stand. An engagement test in Figure 14a follows actual pressure and slip profiles observed in a vehicle for evaluating clutch characteristics and control sensitivities under realistic shift conditions. A release test in Figure 14b demonstrates the breakaway behaviors that are also the subject of Section 3.3. As the actuator pressure drops, the clutch breaks away when its torque capacity becomes below the load level exerted on to the clutch pack. Test 1 and Test 2 exhibit very different breakaway timing even though their pressure profiles are nearly identical. This is due to the variability of static friction as previously discussed in Section 3.3. Clutch component characteristics obtained from programmable clutch test stands generally exhibit far superior correlation with vehicle-level assessment as compared with the standard SAE#2 test data. The application of the enhanced clutch test methodologies in the clutch design process significantly reduces a number of engineering iterations on the right side of the V-model for accelerated system development.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5 SUMMARY

The wet clutch provides critical functions in conventional automatic transmission systems. Its characteristics have a direct impact on fuel economy, shift quality, and drivability. Because clutch behaviors are highly sensitive to geometry and operating conditions, a clutch design engineer is generally required to adjust design features specifically for each application. They face a challenge to balance conflicting requirements in clutch efficiency and coupling behaviors. Although general-purpose CFD tools are commercially available, their applicability to up-front clutch design support is not well established. Accordingly, a clutch engineer is often forced to rely on prior experience and time-consuming trial-and-error-based design assessment. However, the recent advancement in commercial CFD tools, academic, and industry research present strong evidence that truly useful analytical tools will be available in the very near future.

New powertrain systems such as DCT (see Dual Clutch Transmissions (DCT)—Layouts, Clutch Selection, Packaging, Actuation, Manufacturing & Assembly) and HEV powertrain (see Micro, Mild and Full Hybrid, Power Split Configurations) continue to drive innovations in wet clutch technologies. In order to realize higher fuel-economy benefits, the clutch is often expected to function under unconventional operating conditions. For example, both extended clutch slip and severe duty cycles push the clutch components to their thermal limit. An overall structure must be designed to meet a required thermal rejection rate for a given cooling mechanism. A friction material (see Mechanics of Contacting Surfaces) and fluid additives (see Tribological Optimisation in the Powertrain) require ever-increasing thermal durability. This provides the challenges and great opportunities for clutch engineers for innovations.

Lastly, but not least importantly, it is worthwhile to note that the successful implementation of the wet clutch technology in all applications depends on a clutch actuation mechanism (see Automotive Torque Converters) and a control methodology (see AT Control—Actuation Methods & System Integration, Gear Choice, Gear Shift Strategy & Process, Adaptive Features).

ACKNOWLEDGMENTS

The authors would like to thank Gregory M. Pietron of the Ford Motor Company for providing valuable input to improve both technical and editorial quality of this chapter.

RELATED ARTICLES

Micro, Mild and Full Hybrid
Mechanics of Contacting Surfaces
Tribological Optimisation in the Powertrain
Automatic Transmissions—Geartrain Combinations, Components, Design Considerations, Hydraulic System, Packaging, Manuf., Assembly
Power Split Configurations
Dry Clutch
Automotive Torque Converters
Drive Train Noise, Vibration and Harshness
Passive and Active Limited Slip Differentials
Clutch Actuation
Launch Control
AT Control—Actuation Methods & System Integration, Gear Choice, Gear Shift Strategy & Process, Adaptive Features
Dual Clutch Transmissions (DCT)—Layouts, Clutch Selection, Packaging, Actuation, Manufacturing & Assembly

REFERENCES

Aphale, C., Schultz, W., and Ceccio, S. (2010) The influence of grooves on the fully wetted and aerated flow between open clutch plates. *ASME Journal of Tribology*, **132**, 011104.

Audebert, N., Barber, J., and Zagrodzki, P. (1998) Buckling of automatic transmission clutch plates due to thermoelastic/plastic residual stress. *Journal of Thermal Stresses*, **21**(3/4), 309–326.

Berger, E., Sadeghi, F., and Krousgrill, C. (1996) Finite element modeling of engagement of rough and grooved wet clutches. *ASME Journal of Tribology*, **118**, 137–146.

Cao, M., Wang, K., Fujii, Y., and Tobler, W. (2004) A hybrid neural network approach for the development of friction component dynamic model. *ASME Journal of Dynamic Systems, Measurement, and Control*, **126**(1), 144–153.

Cho, J., Katopodes, N., Kapas, N., and Fujii, Y. (2011) CFD modeling of squeeze film flow in wet clutch. Society of Automotive Engineers, 2011-01-1236.

Cho, J. (2012) A multi-physics model for wet clutch dynamics. PhD Thesis. The University of Michigan, Ann Arbor, Michigan, USA.

Davis, C., Sadeghi, F., and Krousgrill, C. (2000) A simplified approach to modeling thermal effects in wet clutch engagement: analytical and experimental comparison. *Journal of Tribology*, **122**, 110–118.

Deur, J., Petric, J., Asgari, J., and Hrovat, D. (2005) Modeling of wet clutch engagement including a thorough experimental validation. Society of Automotive Engineers, 2005-01-0877.

Fanella, R. (1994) Development and Testing of Friction Clutches in *Design Practices: Passenger Car Automatic Transmissions*, 3rd edn, Society of Automotive Engineers, Warrendale, Pennsylvania, USA, pp. 399–409.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fujii, Y., Tobler, W., and Snyder, T. (2001) Prediction of wet band brake dynamic engagement behaviour Part1: mathematical model development. *IMechE Journal of Automobile Engineering*, **215**(D4), 479–492.

Fujii, Y., Tobler, W., Clausing, E., *et al*. (2002) Application of dynamic band brake model for enhanced drivetrain simulation. *IMechE Journal of Automobile Engineering*, **216**(D11), 873–881.

Fujii, Y., Tobler, W., Pietron, G., *et al*. (2003) Review of wet friction component models for automatic transmission shift analysis. Society of Automotive Engineers, 2003-01-1665.

Fujii, Y., Snyder, T., Waldecker, R., *et al*. (2006) Dynamic characterization of wet friction component under realistic transmission shift conditions. Society of Automotive Engineers, 2006-01-0151.

Gott, P. (1991) *Changing Gears: The Development of the Automotive Transmission*, Society of Automotive Engineers, Warrendale, Pennsylvania, USA.

Greenwood, J. and Williamson, J. (1966) Contact of nominally flat surfaces. *Proceedings of the Royal Society*, London, Series A, 295, 300–319.

Hamrock, B. (1994) *Fundamentals of Fluid Film Lubrication*, McGraw-Hill, Inc., New York, USA.

Hirano, T., Maruo, K., Gu, X. and Fujii, T. (2007) Development of friction material and quantitative analysis for hot spot phenomenon in wet clutch system. Society of Automotive Engineers, 2007-01-0242.

INCOSE (International Council on Systems Engineering) (2010) *Systems Engineering Handbook*, Version 3.2.

Ivanovic, V., Herold, Z., Deur, J., Hancock, M. and Assadian, F. (2009) Experimental characterization of wet clutch friction behaviors including thermal dynamics. Society of Automotive Engineers, 2009-01-1360.

Kitabayashi, H., Li, C. and Hiraki, H. (2003) Analysis of the various factors affecting drag torque in multiple-plate wet clutches. Society of Automotive Engineers, 2003-01-1973.

Otani, C. and Kimura, Y. (1994) Analysis of real contact area of a paper-based wet friction material. *Japanese Journal of Tribology*, **39**(12), 1488–1494.

Patir, N. and Cheng, H. (1978) An average flow model for determining effects of three dimensional roughness on partial hydrodynamic lubrication. *ASME Journal of Lubrication Technology*, **100**, 12–17.

Patir, N. and Cheng, H. (1979) Application of average flow model to lubrication between rough sliding surfaces. *ASME Journal of Lubrication Technology*, **101**, 220–230.

Takagi, Y., Nakata, H., Okano, Y., *et al*. (2011) Effect of two-phase flow on drag torque in a wet clutch. *Journal of Advanced Research in Physics*, **2**(2), 021108.

US DOT (Dept. of Transportation). (2007) Systems Engineering for Intelligent Transportation Systems. Publication No. FHWA-HOP-07-069, http://ops.fhwa.dot.gov/publications/seitsguide (accessed 24 October 2013).

Winchell, F. and Route, W. (1961) Ratio changing the passenger car automatic transmission. Society of Automotive Engineers, 311A.

Wu, H. (1971) The squeeze film between rotating films. *Wear*, **47**, 371–385.

Wu, H. (1978) A review of porous squeeze films. *Wear*, **60**(4), 274–281.

Yang, Y., Lam, R. and Fujii, T. (1998) Prediction of torque response during the engagement of wet friction clutch. Society of Automotive Engineers, 981097.

Yesnik, M. (2002) The influence of material formulation and assembly topography on friction stability for heavy duty clutch applications. Society of Automotive Engineers, 2002-01-1436.

Yuan, Y., Liu, A., and Hill, J. (2007) An improved hydrodynamic model for open wet transmission clutches. *Journal of Fluids Engineering*, **129**, 333–337.

Zagrodzki, P. and Farris, T. (1998) Analysis of temperatures and stresses in wet friction disks involving thermally induced changes of contact pressure. Society of Automotive Engineers, 982035.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto090
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Automotive Torque Converters

Donald Maddock
General Motors Company, Pontiac, MI, USA

1 Introduction 1
2 Construction and Operation 2
3 Converter Performance 8
4 Converter Matching 12
5 Design Considerations 15
6 Clutching Devices 23
Acknowledgments 24
References 24
Further Reading 24

1 INTRODUCTION

The torque converter is a hydrodynamic machine used in conventional automatic transmissions to transmit power from the engine to the transmission gear system. In its classical configuration, power flow is accomplished through the action of a moving stream of fluid, with no mechanical connection between input and output shafting. It provides a beneficial dynamic relationship between the torque-speed characteristics of reciprocating internal combustion engines and the tractive effort requirements of automotive vehicles. It also effectively isolates the driveline mass-elastic system from engine torsional disturbances.

1.1 Classification

Torque converters are a type of turbomachine. These devices are distinguished by having rows of rotating blades that exchange energy with a stream of moving fluid. Forces between the blades and fluid influence the motion of the fluid while producing torque on the blade system. In the simplest of turbomachines, there is a single transfer of energy. Pumps and compressors, for instance, convert shaft power to fluid pressure and velocity, whereas turbines do the inverse of this process. Torque converters are a bit more complex machines, with mechanical power transforming to fluid power and back to mechanical power.

Two related devices are prevalent in heavy-duty and off-road vehicles and are generalized with torque converters as hydrodynamic transmissions. While true torque converters have the capability to produce output torque that exceeds input torque, fluid couplings have no capability to multiply torque. Retarders are fluid couplings that have their output member grounded. They function as vehicle brakes.

1.2 Brief history

The torque converter was invented by DrIng Hermann Föettinger shortly before World War I (Voith, 1988). He was employed by the Vulcan ship works in Hamburg, Germany, and was involved in the application of steam turbine engines to naval vessels. Turbines had evolved significantly since their invention in the previous century and were achieving power levels well in excess of 20,000 kW. Gear systems were not yet available to durably match the high rotating speed of the turbine to the much lower optimum speed of the ship's propeller, leading Föettinger to conceive the torque converter to provide the necessary speed reduction. His early prototypes were soon followed by a successful transmission system that utilized separate torque converters to effect normal and reverse rotations of the propeller without needing to reverse the engine. Direction of motion was selected by filling one of the units with working fluid while evacuating the other (Wislicenus, 1947).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Hydrodynamic devices were introduced in passenger cars with the late 1930s Chrysler Fluid Drive and the 1940 General Motors' Hydra-matic. The Fluid Drive used a fluid coupling between the engine and a conventional manual clutch and layshaft (countershaft) transmission, whereas the Hydra-matic combined a fluid coupling with a fully automatic four-speed planetary gearbox. The post-World War years saw a proliferation of planetary automatic transmissions of various sorts and, with them, a number of different torque converter configurations. Some of these designs, particularly those produced by Buick in their line of Dynaflow transmissions, were quite sophisticated, combining several input, output, and reaction blade rows with internal gear sets and one-way clutches (Chayne, 1948). These units provided ratio coverage from vehicle launch to road load cruise conditions without sensible step gear ratio changes. By the early 1960s, however, improved transmission shifting technology made this sort of complex device unnecessary, and the automobile industry generally settled on the simple, three-element Trilok converter, which is in general use nowadays. In the 1970s, fuel economy concerns prompted the inclusion of a wet clutch within the envelope of the hydrodynamic unit, producing a device known as a *lock-up torque converter*. The clutch is externally controlled and, when desirable, directs power directly to the gear system. This eliminates energy loss associated with fluid motion through the blade system.

1.3 Attributes

Torque converters continue to be widely utilized in automotive transmissions because of their advantageous combination of operating characteristics. There is currently no other device that simultaneously allows an unloaded engine idle, smooth ratio changing on acceleration, and good driveline torsional vibration damping. They also transmit large amounts of energy without excessive power densities at any surface or interface.

Converters do provide a degree of challenge to the transmission designer. They are relatively large, heavy components, frequently causing packaging and mass target difficulties. They are expensive to tool and produce and tend to be sensitive to manufacturing variations.

In operation, the converter imposes loads on engine and transmission components that require additional structure and bearing content. Of greater consequence is that, compared to gear systems, even the best torque converter is an inefficient device, generating sizeable quantities of heat and impacting the fuel economy of the vehicle. The heat generation requires that oil-to-water or oil-to-air heat exchanger systems be included in the vehicle. The lock-up clutch, or torque converter clutch, partially overcomes the fuel economy problem but introduces other difficulties of its own.

2 CONSTRUCTION AND OPERATION

2.1 Mechanical arrangement

2.1.1 Mounting and shafting

A typical modern automotive torque converter is shown in Figure 1 and illustrated schematically in Figure 2. Externally, the device appears as a closed, sealed assembly that rotates with the engine crankshaft. The front closure,

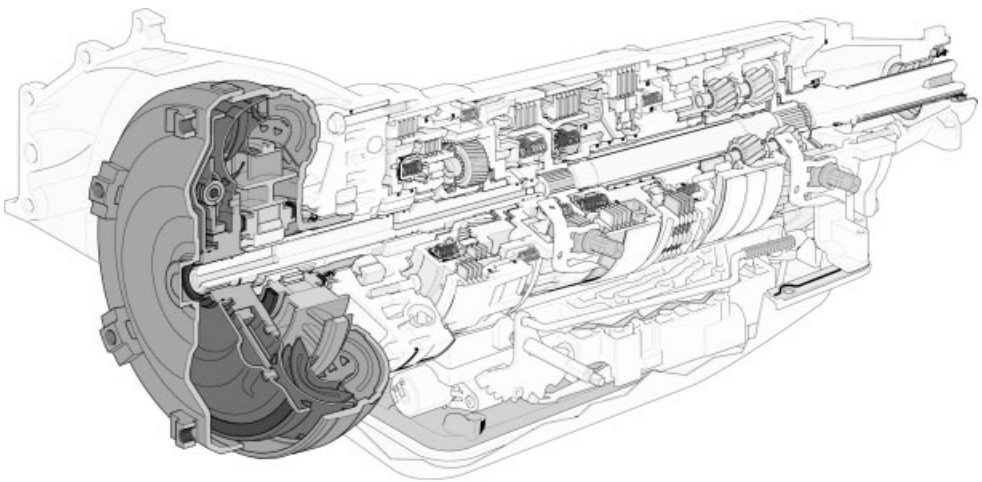


Figure 1. Automotive torque converter. (Reproduced by permission of GM Global Technology Operations LLC.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

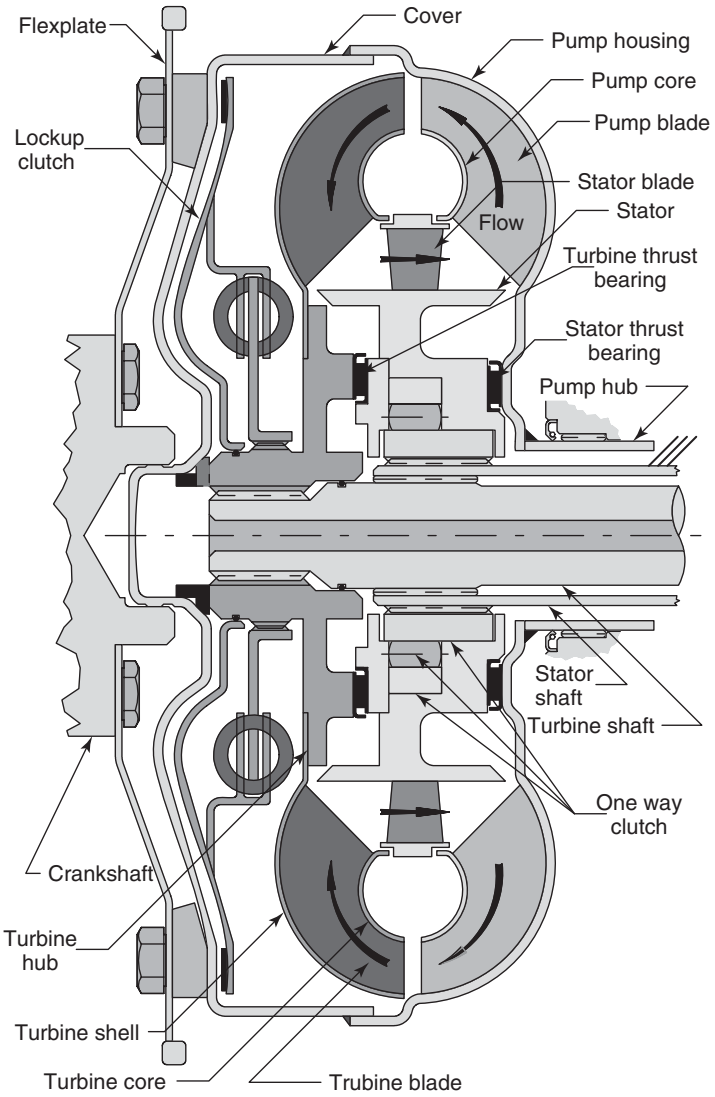


Figure 2. Torque converter schematic.

the cover, is piloted by a precision bore on the crankshaft axis. Power flows from the engine to the cover through a bolted connection to the engine flexplate. The rear of the converter is supported by a bushing or bearing affixed directly or indirectly to the transmission case, and the converter rotates within a partially sealed bell housing formed by the case, engine block, and access closures. Two concentric shafts enter the converter from the transmission. The outermost is fixed against rotation, whereas the inner shaft transmits power from the converter to the transmission gear system. Some transmissions, particularly transverse front wheel drive designs, have a third shaft inside the input shaft to drive the transmission hydraulic pump. Most rear wheel drive transmissions power the pump from the converter hub.

2.1.2 Cooling oil and clutch control

To function properly, the converter must be completely filled with transmission fluid and accommodate a constant through-flow of oil to carry off rejected heat. Two paths connect the converter to the transmission hydraulic system. One feeds cooling oil into the cavity between the converter cover and the lock-up clutch. The second removes hot oil from behind the stator and routes it to the transmission cooler (Figure 3a). Cooling flow thus directed moves the clutch out of contact with the cover surface, activating power transfer through the bladed elements. The clutch is applied by reversing the flow in the two paths. Feeding oil behind the stator raises the pressure behind the lock-up clutch, forcing it into contact with the cover surface

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

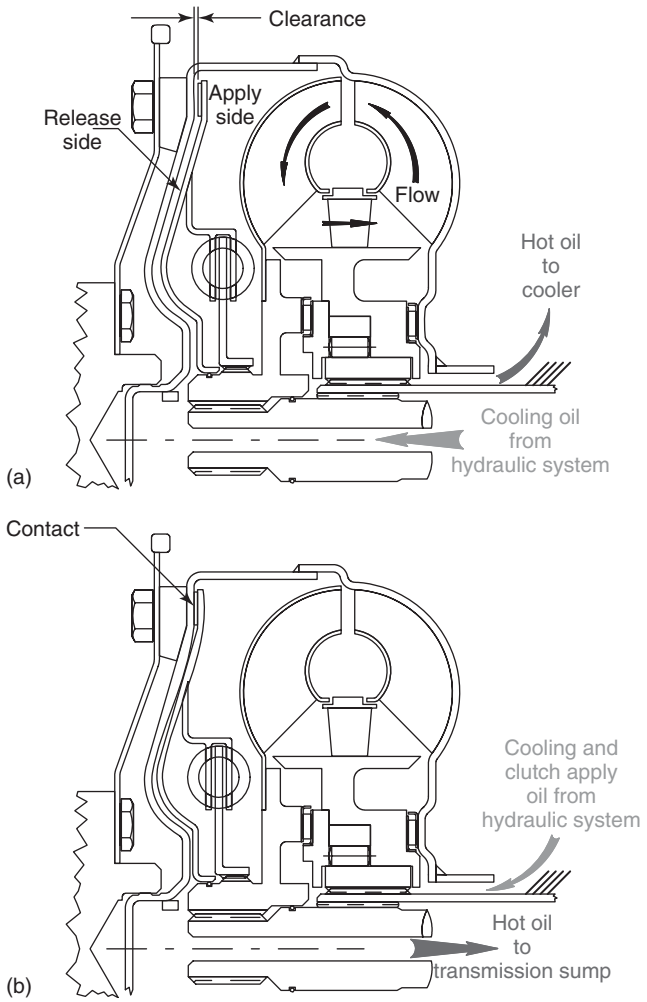


Figure 3. (a) Oil flow with clutch released. (b) Oil flow with clutch applied.

(Figure 3b). Power is then transmitted mechanically, and all hydrodynamic action stops.

In the applied position, the clutch effectively blocks cooling flow through the converter unless a bypass provision is incorporated. This is normally accomplished by adding grooves or a textured surface to the friction material, but some designs provide a third hydraulic circuit in the shafting.

2.1.3 Bladed elements

Three-bladed elements distributed around a roughly elliptical toroidal flow circuit form the hydrodynamic system of the torque converter. The working fluid circulates continuously from element to element, producing torque that delivers engine power to the transmission.

Mechanical power is first converted to fluid power in a mixed flow pump or impeller that forms the rear closure of assembly. The outer wall or housing of the pump is welded to the cover and thus rotates with the cover and flexplate. Working fluid trapped between the pump blades is subject to a centrifugal pressure field that pushes it radially outward between the blades. The stream thus developed must also rotate with the pump blades, and consequently the moving fluid gains angular momentum as it flows outward. This gain in angular momentum results from increased pressure on the leading blade flank, and the integration of this pressure over the blade area and radius produces the torque that loads the engine.

Fluid issuing from the pump continues into the turbine, where the stream's angular momentum is converted back into a mechanical shaft torque that is transmitted to the transmission gear system. The fluid process in the turbine is the inverse of that in the pump. Working fluid conducted inward by the blades looses angular momentum with radius, and this produces a transverse pressure gradient across the blade passage. Increased pressure appears on the trailing side of the turbine blade, and this creates a torque that passes through the turbine shell, turbine hub, and turbine shaft to the gear system.

Rotation of the turbine produces a centrifugal pressure gradient similar to the pump, and this pressure field opposes the flow induced by pump rotation. Consequently, mass flow through the elements decreases as turbine speed increases. The reduction in flow causes the torque-carrying capacity of the converter to diminish as the turbine accelerates and to vanish completely should the pump and turbine speeds become equal. Thus, the pump speed will always exceed turbine speed when the converter is functioning, and the ratio of turbine speed to pump speed, the operating point speed ratio, is a critical variable in the performance of the converter.

The stator or reactor is an axial flow element between the turbine and the pump. It is normally held against rotation by a one-way clutch, such that it overruns freely in the direction of pump rotation but resists motion in the opposite direction. The stator discharges the flow into the pump in the direction of the pump rotation, supplementing the torque absorbed from the engine. This requires that it turn the flow, producing an angular momentum gradient along the blade and a resultant shaft torque.

2.1.4 Materials and construction

Excepting the stator, automotive torque converters are almost universally constructed of stamped low carbon steel. The cover and housing form a highly loaded pressure vessel and thus are relatively thick, usually in the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.0–4.5 mm range. The steel must support considerable cold working during manufacture, allow leak-free, low distortion welding, and retain its yield strength in heat-affected zones. Most pump assemblies are copper brazed, and the high temperatures and extended time at temperature are extremely detrimental to the mechanical properties of unmodified steel. The hot-rolled sheet for these components is often a specialty steel with specific chemistries that are formable, weldable, and temper resistant.

Core rings, turbine shell, and pump and turbine blades are lighter stampings of cold rolled steel sheet. Stock for blades that are to be built into unbrazed assemblies is often bright finished and processed to a quarter hard temper. Parts that will be brazed do not benefit from the higher cold work temper.

Sheet metal blades are usually retained to the shell and core stampings by a tab and slot arrangement. Pump shell tabs are set into blind embossed slots in the housing, as is evident from Figure 1. Turbine shell tabs and all core tabs project through open slots and are rolled over after assembly. In either case, the tab-rolled assemblies are reasonably strong and rigid. Brazing further improves blade retention and greatly enhances the pressure vessel quality of the pump assembly. It also seals incidental blade to blade and slot leaks, enhancing performance characteristics.

Stators are usually diecast aluminum, but some manufacturers use cast magnesium or molded thermoset plastics. With rare exception, the dies or molds are axially drawn, and this places some design restrictions on the blade contour. Both casting and molding permit the blade thickness to be varied rather freely along its length, however, enabling the application of highly advantageous airfoil blade shapes.

2.2 Shaft torques

2.2.1 Angular momentum equations

The essence of converter operation is that the continuously circulating volume of working fluid is accelerated and decelerated by the different blade systems in a manner that advantageous shaft torques are produced. These torques are a function of the size of the unit, the blade shapes, and the rotating speeds of the shafts. Although these relationships are developed by the blade system's complex internal pressure field, they may be illustrated by a simple analysis of the fluid velocities at the outlet of each element.

The shaft torque on any element is equal to the rate of change in angular momentum flux across that element:

$$T_i = \dot{M}_i - \dot{M}_{i-1} \tag{1}$$

In Equation 1, T_i represents the shaft torque on any element i, $\dot{M}_i$ is the angular momentum flux leaving that element at its outlet, and M_{i-1} is the angular momentum flux leaving the immediate upstream element. The angular momentum flux, the rate at which the angular momentum leaves the element, is defined as the product of the mass flow rate $\dot{m}$, the average radius of the stream from the axis of rotation r, and the component of fluid velocity S tangent to the direction of rotation:

$$\dot{M}_i = \dot{m} r_i S_i \tag{2}$$

Equation 2 reflects two assumptions. The first is the reasonably safe premise that the mass flow is identical at the outlet of all elements. This neglects leakage through blade retention slots and also ignores cooling flow, but these are much smaller than the working toroidal circulation. Second, and of greater importance, is the approximation of the outlet velocity profile by an average value of S which, when multiplied by an average value r, produces the same numeric value as integrating the velocity field from shell to core and blade to blade. This is sufficiently accurate for discussion purposes but is of marginal value for quantitative performance analyses.

Combination of Equations 1 and 2 produces the general torque equation:

$$T_i = \dot{m}(r_i S_i - r_{i-1} S_{i-1}) \tag{3}$$

Equation 3 states that the torque on any blade system is proportional to the change in the radius and tangential velocity product across the system.

If converter blades were simple flat radial surfaces with no curvature, the tangential velocity S would simply be the blade velocity at radius r. However, blades of this type would produce unsatisfactory converter characteristics. To produce desirable converter performance, the blades of each element are specifically shaped to alter the tangential velocity as a function of the mass flow. This is accomplished by angling the blades relative to the meridional surface of the torus.

Figure 4 illustrates the effect of pump outlet blade angle on the fluid tangential velocity at a typical operating speed ratio. In this and subsequent diagrams, the symbol U is the blade speed, F the fluid through flow velocity in the meridional plane, W the through flow velocity relative to the blade, and θ the blade angle. The algebraic signs of U, F and W are always positive, and θ is positive when it produces a component of W in the direction of pump rotation. Observing these conventions, the vector geometry produces the relationship:

$$S_p = U_p + F_p \tan\theta_p \tag{4}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

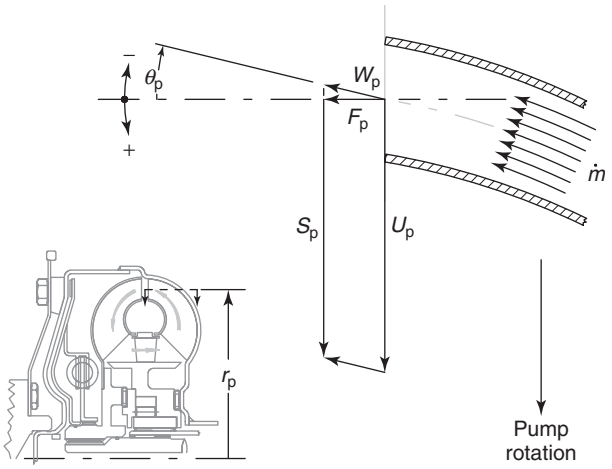


Figure 4. Pump outlet velocities at 0.5 speed ratio.

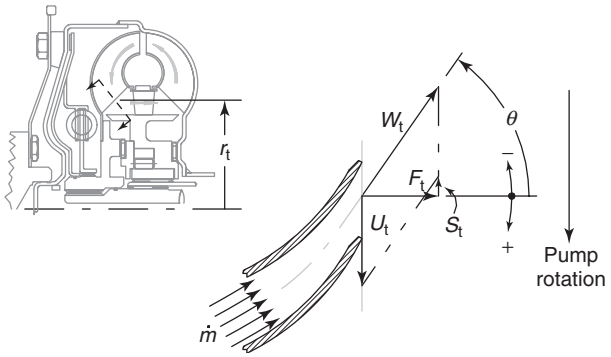


Figure 5. Turbine outlet velocities at 0.5 speed ratio.

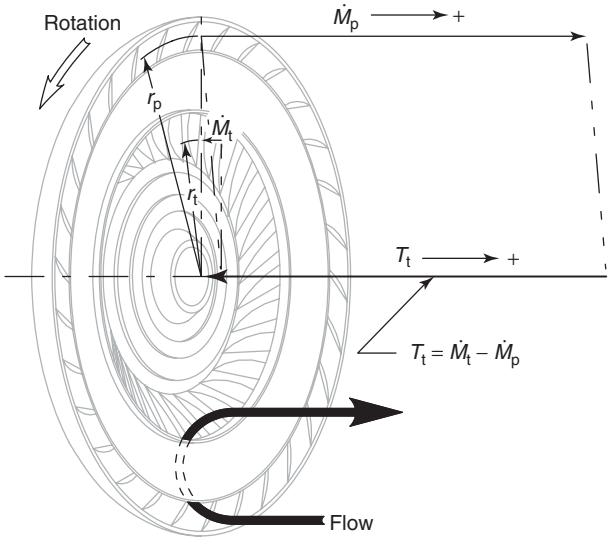


Figure 6. Turbine torque at 0.5 speed ratio.

The subscript p in Equation 4 designates the pump, but the relationship holds generally for all the bladed elements.

Equation 4 is dominated by the blade speed term. The blade speed U_p is the product of the pump angular velocity ω_p and the pump outlet radius r_p:

$$U_p = r_p \omega_p \tag{5}$$

In a typical converter, the pump outlet radius exceeds the turbine and stator radii by 40–50%, and pump speed always exceeds turbine speed when power flows from the engine to the drivetrain. Accordingly, the pump outlet produces the highest tangential velocity in the converter and, as shown by combining Equations 2, 4, and 5, the highest angular momentum:

$$\dot{M}_p = \dot{m} r_p (r_p \omega_p + F_p \tan \theta_p) \tag{6}$$

The turbine blade captures the high momentum flow expressed by Equation 6, and by turning it sharply opposite to the direction of the pump rotation, produces very high shaft torque. The turbine velocities are represented in Figure 5 for a speed ratio of 0.5. Like the pump, the turbine angular momentum is obtained by combining Equations 2, 4, and 5:

$$\dot{M}_t = \dot{m} r_t (r_t w_t + F_t \tan \theta_t) \tag{7}$$

In the conditions illustrated by Figure 5, the effect of the large negative outlet angle exceeds the blade speed, resulting in a negative tangential velocity. In Equation 7, the always negative $F_t \tan \theta_t$ term exceeds the always positive $r_t \omega_t$ term. As the speed ratio increases, however, turbine blade speed will eventually overcome the effect of the negative outlet angle making tangential velocity and angular momentum positive. This typically occurs when the turbine reaches about two-thirds of the pump speed.

The torque exerted on the turbine, as illustrated in Figure 6, is given by subtracting the pump angular momentum from the turbine angular momentum, as in Equation 1. When expanded with the pump and turbine speeds and geometries, this becomes:

$$T_t = \dot{m}[r_t(r_t \omega_t + F_t \tan \theta_t) - r_p(r_p \omega_p + F_p \tan \theta_p)] \tag{8}$$

Equation 8 will always produce a negative value for T_t. The second term, the pump angular momentum, is always a large positive quantity that exceeds the turbine angular momentum, even at high speed ratios. T_t is the torque exerted on the turbine by the transmission, and the negative value indicates that it is opposite in sense from the torque exerted by the engine on the pump. The torque exerted

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

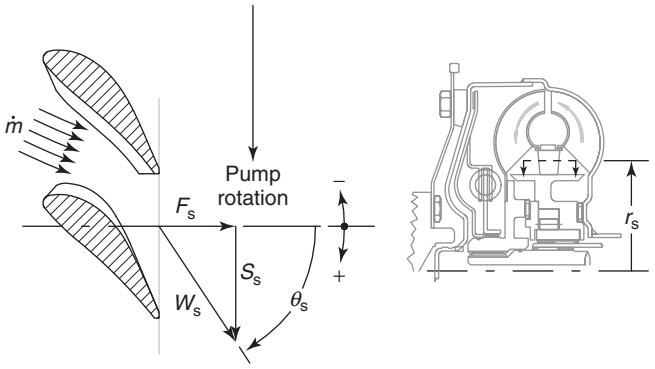


Figure 7. Stator outlet velocities at 0.5 speed ratio.

on the transmission by the converter turbine is opposite in sign and, therefore, has the same sense as the engine torque. The behavior of the turbine angular momentum as speed ratio increases, that is, starting negative and becoming positive, results in transmission input torque (converter output torque) decreasing as speed ratio increases.

The stator receives negative or low positive angular momentum from the turbine and turns the stream back in the direction of pump rotation. This is shown in Figure 7, again for a speed ratio of 0.5. Under these conditions, the stator is held against rotation by the one-way clutch. The stator blade speed is zero, so the angular momentum and torque equations have no $r_s\omega_s$ terms:

$$\dot{M}_s = \dot{m} r_s (F_s \tan\theta_s) \tag{9}$$

$$T_s = \dot{m}[r_s(F_s \tan\theta_s) - r_t(r_t\omega_t + F_t \tan\theta_t)] \tag{10}$$

Stator outlet angles are always positive, so stator angular momentum, as defined by Equation 9 and forming the first term in Equation 10, is always positive. Stator torque is illustrated in Figure 8 for the 0.5 speed ratio condition.

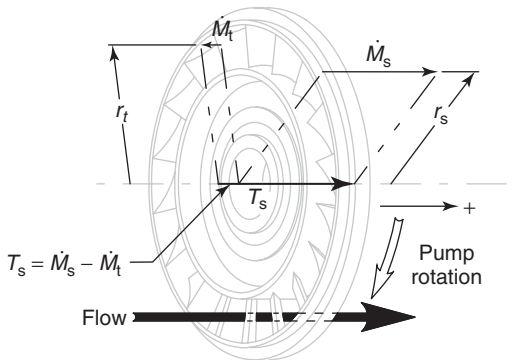


Figure 8. Stator torque at 0.5 speed ratio.

Turbine angular momentum is negative at low and moderate speed ratios but, for this particular converter, becomes positive at speed ratios above 0.66. It continues to increase, and at some higher speed ratio, will exceed the stator angular momentum. At that point, stator torque would become negative, forcing input torque higher than output torque. This being undesirable, the one-way clutch prevents negative stator torque by releasing when stator torque reverses. The stator then freewheels, attaining whatever rotating speed produces no momentum change across the blades. The speed ratio at which the stator overruns is known as the *coupling point*. Above the coupling point, the automotive torque converter losses its capability to multiple torque and thus behaves as a fluid coupling.

Again referring to Equation 1, pump torque is equal to pump angular momentum minus stator angular momentum. At speed ratios below the coupling point, the stator is held against rotation and stator angular momentum is given by Equation 9. The pump torque equation then becomes:

$$T_p = \dot{m}[r_p(r_p\omega_p + F_p \tan\theta_p) - r_s(F_s \tan\theta_s)] \tag{11}$$

Both terms of Equation 11 are always positive, so pump torque is always less than pump angular momentum. This is illustrated in Figure 9.

2.2.2 *Torque multiplication*

As only pump, turbine, and stator torques act on the converter, these torques must add to zero. The angular momentum equations establish the sense and magnitude of

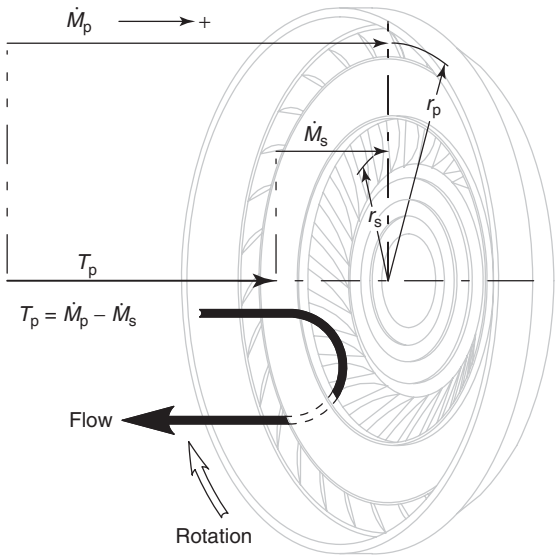


Figure 9. Pump torque at 0.5 speed ratio.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

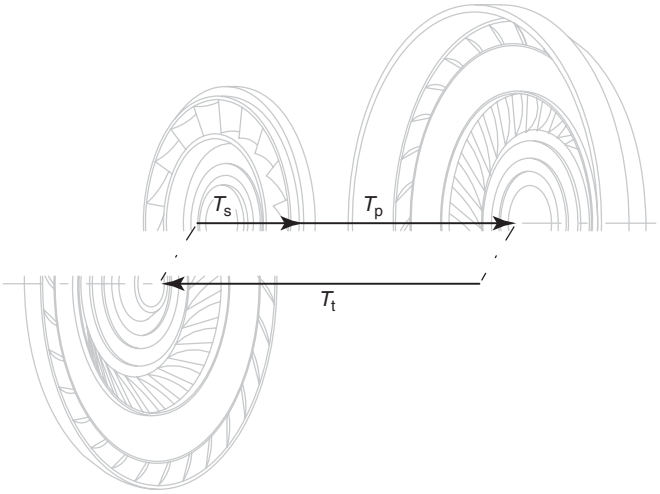


Figure 10. Torque multiplication at 0.5 speed ratio.

these torques, and as illustrated in Figure 10, establish that the magnitude of turbine torque exceeds pump torque by the value of stator torque. This capability is traditionally know as torque multiplication. At the coupling point, stator torque becomes zero and turbine and pump torques become equal. This equality continues at higher speed ratios, as the overrunning one-way clutch prevents any angular momentum change across the stator.

3 CONVERTER PERFORMANCE

The angular momentum equations explain the principles relating the element speeds and fluid mass flow to the shaft torques at a given operating point. However, the operation of a motor vehicle requires that the converter provide suitable relationships between the element torques and speeds under a wide range of conditions. These relationships, known collectively as converter performance, are critical to the performance, drivability, and fuel economy afforded by any powertrain. Further equations based on energy considerations can be used to build a simplified one-dimensional model that predicts converter performance, but these equations rely on coefficients that are not easily derived and consequently do not provide acceptably accurate predictions. Obtaining converter performance is thus an empirical process.

3.1 Performance measurement

3.1.1 Dynamometer equipment

Torque converter speed and torque characteristics are measured on a dynamometer system simulating loading conditions found in service. The converter is enclosed in a fixture that approximates the engine attachment to the converter cover on the input side and the transmission stator shaft, turbine shaft, and pump hub support on the output side. The converter pump is driven by an electric motor, and the turbine torque is absorbed by either a second motor or, on older systems, an eddy current brake. Input and output torques are measured by load cells that are subjected to the reaction torque at the motor and brake frames or by in-line strain gage torque meters in the drive shafts. Stator torque is not directly measured, as it must be the difference between the input and output torques. Transmission fluid is supplied to the converter through the fixture shafting by an external hydraulic power system, and cooling flow is controlled by regulating the converter inlet pressure and outlet back pressure. Inlet oil temperature is established by heaters in the hydraulic sump and an oil-to-water heat exchanger in a separate hydraulic loop. Outlet temperature is measured but not directly controlled, as it is a function of the converter operating point.

3.1.2 Test procedures

Established test procedures for torque converter performance measurement generally involve stabilizing converter operation at a number of operating points and recording pump and turbine speeds and torques at each point. Oil temperature and pressure at the fixture inlet and pressure at the outlet are held constant during the test and outlet temperature and flow rate are recorded at each test point.

The most common performance test is completed by regulating pump torque to a constant value and adjusting turbine speed to sequential values from zero to a maximum. Pump speed and turbine torque are recorded at each point. This is known as an *efficiency test*, and is usually repeated at two input torque levels.

Operation with the turbine stopped and the pump running at speed correlates directly to the vehicle standing with the engine running and the transmission "in gear." Converter behavior under these conditions is critical to idle fuel consumption and vehicle launch performance, so a specific *stall test* is normally run. The turbine shaft is locked against rotation, and the pump torque is regulated to a series of values. Pump speed and turbine torque are recorded at each test point. As no shaft power is removed from the turbine during a stall test, the entire energy input to the pump must be dissipated as heat. With most of this heat absorbed by the working fluid, a cooling cycle is executed between each stall point to prevent excessive oil temperatures.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

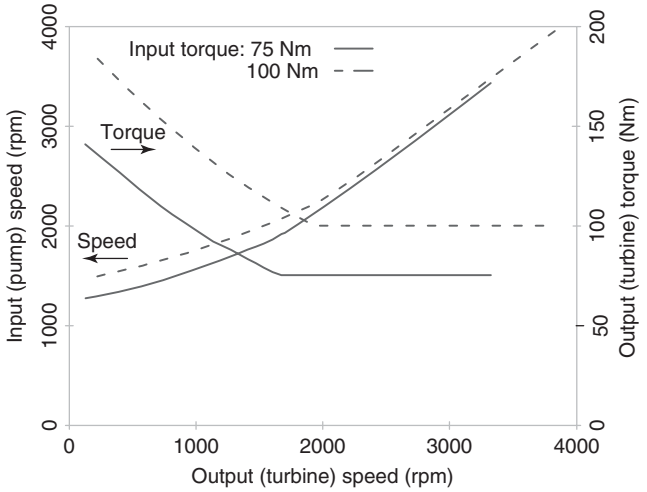


Figure 11. Efficiency test at 75 and 100 Nm.

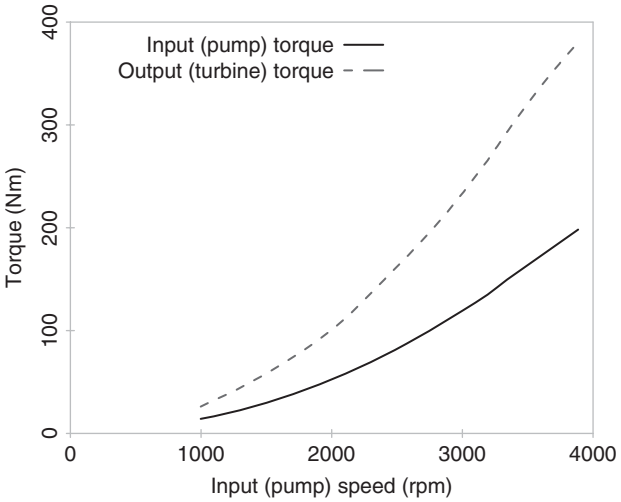


Figure 12. Stall test.

3.2 Performance test data

3.2.1 Measured performance

Typical converter efficiency test data obtained at two different pump torques are plotted on Figure 11. At both torques, the turbine torque is at its maximum at low turbine speed and decreases monotonically as the turbine speed increases. At some elevated speed, the turbine torque becomes equal to the pump torque, indicating that the converter has reached its coupling point and torque multiplication is no longer available. The torques then remain equal as turbine speed continues to increase. Pump speed climbs monotonically to the coupling point, with a gradual increase in slope. Beyond the coupling point, turbine and pump speeds become asymptotic. When the test torque is increased, both turbine torque and pump speed curves move upward and stretch along the abscissa, but the shape of the curves remains essentially similar.

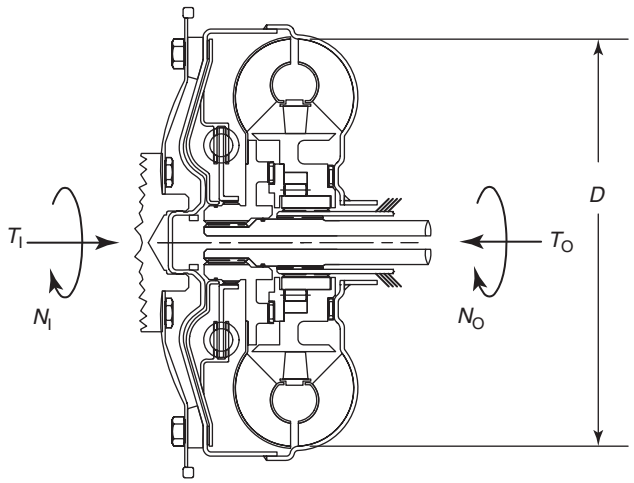
Typical stall test data are shown in Figure 12. As pump speed increases, pump torque and turbine torque increase parabolically. Turbine torque exceeds pump torque at all values of pump speed.

3.2.2 Nondimensional performance

Stall and efficiency test data are combined to extend the efficiency test speed-torque characteristics of the converter to zero turbine speed. While this produces performance information covering the converter's entire operating speed range, the relationships only hold for the arbitrarily chosen constant input torque. However, the data gleaned from an extended test becomes much more useful and revealing when it is nondimensionalized. In this process, the speeds and torques at each test point are converted to dimensionless parameters formed from the test measurements, the density of the working fluid, and a length scale indicating the size of the converter. Formal dimensional analysis allows several formulations of the resulting parameters, but the most useful are the ratio of output speed to input speed, the ratio of output torque to input torque, and a factor combining the input speed, the input torque, the fluid density, and the converter diameter. These dimensionless ratios, illustrated in Figure 13, are known as the *speed ratio*, the *torque ratio*, and the *unit input speed*. In a concession to established practice, the shaft speeds N_I and N_O are measured in revolutions per minute rather than radians per second. Test equipment always exerts a degree of frictional drag on the input and output shafting, and these fixture losses must be deducted from the indicated motor torque and added to the indicated brake torque before the ratios are computed. The shaft torques, T_I and T_O, reflect the torques actually applied to the pump and the turbine. The outside diameter of the converter torus, D, measured over the pump blade outlet at the shell, is customarily used as the length scale. Finally, the fluid density, ρ, is a mild function of the operating temperature and should be obtained at the converter test temperature.

Once the series of converter test speeds and torques are converted to dimensionless ratios, the information may be used to analyze the performance characteristics of a converter design over its entire torque and speed operating ranges, without regard to the size of the unit or the density of the working fluid. The unit input speed parameter, $N_I\ (\rho D^5/T_I)^{1/2}$, is a very informative parameter, as it relates the converter's torque capacity to these quantities. It is a bit

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Dimensional measurements and properties

Symbol	Measurement	Units
N_I	Input (pump) speed	[rpm]
T_I	Input (pump) torque	[Nm]
N_O	Output (turbine) speed	[rpm]
T_O	Output (turbine) torque	[Nm]
D	Converter torus diameter	[m]
ρ	Working fluid density	[kg m^{-3}]
μ	Working fluid viscosity	[kg m^{-1}s^{-1}]

Dimensionless ratios

Symbol	Parameter	Formula
n	Speed ratio	N_O/N_I
t	Torque ratio	T_O/T_I
u	Unit speed	$N_I\,(\rho D^5/T_I)^{1/2}$
E	Efficiency	$N_O T_O/N_I T_I$
Re	Reynold's number	$(\rho T_I/D\mu^2)^{1/2}$

Semi-dimensionless ratios

Symbol	Parameter	Formula	Units
K_I	Input K-factor	$N_I/T_I^{1/2}$	[rpm Nm$^{-1/2}$]
K_O	Output K-factor	$N_O/T_O{}^{1/2}$	[rpm Nm$^{-1/2}$]

Figure 13. Dimensionless ratios.

cumbersome to use for application studies, however, where the converter diameter and fluid are established constants. It is then normally replaced by the semi-dimensionless input K-factor, $N_I/T_I{}^{1/2}$, in which diameter and density are ignored. The input K-factor, or input capacity factor, may be visualized as the speed a given converter would attain when subjected to unit torque. A similarly formed output K-factor, $N_O/T_O^{1/2}$, is useful in studies matching converter characteristics to vehicle road load demands.

The very important converter efficiency, expressed as a percent, is the ratio of the output power, $N_O T_O$, to the input power, $N_I T_I$. Conveniently, it is equal to the product of the speed ratio, N_O/N_I, and torque ratio, T_O/T_I.

Performance data thus reduced are regularly used to model vehicle performance over a large range of engine maps, vehicle characteristics, and driving cycles. It is also used to predict the performance of geometrically similar converter designs of different sizes. However, there are some limitations that must be observed when acquiring data for these application studies and design development.

The dimensionless ratios do not capture the effects of internal mechanical friction among the pump, turbine, and stator and between the stator and ground. Internal pressure gradients create large axial forces on all three elements, and these forces load thrust bearings between the elements. The bearings generate a torque path between the elements that is of little consequence except at very low test torques. The stator one-way clutch similarly generates a small torque to ground when the stator is overrunning. This torque is also insignificant except at very low input torques.

Cavitation becomes an issue when converter charge pressure is low relative to the input power. Torque is developed when the velocity field inside an element produces a pressure drop across its blades. The pressure on the leading blade surface increases, whereas the pressure on the trailing side lessens. Without sufficient charge pressure superimposed over the hydrodynamic pressure field, the trailing side pressure will eventually fall below the vapor pressure of the working fluid. When this occurs, the fluid locally boils, generating a stream of vapor bubbles that propagate downstream. The resulting changes in the effective density of the fluid and disruptions to the local flow field produce significant undesirable deviations in performance.

Friction effects at low torque and the onset of cavitation at high torque define the proper range for acquiring test data for use in subsequent studies. Because both internal friction and the propensity to cavitate are highly dependent on a converter's specific design, the limits must be ascertained during development testing of any new converter. Fortunately, all converters present a very large range of torques that will allow accurate representation of their performance by the dimensionless ratios.

Torque converter performance is relatively insensitive to the large temperature-driven changes in viscosity common to mineral-based transmission fluids. Turbomachines, in general, are relatively indifferent to working fluid viscosity, and most converter test and driving cycles elevate fluid temperatures quite rapidly. However, viscosity effects have been found to become important when designs are resized or "scaled" over a large change in diameters. The ratio termed Reynolds number, $(\rho T_I/D\mu^2)^{1/2}$, in Figure 13 was

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

developed to improve scaling fidelity in cases where viscosity is likely to be important. As input torque and cooling flow temperature are normally constant during efficiency testing, the Reynolds number can be considered a test constant. The best scaling of converter performance data is afforded when the sample converter is tested at the Reynolds number applicable to the ultimate utilization of the test results.

3.2.3 The efficiency plot

The nondimensional performance characteristics of converters are illustrated graphically on the efficiency or speed ratio plot, Figure 14. Because torque ratio, unit speed, input *K*-factor, and efficiency are single-valued functions of speed ratio, it is used as the independent variable. As these curves are dimensionless, they can be constructed from torque and speed data measured at any appropriate input torque. Thus, the torques and speeds from either the high torque test (dashed line) or the low torque test (solid line) from Figure 11, extended to zero speed ratio with stall test data, will produce essentially the same speed ratio plot.

The performance characteristics illustrated on Figure 14 are very typical of modern passenger car and light truck converters. Torque ratio is maximum at stall, usually between 1.5 and 2.5. It decreases monotonically as speed ratio increases, and at the coupling point, approximately at 0.9 speed ratio, becomes exactly 1.0. Above the coupling point, the torque ratio remains constant at unity. The input *K*-factor is nearly constant from stall to about 0.6 speed ratio and then begins to climb. At the coupling point, the slope of the *K*-factor curve increases abruptly. If extended to very high speed ratios, the *K*-factor curve will approach the unity speed ratio abscissa asymptotically. Efficiency is always zero at stall and climbs to a maximum at or slightly before the coupling point. At and above the coupling point, the efficiency exactly equals the speed ratio. The slope of the efficiency curve at stall is exactly equal to the stall torque ratio.

3.2.4 Performance parameters

Torque converter performance data are commonly stored and transmitted in the speed ratio plot graphical format or in an equivalent tabular arrangement. Convenience frequently dictates, however, a shorthand method of describing converter performance. This is most useful when drawing comparisons between various converter designs or when specifying the performance target of a

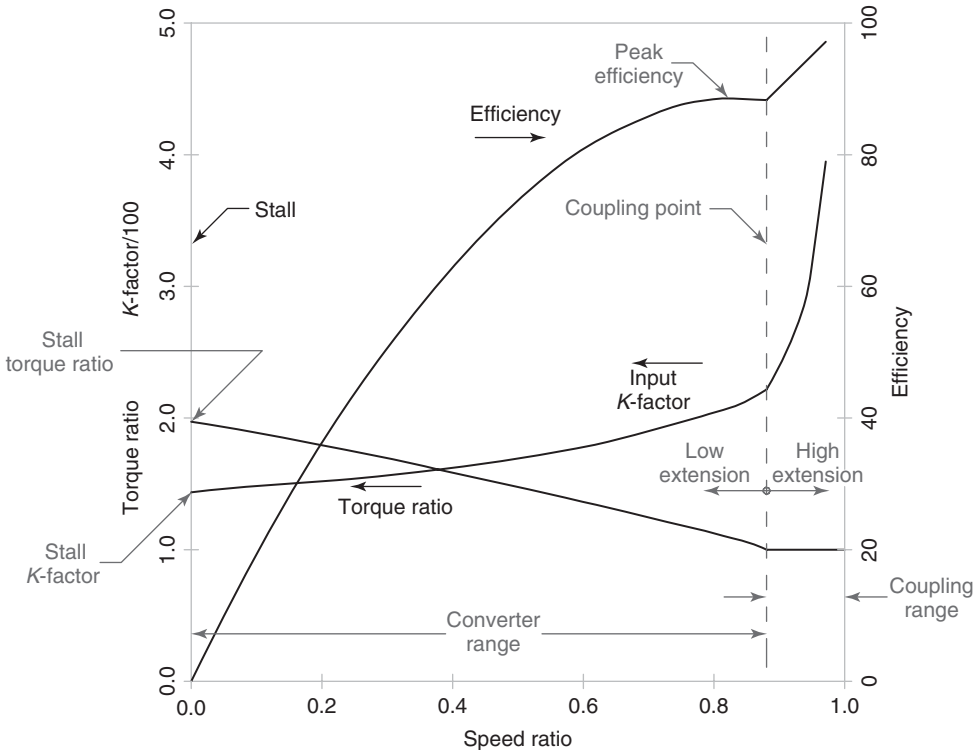


Figure 14. Efficiency or speed ratio plot.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

given application. A traditional terminology, noted on Figure 14, has evolved around several significant features of the speed ratio curve and this terminology is used to convey, somewhat qualitatively, the overall performance of a given unit.

As noted earlier, the stall characteristics of the converter are especially critical. The parameters of consequence at stall are the torque ratio and input *K*-factor, both of which are found at the intersection of their curves with the ordinate axis. The stall input *K*-factor is so important that is commonly appropriated as a model designation for a particular torque converter, and the stall torque ratio of the converter is usually quoted with the *K*-factor. Any reference to the "*K*-factor" and "torque ratio" of a converter may be interpreted as the stall values of these parameters even when the stall designation is not specifically mentioned.

The maximum of the efficiency curve before the coupling point is the "converter efficiency" or "peak efficiency." Its value directly affects the vehicle's fuel economy, particularly during low speed and transient driving.

The relative value of the speed ratio at the coupling point has recently assumed the connotation of "extension." This term was once interpreted in terms of converter output speed at the coupling point, but common usage has evolved to now reflect the speed ratio. It remains a comparative, rather than numerical measure: converters coupling below (roughly) 0.85 speed ratio are "low extension," whereas those coupling above 0.90 are "high extension."

3.2.5 Converter coast characteristics

Torque converters transmit torque from the turbine to the pump when vehicle dynamics drive the turbine faster than the engine. This occurs most frequently when the vehicle is decelerating with the throttle closed. Under these conditions, the speed ratio exceeds unity and flow between the elements reverses. The stator continues to overrun and the converter behaves as a fluid coupling, with input and output torques being equal. Very little torque is transferred until speed ratio exceeds 1.05, but the *K*-factor decreases very rapidly until a speed ratio of 2 is produced. The *K*-factor continues to decrease with further increases in speed ratio but at a much reduced rate.

4 CONVERTER MATCHING

The performance characteristics of the converter have a powerful effect on the performance and fuel economy of the vehicle, so the size and internal geometry of the converter must thus be properly chosen to produce characteristics that will optimize the vehicle attributes. The process of combining engine, vehicle, and converter parameters to attain the best balance of vehicle performance and economy measures is known as *converter matching*. Analysis of the interaction of the prime mover and the converter input is known as *engine matching*; determination of the interaction between the vehicle dynamics and the converter output is *vehicle matching*.

Both graphical and numerical techniques have been developed for matching analyses. The graphical methods predate the general availability of digital computers and are, consequently, limited in quantitative capability. Consequently, they have been largely superseded by highly sophisticated vehicle modeling software that characterizes the engine, torque converter, and gearbox with digital tables of experimentally obtained data. Efficiency test data, obtained at a single test torque and nondimensionalized as noted earlier, provides all the information needed to build the converter tables and allow subsequent simulation of the vehicle system over any route and driving schedule. Transmission producing organizations have large stores of performance data on many diverse converter designs, and the simulation programs allow rapid iteration through many designs to optimize a powertrain package.

All of these studies begin with a selection of converter designs that have a reasonable probability of matching the engine. The transient performance of the powertrain, and consequently the vehicle, during a heavy throttle launch is overwhelmingly determined by the interaction between the torque converter and the engine, so the initial selection of converter characteristics is made primarily on the basis of engine matching. In addition, although the graphical techniques are not quantitatively definitive, they provide a strong insight into the critical powertrain interactions that determine vehicle launch and early in-gear acceleration. Consequently, they continue to be used to guide the converter selection process.

4.1 Absorption curves

To facilitate engine matching, it is necessary to compare the engine's capability to develop torque with the converter's capability to resist that torque. Torque converter speed ratio plots are not convenient formats to facilitate this comparison, but contain all of the information needed to do so. The chart that has been developed to enable engine matching is the plot of "absorption curves," illustrated in Figure 15. These curves are based on the principle that both *K*-factor and torque ratio are single-valued functions of speed ratio. This is equivalent to stating that, at any given speed ratio, a torque converter will produce only one input *K*-factor and only one torque ratio. As the input *K*-factor

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

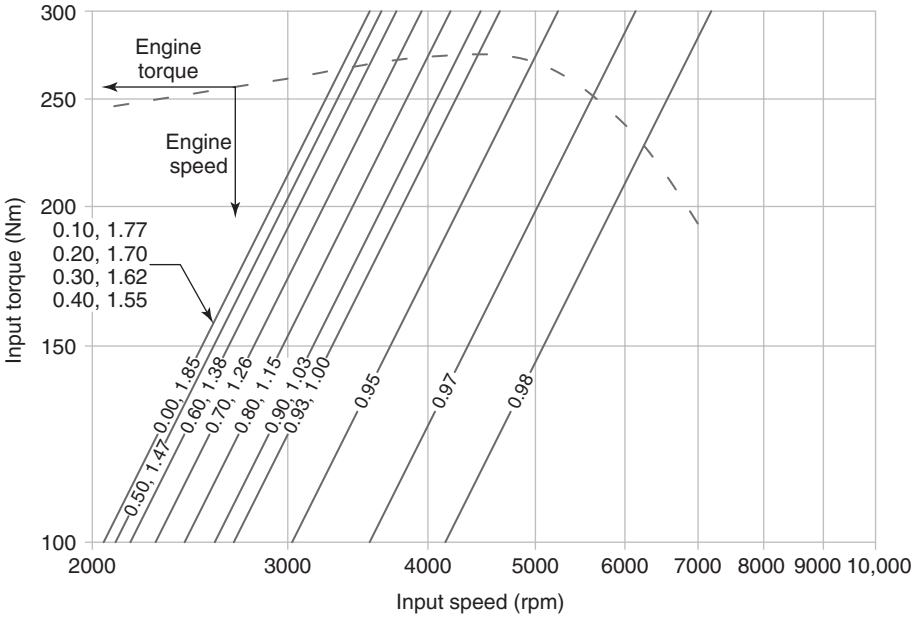


Figure 15. Absorption curves.

is defined as the ratio of input speed to the square root of input torque, the torque absorbed at a constant speed ratio is a parabolic function of the input speed:

$$T_I = N_I^2 / K_I^2 \tag{12}$$

When plotted on a linear grid of torque versus speed, this relationship will appear as a second-order curve, but on the log–log grid of Figure 15, it becomes a straight line with a slope of two. Anywhere on this line, the constant speed ratio and its unique torque ratio will apply. For a different speed ratio, another line will exist, again with a slope of two but its own unique torque ratio. The complete set of absorption curves is generated by figuratively stepping off incremental values along the abscissa of the speed ratio plot, reading the corresponding input *K*-factor, and plotting the corresponding absorption line. These lines are then labeled with the speed ratio and applicable torque ratio. Figure 15 shows several labels assigned to the left-most absorption line. This indicates that multiple speed ratios produce nearly identical input *K*-factors, and consequently superimposed absorption lines. On the speed ratio plot, this appears as a *K*-factor curve that is flat near stall.

As the output torque of the engine, corrected for parasitic loses, is the input torque to the converter, and the engine speed and converter input speeds are identical, the engine torque–speed curve may be plotted on the absorption curve grid. This is also illustrated on Figure 15. The intersections of the engine curve with the individual absorption curves are the match points. At any converter speed ratio, the match point is the condition where the engine's capability to produce torque is exactly equal to the torque converter's ability to resist torque. Consequently, the torque converter will hold the engine at the match point speed, and the engine will produce the match point torque. It follows that during transient engine acceleration, the converter will resist less torque than the engine produces until the match point speed is attained. The difference between the two torques is absorbed by the accelerating system inertia.

An immediate extension of the preceding discussion is that the output speed and torque of the torque converter at each match point can be computed from the absorption line speed ratio and torque ratio. By repeating these calculations for each of the absorption lines, the full output torque versus output speed curve for the combined engine and converter system can be produced. A sample of this type of curve is shown on Figure 16.

4.2 Engine speed regulation

Vehicle performance, by whatever measure, depends on the tractive effort produced at the drive wheels. The primary requisite for maximizing tractive effort, and therefore performance, is to properly position the torque converter absorption lines relative to the engine torque curve. The converter will then regulate the engine speed such that the maximum available engine torque is transmitted to the transmission gear system. Only three key lines must be

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

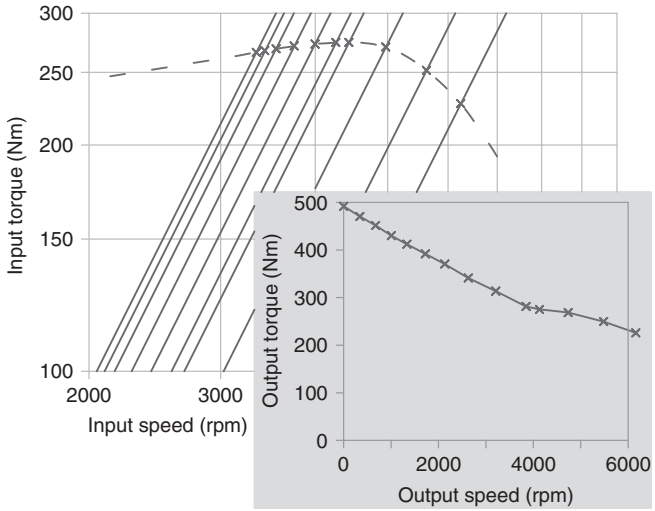


Figure 16. Converter output speed versus output torque derived from engine and absorption curves.

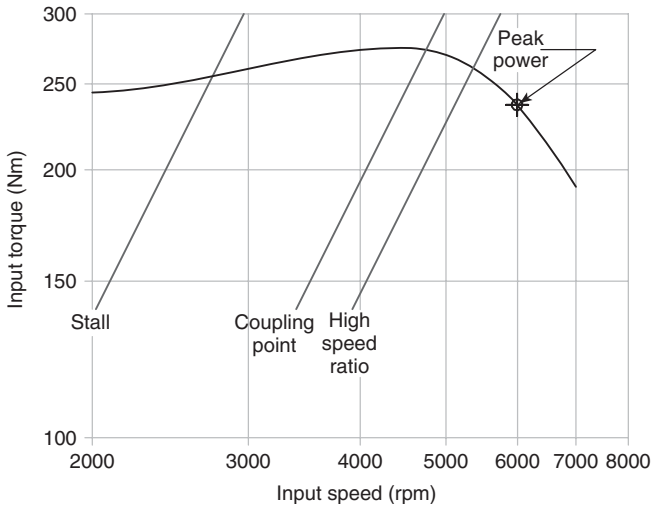


Figure 17. Three-line absorption curve.

considered: the stall line, the coupling point line, and a high speed ratio line. A three-line absorption plot is shown in Figure 17.

4.2.1 *Positioning the stall line—setting the stall K-factor*

Early in the launch event, maximum tractive effort is attained when the engine speed corresponds to the peak engine torque. This occurs when the stall line is positioned to intersect the engine torque curve at its maximum value. The engine speed at the intersection is known as the *stall speed* of the engine–converter combination. It can be verified in the vehicle by locking the brakes, selecting a driving range, and flooring the accelerator. As absorption lines are close together near stall (recall Figure 15), proper selection of stall speed assures good early acceleration. In general, engine speed will then be regulated by the converter to a narrow band near the torque peak until approximately 0.5 speed ratio is exceeded.

Unfortunately, high performance engines having their maximum torque at high engine speeds frequently require some compromise in stall speed. Torque converters are relatively inefficient devices at low vehicle speeds, and consequently a significant percentage of engine power is dissipated as converter heat during launch and acceleration. Engine power is directly proportional to engine speed, so elevating engine speed during launch produces additional converter heating. Not only must this heat be managed by the powertrain cooling system, but the wasted power ultimately represents consumed fuel. Additionally, the time required for the engine to spool up to an elevated stall speed makes the vehicle seem unresponsive, and the additional noise accompanying high engine speeds can likewise be unpleasant. It is thus common practice to limit the stall speed of such applications to approximately 2400 rpm (low power-to-weight vehicles) or 3200 rpm (high power-to-weight vehicles) and accept a decrease in starting tractive effort. Figure 17 shows such a condition, where matching the high engine torque peak would result in heat rejection and drivability problems. Setting the stall speed at 2660 rpm (260 Nm) rather than 4400 rpm (275 Nm) reduces the stall heat rejection rate by 43% at the cost of 5.5% in launch and early acceleration torque.

Positioning the stall line is a graphical analog to set the torque converter stall *K*-factor. The direct effect of this converter performance parameter on the starting tractive effort of the vehicle is the basis for its general acceptance as the single most important converter characteristic.

4.2.2 *The high speed ratio line*

The best tractive effort at the high speed extent of the launch event is attained when peak engine power is transmitted through the torque converter without attenuation. This requires that the converter efficiency be near 100% at the engine power peak. To approach this level of efficiency, the converter must be well into the coupling range before the engine reaches maximum power. As a 3% power loss in the converter can usually be accepted, the general objective is to position the 0.97 speed ratio line at the engine power peak. When less power loss is desired, an even higher speed ratio line can be placed at the peak, but this can exceed the capabilities of many normally sized torque converters.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Unfortunately, torque converter performance data does not always exist for speed ratios above 0.95. In practice, the acceptable high speed tractive effort criterion is usually based on the 0.95 speed ratio line by requiring it to intersect the engine curve well below the power peak. Figure 17 shows the desired condition of positioning the high speed ratio line well below the engine power peak.

4.2.3 The coupling point line

Within the constraints of stall speed and speed ratio at maximum power, it is desirable that torque multiplication be extended over the maximum range of engine speed. This means that the absorption line that represents the coupling point should be as far to the right as possible and intersect the engine curve no lower than the torque peak. The tractive effort during the middle of the launch event (0.5–0.9 speed ratio) will then be maximized, and there will be no unpleasant sag in the vehicle acceleration curve. The distance between the stall and the coupling point lines is proportional to the ratio of the coupling point *K*-factor to the stall *K*-factor. This ratio is sometimes called the *retention* of the converter, and it is qualitatively related to converter extension. It is important to part throttle crowd acceleration and launch.

Unfortunately, converter design features that move the coupling point line to the right also tend to move the intermediate speed ratio lines in the same direction. This is equivalent to imparting slope to the *K*-factor line on the speed ratio plot. *K*-factor slope in the 0.6–0.8 speed ratio range can mildly improve or degrade acceleration, depending on the shape of the engine torque curve, but is surprisingly detrimental to fuel economy. It is most difficult to control in converters that are intended to produce very high or low stall *K*-factors.

4.3 Vehicle perceived performance

Torque converter stall torque ratio and rotating inertia subtly influence vehicle performance in a manner not captured by engine speed regulation analyses. Figure 18 shows the first three seconds of a typical heavy throttle vehicle acceleration curve. Acceleration peaks approximately 1 s after launch and subsequently declines as increasing converter speed ratio reduces torque ratio. The maximum rate of change of acceleration, d^2V/dt^2 or jerk, occurs at about 0.5 s. The driver's perception of the vehicle's performance is greatly influenced by the value of peak jerk.

The numerical value of peak jerk is directly proportional to converter stall torque ratio and inversely proportional to the total engine coupled rotating inertia. The converter generally accounts for about one half of the total inertia, and its contribution is a very strong function of converter diameter. Thus, perceived vehicle performance is enhanced when stall torque ratio is maximized without increasing converter diameter.

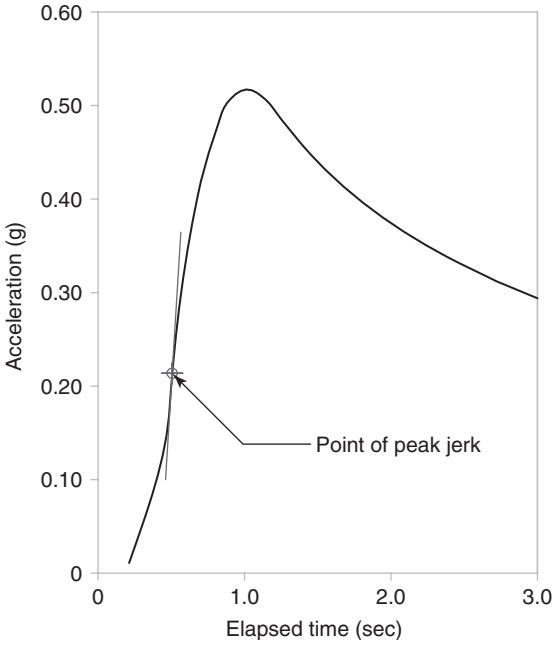


Figure 18. Vehicle acceleration versus time at wide-open throttle.

5 DESIGN CONSIDERATIONS

As noted earlier, the optimum performance and drivability of a vehicle is greatly enhanced by the proper choice of torque converter speed regulation and torque multiplication characteristics. Subsequently, maximizing the efficiency of the converter and fine-tuning of the input *K*-factor line yield significant enhancements to a powertrain's low speed and transient fuel economy. These steps determine the design of the converter hydrodynamic components.

5.1 Shaping the input *K*-factor curve

Proper speed regulation is attained by setting design variables to place the absorption lines at their best location relative to the engine curve, or alternately stated, obtaining the optimum input *K*-factors at the converter's stall, coupling, and high speed ratio points. However visualized, this process establishes the basic shape of the *K*-factor versus speed ratio curve by specifying its values at the three critical points.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

These points on the *K*-factor curve are primarily controlled by the size of the converter, the outlet angle of the pump blade, and the configuration of the stator blade. The influence of each of these features may be considered independently of the others, but their selection for a specific application is strongly dependent on the synergistic relationships between them and their influence on other important converter characteristics.

5.1.1 Diameter effects

Changing torque converter diameter raises or lowers the entire *K*-factor curve without changing its shape. The ratio of the coupling point and high speed *K*-factors to the stall *K*-factor remain constant, and consequently all of the absorption lines move by the same offset while maintaining their relative spacing. Increasing the diameter moves the lines to the left, equivalent to reducing the *K*-factors, whereas reducing the diameter moves the lines to the right, increasing the *K*-factors. This is shown in Figure 19 and is known as *tightening* or *loosening* the converter. The numerical *K*-factor change, as a function of diameter, is derived from the formula for unit speed:

$$K' = K(D/D')^{5/2} \tag{13}$$

Equation 13 applies at any speed ratio when the converter diameter is changed from D to D' while holding the torus design and all the blade angles constant. Computing new converter characteristics in this manner is known as *converter scaling*. It does not change torque ratio or efficiency appreciably, but as illustrated, the large exponent on the diameter ratio strongly effects the *K*-factors.

Two other effects influence diameter selection. Larger converters spread angular momentum changes across increased blade area and radii, thus reducing the leading to trailing side pressure gradient as a function of element torque. Reduced pressure gradient yields higher pressure on the blade trailing side, making the larger converter more resistant to cavitation. It also minimizes the intensity of the turbulent mixing at the outlet end of the blade, suppressing the generation of hydraulic noise.

Converter mass increases roughly with the cube of converter diameter, making thermal capacitance a similar function of diameter. Increased thermal capacitance reduces the converter's tendency to overheat during transient intervals of heavy throttle, low speed operation. Unfortunately, the engine-coupled rotating inertia increases at an even higher rate with diameter, sometimes exceeding a fifth power function. This can have a serious impact on the vehicle's perceived performance.

5.1.2 Pump outlet angle effects

Altering blade designs while holding diameter constant moves both the absolute and relative positions of the absorption lines. A typical absorption curve response to pump blade angle change is shown in Figure 20. These data represent three constant diameter torque converters, identical in all details except pump outlet angle. Increasing the pump outlet angle tightens the converter at stall and, to

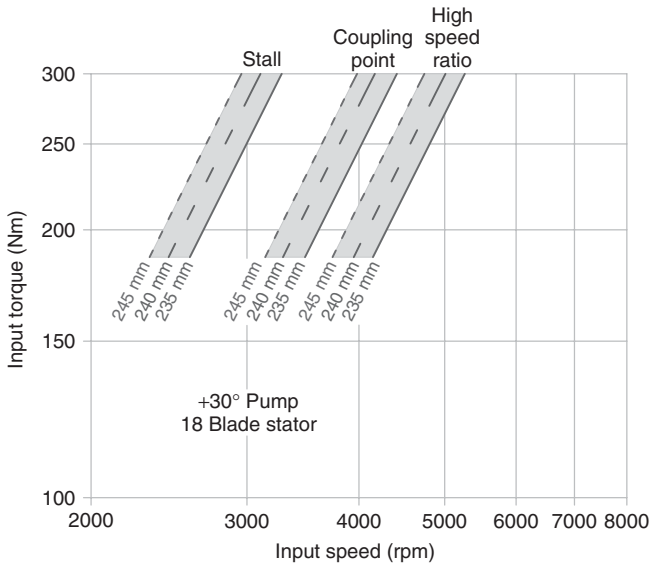


Figure 19. Absorption line response to converter diameter change.

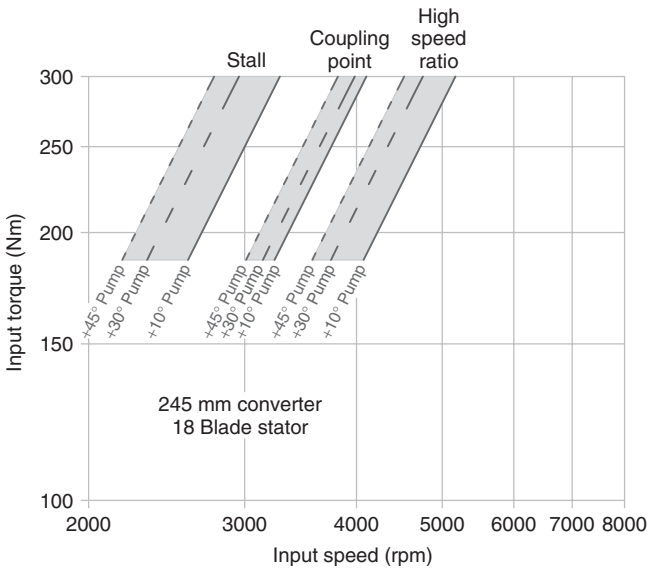


Figure 20. Absorption line response to pump outlet angle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a milder extent, at high speed ratio. The coupling point is even less affected.

The three angles illustrated in Figure 20, +45°, +30°, and +10°, represent the approximate limits of current passenger car practice. Unlike diameter scaling, blade angle changes effect stall torque ratio and efficiency. Increasing pump outlet angle improves efficiency but degrades stall torque ratio to the degree that angles above +45° may produce marginal launch feel. Angles below +10°, although once very common, will produce measurably less than optimum fuel economy.

Pump angle influence trends are not apparent on absorption curves but are effectively demonstrated on the unit speed plot (Figure 21). Here, the independent variable is stall unit speed, the left ordinate is unit speed ratio and torque ratio, and the right ordinate is converter efficiency. Five pump outlet angles are shown: +45°, +30°, and +10° blades of Figure 20 and two additional designs of −15° and −35°. The relationships between outlet angle, stall *K*-factor, stall torque ratio, and converter efficiency are clearly nonlinear but monotonic. Also shown are the ratios of unit speed at the coupling point ($U_{coupling}/U_{stall}$) and high speed ratio point ($U_{0.95}/U_{stall}$). These parameters are identical with the *K*-factor ratios of a given converter, so they are directly proportional to the distance between the respective absorption lines on the log–log plot.

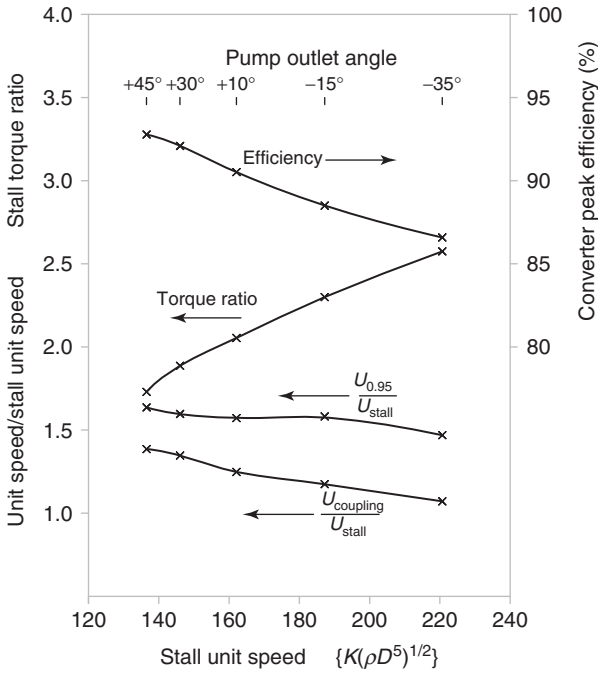


Figure 21. Converter performance response to pump outlet angle.

5.1.3 *Stator outlet angle effects*

Stator effects on absorption lines are highly dependent on the detailed design of the stator airfoil. The stator's short axial length, limited circumference, and wide blade spacing produce a strong synergism between its inlet configuration and thickness distribution that has converter performance influences not captured by outlet angle alone. In general, however, stall *K*-factor is strongly affected by stator outlet angle. Small increases in outlet angle yield large increases in stall *K*-factor, with somewhat smaller increase in the coupling point and high speed *K*-factor. With most airfoil designs, peak efficiency increases with outlet angle, plateaus, and then drops precipitously. Stall torque ratio may behave in a similar manner, or may decrease slowly at first and then more rapidly with increasing angle. Stator families typically have a small range of angles where overall performance is optimized.

Figure 22 illustrates the response of one converter to changes in stator outlet angle. The thin, relatively flat airfoil blade applied to this stator allows significant variation in exit angle without requiring redesign of the entire configuration. When the outlet angle is increased, the blade count is correspondingly reduced to accommodate the increased projection on shell and core circumferential space.

5.1.4 *Turbine outlet angle selection*

Unlike the pump and stator angles, the turbine outlet angle is not used to shape the *K*-factor curve. As the turbine and

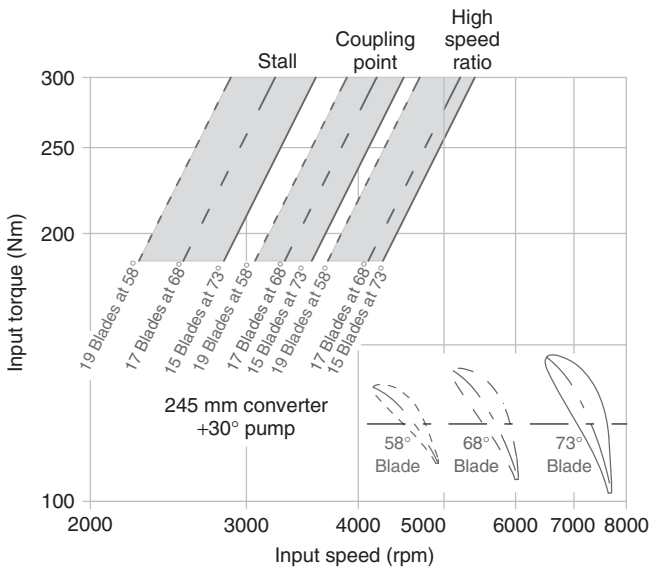


Figure 22. Absorption line response to stator blade count and outlet angle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

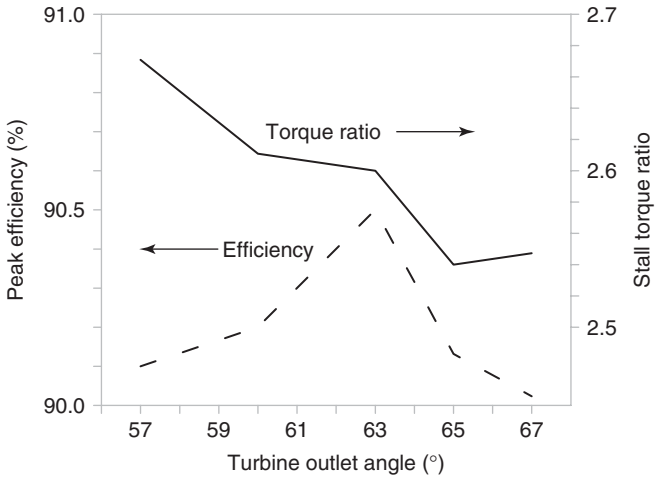


Figure 23. Stall torque ratio and peak efficiency versus turbine outlet angle.

stator torque equations indicate, the torque on both elements increases as the turbine outlet is made progressively more negative. This has such a powerful effect on torque ratio and efficiency that common practice is to set the turbine outlet angle to a very high negative value, usually in the neighborhood of -60°, where these parameters peak. With further increases in outlet angle, the blockage effects of the blade thickness and boundary layer begin to restrict the toroidal flow, causing a reduction in mass flow rate and decline in performance measures. This is illustrated for a typical converter configuration in Figure 23. On this unit, as turbine angle is swept from -57° to -67°, peak efficiency shows a definite maximum at -63° with stall torque ratio already declining. It is very probable that the exact location of the optimum angle for either of these measures is a function of other design features.

5.2 Tuning the torque ratio and efficiency curves

5.2.1 Hydraulic loss

The continuous circulation of working fluid through and between the converter-bladed elements involves a number of processes that dissipate energy. Turbulent eddies are created and dispersed, ultimately producing heat that cannot be recovered. The lost fluid energy accounts for the difference between the converter's input and output mechanical shaft power that directly define the efficiency curve. Decreasing the hydraulic loss at any speed ratio boosts the fluid mass flow. As the mass flow appears in all three torque equations, the torque ratio at that speed ratio increases, and efficiency, being the product of speed ratio and torque ratio, improves.

The equality between mechanical loss and hydraulic loss is best stated in terms of a pressure term rather than the more conventional hydraulic head:

$$N_O T_O - N_I T_I = \dot{m} P_L / \rho \tag{14}$$

The quantity $\dot{m}/\rho$ is the volumetric flow rate, Q, of the working fluid, so Equation 14 may be restated:

$$P_L = (N_O T_O - N_I T_I)/Q \tag{15}$$

In Equations 14 and 15, P_L is numerically the rate of energy loss per unit volume of flowing fluid. It may be visualized as the pressure required to push the working fluid around the torus and through the blade systems.

The pressure loss is traditionally represented as having two components: friction loss and incidence loss. Before the availability of modern computational fluid dynamics (CFD) programs, formulas developed for these components, with the torque equations presented earlier, provided the only means of analytically predicting the performance of new converter designs. With a number of empirical enhancements, they were used with some success for several decades. However, subsequent laser velocimeter and pitot probe measurements revealed that the actual converter flow field is much more complex than represented by these simple models, and they have been largely discarded as quantitative tools. Nevertheless, they continue to provide an effective means of guiding the selection of converter design parameters that minimize hydraulic energy losses, or at least distribute them across the operating range in an optimum manner.

5.2.2 Friction loss

Friction loss models the energy required to move the working flow around the torus without interaction with the blade systems. It was originally formulated as a form of pipe loss, complete with surface roughness, path length, passage hydraulic diameter, and average through-flow velocity terms. This level of sophistication was found to be neither necessary nor effective, and the loss can be represented in a much simpler manner:

$$P_f = c_f(\rho/2)(Q/A)^2 \tag{16}$$

In Equation 16, P_f is the loss attributed to conduit friction, A is a representative toroidal flow area, normally measured at the pump outlet, and Q/A is a representative toroidal velocity. The grouping $(\rho/2)(Q/A)^2$ is thus a kinetic energy term, and the coefficient c_f represents the fraction of the kinetic energy lost in the flow field. Studies have

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

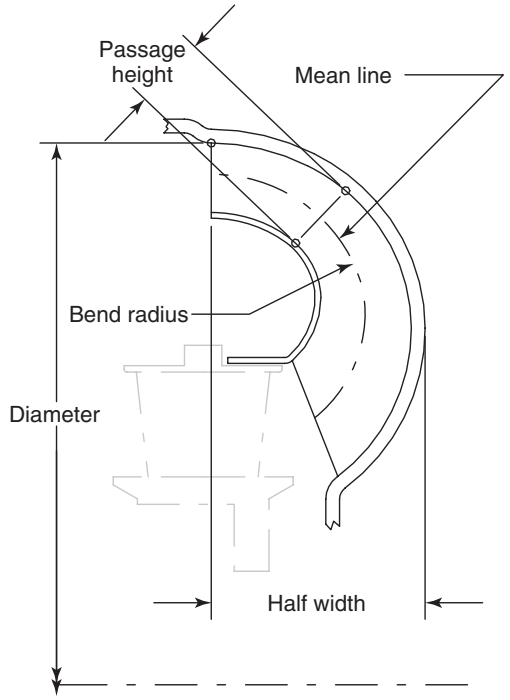


Figure 24. Torus design nomenclature.

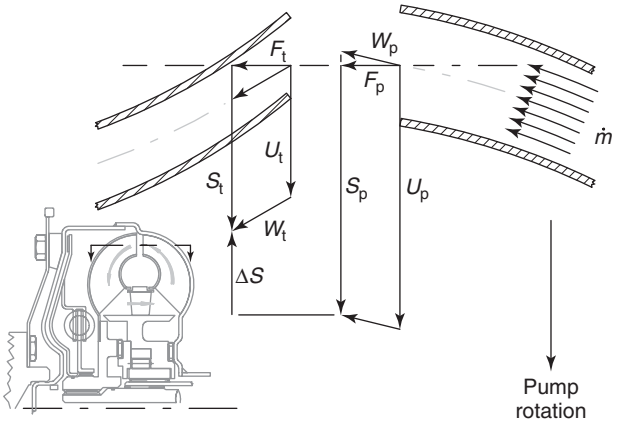


Figure 25. Pump outlet and turbine inlet at 0.5 speed ratio.

shown that c_f is not constant but varies greatly with speed ratio.

Friction loss is strongly affected by torus design features, illustrated in Figure 24. Probably, the most important influence on c_f is the ratio of mean line bend radius to passage height. High passages that are sharply curved do not flow well, and this is equivalent to a high value of c_f. When packaging requirements impress a short half width on the torus, sharply curved passages are unavoidable and performance suffers.

5.2.3 *Incidence loss*

Incidence loss attempts to capture the flow energy that is dissipated when the flow stream approaches an element other than parallel to the blade inlet surface. When this occurs, the blade system must immediately redirect the flow as it moves onboard, and energy is dissipated in this abrupt change in velocity. The condition is illustrated in Figure 25 for the turbine inlet at 0.5 speed ratio. In this illustration, the subscript t indicates velocities just inside the inlet and subscript p refers to velocities at the pump outlet. The rotational speed difference between the two elements is not totally accommodated by the high turbine inlet angle, resulting in a change in tangential velocity, ΔS, immediately onboard the turbine. The incidence loss is defined as the kinetic energy related to this change:

$$P_i = c_i(\rho/2)(\Delta S)^2 \tag{17}$$

The incidence loss coefficient c_i represents the percentage of the energy change that is dissipated as heat. Little can be done to reduce c_i on the pump and turbine, as sheet metal construction does not provide a means of contouring the blade's leading edge. Conversely, cast stator airfoil blades are usually shaped to reduce incidence losses. The common approach to reduce incidence loss on all three elements is to establish blade inlet angles that minimize ΔS.

As the converter is operated over a range of speed and torque conditions, the direction at which the flow stream approaches each of the elements will vary. When the incoming flow is observed from the element's rotating reference frame, the angle of approach relative to the meridonal plane is known as the *relative entrance angle*. The difference between the relative entrance angle and the blade angle is the incidence angle. When the incidence angle is zero, ΔS will be zero and there will be no incidence loss.

Relative entrance angles are established by upstream and onboard velocities at the inlet of each element. This is illustrated for the turbine inlet in Figure 26. Although rotating speeds and the mass flow vary with torque, velocity triangles at a constant speed ratio are geometrically similar. Thus, the relative inlet angles are single-valued functions of speed ratio and may be illustrated as in Figure 27. The data shown are for a particular converter, but the trends are typical of most automotive designs. All three angles demonstrate a noticeable response, with the stator seeing the largest range of angles, followed by the pump and then the turbine. The incidence loss model, Equation 17, suggests that setting the blade inlet angles to match the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

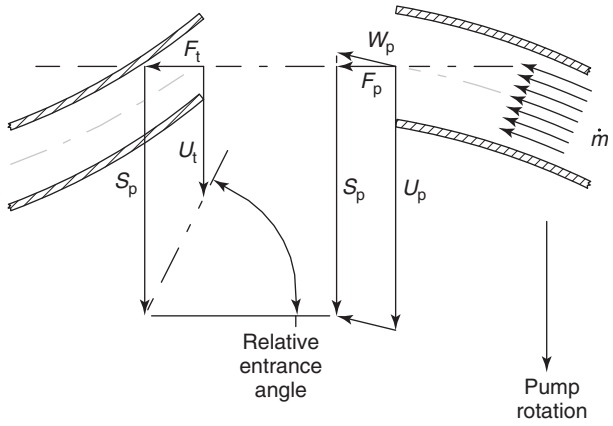


Figure 26. Turbine relative entrance angle at 0.5 speed ratio.

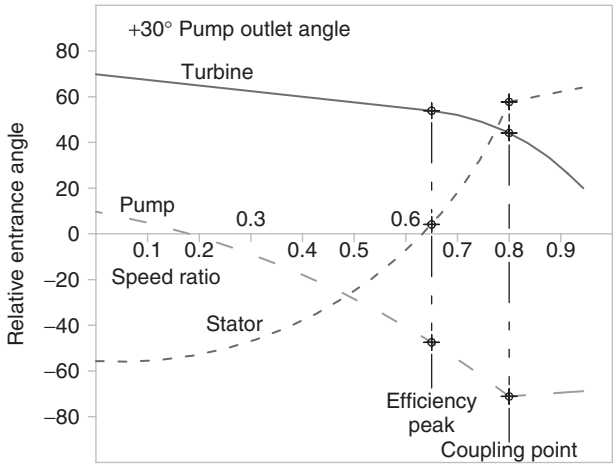


Figure 27. Relative entrance angles versus speed ratio.

low speed ratio-relative entrance angles would enhance the near stall performance, at some sacrifice at high speed ratio. Conversely, blade inlet angle settings at high speed ratio values of relative entrance angle would favor the peak efficiency and coupling range. Experience has proven this to be true, and the concept is used in practice to guide blade designs.

5.2.4 *Pump blade inlet angle*

Figure 27 shows the pump-relative entrance angle ranging from $+10^\circ$ at stall to -50° at the efficiency peak and continuing to -70° at the coupling point. Figure 28 illustrates typical performance compromises that result from a selection of pump blade inlet angles within this range. As expected, high values of inlet angle yield high stall torque ratio but suppress peak efficiency. Moving inlet angles in

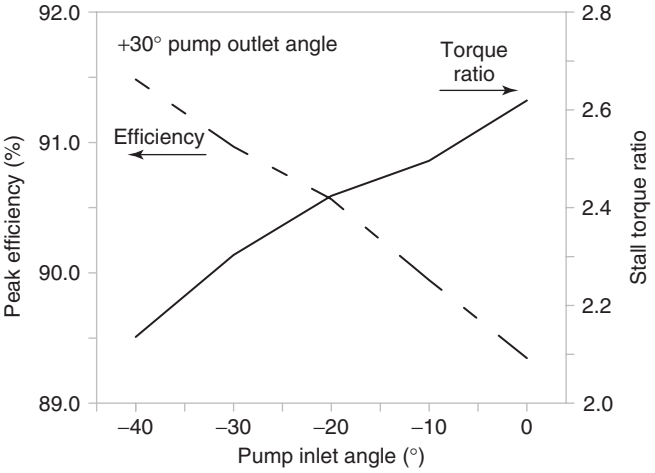


Figure 28. Stall torque ratio and peak efficiency versus pump inlet angle.

the negative direction result in monotonically decreasing torque ratio and increasing efficiency. The unit speed is shifted only slightly over this range of angles. Production pump inlet angles normally lie between -20° and -40° because of the impact of peak efficiency on fuel economy.

5.2.5 *Turbine blade inlet angle*

Figure 27 documents that turbine entrance angle is a very mild function of speed ratio between stall and the efficiency peak, with high values of turbine inlet angle favoring low speed ratios. Most converters respond to high values of turbine inlet angle by producing gains in stall torque ratio approaching 10% with little reduction of efficiency. Consequently, turbine inlet angles are usually set between $+55^\circ$ and $+60^\circ$.

5.2.6 *Stator blades*

The wide variation in stator entrance angle, as is evident from Figure 27, makes good performance at both low and high speed ratios difficult to obtain. Fortunately, stators are essentially nozzles, accelerating the flow field from inlet to outlet. High fluid velocity at the outlet is necessary to develop the optimum element torque and has the desirable propensity to reattach separated flows that are common at high incidence angles. In most converters, the stator torus area, established by the annulus between shell and core, is 10–20% smaller than the pump outlet area, further accelerating the flow. In addition, the airfoil shape provides a powerful mechanism for handling the difficult flow field.

As was noted earlier, the short axial length and small radii of the stator torus severely restrict the size of the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

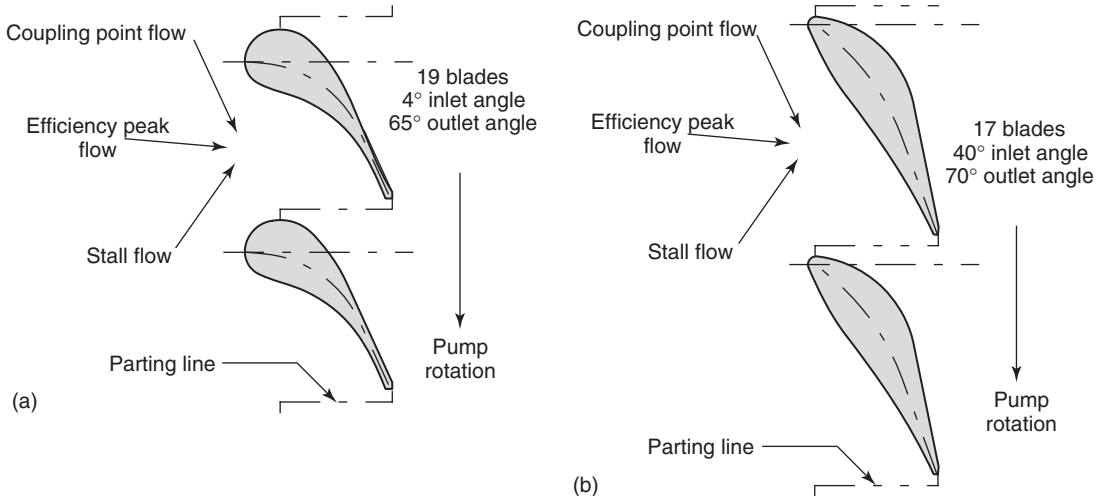


Figure 29. (a) Bullet nose stator blade. (b) High extension stator blade.

stator blade. Furthermore, molding dies require several millimeters of clearance between blades to allow for a parting line, forcing widely spaced blades. Significant velocity gradients are present across the outlet passage, and effective turning is measurably less than the mechanical blade angles. The short blade length causes the airfoil nose configuration and angle to have a strong influence on the outlet velocity distribution.

Two general stator blade forms are in broad use. The bullet nose blade shown in Figure 29a employs a large, bulbous airfoil to fair high negative angle, low speed ratio inlet flow into the stator. The blockage afforded by the airfoil thickness causes inlet flow to accelerate rapidly, reducing separation. The inlet angle is normally near 0°, producing nearly 60° of incidence at stall, but the fairing and early acceleration reduce c_i in Equation 17. The primary benefits of the bullet nose blade are very good stall torque ratio and low stall unit speed. Peak efficiency is only fair, and high speed ratio unit speed is mediocre.

Figure 29b illustrates a second common blade form. A number of different airfoils can be applied to this configuration, but all have a low thickness-to-length ratio and an angle difference between outlet and inlet of less than 50°. With a stall incidence over 90° and zero incidence occurring between 0.7 and 0.8 speed ratios, stall unit speed and torque ratio are definitely sacrificed for superior peak efficiency, coupling point extension, and high speed ratio unit speed. Being relatively thin, the blade makes no concession toward reducing c_i, but does not block high speed ratio flow. As was shown earlier, in Figure 22, this blade can be used effectively to very high outlet angles. This increases the coupling point unit speed, pushing the coupling point absorption line well to the right and increasing the coupling point speed ratio. Consequently, some organizations term this design the "high extension" stator blade.

In the production environment, stator tooling is much less expensive than pump tooling. Consequently, entire families of converters, with a wide range of performance characteristics, can be produced at reasonable cost by combining several stators of different designs with one or two pump assemblies. Figure 30 illustrates the range

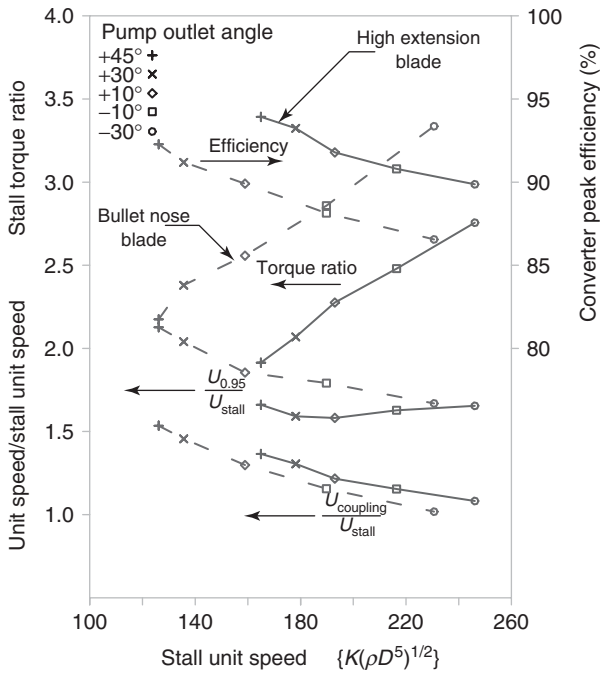


Figure 30. Converter performance response to pump outlet angle—bullet nose and high extension stators.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of performances generated by a typical bullet and high extension stator. Intermediate performance levels can be obtained by adjusting the thickness distribution and angle geometry.

5.3 Pressure vessel and thrust

When converter elements are rotating at speed and developing torque, large pressure gradients are established between and within the elements. These pressures, distributed over the projected areas of the elements, impose sizeable axial forces that must be accommodated. The pressure field is also distributed over the external walls of the converter, contributing to deflections during operation. There exists a synergism between element thrust and the converter deflection that influence the converter's mechanical design and its interface with the engine crankshaft.

5.3.1 Ballooning

The converter cover and pump housing, joined by the seam weld at the outside diameter, form a closed pressure vessel that must contain the internal pressure field and support centrifugal forces on its walls. The weldment's deflection under these combined loads is commonly termed *ballooning*. The pressure field has two components: the imposed static pressure, which is required to suppress cavitation, maintain clutch capacity, and establish cooling flow, and the intrinsic pressure caused by the motion and energy content of the working fluid. The pressure field is quite complicated when the blade systems are adding and extracting energy from the working fluid, but produces the maximum pressure vessel loading at very high speed ratios, when the elements are rotating at similar speeds. The pressure field is then essentially centrifugal, with an offset at the axis equal to the charge pressure. At any radius r, the local pressure P is given by:

$$P = \rho\omega_p^2 r^2/2 + P_C \tag{18}$$

In Equation 18, ρ is the fluid density, ω_p the pump angular velocity, and P_C the charge pressure, assumed to be imposed at the converter centerline. During vehicle operation, P_C is varied by the transmission control system in a typical range 350–900 kPa, and the centrifugal term at the periphery of a moderate size converter at high engine speed can exceed 3700 kPa.

When the parabolic function of Equation 18 is integrated over the relatively large areas of the converter pressure vessel and combined with the centrifugal force on the vessel walls, very significant stresses are produced. These stresses must be controlled to prevent excess elastic deflection and eliminate yielding. The converter is packaged tightly against nonrotating components where running contact would be ruinous, and internal splines and pilots must remain properly engaged. Yielding is unsatisfactory in any pressure vessel, as fatigue failures, with associated oil leakage, are likely if the steel is cycled into the plastic range.

Peak cover stresses are usually located in the corner between the drive lug and the seam weld (Figure 31). This is especially true on clutch-type converters where the sharp right-angle bend and subsequent clutch surface machining cause a relatively thin wall and stress concentration. Cover stresses are controlled by stock thickness and cover shape. Covers which have a relatively steep conical section extending from the inner diameter of the clutch surface to well inboard have been found to have the best ratios of peak stress to stock thickness. Unfortunately, the height of the cone tends to increase the length of the converter assembly. Flat covers, which provide compact packaging, do not possess good stress-to-thickness characteristics and occasionally demand high strength steel.

Pump housing stresses generally peak just outboard of the widest point of the torus of unbrazed pumps or just inside the inner edge of the torus on brazed pumps. Although the rated speed and pressure levels of the pump are always significantly improved by the gusseting effect of brazed blades, yielding, when it occurs, is much more severe. Pump housing stresses are otherwise controlled by stock-thickness selection.

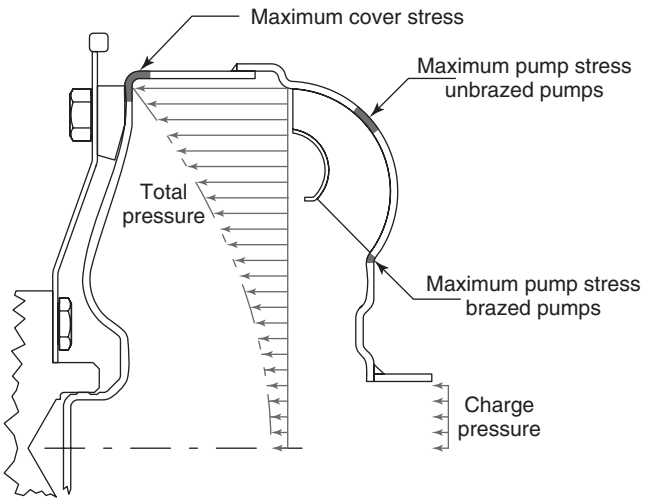


Figure 31. Pressure vessel loading and high stress areas.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5.3.2 *Element thrust*

Converter thrust is the sum of two separate axial forces exerted on the pump-cover weldment. One of these components results from the means used to exchange converter feed and exhaust oils between the converter and transmission. Routing oils through the converter hub subjects the hub area to the oil pressure. This produces a piston effect, pushing the converter toward the engine. The magnitude of this force is dependent only on the area of the hub and the converter charge pressure. It is termed *hub thrust* or *piston thrust*.

The second and more substantial thrust is produced by the integration of the converter's internal pressure field over the projected areas of each element. Element thrusts are substantial, as the projected areas are quite large and the pressure gradients across the elements can be significant. Normally, the turbine and stator thrust toward the converter pump, whereas the pump-cover weldment thrusts toward the engine. The force levels are strongly affected by element blade angles and torus configuration but, for a given design, are proportional to input torque and inversely proportional to converter diameter. As illustrated in nondimensional form on Figure 32, they are maximum at stall and decrease with speed ratio.

In the absence of other constraints, the pump-cover thrust will exactly balance the combined turbine and stator thrusts, with the loads transmitted through internal bearings between the elements. However, the torques being transmitted through the stator and turbine splines produce frictional lock between the elements and their shafts, preventing the stator and turbine from repositioning. The pump hub, however, has no such constraint. The pump's elastic ballooning characteristics allow it to deflect away from the stator, unloading the stator thrust bearing. Should this occur, the pump-cover thrust, equal to the sum of the stator and turbine thrusts, reacts against the engine crankshaft. Combined pump-cover and hub thrusts can easily exceed two metric tons, and this load is exerted against the engine crankshaft thrust bearing.

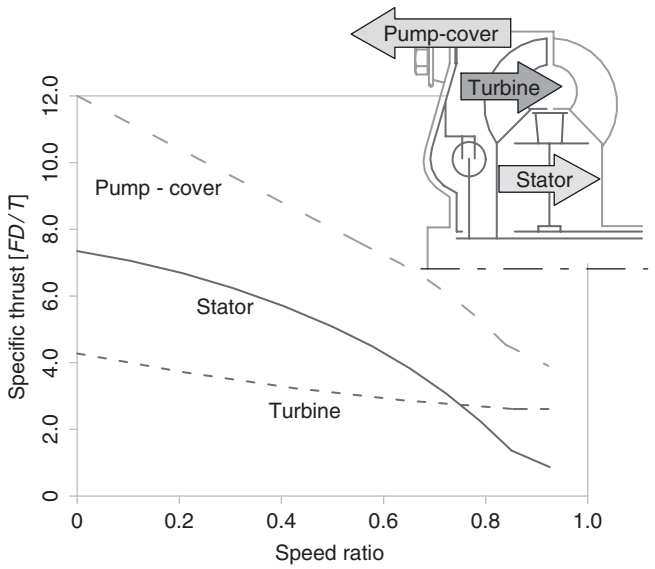


Figure 32. Converter element thrust versus speed ratio for a typical automotive torque converter.

As most engines are not capable of sustaining high crankshaft loads, the "floating pilot" mounting is used. In this arrangement, the cover pilot is allowed to slide freely in the crankshaft bore, whereas the converter position and drive are effected by a flexplate—a drive plate having a relatively soft axial compliance. During operation, pump-cover thrust deflects the flexplate toward the engine until internal contact is reestablished at the stator bearing. Crankshaft load is then determined by the flexplate rate, the displacement permitted by pump ballooning, and the hub thrust. Under even severe conditions, this is seldom more than a quarter of a metric ton. The remainder of the internal thrust is balanced at the thrust bearing.

6 CLUTCHING DEVICES

The torque converter clutch or lock-up clutch was first used in passenger car automatic transmissions in the 1949 Packard Ultramatic and a year later in the Studebaker Automatic Drive. As originally conceived, the clutch was intended to provide mechanical drive to the gear system input shaft during light throttle road load driving. Open converter efficiencies were generally in the 95% region at road load, so a roughly 5% fuel savings could be expected on steady-state highway driving. The designs were large, heavy, and expensive to produce, so with fuel economy not of paramount contemporary importance, the concept shortly fell into disuse.

The modern single-plate clutch configuration was introduced in the 1978 Chrysler TorqueFlite and was followed the next year by similar products from General Motors and Ford. These designs were intended to replicate the road load fuel economy improvements and to extend the gains to low speed, moderate throttle transient operation by implementing very aggressive apply schedules. This eliminated converter losses at lower vehicle speeds, and more importantly, lugged the engine into regions of optimum brake-specific fuel consumption (BSFC).

Mechanical drives have the propensity to pass torsion disturbances from engine-firing periodicity and its harmonics to the vehicle drivetrain and ultimately the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

passenger compartment. The resulting noise and vibration being unacceptable, lock-up clutches must be equipped with torsional isolators, usually a radial arrangement of compression springs, to attenuate the engine-driven excitation. This places a compliance between the engine and transmission inertias, and results in an added normal mode of vibration to the driveline. The natural frequency of this mode must be tuned well below the engine-firing frequency in the regions of optimum BSFC, and this engenders numerous sophisticated isolator designs to provide very high compliance. The added complexity demands significant space in the already cramped transmission bell housing, and this has resulted in torus designs with very narrow half widths. Obtaining creditable hydrodynamic performance from these flat torus converters has proven to be a challenge.

Another very different approach to vibration management is to control the clutch to very low slip speeds just short of full lock-up. The clutch shares torque with the hydrodynamic circuit, and engine torsional vibrations are well isolated from the drivetrain. In these systems, the challenges are to accurately regulate the slip speed and to produce a friction interface that does not self excite. A number of proprietary friction materials, facing textures, and groove patterns have been developed to produce acceptable dynamic friction characteristics.

ACKNOWLEDGMENTS

The author is indebted to Miss Jean Schweitzer, Technical Lead Engineer, GM Powertrain Advanced Torque Converter, for her assistance and support in preparing this article. Performance data and technical information for many of the illustrations are presented courtesy of GM Powertrain.

REFERENCES

Chayne, C.A. (1948) The buick dynaflow drive. SAE Report 89853.

J.M. Voith GmbH (1988) *Hydrodynamics in Power Transmission Engineering,* Heidenheim, p. 18.

Wislicenus, G.F. (1947) *Fluid Mechanics of Turbomachinery,* McGraw Hill, New York, pp. 375–377.

FURTHER READING

Gorskey, R.J. (1955) Buick's twin turbine dynaflow transmission. *SAE Transactions,* **63**, 42–52.

Gorskey, R.J. (1957) The new dynaflow transmission. *SAE Transactions,* **65**, 119–122.

Jandasek, V.J. (1962) Design of a single-stage, three-element torque converter in *Design Practices: Passenger Car Automatic Transmissions,* (4th edn, AE29), SAE International, Warrendale, 2-49–2-69.

Vincent, J.G. (1949) Packard automatic transmission. SAE Report 8985.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto091
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Dual Mass Flywheel

Jürgen Kroll and Ad Kooy
Schaeffler AG, Herzogenaurach, Germany

1 Introduction 1
2 Rotational Irregularity of Piston Engines 2
3 Basic Function and Design 3
4 Simulation and Design 8
5 Further Design Concepts 12
6 Evaluation and Outlook 16
Related Articles 18
References 18

1 INTRODUCTION

The past few years have seen ever tighter restrictions on the permitted CO_2 emissions of automotive combustion engines as a result of EU regulations, and yet further tightening of the rules is on the horizon. Furthermore, from 2014, this regulation will also include light commercial vehicles. In light of this, significantly reducing the fuel consumption of vehicles has become a core task in the automotive industry. Drivers also want more economical vehicles because the costs at the pump have increased continuously and look set to rise even further. As the main load for propulsion is borne by the combustion engine even in a hybrid drive system, the aim is still to optimize the combustion engine further.

The focus is on making it even more efficient. It therefore makes sense to raise the combustion pressure using turbocharging. This permits an increase in the effective power rating with the same displacement (Basshuysen und Schäfer, 2010). The specific torque also increases. As a result, frequently used operating points can be placed in the ranges of low specific consumption. Looking at it another way, this means that the same power rating can be achieved with a lower displacement (downsizing). This effect is also beneficial for consumers, as they can expect equally good, if not better, driving performance despite lower consumption.

As narrow limits are set for the reduction in cylinder capacity from a thermodynamic perspective, in practice, the displacement reduction is achieved by reducing the number of cylinders. We can actually observe that six-cylinder engines are being increasingly replaced by four-cylinder engines and four-cylinder engines by three-cylinder engines. Two-cylinder units are even being used in some cases. Greater specific power rating and greater specific torque are generally used to design the engine features and powertrain such that the operating point of the engine can be shifted to a lower speed at a given driving performance (downspeeding). Figure 1 shows an overview of the development. The left side shows crankshafts for six-, four-, and two-cylinder engines.

Owing to a higher mean pressure, for a turbocharged petrol engine, the ignition pressure currently reaches values up to 120 bar; for a turbo diesel engine, up to 180 bar. However, the increase in the specific torque is associated with a stronger rotational irregularity in the engine. In simple terms: the greater the specific torque, the greater the rotational irregularity. This is expressed in torsional vibrations that run counter to the increasingly demanding expectations of automotive customers for vehicle comfort.

Downspeeding reinforces this finding in as much as lower engine speeds promote vibrations in the powertrain and boom (Figure 2). Without countermeasures, drivers would avoid this operating range and thus counteract the positive consumption effects that are possible. The reduction in the idle speed means that gearboxes tend to rattle at idle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

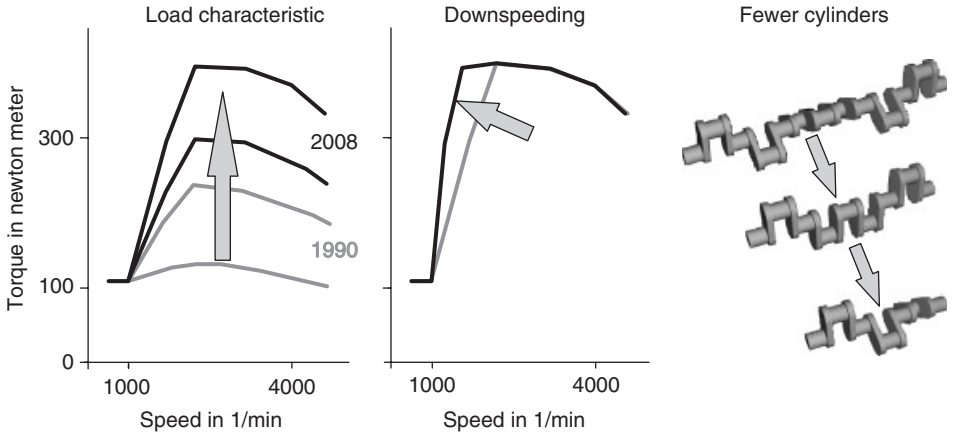


Figure 1. Development trends in engines.

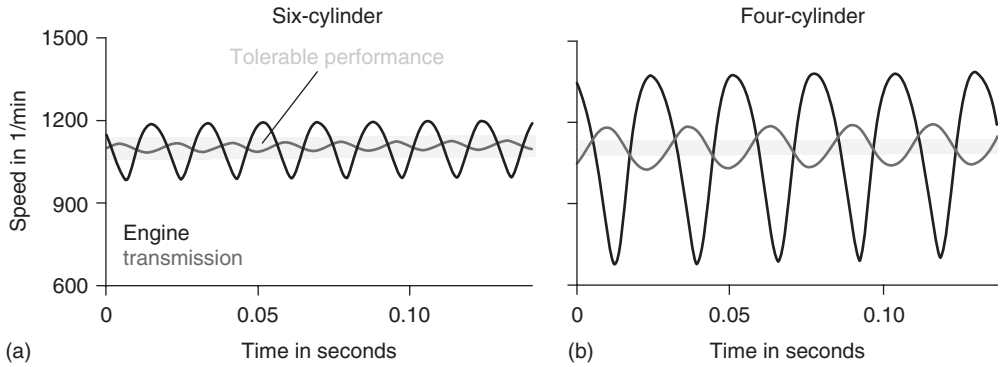


Figure 2. (a,b) Increasing demand for vibration isolation with downsizing and downspeeding.

speed. This loss in comfort has become more important over time because the lower viscosity gear oils used today, aluminum housing and reduced-friction toothed gear pairs only dampen the noise emitted to a limited extent.

To summarize: the possibilities of current engine technology can only be realized with an acceptable result for the customer using a powerful damping system. The conventional concept of a torsion damping clutch disk can no longer meet this requirement as the torsional vibrations that occur in modern engines can no longer be sufficiently absorbed. The following will show that the dual mass flywheel (DMF) fulfils this requirement.

2 ROTATIONAL IRREGULARITY OF PISTON ENGINES

2.1 Tangential force

Owing to the discontinuous way that the piston engine works and the oscillating mass forces, the tangential force that acts on the crank pin is irregular. Over the time of one working cycle, therefore, the actual angular velocity of the crankshaft is not constant (Küntscher and Hoffmann, 2006). The speed fluctuation is given by the degree of irregularity δ. The quieter the engine is to run, the smaller the value for δ must be. The following applies Equation 1:

$$\delta = \frac{\omega_{\max} - \omega_{\min}}{\omega_{\mathrm{m}}} \tag{1}$$

$\omega_{\max}$ indicates the greatest angular velocity of the crankshaft and $\omega_{\min}$ the lowest and ω_{m} the average. The average angular velocity ω_{m} results from Equation 2

$$\omega_{\mathrm{m}} = 2\pi n \tag{2}$$

In this equation, n represents the speed. Constant mean pressures result in approximately constant angular accelerations. As the irregularity is an integral result of this, it is largely dependent on the speed. The smaller the mass moment of inertia of the crank mechanism, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

greater the angular velocity difference $\omega_{max} - \omega_{min}$. As the speed increases, the degree of rotational irregularity quickly becomes very small, and as the speed decreases, the irregularity quickly becomes very high. In principle, this means that if a low degree of irregularity is to be achieved with a slow-running engine, a large flywheel must be used (Küntscher and Hoffmann, 2006). Further factors that influence the degree of irregularity are number of cylinders, pressure progression in the cylinder, geometry, and mass of the moving parts in the crankshaft drive and the combustion process.

2.2 Torsional vibrations and noise development

The powertrain executes partially rotating and partially oscillating movements as well as a swivel movement. The cyclical function of the piston engine accelerates and decelerates the moving power train parts, even when the engine is driven at constant load. This results in mass forces. They are differentiated by internal and external effects. The external effects consist of free forces and torques. They impose movements on the engine that transfer onto the supporting structure in the form of vibrations (Bosch, 2002).

The main cause of these structure-borne noise vibrations are the oscillating mass forces. The excitation as a result of the free mass forces and torques can only be influenced by mechanical balance (that is, the number and arrangement of the cylinders) or by balancer shaft (Lanchester balance).

The second source of structure-borne noise results from the irregular torque output. At lower speeds, it leads to booming or humming with ignition frequency and sometimes even to strong vibrations. This problem has a physical cause and cannot be influenced by the designer. The greater the specific torque of the engine, the more noticeable the humming and vibrating.

The engine also excites the vehicle as a whole, and in particular the gearbox. It swings against the vehicle, leading to rattling. The gearbox is primarily excited with the main excitation order. In a four-cylinder engine, this is the second order. In this context, "order" is the frequency with which an event occurs in relationship to the crankshaft speed (Basshuysen and Schäfer, 2010). A force of the first order changes its size with the crankshaft frequency; a force of the second order with double the crankshaft frequency.

Only with the development of a DMF was vibration isolation against the downstream powertrain achieved. This enabled the direct injection diesel engine for use in passenger cars to be developed on a large scale (Basshuysen and Schäfer, 2010).

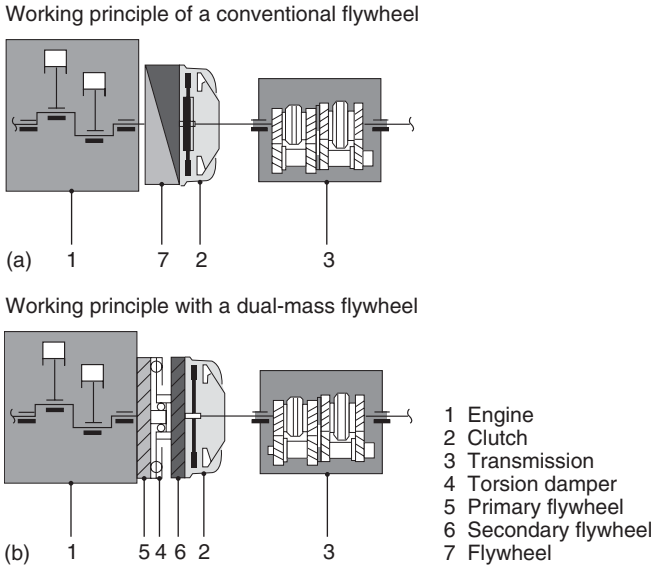


Figure 3. (a,b) Vibration isolation without and with DMF.

3 BASIC FUNCTION AND DESIGN

3.1 Function

The basic principle of the DMF is simple and efficient. With the additional mass on the transmission input shaft, the resonance point, which for the original torsional dampers lay between 1200 and 2400 rpm, is shifted to speeds that are considerably lower than idle speed. This means that the damper works postcritically, and excellent vibration isolation is present even from idle speed. In the usual design with a conventional flywheel and torsion damping clutch disk, the torsional vibrations in the low speed range are forwarded to the gears primarily unfiltered, as shown in Figure 3. This means that the tooth flanks of the gear wheels hit one another, which is perceived as idle rattle (Schaeffler Automotive Aftermarket GmbH, 2012).

In contrast, a DMF uses the spring damping system to filter out the torsional vibrations introduced by the engine and thus relieves the load on the gear components; as a result, the rattling stops. The effect is shown in Figure 4.

3.2 Structure

Figure 5 shows the structure of a standard DMF (Schaeffler Automotive Aftermarket GmbH, 2012). It consists of the primary flywheel (2) and the secondary flywheel (8). The two decoupled inertial masses are connected with one another via a spring damping system and twist-mounted against one another via a deep groove ball bearing or

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

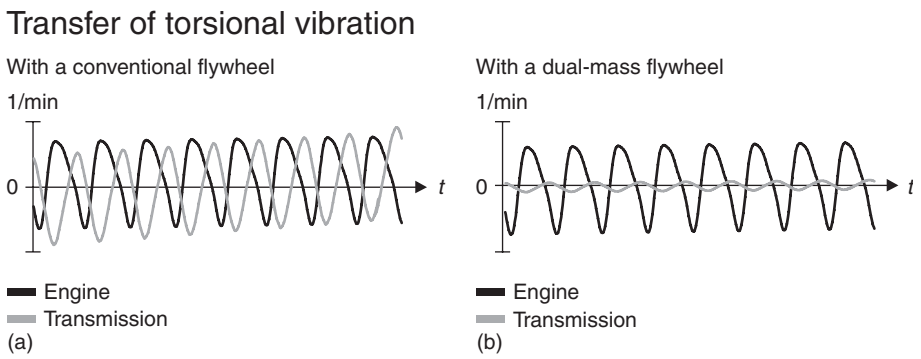


Figure 4. (a,b) Transfer of torsional vibrations from the engine to the gearbox without and with DMF.

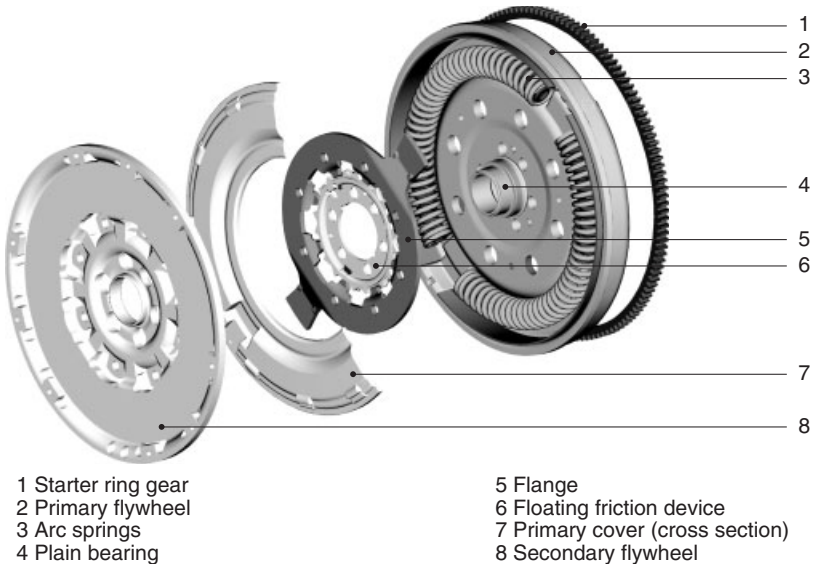


Figure 5. Structure of the DMF.

plain bearing (4). The primary flywheel with ring gear (1) assigned to the engine is bolted tight to the crankshaft. Together with the primary cover (5), it encloses a hollow space that forms the spring channel.

The spring damping system consists of the arc springs (3). They are positioned in guide shells in the spring channel and meet the requirements of an almost ideal torsional damper—at relatively low cost. The guide shells ensure a good guidance. Grease packing in the spring channel reduces the friction between arc spring and guide shell.

The engine torque is transferred from the arc springs via the flange (5). The flange is riveted to the secondary flywheel and, with its flange vanes, reaches between the arc springs. The secondary flywheel increases the mass moment of inertia on the gearbox side. It is fitted with ventilation slots for improved heat removal. As the spring damping system is in the DMF, a rigid design without torsional damper is usually used as a clutch disk.

3.3 Main components

3.3.1 Primary cover, primary flywheel, and arc spring stop

The primary flywheel (Figure 6) is connected to the engine crankshaft. Together with the crankshaft, it forms an inertial unit. Compared with a conventional flywheel, the primary flywheel of the DMF is significantly more flexible, which reduces the load on the crankshaft.

Together with the primary cover, it also forms the arc spring channel. Normally, this is in two parts and limited by the arc spring stops (Figure 7).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 6. Primary cover and primary flywheel.

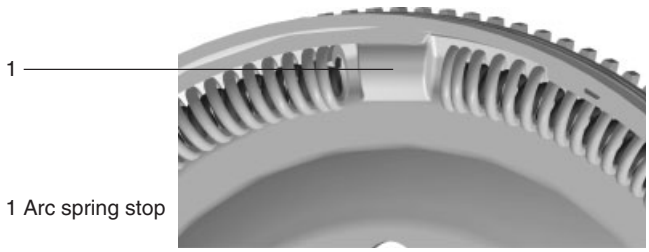


Figure 7. Arc spring stop.

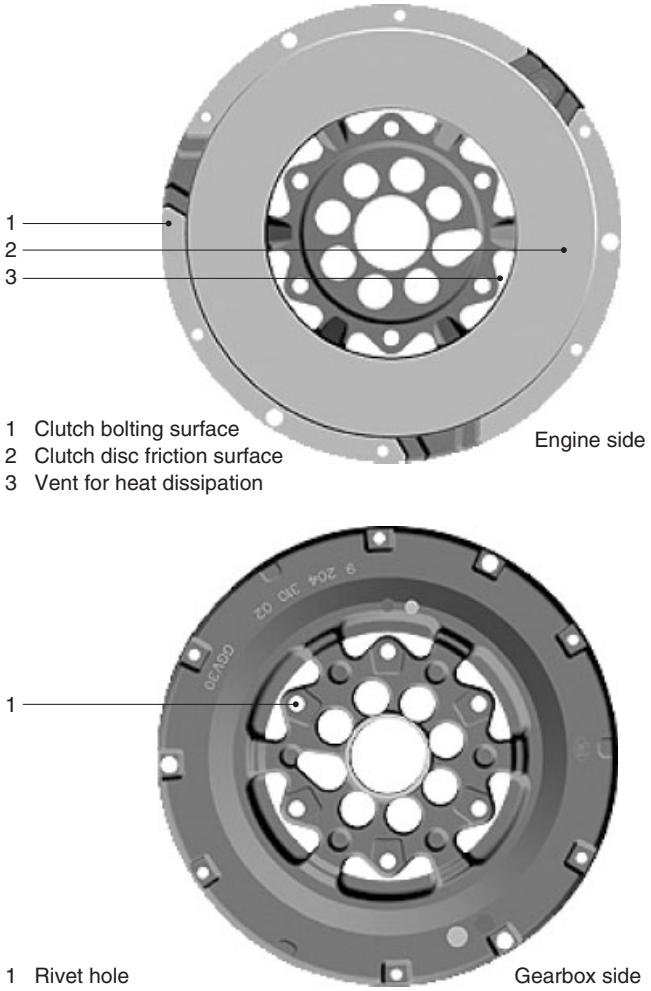


Figure 8. Structure of the secondary flywheel.

3.3.2 Secondary flywheel

The secondary flywheel connects the DMF to the powertrain on the gearbox side. Figure 8 shows the structure. In conjunction with the clutch, the flywheel transfers the modulated torque from the DMF. The cover of the clutch is bolted on at the outer edge. Inside the clutch, a spring mechanism presses the clutch plate against the friction surface of the secondary flywheel when the clutch is engaged. Friction allows torque transfer. The secondary side centrifugal mass consists principally of the secondary flywheel and the flange. The torque is taken up by the arc springs via the flange vanes (see also Section 3.3.4).

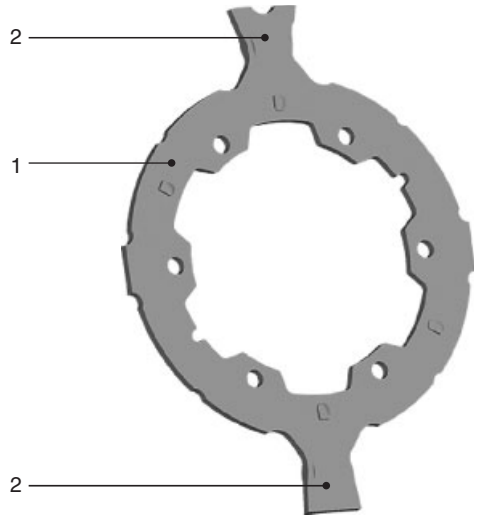
3.3.3 Flange

The flange (Figure 9) transfers the torque from the primary flywheel to the secondary flywheel via the arc springs and hence from the engine to the clutch. It is fixed rigidly to the secondary flywheel and is located with the flange vanes (arrows) in the arc spring channel of the primary flywheel. There is sufficient space between the arc spring stops of the arc spring channel and the arc springs for the primary wheel to move freely against the flange at idle speed.

3.3.3.1 Construction with fixed flange. In the construction with fixed flange (Figure 10), the flange is riveted to the secondary flywheel. The most simple form is the symmetrical flange, with both pressure and drive sides being the same. The admission point of the forces to the arc springs is thus at both the outer and inner areas of the end coil.

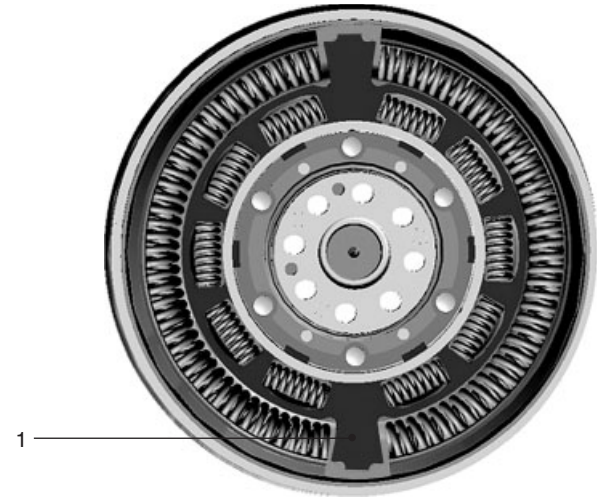
3.3.3.2 Flange with inner damper. On this model (Figure 11), the flange and side plates have spring windows on the inside, and springs sit in these windows. Because the trend for increasing engine torques is continuing, arc spring stiffness has to increase as long as the mounting space is the same. This leads to a deterioration in the vibration isolation. Friction-free inner dampers can counteract

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



1 Flange
2 Flange vanes

Figure 9. Flange.



1 Flange with spring aperture

Figure 11. Flange with spring windows for mounting springs in the inner dampers.

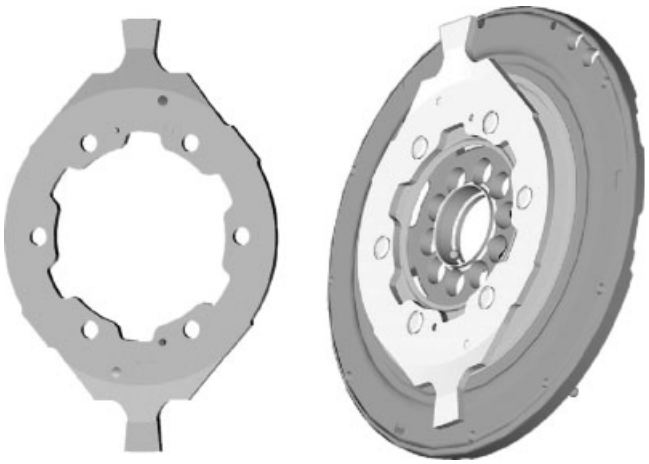
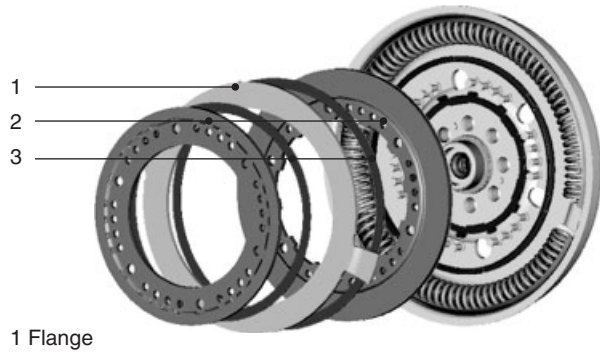


Figure 10. Fixed flange on the secondary flywheel.



1 Flange
2 Retaining panel
3 Facing

Figure 12. Flange with slipping clutch.

this because they improve the drive isolation. They also have a further advantage: at high speeds, the arc springs are pressed strongly outward against the guide shell due to the high centrifugal force. This "deactivates" the coils, meaning that the arc springs stiffen and part of the spring effect is lost. Friction-free inner dampers can maintain a good spring effect because the inward springs are installed straight in the flange. Owing to their low mass and their radially short positioning, these springs are subjected to a substantially lower centrifugal force. In addition, friction in the spring windows is further reduced by the convex curve of the upper edge. This means that as the speed increases there is no increase in the friction and the effective spring rate.

3.3.3.3 Flange with slipping clutch. In contrast to the fixed design, the third type of flange is not riveted to the secondary flywheel. In this case, the flange is shaped as a disk spring. The disk spring is positioned at the edge by two holding plates (Figure 12). The cross section of the retainer is therefore fork shaped. The engine torque is transferred directly because of the friction torque between the retainer and the disk spring. At the same time, the slipping clutch protects the DMF from overloading.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

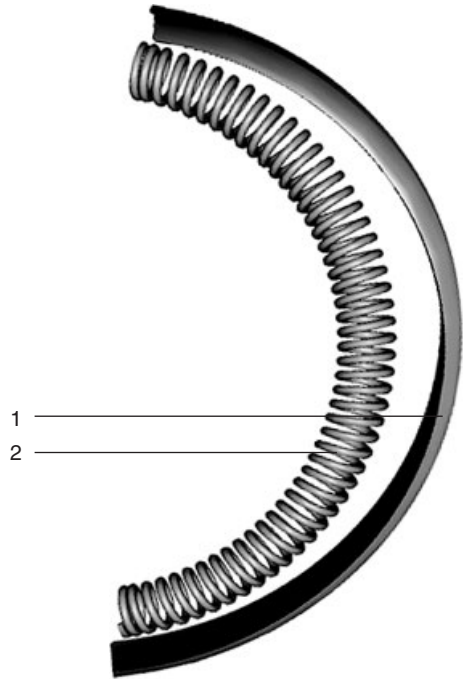


Figure 13. Arc spring with guide shell.

3.3.4 Arc spring

The arc spring is thus named because the coil spring with a large number of coils is installed in a semicircle (arc). This makes the best possible use of the mounting space available. The arc spring rests in the spring channel of the DMF and is supported by a guide shell (Figure 13). In operation, the coils of the arc spring rub along this guide shell generating friction, which is used for damping. The advantage is that the friction increases with the torsion angle. It is thus particularly high at the start and very low for the drive operating point. Stop damping can also be integrated. In order to reduce wear in the arc springs, the sliding surfaces of the arc springs are greased.

As there are countless variants of arc springs, a precisely aligned DMF system can be configured for every vehicle type and every load situation. In addition to single-stage springs, two-stage springs are also used—either as parallel springs (in various models) or as series springs. In practice, the individual types of spring are used in different combinations.

Single Springs. The most simple form of arc spring is the standard single spring (Figure 14).

Single-Stage Parallel Springs. Standard springs today are single-stage parallel springs (Figure 15). They consist of an outer and an inner spring of approximately the same length. Both springs are mounted parallel. The individual characteristic curves are combined to give a set characteristic curve.

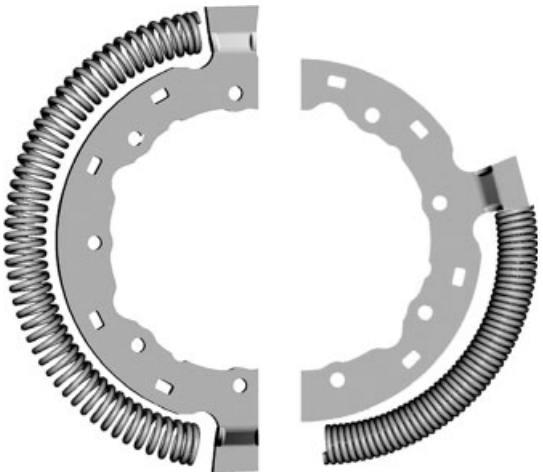


Figure 14. Flange with single spring, right compressed arc spring.

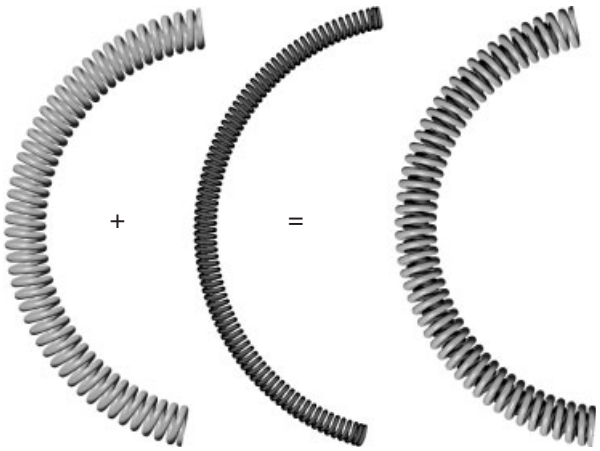


Figure 15. Composition of the single-stage parallel springs.

Two-Stage Parallel Springs. In the case of two-stage parallel springs, the two arc springs also lie one inside the other. The inner spring is shorter and is therefore activated later. The characteristic curve of the outer spring is adapted to the requirements for the engine start. As only the softer outer spring is loaded, the problematic resonance frequency range can be passed more quickly. At higher torque levels up to the maximum engine torque, the inner spring also comes into play. In the second stage, the outer and inner springs work together.

Three-Stage Arc Springs. These arc springs consist of an outer spring and two inner springs, of different strengths, mounted in series. Here, the two concepts of parallel springs and springs mounted in series are used together

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to achieve the optimum level of torsion balance at all engine torques.

4 SIMULATION AND DESIGN

4.1 Simulation

Ever decreasing development times for new car models and increasing cost-savings in the development phase mean that fewer and fewer test vehicles can be built. Simulation technology is therefore becoming increasingly important. The aim is to use simulation technology to optimize the products from the very beginning.

Firstly, all problematic operating points must be defined. The most important powertrain problems are summarized in the overview (Figure 16). These can generally be divided into three groups:

- acoustic problems (idle rattle and body boom);
- tangible problems (shaking);
- strength problems.

4.1.1 Efficient models

As the amount of computer time available is not infinite, efficient models with adequate levels of accuracy must be available for the simulations. To illustrate the vehicle for driving in drive, as described earlier, a model with nine rotating masses is used (Figure 17). Where possible, only the relevant natural modes and powertrain parameters are taken into consideration (Balashov *et al*., 2006). In contrast, all of the force-transmitting elements and excitation objects to be considered are modeled in great detail. The drive can operating state be described sufficiently accurately with this model.

Figure 18 outlines the analytical model, the comparison of measurement with simulation, as well as the target factors, in the same way as the above example for the start operating point, which can also be mapped well using simulation technology.

Operating point	Problem
Idle	Gear rattle
Drive	Gear rattle, boom
Coast	Gear rattle, boom
Engine stop	Gear rattle, clatter
Tip-in and back-out	Surging
Vehicle launch	Judder, rattle, surging
Engine start	Durability, comfort
Sub-idle speed resonance drive	Durability

Figure 16. Powertrain problems dependent on the operating point (from K06; 58).

The engine start is considered as a whole, in the same way as for drive operation. In addition to the DMF parameters, the engine, engine management, starter, and auxiliary equipment are considered. The target factors here are the start area between the actual and determined engine speed and the maximum torque in the DMF. While the evaluation of the start area represents a comfort factor, the target factor "maximum damper torque" is a strength factor. This procedure was developed correspondingly for all further operating points (Fidlin and Seebacher, 2006).

4.1.2 Whole vehicle consideration

When carrying out simulation calculations, the overall system observation is another essential feature. This looks at not only the DMF and the entire clutch system but also the entire powertrain, including engine and engine management. Considering the overall system also allows any problems in the interaction of the torsional vibration dampers with other components in the vehicle to be detected at a very early stage. The engine management, in particular, with idle-controller, smooth running regulator, load change dampers, etc., very often has complex interaction with the DMF.

Figure 19 shows the influence of engine management and starter speed on the engine start. The three diagrams show the engine and gearbox speed signals. First, we see a phase in which the engine is driven by the starter unfired. After a few cycles, the engine is fired up. The engine speed then runs through the DMF resonance, which often presents a problem for the calibrations.

A reduction in the starter speed by 30 rpm would have a considerable negative impact on the engine start (Figure 19, top right). The situation becomes critical if the starter is decoupled too early, for example, before the engine is fired for the first time. The result is unacceptable resonance vibrations that, in an extreme case, can lead to a resonance failure (Figure 19, bottom right).

Interactions between the complex transmission behavior of the DMF and the engine management often have a negative effect on driving comfort even in the idle speed operating state. To achieve perfect isolation with regard to ignition frequency, a degree of play is deliberately provided between the primary and secondary DMF mass. This successfully removes the troublesome idle rattle (Figure 20, left). However, the play and interaction of the set idle speed and the regulator parameters (PI-controller) can also excite low frequency vibrations and thus cause uncomfortable vibrations (Figure 20, right).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

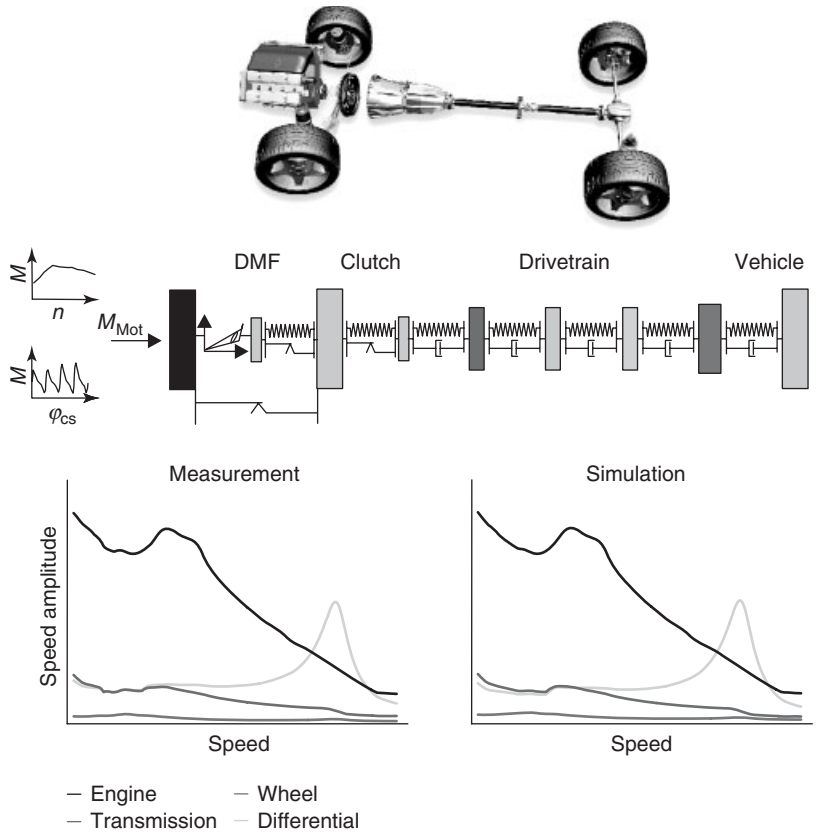


Figure 17. Analytical model, measurement, and simulation for full load in drive.

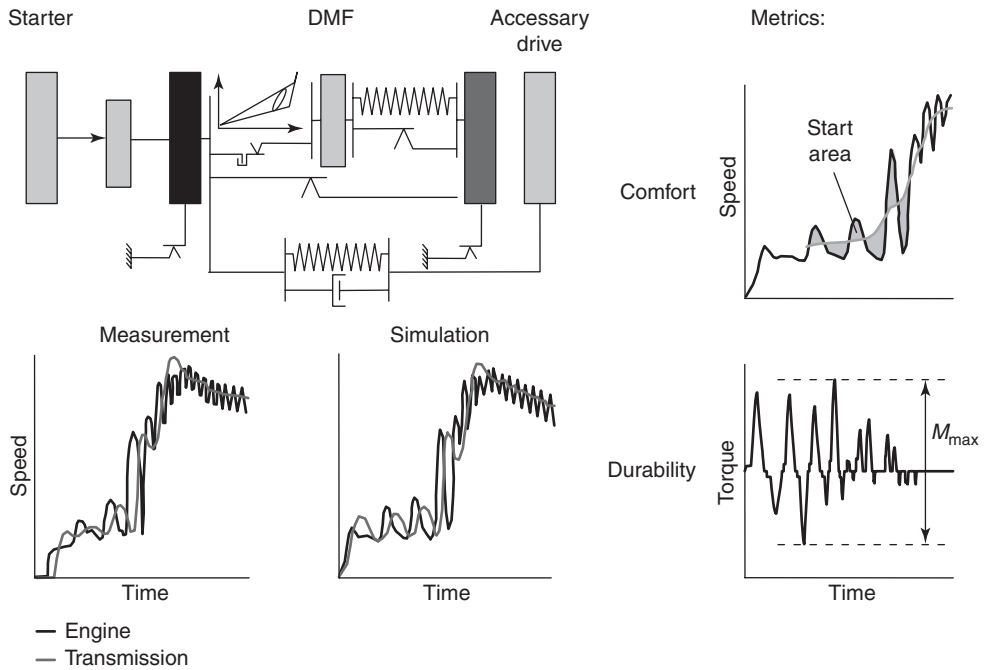


Figure 18. Analytical model, measurement, simulation, and target factors for the engine start operating point.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

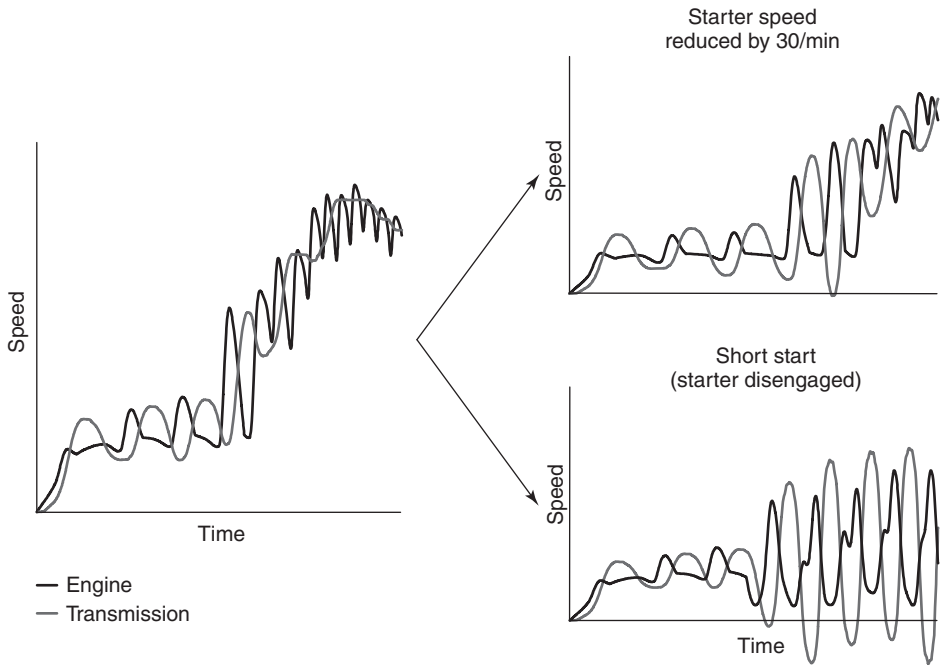


Figure 19. Interaction of the DMF with the starter and engine management in operating point engine start.

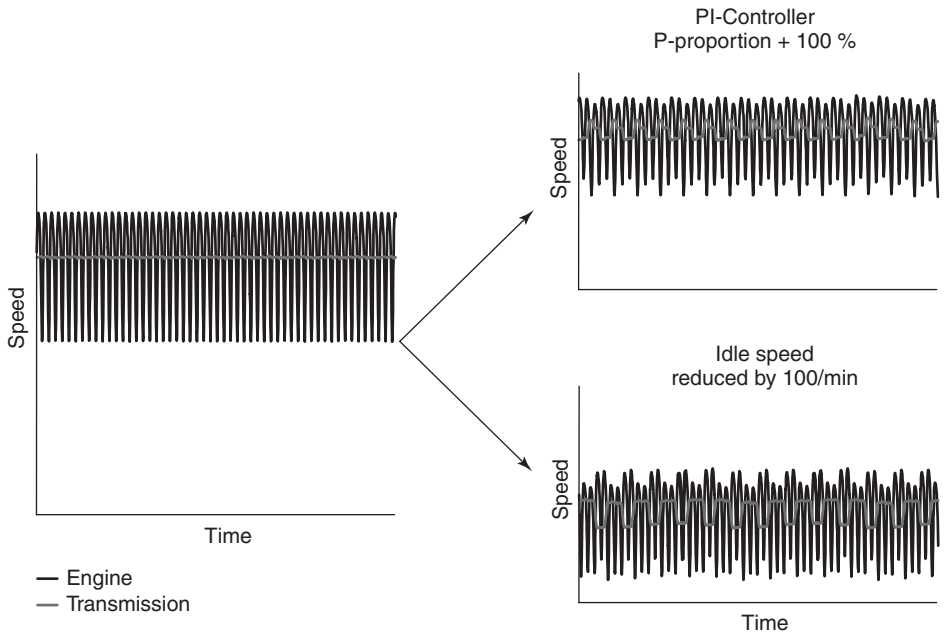


Figure 20. Interaction between DMF and engine management at idle speed.

4.2 Design

4.2.1 Target factors, subjective evaluation, and sensitivity diagrams

In order to objectively evaluate the quality of a design, suitable target factors must be found. The aim therefore is to define the measurable physical factors that must be evaluated or determined, and the extent to which these measuring factors correlate to the subjective feeling.

The determination of the target factors for driving at full vehicle load in drive operating state is shown in Figure 21. Possible problem points here are idle rattle and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

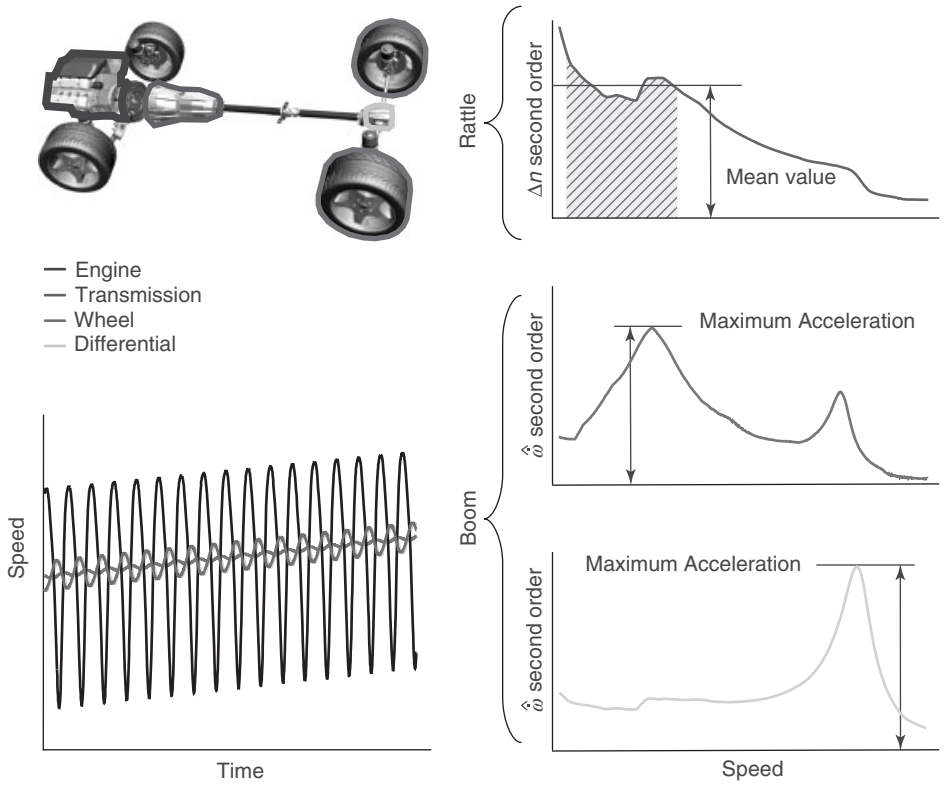


Figure 21. Evaluation of the target factors for rattle and boom at full load in drive.

body boom. To evaluate the quality of the driving in drive, the vehicle is accelerated with maximum torque. The speeds at transmission input, at final drive unit input (for rear-wheel drive vehicles), and at the wheel are measured in high resolution. Depending on the target factor, the signals are differentiated or integrated over time and their amplitudes are shown over the speed of the engine. To evaluate the idle rattle, the average amplitude of the fluctuations in the revolutions per minute of the transmission signal over the critical speed range is evaluated as the target factor (Fidlin and Seebacher, 2006). The body boom correlates accurately enough with the acceleration increases at the wheel and at the final drive unit input. At speeds under 1500 rpm, the maximum resonance magnification at the wheel is the target factor. A powertrain resonance can cause boom noises particularly in rear-wheel drive or four-wheel drive vehicles. The maximum acceleration magnification at the final drive unit input is evaluated as the target factor for this.

Ideally, these measurements are carried out with different DMF designs and various damper systems. This ensures that the largest possible range of measured target factors is covered. The installed systems are evaluated subjectively at the same time as the measurements are carried out. By applying the subjective evaluations over the appropriate target factors, sensitivity diagrams are created.

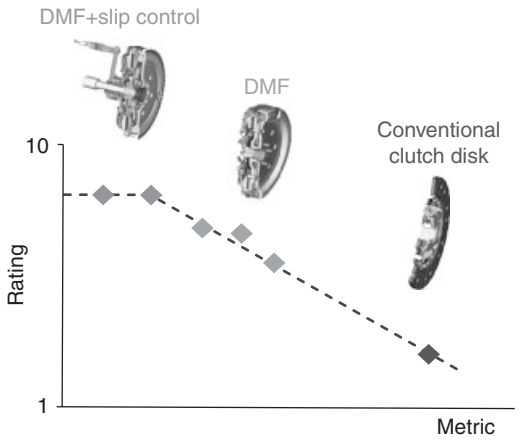


Figure 22. Sensitivity diagram using the example of the target factor humming at low engine speed.

Figure 22 shows the subjective evaluations applied over the target factors using the example of humming at low

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

engine speed. Rating 10 corresponds to driving without noise and rating 1 stands for unbearable body boom. For most customers, rating 8 is the target to be achieved, but at least rating 6 should be obtained.

4.2.2 Chip tuning

Chip tuning is a variant of engine tuning that requires no constructional changes to the engine. To increase performance, only the engine control unit parameters are changed. The modifications are generally aimed at the fuel mixture and the loading pressure of the turbo charger. Vehicle manufacturers occasionally use chip tuning to offer an engine with the same components in different power classes. In these cases, the tapping of the thermal and mechanical reserves is not a problem because these engines are also subject to production trials. There is a difference between this and retrofit chip tuning. The increase in performance and torque targeted here can exceed 30%. There is no check as to whether the powertrain can meet the increased demands in the long term.

Usually, the spring damping system of a DMF is designed specifically for the respective engine. If the maximum torque is in the range mentioned earlier or higher still than the value of the series engine, the safety reserve of the DMF is usually consumed or exceeded (Figure 23). As a result, in normal drive operation, the arc springs can be pressed together completely. This leads to a deterioration in the isolation (noises) or to vehicle shaking. As this happens with ignition frequency, very high load cycles occur that damage not only the DMF but also the gearbox, the drive shafts, and the differential (Schaeffler Automotive Aftermarket GmbH, 2012). The increased wear usually makes slow progress but can sometimes also lead to a sudden DMF failure.

5 FURTHER DESIGN CONCEPTS

5.1 Dual mass flywheel with centrifugal pendulum-type absorber

5.1.1 Basic function

The centrifugal pendulum-type absorber (CPA) is a persuasive option for improving DMF performance further. It is an absorber, that is, a secondary spring mass system outside the power flow. When this type of secondary spring mass system is excited with its resonance frequency, it moves in opposition to the exciting vibrations and thus ideally cancels them out. In a classic absorber connected via a spring (steel spring or elastomer), however, this effect only occurs at one frequency, namely the resonance frequency of the absorber. Two other resonance points arise—above and below the absorber resonance frequency—and have a very disruptive effect. A classic absorber is therefore unsuitable for the applications described here. The CPA works differently. Here, the restoring force of the absorber mass is determined primarily by the dominant centrifugal force and not by the negligible gravitational force (Kroll, Kooy and Seebacher, 2010). However, as this centrifugal force is speed-dependent (in contrast to the constant gravitational force), a speed-adaptive absorber is created, that is, an absorber whose natural frequency shifts as the speed changes. Thus, a fixed order of excitation can be absorbed but not a fixed frequency. This makes the CPA an ideal component for a piston engine as the pendulum is aligned

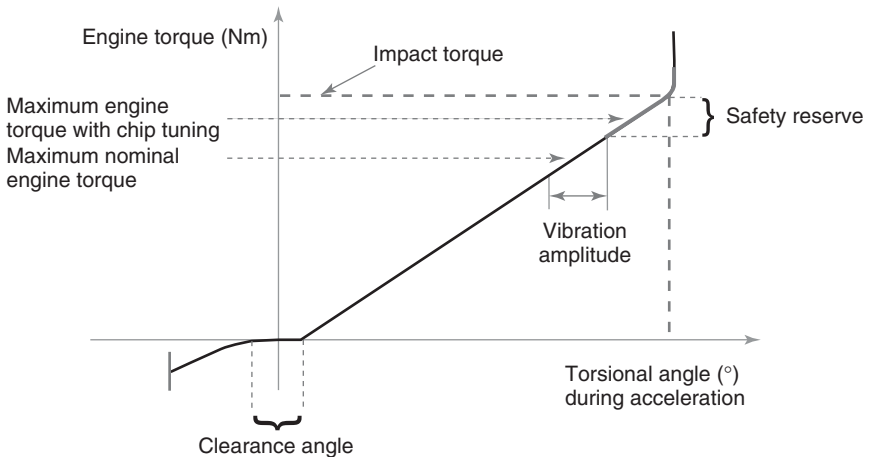


Figure 23. Arc spring characteristic curve, drive side (example).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

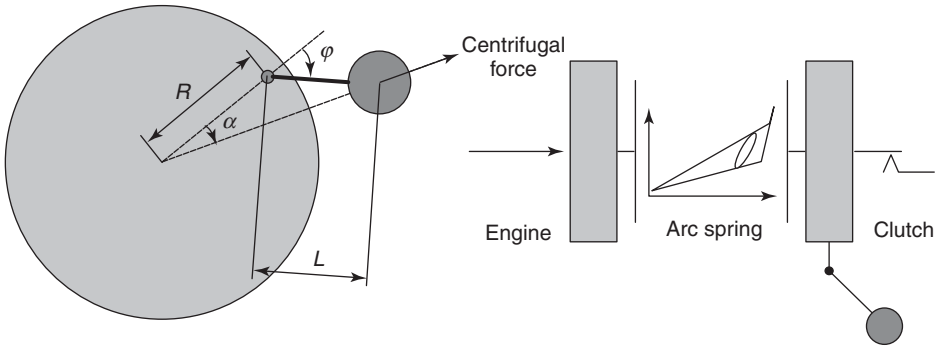


Figure 24. Functional arrangement of the centrifugal pendulum-type absorber.

with the main excitation order and, theoretically at least, can cancel this out. The relevant restoring torque M is calculated from the pendulum mass (m), the radial distance of the pendulum to the center of the secondary flywheel (R), the length of the pendulum (L), the speed ω^2, and the vibration angle. The following applies Equation 3:

$$M = -mRL\omega^2 \sin\varphi \tag{3}$$

Figure 24 illustrates the precise relationship.

5.1.2 Structure

In practice, however, vibrations in the dimensions of a main excitation order cannot be canceled out by the absorber being positioned on the crankshaft or fixed flywheel as the cyclic irregularities of the engine are too great and the vibration angle and size of the absorber masses too small because of the mounting space restrictions. Instead, the CPA is coupled with the secondary side of the DMF, that is, with the element for which the level of the torsional vibrations has already been reduced to 10–20% of the initial value on the engine. As now only the residual vibrations in the ignition frequency have to be compensated, significantly lower pendulum masses and vibration angles are necessary (Reik, Fidlin and Seebacher, 2009). Under these conditions, a CPA can be easily integrated into the mounting spaces usually available. Figure 25 shows the basic form of a DMF with CPA.

The pendulum is a bifilar pendulum, that is, it has two suspension points. The absorber masses are suspended over bolts that move in kidney-shaped tracks in the pendulum masses and in the DMF flange (Figure 26). The absorber arrangement is defined via the design of the tracks (Kooy *et al*., 2002). The masses do not rotate relative to the flange. All points of the pendulum describe the same track curve. A CPA arranged on the secondary side in this way produces a particularly effective vibration isolation with typical improvements in NVH (noise, vibration, and harshness) of two ratings. In experiments, a decoupling of up to 99% was achieved in combination with a DMF with inner dampers.

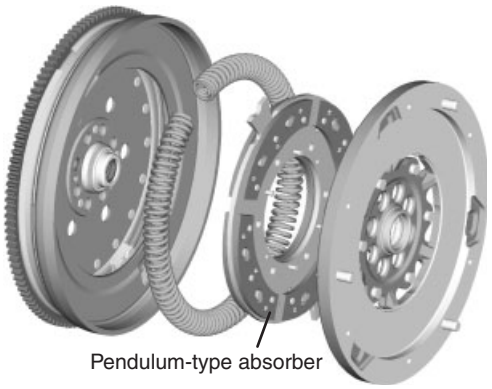


Figure 25. Design of a dual mass flywheel with centrifugal pendulum-type absorber on the secondary side.

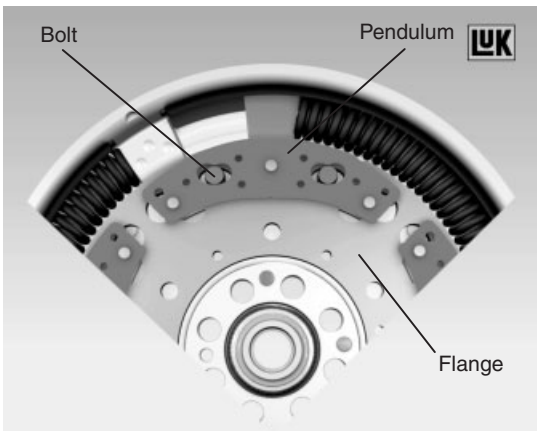


Figure 26. Bifilar centrifugal pendulum-type absorber in DMF.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

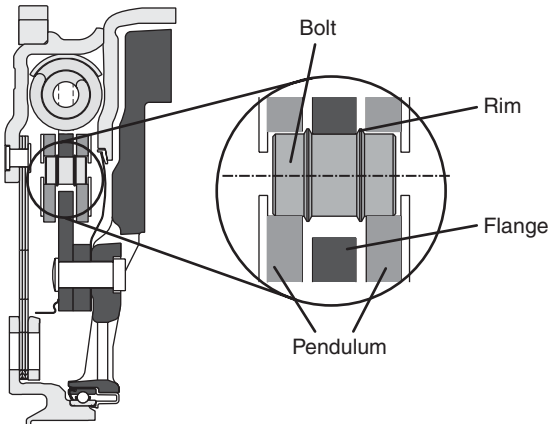


Figure 27. Friction-reduced pendulum positioning.

During the design and production, free vibration of the pendulum must always be ensured.

Friction in the CPA would cause a reduction in the buffered energy, the balance would be destroyed, and the secondary flywheel would start to swing again. Rims on the bolts enable the required, generally friction-free positioning of the pendulums (Figure 27).

The rims prevent contact between the pendulums and the flange and create only minimal friction as the relative speed in the contact area with both is minimal. The rolling friction coefficients must also be observed as, at speeds above 2000 rpm, they reduce the effect of the pendulums because of the high centrifugal force and thus significant rolling friction. Of course, not only the mass but also the radial position of the center of gravity is crucial for the energy absorption capacity.

Compared to the DMF with inner CPA (Figure 28), the axial arrangement is positioning the pendulum radially outward. By that tangentially more mass can be accommodated and the increased radial distance increases additionally the inertia of the pendulums quadratically. Especially, at very low speed, this concept is superior.

The DMF with an internal CPA combines the advantages of the long arc spring damper with a CPA in the same space. With an unchanged start performance, it achieves very good isolation over a wide speed range. In DMFs with external CPAs, the costs of the inner arc spring damper can be optimized using a smaller arc spring damper. The resulting shorter characteristic curves are more disadvantageous at the start but can be compensated in part by the increasingly common automatic starts resulting from the introduction of start/stop systems. The external pendulum means that the very good isolation can still be achieved effectively at low speeds. In a DMF with an external CPA, an inner damper can be integrated in addition to the normal arc spring damper. As a result of the basic isolation of the DMF, which is already good, the pendulums still work fully even at low speeds and thus enable isolation up to 99%.

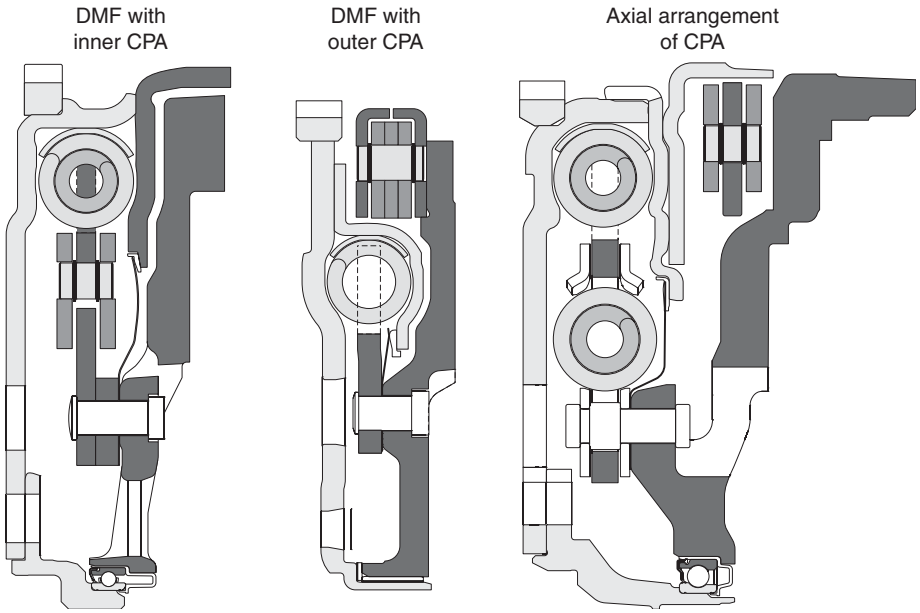


Figure 28. Possible arrangement of the centrifugal pendulum-type absorber in the DMF.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5.1.3 Design

The relatively complex interactions between the powertrain and the pendulum can be described in good quality with a simplified energy model that explains the main influencing factors (Figure 29). The simplified model assumes that the pendulum is positioned with no friction and that its natural frequency is precisely aligned with the respective main excitation order—on a four-cylinder engine this is the second order. Only the main excitation order is considered as excitation as higher orders generally have lower amplitudes and due to the higher frequency, also do not cause as much boom or rattling. For greater clarity, the rotating powertrain is shown as a linear vibration model.

The CPA becomes a linear absorber whose natural frequency automatically adapts to the excitation frequency. The absorber function means that its amplitude increases until the retroactive effect of the absorber has completely calmed the secondary flywheel. In this state, the energy passed into the secondary flywheel via the damper is completely buffered by the absorber.

The vibrational energy is therefore not destroyed but merely buffered, avoiding losses. What is interesting is that the ideal isolation is initially achieved without considering the inert mass of the secondary flywheel. The pendulum merely has to be capable of buffering the vibrational energy completely. Using simulations, we can actually show that a reduced mass moment of inertia of the secondary flywheel is not a direct disadvantage for the vibration isolation under the given condition.

The regulation inherent in the system also leads to an automatic adaptation to different excitations such as under partial or full load and represents a particular strength of the principle. If the excitation is low, the CPA responds with a smaller vibration angle and thus avoids overcompensation. The new optimal balance is set automatically with no additional regulation.

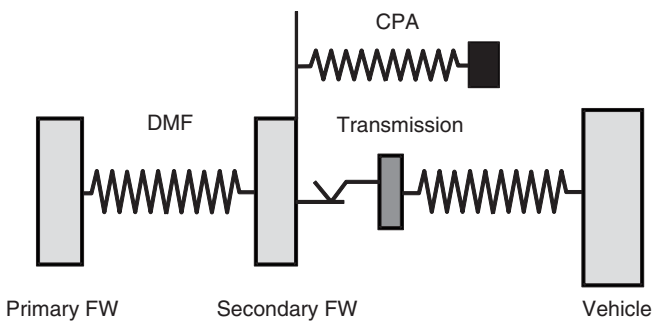


Figure 29. Linear vibration model.

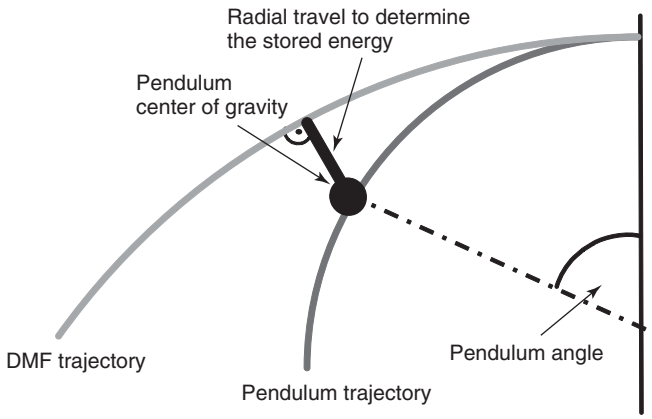


Figure 30. Radial migration of the center of gravity of the pendulum ("VDI_Buchtext," Figure 7). (Reproduced with permission from Kooy, Grahl and Gvozdev, 2011. © Ad Kooy.)

The mounting space and the centrifugal force loads restrict the maximum possible vibration angle of the pendulum considering the strength aspects. Together with the pendulum masses, the vibration angle defines the vibrational energy that can be stored. The vibrational energy is primarily buffered at the end of the track as a type of potential energy, whereby the centers of gravity of the pendulums have migrated inward radially (Figure 30). A larger vibration angle increases this radial path and thus increases the energy that can be stored more significantly than proportional increases in mass.

The maximum energy that can be stored in the pendulum is limited by the maximum vibration angle and increases quadratically with increasing speed. At low speeds close to idle speed, this results in a range in which the CPA cannot completely buffer the residual energy of the secondary flywheel (Zink and Hausner, 2009). The basic isolation of the damper is particularly important here. Arc spring dampers, with their very long springs, are particularly suitable. If the full engine torque is not yet present at speeds approaching idle speed, a two-stage system can achieve further improvements.

In addition to downspeeding, reducing the number of cylinders is a method of saving fuel that is already in widespread use. This leads not only to a significant increase in the degree of isolation demanded by the damper but also to a considerably elongated track geometry of the CPA because of the change in the excitation arrangement (Figure 31).

Identical energy storage capacity can be achieved with identical inertia via an identical radial path. As the number of cylinders reduces, the track width widens and becomes harder to align with existing mounting spaces. Current

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

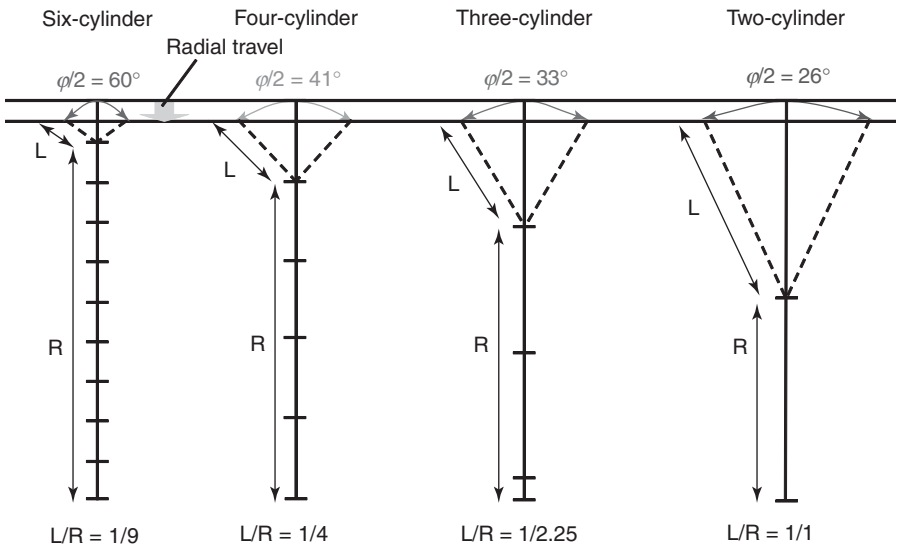


Figure 31. Influence of the number of cylinders on the track geometry of the pendulum with identical energy storage capacity.

systems with CPA also cover the requirements of two-cylinder engines.

5.2 Dual mass dampers for automatic gearboxes

DMF damper technology can be used not only for manual gearboxes but also for CVT (continuous variable transmission) and DCT (double clutch transmission). Here, the secondary mass in the gearbox is created by the conical disk set or the double clutch. In this case, power transfer is not achieved by frictional lockup between the secondary flywheel and the clutch disk, but with form-fit direct drive (spline) from the hub to the transmission input shaft (Figure 32).

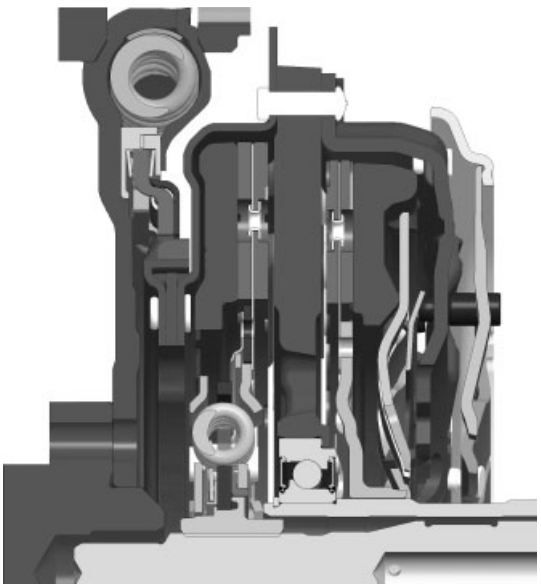


Figure 32. DMF for use with variable gearboxes or direct shift gearboxes.

6 EVALUATION AND OUTLOOK

6.1 Evaluation

The DMF can effectively isolate the downstream powertrain from the torsional vibrations caused by the engine. Comparable results cannot be achieved with a conventional system with clutch disk. The DMF with CPA demonstrates even further improved performance. In some cases, measurements in the vehicle showed an isolation of over 99% based on the rotational irregularity of the engine. This corresponds to an improvement of 2.5 ratings compared with a DMF without CPA. Figure 33 shows the potentials that the various models of this spring mass damper can develop.

This type of vibration isolation is a prerequisite for the successful implementation of downsizing and downspeeding concepts on the market to the extent that otherwise, the increased demands of vehicle drivers for noise comfort in their vehicles could not be met. In this case, however, a very advantageous (from a cost perspective) option for considerably reducing the fuel consumption in real driving operation would remain unused. The consumption simulation in NEDC (Figure 34) shows the potential savings; in each case, the prerequisite is the same excitation

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

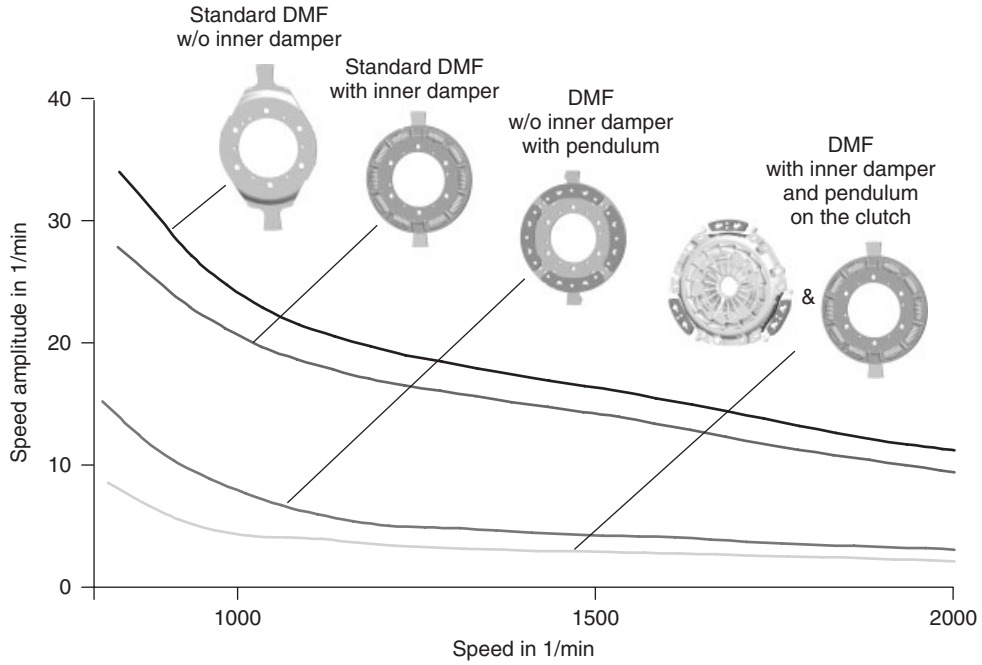


Figure 33. Performance of different DMF concepts.

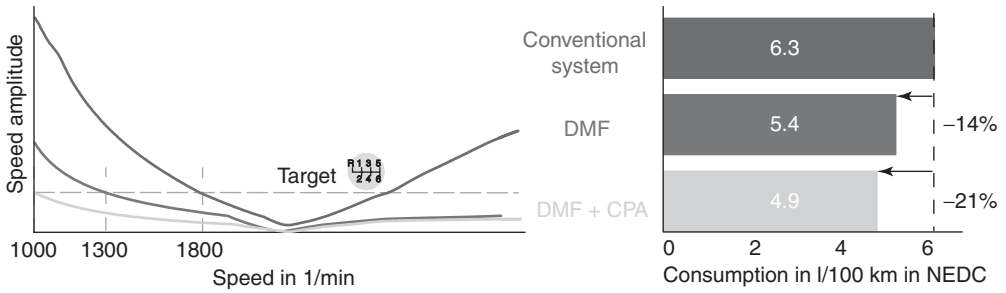


Figure 34. Fuel-saving potential using DMF and DMF with centrifugal pendulum-type absorber.

in the gearbox. Compared with a system with single flywheel, a DMF shifts the driving range by 500–1300 rpm and thus enables a fuel saving of 14%. With a DMF with CPA, a further 7% fuel saving can be achieved by reducing the engine speed by a further 300 rpm.

6.2 Outlook

The further development of the DMF must and will be targeted at meeting the trend for even higher ignition pressures. Current forecasts indicate that 150 bar will be achieved for petrol engines. For diesel engines, up to 240 bar can be expected in the coming years. One approach here would be to modify the arc spring damper. Every improvement in this component makes the work of the CPA easier. One such measure is to reduce the arc spring friction using a better lubricant. A further option is a reduction in the spring rate and the associated reduction in the spring weight by increasing the design stresses.

The performance of the CPA can also be increased further. New pendulum concepts are currently being evaluated. They again indicate a significant step forward in improving isolation. This can be achieved through both a larger pendulum mass and through an improvement in the arc spring damper. Both approaches also reduce the residual energy on the secondary side that cannot be buffered in the lower speed range. This reduces the tendency toward booming noises further.

A further challenge will be in designing the DMF concept for vehicles in particularly price-sensitive market segments.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

RELATED ARTICLES

Engine Management Systems
General Introduction - Basic Definitions, Structure of Part 4
Manual (MT) - Layouts, Design Considerations, Packaging & Manual Shift Mechanism, Manufacturing & Assembly
Dual Clutch Transmissions (DCT)—Layouts, Clutch Selection, Packaging, Actuation, Manufacturing & Assembly
Dry Clutch
Drive Train Noise, Vibration and Harshness
Control Systems and Strategies for Automated Manual and Double Clutch Transmissions
Fundamental Combustion Modes
Zero- and One-Dimensional Methodologies and Tools
Fuels for Engines and the Impact of Fuel Composition on Engine Performance
Torsional Analysis
NVH Considerations in Engine Development
Operating Principles
Engine Performance

REFERENCES

Balashov, D., Burkovski, L., Ferderer, F., *et al*. (2006) Simulation bei Drehschwingungsdämpfern [Simulation for torsional vibration dampers], *ATZ 12/2006 Volume 108*.

Basshuysen, R. and Schäfer, F. (2010) *Handbuch Verbrennungsmotor: Grundlagen, Komponenten, Systeme, Perspektiven, [Combustion Engine Manual: Basics, Components, Systems, Perspectives]*, vol. 49–83, 5th edn, Vieweg+Teubner, Wiesbaden, pp. 991–1034.

Bosch (2002) *Kraftfahrzeugtechnisches Taschenbuch [Vehicle technology pocketbook]*, 24th edn, Friedr. Vieweg & Sohn, Braunschweig/Wiesbaden, p. 436.

Fidlin, A. and Seebacher, R. (2006) *DMF Simulation Techniques*, *8th LuK Symposium*, Baden-Baden.

Kooy, A., Gillmann, A., Jäckel, J., and Bosse, M. (2002) *DMF–Nothing new?* 7. LuK Symposium, Baden-Baden.

Kooy, A., Grahl, U., Gvozdev, M. (2011) Prinzipielle Betrachtungen und Optimierungen zum Fliehkraftpendel, [Basic Aspects and Optimization of Centrifugal Pendulum-type Absorbers], *VDI-Fachtagung Kupplungen und Kupplungssysteme in Antrieben*, Wiesloch bei Heidelberg.

Kroll, J., Kooy, A., and Seebacher, R. (2010) Land in Sight? Torsional Vibration Damping for Future Engines, *9th Schaeffler Symposium*, Baden-Baden.

Küntscher, V. and Hoffmann, W. (2006) *Kraftfahrzeug-Motoren: Auslegung und Konstruktion, [Vehicle engines: Design and construction]*, 4th edn, Vogel Buchverlag, Würzburg, pp. 871–1041.

Reik, W., Fidlin, A., and Seebacher, R. (2009) Good Vibrations—Bad Vibrations, *VDI Conference: Vibrations in Drives*, Leonberg.

Schaeffler Automotive Aftermarket GmbH (2012) Zweimassenschwungrad, Technik Schadensdiagnose, Spezialwerkzeug/Bedienungsanleitung [Dual mass flywheel, technology, damage diagnostics, special tools/operating instructions], Langen.

Zink, M., and Hausner, M. (2009) The centrifugal pendulum-type absorber – application, performance and limits of speed-adaptive dampers, *ATZ, Issue 07/08 2009*.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto092
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Articulated Joints and Couplings—Cardan and CV Joints

Wolfgang Hildebrandt
GKN Driveline, Lohmar, Germany

1 Basic Function of Joints and Driveshafts 1
2 Cardan Joint (Also: Universal Joint/Hooke's Joint) 1
3 Constant Velocity (CV) Joints 3
4 Sideshaft Configurations 12
Related Articles 15
References 15
Further Reading 15

1 BASIC FUNCTION OF JOINTS AND DRIVESHAFTS

Articulated joints are coupling mechanisms for transmitting torque or rotary motion through an angle between an input shaft and an output shaft.

It can be differentiated from

- fixed joints, which only allow an angular displacement during rotation and
- plunging joints, which additionally also allow an axial displacement.

Driveshafts consist of a number of joints connected in series; a parallel displacement between input and output shaft is also possible (Figure 1).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 CARDAN JOINT (ALSO: UNIVERSAL JOINT/HOOKE'S JOINT)

One of the oldest mechanical concepts to transmit torque under rotation is the so-called cardan joint, named after Geronimo Cardano (he used the principle of ring suspension back in the sixteenth century). It consists of two forks, which are connected to input and output side, and a yoke, which is guided in these forks, as shown in Figure 2.

Modern automotive applications include needle bearings between yoke and forks to allow a rolling sliding between yoke ends and fork; the needles are greased/oiled and sealed to protect the configuration against dust, dirt, and water; see Figure 3 as an application example.

2.1 Nonuniform transmission and reaction moments

When transferring a rotation from the first (input) shaft end to the second (output) shaft end of a cardan joint under an articulation angle β, the angular speed ω and the torque T between input side ($_{\text{index}}$ 1) and output side ($_{\text{index}}$ 2) display uneven, periodic rotary oscillations. The nonuniformity of the angular speed as function of the rotation angle α is described in formula 1.

$$\omega_2 = \omega_1 \times \cos(\beta)/[1 - \sin^2(\beta) \times \sin^2(\alpha)] \qquad (1)$$

When transmitting a torque, the so-called secondary torque T_S is generated, which is positioned in the articulation plane and perpendicular to the shaft.

Figure 4 shows in principle the equilibrium of torques at the joint in two different rotation positions.

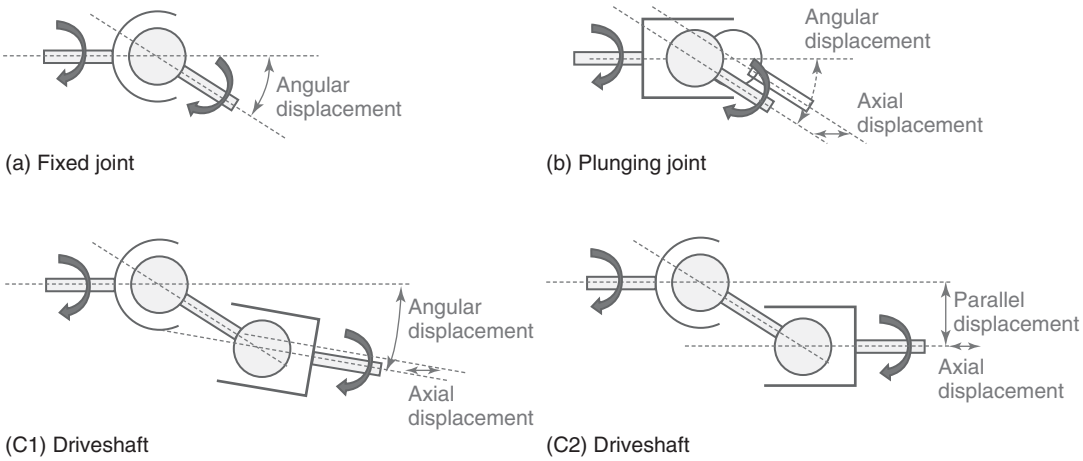


Figure 1. Joints and driveshafts: displacements.



Figure 2. Cardan joint concept.

For further details on kinematics, see, for example, VDI (2003).

2.2 Cardan joint assemblies with constant velocity behavior

The nonconstant velocity behavior of a single joint can be compensated by connecting two units in series, either as a W-bend or as a Z-bend configuration (Figure 5).

For these assembly configurations, it is essential that the connected forks of the joints are positioned in the same plane and furthermore the articulation angles of both individual joints are equal. Only under these conditions, the nonuniformity of the joints can be compensated one against the other.

Figure 6 shows as an application example a front axle configuration of a construction machine. By an external centering of the articulation plane in the axle housing, a constant velocity behavior can be achieved for specific articulation angles.

A constant velocity behavior of two cardan joints in series can also be achieved by *internal* centering. Figure 7 shows as an example a so-called *centered double cardan joint*, in which input and output fork are connected by a ball joint type centering device. This ensures that—for a given angle—for both cardan joints, the individual articulation angles are equal according to a "W-bend" configuration.



Figure 3. Cardan joint application example: propshaft for light commercial vehicle.

2.3 Automotive applications for cardan joints and its derivatives

Automotive applications for cardan joints are limited to those applications, where the impact of nonconstant-velocity behavior can be neglected or where the installation angle is small.

Cardan joints can be found, for example,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

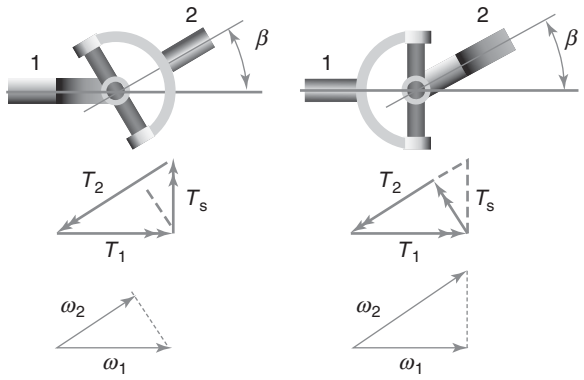


Figure 4. Torques T and angular speed ω for a cardan joint under articulation angle ß.

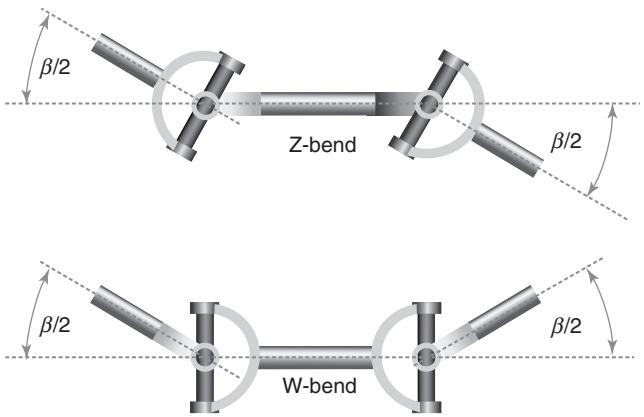


Figure 5. Configurations of cardan joints to achieve constant velocity behavior (Z-bend/W-bend).

- in the steering column of all types of automotive applications
- in propshafts, if installation angles are small
- in all-wheel-driven vehicles with hang-on front drive, where the front axle is designed as live axle.

3 CONSTANT VELOCITY (CV) JOINTS

For constant velocity (CV) joints, the angular speed on input side (ω_1) and output side (ω_2) are equal. Under the assumption, that the power transfer is loss-free, also the input torque T_1 and output torque T_2 must be equal. The necessary torque equilibrium for the joint is fulfilled, when a so-called "half-angle plane" or "homokinetic plane" is introduced, see Figure 8.

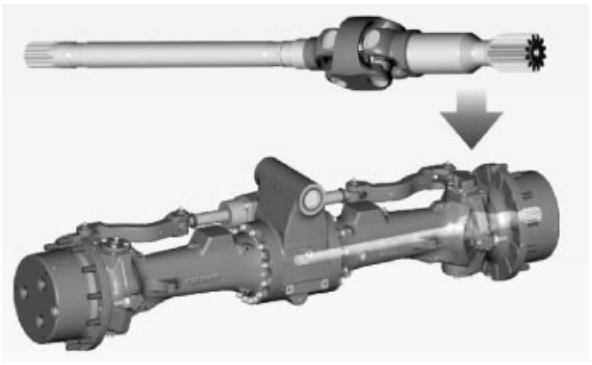


Figure 6. Application of two cardan joints in the axle housing of a construction machine.

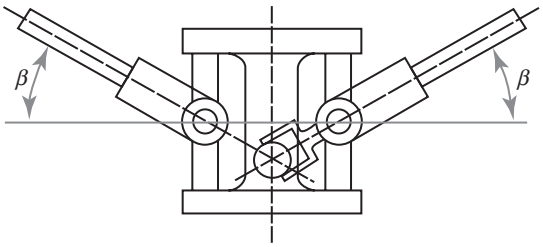


Figure 7. Principle of a centered double cardan joint.

Secondary torques T_{S1} and T_{S2} are—contrary to cardan joints—constant and independent from the rotation position.

$$T_{S1,S2} = T_{1,2} \times \tan(\beta/2)$$

3.1 Constant velocity ball joints

In constant velocity ball joints, balls are used as coupling elements between input side and output side.

The balls are guided in longitudinally elongated ball tracks. CV ball joints can be grouped into two base concepts: *radial separation* or *tangential separation* of input and output tracks (Figure 9).

All relevant developments date from first publications in the first decades of twentieth century:

The *radial separation of ball tracks* was originally proposed by Weiss (1929) (Figure 10), then later optimized as plunge ball joint by Altmann, Enke, and Rothweiler (1964). The balls are loaded perpendicular to the joint center (radius r_1), but—as disadvantage—only half of the balls are used for torque transfer.

The *tangential separation of ball tracks* was proposed by W. Withney in 1908, later significantly further developed by Rzeppa (1934), compare also Figure 11. While the effective

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

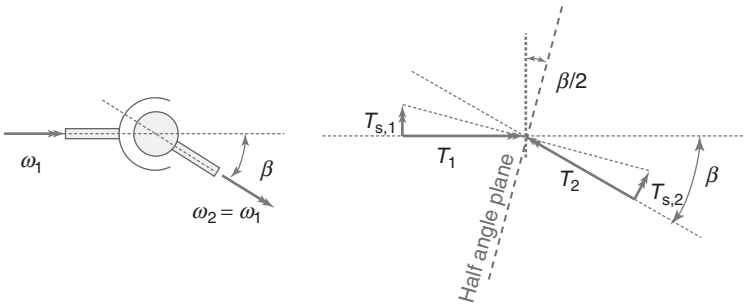


Figure 8. Constant velocity joint and torque equilibrium in the half-angle plane.

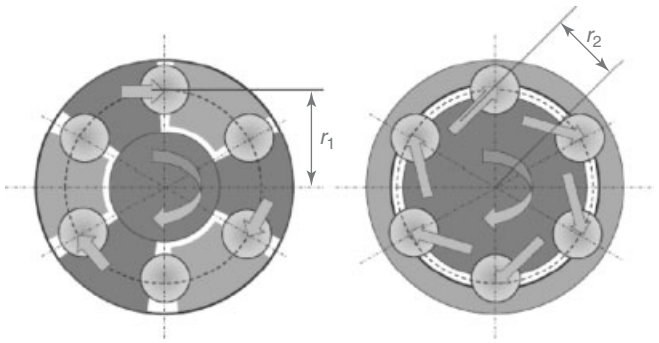


Figure 9. Base concept of CV ball joints; left: radial separation, right: tangential separation.

working radius r_2 for the balls is smaller compared to Weiss' principle, all balls are used for torque transfer.

Today, most of the current CV joints are based on Alfred Rzeppa's ideas and, therefore, will be in the focus of the following considerations.

3.1.1 Ball fixed joints

Ball fixed joints in the design as proposed by A. H. Rzeppa are the most commonly used constant velocity joints in automotive vehicle applications. They are typically positioned on the wheel side of a driveshaft for a front-driven vehicle with a maximum articulation angle of $\sim 47^\circ$.

The principle of a Rzeppa type concept is shown in Figure 12.

The input torque is transferred via the shaft (1) through a spline connection (1b) to an inner race (2), in which balls (3) are positioned in longitudinally extended ball tracks. These balls are also guided in the outer race tracks (4), which by its splined stem (4b) is connected to the wheel hub. The balls (typical number of balls: 6, 8, or even more) are guided in a cage (5), which, on its spherical inner and outer diameter, is fixed between the inner race and the outer race. The whole configuration is greased and sealed against environmental influences by a boot (6).

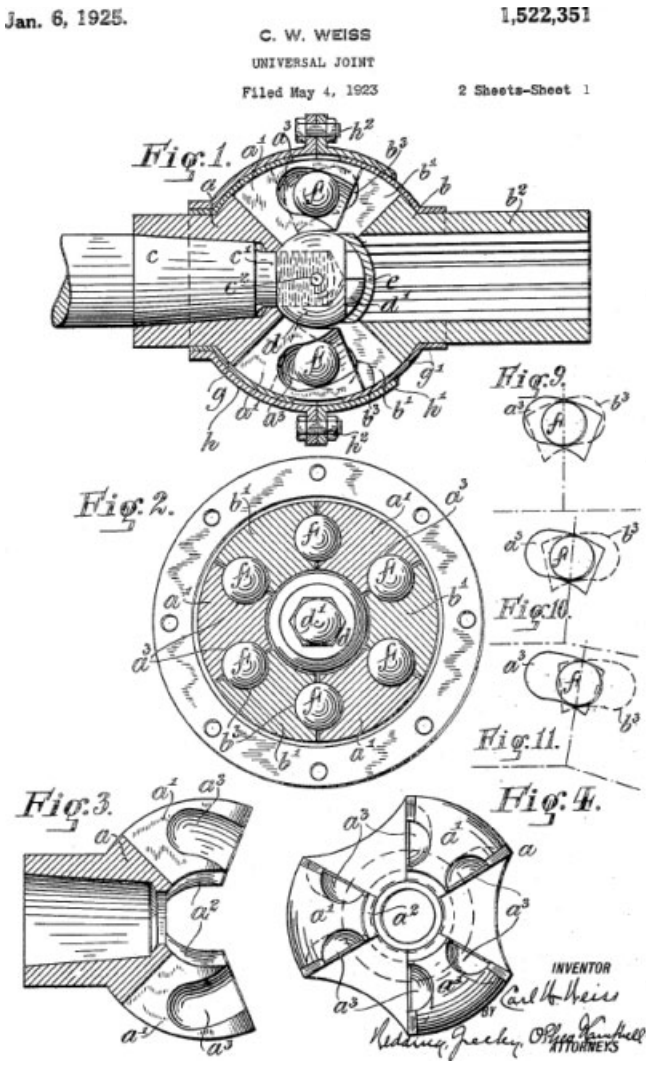


Figure 10. Extract from US patent 1522351. (Reproduced from Weiss (1929)).

All components are made of steel material and are heat treated to achieve a hardness of 57 and more HRC to withstand contact pressures up to 4000 N/mm^2 and more. Tracks and spheres are normally hard-machined (ground

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

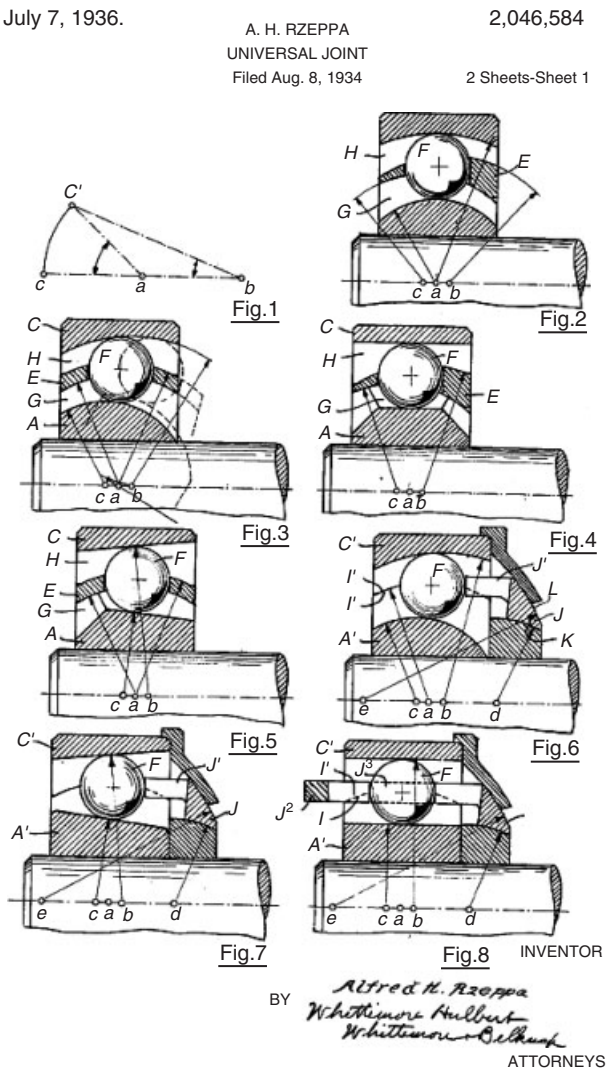


Figure 11. Extract from US patent 2046584. (Reproduced from Rzeppa (1934)).

or hard-turned and hard-milled) to compensate for the influence of hardness distortions; hereby, the rotational backlash of the joint configuration can be limited.

3.1.1.1 Steering mechanism of Rzeppa type joints. An essential feature of the displayed joint is the steering mechanism of the balls; the tracks are axially displaced against the ball plane by an axial offset e_O and an axial offset e_I in the inner race. This results in a mouth opening angle γ of the ball in the track contact (typical values are in the range of 10–18°), as shown in Figure 13.

When the joint is articulated by an angle β (Figure 14) and the longitudinal tracks (i.e., the track offsets e_O and e_I) are arranged symmetrically against the ball plane, each ball—and consequently also the cage plane—is positioned under an angle of $\beta/2$. This half-angle plane—as earlier shown in Figure 8—is essential to achieve a constant velocity behavior of the joint configuration (Figure 14).

3.1.1.2 Joint sizing. Macielinksi (1970) proposed criteria for joint size selection and joint life prediction. The lifetime of the joint is determined by the following relationships:

for a speed of $n < 1000$ rpm:

$$L\,[10^6\text{rev}] = 1,520,360\; n^{0.423}\;[T_{\text{nom}}/T \times A]^3$$

for a speed of $n > 1000$ rpm:

$$L\,[10^6\text{rev}] = 28,245,360\;\;[T_{\text{nom}}/T \times \text{A}]^3$$

where
L: calculated endurance life (revolutions)
A: angle factor with:
$A = [1-\sin(ß)] \times \cos^2(ß)$
ß: articulation angle [°]
n: speed [rpm]
T: torque [Nm]
T_{nom}: Joint application torque [Nm]

The joint application torque can be calculated using the so-called joint capacity factor (JCF).

$$\text{JCF} = m \times \text{pcr} \times {d_\text{B}}^2$$

where
m: number of balls
pcr: pitch circle radius
d_B: ball diameter

The joint application torque is then:

$$T_{\text{nom}} = k \times \text{JCF}$$

where
k: joint type related factor.

3.1.1.3 Longitudinal track forms. The maximum angle of the joint is on one hand limited by the contact of the shaft with the outer race and on the other hand by the support of the balls in the end positions of the outer and inner race tracks. By a modification of the longitudinal track form, a higher angle capability or an increase in the axial safety margin in the longitudinal track can be achieved (Figure 15). A straight track runout was proposed in 1973 (Welschof and Aucktor 1973), allowing a better ability to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

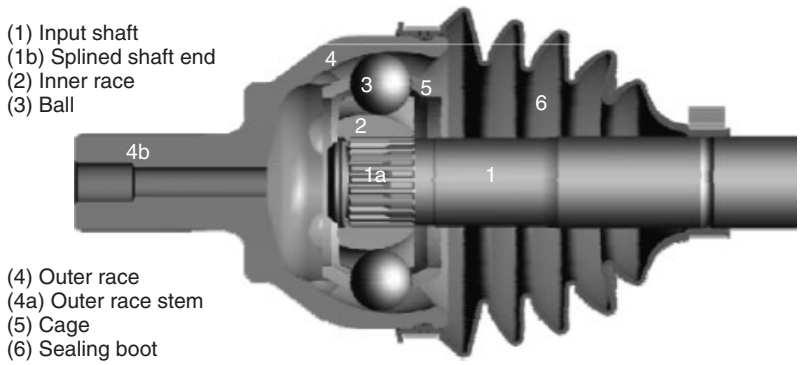


Figure 12. Rzeppa type ball fixed joint.

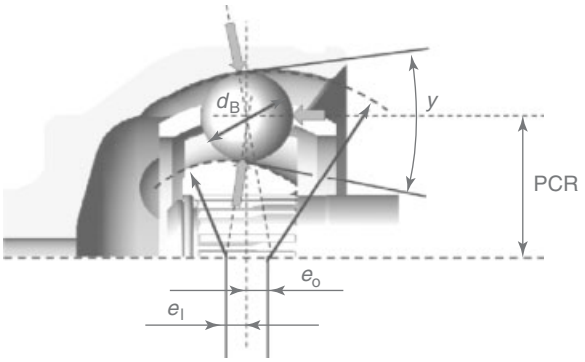


Figure 13. Track steering of a Rzeppa type joint.

apply forming processes for the component manufacture whilst increasing the angle up to ∼50°. Other track form modifications, for example, introducing an angled straight track or an s-shape track, see also Schwärzler and John (1999), allow an increase in the angle to 52° and even more.

3.1.1.4 Transverse track forms. The transverse track form (Figure 16) defines the support of the ball inside the track and has some strong impact on the risk of local stress concentration during operation by a track edge contact.

The transverse track form can be described by the contact angle α and the track conformity value $K = r_T/r_B$.

A round track has a contact angle $\alpha = 0°$ and a conformity value of (typically) $K = 1.002\ldots\ 1.01$ and the advantage of an almost even pressure distribution over the complete track height. A gothic arch track or an elliptical track has the disadvantage that the complete track cross section cannot be applied with contact pressure. They do, however, display a higher tolerance sensibility, that is, the influence of manufacturing tolerances on the rotational backlash of the joint configuration is smaller. Furthermore,

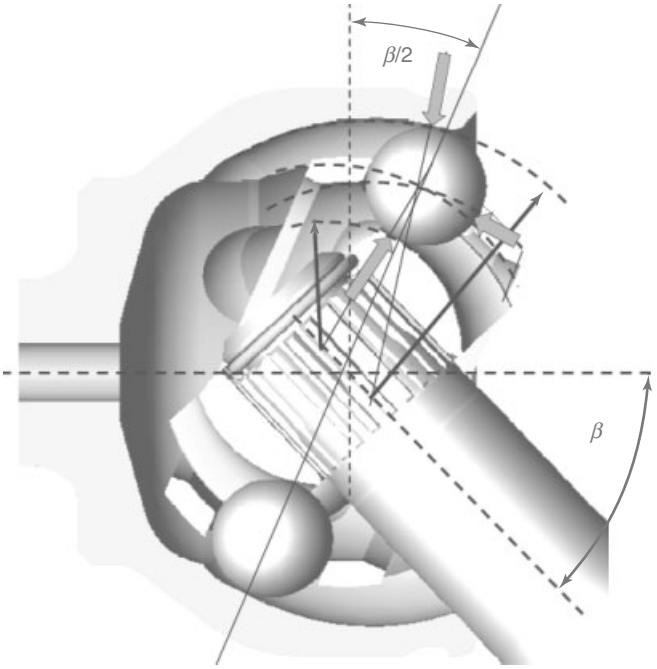


Figure 14. Articulated Rzeppa type joint with half-angle cage plane.

the risk of track edge contacts during high torque operation can be minimized. Typical values for both elliptical and gothic arch tracks are contact angle $\alpha = \sim 35° \ldots 45°$, conformity $K = \sim 1.03\ldots 1.05$.

3.1.1.5 Basic track layout and sphere forces. As shown in Figure 13, during torque transfer, the cage is loaded by an axial force caused by the mouth opening angle γ. As a reaction, the inner race is axially pushed against the cage inner sphere, whereas the cage itself is axially pushed against the outer race inner sphere.

During operation, these forces lead to friction forces in the spheres, which contribute to torque losses during

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

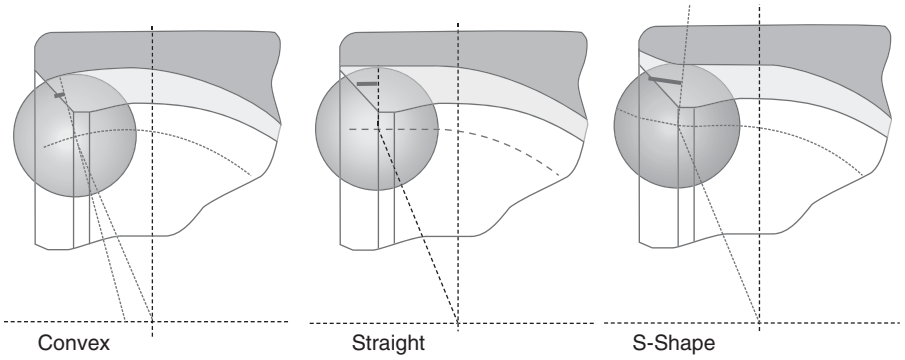


Figure 15. Longitudinal track form runouts in Rzeppa type joints (track safety margins highlighted in red).

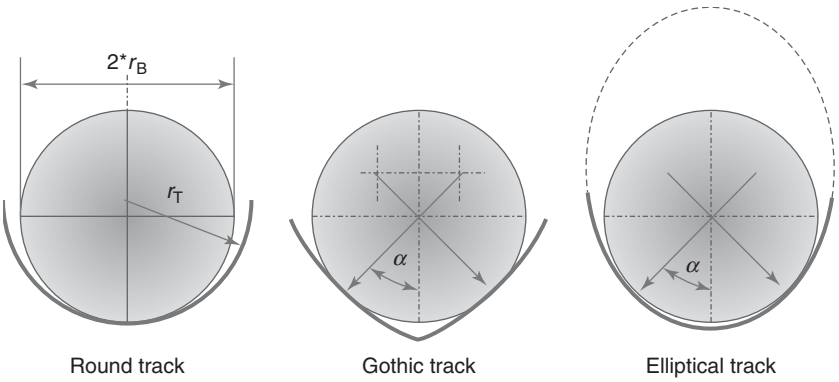


Figure 16. Transverse track forms.

operation and can result in overheating leading to damage of the joint. A reduction of the spherical forces is, to a certain extent, possible by a reduction of the track offsets but negative impact on the ball guidance need to be considered.

An alternative solution to reduce the resulting spherical forces is to arrange the individual tracks in such a form that their axial force components are compensated. By an alternating layout of the track openings, as shown in Figure 17, the axial forces on the spheres can be eliminated. Such opposed track layout can reduce losses in the joint by up to 35% and more, but the negative influence on the maximum achievable articulation angle needs to be considered, as the axial track safety margin becomes smaller. A combination of the opposed track configuration together with changes in the longitudinal track form was proposed (e.g., in Hildebrandt, Horst, and Rickell, 2006), providing a design that combines the advantages of the "opposed track configuration" with an angle capability of up to 52°. Case studies showed an improvement in fuel efficiency by up to 1.3% when considering this technology (in combination with adapted lubrication system) for high installation angle applications (Hildebrandt and Cubert, 2007).

3.1.1.6 Ball guidance by the cage. When rotating an articulated joint, the ball is travelling along the track with an oscillating movement. Owing to the offset of the tracks, the effective mouth opening of the ball varies with the rotational position.

Figure 18 shows the changes in the effective mouth opening angle (the so-called control angle that also considers the effect of the transverse track contact angle) under rotation as function of the articulation angle.

For small angles, the offset layout of the tracks still provides an axial force of the ball against the front cage window plane. With larger articulation angles, the effective mouth opening angle during one rotation can become zero or can even reach negative values, that is, the backward side of the cage window is loaded.

To avoid a locking between a parallel track configuration or to ensure a proper transition between positive and negative control angles, each ball needs to be guided and steered by the cage. Therefore, a narrow tolerance fit

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

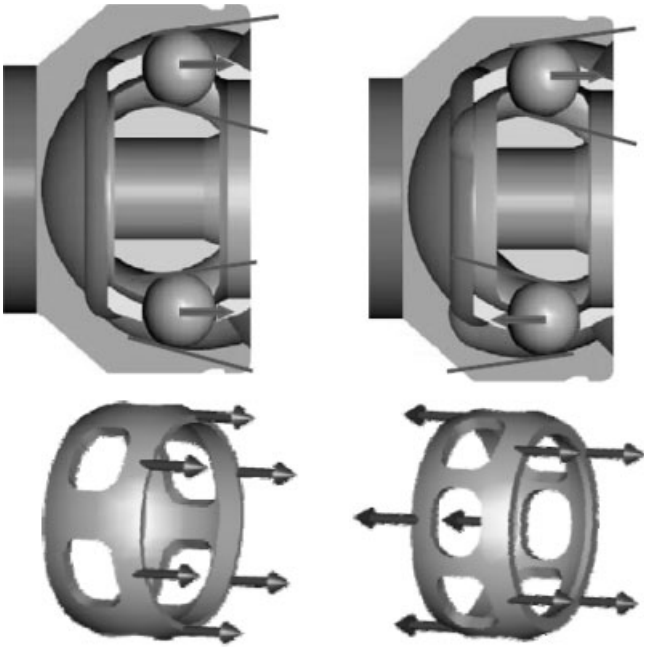


Figure 17. Rzeppa configuration (a) and opposed track configuration (b).

between ball and cage window is required, typically with maximum clearances of only a few 1/100 mm.

3.1.2 Ball plunging joints

Almost in parallel to the early fixed ball joint concepts from Rzeppa, Weiss, and others (Section 3.1.1), plunging joint concepts were also developed. They are typically used as inboard joints in automotive driveshafts. Major design parameters as described earlier (e.g., transverse track form, joint sizing, ball guidance in the cage) are also valid for these joint types.

A major differentiator is the type of steering mechanism, steering mechanism which, besides the articulation capability, must allow an axial movement between input and output.

3.1.2.1 Ball plunge joints with track steering. One of the most popular ball plunge joint with track steering is the VL type/cross groove joint, as shown in Figure 19.

The balls are guided in angled tracks (typically helix angles are in a range of $11^\circ - 16^\circ$), each of the balls is held in a crossing between an inner race and an outer race track.

Owing to an alternating opening of neighbored tracks, the cage is axially balanced in a similar way as described for "opposed tracks" (see Section 3.1.1.5, especially Figure 17).

In normal operation mode, neither the inner race nor the outer race is in axial contact with the cage. An axial plunge between inner race and outer race by a value of p is split into a relative axial displacement of the cage plane in the outer race by a value of $p/2$ and an axial displacement of the cage plane relative to the inner race also by the value of $p/2$. This leads to a very compact assembly of the joint in axial direction; the outer race only needs to allow approximately half of the axial plunge of the whole joint.

Figure 19a shows a flange type joint as a so-called *long plunge version*, which has no plunge limitations by a spherical contact. Such joint concept is typically used in combination with a fixed joint in a front-driven vehicle driveshaft.

Figure 19b shows a VL type joint in a monobloc type configuration with a closed outer race bottom. The plunge capacity of the joint is limited by a sphere contact between inner race and cage; such "inner stop" concept is typically used in rear axle driveshaft, connecting two VL type plunging joints in a floating configuration.

VL type joints typically provide a plunge capacity up to 50 mm and an angle capability of 22°.

3.1.2.2 Ball plunge joints with sphere steering. The DO type joint is a ball plunge joint with an angle capacity of up to 31° and a plunge of up to 50 mm. The base concept goes back to an invention of G. Devos (1964).

The joint consists of an inner race and an outer race, both with parallel track races. A ball cage is guided on its inner diameter on the inner race outer sphere. On its outer diameter, the cage is guided in a cylindrical diameter of the outer race (Figure 20).

By an axial displacement of the inner and outer diameter of the relative to the ball plane, the balls are guided on a half-angle plane during articulation. With this double offset steering mechanism, a constant velocity behavior is achieved.

During plunge, the subassembly of inner race, cage, and balls are moving relative to the outer race. Therefore, the joint is significantly longer than a VL type joint. But as a result of a straight track configuration, when introducing small functional clearance between inner race and cage sphere, the amount of sliding of the ball is reduced significantly. Owing to such "rolling plunge" design, the plunge resistance for a cage steering concept can be much smaller than that for a track steering concept.

The DO type joint is used in front-driven vehicles and can provide benefit for applications, where a high application angle is required, for example, in sport utility vehicles (SUVs).

3.2 Tripod joints

When in the sixties with the Austin Mini, the modern style front-wheel-driven cars started become the industrial trend;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

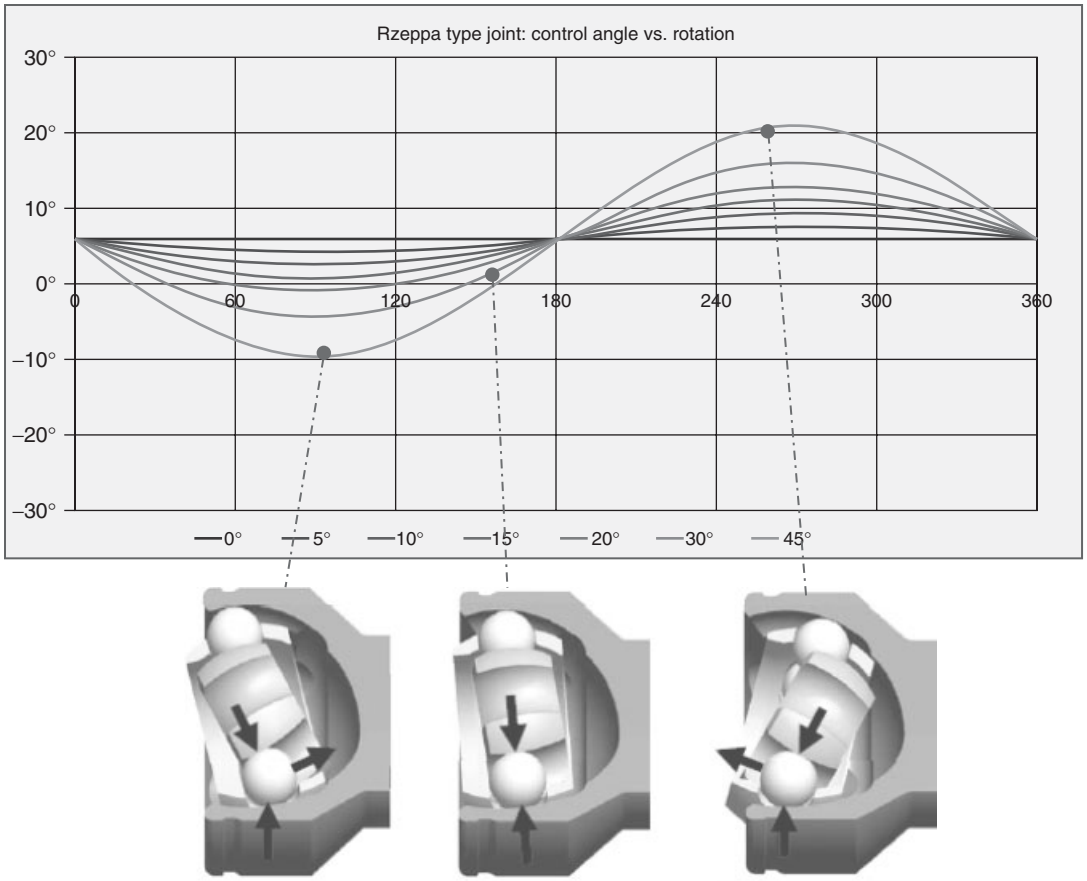


Figure 18. Mouth opening (control angle) under rotation and articulation.

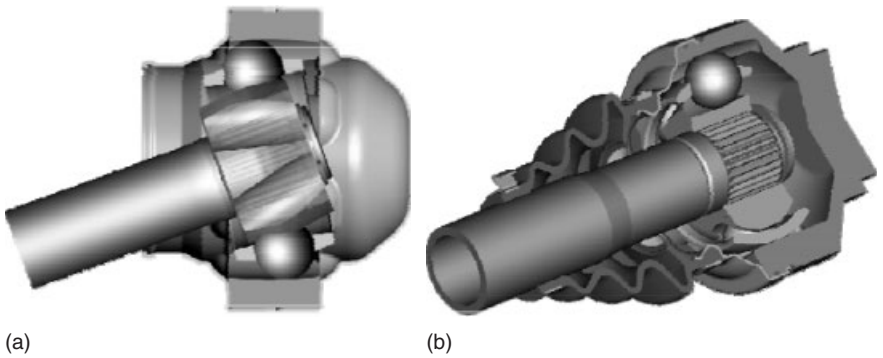


Figure 19. VL type joint; left: flange type long plunge version, right: monobloc type joint with "inner stop" function.

first the tripod joint-equipped shafts were introduced in high volume production, mainly from French car makers.

Tripod joints are based on a different working principle to ball joints. In a tripod joint rollers are used for torque transfer; they are positioned on radially directed arms (pegs) and guided in longitudinal tracks.

3.2.1 Tripod joint driveshafts

Figure 21 shows as an historical example a tripod-equipped driveshaft (extract from historical catalog material (Spicer, 1979)). On the wheel side the so-called GE fixed joint (GE stands for Glaenzer Exterieur) is applied. On the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

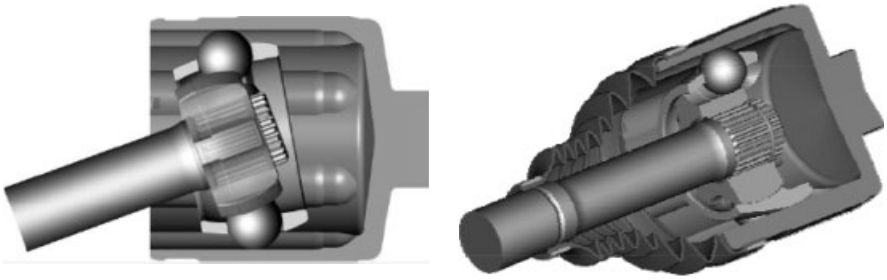


Figure 20. Double offset (DO) type constant velocity joint.

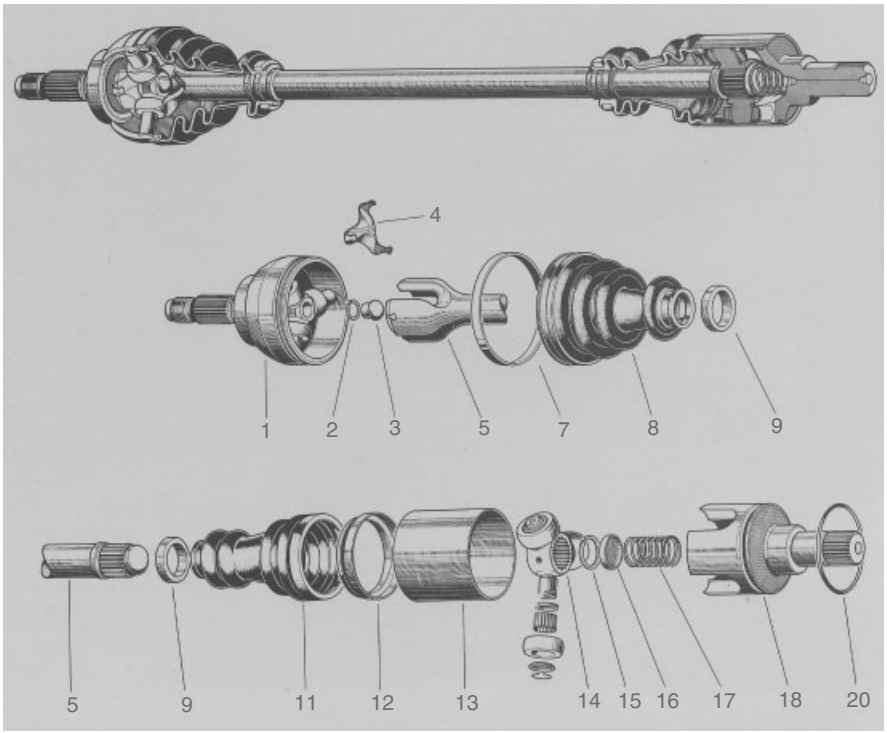


Figure 21. GE and GI type joint/catalog extract. (Reproduced from Glaenzer Spicer (1979). © GKN Driveline.)

differential/inboard side, the so-called GI plunging joint (GI stands for Glaenzer Interieur) is used.

The in board joint GI transfers the torque from the differential into a joint housing (18) that is equipped with three longitudinal tracks. In these tracks, rollers are guided which are positioned on the arms of a tripod body (14). This structure allows a plunge and an articulation of the tripod body against the outer race.

Through the interconnecting shaft (5), the torque is transferred from the GI tripod body to the three longitudinal tracks of the GE shaft end, in which—again—rollers are guided. These rollers are positioned on a three-armed tripod structure, which is connected to the joint outer race end (1). A fixation device (2–4) suppresses axial movements of the GE joint configuration.

Both joints are sealed with boots and its boot clamping devices (9–13 and 20 for GI, 7–9 for GE).

The displayed configuration also includes a preload spring assembly between GI tripod and outer race (15–17) to ensure a proper fixation of the shaft in the differential.

3.2.2 Constant velocity behavior and orbital movement

While the importance of the half-angle plane to achieve a constant velocity behavior was described in Section 3, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

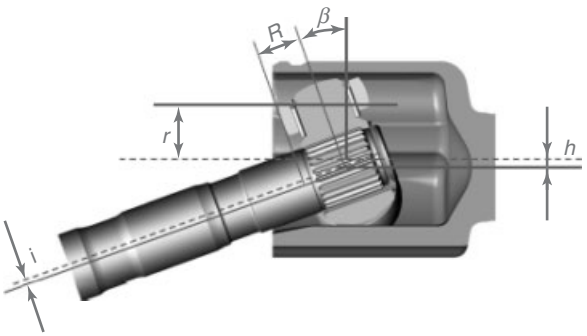


Figure 22. Orbital movement in a tripod type joint.

rollers are articulated by the full bending angle of the joint; therefore, a tripod joint does not have a half-angle plane.

Nevertheless, Orain (1976) could prove that such a concept can have a constant velocity behavior, as long as the center of the tripod body is free to make an orbital movement and the shaft is able to displace parallel to its axis (Figure 22).

This orbital movement takes place three times per revolution; the amplitude of the movement can be calculated using:

$$i = r/2 \times (1/\cos\ \beta - 1)$$

and

$$h = r/2 \times (1 - \cos\ \beta)$$

Here, the value i describes the amplitude of the movement relative of the shaft axis, and the value h refers to the same movement of the tripod center, but measured relative to the track center.

For typical shaft length in automotive applications, a constant velocity behavior of tripod-equipped driveshafts can be assumed.

3.2.3 *Axial force generation*

During articulation of a tripod type joint, the contact force between tripod body and tracks results in an axial force component, which is directed axially along the shaft axis (Figure 23). As a consequence, a tripod joint can generate axial vibrations if the working angles or the friction forces in the contact areas are too high ("shudder noise").

3.2.4 *Plunging tripod joint types*

The major functional properties of a tripod plunging joint depend on the guidance of the rollers on the pegs and on the contact conditions of the rollers in the track, as compared in Figure 24.

For front drive applications, typically, a needle bearing between rollers and pegs is applied to minimize the plunge resistance during an axial movement and reduce the generation of axial forces. Tripod joint concepts with a further degree of freedom between the roller and the peg can improve the rolling behavior of the roller along the track, as sliding friction under articulation is reduced or even eliminated (so-called shudderless joints).

Typical application examples of tripod plunging joints are shown in Figure 25. Figure 25a is a GI type joint for a front-driven vehicle application with a needle bearing support of the rollers directly on the peg, this being a very cost-effective solution for small angle applications. Figure 25b is a so-called AAR (angular adjusted roller) concept, in which the roller consists of an outer ring, an inner ring, and a needle bearing between these. The roller assembly is guided on a spherical end of the peg, which allows an articulation of the roller against the shaft axis and a positioning in the track; hereby, a rolling movement can be ensured even under larger articulation.

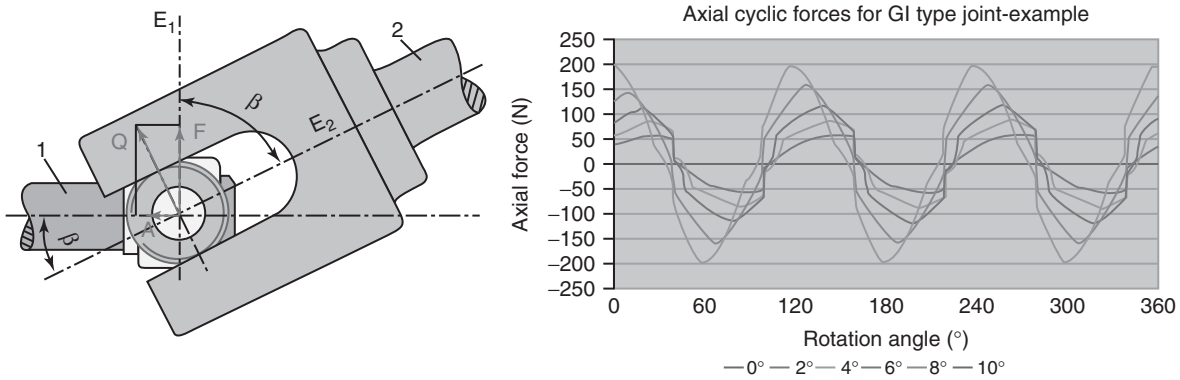


Figure 23. Axial force generation in a tripod type joint.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

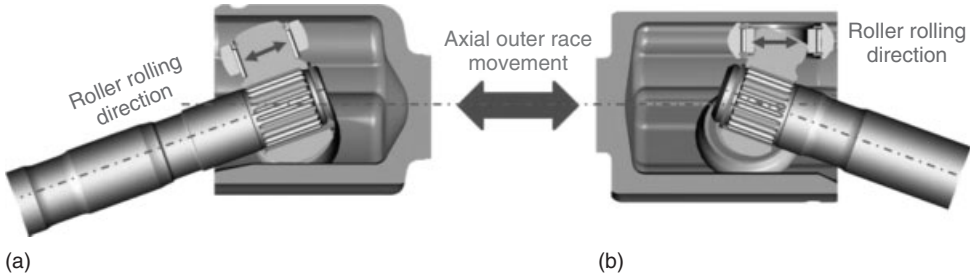


Figure 24. Influence of roller position on plunge resistance.

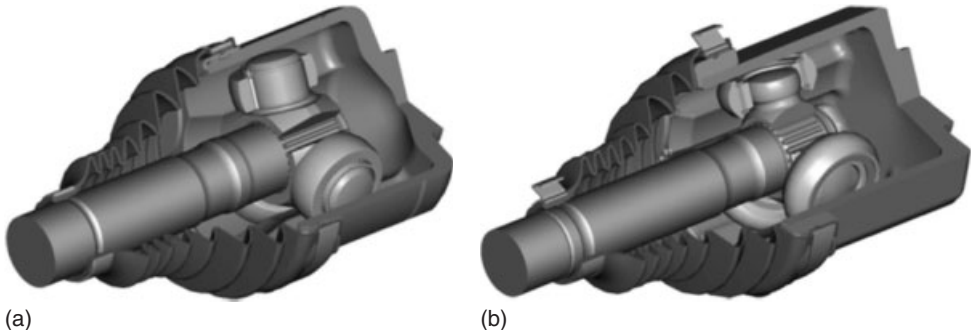


Figure 25. Tripod plunging joint application examples; (a): GI type joint, (b): AAR type joint.

4 SIDESHAFT CONFIGURATIONS

4.1 Selection criteria for driveshafts in vehicle

The joint type selection for automotive driveshaft applications significantly depends on the plunge and angle requirements in the vehicle.

Front-driven vehicles require on the wheel side fixed joint concepts that allow a sufficient steering movement of the driven wheel and small vehicle turning circles. Therefore, maximum angle capabilities of $>45^{\circ}$, in some cases $>50^{\circ}$, are necessary. Preferred choices are fixed joints based on the Rzeppa principle or based on opposed track principle in combination with adapted longitudinal track forms (Section 3.1.1)

On the inboard side, the joint type selection is mainly determined by the plunge-angle characteristics of the joint. Figure 26 shows as an example the kinematics curve of a vehicle application. Depending on the suspension movement, a specific relationship between axial movement and articulation is required. The selected plunge joint concept (here GI type joint) must have a plunge-angle characteristics that allow a free suspension movement without collision of the joint internals. Furthermore, built-tolerances and the assemble ability of the driveshaft in the vehicle need to be considered.

The type of plunging joint also depends on the drivetrain configuration; vehicles with automatic transmission require a low high frequency plunge resistance to reduce the impact of engine/gearbox vibrations during idling; here, tripod type joints are preferred choice.

During change from drive mode to coast mode, a too high rotational backlash of a plunge joint can result in clonk noises caused by contact changes in the torque transferring components. Here, ball plunge joints with a narrow assembly fit can be preferred choice.

The axial force generation of a joint can lead to so-called shudder noises in the vehicle. Especially, for larger installation angle $>5^{\circ}$, shudderless tripod joints (as e.g., the AAR type joint, see Section 3.2.4) can be required to ensure noise-free torque transfer in all driving conditions.

Where for front-driven vehicles, a driveshaft typically consists of an outboard fixed joint and an inboard plunging joint, the driveshafts for rear-wheel-driven (RWD) vehicles can also consist of a series combination of two plunging joints. For high performance RWD vehicles, a combination of two VL type joints can be preferred choice due to a compact axial packaging and very small rotational lashes.

For the rear axle of a hang-on 4WD or for low cost rear-wheel-drive applications, combinations of two plunging joints of GI tripod are known from series production. As here the axial force generation and plunge resistance

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

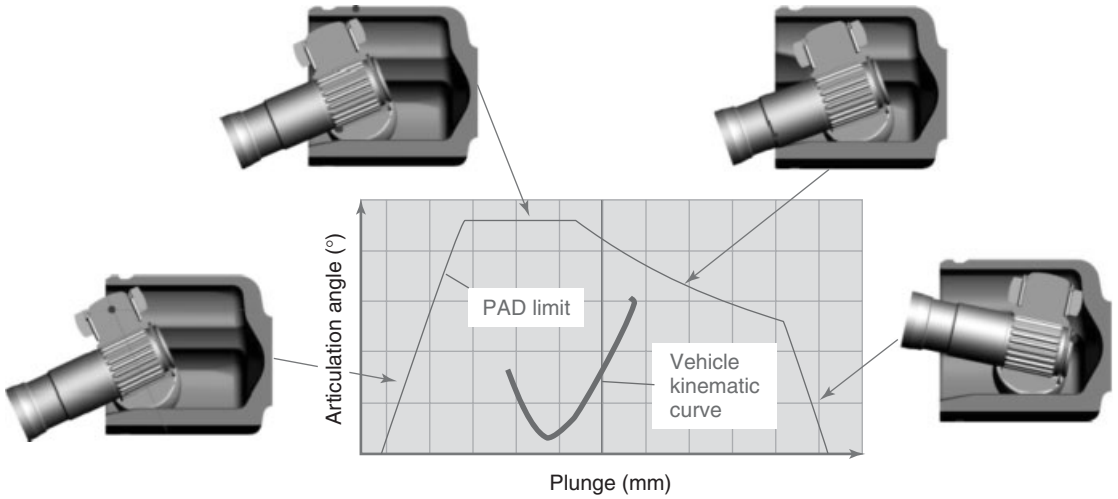


Figure 26. Vehicle kinematic curve in comparison to the PAD (plunge-angle diagram) of a plunge joint.

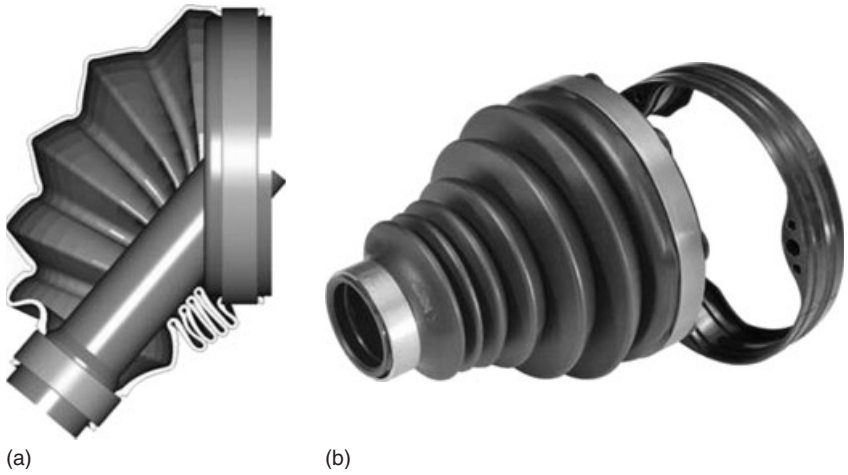


Figure 27. CVJ sideshaft sealing systems.

requirements are lower, concepts without needle bearing between roller and peg are also known.

When combining two plunging joints in a driveshaft, a centering is required to avoid disassembly or operation in an end position. For VL type joints, inner stops can be used (Section 3.1.2.1). For GI type joints, a centering can be achieved by means of springs in both joints (Figure 21).

4.2 Sealing systems

CV joints for automotive applications are typically sealed by a boot that is fixed on its larger diameter on the joint outer race and on its smaller diameter on the interconnecting shaft.

The design of the boot needs to be adapted to the specific joint concept to ensure good contact conditions under maximum articulation (Figure 27a). In case of non-round outer race shapes (e.g., trilobe shape of tripod type joints), an insert can be used as adaptor between the different shapes of the outer race and the boot.

Rubber material is more and more being replaced by thermoplastic elastomers (TPEs) that provide a better durability and improved resistance against environmental influences, and furthermore a reduced system weight and an improved material recyclability.

For front sideshaft applications typical ambient temperatures for the wheel side range from -40°C up to 100°C. For the inboard side, the boot surface temperature can achieve

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

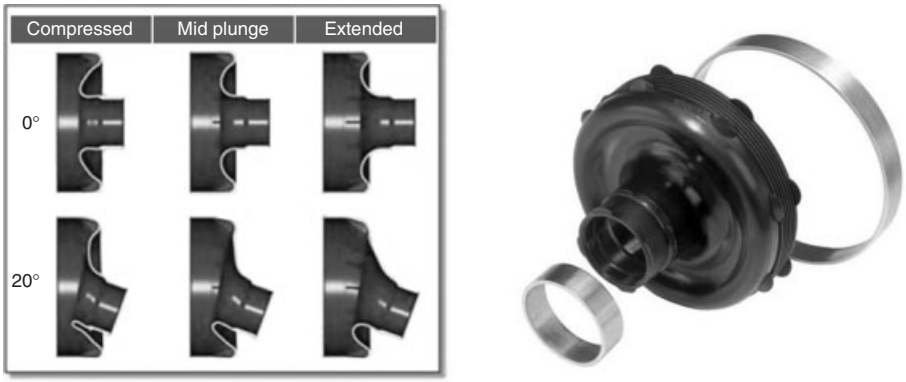


Figure 28. Diaphragm type sealing system for rear sideshafts.

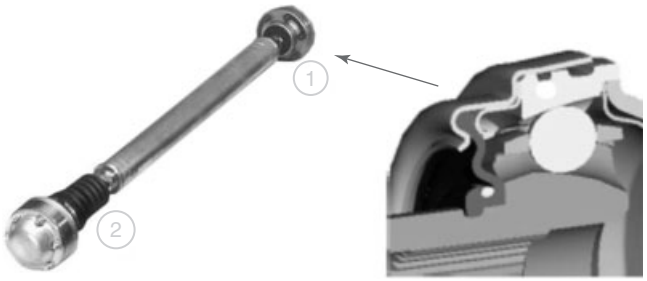


Figure 29. High speed sealing system for a CVJ propshaft application: (1) fixed joint, (2) plunge joint.

peak values of 150°C and more due to heat transfer from exhaust and catalytic converters.

For rear sideshafts, due to the reduced articulation requirements (maximum angle up to 25°), more compact diaphragm type sealing systems with plunge capability of up to 45 mm can be applied (Figure 28).

For high speed applications of constant velocity joints, for example, in propshafts, sealing systems are required to withstand the high centrifugal forces and also high temperature levels, typically caused by heat transfer from the exhaust system.

For these applications, high performance rubbers, HNBR, and silicone are preferred choices, whereas TPE materials so far can be regarded as niche.

Figure 29 shows as an example a propshaft with joint sealing systems for rotational speeds up to 11,500 rpm, articulation angles up to 5°, and 50-mm plunge capability.

4.3 Lubrications for constant velocity joints

In automotive applications, CVJs are typically lifetime lubricated with specially tailored greases. The lubrication and sealing system needs to withstand typical working temperatures of −40 up to 100°C (normal temperature range). For special applications, temperatures can reach up to 150°C (short period: 180°C) either with high internal heat generation of the joints (e.g., due to high installation angles) or due to external heating (e.g., from the exhaust system).

Typically, normal temperature greases are formulated with a lithium soap thickener, mineral oils and special additive packages. High temperature application greases are dominantly based on polyuria thickeners using synthetic or semisynthetic oils.

Owing to high contact pressures of 4000 MPa and more and the oscillating movement of the components, CVJs operate in the mixed or boundary lubrication regimes. Ball joint greases therefore normally use MoS_2 in combination with graphite as part of the additive package (but note that for tripod type joints, greases with MoS_2 can have negative impact on durability and function).

The friction coefficient of the grease has a strong influence on joint performance in the vehicle. For plunge joint, axial force generation and plunge resistance can be significantly reduced and can contribute to vehicle noise and vibration optimization. For ball type joints, a reduction of joint losses (in the range of −30% and more) with corresponding impact on vehicle fuel consumption and CO_2 emission can be considered.

4.4 Interfaces

The differential side of a driveshaft typically uses a splined connection between plunge joint outer race and differential side gear. For transverse engine installations, additional linkshafts or long stem outer races can be used to avoid unequal sideshaft length for left and right side. In these cases, an intermediate bearing is applied which supports the outer race via a bracket against the engine/gearbox housing.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

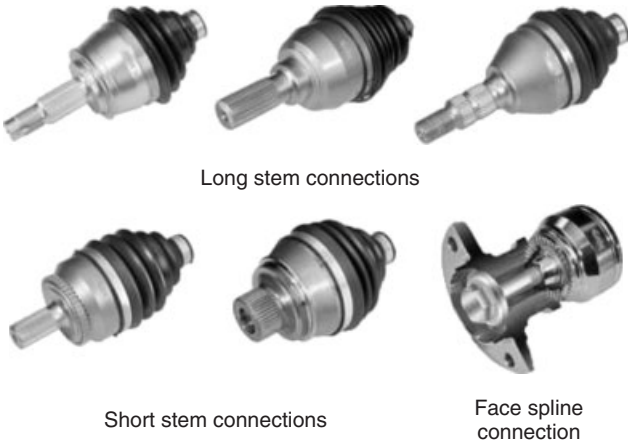


Figure 30. Wheel side driveshaft interfaces.

On the wheel side joint, the connection system is mainly determined by customer preferences. Figure 30 gives an overview of commonly used interface types between outboard joint and wheel hub unit. Both long and short stem connections use a radial spline for torque transfer to the wheel hub unit; a face spline connection uses an axial spline and provides a more compact packaging and assembly advantages.

RELATED ARTICLES

Propshafts (Driveshafts)

REFERENCES

Altmann, W.E., Enke, K., and Rothweiler, A.F. (1964), Sliding Joint, US patent 3298200, filed 1964, USA.

Devos, G. (1964) French patent 1418233, October 8, France.

Glaenzer Spicer (1979) *Catalogue material*.

Hildebrandt, W. and Cubert, J.-M. (2007) *Das neue Gleichlaufgelenk 'countertrack' für den Einsatz als radseitiges Festgelenk im Antriebsstrang*. 16th Aachener Kolloquium Fahrzeug- und Motorentechnik; S. 1651–1666, Aachen, Germany.

Hildebrandt, W., Horst, J., and Rickell, R. (2006) New constant velocity fixed joints for front wheel drive cars ('Neue Gleichlaufgelenke für den Frontantrieb'). *ATZ Automobiltechnische Zeitschrift*, **108**, 356–363. Springer, Germany

Macielinski, J.W. (1970) 'Propeller Shafts and Universal Joints - Characteristics and Methods of Selection', Session 3, Paper no. 29 in *Proceedings of the Institution of Mechanical Engineers, Conference Proceedings September 1969* vol. 184, no. 9, 516–554, Institution of Mechanical Engineers, USA. DOI: 10.1243/PIME_CONF_1969_184_308_02

Orain, M. (1976) *Allgemeine Theorie und experimentelle Forschung der Gleichlaufgelenke*, Glaenzer Spicer, Germany.

Rzeppa, A.H. (1934) *Universal Joint*. US patent 2046584, filed Aug 8th, USA.

Schwärzler, P. and John, F. (1999): *Constant velocity universal joint*, US Patent 6319133.

VDI (2003) *Cardan Shafts and cardan shaft lines: Homokinetic mechanisms*, Standard 2722, August, Verband Deutscher Ingenieure/Association of German Engineers, Germany.

Weiss, C.W. (1929) *Universal Joint*. US patent 1522351, filed May 4th, USA.

Welschof, H-H. and Aucktor, E. (1973) *Constant Velocity Joint*. US Patent 3879960, filed Oct 25th.

FURTHER READING

There are two excellent texts listed below that the interested reader will find useful for further information. These give a both a wide overview and good detail on the topic of CVJ technology and are recommended as a supplement to the introduction provided here.

Schmelz, F., Graf v. Seherr-Thoss, H.-C., and Aucktor, E. (2006) *Universal Joints & Driveshafts; Analysis, Design, Applications*, 2nd enlarged Edition, Springer-Verlag, Berlin Heidelberg, Berlin, Germany.

Pierburg, B. and Amborn, P. (1998) *Constant-Velocity Driveshafts for Passenger Cars: Driveline systems, joints, interconnecting and longitudinal driveshafts*; GKN Löbro – Landsberg / Lech; Verlag Moderne Industrie, (Die Bibliothek der Technik; Bd. 170).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline International GmbH
DOI: 10.1002/9781118354179.auto095
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Drivetrain Noise, Vibration, and Harshness

H. Rahnejat[1], S. Theodossiades[1], P. Kelly[2], and M.T. Menday[1]

[1]*Loughborough University, Loughborough, UK*
[2]*Ford Motor Company, Cologne, Germany*

1 Introduction 1
2 Clutch Take-Up Judder 2
3 Clutch In-Cycle Vibration (Whoop) 3
4 Driveline Shuffle and Clonk 5
5 Transmission Rattle 8
6 Closure 11
References 11

1 INTRODUCTION

There is a plethora of noise, vibration, and harshness (NVH) concerns in the drivetrain system. These include low frequency, relatively large displacement inertial dynamics such as clutch take-up judder, which is the first torsional mode of the driveline system (Rabieh and Crolla, 1996; Centea, Rahnejat, and Menday, 2001). Another low frequency rigid body oscillation is driveline shuffle, which is the first torsional vibration mode, coupled with fore-and-aft oscillations of the vehicle, known as *shunt* (Farshidianfar *et al*., 2002). These concerns may be regarded as harshness, in the frequency range 3–10 Hz, depending on the drivetrain system, the driveline gearing, and the vehicle mass.

There are also vibration concerns at higher frequency range. One is the clutch in-cycle vibration, referred to in the industry as whoop; the resulting radiated sound during clutch movement (Hasebe and Aisin Seiki, 1993). Kelly, Rahnejat, and Biermann (1998) and Kushwaha *et al*. (2002) showed that the whoop phenomenon is as the result of the combustion fundamental frequency (half engine order vibration in the case of a four-stroke engine) from the firing of cylinders close to the flywheel. This induces conical motion of the flywheel, causing it to impact the clutch system during the clutch actuation process. The ensuing vibration of the clutch system is in the frequency range 150–250 Hz and is felt by the rough oscillations transmitted from the flywheel impacts and along the clutch cable to the clutch pedal as well as emanating as a whoop sounding noise from the driver's foot-well area. Hydraulic actuating clutch systems are also not immune from this phenomenon as the stick–slip oscillations of the actuator induce the same problem.

Another NVH problem is the transmission rattle, which is the result of impacting of gear teeth pairs. This phenomenon usually occurs in the case of unselected (unengaged) gears (Wang, Manoj, and Zhao, 2001; Theodossiades, Tangasawi, and Rahnejat, 2007; Tangasawi, Theodossiades, and Rahnejat, 2007), although contact separation or transmission error in engaged gear pairs can also contribute (Kim and Singh, 2001; De la Cruz, Theodossiades, and Rahnejat, 2010). Gear rattle has a broadband nature as perceived, because the resulting vibration at the impact sites travels along the structural members as a structure-borne wave and excites their modal behavior. As the drivetrain components are lightly damped, such impulse loading can result in structure-borne noise output. There is also clearly some airborne noise as well from the impact site itself, which can be heard in the absence of significant engine noise, for example, when the vehicle is idling. The nature of rattle is repetitive and at a given transmission speed may be regarded as a steady-state noise and vibration

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

problem. It may easily be confused with diesel engine knock.

In contrast, some other NVH concerns have a transient nature and occur as the result of driver behavior or road input. These include driveline *clonk*, which is an elasto-acoustic response of the driveshaft tubes (Menday, Rahnejat, and Ebrahimi, 1999). This phenomenon occurs when a high energy deformation wave travels from an impact site along the driveline, causing elastic deformation of the hollow thin-walled driveshaft pieces and excite their modal responses (Theodossiades *et al*., 2004). There is coincidence of some of the many structural natural modes of the tubes with their internal acoustic natural modes. This coincidence is referred to as *elasto-acoustic coupling*, leading to noise radiation, usually at quite high frequencies (500–5000 Hz), and emanating as a disconcerting metallic sound. Therefore, like many other NVH concerns, the term *clonk* or *clunk* is chosen onomatopoeically (Menday, Rahnejat, and Ebrahimi, 1999; Biermann and Hagerodt, 1999). The traveling wave is often caused by the impact of teeth pairs of engaged gears under impulsive conditions, such as with sudden clutch actuation, abrupt throttle tip-in or back-out actions, or misalignment of driveshaft pieces from coast to drive condition or as the result of road input. Any misalignment of shafts can also cause contact separation and impact on highly loaded hypoid gears of the differential, giving rise to a tonal sound output, referred to as *axle whine* (Kim and Lee, 2007; Koronias *et al*., 2010).

The impulsive action often leads to teeth pair impact, sometimes severe enough to cause torque reversal, thus leading to severe clonk conditions. Clonk can also take place with driveline shuffle and vehicle shunt (Krenz, 1985). With lower impulsive energy, the severity of clonk is attenuated and a sound more like a *thump* is heard. Side-to-side motion of the driveline pieces about the center bearing of the driveshaft or lateral motion of the driven axle can also cause another vibration concern known as *shudder*.

There are, therefore, many NVH concerns in the drivetrain system. These are mostly as the result of engine order vibration input, unexpected driver actions, road input, lash zones, and low structural damping. The situation has been exacerbated in recent years with the trend toward high output power-to-weight ratio vehicles and use of compact transmissions, differentials, and hollow driveshaft pieces. In this chapter, some of the aforementioned drivetrain NVH problems are discussed.

2 CLUTCH TAKE-UP JUDDER

Clutch judder is a low frequency phenomenon that occurs when the pull away gear has been selected and the clutch pedal operated, bringing the clutch into progressive engagement. The oscillations are readily felt at the driver's seat, this being a suitable location to measure the problem.

The role of the clutch in the slip maneuver is to synchronize the rotational speed of the engine/flywheel with the transmission shaft. When there is no relative angular velocity and the clutch is fully clamped, the clutch pedal is released and the engine torque drives the vehicle forward. During the transition from a fully disengaged to a fully engaged state, a longitudinal body vibration may be felt. This starts as a torsional vibration in the driveline and is reacted at the engine and body mounts before transmitting to the driver's seat. The judder mode is lightly damped and may continue for as long as the clutch slips, and it would be necessary to disengage the clutch in order to stop the judder vibration and prevent the clutch overheating.

Fore-and-aft driver's seat longitudinal acceleration is an adequate indicator of judder and this can be readily measured and subjectively correlated. It may be directly compared with the dynamic behavior of the engine mounts. Clutch judder should not be confused with two other similar low frequency vehicle vibrations; namely

- *Brake judder*, which occurs during braking and at a higher frequency.
- *Shuffle*, which occurs after application of the throttle and with the clutch fully engaged (i.e., it is not related to the friction contact between flywheel and clutch disk).

Both shuffle and judder occur when the first torsional driveline eigen-mode has been excited. For a rear-wheel-drive vehicle, this mode has a maximum amplitude response at the flywheel and has an in-phase response of minimum amplitude at the rear axle. The mode is reacted at the tire-road contact patch; consequently, the low frequency longitudinal oscillations of the vehicle occur.

As judder is a modal phenomenon of the driveline system, a multibody system analysis approach should be undertaken to simulate the conditions. One such model is reported by Centea, Rahnejat, and Menday (2001) for a cable-operated clutch system. The model showed that the propensity to judder was directly related to the slope of kinetic coefficient of friction μ with slip velocity, when its value is negative: $n = \frac{\mathrm{d}\mu}{\mathrm{d}v} < 0$. The larger the negative value of n is, the greater is the judder response. Figures 1 and 2 show increasing oscillations before fully clamped clutch with an increasingly negative slope n. Simulations and experimentation have also shown that the potential for judder is greatly reduced if the slip gradient of the clutch lining friction coefficient is held to be positive or at least zero for its useful service life. The challenge is to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

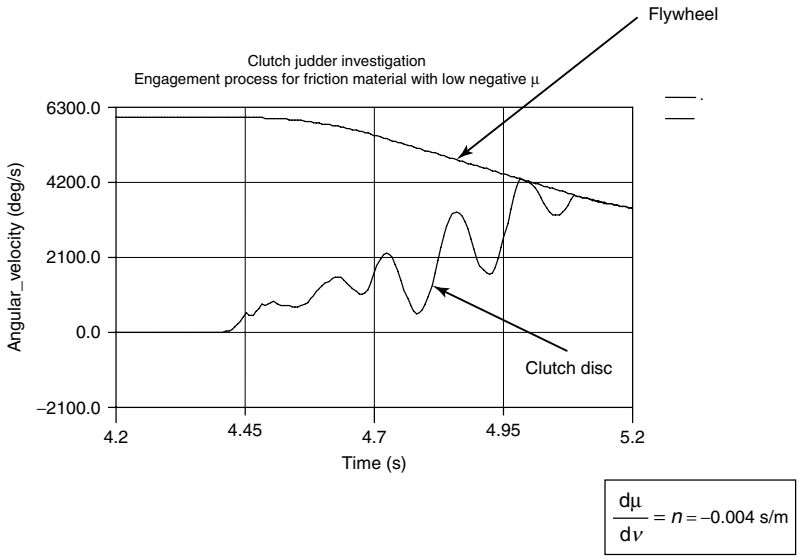


Figure 1. Take-up judder oscillations for low negative friction gradient with slip speed. (Reproduced from Centea *et al*. 2001. © Elsevier.)

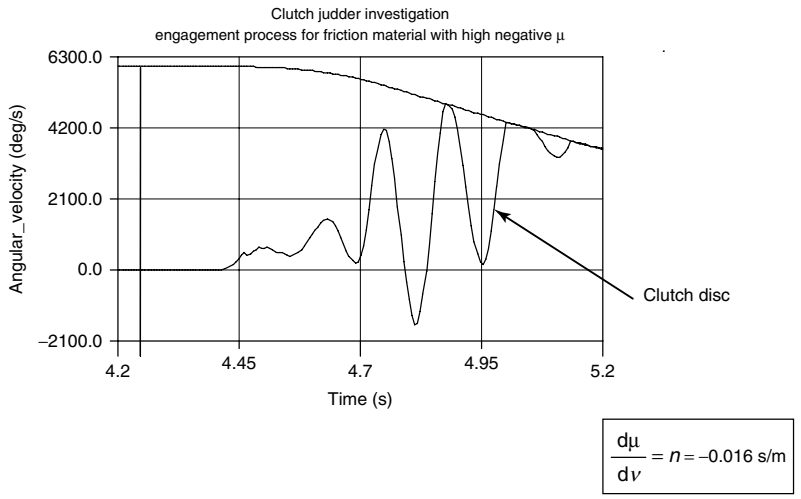


Figure 2. Exacerbated take-up judder with an increasingly negative friction gradient with slip speed. (Reproduced from Centea *et al*. 2001. © Elsevier.)

design a clutch friction disk that maintains these properties throughout service life.

3 CLUTCH IN-CYCLE VIBRATION (WHOOP)

As already mentioned in Section 1, during the clutch pedal movement and with some vehicles (especially those with diesel engines), a tactile clutch pedal vibration is felt, accompanied by a disagreeable noise. Therefore, the phenomenon occurs during the clutch transition state (from engagement to disengagement and vice versa), thus the reason for the chosen name. The accompanying noise is onomatopoeically termed as *whoop*.

Whoop is a transient dynamic phenomenon, occurring during the engagement and disengagement of the drive (engine) with the driven system (the drivetrain system). Figure 3 shows the measured response from the clutch pedal and the driver's foot-well area.

With all NVH phenomena, including clutch whoop, it is essential to determine the underlying cause of the problem

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

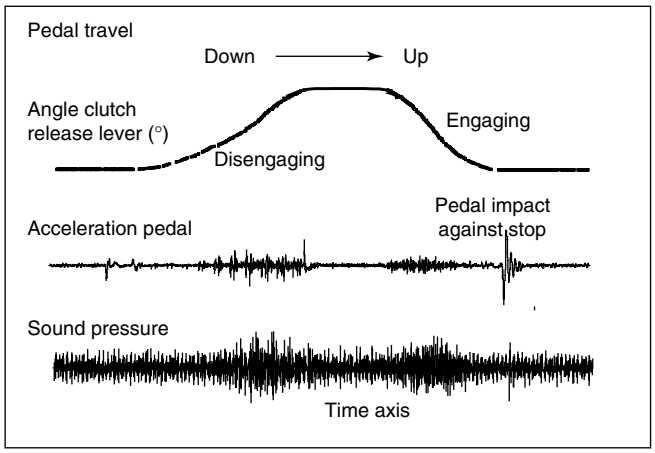


Figure 3. Clutch whoop response during clutch movement. (Reproduced from Kelly, Rahnejat, and Biermann, 1998. © John Wiley & Sons.)

as a prelude to devising a root cause solution. In many cases, however, a root cause solution may be difficult to achieve or may contravene the overall vehicle development strategy, such as those resulting in an increase in the mass or inertia of components. Therefore, a palliative approach may be adopted, which requires the determination of the paths of noise and vibration propagation. Figure 4 shows the path of vibration transmittance from the engine, through the crankshaft–clutch system interface into the vehicle cabin. It can be seen that as the cylinders 3 and 4 are nearer to the flywheel–clutch interface, when ignition occurs,

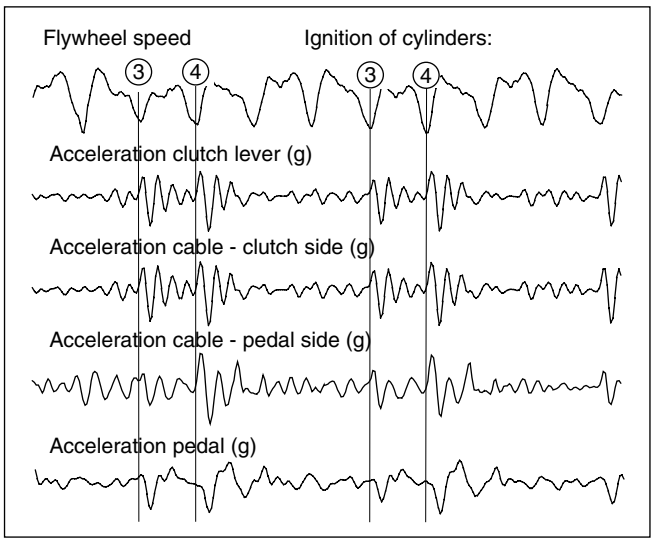


Figure 4. Clutch whoop vibration transmission path. (Reproduced from Kelly, Rahnejat, and Biermann, 1998. © John Wiley & Sons.)

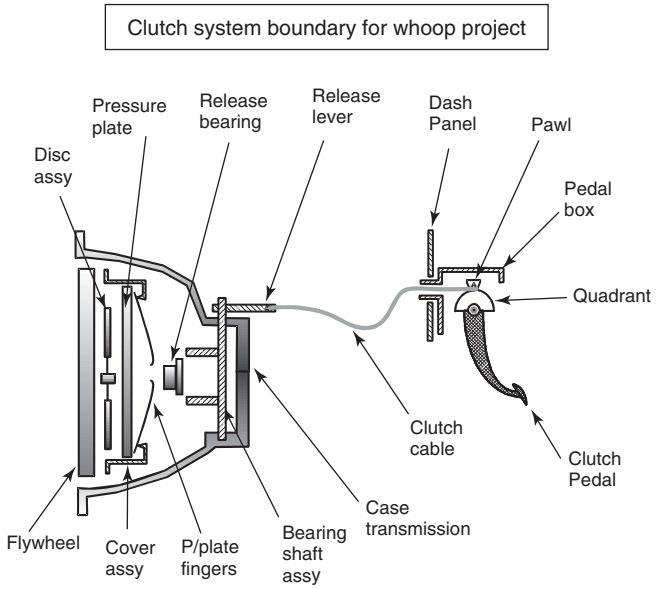


Figure 5. The boundary for the clutch whoop investigation.

increased vibration levels are noted along the transmission path to the clutch pedal.

To implement any root cause or palliative measure, it is necessary to undertake a fundamental study with a prescribed boundary of investigation. For the clutch whoop problem, the system of interest is shown in Figure 5. Clearly, any solution to an NVH problem should not exacerbate other NVH concerns (some of which are described in this chapter). It should also not result in any loss or deterioration of system function. Therefore, traditionally a large mass-stiffness damper known as the *Diehl-fix* has been added to the release lever to attenuate the traveling wave from the flywheel–clutch interface before reaching the clutch cable in cable-operated clutch systems. However, this mass damper often weighs 1–2 kg and occupies a volume of 100–200 mm^2 at a unit cost of $7. As already noted, lightweight and compact power trains are now regarded as essential design features, as well as reduced costs. Thus, other solutions are sought. Therefore, a fundamental study necessitates the development of detailed models for both an in-depth understanding of the phenomenon and parametric simulation studies, rather than expensive physical prototype testing.

A multibody dynamic model, comprising a four-stroke four-cylinder engine and a cable-operated clutch system, is reported by Kushwaha *et al*. (2002). The model includes combustion pressure, firing order of engine cylinders, crankshaft engine bearings, and crankshaft flexibility. Details of the model are beyond the scope of this chapter. Readers should refer to Kushwaha *et al*. (2002).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

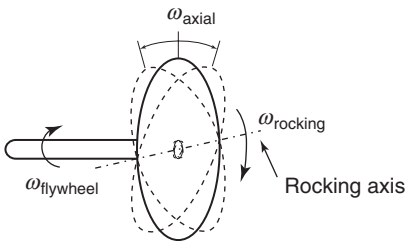


Figure 6. Conical whirl of the flywheel due to crankshaft flexibility.

With crankshaft flexibility included in the model, the applied combustion forces from each cylinder tend to induce combined torsional-bending deflections of the crankshaft. This leads to the conical whirl of the flywheel (Figure 6). This is known as *flywheel nod*.

For a four-stroke four-cylinder engine with a fairly rigid crankshaft, the spectrum of oscillations superimposed on the steady-state angular velocity comprises second engine order (twice the crankshaft angular velocity) and its full integer harmonics (Rahnejat, 1998). However, with increasing flexibility due to the use of lighter materials, spectral contributions due to combustion fundamental at half engine order and all its multiples remain resident on the crankshaft angular velocity. These affect the flywheel motion with its axial nodding causing an impact with the clutch system during the engagement and disengagement processes. The model by Kushwaha *et al*. (2002) predicts flywheel movement within 9% of that measured from the flywheel surface using a series of proximity devices. Figure 7 shows the spectral content of the flywheel axial

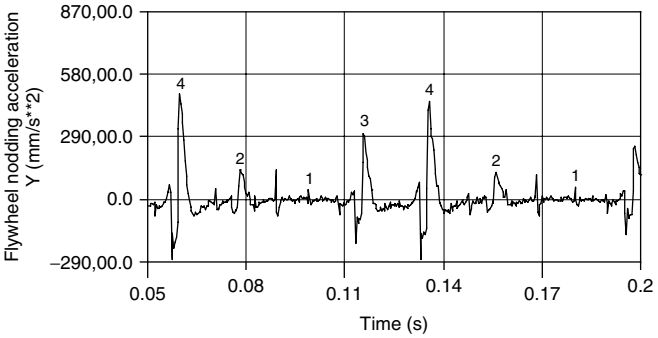


Figure 8. Increased flywheel nodding acceleration with firing of cylinders 3 and 4 proximate to the flywheel. (Reproduced from Kushwaha *et al*. 2002. © Sage/Institution of Mechanical Engineers.)

nodding motion. The increased flywheel movement due to the firing of third and fourth cylinders is sufficient to account for the clutch whoop problem (Figure 8). Note that for a four-cylinder four-stroke engine, the impulsive second order is main engine torsional signal.

4 DRIVELINE SHUFFLE AND CLONK

Shuffle is an uncomfortable tactile low frequency, fore-and-aft, lightly damped vibration of the vehicle. A shuffle mode may continue for several cycles. It can occur following a rapid throttle tip-in or back-out. A tip-in is a fairly rapid actuation of the throttle, whereas back-out is a rapid release of the throttle. In either case, the power flow along the drivetrain changes direction on account of ensuing

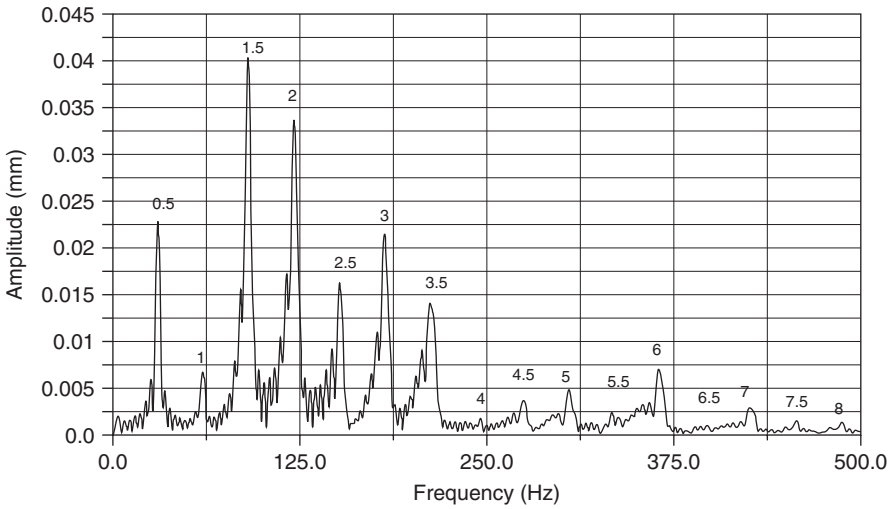


Figure 7. Spectrum of vibration harmonics of engine order in flywheel nodding motion. (Reproduced from Kushwaha *et al*. 2002. © Sage/Institution of Mechanical Engineers.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

acceleration or deceleration (the drivetrain is defined as the torque path between the flywheel and the driven road wheels). The result is the application of an impulse, causing system oscillations, which are torsional for the drivetrain, accompanied and coupled with longitudinal motions of the vehicle (shunt). These occur at the lowest rigid body natural frequency of the system. The extent of oscillations depends on the stiffness and damping of the drivetrain, and in the case of the vehicle body, the longitudinal stiffness of the driven wheels' contact patch in series with the fore-and-aft stiffness of driven axle's suspension bushings. The oscillations are also affected by the lash rate in the torsional system (i.e., the drivetrain). Any rapid changes in the stiffness characteristics of the system can exacerbate the system response through impacts in the various lash zones. Therefore, cycles of shuffle may be accompanied by audible metallic clonk noises as the torque traverses the drivetrain lash during each cycle of shuffle (Krenz, 1985; and Arrundale *et al*., 1998).

Arrundale *et al*. (1998) reported a multibody driveline model, comprising elastic thin-walled driveshaft tubes. Throttle tip-in input from a four-cylinder four-stroke 1.8 L engine was used to drive the model. The results are shown in Figure 9, with low frequency cycles of shuffle at 5 Hz. In fact, shuffle response is often in the range 3–8 Hz, depending on the vehicle inertial properties and system stiffness. Hurried driver behavior in the form of rapid throttle action or release of clutch often introduces a sharp change in the input energy to the drivetrain system. Some of this energy is expended in vehicle shuffle, while the remainder causes impact in drivetrain lash zones and travels back and forth (in severe cases) along the drivetrain. The energy is often sufficient to excite modal response of lightly damped hollow structures such as the modern driveshaft tubes. These sharp high frequency oscillations can be seen with each cycle of shuffle response in Figure 9.

The clonk vibration signal was obtained by a laser vibrometer placed at a suitable distance near to a hollow driveshaft tube as the structural wave passes through it. The induced vibrations were measured by the vibrometer from a suitably reflective tape attached to the driveshaft tube. At the same time, the clonk sound can also be measured as sound pressure variation by a microphone directed toward the same target. A typical clonk noise signal is shown in Figure 10 (Gnanakumarr, 2010).

The signal comprises a ramp up duration of 80–200 ms, followed by an impact time of a few milliseconds and a much longer decay period as shown in the figure. It is

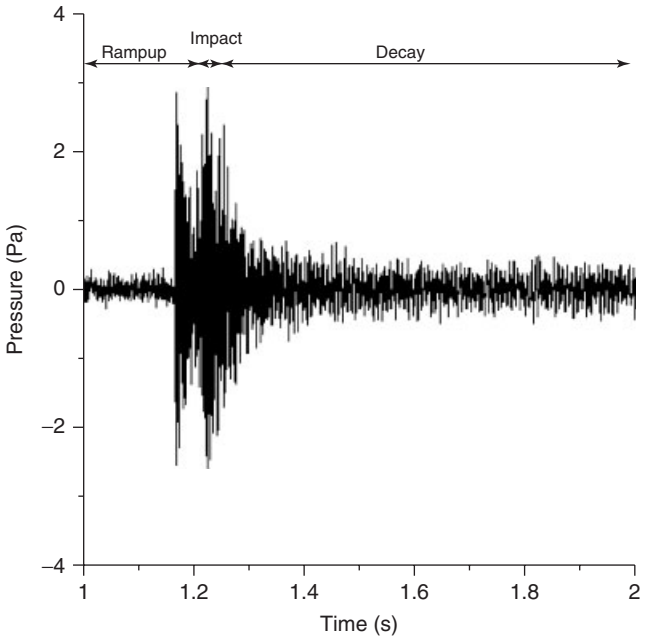


Figure 10. A typical clonk sound signal. (From Gnanakumarr 2010. Reproduced by permission of Woodhead Publishing Limited.)

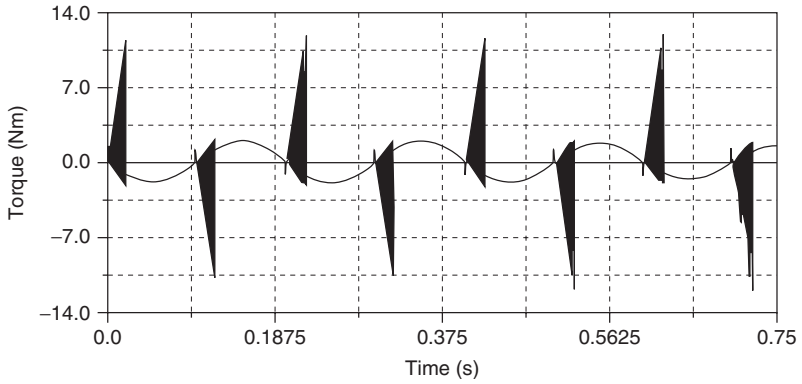


Figure 9. Cycles of shuffle with high frequency impulsive action through zones of low lash rate. (Reproduced from Arrundale *et al*. 1998. © ISATA, 1998.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

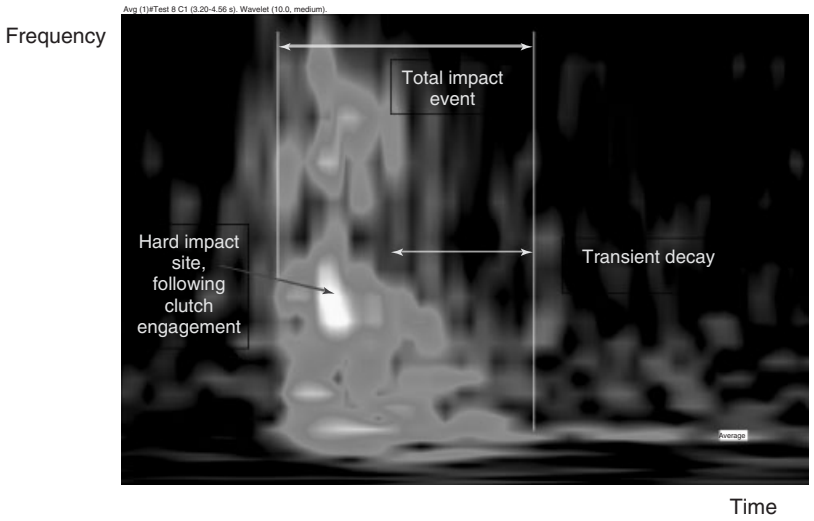


Figure 11. Wavelet of a typical clonk signal.

useful to look at the content of such signals in time frequency domain. Such a representation is known as a *wavelet*. Figure 11 shows a typical wavelet of a clonk vibration signal. The ordinate of the figure is frequency and its abscissa is time. Note that at the instance of impact the signal contains a broadband of frequencies; typically in the range 500–5000 Hz. Thus, the resulting wave front traveling along the drivetrain can excite a large number of modal responses of the system components, particularly those of hollow thin-walled structures, such as the driveshaft tubes that have a dense population of natural frequencies/mode shapes.

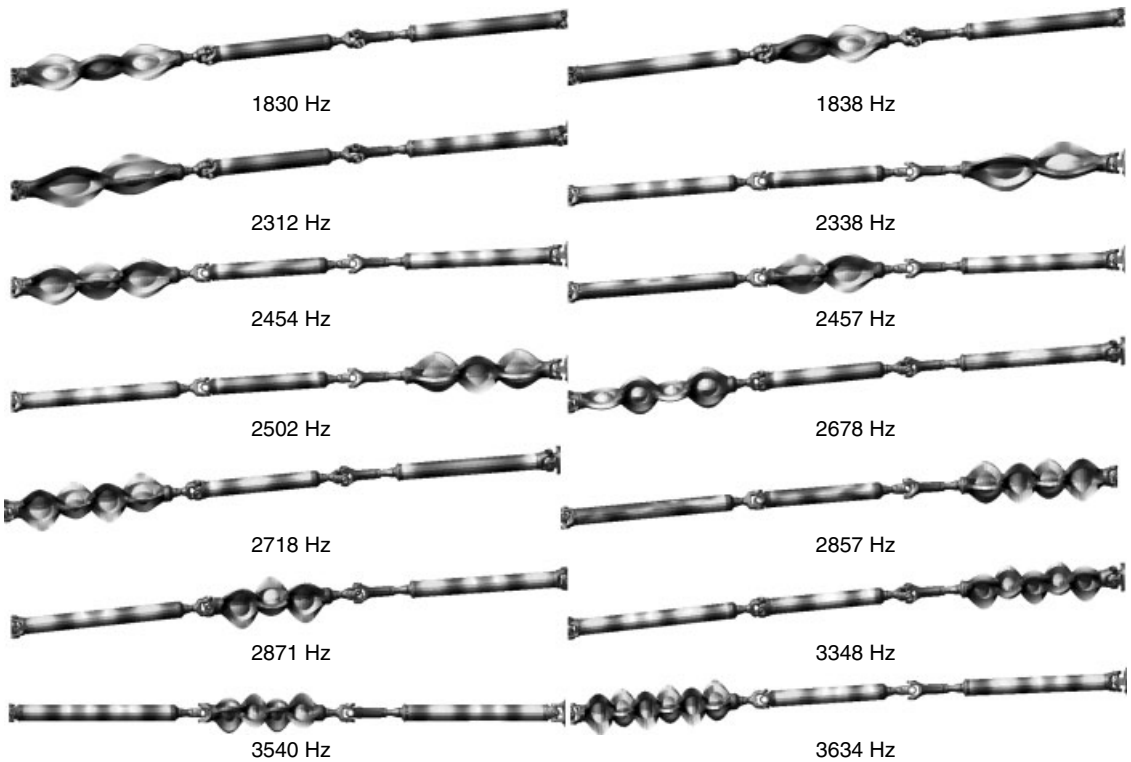


Figure 12. A large number of flexural acoustically active modes (breathing modes). (Reproduced from Theodossiades *et al*. 2004. © Sage/Institution of Mechanical Engineers.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Theodossiades *et al*. (2004) showed that some of the excited structural modes of the driveshaft tubes are efficient noise radiators, giving rise to the clonk noise, which is metallic and accelerative in nature. These are generally the modes that are combined high frequency (many half-wavelength oscillations axially and around a tube's circumference) torsional and bending modes, usually referred to as *flexural modes*. The efficient noise radiating modes among the flexural responses are known as the *breathing modes* (Figure 12). There is also ringing noise during the decay of the clonk response.

There have been traditionally many palliative measures in order to take away the associated sharp metallic nature of the clonk noise, while still maintaining driveline's lightweight. These palliations have included use of sound absorbent media, such as foam filling the tubes, or use of a polymeric sandwich structure. The former can be an irritant substance and in some cases carcinogenic. Thus, its use is now prohibited in many places through directives and/or by legislation. The latter is considered to be expensive for high volume mass manufacture.

Gnanakumarr *et al*. (2006) described various methods of palliation of high frequency sharp metallic clonk response. In particular, they described the use of cardboard inserts fitted into the driveshaft tubes that reduce the sharpness of high frequency metallic noise contributions (above 2500 Hz) as shown in Figure 13 (note that higher noise level is represented by the dark tone in this figure within the lower rectangular region). The cardboard inserts break up the traveling standing structural waves, as well as attenuating the sound propagation by absorption. The secondary events are low frequency contributions because of the decaying ringing noise in the hollow cavity of the tubes.

5 TRANSMISSION RATTLE

Gear rattle is associated with the characteristic structure and airborne noises that are radiated to the environment by the impact of unselected (unengaged) transmission gear pairs (Sakai *et al*., 1981; Wang, Manoj, and Zhao, 2001). The phenomenon occurs at low teeth impact forces and is qualitatively similar to the sound produced by marbles in a shaken metallic can. Therefore, it has a distinct repetitive sound quality differentiating it from other noise NVH concerns (Dogan, Ryborz, and Bertsche, 2006). Rattle is induced by engine order vibrations (Rahnejat, 1998) in the presence of backlash in meshing teeth pairs. It is particularly noticeable in diesel vehicles with higher output torque fluctuations, where the intensity of rattle noise has found to be directly affected by the resulting engine order torsional vibrations (Tangasawi, Theodossiades, and Rahnejat, 2007; Theodossiades, Tangasawi, and Rahnejat, 2007).

Transmission rattle is a broadband NVH phenomenon as perceived by vehicle occupants and other road users alike (Figure 14 shows the response from a bearing cap of a transmission housing). It can cause unnecessary concern for the vehicle owner. Its spectral composition comprises engine order vibration, meshing frequencies of loose gear pairs, and structural modal response of the transmission casing. The higher band frequency due to structural response is because of traveling waves from the impact sites along the transmission output shaft and transmitted though the bearings onto the casing. Transmission rattle manifests itself under various operating conditions. In general, these are

- Idle/neutral rattle: The vehicle is idling and as the name indicates without any gear selected. The usual engine speed is in the range 600–1000 rpm.

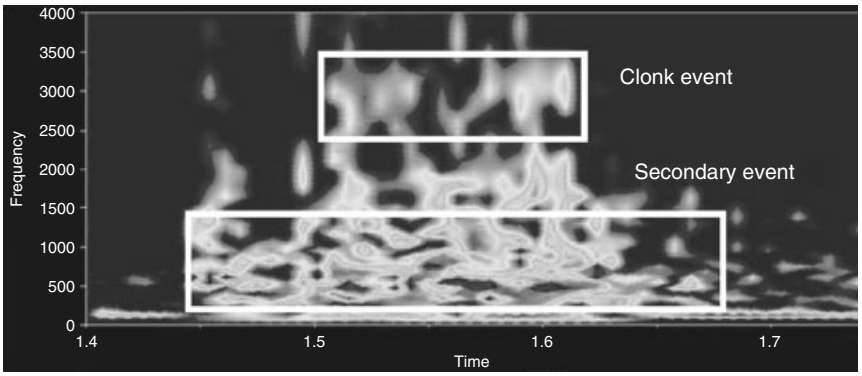


Figure 13. Attenuated clonk noise by cardboard inserts. (Reproduced from Gnanakumarr *et al*. 2006. © Inderscience Enterprises Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

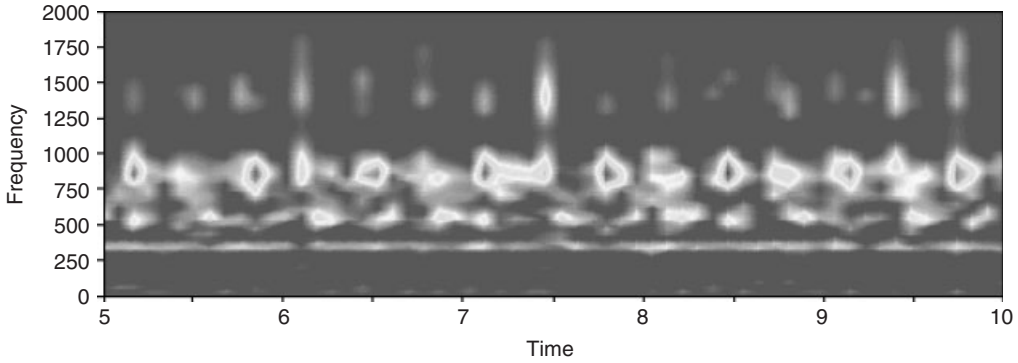


Figure 14. Wavelet of idle gear rattle from 2 L diesel engine vehicle.

- Creep rattle: The engine speed is low (600–1000 rpm) with vehicle driven in first or second gear.
- Drive rattle: The engine speed varies between 1000 and 2000 rpm with second, third, or fourth gear selected and with partially loaded throttle.
- Over-run rattle: The engine speed is between 2000 and 4000 rpm with second gear engaged and throttle at over-run position.
- Coast rattle: With floating throttle and the engine speed between 4000 and 2000 rpm and second gear engaged.

Various methods have been used in order to attenuate the effect of rattle. These include backlash eliminators, clutch predampers, controlled slip clutches, and dual mass flywheel (DMF), the last of which has been found to be the most effective, but a costly palliative.

Essentially, a DMF integrates the function of a usual solid mass flywheel with that of a tuned torsional spring-damper, often in the form of an arc spring (Figure 15). In this configuration, the primary flywheel is bolted onto the engine crankshaft and is free to rotate relative to the secondary flywheel. An arc spring transmits the torque between the primary flywheel and a flange, which is riveted to the secondary flywheel. With a positive load, the arc spring is pushed by the end stop on the primary side against the flange on the secondary side. With a negative load, the arc spring is pushed in the opposite sense by the flange against the end stop. Long travel of the spring at relatively low frequency achieves good isolation, while still maintaining sufficient stiffness to transmit the maximum engine torque. In this manner, spring stiffness can be tuned to act as a band-pass filter, removing or attenuating the

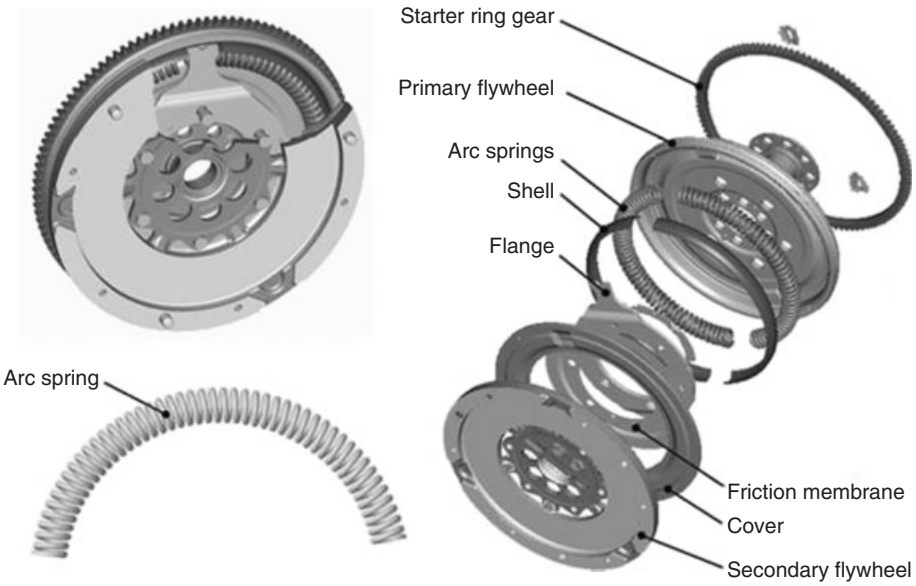


Figure 15. Components of an arc spring dual mass flywheel.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

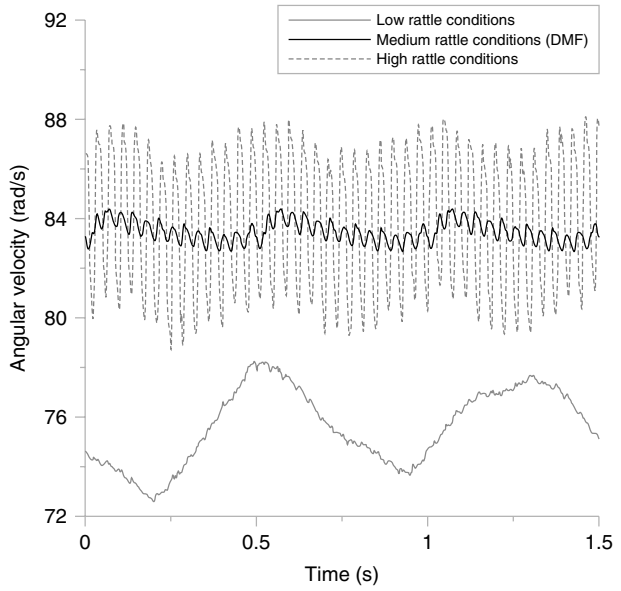


Figure 16. Torsional oscillations resident on the transmission input shaft. (Reproduced from De la Cruz *et al*. 2010. © Institution of Mechanical Engineers.)

significant engine order vibration contributions, depending on the engine type (number of cylinders and combustion signature). Therefore, lower amplitude oscillations are achieved, resident on the transmission input shaft. Note that these engine torsional vibrations are found to be the main contributory factor in transmission rattle (Tangasawi, Theodossiades, and Rahnejat, 2007; Theodossiades, Tangasawi, and Rahnejat, 2007; Dogan, Ryborz, and Bertsche, 2006).

The damping characteristics of the DMF can be further enhanced by attaching additional components to the arc spring, thereby creating friction between the spring coils and the flange. These components are usually referred to as *shoes* with an appropriate wedge shape to entrain grease, which is filled into the DMF enclosure, into the contact. Some vibration is therefore attenuate through viscous shear and drag introduced by the grease. Further description for DMF is given by Kelly *et al*. (2010).

All the palliative methods essentially attenuate the engine torsional vibrations resident on the transmission input shaft, thus reducing the impulsive nature of input between conjugate teeth pairs of unselected gears (Figure 16; De la Cruz, Theodossiades, and Rahnejat, 2010).

In Figure 16, the prevailing rattle condition is obtained from the transmission input shaft under various drivetrain configurations. The response, indicated by medium rattle is when a DMF is used, that for low rattle condition makes use of a clutch slip device as a predamper and the high rattle condition corresponds to no palliation (a solid mass flywheel is used and with no form of palliation). Gear rattle is easily noted and its intensity is often subjectively rated by NVH engineers. However, it is necessary to have an objective function, determining an acceptable level for the rattle noise. To achieve this, it is necessary to undertake quite in-depth numerical analysis, such as those reported by Tangasawi, Theodossiades, and Rahnejat (2007) for idle rattle condition and De la Cruz, Theodossiades, and Rahnejat (2010) for creep rattle. The former uses the *rattle ratio* for its objective function. This is the ratio of inertial torque of an idling gear (loose gear) to that of the resistive drag torque (caused by friction between the loose gear and its retaining output shaft, as well as some small contribution due to flank viscous friction between the gear meshing teeth pairs with their conjugate pinion teeth). Seaman, Johnson, Hamilton (1984) defined the *rattle threshold* as the angular acceleration at which an idle gear's inertial torque exceeds the drag torque (i.e., a rattle ratio exceeding unity). Rattle ratio can be regarded as an indication of rattle, but is not the best metric which can be used. A better ratio is defined by De la Cruz, Theodossiades, and Rahnejat (2010) as the *impulsion ratio*. This is the ratio of drive torque over the drag torque for a loose gear. Therefore, any torque reversal as the result of severe impact conditions can also be ascertained (Figure 17).

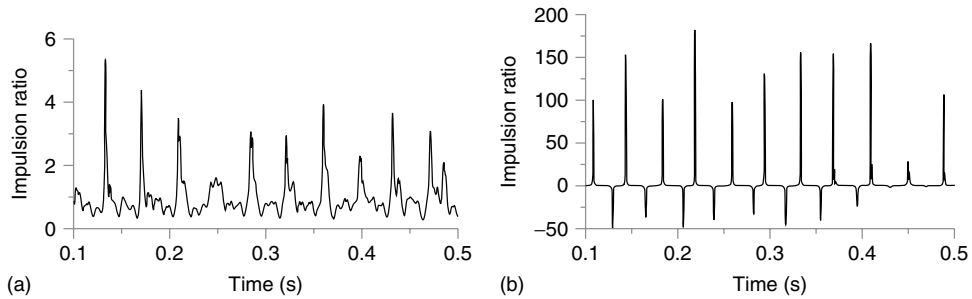


Figure 17. Impulsion ratio corresponding to conditions of (a) low rattle and (b) severe rattle. (Reproduced from De la Cruz *et al*. 2010. © Sage/Institution of Mechanical Engineers.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6 CLOSURE

In this chapter, a number of drivetrain NVH phenomena are described. Some aspects of analysis and palliation measures are also noted. In general, with an increasing trend toward lighter power train construction and higher output power, the propensity for occurrence of these phenomena and others are on the increase. In particular, there would be an increasing effort in power train hybridization as well as development of electric vehicles. These are likely to further introduce NVH issues. As a result, the need for in-depth investigation, particularly numerical analysis, would be increasing in the future.

REFERENCES

Arrundale, D., Hussain, K., Rahnejat, H., and Menday, M.T. (1998) Acoustic response of driveline pieces under impacting loads (Clonk), *Proc. 31st ISATA*, Dusseldorf: 319–331.

Biermann, J.W. and Hagerodt, B. (1999) Investigation into the clonk phenomenon in vehicle transmission—measurement, modelling and simulation. *Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics*, **213** (1), 53–60.

Centea, D., Rahnejat, H., and Menday, M.T. (2001) Non-linear multi-body dynamic analysis for the study of clutch torsional vibrations (judder). *Applied Mathematical Modelling*, **25** (3), 177–192.

De la Cruz, M., Theodossiades, S., and Rahnejat, H. (2010) An investigation of manual transmission drive rattle. *Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics*, **224** (2), 167–181.

Dogan, S.N., Ryborz, J., and Bertsche, B. (2006) Design of low noise manual automotive transmission. *Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics*, **220**, 19–95.

Farshidianfar, A., Ebrahimi, M., Rahnejat, H., and Menday, M.T. (2002) High frequency torsional vibration of vehicular driveline systems in clonk. *International Journal of Heavy Vehicle Systems*, **9** (2), 127–149.

Gnanakumarr, M. (2010) High frequency impact-induced phenomena in driveline clonk, in *Tribology and Dynamics of Engine and Powertrain* (ed. H. Rahnejat) Chapter 30, 914–927, Woodhead Publishing, Oxford-Cambridge-Philadelphia-New Delhi. ISBN 978-84569-361-9.

Gnanakumarr, M., King, P.D., Theodossiades, S., and Rahnejat, H. (2006) Methods of palliation for high frequency elasto-acoustic response of truck drivetrain system. *International Journal of Heavy Vehicle Systems*, **13** (4), 253–262.

Hasebe, T. and Aisin Seiki, U. A. (1993) Experimental study of reduction methods for clutch pedal vibration, *SAE paper Number 932007*.

Kelly, P., Pennec, B., Seebacher, R., *et al.* (2010) Dual mass flywheel as a means of attenuating rattle, Chapter 28: 857–877, in *Tribology and Dynamics of Engine and Powertrain*, Woodhead Publishing ISBN 978-84569-361-9, Oxford-Cambridge-Philadelphia-New Delhi.

Kelly, P., Rahnejat, H., and Biermann, J. W. (1998) Multi-body dynamics investigation of clutch pedal in-cycle vibration (whoop), *Transactions of IMechE Conference on Multi-body Dynamics: New Methods and Applications*: 167–178.

Kim, S.J. and Lee, S.K. (2007) Experimental identification of a gear whine noise in the axle system of a passenger van. *International Journal of Automotive Technology*, **8** (1), 75–82.

Kim, T. and Singh, R. (2001) Dynamic interactions between loaded and unloaded gear pairs under rattle conditions, *SAE technical paper*, 2001-01-1553: 1934–1943.

Koronias, G., Theodossiades, S., Rahnejat, H., and Saunders, T. (2010) Axle whine phenomenon in light trucks: a combined numerical and experimental investigation. *Proceedings of the IMechE, Part D: Journal of Automobile Engineering*, **225** (7), 885–894.

Krenz, R. (1985) Vehicle response to throttle tip in:tip out, *SAE paper No. 850967*.

Kushwaha, M., Gupta, S., Kelly, P., and Rahnejat, H. (2002) Elasto-multi-body dynamics of a multicylinder internal combustion engine. *Proceedings of the IMechE, Part K: Journal of Multi-body Dynamics*, **216** (4), 281–293.

Menday, M.T., Rahnejat, H., and Ebrahimi, M. (1999) Clonk: an onomatopoeic response in torsional impact of automotive drivelines. *Proceedings of the IMechE, Part D: Journal of Automobile Engineering*, **213** (4), 349–357.

Rabeih, E.M.A. and Crolla, D.A. (1996) Intelligent control of clutch judder and shunt phenomena in vehicle drivelines. *International Journal of Vehicle Design*, **17** (3), 318–332.

Rahnejat, H. (1998) *Multi-body Dynamics: Vehicles, Machines and Mechanisms* PEP (IMechE, UK) and, SAE joint publishers, ISBN 0-7680-0269-9, Warrendale, PA, USA.

Sakai, T., Doi, Y., Yamamoto, K., *et al.* (1981) Theoretical and Experimental Analysis of Rattling Noise of Automotive Gearbox, *SAE Technical Paper*, 810773.

Seaman, R, Johnson, C., and Hamilton, R. (1984) Component Inertial Effects on Transmission Design, *SAE Technical Paper*, 841686.

Tangasawi, O., Theodossiades, S., and Rahnejat, H. (2007) Lightly loaded lubricated impacts: idle gear rattle. *Journal of Sound and Vibration*, **308** (3–5), 418–430.

Theodossiades, S., Gnanakumarr, M., Rahnejat, H., and Menday, M. (2004) Mode identication in impact-induced high-frequency vehicular driveline vibrations using an elasto-multi-body dynamics approach. *Proceedings of the IMechE, Part K: Journal of Multi-body Dynamics*, **218**, 81–94.

Theodossiades, S., Tangasawi, O., and Rahnejat, H. (2007) Gear teeth impacts in hydrodynamic conjunctions promoting idle gear rattle. *Journal of Sound and Vibration*, **303** (3–5), 632–658.

Wang, M., Manoj, R., and Zhao, W. (2001) Gear rattle modelling and analysis for automotive manual transmissions. *Proceedings of the IMechE, Part D: Journal of Automobile Engineering*, **215** (2), 241–258.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto096
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Open Differentials

Joseph Palazzolo
GKN Driveline, Auburn Hills, MI, USA

1 Introduction 1
2 Components 1
3 Operating Principle 4
4 Planetary-Style Differentials 6

1 INTRODUCTION

In the traditional drivetrain system for automotive passenger vehicles, there is a need for a device that can divide torque to the driving wheels equally from side to side. This mechanical device is known as a *differential*, specifically referred to as an *open differential* or, at times, a conventional differential. The open differential allows for a single input from the powertrain to be divided to the two outputs, typically the wheels. The open differential still accommodates the requirement of speed differences experienced in turn maneuvers while providing equal torque to the outputs. This equal torque balance of the conventional open differential can create drive force difficulties in the vehicle if one wheel has lost tractive effort as experienced with one wheel on a low coefficient of friction surface such as snow or ice. This is further discussed in Passive and Active Limited Slip Differentials regarding limited slip differentials.

2 COMPONENTS

For the sake of simplicity, this chapter discusses the most common-style axle differential, which is a bevel gear style. The bevel gear-style open differential consists of the following principal components, which are shown in Figure 1:

differential carrier;
differential pin;
differential pinion gears;
differential side gears.

Figure 1 is a two-pinion differential with squareback side gears. The large window area in the differential carrier is to allow for the assembly of the differential internal gears. There is a ring gear mounting flange with mounting holes for the fasteners. These mounting holes can be either clearance holes for the fasteners or threaded holes, in the case of transaxle style mounting structures.

There are other components, such as side gear washers and shims, pinion gear washers and shims, and differential pin retention fastener, but the main components to understand the operation are the four listed earlier.

The torque path through the differential is shown in Figure 2 and the components are further described later. The torque is applied through the ring gear bolted flange to differential case. The flow from the differential case to the wheels is outlined on the right. Progressing downward in the chart is when the engine is driving the wheels and upward is when the wheels are back driving the input (hypoid gear) that occurs during coast conditions.

2.1 Differential carrier

The differential carrier is usually bolted or welded to the input gear. The input gear is typically the hypoid ring gear in a traditional axle arrangement or it is the final drive output gear for transaxle systems. The torque from the powertrain is transferred to the differential carrier via this

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto097
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

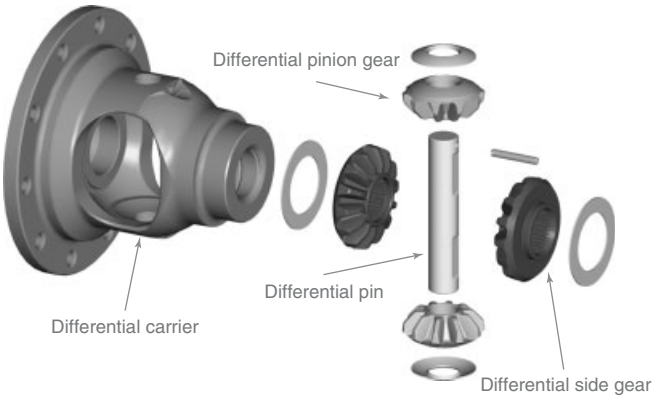


Figure 1. A typical open differential exploded view.

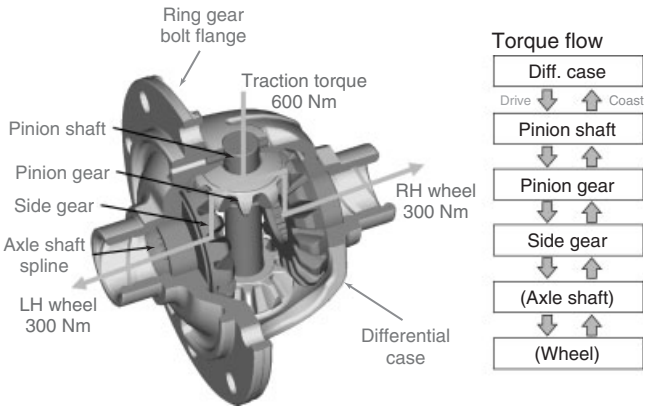


Figure 2. An open differential showing the powerflow.

ring gear and its attachment. The ring gear mounting flange must be rigid enough to support the thrust loads produced from the hypoid gear arrangement, whereas the differential case is supported on a set of bearings to allow it to freely rotate within the axle housing assembly. This is the single source of input torque and speed for the differential. In most applications, the differential case will be a single piece, machined casting with "windows" that allow access for machining and assembly. In other cases, specifically, when there are three- or four-pinion gears, the differential case may be a two-piece design that is split in order to allow machining and assembly of the individual case halves. The differential case is normally a cast iron construction that is machined after casting. Some high performance applications utilize forged aluminum as well for weight savings, but these are very special cases. The differences between single- and two-piece differential housings are illustrated in Figure 3. The two-piece version can be machined and assembled from the opening created when the housing halves are apart. Even though this diagram

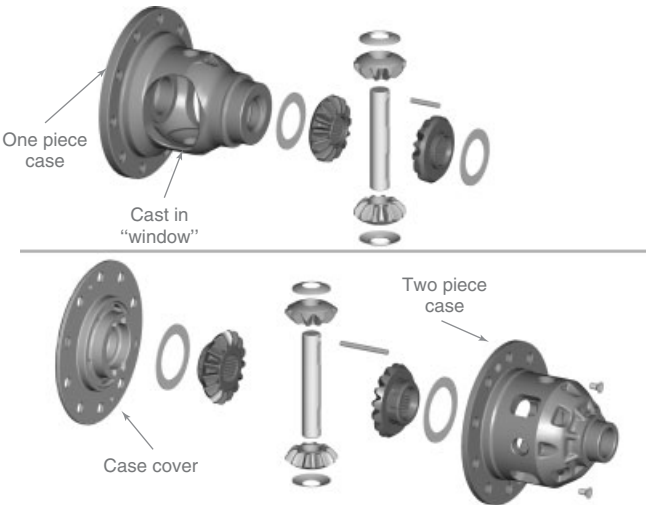


Figure 3. Illustration of a single-piece and a two-piece differential housing.

illustrates two-pinion differentials, the main reason for a two-piece housing is to allow for more than two pinions.

2.2 Differential pin

The differential pin is the next component to receive torque from the powertrain. It is located in the differential carrier and supports the differential pinion gears. Most standard open differentials have a single pin to locate the two-pinion gears. There are some applications that utilize a "T"-shaped, multipiece differential pin to support four-pinion gears. The differential pin not only locates the pinions and transfers torque but, in semi-float axle applications (see Axle Systems for more details), also acts as the thrust reaction member of the axle shafts. In the semi-float arrangement, the differential pin also acts as a mechanical method to retain the axle shaft retention device, which is typically a thick C-shaped steel washer as shown in Figure 4. The left side axle shaft illustrates the C-washer inside the side gear pocket with a clearance on the end of the shaft to the differential pin, whereas the right axle shaft illustrates the axle shaft motion being restricted by the differential pinion shaft. The differential pins are normally made out of heat-treated steel with special coatings to act as a bearing and wear surface for the pinion gears.

2.3 Differential pinion gears

The differential pinion gears ride on the differential pin and transfer power from the pin to the differential side gears. The pinion gears are normally smaller than the side gears.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto097
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

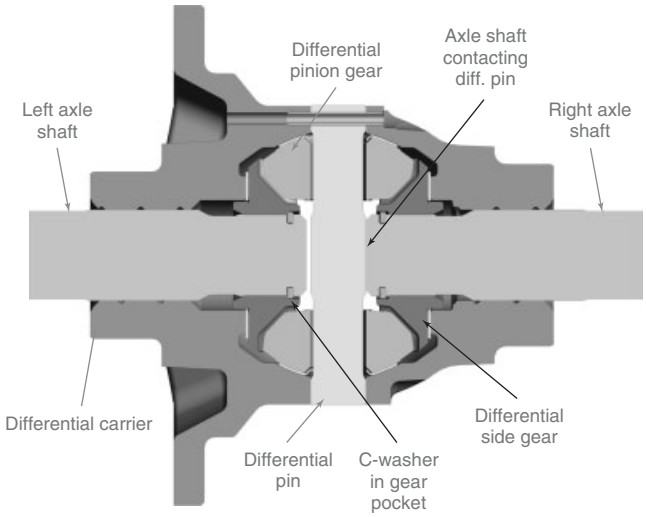


Figure 4. A typical two-pinion open differential.

The normal convention in gearing is to utilize the term *pinion* to describe the smaller of the two gears in mesh and the term *gear* is for the larger of the gear pair. A typical number of teeth a pinion gear is in the range of 8 to 10. The pinion gears mesh with the differential side gears and are perpendicular to the axis of rotation. This change in axis is one of the features of a bevel gear set. This bevel gear tooth arrangement utilized in differentials is a straight side gear geometry as opposed to spiral. The main reason for this is that the side gears and pinion gears in normal, straight-ahead driving conditions do not turn relatively. This is further explained in Equation 2.

The number of pinions is typically two or four, but there are some applications where three are used. Figure 5

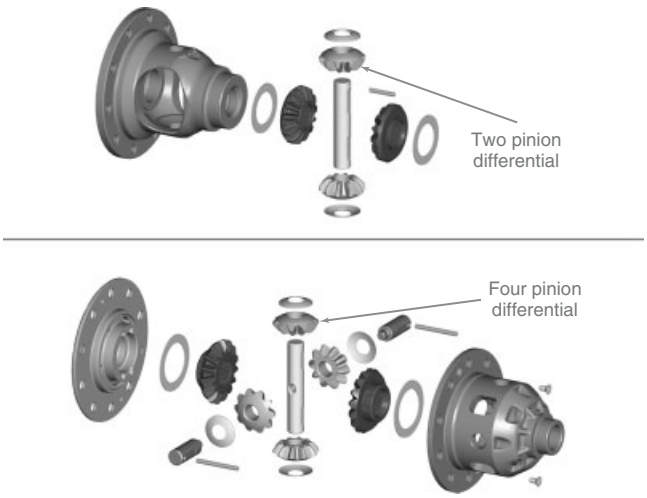


Figure 5. Illustration of differences between two- and four-pinion differential cases.

illustrates the typical differential case differences between a single-piece differential case and a two-piece differential case. The upper exploded view is a typical two-pinion differential case, whereas the lower view is of a four-pinion differential case. The main reason for a two-piece differential case is to allow for more than two pinion gears. The main reason to increase the number of pinions is to increase the torque-carrying capacity within a given packaging environment. Of course, there are manufacturing costs and weight concerns with adding more pinions that need to be considered. Most four-pinion differentials do not allow for "C-washer"-style axle shaft retention as well. In these situations, an outboard axle shaft retention method is utilized. There is a relatively simple equation that can be used to determine a suitable tooth combination.

$$2\,\frac{N_{\text{side gear}}}{N_{\text{p}}} = \text{whole number} \tag{1}$$

where the following symbols are:

$N_{\text{side gear}}$ = number of teeth of the side gear;
N_{p} = number of pinions (typically, 2, 3, or 4).

Equation 1 describes the tooth combination constraints that must be observed in order to assemble a differential with multiple pinions. From this equation, it can be quickly seen that a 14-tooth side gear can be a two- or four-pinion differential but never a three-pinion unit. While a 15-tooth side gear can be either a two or three pinion differential. A typical computer software package that is used to design the actual tooth profiles will adjust the face profiles, such that the gears will operate correctly but the designer must input the correct tooth combination.

2.4 Differential side gears

The other component that makes up the bevel gear set is the differential side gear and is the remaining piece of the torque flow puzzle as shown in Figure 6. This differential has a side gear with a squareback along with pockets for a C-washer-style retention along with spiral grooves machined in the differential housing to aid in lubrication of the axle shaft journal. This gear is the last component within the differential to transfer the torque and speed to the axle shaft. The connection method is typically an internal spline. Side gears normally have a tooth count in the range 11 to 16. There are differential-style bevel side gears that can be of either a squareback or a roundback face. There are many different reasons for each but that level of detail is outside the scope of this text.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto097
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

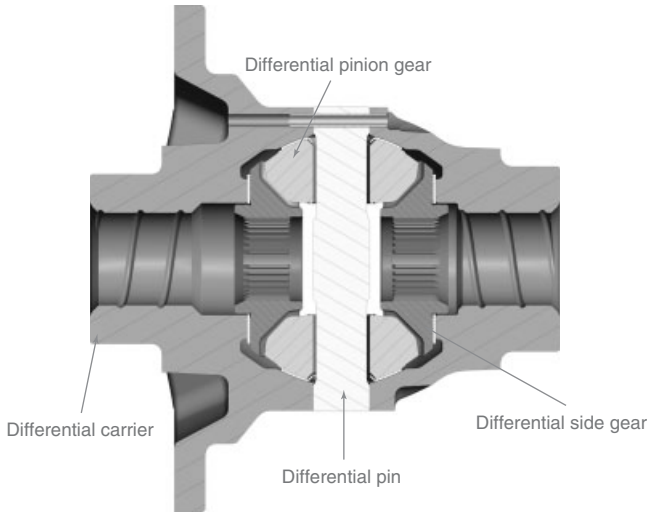


Figure 6. A cross-sectional view of an open differential.

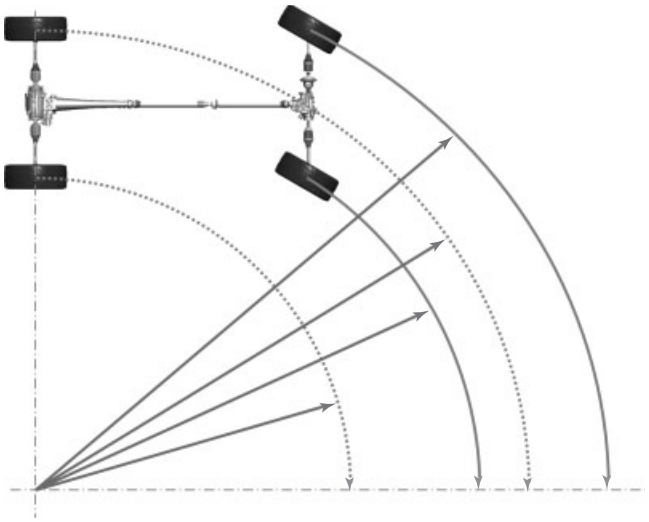


Figure 7. Vehicle cornering.

3 OPERATING PRINCIPLE

The open differential allows for speed differences across the outputs. These potentially varying rates are required to allow the tires to rotate independently during turn maneuvers or unequal tire sizes, pressures, and the like. The primary function of the open differential is to split torque between the two outputs. This provides uniform drive-force distribution across the outputs of the differential.

3.1 Speed relationship

Equation 2 describes the mathematical speed relationship of the outputs (left and right) for a typical axle differential. This simplistic equation works for any units, so one can substitute speed in revolutions per minute, miles per hour, kilometers per hour, and any other units as long as they are consistent. Assuming that a vehicle is traveling in a straight line and the wheels are rotating at the same speed, one can quickly see that the output connections are traveling at the same speed as the inputs.

$$2 \times \omega_{\text{in}} = (\omega_{\text{left}} + \omega_{\text{right}}) \tag{2}$$

where the following symbols are:

ω_{in} = input speed;
ω_{left} = left wheel speed;
ω_{right} = right wheel speed.

Example: 2×10 rpm = (10 rpm + 10 rpm).

Figure 7 shows that all four wheels travel through a different radius of curvature during a turn maneuver. The front differential (right side of the diagram) averages the front wheel speeds, whereas the rear differential (left side of the diagram) averages the rear wheel speeds. As the rear wheels travel in a smaller radius trajectory than the front, there is also a difference in the average of the front and rear speeds.

Even though this seems very simplistic in this form, the function of the differential can be misleading when in a vehicle on a service hoist. Imagine if the input to the differential is held stationary (which is what happens when the vehicle is in park) and rotate one wheel forward—so ω_{in} is zero and ω_{left} is some value, say 10 rpm—then ω_{right} must be negative 10 rpm. Therefore, one wheel spins forward and the opposite wheel will spin in the opposite direction at the same speed. This satisfies the equation and is mechanically a correct behavior of the differential, but can be confusing to the person spinning the tires.

3.2 Torque relationship

The open differential allows for speed differences between the outputs while maintaining a torque balance. Most engineers like to refer to the output torque as a fixed split ratio in percentages. Typically, it is referred to as a *50/50 differential*, as in most conditions, the open differential will split the input torque evenly between the outputs. There are some situations where this torque balance feature of the hardware can cause undesired performance and that the reader may have already experienced in driving maneuvers. Imagine that one wheel is on ice, whereas the other is on the dry road surface. The torque that can be reacted by the slippery ice surface is very low. The differential will

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto097
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

limit this torque to the other wheel as well. This balance means that the wheel with good traction may not have enough torque to propel the vehicle forward. Therefore, the vehicle is now with one wheel spinning and the other idled. Equation 3 illustrates the total available driving in the vehicle.

$$T_{\text{in}} = T_{\text{min}}\{T_{\text{left}}, T_{\text{right}}\} \times 2 \tag{3}$$

where the following symbols are

T_{in} = input torque;
T_{min} = the smaller of the output torques;
T_{left} = left output torque;
T_{right} = right output torque.

Further examination of Equation 3 shows that if one of the output torques is very small, then twice a small number is still a small value. The open differential in this situation is "traction limited" and cannot deliver more torque than can be reacted by the spinning wheel.

One method that can be utilized to visually depict an open differential with one wheel spinning is a simple pulley arrangement with different masses on either end of the rope as shown in Figure 8. The smaller mass represents the wheel on a slippery surface, whereas the larger mass represents the wheel on a good adhesion surface. The open differential has no means to counteract this pictorial mass difference and, therefore, the side with more mass indexes downward freely.

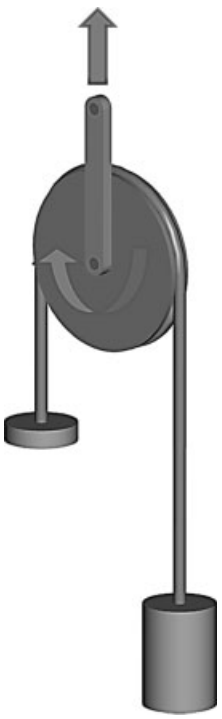


Figure 8. Open differential—pulley analogy.

There are a few methods to counteract this situation. One common method is to increase the torque on the spinning wheel. This can be easily achieved by applying a braking force to that wheel. The driver can easily perform this maneuver by lightly depressing the brake pedal while accelerating. This will create a reaction torque at the spinning wheel that will be *biased* to the opposite via the differential. This braking force can be automatic from the modern-style traction control system that applies brake pulses to the spinning wheels for the driver. Another method is to provide a mechanically or electrically actuated device in parallel across the wheels to combat this torque and speed differences. This is further discussed in Passive and Active Limited Slip Differentials.

3.3 Torque bias ratio

The phrase *biasing torque* across the differential is used to describe the phenomenon when torque is transferred from one of the outputs to the other. There is even a need to refer to the amount of change of that torque from the outputs and often called *torque bias ratio* (*TBR*).

$$\text{TBR} = \frac{T_{\text{high}}}{T_{\text{low}}} \tag{4}$$

Equation 4 quickly begs the question, how can the open differentials have higher torque on one output when compared to the other when the differential is supposed to balance torque? If the open differentials were frictionless and ideally spun freely without any resistance, then the TBR would be unity. The mechanical components do not perform ideally and there is some friction in the system along with inefficiency in the torque transfer. In Passive and Active Limited Slip Differentials, this friction is further discussed as a means to increase vehicle performance.

On the basis of the gear geometry and relationship between the side and the pinion gears along with the interface of the pinions to the differential pin, there are thrust forces exerted between them that create a friction torque during speed difference events. The friction force generated between the side gear and its washer is mathematically expressed in Equation 5.

$$\text{F}_{\text{gw}} = \mu_{\text{sg}} \times \left(\frac{\sin\alpha_{\text{g}} \tan\theta_{\text{g}}}{R_{\text{sg}}}\right) \times \left[\frac{(\text{DG}_{\text{od}} + \text{DG}_{\text{id}})}{4}\right] \tag{5}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto097
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where the symbols are as follows:

F_{gw} = friction between side gear and its washer;
μ_{sg} = coefficient of friction between the side gear and its washer;
α_g = side gear pitch cone angle (°);
θ_g = side gear pressure angle (°);
R_{sg} = side gear pitch circle radius;
DG_{od} = outside diameter of the side gear washer;
DG_{id} = inside diameter of the side gear washer.

A typical value range for the side gear pitch cone angle is 50–60° and the pressure angle is 20–35°.

The side gear thrust force in combination with friction between the side gear back face and the side gear washer creates a resistance to speed difference between these surfaces. A typical value for the coefficient of friction (μ) of these steel components is 0.13. There is also friction present between the differential pinion gear bores and the differential pin shaft along with the differential pinion gears and the pinion gear washers. The friction force between the pinion gear and the differential pin is expressed mathematically in Equation 6.

$$F_{ps} = \mu_{ps} \times \left(\frac{1}{R_{pg}}\right)\left(\frac{D_{ps}}{2}\right) \tag{6}$$

where the symbols are as follows:

F_{ps} = friction between the pinion shaft and the pinion gear bore;
μ_{ps} = coefficient of friction between the differential pin and the pinion bore;
R_{pg} = pitch circle radius of pinion gear;
D_{ps} = pinion shaft's outside diameter.

The last frictional area to represent mathematically is the friction between the pinion gear and its washer.

$$F_{pw} = \mu_{ps} \times \frac{(\tan\theta \cos\alpha_s)}{(R_{pg}\ (\sin\alpha_s + \mu_{ps}\ \cos\alpha_s)} \times \left[\frac{(DP_{od} + DP_{id})}{4}\right] \tag{7}$$

where the symbols are as follows:

F_{pw} = friction between the pinion gear and its washer;
μ_{ps} = coefficient of friction between the pinion gear and its washer;
θ = pinion gear pressure angle;
α_s = side gear pitch cone angle;
DP_{od} = pinion washer's outside diameter;
DP_{id} = pinion washer's inside diameter.

Now that all of the friction components are expressed, their effects on TBR can be analyzed. Each of the friction components has been separately expressed along with its own coefficients of friction. More often than not, the coefficient of friction is same for all the interfaces but one could, theoretically, design different frictional interfaces to adjust the differential behavior. Equation 8 describes the TBR in terms of the frictional force relationships described earlier.

$$TBR = \frac{(1 + F_{gw})[1 + (F_{ps} + F_{pw})]}{(1 - F_{gw})[1 - (F_{ps} + F_{pw})]} \tag{8}$$

4 PLANETARY-STYLE DIFFERENTIALS

As mentioned earlier, bevel-style differentials are the most common for side-to-side even torque split differentials applications. It is also possible to achieve the same mechanical properties utilizing planetary gear arrangements. The main difference with planetary arrangements is that they are typically narrower and larger in diameter when compared to bevel-style differentials. Just like bevel differentials, planetary differentials utilize straight cut or spur gears. There are three main elements of the planetary—the ring gear, the sun gear, and the planet carrier. The ring gear has internal spur gear teeth, whereas the sun gear has external spur gear teeth. The carrier locates the planets and allows them to spin independent of the carrier—much like the bevel-style pinion spinning on the differential pin. In order to achieve the torque split requirements, an idler planet is inserted to maintain the correct direction of rotation during speed difference maneuvers. The additional planet in the mesh reverses the direction of rotation and is often referred to as a *compound planetary arrangement*.

The torque ratio is balanced as long as the number of teeth on the sun gear is half the number of teeth on the ring gear. The input is the ring gear, whereas the sun gear and the carrier gear are the outputs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto097
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Passive and Active Limited Slip Differentials

Joseph Palazzolo
GKN Driveline, Auburn Hills, MI, USA

1 Introduction 1
2 Passive Limited Slip Differentials 2
3 Active Limited Slip Differentials 12
4 Summary 15
References 15
Further Reading 15

1 INTRODUCTION

On the basis of the traction limited behavior of an open differential, as described in Basic Open Differentials, there has been a need to develop a differential device that will overcome the shortcomings of the open differential. This device is referred to as a *limited slip differential* (*LSD*). At the same time, it is preferred for the LSD to mimic the open differential performance in normal driving conditions and only intervene during excessive wheel slip events. There are a numerous types and arrangements of LSDs, each having its own unique performance characteristics. These characteristics can be considered advantageous and at times detrimental to the vehicle performance or behavior. One must keep in mind that the main reasons for a type of LSD can be to overcome the distinctive traction limited characteristic of an open differential or to deliver a higher performing vehicle driving behavior. With so many variants of these devices in the market, it is best to categorize them as two basic types: active and passive. The active devices have some sort of electronic control system typically with feedback, whereas the passive devices respond purely mechanically. The different types of devices that are in either the production vehicles or the most commonly encountered vehicles in the aftermarket will be reviewed in this chapter. Please keep in mind that there are many devices that have a similar function under different brand names in the aftermarket. The basic principles are the same and will be described here.

The LSD is unique, with two potential torque paths through the device. The primary torque path is through the open differential. Therefore, the torque is from the ring gear flange to the differential pin, then to the differential pinion gears, and finally the side gears as reviewed in Basic Open Differentials. The secondary path is utilized when the open differential portion alone cannot transfer the power based on the traction limited behavior of the open differential. The speed or torque difference within the differential, which is caused by a slip event, requires additional means to transfer the power, which is the limited slip device. During this event, the secondary path typically flows through a clutch pack to the differential gears. The key differentiator in the various style devices is the means utilized to engage the clutch pack. The limited slip mechanism is reacting to the speed difference from one output of the differential to the other. The mechanism will always try to transfer torque from the faster spinning wheel to the slower spinning wheel in order to equalize the speeds across the outputs. The device is basically diverting speed to the slower wheel in an effort to balance the output speeds. This explains why the phrase "limited slip" is used to describe the devices.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 PASSIVE LIMITED SLIP DIFFERENTIALS

The term *passive* refers to the fact that the devices in this category are differentials that are purely mechanically engaged without any electrical feedback from the vehicle or driver directly. There are basically two styles of mechanical self-actuation methods employed for passive style LSDs. These differentials respond to either a speed or a torque difference from the outputs of the differential or at times, a combination of both. As such, these units are commonly referred to as *speed sensing* or *torque sensing*. The term *sensing* can be a little misleading, as these devices do not have any electrical sensors but rather mechanically sense a speed or torque difference. On the basis of the speed or torque difference, the LSD will react accordingly to a typically predetermined performance characteristic.

As these units are purely mechanical devices, they can cause some other vehicle system concerns based on this interaction. Imagine that there is an event in the vehicle where a speed difference is preferred (such as an antilock brake system (ABS) event), and the passive LSD will not allow that speed difference to occur freely. In this situation, there can be a system characterization issue and additional ABS tuning involved in order to calibrate for this interaction. It is possible to have a passive LSD function with ABS or stability control; the development and vehicle tuning engineers just need to understand the mechanical behavior of the differential and adjust the brake system logic accordingly. With the proper attention allotted to the total system interaction, both systems, ABS and passive LSD, can work in harmony.

Through examination of Figure 1, it is evident that there are two distinct potential paths of torque flow through the differential. The normal path is through the traditional open differential style gears. There is an alternative path through the limited slip device that governs the unit when the bevel gear arrangement is no longer adequately controlling the slip across the unit. This parallel path of power is where the different style limited slip devices are installed. A generic review of the more common styles of devices in the market is included later in this section.

2.1 Limited slip differential construction

There are two means to mechanically link an LSD device. It is common to describe the performance of the LSD by referencing the two outputs of the differential, which are the wheels for an axle application. This common convention is a great method to describe the generic function, but skips some of the important mechanical behavior that is required to design and analyze the limited slip device.

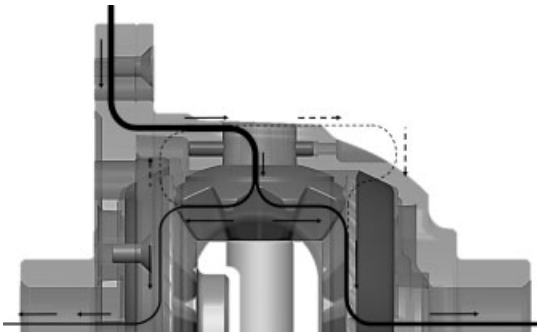


Figure 1. A phantom view of a typical limited slip differential.

If the LSD device is truly connected across the outputs of the differential, it is referred to as a *shaft-to-shaft style LSD*. Even though this is the most common method used to pictorially describe the function, it is not very common in actual hardware application. The vast majority of production vehicles do not use this style construction, as it is mechanically difficult to link all of the components together and adds more complexity with additional concentric shafts. There are still some basic concepts that need to be reviewed with this style architecture and will help explain the alternate, more common, construction method of shaft-to-housing. For illustrative purposes, it is assumed that the vehicle is traveling in a right-hand turn maneuver. In this event, the speed of the left wheel is greater than the speed of the right wheel, $\omega_L > \omega_R$. The kinematics is reviewed in Basic Open Differentials. The LSD device will transfer torque from the faster spinning left wheel to the slower spinning right wheel. The device will basically try to match the speed across the differential's outputs by diverting the torque. As the bevel gear style differential will split the torque equally and the LSD device will add torque to the slower right side wheel, the following equations describe the torque potential in the system.

$$T_L = \frac{T_{in}}{2} - T_c \tag{1}$$

$$T_R = \frac{T_{in}}{2} + T_c \tag{2}$$

where the following symbols are

T_L = torque of the left wheel
T_R = torque of the right wheel
T_{in} = input torque
T_c = torque of the LSD clutch

If a left-hand turn maneuver is experienced, the right wheel is the faster spinning wheel and the above-mentioned

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

equations are similar, just needing to exchange the signs of operations, adding LSD torque to the left wheel and subtracting for the right wheel.

The other mechanical method to configure the LSD and the most common is actually with the LSD clutch device mechanically connected between the side gear and the differential case. With this configuration, the above-mentioned equations need to be slightly modified but still govern the torque transfer. In this case, the substitution of $T_c/2$ in place of T_c yields the following: Right-hand turn ($\omega_L > \omega_R$):

$$T_L = \frac{T_{in}}{2} - \frac{T_c}{2} \quad (3)$$

$$T_R = \frac{T_{in}}{2} + \frac{T_c}{2} \quad (4)$$

As discussed earlier, for the shaft-to-shaft configuration, exchanging the signs of operation for a left-hand turn and these equations describe the torque transferred. Even though the more common architecture, with the LSD between the side gear and the differential carrier, is easier mechanically to design, there is a trade-off in the amount of LSD clutch torque transferred, which is half of what it could be for a given size clutch. This just means that the LSD device needs to be sized accordingly for the configuration. This is one of the main reasons that the LSD clutch pack is often two separate packs behind each of the differential side gears.

2.2 Preload

Preload is a term used to describe a fixed torque that must be overcome before the differential freely allows a speed difference. Typically, preload is achieved from a mechanical spring. In Figure 2, a coil spring is shown to represent preload. On the basis of the $T_c/2$ reviewed earlier, there are two clutch packs in the device. Another method to explain preload is that there will be a fixed torque between both outputs up to the preload value independent of road surface conditions. This fixed torque is basically a torque that must be overcome before the differential unlocks, which will allow a speed difference to occur. If a speed difference is required and the torque is below the preload value, then the differential will mimic the performance of a mechanically locked differential, which will not allow speed difference. Typically, the mechanical torque preload is in the 50–150 N-m range before the LSD will allow differentiation within the unit. Preload is a great feature to aid with traction and handling but needs to be integrated with care. If too much preload is in the system, then the tires may not be allowed to easily differentiate in turn maneuvers at low speeds. The tires overcoming this excessive preload during a turn event may cause wind-up across the differential, and the driver will surely notice this driveline wind-up. Conversely, at high speeds, there may be a positive yaw-damping effect associated with higher preload. The ideal condition would be to have the flexibility to vary preload torque based on the vehicle speed and turn events. Figure 3 shows the torque transferred from the outputs of the differential. The solid line shows the performance of a traditional open differential; as the open differential typically has a 50/50 torque split output, the line has a slope of 45°. The effect of preload shifts the curves by the amount of the preload and is shown with dashed lines, but the slope of the curve is constant. The upper right quadrant of the graph represents a drive condition with left and right turns, whereas the lower left quadrant represents a coast condition with left and right turns.

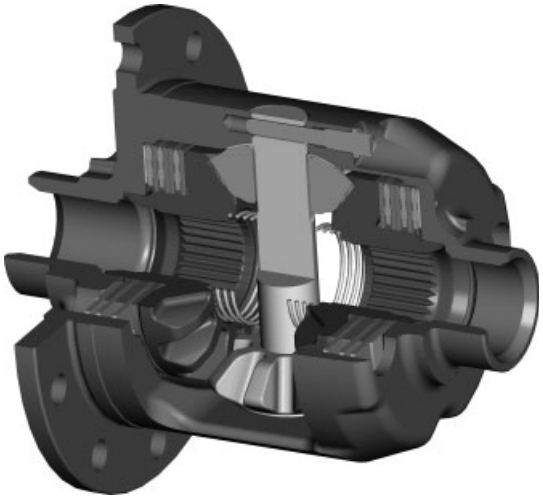


Figure 2. A model of a typical limited slip differential.

2.3 Speed-sensing differentials

As the name suggests, a speed-sensing LSD is a device that responds to excessive wheel speed difference across the unit. As the outputs of the differential are mechanically attached to the wheels, the device is responding to excessive tire slip or speed differences of the wheels. There is a characteristic locking torque transferred across the differential in relation to this speed difference as shown in Figure 4. The horizontal axis of this graph is the speed difference, whereas the vertical axis is torque transferred. There are two curves plotted showing the difference between a digressive (shown as a dashed line) and progressive locking (shown as a solid line) behaviors of the differential. One needs to be aware that the speed difference is across the device and not absolute vehicle speed that is being referenced.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

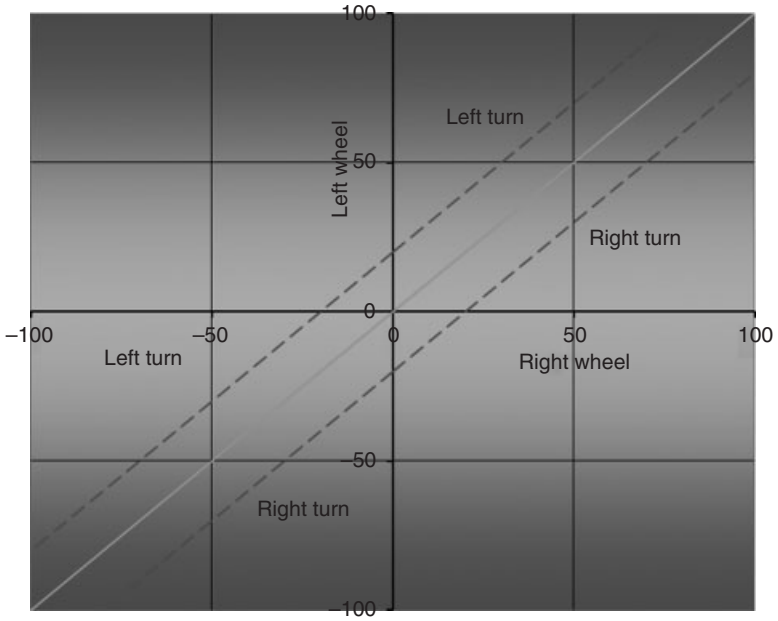


Figure 3. Preload graph. (Reproduced by permission of Joe Palazzolo.)

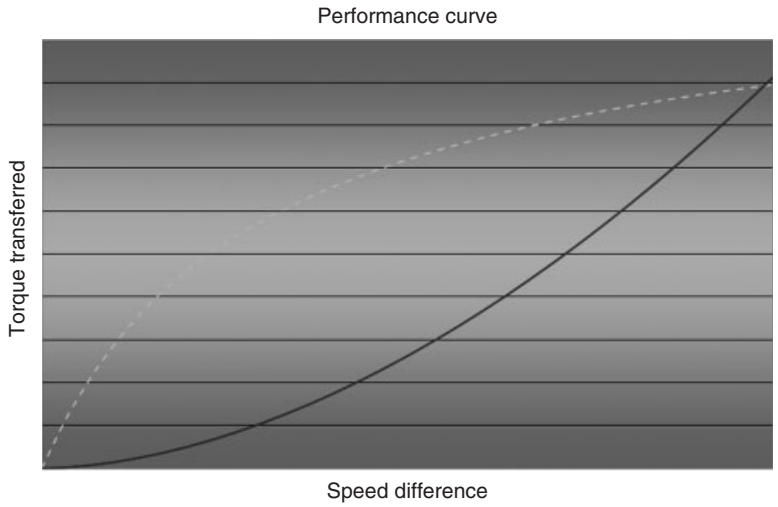


Figure 4. Performance of a speed-sensing differential. (Reproduced by permission of Joe Palazzolo.)

Most speed-sensing LSDs employ some sort of pump device that reacts to the speed difference to allow the pumping device to generate pressure. This pressure is then applied to a piston that applies a compressive force to a clutch pack. By the nature of the pump device, this style LSD is reactive to wheel slip after the slip has occurred. Not all of the devices employ a clutch pack, but it is most common to have a clutch pack. The different technologies will be reviewed in the following sections.

2.3.1 Viscous control technology

The viscous control technology uses fluid friction in order to transfer torque during slip events across the differential in a digressive performance curve as shown earlier. The Harry Ferguson Limited Research Department first discovered and applied the fluid dynamics properties of a unique fluid that resists motion when it experiences shear stress. In the absence of any shear stress, the fluid returns to it normal state. The fluid that is utilized is silicon based with a typical

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

kinematic viscosity range between 50,000 and 500,000 centistokes. The silicone fluid has very stable physical properties over a wide range of operating temperatures, especially when compared to traditional mineral-based oils. The fluid used for viscous control devices is not shared with the rest of the differential and is enclosed in a separate sealed housing. These units are filled for life and the fluid, therefore, does not require any service interval over the lifetime of the vehicle. In this sealed chamber, there are two sets of steel plates that are mechanically attached to the two elements of the differential. When there is a speed difference across the differential, these steel plates resist the motion through the fluid dynamics of the silicone fluid. This resistance from the fluid shear creates a torque across the plates and, therefore, the slip speed difference is minimized based as a function of the fluid viscosity, fluid fill amount, temperature, and slip speed. Figure 5 is a cutaway of a typical viscous, passive LSD by GKN. Notice the separately sealed chamber and viscous plates on the right side of the differential. As the device does not rely on mechanical friction and also requires a separate sealed chamber, the biasing device is nonsymmetrical.



Figure 5. Viscous differential style LSD.

2.3.1.1 Viscous coupling. The viscous coupling style control devices can operate in two potential modes during speed difference events. The first mode is normal viscous operation and follows a digressive curve, as shown in Figure 4. During slip events, that device will transfer torque per the characteristic curve and the amount of torque delivered is adequate for most situations. There can be a second mode of operation that is technically called *self-induced torque amplification* or more commonly referred to as *hump mode*. The term *hump mode* refers specifically to the fact that the shape of the torque transferred versus speed difference curve resembles a camel hump. As shown in Figure 6, the torque transferred is on the vertical axis with absolute time on the horizontal axis. The slip speed is held constant during this event. After several seconds, the torque-transferred curve sharply increases and transferred full torque in order to overcome the vehicle situation. This "hump mode" is actually an overheating protection mechanism to shelter the viscous device from excessive heat and wear. This increase in torque allows the vehicle to overcome the slip condition that requires additional torque when compared to the first mode of operation. The viscous coupling is a device that has a fixed volume of fluid and operating volume.

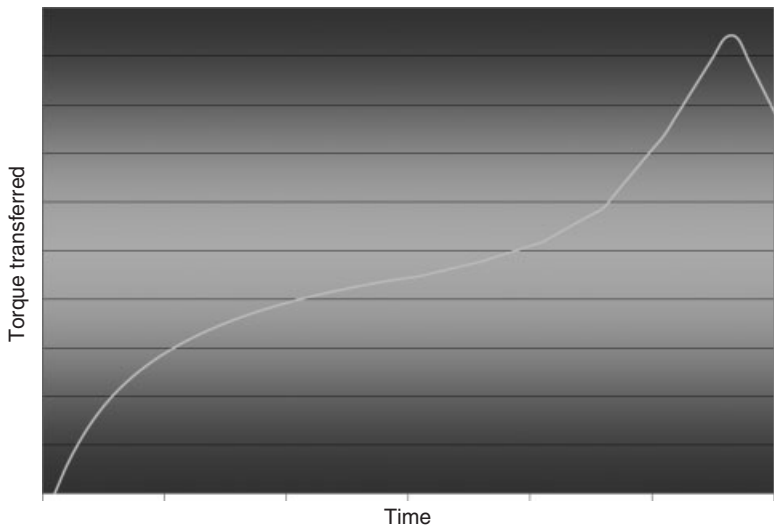


Figure 6. Viscous "hump mode" graph. (Reproduced by permission of Joe Palazzolo.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 7. Cutaway of a ViscoLok differential.



Figure 8. The hydraulic pump limited slip differential. (Reproduced by permission of Joe Palazzolo.)

2.3.1.2 ViscoLok technology. The term *ViscoLok* refers to a unique fluid shear pump that utilizes silicone fluid as the pumping fluid. The main concept is similar to that of the viscous coupling but there is a significant difference. This unit has a single pump mechanism that shears the fluid during speed difference events. For this device, the fluid volume is not fixed as compared to the viscous coupling. As the fluid pump shears the fluid, the temperature and pressure increase. This increase in pressure is applied to a piston that, in turn, applies a wet clutch pack. This allows for a higher torque transferred in a given package volume and utilizes traditional clutch plate technology with a unique speed-sensing application. As seen in Figure 7, the viscous shear pump is the plastic disk on the left of the piston. The pump operates in a separate chamber and the piston reacts to the increased pressure to apply the multiplate clutch pack.

2.3.2 Hydraulic pump technology

Another common speed-sensing LSD is configured with a hydraulic pump mechanism in combination with a clutch pack. In this style device, the hydraulic pump creates the fluid flow based on the speed difference across the device. The most common style pump is a gerotor style arrangement as shown in Figure 8 that is driven off of the speed difference of the wheels. The pump pressure applies a clutch pack to provide the limited slip functionality. The inner and outer rotors can be seen on the left of the differential and share common oil with the differential. This fluid flow is directed to a hydraulic piston that in turn applies a clutch pack. This style device is said to have a progressive locking characteristic based on the shape of the torque transferred versus speed difference performance curve. There is even a variant of this device that has integrated electronically controlled valve that regulates the apply pressure on the clutch pack. The hardware architecture is the same as the mechanical device with the addition of the pressure control valve.

2.3.3 Flyweight technology

This multiplate LSD technology is unique when compared to the others discussed earlier, in that the engagement mechanism is triggered through speed difference but makes use of centrifugal flyweight technology. The flyweights are triggered to centrifugally advance at approximately 100 rev/min speed difference. The flyweight engages a cam ramp mechanism that is located behind one of the side gears, which applies force to a clutch pack. The two engagement flyweights are oriented such that they allow the device to engage in either forward or reverse direction of travel of the vehicle. The clutch pack is located between the cam plate and the differential housing. There is also a reaction block in between the side gears to transfer the load to the opposite gear. This indirect force transfer compresses the additional clutch pack on the opposite side of the LSD for added torque-carrying capacity. This LSD also has a supplementary flyweight mechanism that utilizes centrifugal force again to temporarily bypass the LSD function above speeds of approximately 20 mile/h (32 km/h). This is a unique feature, which allows for LSD functionality at lower vehicle speeds and the unit mimics the performance of an open differential at speeds above 20 mile/h. On the basis of this speed-dependent functionality and the fact that the LSD function is designed to be restricted based on vehicle, some engineers will refer

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 9. The governor lock limited slip differential. (Reproduced by permission of Joe Palazzolo.)



Figure 10. Cone-style limited slip differential.

to this device as the governor lock or gov-lock for short. As shown in Figure 9, the cylindrical mechanism of the flyweight-style LSD is the engagement device, whereas the longer more oval-shaped device above the cylinder is the governor mechanism that disables the limited slip functionality at higher speeds.

2.4 Torque sensing

The term *torque sensing* refers to the fact that this classification of LSD devices responds to torque differences across the outputs of the differential. These devices are also influenced by not only the speed across the outputs but also the internal torque reaction and separating forces within the bevel differential side gears. There are many different styles of devices in the category and the technologies will be reviewed with some of the governing torque-transfer characteristics.

2.4.1 Cone clutch-style friction technology

The basic function of the cone-style LSD is to mechanically limit the speed difference exhibited across the outputs through a tapered ring and mechanical friction. The side gear separating forces during speed difference events are applied to the tapered ring. The tapered ring acts as an energy-absorbing device and resists this speed difference motion to yield the limited slip effect on the differential. Figure 10 shows the typical cone-style LSD with the tapered rings between the differential side gears and housing. There is also a coil-type preload spring installed between the side gears.

The bias torque or locking torque calculation method is similar to that of an open differential with the addition of the friction clutch from the tapered clutch rings. Equations 6 and 7 from the open differential (see Basic Open Differentials) describe the frictional force generated from the interface between the pinion shaft and bore and the pinion gear and its washer, respectively. Those equations are repeated below for reference:

$$F_{ps} = \mu_{ps}\left(\frac{1}{R_{pg}}\right)\left(\frac{D_{ps}}{2}\right) \tag{5}$$

where the symbols are as follows:

F_{ps} = friction between the pinion shaft and the pinion gear bore;
μ_{ps} = coefficient of friction between the differential pin and the pinion bore;
R_{pg} = pitch circle radius of pinion gear;
D_{ps} = pinion shaft outside diameter.

$$F_{pw} = \mu_{ps}\frac{(\tan\theta\ \cos\alpha_s)}{[R_{pg}(\sin\alpha_s + \mu_{ps}\ \cos\alpha_s)]}\left[\frac{(DP_{od} + DP_{id})}{4}\right] \tag{6}$$

where the symbols are as follows:

F_{pw} = friction between the pinion gear and its washer;
μ_{ps} = coefficient of friction between the pinion gear and its washer;
θ = pinion gear pressure angle;
α_s = side gear pitch cone angle;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

DP_{od} = pinion washer outside diameter;
DP_{id} = pinion washer inside diameter.

The additional frictional force that needs to be accounted for is the force resulting from the side gear and the tapered friction ring. The equation is very similar to Equation 5 from the open differential (see Basic Open Differentials) with the additional friction from the tapered clutch. The new equation is mathematically described as follows:

$$F_c = \mu_c \left(\frac{\sin \alpha_g \ \tan \theta_g}{R_{sg} \ \sin \beta_c} \right) \left[\frac{(DG_{od} + DG_{id})}{4} \right] \quad (7)$$

where the symbols are as follows:

F_c = friction between the side gear and the tapered clutch;
μ_c = coefficient of friction between the side gear and the tapered clutch;
α_g = side gear pitch cone angle (°);
θ_g = side gear pressure angle (°);
R_{sg} = side gear pitch circle radius;
β_c = tapered ring angle (°);
DG_{od} = outside diameter of the side gear washer;
DG_{id} = inside diameter of the side gear washer.

With the above equations developed, the torque bias ratio can be described in terms of the friction forces as the following:

$$TBR = \frac{(1 + F_c)}{[1 + (F_{ps} + F_{pw})]}(1 - F_c)[1 - (F_{ps} + F_{pw})] \quad (8)$$

There is an added benefit of the addition of a clutch device in an LSD in the ability to preload the clutch. Typically, the preload is achieved using a mechanical spring element in the system that applies a known load to the friction surface. This known load acts to lock the differential up to the point of its torque capacity from preload. This is analogous to the force required to impart motion to a block on a surface; the effect of preload is increasing the weight of the block. In order to understand the initial preload, the resulting torque from the interfaces is described as follows:

$$T_{tc} = \left(\frac{F_k}{\sin \beta_c} \right) \left[\frac{(\varphi_{od} + \varphi_{id})}{4000} \right] (\mu_c) \quad (9)$$

where the new symbols are as follows:

T_{tc} = friction torque of a single cone clutch (N-m);
F_k = preload spring force at assembly (N);
φ_{od} = friction clutch outer diameter (mm);
φ_{id} = friction clutch inner diameter (mm).

$$T_{cg} = F_k \left[\frac{(\gamma_{od} + \gamma_{id})}{4000} \right] (\mu_{gc}) \quad (10)$$

where the new symbols are as follows:

T_{cg} = friction torque of a single side gear and the cone clutch (N-m);
γ_{od} = outside diameter of the friction surface of the retainer (mm);
γ_{id} = inside diameter of the friction surface of the retainer (mm);
μ_{gc} = coefficient of friction between the side gear and the cone clutch retainer.

The clutch torque can now be described as follows:

$$T_c = 2T_{tc} + 2T_{cg} \quad (11)$$

Finally, the initial preload torque is:

$$T_i = T_c(1 + TBR) \quad (12)$$

2.4.2 *Multiplate friction technology*

Multiplate LSDs are called such, as they utilize a clutch pack that consists of a series of multiple friction clutch and steel reaction plates in order to provide the limited slip functionality. There are many different types of friction materials that are used and each has its advantages and disadvantages. These materials can be sintered bronze, bonded paper, carbon fiber, and even molybdenum coatings to name a few. Just like the cone-style device, these style units react to the separation force of the bevel gears from the torque in the system along with the speed difference across the bevel gears.

There are two broad categories with further subcategories within each shown in Figure 11. TL stands for Traction-Lok, whereas LOM stands for Lok-O-Matic.

2.4.2.1 External pressure. The first category is the external pressure style where the preload apply force is developed external of the bevel differential gear set. The spring is between the differential housing and the back face of the side gears. The preload torque relationship is given in Equation 13 and this is a trapped torque across the differential. When this initial preload torque is overcome in the vehicle, the reaction torque of the mechanical device will govern the performance characteristic of the unit.

$$T_c = (r_2)\,\mu\,(F_k) \left(\frac{n}{1000} \right) \quad (13)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

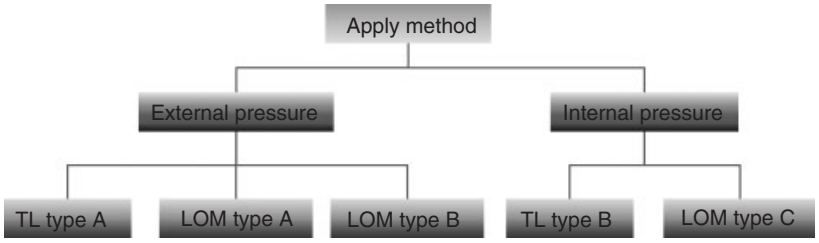


Figure 11. A multiplate LSD summary chart.

where the new symbols are as follows:

$r_2 = \left[\frac{(RC_{od}+RC_{id})}{2}\right]$ = friction radius of the clutch plate (mm);
F_k = total spring force on one side of LSD (N)
RC_{od} = friction clutch outside radius (mm);
RC_{id} = friction clutch inside radius (mm);
n = number of friction surfaces on one side of the LSD.

In addition, the preload or initial torque is simply twice the clutch torque.

$$T_i = 2T_c \tag{14}$$

This initial torque applies to both the external pressure-style TL and LOM units.

2.4.2.2 Internal pressure. The internal pressure-style units have the preload spring packaged mechanically in between the differential side gears. The initial or preload torque is described in Equation 15 and now introduces a new term *transfer ratio*.

$$T_i = T_c(1 + R'_t) \tag{15}$$

where R'_t is the transfer ratio and defined accordingly in the following sections for the TL and LOM-style devices. The clutch torque is the same as Equation 13.

2.4.2.3 Trac-Lok style with preload. The traction lock-style LSD is often abbreviated as Trac-Lok or just TL. It consists of a traditional-style open differential, bevel gear arrangement with a series of flat clutch, and reaction plates between the differential side gears and the differential housing. As this arrangement reuses much of the open differential geometry, it is the most common style in production. The clutch plates are typically preloaded mechanically with a spring. This preload can be from either Belleville or helical coil-style compression springs between the clutch pack and the side gears. This is referred to as *external* pressure, as the preload force is exerted from outside of the differential gears to the clutch pack. There is also an internal pressure style where the preload force and spring is mechanically in between the differential gears. The internal style arrangement is the most common in production with a couple of minor variants on spring style, mainly a series of compression springs or an s-shaped spring.

2.4.2.3.1 External pressure-style (TL type A). The TL style with external pressure for torque transfer has a preload torque as described earlier along with an additional clutch pack force based on the bevel gear differential specific geometry. This additional force happens when the side gears are moving relative to one another. In Figure 12, there is a typical TL-style LSD, such that the preload of the clutch pack is achieved typically from Belleville-style spring washer between the differential housing and the clutch pack represented by the arrows. This is referred to as *external* pressure, as the apply force is generated outside of the differential gears.

The locking ratio percentage is based on the gear geometry as follows:

$$f_1 = (r_2)\,\mu \left(\frac{\tan\theta_g \sin\alpha_g}{r_1}\right) n \tag{16}$$

where r_1 is the operating radius of the side gear (mm) and expressed as follows:

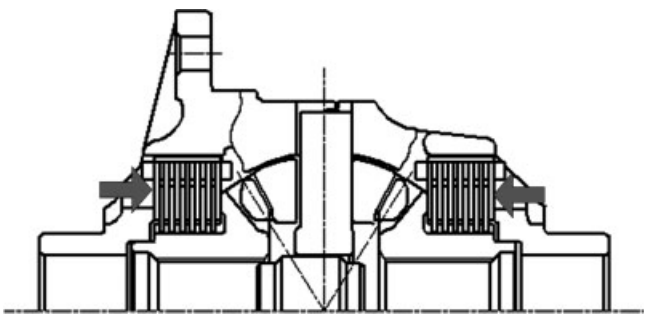


Figure 12. Cross-section drawing of a traction lok limited slip differential.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

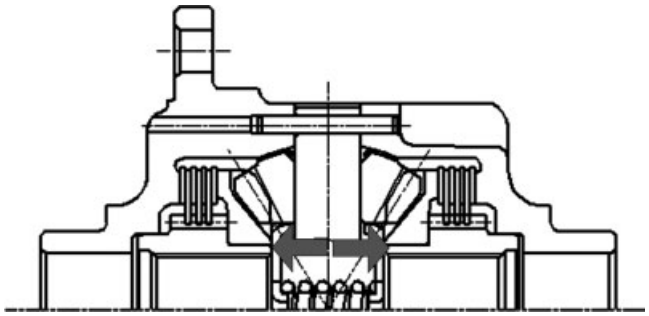


Figure 13. A cross-section drawing with the clutch pack preload in *internal* pressure.

$$r_1 = \frac{(R_1 + R_2)}{2} \tag{17}$$

with R_1 and R_2 as the operating contact radii of the side gear. Now the transfer ratio, R_t^1 can be expressed below.

$$R_t^1 = \frac{(1 + f_1)}{(1 - f_1)} \tag{18}$$

2.4.2.3.2 Internal pressure-style (TL type B). For the internal pressure style, the clutch torque equation is the same as Equation 13, but the initial torque relationship must consider the transfer ratio gradient term. In Figure 13, the clutch pack preload is achieved from the coil-style spring that is located in between the side gears. On the basis of the position of this spring being within the differential gears themselves, it is referred to *internal* pressure. The transfer ratio, R_t, term is based on some of or all of the following depending on the style of device; differential gear geometry, ring gear torque, and the cam geometry if the differential is a cam-style unit. In this context, $R^2{}_t$ is the transfer ratio term and is equal to

$$\frac{(1 + f_2)}{(1 - f_2)} \tag{19}$$

$$f_2 = \mu\ (r_2)\ (n) \left[\frac{\left(\tan\ \theta_g\ \sin\ \alpha_g\right)}{r_1} \right], \quad \text{similar to 16}$$

$$R_t^2 = \frac{\left\{ T_c + \left[\frac{(R'_t \times T_r)}{(1 + R'_t)} \right] \right\}}{\left\{ \left[\frac{T_r}{(1 + R'_t)} \right] - T_c \right\}} \tag{20}$$

where the new variables are as follows:

T_r = ring gear torque (N-m).



Figure 14. A cutaway LOM-style differential.

2.4.2.4 Lok-O-Matic style with preload. The term *LOM* refers to a style of LSD that has an inner differential housing that is nested within the outer ring gear housing as shown in Figure 14. This inner housing has a cam profile interface with a mating cam profile on the differential pin. There is another level of intricacy, as the cam surface interface with the ring gear-supported differential housing and the differential-supported inner housing. On the basis of the number of these pressure ring lugs and geometry, the pressure applied to the clutch pack can be tailored and even different from drive torque to coast torque. These cam profiles can add ring gear force to the clutch plate apply force and alter the performance characteristic of the LSD.

Just like the TL style, the clutch plate apply force can be external or internal with the addition of the ring gear torque influence. Therefore, the LOM-style device not only gives the design engineer the ability to tailor torque bias ratio based on the number of clutch plates and preload force but also adds cam angle and direction of torque application. There are three types of LOM units that will be described; the first two use external pressure for preload—one with side gear thrust adding to the bias torque and the other without, whereas the third unit utilizes internal pressure for preload along with ring gear torque. Some of the equations that describe the torque transfer ratio are the same as the TL style and will be noted accordingly. As shown in Figure 15, the inner differential housing is split and will separate based on the reaction forces from the cam surfaces and the differential pins. These cam-separation forces are depicted in this figure based on cam angle. This force reaction can be tailored during the design process in order to achieve a range of torque bias ratios.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

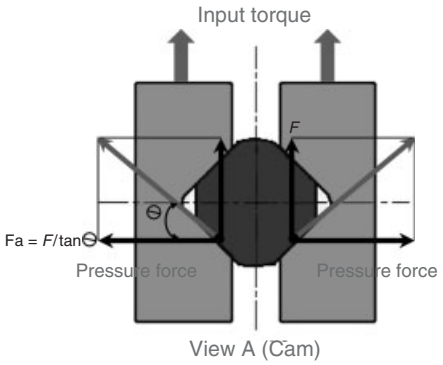


Figure 15. Cross-sectioned LOM differential cam mechanism.

2.4.2.4.1 LOM with external pressure and side gear force (LOM type A). This style LSD utilizes pressure applied to the clutch pack from outside of the bevel differential gears. This force is generated between the clutch pack and the ring gear support housing as shown in Figure 16. The LSD is mechanically arranged such that there is a clearance between the side gears and the ring gear support housing. This axial clearance allows the side gear to also separate and contribute additional clamp load on the clutch pack. Mechanical preload is achieved with Belleville-style spring washers that are located between the outer differential housing and the clutch pack and represented by the arrows.The locking ratio is described mathematically as follows:

$$f_3 = \mu\,(r_2)\,(n)\,\frac{(\tan\,\theta_g\,\sin\,\alpha_g)}{r_1 + (\tan\alpha/r_3)} \tag{21}$$

where the new variables are as follows:

$\alpha =$ cam angle (degrees);
$r_3 =$ operating radius of cam (mm).

$$R_t^3 = \frac{(1+f_3)}{(1-f_3)} \tag{22}$$

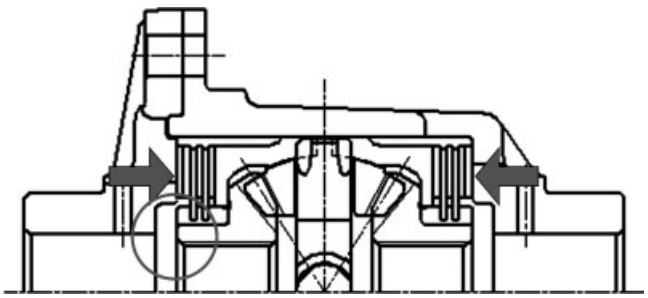


Figure 16. LOM type A.

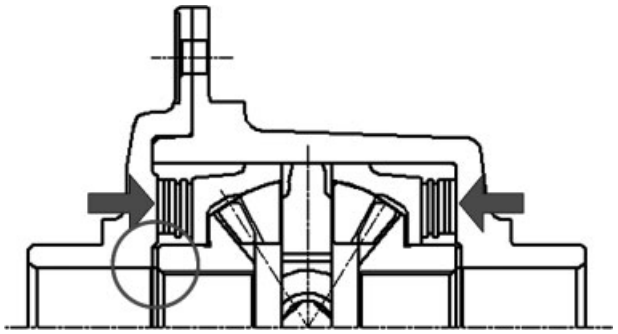


Figure 17. LOM type B cross-sectional view.

2.4.2.4.2 LOM with external pressure without side gear force (LOM type B). This style LOM LSD is similar to LOM type A except that there are thrust washers in place that limit the interaction of the differential side gear separation forces into the torque bias of the differential, as shown in Figure 17.

The locking ratio is similar to that of the LOM type A except that the side gear separating forces are omitted from the analysis, as there is a thrust washer between the side gear hub and the differential case, which positively locates the side gear relative to the clutch pack.

$$f_4 = \mu\,(r_2)\,(n)\left(\frac{\tan\alpha}{r_3}\right) \tag{23}$$

$$R_t^4 = \frac{(1+f_4)}{(1-f_4)} \tag{24}$$

2.4.2.4.3 LOM with internal pressure (LOM type C). This internal pressure-style unit is similar to the TL style with the addition of the cam profile forces during slip events. On the basis of that, the device is a combination of the TL and LOM performances and is mechanically shown in Figure 18. The preload force generated from within the differential side gears via coil springs. The coil springs are installed and compressed at assembly of the differential.

The equations are already described earlier but need to be applied differently. The force from these cam surfaces further applies the clutch pack. The locking ratio is the same as Equation 21 for f_3; the transfer ratio is as follows:

$$R_t^5 = \frac{\left\{T_c + \left[\frac{\left(R_t^3 \times T_r\right)}{\left(1+R_t^3\right)}\right]\right\}}{\left\{\left[\frac{T_r}{\left(1+R_t^3\right)}\right] - T_c\right\}} \tag{25}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

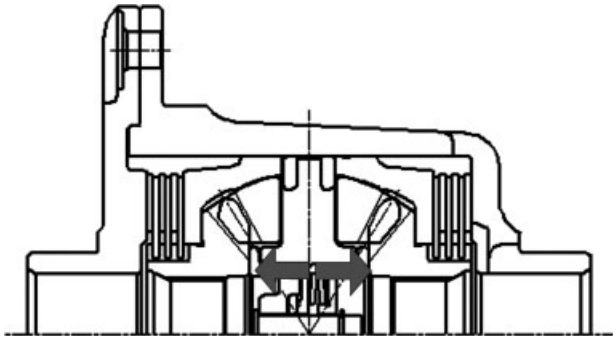


Figure 18. An LOM type C cross-sectional drawing with internal pressure.

2.5 Helical differentials

The helical-style LSDs utilize parallel- or crossed-axis helical gears instead of traditional bevel gears for the differential function as shown in Figure 19. The traditional bevel differential gears are replaced with helical side gears along with the mating element gears. Close inspection reveals that the element gears are supported on the outside diameter of the gear teeth instead of separate shafts and bearings. All of the mechanical friction developed for these gears is what provides the bias within the differential. With this style of arrangement, the torque is proportioned based on the internal friction and thrust forces developed using the helical gear geometry. These devices exhibit a mechanical control-style system between the input drive source and the two outputs. The control system is from simple mechanical friction and applies force at a distance, which creates a torque resistance. The apply force is generated from the input torque from the ring gear to the differential housing. On the basis of the coefficient of friction and gear geometry, the locking torque is developed. As the thrust forces change direction from drive to coast, it is possible to alter the bias torque using low friction washers to tailor the performance for torque transfer during drive events and open differential functionality during braking events.

There are four main areas of purposely designed friction in this LSD and are as follows: side gear to differential housing, side gear face to side gear face, element face to differential housing, and lastly side gear to element gear mesh. Of all these friction forces, the largest contributor is the side gear reaction forces with the respective element gears. This is not to say that the other frictional aspects are negligible. The specific mathematical derivation of expression of this device is quite lengthy and readily available in the literature.

The torque bias ratio of (Chocholek, 1988) this style can be as high as 6.0; however, for typical original equipment applications, the TBR is kept below 3.0 : 1. Higher values are possible and typically used for purpose-built racing applications. On the basis of the gear geometry and subsequent thrust forces, it is possible for the TBR to be a high value in the drive condition and a lower value in the coast condition. This variation in TBR allows the helical differential to mimic an open differential performance during coast conditions such as an ABS event.



Figure 19. A typical helical-style limited slip differential.

3 ACTIVE LIMITED SLIP DIFFERENTIALS

Active LSDs share many of the same components as the passive units with the added ability to electronically control the locking torque across the differential outputs. As the differential is electronically controlled, the term electronic limited slip differential (eLSD) is commonly used to describe these differentials. This electronic control system allows the unit to act as an open differential or have a locking torque, up to the clutch pack torque capacity. This locking torque can be adjusted based on the control system. For example, there may be a desire to have the active differential that maintains a clutch pack preload of 200 N-m in straight-line driving with an immediate reduction to 50 N-m when the vehicle encounters a turn maneuver. This electronic activation can be triggered independent of vehicle or wheel speeds or torques. The devices can be preemptively engaged before excessive wheel slip. These

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

style devices are typically found on high end modern vehicles and offer additional benefits over passive devices.

The electronic system would typically have a steering wheel angle sensor that would communicate to the electronic control unit that a turn maneuver is executed and the system would respond accordingly. There are many different scenarios and software algorithms that are developed when this type of hardware is available. The main advantage is the ability to control bias torque independent of slip speed. Of course, we cannot forget that the device is still diverting torque from the fast wheel to slower wheel or providing a preload torque before the eLSD allows for speed differences.

3.1 Vehicle system interaction

With any of these devices installed in a vehicle, there are certain system interactions that must be considered. There can be a trade-off between pure straight-line acceleration and traction maneuvers versus high speed cornering behavior. How the device interacts during ABS events or electronic stability control events needs to be considered. The vehicle sensor set that is available has a very important role in the overall system refinement. Most modern vehicles have the following sensors and signals available: four wheel speeds, steering wheel angle, vehicle yaw rate, throttle position, ABS/ESC active, clutch position, and off-road switch. With these sensors in place, the eLSD control software algorithm can be structured to accommodate vehicle maneuvers that include straight-line acceleration, high speed stability, and vehicle dynamics corrections while in a turn.

3.2 Direct actuation style differential

The first eLSD application subcategory is an actuation method that can be achieved either hydraulically or electromechanically with the main feature being the ability to engage the clutch pack independent of vehicle or wheel speeds. This direct actuation method allows the driver to select full lock mode and also allows the vehicle design engineer the ability to preemptively engage the LSD before any slip events. As shown in Figure 20, the direct actuation eLSD utilizes a separate electric motor that can independently drive a ball ramp mechanism in to apply the clutch pack. The ball ramp mechanism converts rotational motion from the electric motor to axial motion to apply the clutch pack.

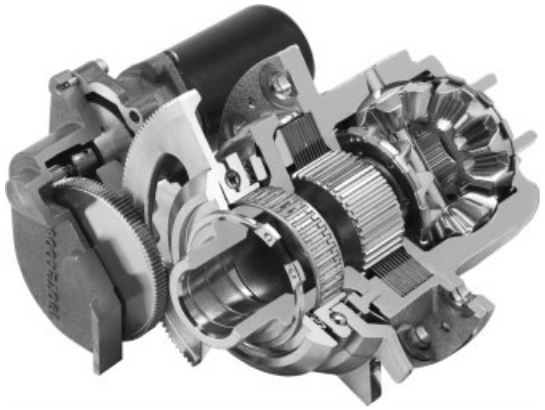


Figure 20. Direct actuation via electric motor-driven ball ramp mechanism.

3.3 Indirect actuation style differential

The indirect method is such that a pumping device or electromagnetic coil with a pilot clutch will apply the clutch pack but requires a small amount of wheel rotation to engage. The difference between this and the direct method is a slight delay in vehicle response. Depending on the vehicle conditions, this slight delay may be considered negligible but there is a performance difference.

3.4 Electronic torque-vectoring differential

With all of the above-mentioned style devices, the main functionality is to provide a locking effect across the differential based on the torque capacity of the clutch along with the performance of the apply system. Typical LSDs basically can vary the locking torque across the outputs or make the outputs travel at the same speed. Electronic torque vectoring (ETV), on the other hand, has the ability to change the output speeds of the differential relative to each other. With an ETV system as the differential, there is an additional gear-ratio mechanism that is electronically controlled that can either increase or decrease the speeds of the outputs. In Figure 21, there are ETV modules bolted to either side of a traditional open differential axle. There are two separate electric motors to actuate the clutch packs that act as brakes across the planetary style gearing. This planetary style gearing is what provides the ratio offset across the outputs. The effect of trying to change the speed of the wheels causes a torque reaction that preloads the suspension and helps the vehicle turn more sharply as compared to an open differential. The engagement device is typically actuating a clutch pack arrangement that acts as a brake to slow down one of the gear train elements to provide the necessary speed relationship. An ETV is far superior

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 21. A cutaway display of an entire rear axle. (Reproduced by permission of Joe Palazzolo.)



Figure 22. A sectioned electronic locking differential.

to an eLSD in most vehicle dynamic events; however, for a pure traction event such as straight-line hill climb with ice on one wheel and asphalt on the other, the ETV system will not perform, as well as typical LSDs. This is probably the only maneuver where ETV is not superior to an LSD.

3.5 Electronic locking differential

The locking differential-style units are unique in that they allow for a rigid mechanical lock across the differential. This is typically achieved with some sort of sliding collar or positive engagement face spline. These devices are electronically actuated but require the driver to select for the true lock-mode functionality. These units are intended for off-road usage, as, when engaged, they do not allow for any speed difference across the outputs of the differentials. As they are a rigid mechanical coupling across the wheels, if one wheel encountered a significant reduction in tractive effort, the device is capable of delivering total drive torque to the other wheel. The drive torque is not limited to the capacity of a clutch pack or other device, as described earlier for limited slip devices. This is why the term *locking* or *lockers* is used to describe this arrangement. A very important design issue is to make certain that the axle system can support delivery of 100% of total drive torque to one wheel, if one of these devices is installed in the vehicle. There are also pneumatic-style actuated locking differentials available in the aftermarket to support the

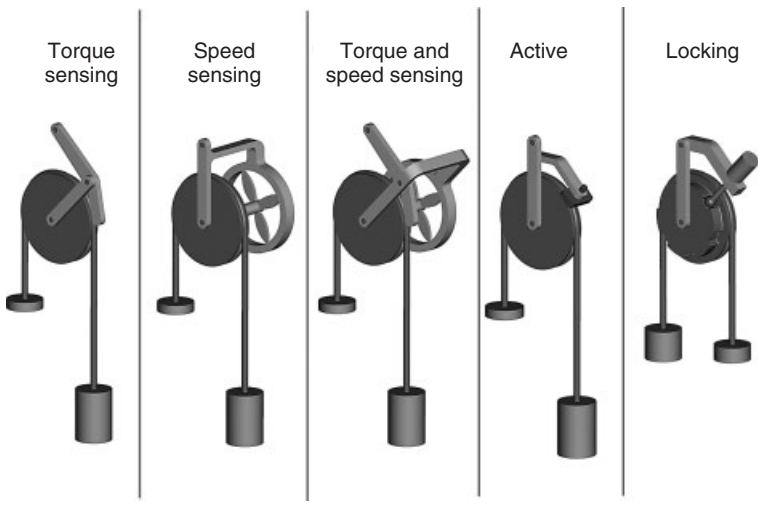


Figure 23. Limited slip differentials—pulley analogy.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

off-road enthusiast, whereas original equipment-installed units are electrically actuated. Figure 22 is a cutaway of an electronically controlled locking differential that shows the electromagnetic coil for actuation of the mechanism. The torque-carrying portion is achieved via dog clutch-style tooth engagement. This provides for a rigid mechanical lock across the outputs of the differential. With this rigid lock, the differential function is bypassed and the unit acts as a solid unit when the system is engaged. As dog clutch teeth are employed for the torque transfer, the speed difference across the outputs must be relatively low in order for the system to engage.

4 SUMMARY

There are many different styles of LSDs that are available on the market. These range from passive and active devices, along with true locking and even vectoring units. The different styles of LSDs that are speed sensing and torque sensing can be pictorially represented with a simple pulley analogy. The traditional pulley with equal weights would represent an open differential with homogeneous coefficient of friction on both output wheels. The differential-style devices are summarized in Figure 23. The torque-sensing differential can be represented by mechanical friction slowing the pulley motion, whereas speed sensing is with the pulley attached to a fan blade. These can be combined for torque and speed sensing. An active system can be represented with a controlled brake on the pulley. Finally, the locking differential has a cogged positive lock in the pulley.

REFERENCES

Chocholek, S.E. (1988) The development of a differential for the improvement of traction control. C368/88, IMechE.

FURTHER READING

Palazzolo, J. (2010) *High-Performance Differentials, Axles & Drivelines*, ABC Books, UK. ISBN: 1934709026

Palazzolo, J. (2013) *Ford Differentials, How To Rebuild the 8.8 and 9-inch*. ISBN:978-1-61325-038-9

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto098
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Clutch Actuation

Joško Deur[1] and Vladimir Ivanović[2]

[1] *University of Zagreb, Zagreb, Croatia*
[2] *Ford Research and Advanced Engineering, Dearborn, MI, USA*

1 Introduction 1
2 Various Design Concepts of Clutch Actuation System 2
3 Main Features of Static and Dynamic Behaviors of Clutch 10
4 Clutch Actuation Control System 15
5 Conclusion 17
Acknowledgment 17
References 17

1 INTRODUCTION

The clutch actuator is a mechatronic system that provides means of controlling a variable that directly influences the clutch torque. For widely used plate clutches (see examples in Figure 1a, c, and d), the influencing variable is the clutch normal force F_n, which relates to the clutch torque τ_c in the proportional manner (Figure 2, see Dry Clutch and Clutch Wet):

$$\tau_c = \frac{2}{3}\frac{r_o^3 - r_i^3}{r_o^2 - r_i^2} N_f \mu(\omega_s, \vartheta, F_n) \cdot F_n(u) \qquad (1)$$

where N_f is the number of active friction surfaces, r_o and r_i the outer and inner radii of friction plate, respectively, and μ the friction coefficient. Apart from the precise normal force control by means of actuator control variable u (e.g., voltage or position), the clutch torque control law needs to include adaptation with respect to friction coefficient variations because of changes of clutch slip speed ω_s, friction interface temperature ϑ, and the normal force F_n itself. Here, the interface temperature ϑ should be estimated online based on a clutch thermal model fed by the dissipated power $\tau_c\omega_s$ and a measured "heat sink" temperature.

For band clutches (Figure 1b), the clutch torque is again adjusted by means of force control; however, in that case, this is the band pushing force, which corresponds to the spatially distributed band-drum normal force (Fujii, Tobler, and Snyder, 2001). On the other hand, some electromagnetic clutches do not include moving parts for controlling the normal force; however, the clutch torque is rather dependent on the solenoid current-controlled magnetic flux density in a "smart" magnetic medium such as magnetorheological fluid (MRF) or magnetic particles (MPs).

The clutch actuation systems can be divided with respect to several aspects listed in Figure 3. Firstly, it matters if whether the actuator is designed for a wet or a dry clutch, and plate or band clutch. Secondly, the supervisory clutch control task determines whether the clutch operates in an intermittent mode (typically, clutch engagement; see Automatic Transmissions - geartrain combinations, components, design considerations, hydraulic system, packaging, manuf., assembly) or a continuous-duty mode (e.g., torque control in driveline applications; see Axle Systems), which has influence on both clutch and clutch actuator designs. Thirdly, there are many actuation design concepts, starting from the conventional hydraulic and pneumatic systems, through more recent electromechanical and electromagnetic systems, to emerging MRF and MP concepts. Fourthly, the design concepts can be divided into different categories with respect to the actuator transmission type, which include direct actuation, lever actuation, ball ramp systems,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

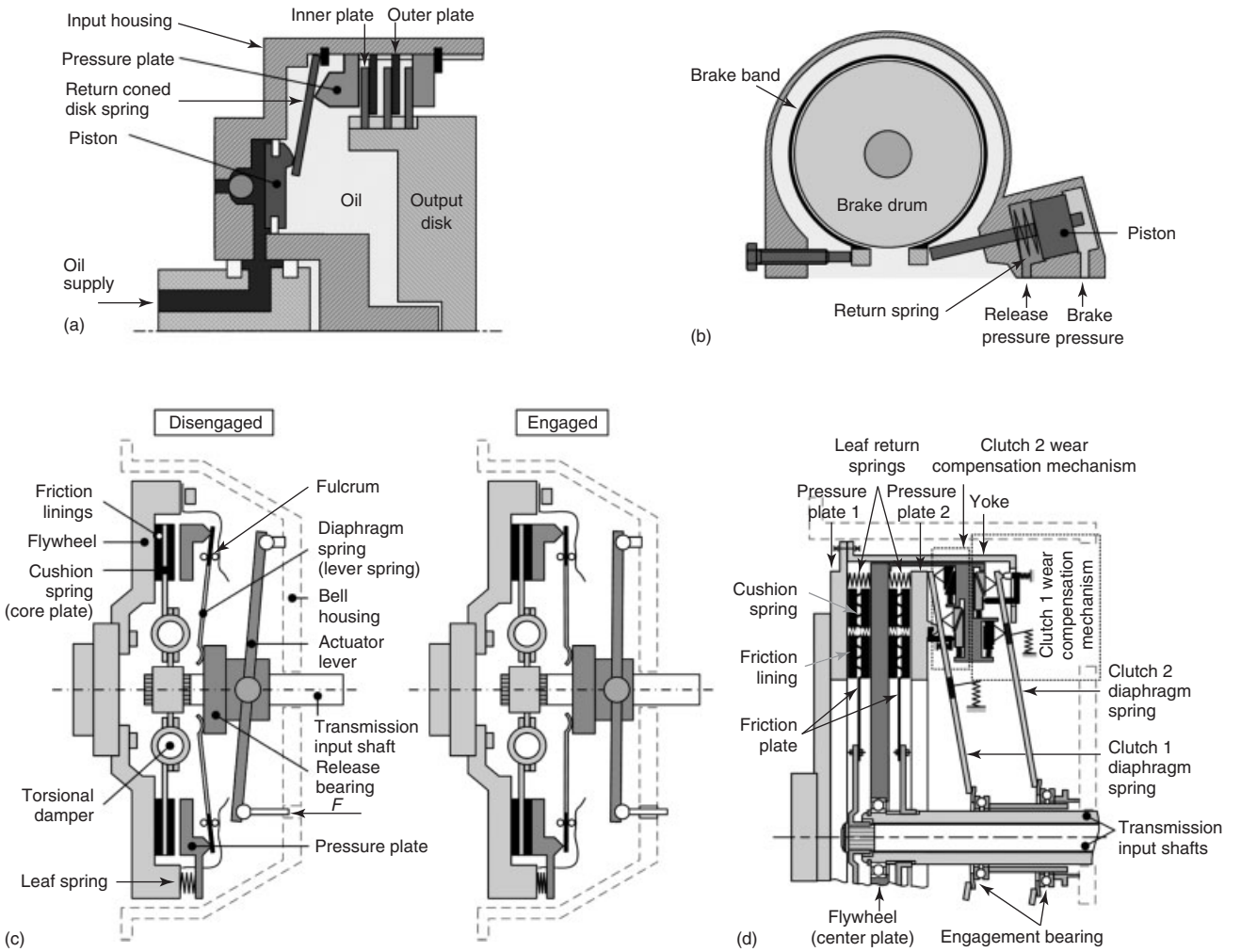


Figure 1. (a) Schematics of wet clutch. (Adapted from Fachkunde Kraftfahrzeugtechnik (n.d.)); (b) band clutch/brake. (Adapted from Fachkunde Kraftfahrzeugtechnik (n.d.)); (c) AMT dry clutch. (Modified from ZF Friedrichshafen AG (n.d.)) and (d) DCT dry clutch. (Adapted from Ivanović *et al.*, 2012b)

ball screw drives, and worm gears. Finally, the clutch control system can be designed according to two basic concepts: actuator force/torque control and actuator position control, whereas some actuators have combined the features of force/torque and position controls.

From the standpoint of clutch actuator design and control system synthesis, it is important to understand various factors that influence the clutch static and transient behaviors. For instance, the steady-state accuracy of clutch torque control is affected by mechanical friction and magnetic hystereses. On the other hand, the clutch transient performance is influenced by the clutch clearance and the actuator backlash, as well as the actuation system structural compliance. The thermal effects have direct influence on the clutch friction coefficient and, thus, on the clutch torque, and in some cases (e.g., dry clutches and/or position-controlled systems), the clutch pack thermal expansion can affect the steady-state accuracy. Finally, the friction material wear tends to increase the clutch clearance, and at the same time, it affects the friction coefficient behavior.

2 VARIOUS DESIGN CONCEPTS OF CLUTCH ACTUATION SYSTEM

2.1 Hydraulic system

2.1.1 Structure of hydraulic actuation system

Figure 4 shows the functional diagram of an electrohydraulic control system. Mechanical energy supplied by the internal combustion engine (ICE) or an electric motor (EM) is converted into hydraulic energy through a positive

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

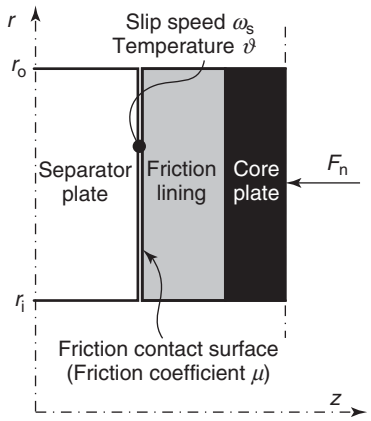


Figure 2. Definition of main clutch quantities. (Reproduced from Ivanović, 2012. © Sage.)

displacement pump of gerotor, vane, gear, or axial piston design (Merritt, 1967), which is stored in a hydraulic accumulator (Figure 6d). Depending on the target application, control valves, which are operated through an electronic control unit, are used to control flow direction, flow rate, and/or pressure of the hydraulic energy. The hydraulic energy is finally converted back into the mechanical energy by a hydraulic piston connected to the clutch pressure plate (Figure 1a). Note that the fluid flow and/or pressure can also be controlled directly by controlling the pump speed or volume. This efficient approach can be applied for systems with one or multiple clutches if only one is actuated at the time. The complexity and cost of the system are reduced in this way, because the hydraulic accumulator and pressure/flow control valves (FCVs) are eliminated and consequently simpler system of oil filtration is required.

2.1.2 Description of main control elements

The time response and accuracy of hydraulic actuation systems are generally defined by the dynamics of pump, accumulator, control valves, line flow, and clutch pressure plate piston and oil properties (e.g., Merritt, 1967; Watechagit and Srinivasan, 2003). The key elements are electronically controlled pressure reducing valves (PRVs) that are used as either (i) direct large-flow clutch actuation valves or more often (ii) low flow signal-level devices that

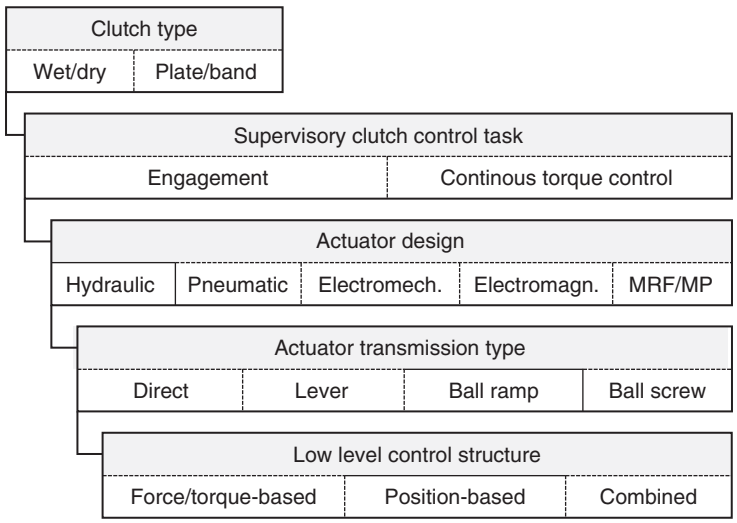


Figure 3. Clutch actuation division chart.

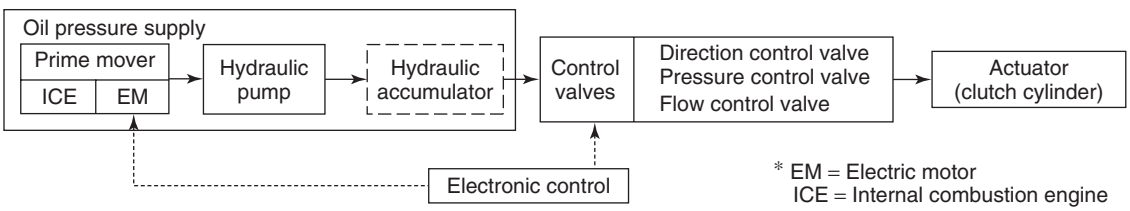


Figure 4. Electrohydraulic control system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

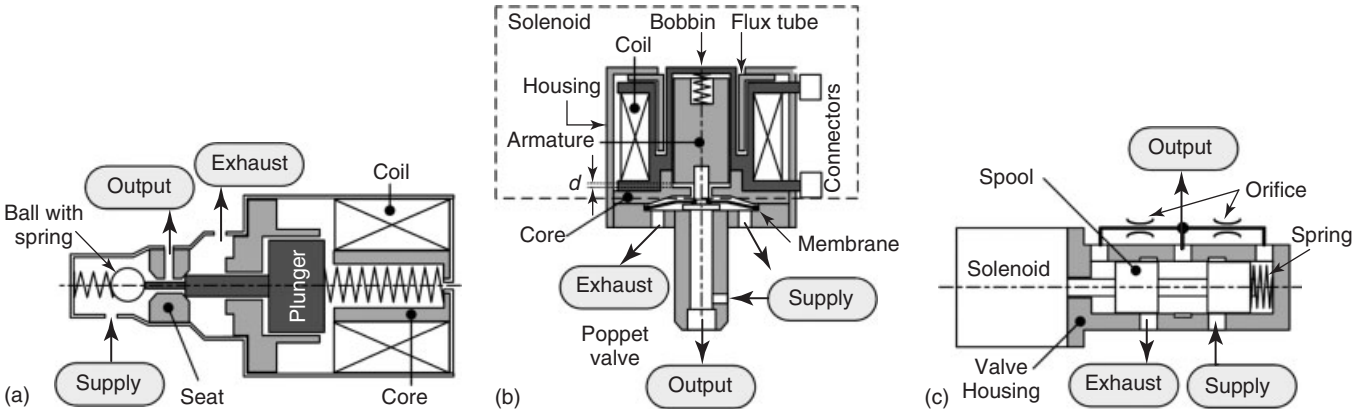


Figure 5. Hydraulic solenoid valves: (a) PWM solenoid. (Modified from Cho, 2002. Copyright © 2002 SAE International. Reprinted with permission); (b) VBS-variable bleeding solenoid. (Modified from Holmes and McKenna, 2001); and (c) VFS-variable force solenoid.

provide operating pressure signal for the main pilot operated three-way spool-type large-flow PRV (Merritt, 1967). Figure 5 shows the schematics of three typically used three-way PRVs: (a) pulse-width-modulated solenoid (PWMS), (b) variable bleed solenoid (VBS), and (c) variable force solenoid (VFS).

The PWMSs (Figure 5a; see, e.g., Cho, Oh, and Lee, 2002) are mostly used at the signal level, but can also be used as the direct actuation valves. They do not include control pressure (p_c) feedback. Depending on the solenoid state (energized or de-energized), the control port is connected with either the supply or the exhaust port, respectively. The control pressure is controlled by the duty-cycle ratio of a PWM signal with the period of around 16 ms. As such, the control signal is characterized by oscillations and it is sensitive to the supply pressure variations.

The VBSs (Figure 5b) are used as signal-level devices. The control pressure is adjusted by controlling the oil bleeding based on a balance between the solenoid and the pressure feedback force on the valve element (ball, conical, or flat as in Figure 5b; (Holmes and McKenna, 2001). The solenoid force is controlled by varying the magnetic flux through the air gap d by means of controlling the solenoid current based on the duty-cycle adjustment of PWM armature voltage signal. In terms of control pressure oscillations, the VBSs are better than PWMSs; however, they are still sensitive to the supply pressure oscillations.

The VFSs (Figure 5c) are based on a balance between the solenoid and the pressure feedback force on the valve spool. The solenoid force is controlled in the same manner as on VBSs, that is, through the duty cycle of PWM solenoid voltage signal with a high carrier frequency (e.g., 200 Hz), which provides virtually ripple-free control pressure (Lee *et al*., 2010). They can be designed as low flow or large-flow valves and can be, therefore, used at signal or direct actuation level, respectively.

2.1.3 General facts and application examples

The hydraulic/electrohydraulic actuators have been broadly used for the purpose of automotive clutch actuation because of large output/weight ratio, fast response, easy conversion of engine/EM output into hydraulic power elsewhere in the vehicle where space is not critical, safety, high reliability, good packaging, and the best ability for clutch torque closed-loop control because of direct relation between the hydraulic pressure and the clutch pack normal force (Yoshioka *et al*., 1985; Turner and Ramsay, 2004; Francis, Haselton, and Pritchard, 2006). Therefore, in some application fields such as torque converter automatic transmission (AT), hydraulics have been used exclusively. However, the main disadvantage of these systems is relatively low efficiency mainly because of internal losses and poor duty cycle when the pump is driven by a mechanical drive. According to Turner and Ramsay (2004), the hydraulic actuation systems may represent 50% of the total transmission losses. This loss share can be reduced if the pump is driven by an EM, but losses associated with leakage and flow still remain. Another disadvantage of these systems is sensitivity to the operating temperature as a consequence of change of oil viscosity and valve solenoid coil resistance, which directly affects the response time (it typically reduces with decreasing temperature).

Figure 6 shows examples of electrohydraulic clutch actuation systems categorized according to the field of application. Wet clutches in conventional *torque converter ATs*

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

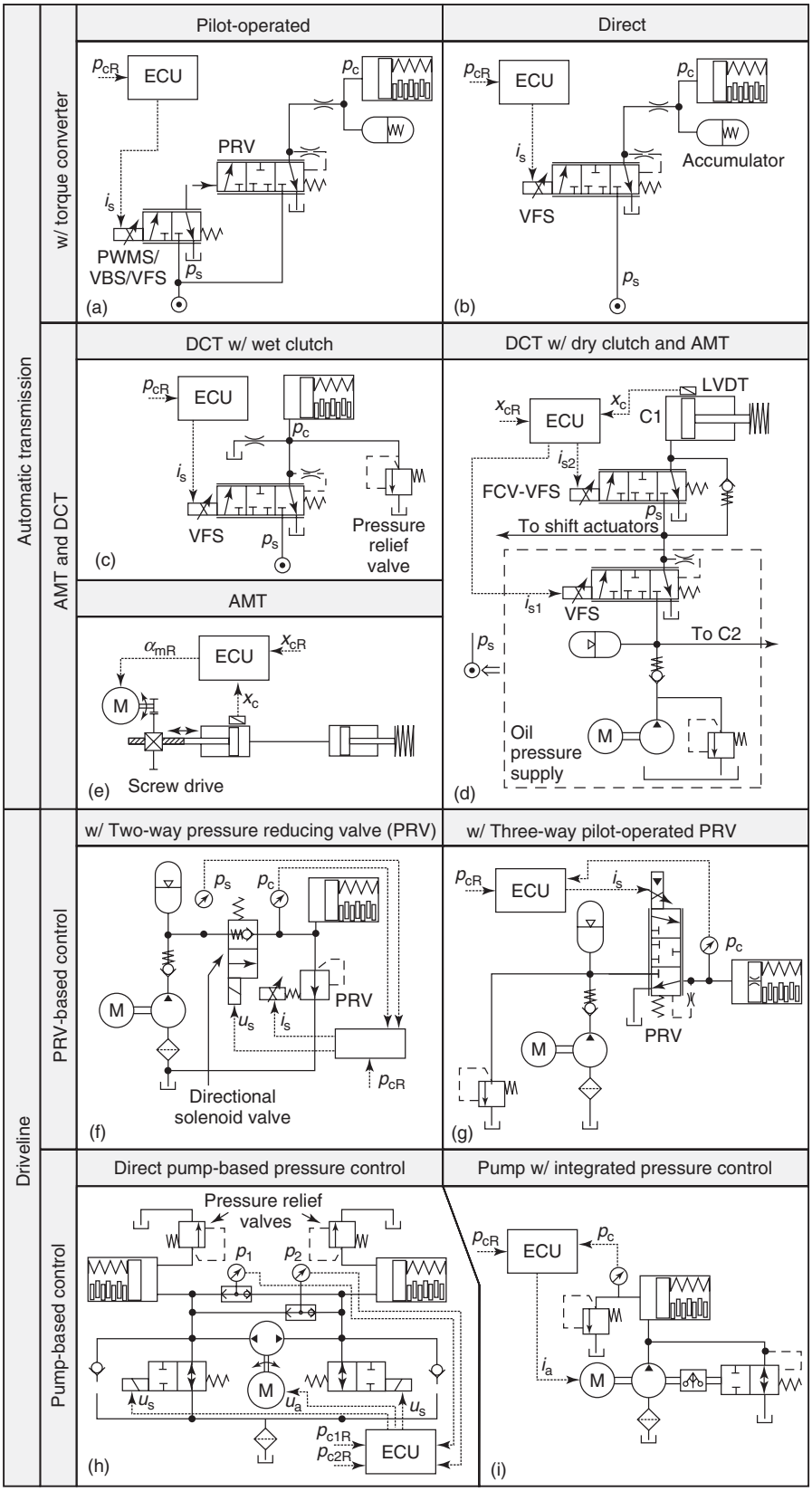


Figure 6. (a–i) Electrohydraulic actuation systems for different applications.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

(Figure 6a and b) are usually controlled in the open-loop manner using three-way PRVs. The clutch pressure is not measured because of the cost reduction and packaging reasons. The closed-loop pressure control is, though, a subject of research activities (see, e.g., Zheng *et al*., 2009). An accumulator is used in parallel with clutch actuator cylinder in order to provide smooth engagement after clutch filling phase by avoiding large pressure overshoot, oscillations, and instability. The clutch pressure is controlled by pilot-operated PRVs (Figure 6a; Lewis and Bollwahn, 2007; Watechagit and Srinivasan, 2003) or emerging direct large-flow three-way PRVs, that is, VFSs (Figure 6b; Kondo *et al*., 2007). The PRVs are always of spool type with internal pressure feedback loop similar to the VFS in Figure 5c with difference that the spool control force is provided by pressure signal instead of the solenoid (Merritt, 1967; Watechagit and Srinivasan, 2003). The pilot-operated PRVs are operated by pressure signal provided by an additional low flow pilot PRVs (Figure 6a) such as PWMS, VBS, or VFS.

The direct three-way PRVs (large-flow VFS) are also used for controlling *wet clutches* in *dual clutch transmissions* (DCTs; see Dual Clutch Transmissions (DCT) - layouts, clutch selection, packaging, actuation, manufacturing & assembly), as shown in Figure 6c (Mustafa *et al*., 2010; Balau, Caruntu, and Lazar, 2011). Instead of using the hydraulic accumulator, this system uses a small discharge (constant bleeding) orifice. The oscillations and instability can be avoided in this way, because the PRV spool does not need to cross the dead zone to reduce the clutch pressure after the filling phase.

In order to avoid a significant influence of mechanical friction hysteresis (Section 3), the hydraulic actuation systems for *dry clutch control applications* (Figure 6d) are based on closed-loop control of clutch engagement bearing position using three-way proportional FCVs similar to the VFS in Figure 5c but with the absence of control pressure feedback. These systems can be found in *DCTs* with normally open dry clutches (Figure 1d; Hadler *et al*., 2008) and AMTs (*automated manual transmissions*; Figure 1c; see Automated Manual Transmissions (AMT) - system design considerations, clutch operation, shift actuation alternatives) with normally closed dry clutches (Montanari *et al*., 2004). Figure 6e shows another concept of electrohydraulic actuator for AMT clutches (Oberlack and Reul, 2006) that combines the features of electromechanical and hydraulic actuation systems. A brushless DC motor (Section 2.3) drives a hydraulic piston through a ball screw. The piston is connected through a hydraulic line to another piston that actuates the clutch. The clutch piston position is measured and fed back to the control unit, which controls the driving EM. A pressure-based feedback can also be implemented, as needed. Similar actuation system has been proposed for driveline clutches by Francis, Haselton, and Pritchard (2006). Note that this actuation concept is in essence related to pump-based control.

Next to the ATs, wet clutches are widely used in *driveline applications* such as active limited slip differentials (ALSDs; see Passive and Active Limited Slip Differentials) or torque vectoring differentials (TVDs; see Axle Systems). In terms of hydraulic control system, they can be divided into PRV-based systems (Figure 6f and g) and pump-based systems (Figure 6h and i). PRV-based control is similar to the previously explained AT clutch control. However, in order to provide better accuracy and favorable response time for a small number of clutches (one or two), the clutch actuation pressure is controlled in the closed-loop manner. Figure 6f and g shows implementation variations with a two-way PRV (Ross *et al*., 2007) or a VFS (Morselli *et al*., 2003), respectively. Note that the system in Figure 6g includes a small discharge orifice bored in the clutch piston for the same reason as explained earlier for the DCT wet clutches.

The pump-based system in Figure 6h shows the clutch control system for a TVD with two superimposed clutches (Sackl, Eibler, and Linortner, 2008), where one clutch is actuated at the time. The positive displacement pump is driven by a four-quadrant speed-controlled electric servomotor, thus providing the possibility of closed-loop pressure control. The clutch actuator pressure is measured by two redundant sensors. The system air bleeding is provided by pressure relief valves at specific instants, for example, when pressure spikes are filtered by opening the valves. Another pump-based pressure control system is shown in Figure 6i (Nilsson *et al*., 2011). In this case, the pump controls the pressure by commanding the pump DC motor armature current reference (i.e., the motor/pump torque). System air bleeding is again realized through the pressure relief valve.

2.2 Pneumatic system

The basic principle of pneumatic systems is very similar to the hydraulic systems. The difference is related to the working medium, which, in this case, is the compressed air. The air is an easily compressible gas, which introduces significant compliance in the system dynamics. Its density is significantly influenced by temperature, which affects volumetric flow through the control valves, and thus the overall system response. This nonlinear behavior coupled with nonlinear friction effects in pneumatic cylinders and valves, as well as variations of the accumulated air pressure, makes the pneumatic systems demanding in terms of control. Note also that the working pressure used in pneumatic systems is by more than one order of magnitude

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

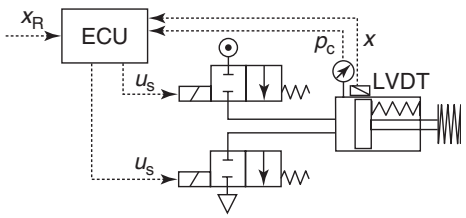


Figure 7. Electropneumatic actuation system.

lower than in hydraulic systems, thus making them inconvenient where high actuator force (i.e., high force density) is needed.

The pneumatic systems are used for transmission clutch actuation purposes on commercial vehicles (e.g., trucks or buses) because of available compressed air supply. The transmissions are of AMT type with normally closed dry clutches (Figure 1c) for which an accurate and robust release bearing position is required (Montanari *et al*., 2004; Langjord and Johansen, 2010). Figure 7 shows a typical electropneumatic clutch actuation system (Langjord and Johansen, 2010). It is based on cheap and simple ON/OFF solenoid valves. More expensive and larger servo valves can also be used, but even simple ON/OFF valves with proper control can give an acceptable system behavior. Note that the cylinder pressure and position are measured and fed back to the control unit. The control unit outputs the voltage signal for the control valves.

2.3 Electromechanical system

The electromechanical systems comprise an electric motor (EM) for the clutch actuation. Both linear and rotary motors can be used. While the linear motor-based systems are in the research/development stage (Wheals *et al*., 2009), their rotary motor-based counterparts have found many production applications, some of which are described below. Besides the EM, the electromechanical actuation systems comprise an additional device to transform rotational motion of the motor into linear motion at the clutch pressure plate, and it can be of ball ramp, screw drive, or lever-based design (Figure 8). Their main advantages when compared to the electrohydraulic systems (Section 2.1) include (Turner and Ramsay, 2004) a high efficiency (up to 95%), excellent motor control accuracy, and a great modularity and ease of implementation in existing transmission or driveline hardware. The main disadvantages of the electromechanical systems relate to smaller force/torque density and a significant influence of the actuator transmission friction, backlash, and/or compliance effects on the overall control performance (Section 4).

2.3.1 Ball ramp mechanism

Ball ramp-based system for an ALSD is illustrated in Figure 8a (Gassmann and Barlage, 2004; Ivanović *et al*., 2012a). A geared permanent-magnet DC motor is used to engage the clutch through a ball ramp mechanism that converts the motor torque into a high clutch-pack normal force. By combining bevel gears with ball ramp, a very high reduction ratio between the pressure plate displacement and the driving motor angle can be achieved (<0.2 mm/rev). The ball ramp mechanism consists of input and output disks with oppositely arranged grooves with defined slope (ramp), and balls placed in the grooves. Here, the slope is constant along the groove length, whereas in some actuator designs, a nonlinear slope may be used in order to provide quicker response during transition through clutch clearance zone. The ball ramp system is a part of the differential case, but it does not rotate together with the case. This is achieved by separating the ball ramp system from the case by three needle bearings. The motor is equipped with an electromagnetic brake in order to improve the system efficiency by holding the de-energized motor in a desired position for specific control tasks (e.g., locking the differential).

The ball ramp input disk is driven by the DC motor. The output disk rotation is constrained by a spring-damper system fixed to the differential housing, which allows for linear motion and only small amount of rotational motion during the clutch disengagement. The role of the damper is to reduce stress on the ball ramp elements during the clutch disengagement interval, when the output disk is pushed back by the coned disk return spring. The output disk is axially connected, but not physically bonded to the clutch press plate, that is, the actuator can only push the press plate but cannot pull it during the disengagement phase. The ball elements of the ball ramp assembly are characterized by significant compliance (Ivanović *et al*., 2012a), which introduces load-dependent rolling resistance (friction) that affects the efficiency and also control performance (Section 3).

2.3.2 Ball-screw mechanism

An alternative to the ball ramp mechanism is a highly efficient and stiff ball screw or planetary lead screw, as illustrated in Figure 8b (Turner and Ramsay, 2004). Note that the reduction ratio in this case (typically 1 mm/rev) is significantly lower when compared to the ball ramp system. Therefore, this type of actuator may be considered as a quasi-direct actuation, that is, together with the motor, it represents an electric cylinder. In order to boost the reduction ratio, and consequently reduce the motor

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

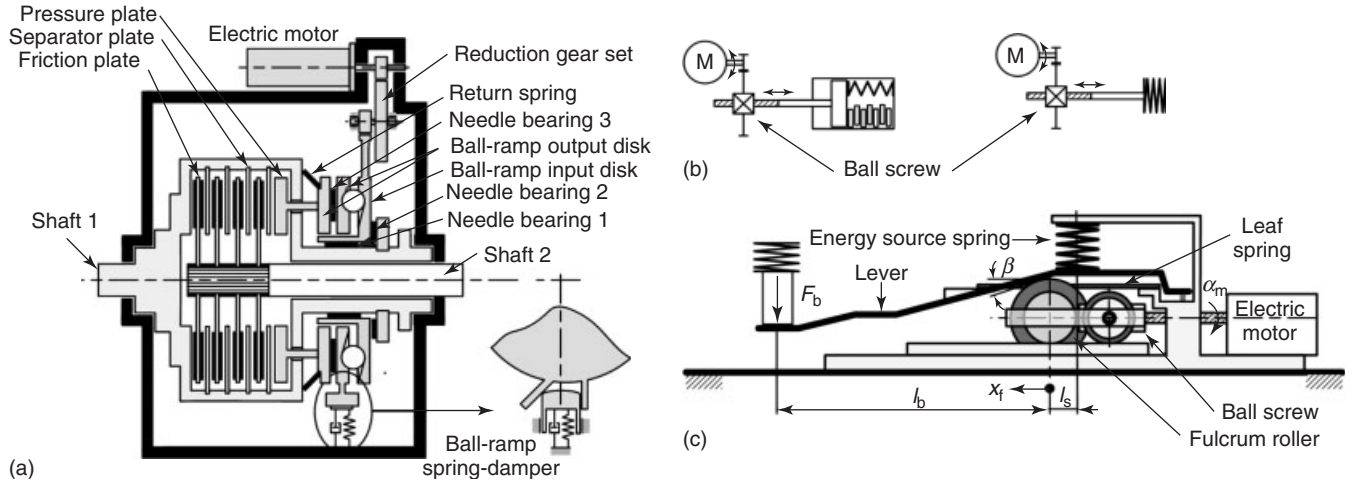


Figure 8. Electromechanical clutch actuation systems based on (a) ball ramp mechanism. (Modified from Ivanović, 2012a. © Sage); (b) ball screw; and (c) lever. (Reproduced in part from Ivanović, 2011. © ASME).

power/size, the electric cylinder may be combined by another mechanism such as lever system (Section 2.3.3).

2.3.3 Lever-based mechanism

Figure 8c illustrates a lever-based electromechanical actuator for the DCT dry clutch in Figure 1d (Wagner *et al*., 2009; Ivanović *et al*., 2011). The actuator consists of a lever, a brushless DC motor connected to a precise ball screw, a cart with lever fulcrum roller and additional auxiliary rollers, a leaf spring, and a pair of preloaded helical energy source springs. The motor drives the ball screw, which is directly connected to the cart that holds the fulcrum roller. The lever is leaned onto the fulcrum roller and loaded by the energy source spring force F_s and the engagement bearing load F_b. As such, the lever maintains torque equilibrium of these forces acting through their respective lever arms. The higher the fulcrum position x_f is, the higher is the engagement bearing force F_b. The lever has a specific nonlinear profile providing that the lever moves toward the dual clutch assembly when pushed by the fulcrum roller. Hence, the lever at the same time moves radially with respect to the screw drive center line and rotates around fulcrum roller, thus maintaining the torque equilibrium.

The driving EM provides torque required to oppose the axial component of the fulcrum–lever contact force (defined by angle β), the drive acceleration torque, and friction in motor, ball screw, and fulcrum rollers. For safety reasons, a passive clutch opening must be provided in case of system/power supply failure. This is provided by a sufficiently large value of the fulcrum–roller contact angle β for any actuator position. If the lever rotational angle would significantly differ from its initial value, this condition could not be met. The lever geometry and parameters are, therefore, optimized, so that the lever angle deviations are minimal, and also to provide close-to-linear relation between the actuator motor position and the clutch normal force.

Owing to the nonlinear lever geometry, the motor torque versus actuator position static curve is rather flat, but it includes a wide hysteresis because of actuator friction (Ivanović *et al*., 2011). As such, the system may be regarded as a ball (roller) ramp system, with a movable roller, elastically suspended output plate, and a nonlinear (ball) roller ramp contact, shaped in order to meet specific requirements.

2.4 Electromagnetic system

Electromagnetic systems utilize solenoids to generate magnetic flux circulating through the surrounding ferromagnetic yoke and armature parts, and an air gap between them. The magnetic flux generates the magnetic force acting on the armature that carries the piston and is suspended by a return spring. The larger the solenoid current is, the larger is the magnetic flux and the magnetic force. The flux versus current static curve includes a magnetic hysteresis (Section 3). For the common case of DC current magnetization, the magnetic force increases as the air gap reduces. The force response time depends on the solenoid time constant and also on the time constant related to the eddy current effect (Section 2.5).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

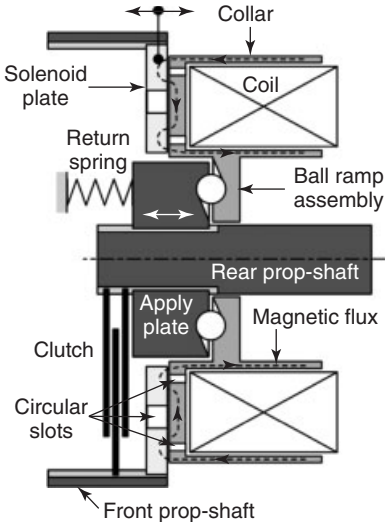


Figure 9. Electromagnetic ball ramp clutch actuation systems. (Adapted from Hrovat *et al*., 2000. © Taylor & Francis/CRC Press.)

The electromagnetic actuation systems can be direct or indirect. In the case of direct actuation, the electromagnetic force is directly associated with clutch normal force by appropriate clutch design (Kunii *et al*., 2005). The same reference proposes the use of an additional solenoid, the so-called search solenoid, which provides online reconstruction of the magnetic flux for the purpose of compensation of the flux/force variations with the air gap change.

Figure 9 shows an example of the indirect wet clutch actuation, which is used for torque control of a wet clutch in a 4 WD system on-demand transfer case (Hrovat, Asgari, and Fodor, 2000). The clutch is used to control the slip speed between the front and rear prop shafts. The system comprises a solenoid surrounded with a collar and a solenoid plate, ball ramp assembly with a return spring (cf. Section 2.3), and a wet clutch assembly. The solenoid is fixed to the stationary housing, whereas the collar is connected to the ball ramp input disk. The ball ramp output disk (the apply plate) and the solenoid plate are connected by spline couplings to the rear and front prop shafts, respectively. For the de-energized solenoid, the collar rotates with the rear prop shaft speed. When the solenoid is energized, a magnetic flux is established through the collar and solenoid plate and the corresponding magnetic force is generated. Note that the collar and the solenoid plate comprise circular slots that properly define the path of strong magnetic flux between the collar and the plate.

When a slip speed occurs between the two prop shafts because of the tendency of rear axle spinning, the generated magnetic force between the collar and the solenoid plate increases sliding friction force between them, thus reducing the collar speed. This induces a relative speed between the collar and the apply plate, which causes the apply plate to move axially, thus pressing the clutch pack and developing the clutch torque that reduces the rear prop-shaft speed. The clutch normal force is proportional to the friction torque (i.e., the magnetic force) between the collar and the solenoid plate as long as the relative speed between them exists. A mechanical hysteresis can be expected in this relationship because of the ball ramp friction (Section 2.3).

2.5 Magnetorheological fluid (MRF) and magnetic particle (MP) systems

A special feature of MRF/MP clutches is that their actuation system is fully electromagnetic. That is, the clutch is controlled by changing the rheological/magnetic properties of MRF/MP medium by means of solenoid current control. This greatly simplifies the actuator design and improves the control performance, as there are no moving actuation parts and related friction, backlash, and compliance effects.

There are three basic types of MRF/MP clutches: plate, bell-shaped, and bevel clutches (Lampe, 2000). The plate and bell-shaped clutches are similar to the plate and band friction clutches (shown in Figure 1), respectively, with the main difference that the gaps between the plates and the drum and case, respectively, are filled by the MRF/MP medium. The bevel clutch may be regarded as a "combination" of plate and bell-shaped clutches. Figure 10 shows a bell-shaped MP clutch used as a soft starting

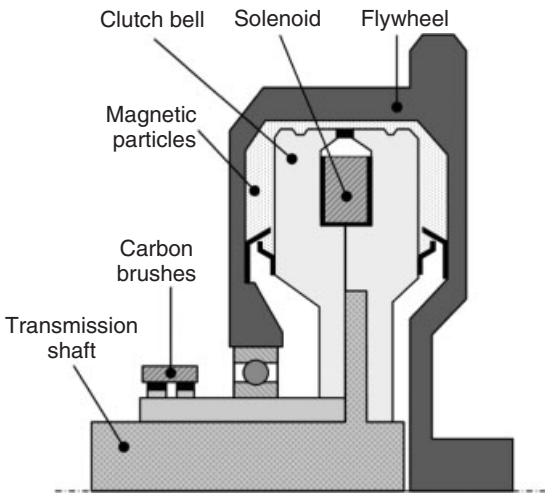


Figure 10. Magnetic particle clutch. (Modified from Sakai, 1988 and Fachkunde Kraftfahrzeugtechnik, n.d.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

device for a continuously variable transmission (Sakai, 1988). Changing the solenoid current I results in a variable magnetic flux density B in the MP (or MRF) medium, which gives the possibility of adjusting the shear stress $\sigma(B)$ and, correspondingly, the clutch torque $\tau_c(B)$ (see Lampe, 2000 for torque equations for the different clutch types). The clutch is designed based on the well-known electromagnetism laws and known MRF/MP medium static curves $\sigma(B)$ and $B(H)$, where the magnetic field H is directly related to the solenoid current (similarly as with the electromagnetic clutch).

On the basis of the notes from Rabinow (1948), Wheals *et al*. (2004), and Kawai *et al*. (1988), it appears that the main advantages of the MRF clutches (where micron-sized MPs are suspended in oil) include smooth operation, less emphasized particle wear and related torque loss, less emphasized abrasiveness (e.g., with respect to seals), and good oxidation and water adsorption resistance. On the other hand, the MP clutches should have significantly smaller drag torque (because of the absence of viscous torque component), the particles could resist approximately twofold larger temperatures than MRF (which results in a larger thermal capacity), and there is no MRF's specific effects such as fluid thickening. In order to significantly mitigate the MR fluid thickening and particle settling effects, proper oil additives are used according to Carlson (2001) and Jolly, Bender, and Carlson (1998).

The MRF/MP gap should be as small as possible to reduce the current and, thus, the power consumption. However, for high speed clutches with a narrow gap, the centrifugal particle settling can occur, thus leading to a loss of torque consistency. According to the theoretical and experimental analyses presented by Lampe, Thess, and Dotzauer (1998) and Lampe (2000), the centrifuging effect may be avoided by increasing the MRF thickness to approximately 3 mm.

The clutch torque response speed is predominantly affected by (Lampe, 2000) (i) the current delay due to the solenoid inductance and (ii) the additional delay of magnetic flux response due to the effect of eddy currents (note that the MRF/MP delay typically equals a couple of milliseconds and may be neglected). As shown by Lampe (2000) and Deur, Herold, and Kostelac (2009), the current delay can be substantially reduced by a short duration, controlled forcing of solenoid voltage in a similar manner as done for reducing the DC motor armature current delay. The eddy current effect on torque response delay can effectively be overcome only by a slotted design of the magnetic core. The slot width should be less than 3 mm to reduce the eddy current-related response time constant well below 100 ms (see Lampe, 2000; Deur, Herold, and Kostelac, 2009).

3 MAIN FEATURES OF STATIC AND DYNAMIC BEHAVIORS OF CLUTCH

Starting from the clutch actuator design features elaborated in Section 2, this section describes the main effects that influence the steady-state and transient behaviors of clutch torque control system.

3.1 Main effects affecting accuracy of clutch torque control

3.1.1 Clutch clearance

The clutch clearance, as well as the backlash in the actuator components, should be as small as possible to reduce the pure delay of clutch torque response during the clutch engagement. On the other hand, the clearance should be large enough to avoid significant drag torque of wet clutch and safely provide zero drag torque of dry clutch. The conflicting requirements are obviously more difficult to be met for wet clutches, as (i) the viscous drag torque is inversely proportional to clearance and (ii) the fluid film squeeze resistive force (Section 3.1.5) opposes the actuator force and increases the pure delay. On the other hand, in the case of dry clutches, the friction lining is thicker and wears faster. This would result in a progressive increase of clearance, and consequently in increase of the response pure delay and change of the control static curve (particularly for the lever-based actuator). To avoid/mitigate these effects, the dry clutches are equipped with a wear compensation mechanism (Figure 1d), which keeps the clearance below a predefined value (typically, around 0.5 mm). The MRF/MP clutches have a fixed and relatively large value of clearance/gap (Section 2); however, this does not have any significant influence on torque response delay.

Regardless of whether the wear compensation mechanism is present in the particular friction clutch, the control algorithm should include a routine for fast passing through the clearance/backlash zone and the so-called soft landing, where the clearance parameter should be identified by an auto-tuning routine (Section 4). For the electromagnetic actuator and also the lever actuator to some extent, additional compensation/mapping is needed to accurately provide the demanded normal force in the presence of clearance-related actuator position offset variations (Section 2).

3.1.2 Friction hysteresis

The friction hysteresis inherently appears in the force/torque-controlled clutch actuation systems. Here, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

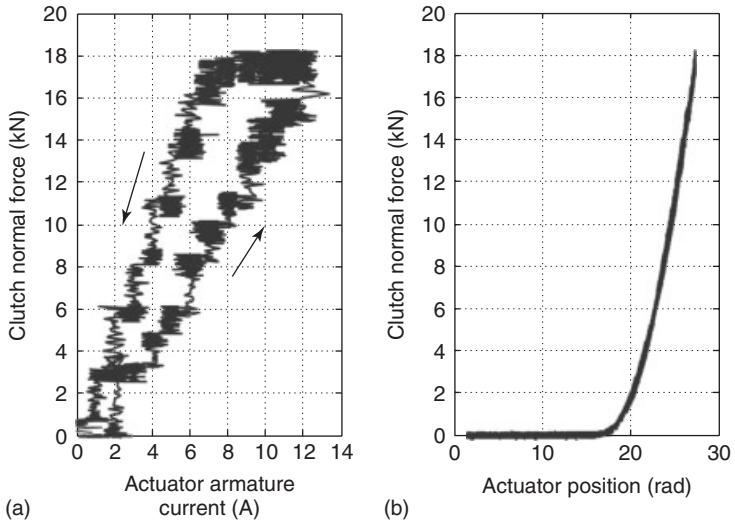


Figure 11. (a and b) Static characteristics of ALSD ball ramp actuation system. (Adapted from Ivanović, 2012a. © Sage.)

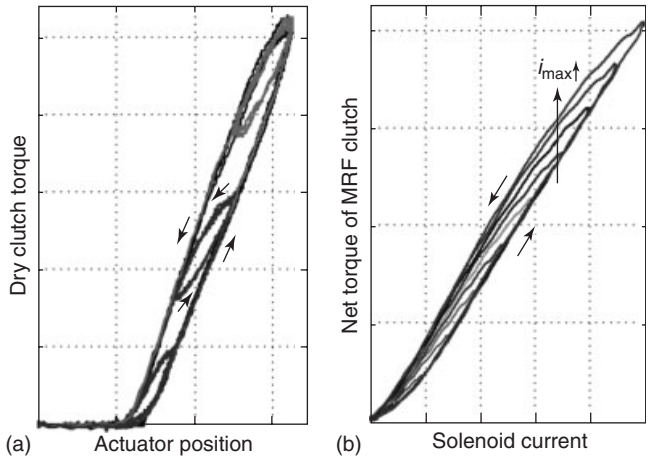


Figure 12. Recorded actuator hystereses for lever-based dry clutch (a) and MRF clutch (b).

actuator can be regarded as a force/torque source device, and any friction resistance in the actuator reduces the net normal force. This results in a clockwise hysteretic dependence between the actuator force (or current, voltage, and pressure) and the clutch normal force F_n, as illustrated in Figures 11a and 12a on the examples of ball ramp wet clutch actuation (Ivanović *et al*., 2012a) and lever-based dry clutch actuation (Ivanović, Deur, and Tseng, 2012b), respectively. As the ball ramp load corresponds to the normal force (Section 2), the ball ramp friction grows with the normal force, that is, the hysteresis width increases with F_n (Figure 11a). In the presence of dynamic friction effects (Armstrong-Hélouvry *et al*., 1994), the friction hysteresis may become narrower for the narrower-operating region (Figure 12a).

Although the friction compensation algorithms have been widely applied in *closed-loop*, speed- and position-controlled servo systems (Armstrong-Hélouvry *et al*., 1994), they cannot be equally effectively applied in an *open-loop* force/torque control system. This is due to the fact that the clutch actuation mechanism is often in or close to the zero-speed (stiction) region, which is relevant for force control and where it is difficult to estimate friction. Gassmann and Barlage (2004) propose a robust (nonmodel-based) friction compensation method based on the injection of oscillatory signal (the so-called dither) into the actuator DC motor armature voltage reference. They demonstrate that the hysteresis width can be significantly reduced when the dither signal is applied.

One of the major advantages of the position-controlled clutch actuation systems is that they are not prone to friction hysteresis, provided that a part of the actuation mechanism upstream the element with dominant friction is rather stiff. Figure 11b indicates that the same ball ramp actuator has a hysteresis-free static curve when it is expressed in the actuator motor position instead of the motor current. Therefore, the position-controlled system has a good potential for favorable clutch torque control accuracy, provided that the normal force versus position map in Figure 11b is independent of operating parameters (e.g., oil temperature) or that it can be adapted online. Such a control system may be extended with a friction compensator for further performance improvement (e.g., reduction of an initial pure delay due to the friction-induced standstill interval).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.1.3 Magnetic hysteresis

Magnetic hysteresis $B(H)$ is an inherent property of the ferromagnetic materials. In the clutch actuation systems, it manifests itself in a similar manner as the friction hysteresis, which is illustrated by the comparison of lever-based and MRF clutch actuator static curves in Figure 12. The MRF clutch shows a significant effect of narrowing hystereses (Figure 12b). Although this effect can be rather accurately modeled (Deur, Herold, and Kostelac, 2009), the effectiveness of model-based compensation may be doubtful, particularly under the (quasi)steady-state conditions. Therefore, it is of importance to design the electromagnetic clutches (including the MRF/MP ones) using the soft ferromagnetic materials having a narrow hysteresis.

3.1.4 Structural compliance in clutch actuation mechanism

The force/torque-controlled actuation system should be rather stiff, in order to transfer the generated actuator force to the clutch normal force without generating axial/torsional vibrations. It should be noted that although the actuator friction typically acts as a strong vibration damper, notable vibrations may still be excited for abrupt transient (e.g., clutch landing, see Section 3.2) if the actuation mechanism is not stiff and the control strategy is relatively crude (soft landing and/or active damping routines are absent).

On the other hand, for position-controlled or combined actuation systems, there should be compliance downstream the element with dominant friction (cf. Section 3.1.2). This is because the actuator motor position control is based on the actuator stress–strain curve (Figure 11b), which cannot then be very stiff in order to avoid a very narrow range of motor position control. Examples of compliance elements in position-controlled and combined systems are ball ramp compliance (Figure 8a) and diaphragm spring support point compliance (Figure 1d).

3.1.5 Fluid film squeeze resistance

During the wet clutch engagement, the fluid first need to be squeezed from the clutch before the asperity contacts are established and the clutch torque is developed. The fluid film squeeze process is characterized by generation of the fluid resistive force that should be considered in the clutch actuator design stage. Assuming an ungrooved, nonpermeable plate clutch (Figures 1a and 2), the fluid resistive force is given by (Berger, Sadeghi, and Krousgrill, 1997):

$$F_{fs} = -\underbrace{12\eta Q(r_\mathrm{i}, r_\mathrm{o})\frac{1}{h^3}}_{b_\mathrm{fs}(h)}\frac{\mathrm{d}h}{\mathrm{d}t} \tag{2}$$

where

$$Q(r_\mathrm{i}, r_\mathrm{o}) = \frac{r_\mathrm{o}^4 - r_\mathrm{i}^4}{16} - \frac{r_\mathrm{i}^2(r_\mathrm{o}^2 - r_\mathrm{i}^2)}{8} - \frac{(r_\mathrm{o}^2 - r_\mathrm{i}^2)^2}{16\ln(r_\mathrm{i}/r_\mathrm{o})} - \frac{r_\mathrm{o}^4 - r_\mathrm{i}^2 r_\mathrm{o}^2}{8}$$

where h is the effective fluid film thickness, $\mathrm{d}h/\mathrm{d}t$ the fluid film squeeze speed, and η the fluid dynamic viscosity. For the clutch with the number of fluid gaps denoted by N'_f (note that $N'_\mathrm{f} = N_\mathrm{f}$ is valid for the typical case of double-sided friction plate), the fluid film thickness and speed may be calculated as $h = (x_\mathrm{pp,max} - x_\mathrm{pp})/N'_\mathrm{f} + h_\mathrm{min}$ and $\mathrm{d}h/\mathrm{d}t = v_\mathrm{pp}/N'_\mathrm{f}$ (Ivanović *et al*., 2012a), respectively. Here, the variables x_pp and v_pp denote the press plate position and speed, respectively, and $x_\mathrm{pp,max}$ and h_min are the press plate position and the film thickness at the rated normal force F_n (note that owing to asperity roughness, there is a residual $h_\mathrm{min} \neq 0$ even for large/rated F_n).

Ivanović *et al*. (2012a) have experimentally identified the damping factor $b_\mathrm{fs}(h)$ in Equation 2 for the active differential, grooved wet clutch from Figure 8a, and shown that it accurately satisfies the inversely proportional relation $b_\mathrm{fs}(h) \sim h^{-3}$, with the h_min identified to be 12 μm. However, the damping factor magnitude was fourfold smaller than the one predicted by Equation 2, which was explained by the neglected fluid flow through the grooves and the friction material permeability. The fluid resistive force F_fs has been found to be two order of magnitudes lower than the rated clutch normal force, which is explained by a large rated normal force (40 kN), relatively low maximum press plate speed (approximately 2 mm/s), and a large number of fluid gaps ($N_\mathrm{f} = N'_\mathrm{f} = 20$). This means that the resistive force F_fs may be neglected from the standpoint of actuator design of the particular ALSD clutch. Of course, this conclusion may not hold for other wet clutches, such as those used in ATs.

3.1.6 Friction coefficient variation

The experimentally identified, multidimensional dependence of the clutch friction coefficient μ in Equation 1 is shown in Figure 13 for the case of ALSD wet clutch (Ivanović *et al*., 2012a). The clutch's separator plates are made of hardened steel and the friction plates contain friction material made of woven carbon fabric to which a phenolic resin has been applied. The identification results indicate that the coefficient of friction (COF) varies in the range from 0.06 to 0.12, which points to the boundary lubrication operating conditions (Hamrock, Schmid, and Jacobson, 2004). The COF decreases with rise of the clutch friction interface temperature ϑ and the normal force F_n. In the low slip speed range, it predominantly shows a rising

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

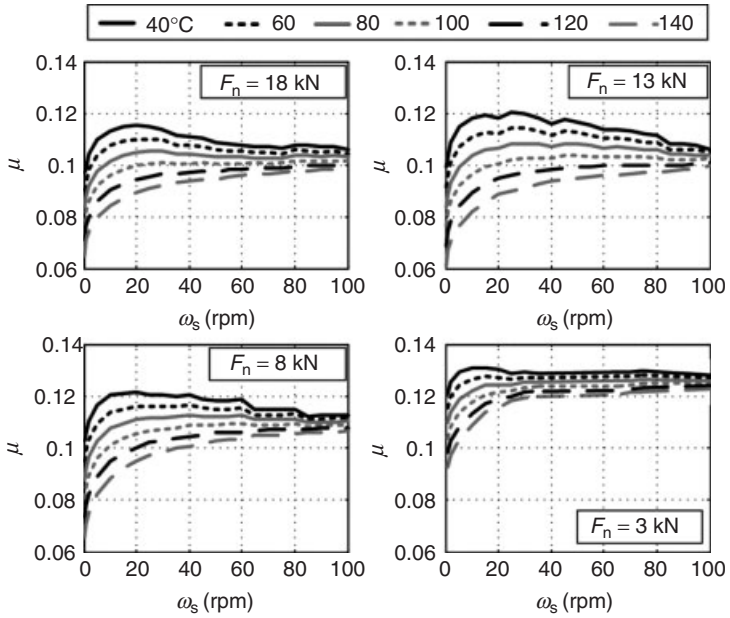


Figure 13. Static curves of ALSD wet clutch friction coefficient. (Reproduced from Ivanović, 2012a. © Sage.)

dependence on the slip speed ω_s. Exceptionally, this dependence can have a negative gradient at low temperatures. The negative gradient is generally undesirable, as it relates to a negative friction damping coefficient that can cause judder vibrations (Crowther *et al*., 2004).

The COF typically reduces with the clutch wear. This property can be used for the purpose of clutch wear monitoring for those actuators that are not equipped with the position sensor (Fei *et al*., 2008), for example, the hydraulic ones used in the AT applications. Moreover, the clutch/fluid wear is usually associated with the increasing sensitivity to negative COF gradient related to the judder appearance.

The dry clutch COF usually varies in the range approximately from 0.3 to 0.4. It generally shows similar trends in terms of dependence on slip speed, temperature, and normal force as in the case of wet clutch, including the likely appearance of negative gradient at low slip speeds and low temperatures (Deur *et al*., 2012).

3.1.7 Clutch pack thermal effects

During the clutch engagement, the clutch dissipates the power $\Phi_f = \tau_c \omega_s$ (Figure 2 and Equation 1), which is converted into the heat flux that warms up the clutch. As the clutch pack heats up, it thermally expands, thus potentially affecting the accuracy of normal force control. The effect occurs with the position-controlled actuation systems (and to a lower extent, with the combined system), because for a given actuator/press plate position, the temperature rise-related thermal expansion drift causes the normal force rise and potentially large clutch torque control errors (Deur *et al*., 2012). The problem is more emphasized if the thermal expansion is stronger (e.g., for dry clutches) and if the actuation mechanism is stiffer. It can be overcome by estimating the thermal expansion-related overlap and correspondingly reducing the actuator position reference to keep the normal force equal to the demanded force (Section 4).

The thermal expansion compensation, thus, requires estimation of the clutch interface temperature ϑ (the separator plate temperature for wet clutch and the flywheel/press plate temperature for dry clutch). The temperature information is also needed for friction coefficient estimation, as well as for the clutch temperature monitoring. The temperature estimator is based on online simulation of a lumped parameter clutch thermal model (open-loop estimation). Although the heat transfer concerns multiple distinctive thermal masses (see the example of a wet clutch in Figure 14), the model may be described by a single differential equation (Ivanović *et al*., 2012a):

$$C_{sp}\dot{\vartheta} = \Phi_f - H_e(\omega_s, \tau_c)(\vartheta_c - \vartheta_{oil}) \tag{3}$$

where $C_{sp} = m_{sp} c_{steel}$ is the separator plate thermal mass and ϑ_{oil} the measurable oil sump temperature. The heat transfer factor H_e is a rising (typically linear) function of the slip speed ω_s, which accounts for the clutch fluid pumping

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

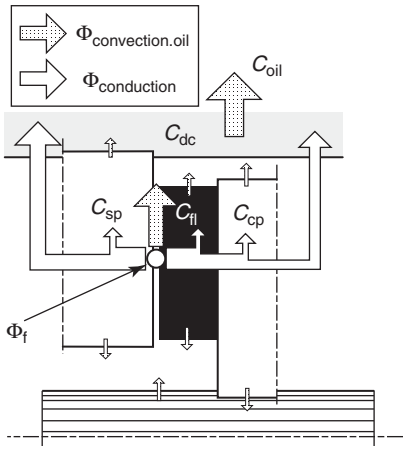


Figure 14. Schematic diagram of wet clutch thermal model. (Reproduced from Ivanović, 2012a. © Sage.)

effect. The factor H_e is also found to be linearly dependent on the clutch torque τ_c, which is predominantly due to the reduced-order model structure. For a favorable model accuracy, the factor H_e may also include an empirically tuned time-variant lag dynamics. The dependence on τ_c and the lag dynamics diminish if the full, third-order model is designed according to the scheme in Figure 14 (Ivanović *et al*., 2012a). However, parameterization of such a model is more complex and it may turn out to be less accurate than the simple model (3).

For the dry clutch, the press plate temperature ϑ_{pp} can be significantly higher than the flywheel temperature ϑ_{fw} (owing to a smaller press plate mass), and the yoke temperature ϑ_{yo} is typically significantly lower than the flywheel temperature and higher than the bell-housing temperature ϑ_{bh}. Therefore, the model should include three state variables to be estimated (ϑ_{pp}, ϑ_{fw}, and ϑ_{yo}, with the former two being equal to the friction interface temperatures) (Hoić *et al*., 2012), whereas the bell-housing temperature ϑ_{bh} can be either measured or estimated from external, power train temperatures. The thermal expansion model is also generally more complex for the dry clutch when compared to the wet clutch.

3.2 Main features of clutch torque transient response

Figure 15 shows the ALSD wet clutch response for a crude, torque-based clutch control, characterized with the maximum (uncompensated) clutch clearance and the absence of soft landing intervention. This worst-case/low performance scenario is intentionally used to clearly illustrate the main effects of the clutch response and give motivation for control system improvements described in Section 4. The response with respect to actuator DC motor armature current reference step is divided into four phases, for which the main influencing effects are listed in the legend of Figure 15 (Ivanović *et al*., 2012a).

The initial, *pure delay phase* is predominantly determined by the clutch clearance, and also with the motor speed limit, and to a lower extent with return spring, friction, and fluid squeeze loads. The *clutch activation dynamics* is characterized by an abrupt lag dynamics with a significant clutch torque overshoot and a certain oscillatory behavior. The overshoot and oscillatory behavior

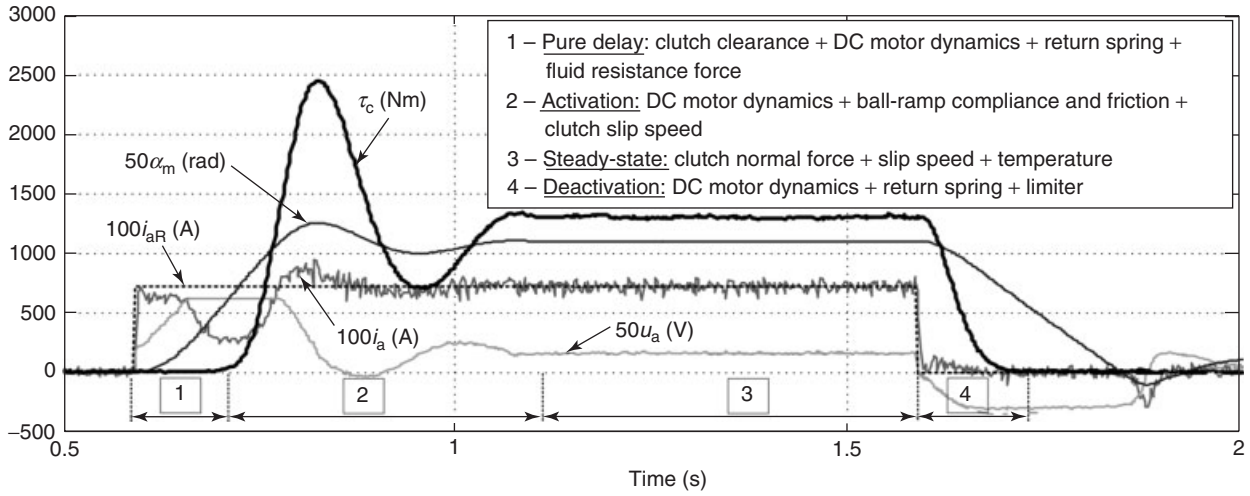


Figure 15. Active differential wet clutch response for crude, torque-based clutch control. (Reproduced in part from Ivanović, 2012a. © Sage.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

are due to the uncontrolled high speed landing in combination with the actuator inertia and compliance effects. The response damping is relatively good because of the significant amount of friction. It should be noted that the response can be significantly slower for very small clutch slip speeds, because during the transient, the clutch can then become locked, and thus uncontrollable (Ivanović *et al*., 2012a).

The steady-state clutch torque is determined by the developed normal force and the clutch friction coefficient, Equation 1, as well as by the actuator mechanism friction hysteresis (Figure 11a). The clutch torque response during the *clutch deactivation phase* is relatively fast, as the actuator motor torque is zero and the return spring is strong enough to quickly disengage the ball ramp actuator. The motor position response back through the clearance region is somewhat slower and shows some "bouncing" when "hitting" the mechanical limiter position.

4 CLUTCH ACTUATION CONTROL SYSTEM

On the basis of the insights and control/estimation ideas described in Section 3, this section presents details of the two distinctive clutch control concepts. Without a loss of generality, the presentation utilizes the example of ALSD wet clutch (Figure 8a).

4.1 Force/torque-controlled system

Here, the actuator directly commands a variable that is related to the clutch normal force (or flux density for the MRF/MP clutches). This can be armature voltage, solenoid voltage, or fluid flow. In order to provide more accurate and generally faster control, it is more appealing to command the variable that is directly proportional to the actuator torque/force, and that would be the armature current, solenoid current, or fluid pressure. However, such variable should be sensed to be fed to the corresponding controller, which increases the actuator price and is, thus, rather prohibitive for multiclutch systems such as ATs.

The open-loop force/torque controller may be divided into three subsystems: (i) dynamic subsystem for reference shaping including clearance compensation and soft landing, (ii) actuator command mapping, and (iii) friction compensation. In the simplest, purely open-loop applications (e.g., those in ATs) there is no feedback information from the actuation system, so that the clutch engagement can solely be controlled by shaping the time response of reference variable (e.g., valve solenoid voltage). In the more demanding, continuous torque control application (e.g., driveline system clutches or startup clutches), the clutch actuators are usually equipped with sensors such as motor position/speed or sometimes current/pressure sensors. This facilitates the application of auto-tuned clearance compensation and soft landing control intervention, where both rely on the "bite"-point detection based on observing a sudden peak in actuator motor armature voltage/current or speed when the asperity contact is established. In this case, the clutch torque demand can be transformed to the actuator command mapping block without shaping. The transformation map may also include the clutch/oil sump temperature input, because it notably determines the torque control accuracy. Separate maps can be used to correct for the influence of other factors such as clutch slip speed, motor winding temperature, and clutch wear. The maps are identified experimentally for a wide range of conditions, that is, some significant calibration effort is needed. Finally, the actuator command (the armature voltage u_a in the case of ALSD clutch) can be perturbed by a dither signal to reduce the effect of actuator transmission friction (Gassmann and Barlage, 2004).

The lever-based actuator (Figure 8c) may be regarded as an indirect force-source system, so that the force-controlled structure can conveniently be applied to it, as well. The fact that the engagement bearing force does not solely depend on the actuator command—the fulcrum/motor position, but also on the lever angle depending on the engagement bearing position/load, does not have a major consequence on the force-controlled system structure since the "combined force/position control effect" can be captured by the proper multi-dimensional actuator command mapping. Alternatively, the physical model-based mapping may be used.

4.2 Position-controlled system

The main disadvantage of force/torque-controlled system relates to difficulties with friction compensation, which can result in significant static and transient errors in actuators with dominant transmission friction effects (e.g., electromechanical actuators, Sections 2 and 3). This can be overcome using the position-controlled system structure, which is shown in Figure 16 based on Ivanović *et al*. (2009) and Deur *et al*. (2012). Using the estimated clutch coefficient of friction μ (COF, Section 3), the demanded torque τ_{cR} is transformed to the normal force demand F_{nR}. In order to avoid the algebraic loop, the COF-map's input F_n should be obtained by filtering F_{nR} or directly from τ_{cR} using a prescribed/constant COF. The normal force demand

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

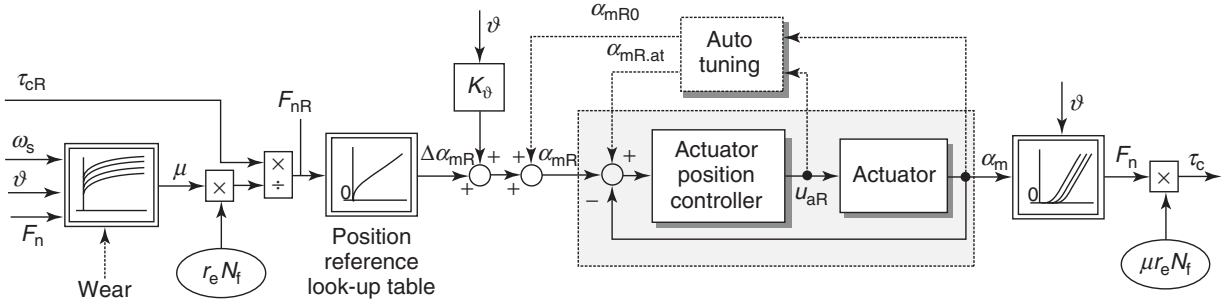


Figure 16. Block diagram of position-controlled system. (Adapted with permission from Ivanović, 2009. © ASME.)

F_{nR} is referred to the actuator motor position using the empirically determined 1D map (Figure 11b). A 2D map can alternatively be used if the transmission compliance characteristic $F_n(\alpha_m)$ is significantly influenced by the (oil) temperature.

The obtained, net actuator position command $\Delta\alpha_{mR}$ is corrected by the position offset calculated from the clearance compensation auto-tuning routine. As the thermal expansion-related position offset may be regarded as a negative clearance (or overlap), the thermal expansion effect may be compensated for by adding the actuator position reference offset calculated as a product of the estimated friction interface temperature ϑ and a negative value of the thermal expansion gain K_ϑ. The gain K_ϑ can be obtained by test rig-based experimental identification, that is, from the gradient of actuator position versus temperature curves for the constant/regulated normal force. Alternatively, the corrective, slowly changing, thermal- and other disturbance-induced position offsets can be calculated from a slow integral-type clutch torque controller, which utilizes the engine torque observer during the intervals of reliable engine torque estimation (Deur *et al*., 2012).

The auto-tuner estimates the clutch clearance and tunes the corresponding actuator position reference offset α_{mR0}. It is executed sporadically when the clutch is not in use (e.g., every time the engine is turned off). The clearance identification is based on ramping up the actuator position reference $\alpha_{mR,at}$, and detecting/holding the position at which the armature voltage suddenly rises, indicating the motor torque rise at the "bite" point.

The experimental response in Figure 17 illustrates the effectiveness of the laboratory-tested actuator position-based clutch torque control system in both small- and large-signal operating modes. The fast, aperiodic, and steady-state accurate response is achieved in different operating modes and operating points without using friction compensator. The response is very repeatable and steady-state

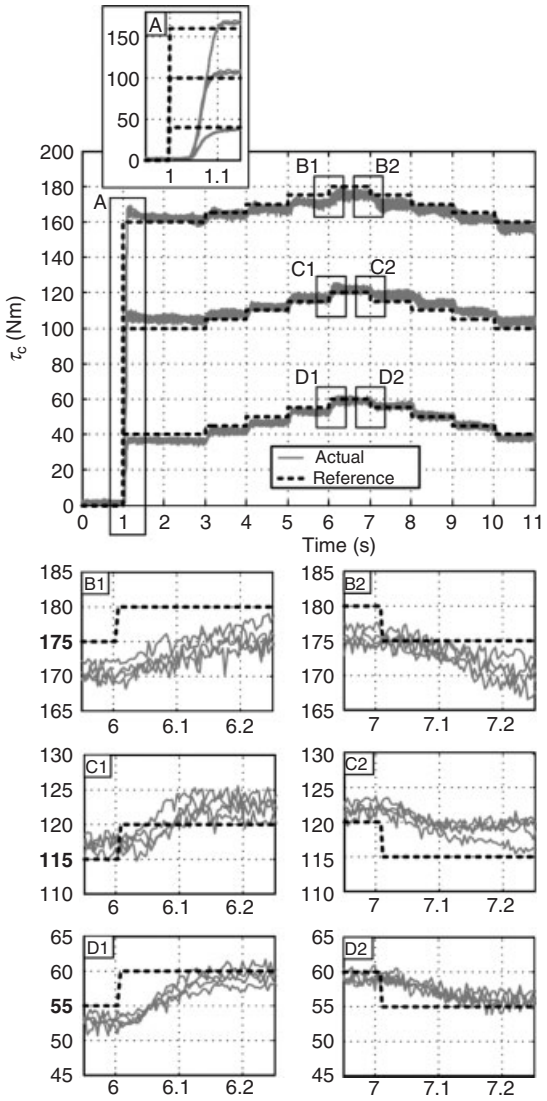


Figure 17. Experimental verification results for position-controlled system. (Adapted with permission from Ivanović, 2009. © ASME.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

torque error is lower than 10% for the considered case of cool oil.

5 CONCLUSION

In parallel to the introduction of new concepts of clutch-based automatic transmission (AT) and active driveline systems, there have been significant research and development efforts in designing novel clutch actuation systems. As most of the new transmission and driveline systems are based on a single or dual clutch, where often a continuous clutch torque control is required, there have been strong opportunities and demands for developing more advanced clutch actuation systems including feedback control subsystems. The traditional electrohydraulic actuators have, therefore, been facing a strong competition, mostly from various designs of electromechanical actuators, and also from electromagnetic and sometimes MRF/MP actuators.

The hydraulic actuators remain to be exclusively used with torque converter ATs, because of a superior force density and favorable packaging features for such multi-wet clutch systems. The electromechanical actuators gain their presence in various driveline applications and in some specific transmissions such as those based on dry clutches. This is because of their good modularity/ease of assembly, high efficiency, and excellent control features of EMs. The direct electromagnetic actuators have a low force density, and they should, therefore, be used in high speed/low torque clutch applications (e.g., active differentials with geared stationary clutches/brakes) or they should be combined with a mechanical transmission (e.g., ball ramp mechanism). The MRF/MP clutches offer a strong potential in terms of great simplification of mechanical system and excellent control performance; however, they are still rarely applied in automotive drive trains because of the inferior torque density compared to friction clutches. To successfully compete with the electromechanical actuators in the demanding applications, the electrohydraulic actuators have been evolved toward more efficient designs (variable-flow pumps are employed instead of valves) and improved control features (pressure and/or position feedback is utilized).

The success of modern clutch actuation systems is strongly associated with the performance of clutch torque control strategies. On the other hand, the control system design and calibration largely depends on the quality of experimental characterization and modeling of the clutch actuator dynamics, as well as on the choice and characteristics of sensors. Hence, to achieve a favorable performance-to-cost ratio, the design of the overall clutch actuation system requires a fully mechatronic approach from the early stage of development.

ACKNOWLEDGMENT

It is gratefully acknowledged that the authors' research work on the electromechanical, MRF, and MP clutch actuation systems, whose results are partly presented in this chapter, has been supported by the Jaguar Cars Ltd and the Ford Motor Company.

REFERENCES

Armstrong-Hélouvry, B., Dupont, P., and Canudas de Wit, C. (1994) A survey of models, analysis tools and compensation methods for the control of machines with friction. *Automatica*, **40**, 419–425.

Balau, A.-E., Caruntu, C.-F., and Lazar, C. (2011) Simulation and control of an electro-hydraulic actuated clutch. *Mechanical Systems and Signal Processing*, **25**, 1911–1922.

Berger, E.J., Sadeghi, F., and Krousgrill, C.M. (1997) Analytical and numerical modeling of engagement of rough, permeable, grooved wet clutches. *ASME Journal of Tribology*, **119**, 143–148.

Carlson, J.D. (2001) What Makes a Good MR Fluid? *8th International Conference on Electrorheological (ER) Fluids and Magneto-rheological (MR) Suspensions*, Nice, France, www.lordcorp.com (accessed 30 August 2013).

Cho, B.-H., Oh, J.-S., and Lee, W.-H. (2002) Modeling of pulse width modulation pressure control system for automatic transmission. SAE paper #2002-01-1257.

Crowther, A., Zhang, N., Liu, D.K., and Jeyakumaran, J.K. (2004) Analysis and simulation of clutch engagement judder and stick–slip in automotive powertrain systems. *Proceedings of the Institution of Mechanical Engineers, Part D: J. Automobile Engineering*, **218**, 1427–1446.

Deur, J. Herold, Z., and Kostelac, M. (2009) Modeling of Electromagnetic Circuit of a Magnetorheological Fluid Clutch. *Proceedings of 2009 IEEE Multi-conference on Systems and Control*, St. Petersburg, Russia.

Deur, J., Ivanović, V., Herold, Z., *et al*. (2012) Dry clutch control based on electromechanical actuator position feedback loop. *International Journal of Vehicle Design*, **60**, 305–326.

Fachkunde Kraftfahrzeugtechnik (n.d.) 27. Auflage, Verlag Europa-Lehrmittel (Croatian translation). ISBN: 953-6054-95-7.

Fei, J., Li, H.-J., Qi, L.-H., *et al*. (2008) Carbon-fiber reinforced paper-based friction material: study of friction stability as a function of operating variables. *Journal of Tribology*, **130**, 1–7.

Francis, P., Haselton, D., and Pritchard, L. (2006) Pre-emptive torque management™ (PTM)™. SAE paper #2006-01-0817.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fujii, Y., Tobler, W.E., and Snyder, T.D. (2001) Prediction of wet band brake dynamic engagement behaviour, Part 1: mathematical model development. *Proceedings of the Institution of Mechanical Engineers (IMechE)*, **215** (D), 479–492.

Gassmann, T. and Barlage, J.A. (2004) Electronic torque manager (ETM): an adaptive driveline torque management system. SAE paper #2004-01-0866.

Hadler, J., Metzner, F.-T., Schäfer, M., *et al*. (2008) The seven-speed dual clutch transmission from Volkswagen. *ATZ*, **110** (11), 26–33.

Hamrock, B.J., Schmid, W.R., and Jacobson, B.O. (2004) *Fundamentals of Fluid Film Lubrication*, second edn, Marcel Dekker, Inc., New York-Basel

Hoić, M., Deur, J., Herold, Z., and Ivanović, V. (2012) Modeling of Dual Dry Clutch Thermal Dynamics. *2012 Powertrain Modelling and Control Conference*, Bradford, UK.

Holmes, G.R. and McKenna, M.D. (2001) Proportional variable bleed solenoid valve with single adjustment pressure calibration and including poppet valve seal ball. Patent No. US 6,305,664 B1,

Hrovat, D., Asgari, D., and Fodor, M. (2000) Automotive mechatronic systems in *Mechatronic Systems, Techniques and Applications: Vol. 2—Transportations and Vehicle Systems* (ed. C.T. Leondes), Gordon and Breach Science Publishers, Amsterdam, pp. 1–98.

Ivanović, V., Deur, J., Hancock, M., and Assadian, F. (2009) A Closed-Loop Strategy of Active Differential Clutch Control. *Proceedings of 2009 ASME Dynamic Systems and Control Conference (2009 DSCC)*, Hollywood, CA.

Ivanović, V., Deur, J., Herold, Z., *et al*. (2012a) Modelling of electromechanically actuated active differential wet-clutch dynamics. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, **226** (4), 433–456.

Ivanović, V., Deur, J., Milutinović, M., and Tseng, H.E. (2011) Dynamic Model of Dual Clutch Lever-Based Electromechanical Actuator. *Proceedings of 2011 ASME Dynamic Systems and Control Conference: Automotive and Transportation Systems*, Arlington, VA.

Ivanović, V., Deur, J., and Tseng, H.E. (2012b) Bond Graph Model of Electromechanical Actuation System for a Dry Dual Clutch. *Proceedings of 10th International Conference on Bond Graph Modeling and Simulation*, July 8–11, Genoa, Italy.

Jolly, M.R., Bender, J.W., and Carlson, J.D. (1998) *Properties and applications of commercial magnetorheological fluids*. SPIE 5th Annual International Symposium on Smart Structures and Materials, San Diego, CA.

Kawai, N., Honma, K., Takigawa, H., *et al*. (1988) Production, compaction and application of metal powders. *Metal Powder Report*, **43** (1), 21–25.

Kondo, M., Hasegawa, Y., Takanami, Y., *et al*. (2007) Toyota AA80E 8-speed automatic transmission with novel powertrain control. SAE paper #2007-01-1311.

Kunii, R., Iwazaki, A., Sekiya, S., *et al*. (2005) Development of direct electromagnetic clutch system. SAE paper #2005-01-0551.

Lampe, D. (2000) *Untersuchungen zum Einsatz von magnetorheologischen Fluiden in Kupplungen*. Dissertation (Ph. D. Thesis). TU Dresden.

Lampe, D., Thess, A., and Dotzauer, C. (1998) MRF-Clutch-Design Considerations and Performance. *The 6th International Conference on New Actuators*, *Actuator 1998*, Bremen, Germany.

Langjord, H. and Johansen, T.A. (2010) Dual-mode switched control of an electropneumatic clutch actuator. *IEEE-ASME Transactions on Mechatronics*, **15** (9), 1–8.

Lee, G.S., Sung, H.J., Kim, H.C., and Lee, H.W. (2010) Flow force analysis of a variable force solenoid valve for automatic transmissions. *Journal of Fluids Engineering*, **132**, 1–7.

Lewis, C. and Bollwahn, B. (2007) General motors hydra-matic & Ford new FWD six-speed automatic transmission family. SAE paper #2007-01-1095.

Merritt, H.E. (1967) *Hydraulic Control System*, John Wiley & Sons, Inc., New York

Montanari, M., Ronchi, F., Rossi, C., *et al*. (2004) Control and performance evaluation of a clutch servo system with hydraulic actuation. *Control Engineering Practice*, **12**, 1369–1379.

Morselli, R., Zanasi, R., Cirsone, R., *et al*. (2003) Dynamic modeling and control of electro-hydraulic wet clutches. *IEEE-Intelligent Transportation Systems*, **1**, 660–665.

Mustafa, R., Kassel, T., Alvermann, G., and Kucukay, F. (2010) Modelling and Analysis of the Electro-Hydraulic and Driveline Control of a Dual Clutch Transmission. *Proceedings of the 2010 FISITA*, 2010-SC-O-18.

Nilsson, J., Herven, D., Dahlström, P., and Severinsson, L. (2011) A hydraulic pump assembly. Patent application WO 2011/043722 A1.

Oberlack, N. and Reul, A. (2006) Electro-hydraulic actuators for control applications in clutches and transmissions. *ATZ*, **108** (7–8), 570–578.

Rabinow, J. (1948) The magnetic fluid clutch. *AIEE Transactions*, **67**, 1308–1315.

Ross, C.S., Carney, C.E., Schanz, T., and Gaffbey, F. (2007) Development of an electronically-controlled limited-slip differential (eLSD) for FWD applications. SAE paper #2007-01-0925.

Sackl, W., Eibler, G., and Linortner, T. (2008) Torque vectoring with electro-hydraulic actuation. *ATZ*, **110** (12), 20–26.

Sakai, Y. (1988) The "ECVT" electro continuously variable transmission. SAE paper #880481.

Turner, A.J. and Ramsay, K. (2004) Review and development of electromechanical actuators for improved transmission control and efficiency. SAE paper #2004-01-1322.

Wagner, U., Bührle, P., Müller, B., *et al*. (2009) Dry double clutch system: innovative components for highly efficient transmissions. *ATZ*, **111** (11), 28–33.

Watechagit, S. and Srinivasan, K. (2003) Modeling and simulation of a shift hydraulic system for a stepped automatic transmission. SAE paper #2003-01-0314.

Wheals, J.C., H. Baker, K. Ramsey, and W. Turner (2004) Torque vectoring AWD driveline: design, simulation, capabilities and control. SAE paper # 2004-01-0863.

Wheals, J.C., McMicking, J., Shepherd, S., *et al*. (2009) Proven high efficiency actuation and clutch technologies for eAMT™ and eDCT™. SAE paper #2009-01-0513.

Yoshioka, S., Ononkuchi, I., and Inoue, N. (1985) Trend of automobile's electro-hydraulic control. Electronics Research Institute Report, 1–19.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ZF Friedrichshafen AG (n.d.) Clutch Systems, http://www.zf.com/media/media/en/document/corporate_2/downloads_1/flyer_and_brochures/cars_flyer/kupplungssystemefrpkwbis800nm.pdf (accessed 30 August 2013).

Zheng, Q., Kraenzlein, J., Hopkins, E., *et al*. (2009) Closed loop pressure control system development for an automatic transmission. SAE paper #2009-01-0951.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto102
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Launch Control

Andrew P. Harrison
Drive System Design Ltd, Warwickshire, UK

1 Introduction 1
2 Main Text 1
Glossary 7
Reference 7

1 INTRODUCTION

Launch control means many things to different people. It can mean simply the method by which a vehicle accelerates from rest or a dedicated function designed to extract the maximum possible acceleration from a vehicle in a race or simulated race start event.

The manner in which a vehicle moves at low speed can be a dominant feature in determining both the subjective and the objective performances of a vehicle. Traditional automatic transmissions employed a torque converter as their launch device (*see* Automatic Transmissions—Geartrain Combinations, Components, Design Considerations, Hydraulic System, Packaging, Manuf., Assembly). Tuning of low speed attributes including creep, torque multiplication, response, and feel was traded against cost, efficiency, package, and weight through careful manipulation of the component design (*see* Automotive Torque Converters). The continuous search for better efficiency has led to the development of new transmission concepts (*see* General Introduction—Basic Definitions, Structure of Part 4), many of which employ friction clutches as the launch device. By combining a friction clutch with an electronic control system, it is possible to maintain many of the desirable attributes of the torque converter while improving overall system efficiency and also introducing additional features and modes of operation.

This chapter discusses first the attributes most desirable of a modern launch control system and then moves on to discuss common control techniques applied to friction clutches in the modern automobile.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto103
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 MAIN TEXT

In this chapter, "launch control" is used as an umbrella term including all aspects of the control of a vehicle moving at low speed and directly or indirectly influenced by the launch device. Detailed discussion is focused on the subjects of low speed maneuvering (creep and hill hold), vehicle launch for maximum subjective feeling (pullaway control), and vehicle launch for maximum acceleration (performance launch control).

2.1 Torque converters and their role in defining the ideal pullaway characteristic

The torque converter has been the main launch device for automatic transmissions (*see* Automatic Transmissions—Geartrain Combinations, Components, Design Considerations, Hydraulic System, Packaging, Manuf., Assembly) for many decades and therefore is for many drivers their first or only experience of an automated launch device. Two inherent features of a torque converter, which are seen by many as desirable attributes to be maintained in any alternative concept, are that of creep and hill hold.

2.1.1 Creep

The term *creep* is used to describe the tendency of a torque converter-equipped vehicle to want to move at a low speed whenever drive is engaged. Holding the vehicle at rest or travelling at a speed below the creep speed of the vehicle requires the driver to apply the brakes or select neutral.

Creep is seen as a positive attribute of a torque converter, assisting the driver in maintaining control of a vehicle during very low speed maneuvers by facilitating the use of a single pedal to control the vehicle motion. Therefore, while not universal in its implementation, it is usual to design the control system for a launch clutch such that it partially engages the launch clutch in this condition and simulates the creep behavior of a torque converter-equipped vehicle.

A negative side effect of the creep torque applied by the torque converter is the additional engine load, which can significantly increase idle fuel consumption and emissions. Therefore, many systems have been developed to reduce this load by temporarily selecting neutral when the brake pedal is applied. One such early system and its operation are described by Forster and Eltze (1978). In this system, the transmission temporarily selects neutral via a switch mounted on the brake pedal activating a servo valve, which in turn removes the control pressure from the clutch. Modern systems recreate this functionality through electronic control of clutch apply pressure.

It is desirable that a launch clutch that features simulated creep, be it mechanically or electronically controlled, should maintain this feature and reduce the load on an engine because of simulated creep torque when the brake is applied.

2.1.2 Hill hold

Hill hold is used to describe the ability of a vehicle to hold stationary on an uphill gradient without the brakes being applied.

The creep torque inherent of a torque converter is a function of the slip across the converter (*see* Automotive Torque Converters) and therefore changes with both the engine and the vehicle speed. A vehicle in creep on a level road will achieve a constant speed when the slip reduces sufficiently that the wheel torque generated by the torque converter is equal to the road load experienced by the vehicle.

Increasing the gradient increases the road load and therefore equilibrium is achieved at a lower vehicle speed. Eventually, a gradient can be achieved at which the creep torque is no longer sufficient to overcome the road load and the vehicle is held at rest.

Further, increasing the gradient will eventually cause the vehicle to start to roll backwards, increasing the slip across the converter and in turn increasing the torque at the wheels. Although on this higher gradient the torque converter is unable to prevent the vehicle rolling back, it will still provide some significant restoring torque helping the driver to maintain control and/or launch the vehicle on the gradient.

The gradient on which the vehicle will fail to accelerate from rest and the gradient on which the vehicle will start to roll back are significantly different values. This hysteresis is created by the tire static rolling resistance and the static friction in the driveline. It is the ability of a torque converter-equipped vehicle to hold on this range of gradients, which is described as its inherent hill hold capability.

2.2 Launch control for maximum acceleration performance

In motorsport applications, the term *launch control* is often used to describe a system that, through a combination of traction control and clutch torque capacity regulation, attempts to achieve the maximum possible acceleration of a traction-limited vehicle. Seen as a driver aid, capable of achieving significant improvement in both ultimate vehicle performance and repeatability of race starts, such systems have been banned from Formula 1 since the start of the 2008 season.

Similar systems applied to road cars may be considered to be enhancing both performance and safety of the vehicle.

Vehicle performance is improved because of both the maximum vehicle acceleration and the maximum gradeability of the vehicle being increased.

Vehicle safety is increased by allowing an unskilled driver to accelerate away from a hazard in the minimum period of time.

2.2.1 Maximizing available wheel torque

The ideal launch control system will maintain the vehicle at the traction limit from the moment the launch is initiated until a vehicle speed is achieved beyond which insufficient power is available.

With a friction clutch launched vehicle, maximum wheel torque is available when the transmission is in its lowest gear (maximum torque multiplication) and the engine is at peak torque engine speed. Therefore, the simplest approach to maximize the wheel torque available to the launch control system is to control the engine to the speed at which it is capable of delivering maximum engine torque. This approach is often used on low powered vehicles and particularly those which are not traction limited during launch.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto103
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

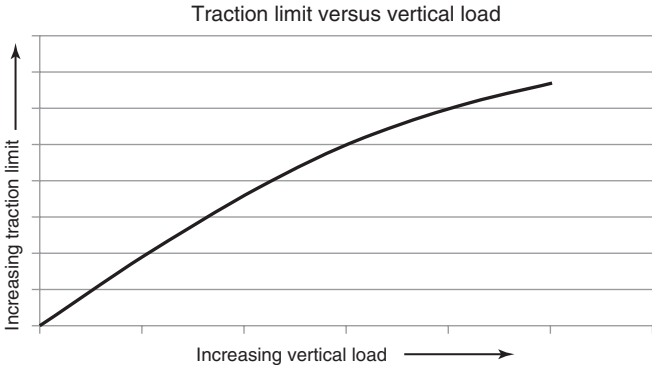


Figure 1. Trend of increasing traction limit with increasing vertical load.

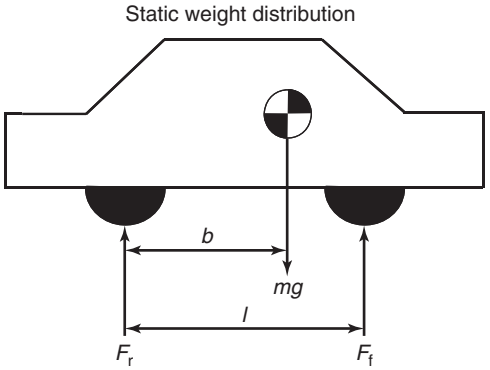


Figure 2. Variables used in calculating the static weight distribution of a vehicle.

However, a number of situations arise where it may be advantageous to set an alternative engine speed setpoint.

If sufficient torque is available to exceed the traction limit at engine speeds below the maximum torque engine speed, then a significant reduction in clutch slip energy can be achieved without impact on vehicle performance by lowering the engine speed setpoint.

If insufficient torque is available to exceed the traction limit, then additional torque may be generated by raising the engine speed to a level above the maximum torque speed before launch, and then transferring the additional kinetic energy stored in this rotating inertia to the vehicle for an additional acceleration effort during the launch event. This technique can be useful both for low powered vehicles and also as a method for compensating for initial lag on turbocharged vehicles where the steady state torque delivery may not be available until a time period after the launch event has been initiated.

2.2.2 *Static weight distribution*

The maximum tractive effort a tire is able to support is a function of the vertical load applied to it. Increasing vertical load results in increasing traction until the tire eventually becomes overloaded. Figure 1 shows a typical trend in increasing traction limit with increasing vertical load.

When a vehicle is at rest, the load on the driven axle is a function of the total mass of the vehicle, the location of the center of gravity, and the length of the wheelbase as shown in Figure 2.

The static axle loads are given by the equations:

$$F_f = \frac{mg \times b}{l} \tag{1}$$

and

$$F_r = mg - F_f \tag{2}$$

where

F_f = Static front axle load (N)

F_r = Static rear axle load (N)

m = vehicle mass (kg)

g = acceleration due to gravity ($N \cdot kg^{-1}$)

b = position of center of gravity ahead of rear axle (m)

l = length of wheelbase (m)

Table 1 shows the typical range of front to rear static weight distribution distributed by vehicle type and powertrain layout. The values are taken from various sources across the internet, and where sources do not agree, an average of the published values has been taken.

This variation in static weight distribution explains the difference in traction experienced by many owners in conditions of low friction and hence low rates of acceleration. Drivers of front-engine rear-wheel-drive vehicles will often be stranded unable to climb a gradient when drivers of similarly equipped front-engine front-wheel-drive vehicles are able to continue their journey.

2.2.3 *Dynamic weight transfer (steady state)*

When a vehicle accelerates, the tractive effort is created at the contact between the tires and the ground. However, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto103
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Typical front/rear weight distribution in modern road cars

Powertrain Layout	Example Vehicle	Front Axle Weight (%)	Rear Axle Weight (%)
Front-engine, front-wheel drive	Ford Focus RS	63	37
	Saab 9-5	60	40
	Corsa GSI	63.4	36.6
	Volkswagen Scirocco	64	36
Front-engine, rear-wheel drive	Mazda MX-5	50	50
	Toyota GT86	53	47
	Mercedes SLS	47	53
	Corvette Stingray	50	50
Front-engine, four-wheel drive	Volvo V70R	54	46
	VW Golf R	60	40
Mid-engine, rear-wheel drive	Lotus Elise	38	62
	Lotus Evora	38.7	61.3
	McLaren MP4-12C	42.5	57.5
	Ferrari 458	42	58
	Porsche Cayman S	45.3	54.7
	Renault Clio V6	40.4	59.6
Mid-engine, four-wheel drive	Audi R8	44	56
	Bugatti Veyron	44	56
	Lamborghini Aventador	43	57
Rear-engine, rear-wheel drive	Porsche 911 GT2	37	63
Rear-engine, four-wheel drive	Porsche 911 Carrera 4	40	60

center of mass of the vehicle will usually reside at some height above this ground plane and hence a moment arm is created, which must be balanced, in steady state, by weight transfer between the two axles. In the case of acceleration in the forwards direction, weight will be transferred from the front to the rear axle.

The weight transfer to the rear of a vehicle during acceleration is given by taking moments about the front axle contact patch as defined in Figure 3:

$$\Delta F = \frac{ma \times h}{l} \tag{3}$$

where

ΔF = Weight transfer toward rear axle (N)

m = vehicle mass (kg)

a = longitudinal acceleration of vehicle ($\mathrm{m \cdot s^{-2}}$)

h = height of center of gravity above ground plane (m)

l = length of wheelbase (m)

This dynamic weight transfer leads to the traction advantages seen by rear-wheel-drive vehicles when compared to front-wheel-drive vehicles at high rates of acceleration and also explains why, as acceleration levels increase, the traction benefit of four-wheel-drive systems is progressively reduced. This is being best illustrated when considering the top classes in the sport of drag racing where the fastest accelerating vehicles in the world can be found. Top fuel dragsters and closely related classes run as rear-wheel-drive vehicles. The rate of acceleration is sufficient that the weight transfer to the rear axle completely unloads the front axle and often the vehicle can be seen to be carrying its front axle clear of the ground for the initial section of the race.

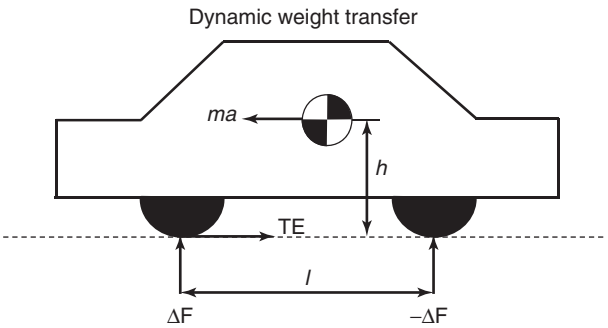


Figure 3. Variables used in calculating the dynamic weight transfer of a vehicle.

2.2.4 *Dynamic weight transfer (transient)*

A further consideration important when designing a launch control system is the influence of the design of the tires

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto103
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and vehicle suspension system on the initial weight transfer. The weight on the driven axle is transferred to the tires via the vehicle suspension, and therefore as the vehicle transitions from its static to dynamic condition the suspension system is required to support the additional vertical load. During the initial transient, any additional vertical load is supported only via linkage forces created by features such as anti-squat geometry (*see* General Introduction—Basic Definitions, Structure of Part 4) and by forces generated in the suspension dampers because of the rate of deflection of the suspension. Further forces are then generated by the suspension springs and tire sidewalls as the suspension deflection approaches its steady state condition and the full weight transfer is established.

2.2.5 Aerodynamic downforce and its influence on traction

Formula 1 racecars are generally accepted as being able to drive upside down on the roof of a tunnel at speeds in the region of 130 km/h. With a mandated minimum rear axle load of 342 kg, this indicates that the total combined downforce effect of the wings and bodywork at this speed must be equivalent to at least 342 kg. Assuming that the lift coefficient of the vehicle is constant (*see* Fundamentals, Basic Principals in Road Vehicle Aerodynamics & Design), this means that at a speed as low as 70 km/h, well within the traction-limited launch window, a Formula 1 car can already be creating an additional 100 kg or 30% of static axle load. Referring back to Figure 1, we can appreciate that this has a significant impact on the available traction and must be considered in determining the traction limit.

By contrast, most road vehicles generate low levels of lift (negative downforce) and therefore effectively reduce the load on the drive axle as the vehicle accelerates. Even the most extreme road cars only produce moderate levels of downforce. One of the highest downforce producing road cars currently in production is the Ferrari Enzo, and this is only claimed to generate 760 kg of downforce at 300 km/h. Comparing this back to the Formula 1 car example quoted earlier, and making the same assumption regarding constant lift coefficient, this equates to just 40 kg of downforce at 70 km/h and continues to be negligible until a speed beyond that at which the vehicle ceases to be traction limited.

2.2.6 Integration with traction control systems

Many modern vehicles are equipped with some form of traction control system designed to reduce wheelslip by modulating engine torque, brake apply torque, or a combination of both (*see* The potential for handling improvements by global chassis control). Integration of clutch and traction control systems is a key aspect of achieving an optimum launch in a modern vehicle and it is the traction control system that becomes fully responsible for achieving the best possible rate of acceleration once the launch clutch is fully engaged.

2.3 Pullaway control considerations

In this chapter, pullaway control is used to differentiate from launch control in that pullaway control is used to describe the process of accelerating a vehicle from rest at some level below the maximum possible. In this condition, the priorities move from outright performance to achieve a balanced compromise among emissions, acceleration feel, durability, and performance.

2.3.1 Acceleration feel

Achieving a successful subjective feeling during vehicle launch can be a compromise between initial response and ultimate acceleration. Unlike the performance launch, the control system will usually not be forewarned of an impending acceleration request (via a launch mode button, double foot start, etc.) and therefore the usual conditions for the start of a pullaway event will be vehicle stationary, foot brake applied, accelerator pedal released, engine speed at idle, engine load at creep torque, or lower. The torque available at this engine speed will usually be able to support only a modest rate of acceleration and therefore the engine speed must be allowed to rise in order that sufficient torque becomes available to meet the driver's demand for acceleration. However, any torque used to accelerate the engine is not available to accelerate the vehicle. Therefore, a compromise must be found between the immediacy of response and the ultimate rate of acceleration achieved.

Achieving the correct compromise is a key skill of the vehicle development engineer. However, a key attribute used by many vehicle manufacturers to assess the response of a vehicle to a driver demand is that of the "4 s distance." This is defined as the distance traveled by a vehicle within the first four seconds elapsed time after the driver demand for acceleration is received. Maximizing this distance will involve finding the optimum balance between engine torque used initially to accelerate the engine to a speed at which greater torque is ultimately available and engine torque used to immediately accelerate the vehicle.

2.3.2 Comfort

Where the situation or vehicle type demands a comfort-oriented launch, the goal of the pullaway control system is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto103
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to transition from the start condition to the driver-demanded rate of acceleration with the least disturbance to either the driver or the passengers while completing this transition in an acceptable and safe period of time.

As with shift control (*see* General Introduction—Basic Definitions, Structure of Part 4), many attempts have been made to objectively quantify what is a comfortable pullaway. However, the general acceptance is that maximizing comfort is achieved by minimizing jerk through the pullaway event.

2.3.3 Sportiness

Achieving a sporty pullaway feel is in many ways about creating the opposite attributes to that of a comfort-focused pullaway. An immediate response, noticeable with a sudden change in acceleration, will usually be perceived as sporty. Similarly, if this wheel torque is applied in a sudden and aggressive manner, the tires will be more prone to slipping. While wheel slip is generally to be avoided for maximum performance, a tire struggling for grip will generally be perceived by most drivers as an attribute of a sporty launch.

2.4 Clutch control methodologies

2.4.1 Introduction

Automated launch clutches have been in existence for many years. Many early clutches had no electronic control and regulated the clutch apply force via mechanisms such as centrifugal force and manifold vacuum. These early systems remain relevant today as their operation forms the foundation for the control algorithms incorporated into many modern electronic control systems.

2.4.2 Modern electronic control methods

Modern launch clutches are controlled via increasingly complex control algorithms running inside an electronic control unit. Actuation is via either electrohydraulic or electromechanical means. Electrohydraulic systems are generally found on higher performance systems and control the capacity of the launch clutch using hydraulic fluid as the power transfer means. Electromechanical systems control the capacity of the launch clutch via electrically driven actuators acting directly onto the launch clutch actuation system.

Development of launch control systems has typically been an evolutionary process with the early passive systems giving way to electronic systems, which mimicked the operation of the passive systems. These systems have then evolved to become more and more complex bringing the benefits of both increased functionality and increased robustness of performance in service.

Launch control systems control two actuators, the engine torque and the clutch torque capacity. Two basic control targets are defined for the system, achieving a desired engine speed profile and achieving a desired vehicle acceleration profile.

For a viable system, each one of the actuators must be assigned to each one of the control outputs. In reality, both systems must closely interact to achieve the necessary system-level performance. Selection of the appropriate allocation of targets occurs not only because of engineering principles but also because of commercial and political constraints within engineering companies, and both techniques have found widespread application within the industry. The strengths and weaknesses of each approach are discussed in the following sections.

2.4.2.1 Engine speed control via clutch regulation. The most common methodology for launching a vehicle in use today involves the launch clutch being allocated the task of controlling the engine speed to a target profile. The engine torque output is a function of driver demand and engine speed as per a manual vehicle. When the driver demand is increased, the engine speed rises in response to the increase in engine torque. The launch clutch then responds to this rise in engine speed by increasing its torque capacity and hence the reaction torque on the engine. In turn, this reaction torque is transferred to the wheels via the transmission and driveline and results in a change in vehicle acceleration.

When a traction event is detected by the traction control system, a torque decrease is requested from the engine, the speed falls, and the clutch control responds by reducing its torque capacity thereby reducing the torque at the wheels.

This method allows the launch control function to be developed largely independently from the other vehicle systems and generally requires the lowest level of integration and change from the other vehicle systems to support its introduction. However, the discussion of the multiple stages between detecting a traction event and the system reaction being achieved highlights how significant improvement can be made by interfacing the traction control directly to the clutch capacity and by directly controlling engine speed via engine torque regulation. However, it is this method of control which dominates in the market place. This is due to the majority of vehicle architectures being established for either manual or torque converter AT-equipped vehicles, and hence introducing a launch clutch using this methodology requires the minimum of revision to the base vehicle architecture.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto103
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.4.2.2 Engine speed control via engine torque control. With this methodology, the control of engine speed is achieved by manipulating the engine torque output and the launch clutch is directly responsible for controlling the vehicle acceleration via wheel torque control. Intervention from external systems such as traction control can be achieved directly through commands to the launch clutch and the system performance is robust to variations in the engine torque response. This approach allocates each of the actuators the role in which it has direct control and therefore the role that allows the powertrain to deliver the highest possible level of performance.

The challenge of this approach is that it carries little commonality with the operation of a conventional vehicle and therefore requires a dedicated vehicle architecture. For this reason, this approach is usually only found on vehicles where the powertrain is solely or primarily equipped with an automated launch clutch.

2.4.3 Special functions

A key characteristic of a launch clutch is its ability to regulate torque independent of slip speed up to the point where slip is reduced to zero. When combined with an electronic control system, this facilitates both the maintenance of key driving features carried forward from torque converter-equipped transmissions and often the enhancement of these features to a higher level of robustness or performance.

2.4.3.1 Hill hold. First introduced in Section 2.1.2, hill hold is used to describe the ability of a vehicle to hold stationary on a gradient. Combining a launch clutch control system with a directional speed sensor allows the hill hold capability of a vehicle to be enhanced to prevent rollback on gradients up to and including that possible with the engine torque available at the current engine speed.

2.4.3.2 Creep control. Introducing an electronically controlled launch clutch facilitates the introduction of closed loop creep control functionality utilizing output shaft speed sensing, wheel speed sensing, or a combination of both. In this condition, the control system is able to offer consistent vehicle acceleration and speed control across a range of road gradients and independent of engine speed setpoint.

GLOSSARY

Launch device
Torque converter
Friction clutch
Speed
Acceleration
Jerk

REFERENCE

Forster, H.-J. and Eltze, U. (1978) Brake Operated Low Ratio Release. US Patent No. 4105101.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto103
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Control Systems and Strategies for Automated Manual and Double Clutch Transmissions

Bruno Müller, Götz Rathke, Michael Reuschel, Uwe Wagner, Stefan Winkelmann, Marco Grethel, and Jürgen Gerhart

LuK GmbH & Co. KG, Bühl, Germany

1 Introduction 1
2 System Architecture 1
3 Hardware 6
4 Software 12
5 Technology Drivers for Future Developments 16
References 17

1 INTRODUCTION

Over the past few years, automated transmissions that build on manual transmission technology and generally use similar technologies have changed the world of transmission technology considerably. The most well-known representatives of this include automated manual transmission (AMT) and double clutch transmission (DCT). Figure 1 shows the transmission concepts with basic design examples. They offer the driver more than a considerable increase in comfort; thanks to the automatic shift point selection that can be designed optimally, consumption can also be reduced.

The first series application of a DCT was in the six-speed DCT from Volkswagen in 2003 (known internally as *DQ250*), which was based on an oil-cooled, that is, "wet," double clutch. The potential of this transmission concept for fuel consumption and response time enabled Volkswagen to gain new market shares. As a result, further applications for the technology developed quickly.

Yet despite its benefits, the AMT has remained a niche product since its launch. It is a different story for the DCT, for which high sales opportunities are predicted, particularly on the European and the Chinese market. DCTs receive additional thrust from the hybridization of the powertrain. The integration of electric motors in and on existing transmission systems is an important focus for development worldwide. The cost-intensive hybrid drives promise considerable consumption potential, particularly when used in urban regions with frequent stop-and-go traffic. Owing to their effective competitive marketing, they are regarded by the general public as a serious alternative to classic powertrains. The additional possibilities of the electric motor integrated in the transmission can be implemented particularly cleverly in connection with a DCT. Details of the system architecture, hardware, software, and future developments of AMT and DCT are discussed later.

2 SYSTEM ARCHITECTURE

2.1 System components

From the driver's perspective, the automated transmission is responsible for setting the required speed safely and comfortably at all times. To do this, the transmission system must guarantee a suitable wheel torque progression through the interaction of multiple components (for more details on construction and operation see Automated

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

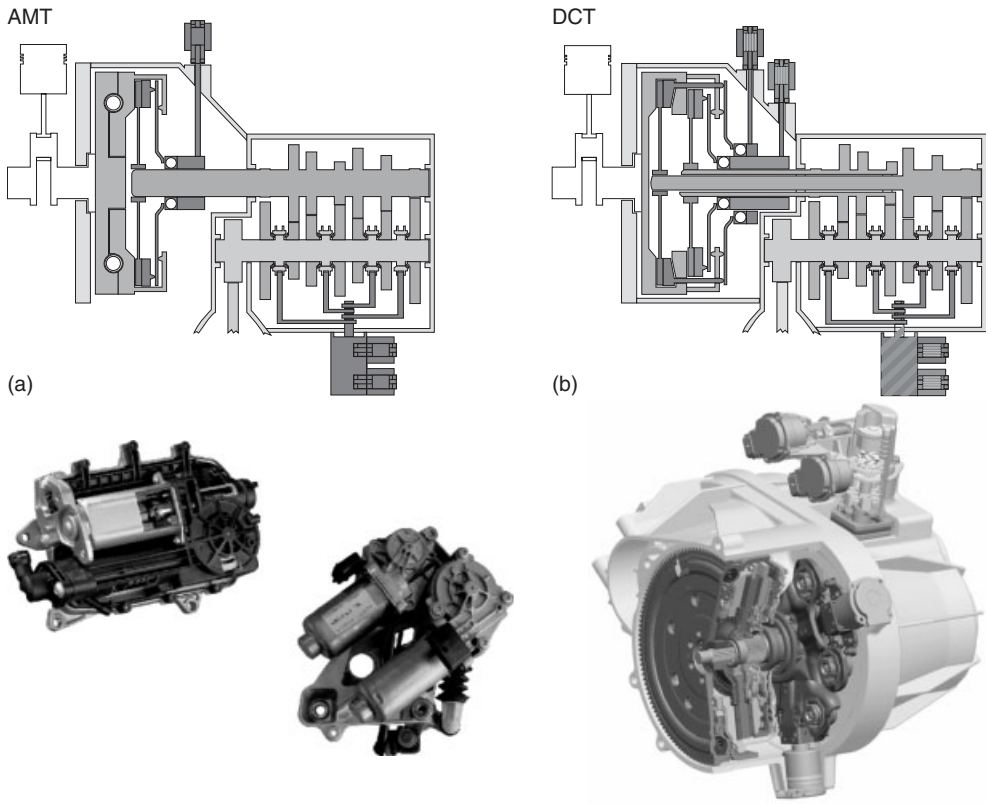


Figure 1. Transmission concepts and design examples for (a) AMT and (b) DCT. (Reproduced with permission from Schaeffler.)

Manual Transmissions (AMT)—system design considerations, clutch operation, shift actuation alternatives and Dual Clutch Transmissions (DCT)—Layouts, Clutch Selection, Packaging, Actuation, Manufacturing & Assembly). Figure 2 shows this basic interaction.

2.1.1 Controller and actuator

To fulfill this requirement, the control system actuates the clutches to set a driving torque at the transmission input and, using the shift mechanism, selects the appropriate transmission ratio in the transmission. Depending on the design, the automated transmission also controls the clutch cooling for wet clutch or clutches or the actuation of a parking lock.

Irrespective of the specific design of the system, clutch and transmission actuation follow basic functional structures. This means that first of all, the energy required for clutch operation must be provided. Electric motors supplied with power by the vehicle battery or hydraulic pumps driven directly by the engine are used for this purpose.

Actuators convert the energy supplied into actuation travel or an actuation force. Thus, electromechanical actuators convert the rotational movement of the electric motor into a longitudinal movement. The ratio can be purely mechanical with the aid of a lever or hydrostatic with the aid of master and slave cylinders with hydrostatic lines. If the energy is provided via hydraulic pumps, the actuators are valves that control the pressure provided by the pump and distribute it to the downstream elements of the actuation system.

Finally, the actuation system establishes the connection between the actuators and the clutch and creates the clamp force in the clutch. In electromechanical systems, actuation is via engagement bearings; in hydraulic systems, generally via rotating slave cylinders.

Via the clamp force, the clutch defines the torque transferred from the engine to the transmission. In contrast to dry disk clutches, wet-running multidisk clutches require an additional element for clutch cooling.

The control system and software of the automated transmission require information from the functional structures to enable them to control the transmission system safely and comfortably (Figure 3). Sensors are generally used to observe the behavior and state of the components. Where the use of sensors is not possible

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

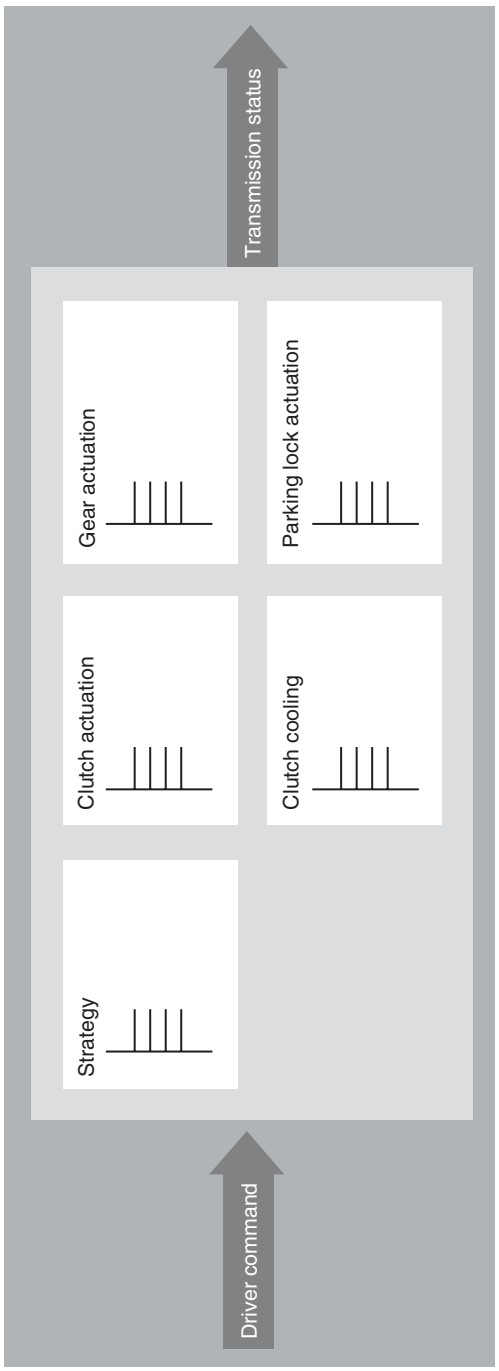


Figure 2. Basic structure of the transmission control. (Reproduced with permission from Schaeffler.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

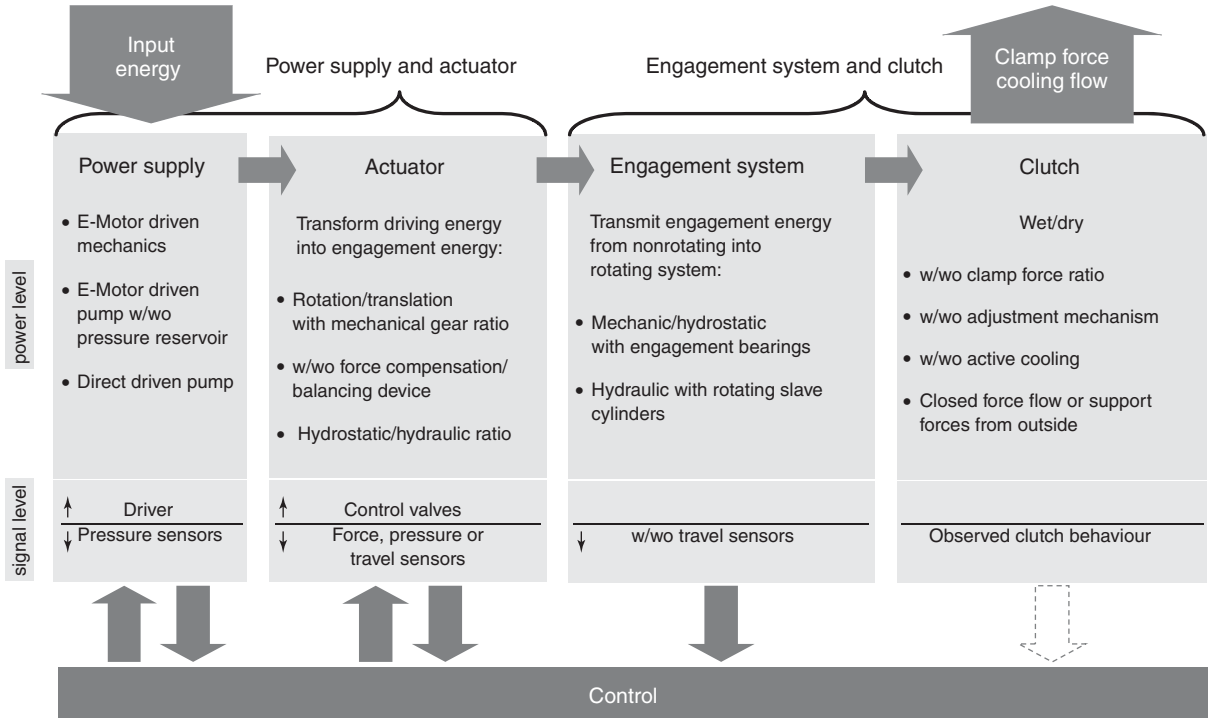


Figure 3. Functional structure of the clutch control. (Reproduced with permission from Schaeffler.)

for technical or cost reasons, software strategies must continuously assess the system behavior in real time based on models.

For the control of the transmission actuation, the same structures apply concerning power supply and actuator mechanisms as for the clutch actuation. However, the actuators have a direct effect on the transmission gearshift rails, set the gears, and actuate the parking lock—where this function is provided for in the system.

2.1.2 Designs

The requirements for function and quality of the system and the boundary conditions of requirement management (Pohl, 2008) determine the specific design of an automated transmission (Figure 4).

Basic decision criteria for the system design are low CO_2 emissions with simultaneous high comfort, reliable function at all times and a long life. These are supplemented by high demands in terms of mounting space requirements, system weight, and component complexity. More recent criteria also include the possibility for hybridization and/or stop–start capability of the system (Steiger *et al*., 2013)

The various criteria are weighted in the context of the overall vehicle design in order to determine the importance of the advantages and disadvantages of each design variant. This applies both to the clutch concept (wet multidisk clutch and dry friction disk clutch) and to the selection of the actuating mechanism and actuation.

For example, a wet-running multidisk clutch is used primarily for vehicles with a high mass and engine power, as the wet-running clutch can dissipate the frictional power that arises more easily. This has a positive effect on service life and reliability. At the same time, however, the clutch cooling required can increase energy consumption.

In contrast, a dry friction disk clutch is distinguished by high energy efficiency—particularly if it is equipped with an electromechanical actuation. However, the design variant may reach mounting space and life limits more quickly (Wagner *et al*., 2009).

2.2 Design and concept landscape

In addition to the selection of the optimum mechanical components for the particular use case, the integration of the automated transmission into the electronic vehicle architecture must also be considered. As, to a large extent, the mechatronic system takes over the responsibility for safe function and proper use from the driver of the vehicle, the requirements for functional safety must be made an integral part of the system design. For this purpose, ISO 26262

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

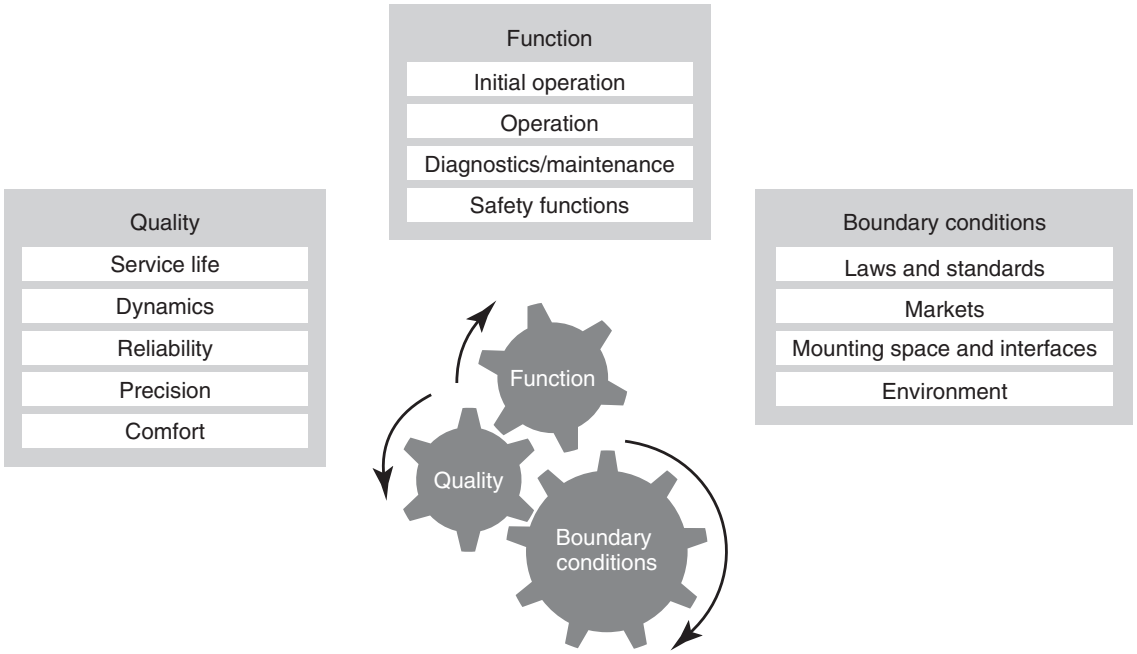


Figure 4. Decision criteria for the system design. (Reproduced with permission from Schaeffler.)

defines a procedure module together with required activities and work products as well as methods to be applied in development and production.

2.2.1 Central versus local control system

An important question in the system design is whether and to what extent the intelligence of the manual transmission should be distributed. Highly integrated control units with sensors and actuators that connect all tasks and all requirements (temperature, fatigue limit, and functional safety) in one device have the advantage of reduced system and assembly costs. They also provide the option of a complete system test before installation in the transmission unit. The reduction in the number of cables and contact and connector transitions also helps to provide increased system reliability (Stark and Schuch, 2010).

The alternatives are architectures in which a main control unit communicates with intelligent sensors and actuators via a bus system. This enables high resolution signals to be processed directly as well. Requirements from functional safety can also be distributed to the intelligent actuators.

2.2.2 Functional integration in the powertrain management

Automated transmissions exchange data and commands intensively with the engine control unit as well as the chassis and safety systems [electronic stability program (ESP), antilock braking system (ABS)]. To achieve high driving comfort at all times, the interfaces between the systems must be precisely aligned. Examples are the required reduction in engine torque during gearshifts for an AMT, or the required precision of the engine torque signal: in the AMT, the torque intervention ensures that the engine speed does not increase in the period in which the clutch is open during a gearshift. The torque calculated by the engine control unit is used by the transmission control system to compare the models for calculating the torque transferred from the clutches. These models run internally. If an automated transmission is integrated in a hybrid powertrain, this results in additional mechanical, electric, and electronic interfaces that have to be optimized as part of the boundary conditions and criteria described.

2.3 Integration into the vehicle

Irrespective of the specific design, an automated transmission is only one part of the overall vehicle system. The high complexity and the strong interaction between the subsystems require intensive alignment and testing of the subsystems as well as the overall system. Ultimately, comfort, performance, and NVH aspects can only be assessed in the vehicle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.3.1 Vehicle alignment

In the vehicle, the effects of the automated transmission are directly evident to the driver: the clutch torques transferred determine the vehicle acceleration. The shifting characteristic curves define when each gear is activated and therefore have a considerable influence on the characteristics of the vehicle. Thus, the alignment of the automated transmission, that is, the development and parametrization of control unit software functions are very important.

The alignment defines the comfort or sportiness of the vehicle ride, as well as having a direct influence on the load for the individual transmission components and on CO_2 emissions. Thus, starts with very high engine revolutions per minute are deemed to be sporty. However, they increase the load on the start-up clutch considerably. The continuous alignment of actual load and load assumed during the design of the system is an important part of the development process.

2.3.2 System and subsystem tests

Validation, that is, proving the durability of the transmission system and its subsystems, is not based solely on real vehicle tests. In practice, numerous calculations and virtual simulations are used. For example, load spectrums with which the endurance requirements can mapped nested on a time basis are derived based on individual vehicle measurements or design data. Environmental influences such as temperature or humidity are considered in accordance with ISO 16750.

3 HARDWARE

To implement AMT or DCT controls, numerous hardware components have to work together. These include several design variants of electric motors, control units, sensors, and, depending on the design, an electrohydraulic, electromechanical, or hydrostatic actuating mechanism. Depending on the combination and design of these components, various concepts can be considered and types of actuation systems developed that satisfy not only the specific requirements of the OEMs but also the individual applications. Regardless of how different the transmission actuation systems are individually, all systems contain electric motors, control units, and sensors.

3.1 Electric motors

In principle, direct current servo motors are used in the actuators for clutch and transmission control. They contain coils and permanent magnets and have to be commutated so that the motor coils can be controlled correctly in accordance with the angle of rotation.

In conventional direct current motors (DC motors), the stator contains the magnets and the rotor the copper coils. The current for controlling the motor is supplied to the coils from outside via the mechanical commutator (carbon brushes). The commutator constantly reverses the polarity of the motor coils depending on the angle of rotation. This causes the motor to rotate in a specific direction. Therefore, there is no need for a commutation sensor in DC motors.

For brushless direct current motors (BLDC motors), the arrangement is exactly the opposite: the rotor contains the magnets, the stator the copper coils. This means that the coils can be contacted directly. The commutation is electronic rather than mechanical and via the control unit. An angle sensor that records the angle of rotation and transmits it to the control unit is generally required. This is evaluated in the commutation module and the motor coils are controlled according to the angle of rotation in the power stage. Owing to the advantages of the BLDC motor (including increased efficiency, increased energy density, lower inertia, lower EMC emission, and the possibility of optimizing the speed/torque characteristic), this type of direct current motor is often used in clutch/transmission actuators. To increase the power density further, rare-earth magnets are generally used nowadays instead of ferrite magnets. The costs of rare-earth magnets have risen exorbitantly since 2010, meaning that the benefits of the high power density must always be weighed against the costs.

The rotary sensor is also no longer necessary. In this situation of control without sensors, the angle of rotation is determined via the voltage and current signals. The motor start-up and fast changes in the motor speed are particularly critical as they can only be captured very slowly or very imprecisely via the current/voltage measurement. Therefore, at present, sensorless motors are only used in a few dynamic drives, such as pump drives.

During the relative measurement of the rotor position, angle increments are sent to the control unit in the form of pulses so that the control unit can recognize whether and how the motor is rotating. The motor speed is also derived from this angle information. Usually, multiple digital Hall sensors, which detect the magnetic field of a sensor magnet, are used. These sensors enable good dynamic and precise control of the rotor position to be realized as part of the sensor resolution.

With electromechanical actuators, it is important to know precisely where an actuator is. If an incremental sensor is used, an actuator must move to a reference point to determine the absolute position. Starting from this point,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

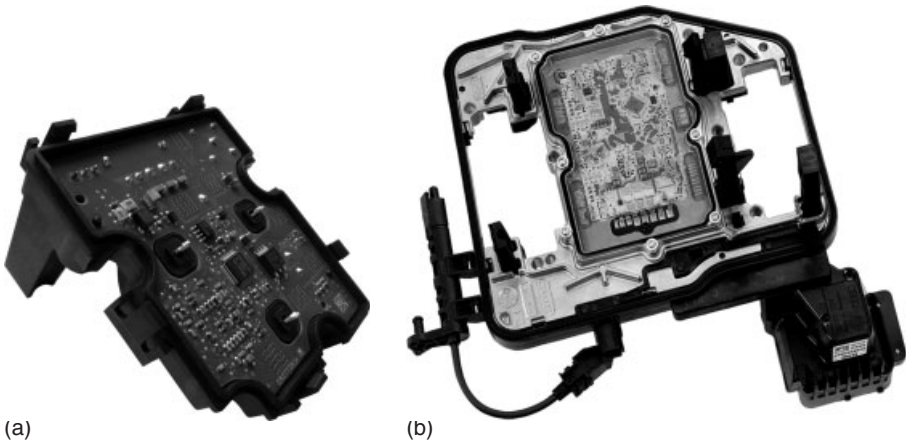


Figure 5. Examples of integrated control units: (a) local actuator control unit (LCU) in printed circuit board technology. (Reproduced with permission from Delphi. © Delphi.) (b) Integrated transmission control unit in hybrid technology. (Reproduced with permission from Continental. © Continental AG.)

the control system counts the sensor pulses and, if the sensor resolution and mechanical ratio are known, can continuously calculate the precise position of the actuator. After a control unit reset, however, the stop position has to be retaught.

If the reference position and referencing operations are not desired, absolute sensors are used. The sensor immediately shows the current angle position or actuator position of the motor even after a control unit reset.

3.2 Control units

Control units are usually customer-specific solutions. Stand-alone units (control units mounted on the chassis), local control units (control units mounted on the transmission), or integrated control units can be used (Stark and Schuch, 2010). Figure 5 shows an example of an integrated transmission control unit. Stand-alone units generally do not contain any sensors. They are not installed in or on the transmission, but as a separate electronics box in a suitable installation space in the vehicle. Attachment units have fixed connections to the transmission and are subject to increased vibration and temperature loads. Sensors and actuators are connected via an electrical harness. The mechatronics solutions of integrated transmission control systems represent the highest level of integration (Stark and Schuch, 2010). Here, the mechanics, hydraulics, electronics, sensors, and actuators are integrated in one unit.

The control units must satisfy high requirements in terms of vibration and temperature resistance. In printed circuit board technology, control units can be used on a conventional fiber glass epoxy resin up to an ambient temperature of $\sim$125°C. Temperatures of more than 150°C and vibration loads of up to 50 g are much too high for most housed semiconductor components (Stark and Schuch, 2010). Thus, design solutions with ceramic interconnect devices and unhoused semiconductor components are used. Thick film substrates, LTCC (low temperature co-fired ceramic) multilayer circuits, and DBC substrates (direct bonded copper) are available. Figure 6 shows the different substrate technologies.

As an alternative to central control units, it is also possible to distribute the intelligence to multiple local control units. This provides a technology that opens up completely new paths for system architecture. This applies above all in combination with an electromechanical clutch and gearshift actuation (Müller *et al.*, 2010).

Where possible, the sensors of transmission control systems are integrated directly in the control system electronics. This eliminates the need for contact interfaces that require sealing and costly cabling work. It also avoids the risk of electromagnetic interference engaging on the line between the sensor and the control unit. However, the sensor integration is usually associated with high mounting space requirements in the control unit as the control unit has to be directly connected to the point to be evaluated. Figure 7 shows an example of the structure of a control unit.

Multiple control units are connected with a communication bus via a communication interface. The most common form is the CAN bus. Future developments of control units with FlexRay are conceivable. The LIN bus is too slow for the applications in the area of actuator control of a double clutch (Mayer, 2006).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

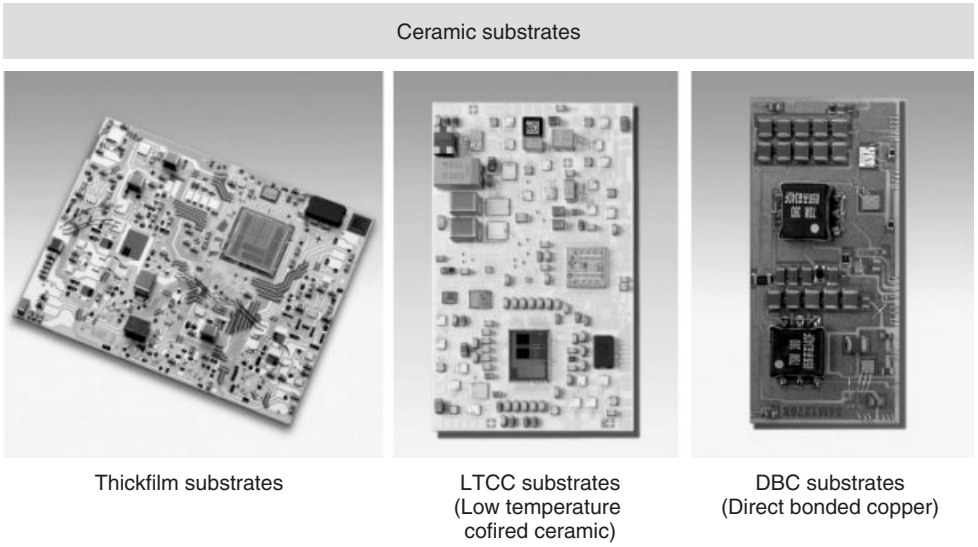


Figure 6. Ceramic technologies for interconnect devices. (Reproduced with permission from Stark and Schuch, 2010. © Continental AG.)

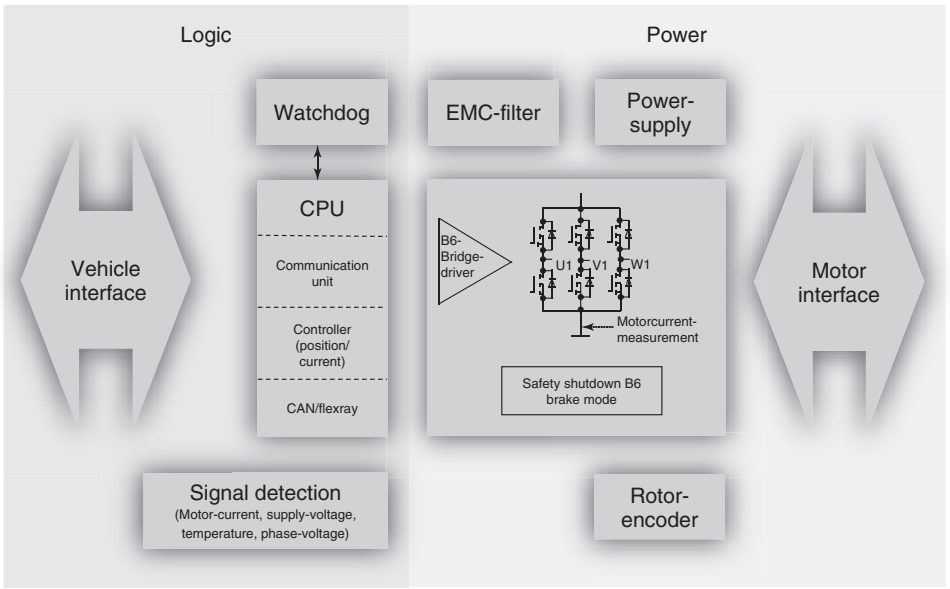


Figure 7. Example block diagram of a control unit. (Reproduced with permission from Schaeffler.)

3.3 Sensors

To increase robustness against wear and mechanical tolerances, noncontact sensors are generally used. Position, angle, torque, and speed sensors are used, as well as pressure sensors. The information from pressure sensors from the hydraulic line between the actuator and the clutch can be used to calculate the clutch torque transferred. Position sensors convert the position of the release bearing (measured in the actuator or on the slave cylinder) into an electrical signal using the magnetic or inductive principle. As this position is not proportional to the clutch torque over life, temperature, and other influencing factors, a software adaption must take place to calculate the torque information from the position information.

Speed sensors measure the transmission input speed, the transmission output speed, and the wheel speeds.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.4 Designs

There are various concepts with different actuators for the shift operation and for actuating the clutches for the DCT and AMT design: electrohydraulic actuating mechanisms and electromechanical actuating mechanisms. The initial costs for an electrohydraulic actuating mechanism are higher than those for an electromechanical one. However, as the number of actuation points and power requirements increase, the hydraulic solution becomes less expensive as compared to an electromechanical one. In general, the electromechanical actuating mechanism has a lower power input, which ultimately has a positive effect on a vehicle's fuel consumption.

3.4.1 Electrohydraulic actuating mechanism

Hydraulic systems are known for being able to achieve movements quickly and precisely at high loads (Drexler *et al*., 1988). The high power density of the function components also make them interesting for vehicle applications. Thus, hydraulic actuator systems have proved themselves in automatic transmissions for many years.

However, internal leakages, the low temperature behavior, and the wide speed range of pumps driven by combustion engines require technical tricks to ensure problem-free and energy-efficient function in modern vehicles. The simple and cost-effective expansion of a functionality of an existing hydraulic system is in contrast to the high initial expense.

Common hydraulic systems generally contain the following components: delivery pump, valves for distributing the volume flow, that is, for cooling the clutch and valves for varying the pressure level, and thus the force in the actuator. The valve functions are generally grouped in assemblies known as *valve chests* or *hydraulic blocks*. In these assemblies, the pressure oil flows from valve to valve in channels or bores. The valve chests are generally placed inside the transmission housing and are thus protected against many environmental influences. If they are mounted on the outside of the transmission, corrosion protection and sealing to the outside present special requirements.

The method of provision of the hydraulic energy plays a key role particularly with regard to the efficiency of the hydraulic systems. Classical pumps driven by combustion engines are characterized by the shortest conversion routes from the rotation energy of the drive motor emitted to the hydraulic energy for the actuator system. However, the fact that the pump displacer size is defined by the minimum required flow rate at relatively low engine speeds and thus supplies too much oil at higher engine speeds is disadvantageous from an energy perspective.

A better demand orientation for energy provision can be realized using a pump drive form that is independent of the combustion engine speed (e.g., via an electric motor). However, some of this energy advantage is consumed by the longer efficiency chain. For further energy provision concepts, the demand situations must be considered in more detail, as the requirements for the volume flow source can be considerably different even with identical or comparable hydraulic power. This is shown by common demand situations such as low volume flow at high pressure and high volume flow at low pressure. There must also be an analysis of which demands occur permanently, frequently, or rarely.

The necessity of components for energy storage (accumulator) and conversion of pump characteristics (ejector or hydrotransformer) can be derived from this knowledge. To achieve the best possible demand orientation at optimized efficiency, modern supply sources use mixed concepts for hydraulics rather than just one pump. Two examples on the market also show this: Volkswagen uses powerpack hydraulics for its seven-speed DCT DQ200. This involves a small, electrically driven high pressure pump with a pressure accumulator to provide the system with high volume flows at short notice (Hadler *et al*., 2008). Mercedes-Benz has equipped its seven-speed DCT 7G DCT with a minimized, permanently driven main pump and an additional electrical pump (Wörner *et al*., 2011).

The logical processes for controlling the actuators are represented via a number of valve types and valve functions. In addition to the requirements for direct valve function (valve characteristic, dynamics, and hysteresis), particular attention must be paid to reducing the valve losses (including leakage) during the design of the valves.

Therefore, there is a clear trend away from pilot operated toward directly controlled valves. A further fundamental differentiation in the valves is in the selection of the material for the pistons and the direct valve environment (valve bore). Direct valves work with coated aluminum pistons that are operated directly in a valve chest made of aluminum casting. In contrast, cartridge valves have steel pistons that are operated in a steel sleeve. In turn, these steel sleeves are installed in a valve chest or a hydraulic block.

Interfaces are required both for the clutch actuation and the shift actuation. These can be realized hydraulically or mechanically. In the latter case, linear, swivel, or rotation motors are integrated in the hydraulics. Depending on the interface, pressure or path sensors are used to provide the transmission control unit with feedback about the current

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

function execution. Thus, the output can be controlled directly or adaptively or regulated actively via pressure or path sensors.

3.4.2 *Electromechanical actuating mechanism*

The electromechanical actuating mechanism is deemed to be a particularly efficient actuation system. To actuate a DCT, an average power input from the vehicle electrical system lower than 25 W can be realized in a representative cycle. This can reduce the fuel consumption by 1–2% compared with conventional actuation systems. When designing and selecting the elements that transfer power, attention must be given to ensure that the holding currents and effort for ratio adjustment are minimized. If new electric motors, sensors, and control units or even new mechanical elements are to be used, the development costs for an application-specific actuating mechanism are relatively high. In the meantime, however, there is a large set of existing components for electromechanical solutions as well.

In addition to the electric motors, sensors, and control unit, an electromechanical system needs actuator specific software, for example, referencing, plausibility, and safety strategies, such as fault detection and replacement strategies. To convert the rotary motion of the motor into an axial engagement motion, simple spindle nuts or more complex elements such as a ball screw drive or a planetary roller spindle drive can be used. Further mechanical elements are transmission ratio elements such as spur gears, crown gears, or bevel gears, worm gear transmission, or lever ratios. Depending on the mounting space and characteristics, compensation springs that level a nonlinear load progression can also be helpful to support the electric motor.

Depending on the control task, the actuation systems are equipped with additional sensors. The closer the sensor is placed to the effective working point, the easier the control task generally is, but the more difficult the design integration and connection of the sensors. Therefore, the approach is generally to use the simplest and fewest sensors—in the best case, only the sensors in the electric motor—which in turn requires more software intelligence.

In the past, electromechanical transmission control systems were controlled by control units mounted on the chassis. However, to improve the separation of the transmission system from the vehicle, there is a trend toward attached control units or integration of the control system in the transmission, as is already standard for automatic transmissions. Further elements, such as an engagement lever and release/engagement bearings or gearshift forks, are required as an interface with the transmission in order to guide the actuation movement and force to the required effective working point.

Various designs in current vehicle models provide convincing evidence of the diversity of an electromechanical actuating mechanism. In the development of the AMT for the VW up compact car therefore, particular attention was given to design the required actuation systems for gear selection and clutch control such that mounting space and weight were optimized (Schäfer, 2011). The gearshift for engaging gears consists of two independent electric motors. One selects the gearshift rail and the second one engages or disengages the gears by a conventional gearshift shaft with switching finger. The actuator for opening and closing the clutch is operated by a further electric motor. As a mechanical compensation spring that compensates the counterforce of the clutch spring is used, the electric motor only consumes the frictional forces during opening and closing operations.

Two DC motors are also used for the Opel Easytronic gearshift module. The selection axis is designed for short reaction times and the shift axis for high shift loads. The gear actuator finger is driven by a worm gear and a shift elasticity. The shift elasticity consists of springs that are pulled up under torque. It is required to reduce the torque and force peaks that occur on the synchronization or another stop due to the high inertia of the motors and to improve the controllability of the synchronous force. The motors are premounted to a gear actuator module in an aluminum housing (Figure 8).

For a DCT with an electromechanical clutch actuating mechanism, it is obvious that the DCT gearshift actuation has an electromechanical design (Kimmig *et al*., 2010). An example of this type of gearshift actuation is the LuK Active Interlock actuator (Figure 8). The special design of the

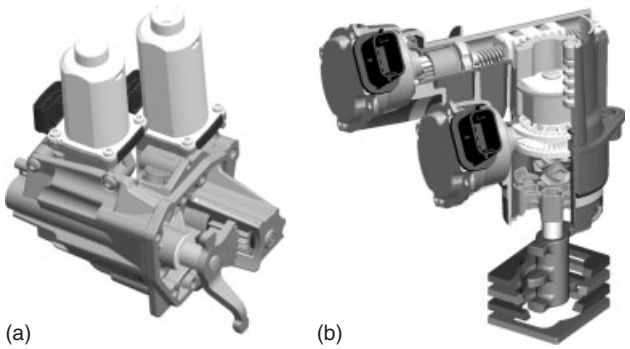


Figure 8. Gearshift modules. (a) Conventional gearshift module for AMTs; (b) active Interlock actuating mechanism for double clutch transmissions. (Reproduced with permission from Schaeffler.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

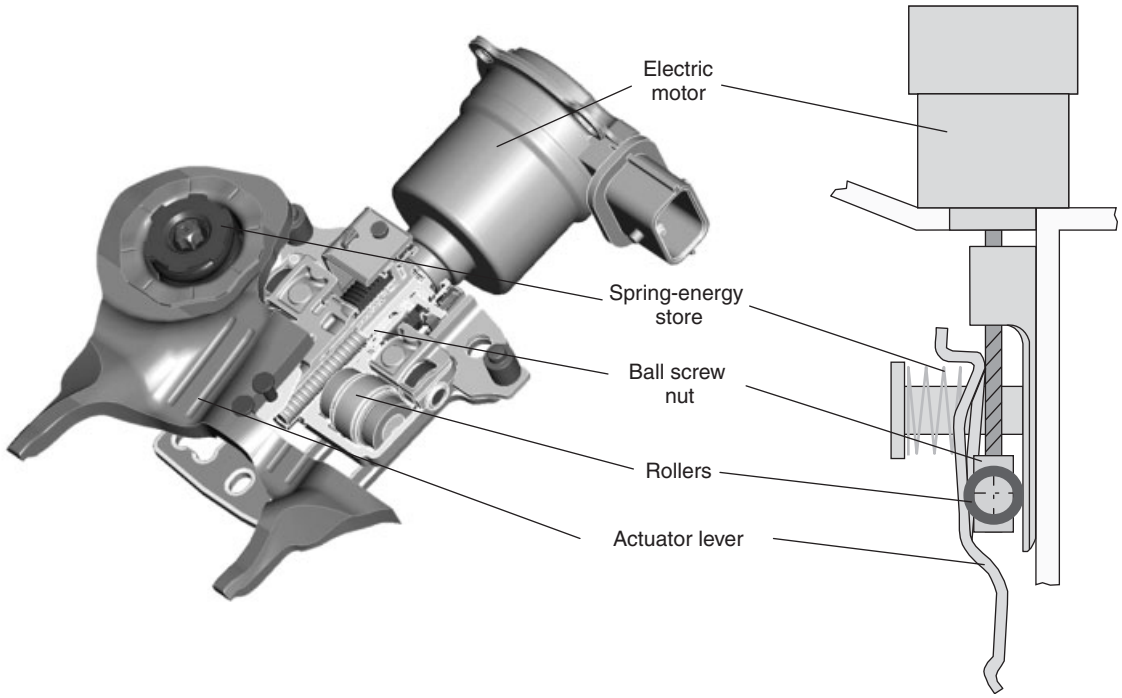


Figure 9. Electromechanical lever actuator for clutch actuation. (Reproduced with permission from Schaeffler.)

shift and eject fingers means that it is possible to preselect and engage the gears in both partial transmissions in any combination. The finger geometry together with the inner shaft rail geometry forms the so-called active interlock. It prevents mechanically impermissible gearshifts meaning that the expensive sensors on the gearshift rails are no longer necessary.

A further design principle uses an electromechanical lever actuator for clutch control (Figure 9). This creates the force required for closing the clutch via a stored-energy spring. The force takes effect at the outer end on the engagement lever designed as a rocker (Kimmig *et al*., 2010). An electric motor screwed onto the transmission housing creates an adjusting movement, via a ball screw drive, in tension rollers arranged between the lever and the transmission wall. These tension rollers form the central lever bearing point. Thus, the bearing point and therefore the effective lever ratio can be changed via the electric motor. For the electric motor, this results in a lower, almost constant force level and thus enables a considerable reduction in the motor size for the clutch actuator.

The use of an actuating mechanism with controller drums also promises low energy consumption (Faust, Bündner, and DeVincent, 2010). Two electric motors that actuate one controller drum each via two-stage ratios are used to shift the gears. Both controller drums are identical and each has only one groove for moving the gearshift forks. The use of this principle means that no additional mechanical lock is required for preventing multiple gears being engaged simultaneously in the same separate subtransmission in the event of a malfunction. The controller drums are controlled by motors integrated in a mechatronics unit via a cascade of gears. In addition to the transmission control system with power stages and motors for gearshift, the mechatronics unit also contains the power stages for the motors on the lever actuator.

A hydrostatic line may also be used to actuate clutches. An electromechanical master cylinder moves a closed-oil column that in turn actuates a slave cylinder, which then actuates a clutch, for example, via a release/engagement bearing. The advantages of a hydrostatic connection are the high flexibility of the wiring, low elasticity to minimize the actuation losses, and a ratio that can be adjusted via the cylinder sizes. The temperature-specific volume expansion of the fluid is a disadvantage, which is why a connection to the reservoir has to be established regularly to enable volume compensation.

The new hydrostatic clutch actuator (HCA) provides a further powerful component for automated clutch control. This is shown in Figure 10. Owing to the extensive optional equipment, this actuating mechanism can be tailored for a wide range of applications. The design selected also allows good scalability of the components. Use as an individual actuator for clutch-by-wire applications or for controlling

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

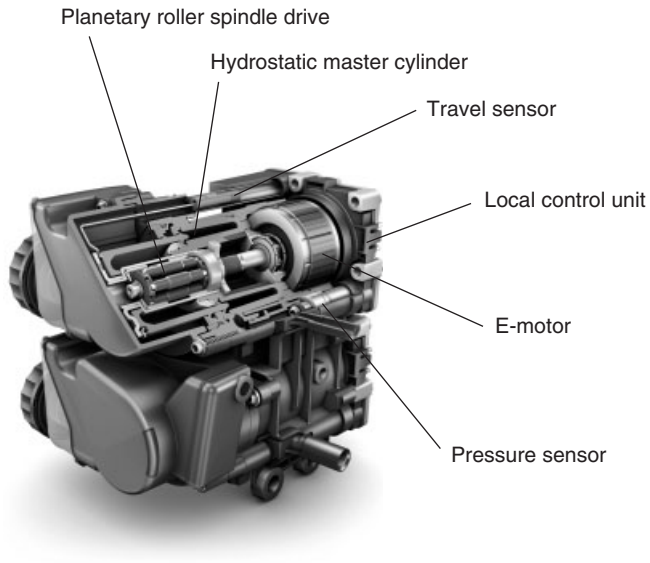


Figure 10. Hydrostatic clutch actuator. (Reproduced with permission from Schaeffler.)

hybrid clutches is also possible, in addition to use in DCTs. The use of a planetary roller spindle drive with very high ratio means that further ratios are not necessary. A large master cylinder in the form of an annular piston can also be used. This means that the actuator can have a very compact design. This type of hydrostatic actuator is used, for example, in the Honda "Intelligent Double Clutch Drive" hybrid system. This system has been optimized for very low fuel consumption. At the same time, this clutch actuator offers a previously unattained precision for dry and wet clutches from 100 to 500 Nm. Thanks to the target path regulation on the clutch, a lower hysteresis and better micro-control behavior are achieved than with a pressure control that could also be represented with this actuating mechanism. Therefore, the HCA is also suitable for very demanding control tasks.

4 SOFTWARE

4.1 Task

In modern vehicles, the software or function control has a key role. The aim is always to control a number of different actuators reliably, robustly, and precisely (Küpper, Serebrennikov and Göppert, 2006). Because these actuators take over complex and extensive tasks in automated transmissions or in the chassis control, the higher level functional logic is particularly important. Through, for example, appropriate control logic and adaption routines, this functional logic must provide a consistently high comfort over the entire life. At the same time, relevant diagnostic and safety tasks must be fulfilled to ensure that at all times, hardware faults are detected reliably and no safety critical driving situations can occur. A new task is the definition of interfaces for interactions with modern communication and entertainment systems.

The following insight into the software and function modules of automated transmission systems focuses on AMTs with single clutch (AMT) and DCTs. However, many aspects can also be applied to torque converters and hybrid systems. For a deeper insight, special aspects of the functional logic of the DCT will be addressed. In particular, the process of a gearshift from first to second gear under traction will be considered.

4.2 Software structure

4.2.1 Embedded software

Embedded software is used in systems that are generally characterized by restrictive boundary conditions such as low memory capacity, low CPU timing, or CPUs with a low bit rate (16-bit or 32-bit). The main causes are firstly, the multitude of up to 100 control units in one vehicle, and secondly, the high cost pressure for vehicle manufacturers.

The software is often programmed manually in the classic way using ANSI C, supported by corresponding structogram editors. However, code generators are also used in numerous projects. The main reason for this is that programs such as Matlab/Simulink are used in the development of relevant functional algorithms. These are well suited to mapping control-oriented functions. Using appropriate code generators, such as Targetlink, these graphical models or structures can be transferred into executable code with little effort. Requirements of this type of code, such as the avoidance of floating point operations or the definition of application accesses, can thus be implemented easily. Further benefits include direct consistency between the simulation model and the related code.

4.2.2 Concept

Core requirements of the architecture or the software concept are reusability and simple maintenance. This is guaranteed by a modular structure. At the same time, high reusability increases the degree of maturity of the functions. If attention is given to good maintainability from the beginning of development, this ensures that existing functions can be extended without problems. The example

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

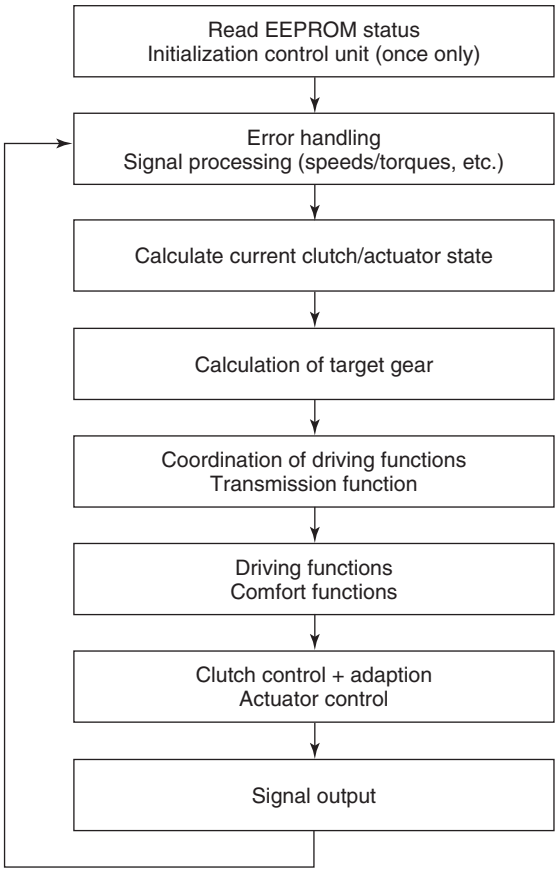


Figure 11. Signal flow diagram of the actuation of a double clutch transmission. (Reproduced with permission from Schaeffler.)

of a signal flow diagram for controlling a parallel shift transmission in Figure 11 illustrates the principle structure of the software and its main functions.

4.3 Function modules

A distinguishing feature of modular function concepts is that certain functional tasks are clearly assigned to certain functions or software modules. Therefore, we can describe the functions of an automated transmission using individual function modules.

4.3.1 Gear selection strategy

Automatic or automated transmissions usually have six to nine gears. Therefore, one of the basic functions is to select a suitable gear that best corresponds to the driver input. This task must also be seen from the perspective that the large gear spread often provides the possibility of realizing one and the same driving situations in different gears.

A sensible or appropriate driving gear is oriented not only around the engine speed range defined as permitted but also around the demands of the driver with regard to dynamics and operating efficiency. Both aspects can be determined by evaluating driver actions on the accelerator pedal, brake, selector lever, or—depending on the equipment—shift paddle. One simple option is to define the shift points on a gear-specific basis via the accelerator pedal angle and the driving speed. In many concepts, this basic information is refined via dynamic driver recognition, the incorporation of GPS data, and the recording of other environmental conditions such as traffic density. As a consequence, the determination of the correct driving gear or the operating strategy has become a defining feature of vehicle brands. The situation is similar for the preselection gear that is engaged on the inactive shaft. Company-specific different concepts are also effective here.

4.3.2 Coordination and protection functions

The appropriate driving function must be selected based on the driving situation and the corresponding target gear. This selection is made in a central coordination function that also guarantees that the clutch and transmission controls act in alignment with each other. In parallel to this coordination, relevant monitoring functions ensure that there is no damage to the transmission hardware or other operating-relevant functions. For example, a hill-start situation with the foot brake applied is interpreted as misuse. A reducing engine torque intervention can signal to the driver that this state can damage the vehicle. Further optical, haptic, or acoustic signals to the driver are possible.

4.3.3 Driving functions

A whole range of driving functions are available for an automated transmission. These range from creep functions, launch and driving functions, and the shift function to stopping functions. They are supplemented by efficiency functions such as start–stop systems and coasting functions.

A central function for DCTs is a shift operation with no interruption in tractive power. This must be realized quickly and with comparable comfort to that of an automatic torque converter transition. As an example for the various shift types, Figure 12 shows a shift under traction from first to second gear. Owing to high driver sensitivity at low speeds and the typically high gear spread, this transmission stage is vitally important.

The shift contains the following phases: in phase 1, any existing excess torque on the clutch is reduced and driving takes place specifically with slight clutch slip of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

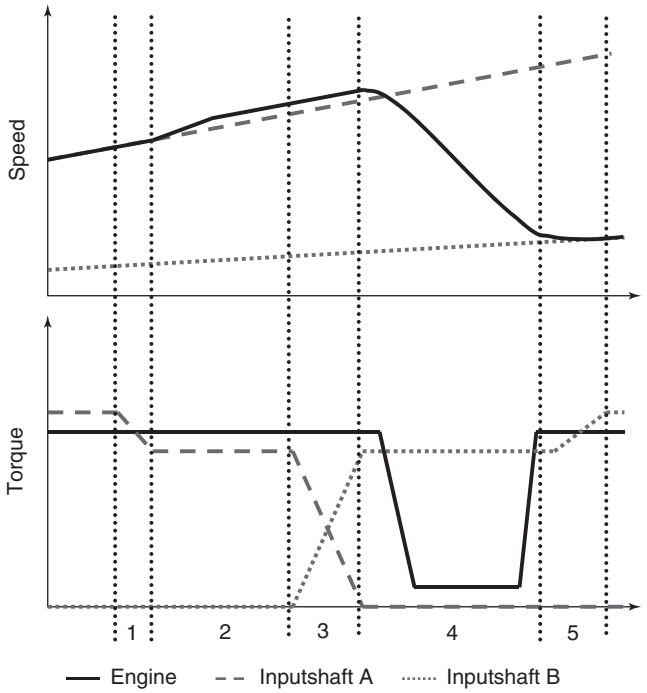


Figure 12. Shift phases (example upshift). (Reproduced with permission from Schaeffler.)

approximately 10–20 rpm. This means that the clutch torque transferred can be determined relatively precisely before the shift begins. This precision is helpful for the subsequent torque transfer.

Provided the required target gear has already been preselected, phase 2 is not necessary. If this is not the case, there is a waiting period until the required target gear is actually engaged. In phase 3, there is then an overlapping in which the clutch torque is overlapped in ramp form from the outgoing to the incoming clutch. This torque overlapping phase cannot be perceived by the driver. It is important that the total torque corresponds to the respective driver input torque. This total torque is adjusted accordingly for accelerator changes during the shift. The torques of the outgoing and incoming clutch can be corrected by observing the clutch slip during this time.

In phase 4, the speed is finally adjusted to the target gear. This generally takes place via an engine intervention. Reducing the engine torque results in a corresponding reduction in engine speed. Therefore, this phase can be perceived as a shift acoustically and optically on the rev counter. The engine intervention is withdrawn promptly before the engine speed approaches the target shaft speed. In this context, the difficulty for the control system is that at the end of the speed adjustment phase, the clutch torque must match the engine torque and the engine speed must be close to the transmission input speed. The result is that the engine speed runs gently into the target shaft speed. In phase 4, an additional boost torque can be set on the clutch. This provides the impression of acceleration during the shift. In phase 5, the clutch moves to excess torque again.

4.3.4 Comfort functions

Dry automated transmission systems are highly efficient. However, this advantage is generally at the cost of lower transmission damping properties. In order to still achieve high driving comfort despite this, additional comfort functions must be provided. These include a controlled isolation system. In applications with no dual mass flywheel, this ensures that in certain operating ranges in which strong engine excitation would occur, a minimal slippage leads to a decoupling of vibrations. Slippage values of 10–30 rpm are typical for this operating range. In transmission systems with dry clutches, the effect of partial slippage damping is used for this.

A further measure for increasing comfort is the use of a load-reversal damping function. This is implemented such that the clutch torque is slightly greater than the engine torque in driving operation. In the case of an abrupt loadchange, for example, due to sudden acceleration, the excessive powertrain torques that occur are reduced by short-term slippage. This function supplements the load switch damping of the engine control unit.

In certain cases, judder of the clutch can occur in automated transmissions. This judder can be both geometric and friction-induced judder. To dampen both types, damping functions are being developed that lead to a reduction in the vibration amplitudes through a corresponding modulation of the clutch torque.

4.3.5 Hardware-related functions

The hardware-related control functions ensure that the transmission actuating mechanism is controlled reliably and such that it can be reproduced. In clutch control, the actuator is controlled in accordance with the adapted clutch characteristic curve. Depending on the clutch type, corresponding compensation in the sense of pilot controls can be executed. In a typical case, for example, speed or temperature effects are compensated in whole or in part. Selected points on the clutch characteristic curve in Figure 13, such as the touch point/bite point and frictional coefficient, are adapted. The touch point can take place both via the evaluation of torque balances and via the evaluation of changes in speed on the inactive shaft.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

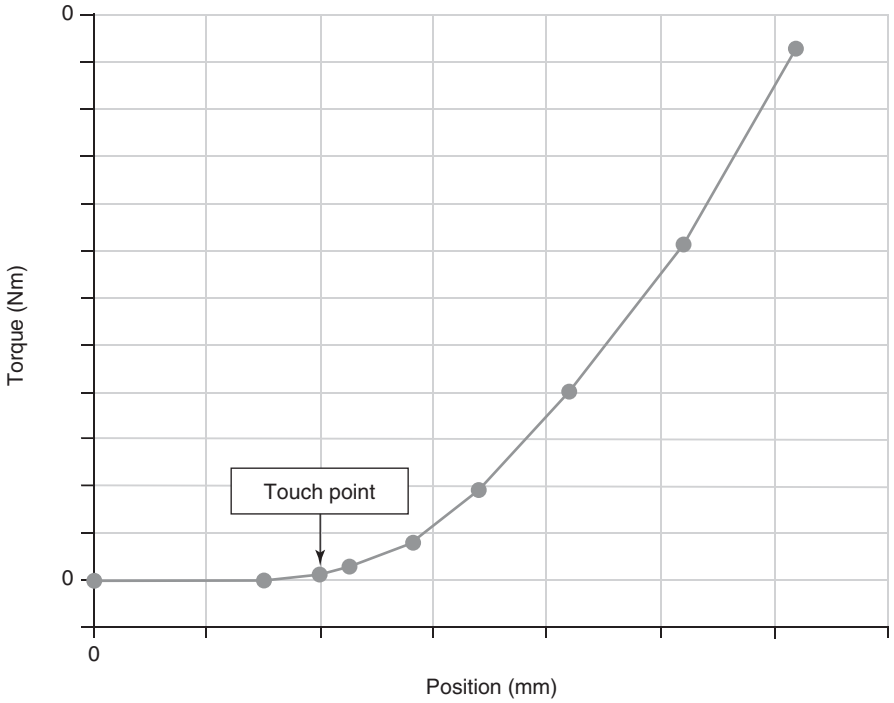


Figure 13. Illustration of a clutch characteristic curve (dry clutch). (Reproduced with permission from Schaeffler.)

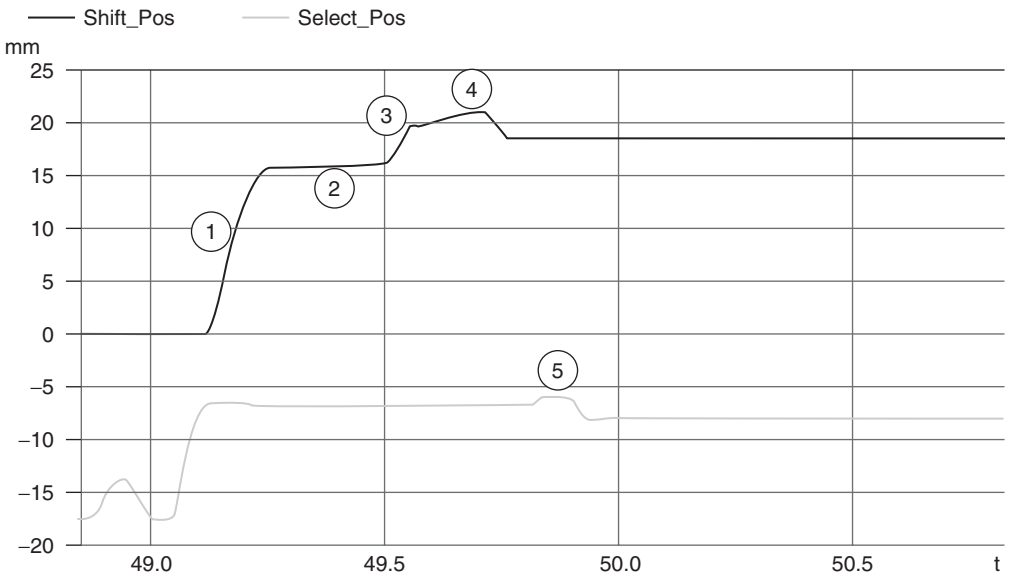


Figure 14. Actions in the transmission when engaging a gear. (Reproduced with permission from Schaeffler.)

In the same way as the shift operation already described, Figure 14 describes the actions required in the transmission when the second gear is engaged.

In the initial state, the second subtransmission is in neutral. In a first step (1), the switching finger is moved to the synchronous position. This takes place at a controlled speed, whereby the speed is reduced before the adapted synchronous position is reached.

The speed is also controlled during the synchronization phase. However, an upper force limit must not be exceeded.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

As soon as the synchronization has taken place, the gear stop position is approached with position control (3). To verify this position, a sensing operation takes place in the shift direction to check or correct the adapted values.

As incremental sensors are generally used for cost reasons, it is important to check the reliability of this actuator position. This degree of reliability is reduced, for example, if the sensors detect a Hall error or an unexpected blockade situation. Depending on the operating situation, after this type of reduction in reliability, a validation of the transmission position is requested via referencing.

A further hardware-related function is the actual actuator control. With actuators with electric motors, movement is generally in a position-controlled or speed-controlled state. With hydraulic valves, a current-controlled or position-controlled mode is generally used. Depending on the actuator concept, a voltage-controlled mode can also be used for switching valves, for example.

4.3.6 Safety concept

To prevent critical operating states from arising, a safety concept for transmission control systems is state of the art. This is based on the requirements of ISO 26262. The aim is to ensure, for example, that the vehicle cannot start driving itself, that is, a clutch torque build-up is prevented if no driver presence is safely detected.

4.3.7 Diagnostics concept/error handling

Error handling is often based on a modular, generic concept, whereby the error variables in the sense of DTCs (detection trouble codes) are limited to a few error routines. A majority of the DTC routines are typically used for all vehicles (DTCs for engine signals or actuator errors).

In many cases, there is a differentiation between customer-specific and customer-independent routines in the signal recording routines. The aim of this is to ensure that no unnecessary overhead is created. When an error is detected, a corresponding status bit is usually set so that the driving/comfort functions that access it can access corresponding replacement values. Replacement strategies only have to be used in exceptional cases, for example, if certain signal information or hardware components are no longer available.

4.3.8 End-of-line routines

To ensure that automated transmissions drive reliably and comfortably at all times from the beginning of production, certain transmission-relevant parameters are adapted during the production process. Special end-of-line routines learn these parameters. These routines can also be used subsequently for teaching in a garage if, for example, hardware components have been replaced.

4.3.9 Hardware drivers

To control the corresponding actuators (electric motors, valves, etc.), the control unit on which the software is operated must have corresponding hardware drivers. Often, valve drivers are used that generally modulate a high frequency voltage. The self-adjusting currents are actuator-dependent and range from just a few amperes up to 100 A. At high currents in particular, this represents a high demand on the load capacity of the vehicle electrical system. This is one of the reasons why vehicle electrical systems are subject to permanent further development.

5 TECHNOLOGY DRIVERS FOR FUTURE DEVELOPMENTS

In Europe, the general trend toward automation and the availability of the manual transmission as a basic technology have led to intense developments in recent few years, in particular for DCTs. If the automated forms of manual transmission are also to continue to expand and compete with classic automatic transmissions, however, further improvements must be implemented in the areas of efficiency (including weight and mounting space), comfort, and costs. The same applies for the start/stop capability and the capacity of the systems for hybridization. The actuating mechanisms of these transmissions always play a key role. This leads to the technology drivers listed below for the actuating mechanisms.

To reduce the power input of the actuating mechanisms, on-demand systems will prevail. These are systems that adapt their power input to the required actuator power. In the case of hydraulic actuators, this means that the pumps must be controlled on a demand basis. Furthermore, the leakages must be reduced by reducing the gap and/or the nominal working pressure. In the case of electromechanical actuators, nonlinear ratios offer a starting point for reducing the base powers and thus enabling the construction of servo motors (including the electronic power unit) that are smaller and thus more efficient.

However, in order to construct these smaller and more efficient servo motors, the degree of integration must increase further. For electromechanical actuators, this means that additional mechanical functions will be used in an attempt to reduce the need to install electric motors. It is also to be expected that the mechatronic system components engine, control unit, sensors, and mechanics

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

will be integrated more deeply. In a further step, it may even be possible to integrate the clutch actuating mechanism in the clutch. These types of solutions are currently being investigated based on dry double clutches.

One very important cost driver for DCT actuating mechanisms are rare-earth materials. These materials are used for magnets with high power density for sensors, electrohydraulic valves, or electronically commutated electric motors. The market prices for rare earths have risen exorbitantly in recent years. Therefore, motors that use as little magnetic material as possible are increasingly being used. Sensors that can be constructed without this magnetic material are also being developed. Further cost reductions for sensors can be achieved through standardization and integration in the TCU (transmission control unit). Furthermore, more intelligent software should reduce the number of sensors used. However, it is not just the electronics that are cost drivers—the mechanics and hydraulic modules are also cost drivers. For these modules too, attempts must be made to reduce costs via standardization or module solutions and thus an increase in quantities.

Current development efforts around increasing comfort, additional functions, hybridization, and start–stop capability place additional requirements on the performance and functionality of the actuating mechanisms. Initially, this means that the TCUs and in particular, their controllers, are becoming more powerful. However, for functional enhancements, additional actuator elements for controlling the additional control elements, for example, of hybrid clutches are required. This will increase the complexity of the actuator system. For start/stop and hybrid operation in particular, actuators are required that can be operated independently of the combustion engine. At least one energy storage device must be provided to allow actuator operation to restart the engine.

However, the greatest challenge is in covering the increasing complexity through increasing performance requirements and the reduction of cost, weight, and mounting space.

REFERENCES

Drexler, P., Faatz, H., Feicht, F., *et al*. (1988) *Planning and Design of Hydraulic Power Systems*, 1st edn, Mannesmann Rexroth, Lohr am Main.

Faust, H., Bünder, C., and DeVincent, E. (2010) Doppelkupplungsgetriebe mit trockener Kupplung und elektromechanischer Aktuatorik [Dual clutch transmissions with dry clutch and electromechanical actuating mechanism]. ATZ 4/2010, 270.

Hadler, J., Metzner, F., Schäfer, M., *et al*. (2008) Das Siebengang-Doppelkupplungsgetriebe von Volkswagen [The Volkswagen seven-gear dual clutch transmission]. ATZ 6/2008, 513.

Kimmig, K., Bührle, P., Henneberger, K., *et al*. (2010) Success with Efficiency and Comfort. LuK Symposium 2010, proceedings, 153.

Küpper, K., Serebrennikov, B., and Göppert, G. (2006) Software for Automized Transmissions. LuK Symposium 2006, proceedings, 155.

Mayer, E. (2006) Serielle Bussysteme im Automobil [Serial bus systems in automobiles]. elektronik industrie 6/2006, 70.

Müller, B., Kneissler, M., Gramann, M., *et al*. (2010) Advance Development Components for Double-Clutch Transmissions. LuK Symposium 2010, proceedings, 172.

Pohl, K. (2008) *Requirements Engineering*, 2nd edn, dpunkt Verlag, Heidelberg.

Schäfer, M. (2011) Souveräne Kraftentfaltung [Superior power development]. ATZextra 9/2011, 44.

Stark, R., Schuch, B. (2010) Innovative Technologies for Transmission Control Units. LuK Symposium 2010, proceedings, 191

Steiger, S., Treder, M., Neuberth, U., *et al*. (2013) *Innovative Weiterentwicklungen bei trockenen Doppelkupplungssystemen [Innovative and Advanced Developments for Dry Double Clutch Systems]*, VDI-Fachtagung Kupplungen und Kupplungssysteme in Antrieben, Karlsruhe.

Wagner, U., Bührle, P., Müller, B., *et al*., (2009) Dry double clutch systems. ATZ 11/2009, 826.

Wörner, R., Damm, A., Eberspächer, R., *et al*., (2011) Effiziente Front-Quer-Getriebe von Mercedes-Benz [Efficient front-transverse transmissions from Mercedes-Benz]. ATZ 12/2011, 915.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto104
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

AT Control—Actuation Methods and System Integration, Gear Choice, Gear Shift Strategy and Process, Adaptive Features

Takashi Shibayama
JATCO Ltd., Fuji City, Japan

1 Introduction 1
2 At Control System 1
3 At Shift Control Procedure 2
4 Basic Concept of at Shift Schedule Design 3
5 Various Additional Control for at Shift Schedule 7
6 Control for Good at Shift Feeling 7
7 Engine—At Integrated Shifting Control 10
8 Example of Detail Time Chart for Shifting Control 11
9 Other at Control 11
Acknowledgments 13
References 13
Further Reading 14

1 INTRODUCTION

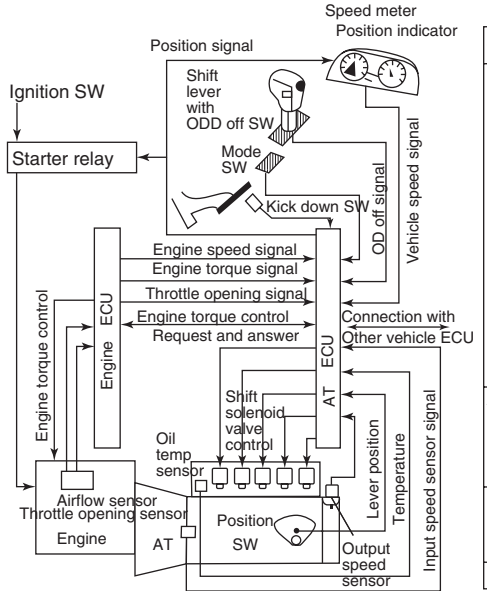
In order to understand the basic concept and outline of AT control, the following topics are discussed. Torque converter clutch control, idle neutral control, idle stop control, and launching clutch control are also briefly discussed in this chapter.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 AT CONTROL SYSTEM

Figure 1 shows a typical AT control system that has several sensors, actuators, and communication interfaces. The most important sensors are speed sensors by which the AT ECU (electronic control unit) can detect gear ratio, slip ratio of the torque converter, vehicle speed, etc. (Shinohara *et al*., 1989) The AT ECU makes gear ratio changes by detecting vehicle speed and engine throttle opening, according to the predetermined shift schedule. The position sensor is used for changing shift schedule for L range, sensing R range, etc. Most AT control systems include a kick down switch for detecting the driver's intention for rapid vehicle acceleration and selecting a lower gear ratio to maximize performance. The intake airflow sensor of the engine is used for calculating engine crank shaft torque, which is very important for providing the proper hydraulic pressure to control shift smoothness and preventing unwanted slippage of the friction clutches and brakes. A temperature sensor is used for detecting ATF temperature, which is an important parameter for maintaining consistent shift feel at varying operating temperatures. The sensor can also be used to detect over heating situation and initiate a "fail safe" mode to protect the transmission from damage. Also for low ATF (automatic transmission fluid) temperatures, like just after engine starting, some adjustment for the ATF viscosity change can be compensated by this sensor.

The table in Figure 1 shows typical devices used for AT control. Communication between the engine and the AT is essential to ensure smooth and quick shifts. During shifting, engine torque control is very effective to improve shift



Category	Device	Major act of the device
Sensor	Engine speed sensor	To detect torque converting slipping ratio
	AT input speed sensor	To detect gear ratio of AT
	AT output speed sensor	To detect gear ratio of AT
	Vehicle speed sensor	To detect wheel slip
	Position sensor	To detect P.R.N.D.L opening range
	OD off SW	To inhibit over drive range
	Mode SW	To switch mode, for example, economy or sporty
	Kich down SW	To detect driver's high intention for acceleration
	Engine throttle opening	To detect driver's intention for acceleration
	Engine airflow meter	To detect engine output torque
	Temperature sensor	To detect AT oil temperature to know viscosity of oil
	Other sensor	It depends on the purpose of some additional special control For example, hydraulic pressure sensor, SW, g sensor, ect.
Actuator	Solenoid value	To execute AT shifting by controlling hydraulic pressure
	Engine torque adjusting device	To adjust engine torque during AT shifting
	Indicator lamp	To indicate P.R.N.D.L range status to driver
	Starter relay	To allow engine starting during P.N range only
Communicater	Engine - AT communicator	To agree engine torque adjust during AT shifting
	Other vehicle ECUs - AT communicator	To make communication with ADS, ASCD, NAVT, ect. for some additional special control
	*Recently, both above communications are usually done by CAN network	

Figure 1. A typical AT control system. (Reproduced by permission of Jatco, Ltd.)

quality. Finally, the vehicle CAN network enables good communication with other vehicle ECU's such as ABS (antilock brake system) and NAVI (automotive navigation system). The AT ECU can even change shift schedule by watching this information.

3 AT SHIFT CONTROL PROCEDURE

Figure 2 shows a typical 1 to 2 shift control process. The AT ECU calculates vehicle speed based on output speed sensor signal and calculates throttle opening based on

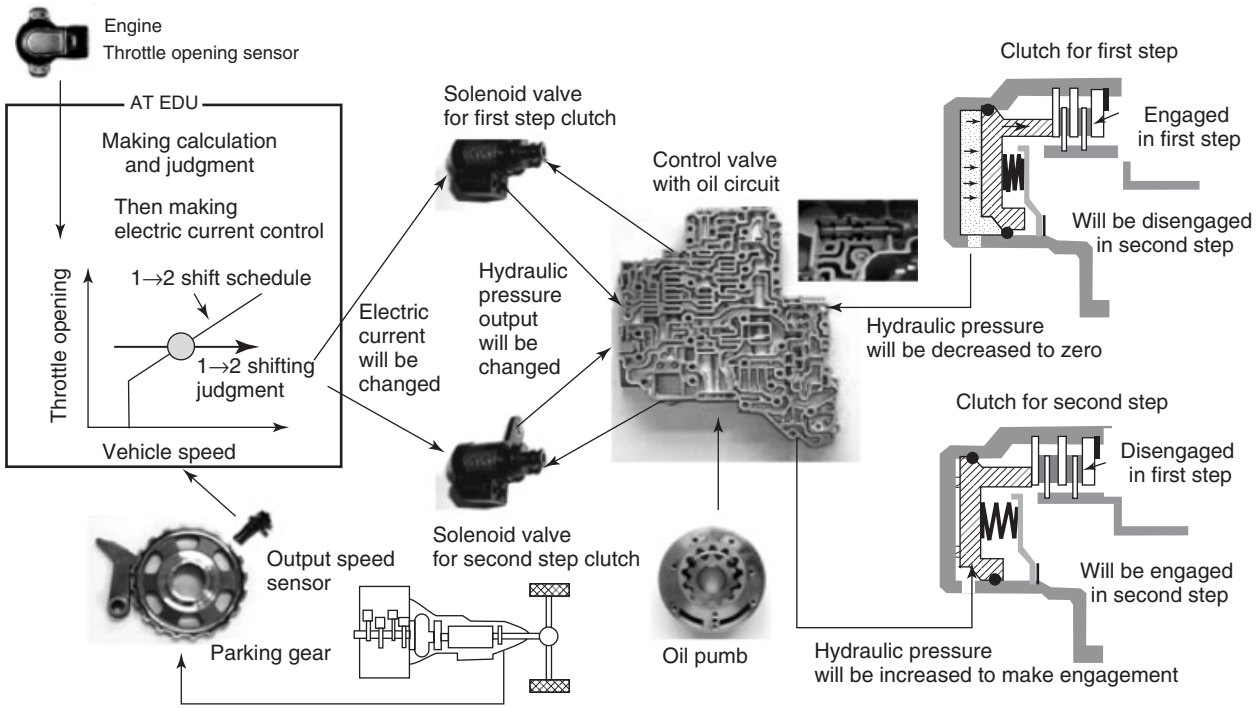


Figure 2. A typical 1 to 2 shift control process. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

throttle sensor signal. During acceleration, vehicle speed will increase and when the speed exceeds the 1 to 2 shift line, the shifting decision will be made. Then, the AT ECU will change the electric current of solenoid valves, thereby changing their hydraulic pressure output. This hydraulic pressure output from the solenoids is then directed through the oil circuit in the control valve body to the shifting clutches. By this process, first clutch status will be changed from engagement to disengagement, whereas second clutch status will be changed from disengaged to engaged.

4 BASIC CONCEPT OF AT SHIFT SCHEDULE DESIGN

For making an adequate shift schedule, designers have to understand the torque converter characteristics. Figure 3a shows the conversion characteristics as a function of throttle opening. A throttle opening of 8/8 is the widest opening, and 0/8 is the idle condition. By its hydraulic conversion characteristics, the torque converter will produce a high output torque when the vehicle stationary similar to that of an electric motor.

This torque converter output torque forms the input torque to the transmission gear train that consists of planetary gears and clutches to make several shift steps as discussed in (see Automatic Transmissions—Geartrain Combinations, Components, Design Considerations, Hydraulic System, Packaging, Manuf., Assembly). Figure 3b shows the gear train conversion characteristics, in the first step and second step as an example.

Now, as an example, consider design for a 3/8 throttle 1 to 2 shift point. As is shown in Figure 3c, there is a difference in output shaft torque between the first step and the second step. In the first step, the torque is high initially but quickly decreases as engine speed and output speed. On the other hand, in the second step, a higher speed is available but the output shaft torque continues to decrease. It becomes obvious that the cross point is the best shift point to provide the most powerful acceleration. Figure 3d shows the cross point shifting schedule strategy.

Upshift schedules are investigated considering the above-mentioned cross point. Figure 3e shows a typical first to second upshift schedule. In addition, Figure 3f shows a typical upshift schedule. Here, throttle opening is popularly used for calculating engine torque as well as detecting the driver's acceleration intention. Vehicle speed is used instead of output shaft speed, which is more apparent when the test driver test drives a vehicle.

Figure 4 shows various shift schedule adjustment issues. Figure 4a shows adjustment for shift feeling improvement avoiding hesitation or harsh shift. Actually, the shift schedule is usually set before the cross point, because deviation torque caused by engine inertia energy has to be considered to avoid harsh shift during shifting.

During 1 to 2 shifting (or any upshift), as engine speed goes down, an inertia torque is generated because of the change in engine rotational speed. The additional torque is added onto the second gear output torque as shown in the figure. Thus, it is popular to set shift points a little bit before the cross point. Figure 4a shows a typical upshift point design concept.

Figure 4b shows downshift schedule adjustment to avoid busy shift. As is shown in the figure, usually the hysteresis between upshift and downshift is provided. This hysteresis is necessary to guard against frequent up- and downshifting at or near the shift line with slight throttle opening changes. Such frequent up- and downshifting is called *busy shift* or *shift hunting*, which is most often seen when going up a grade or hill.

The right side of Figure 4b shows an example of such busy shifting risk. In the figure, road load line is added, for example, 0% road load, 3% road load, and 6% road load. As is shown in the figure, a wider throttle opening is required on a steeper hill, in order to maintain a fixed vehicle speed. On a steep hill, the greater throttle requirement to maintain vehicle speed can easily place the throttle in between the upshift and downshift lines. A slight movement of the accelerator pedal can trigger a downshift or upshift repeatedly. Greater hysteresis can minimize this shift busyness or hunting. This greater hysteresis enables wider throttle opening before a downshift is triggered and also tends to inhibit upshifting until hill climbing is over.

Fuel economy is another important issue for shift schedule adjustment. Figure 4c shows the fuel economy map of an engine. There exists an optimal operation point for fuel economy. Therefore, designers always check their shift schedule to aim at frequent usage of this optimal area. Sometimes, the investigation result tells the designer to use a higher engine speed to get good fuel economy. However, it depends on characteristics of the engine. Usually, a low engine speed is good for fuel economy.

Shift schedule adjustment for wide throttle opening is another important issue. When the throttle opening is wide, the driver is asking for good acceleration; therefore, in a wide throttle opening condition, acceleration performance gets priority. On the other hand, in the small throttle opening area, fuel economy is most important. Therefore, such investigations using engine fuel economy maps are often done for the small throttle opening area.

In actual traffic, drivers sometimes want to make full power acceleration utilizing engine max power. For such a purpose, a kick down area is usually adopted. There

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

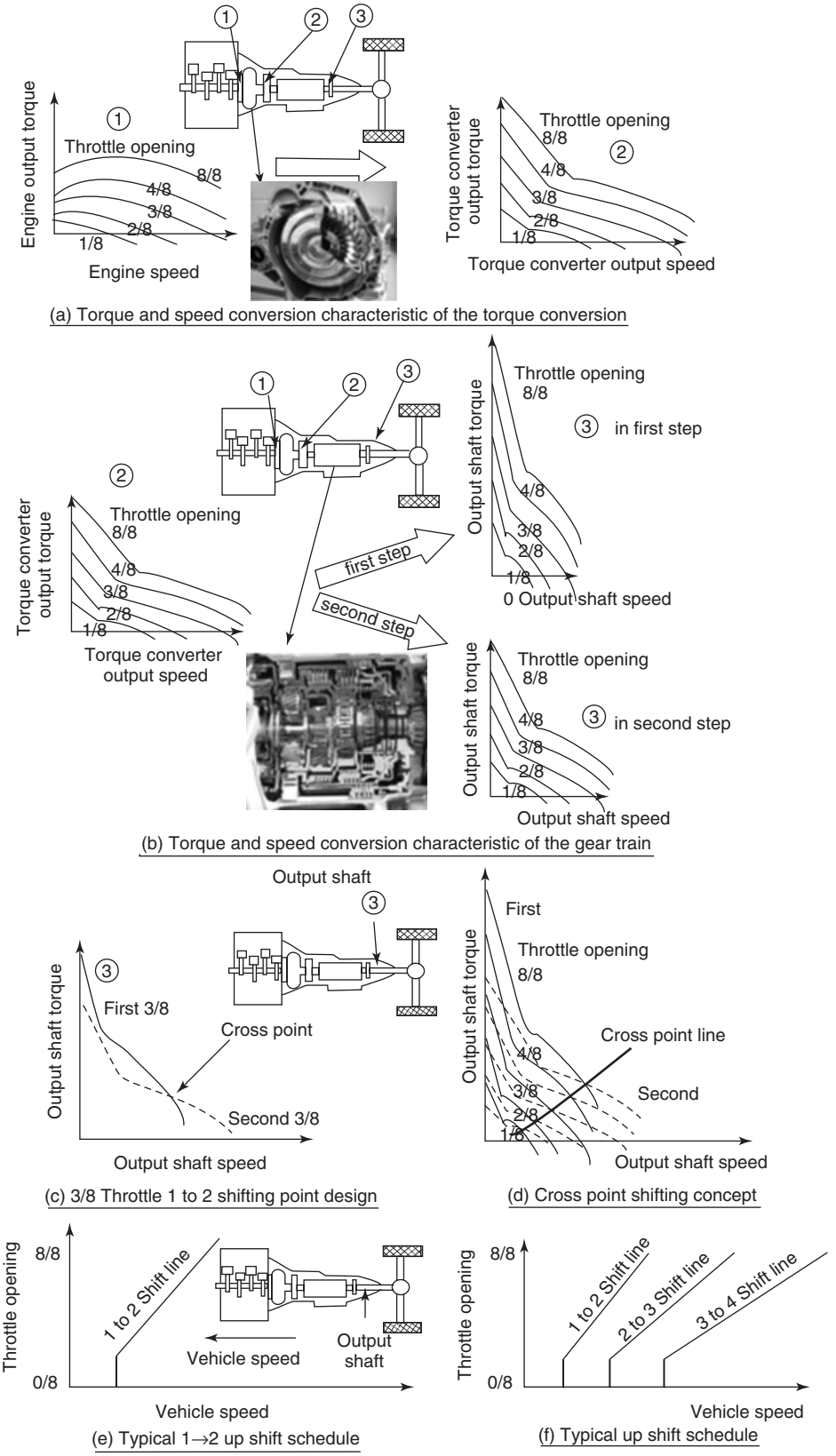


Figure 3. (a–f) Investigations for producing AT shift schedule. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

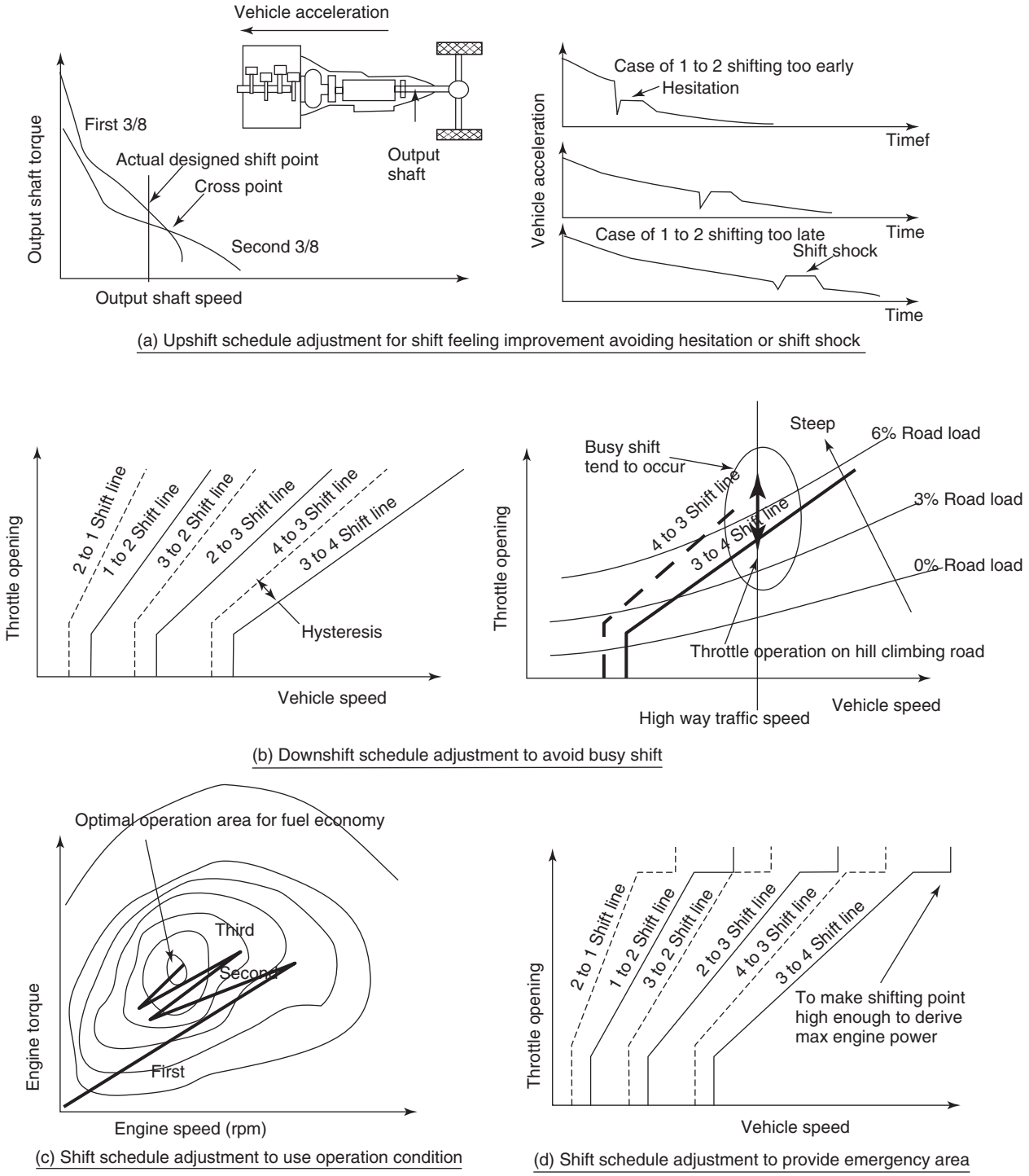


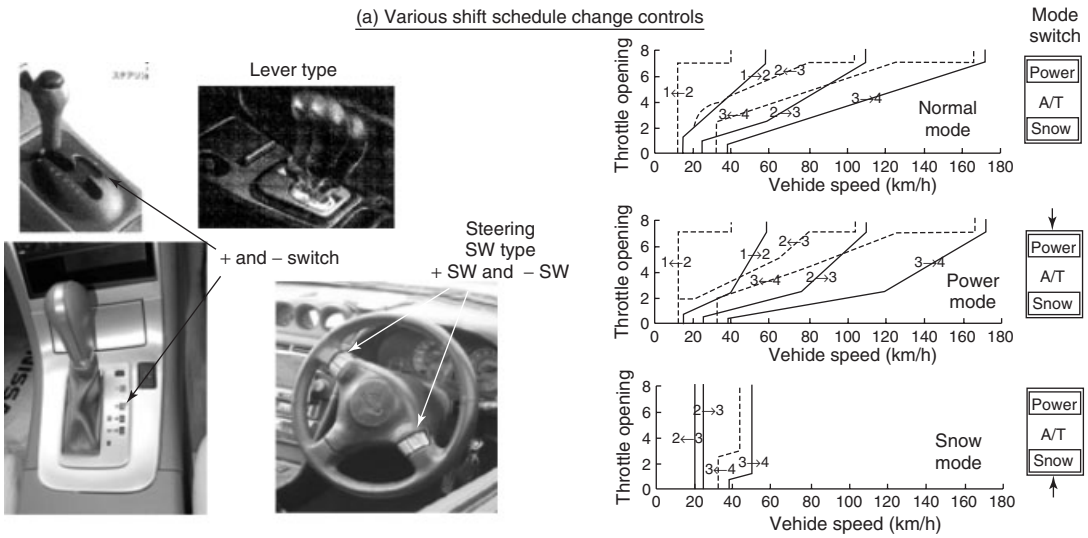
Figure 4. (a–d) Various shift schedule adjustments. (Reproduced by permission of Jatco, Ltd.)

are two ways to accomplish this: one is to watch only throttle opening. The other one is to facilitate a kick down detecting switch, which has some resistance force (or "detent" feel) to push on. By the resistance force, the driver can notice where the limit of normal control is and from where kick down control begins. The switch is a convenient device for drivers, some systems make a clicking sound when the switch is stepped on, thereby providing feedback

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

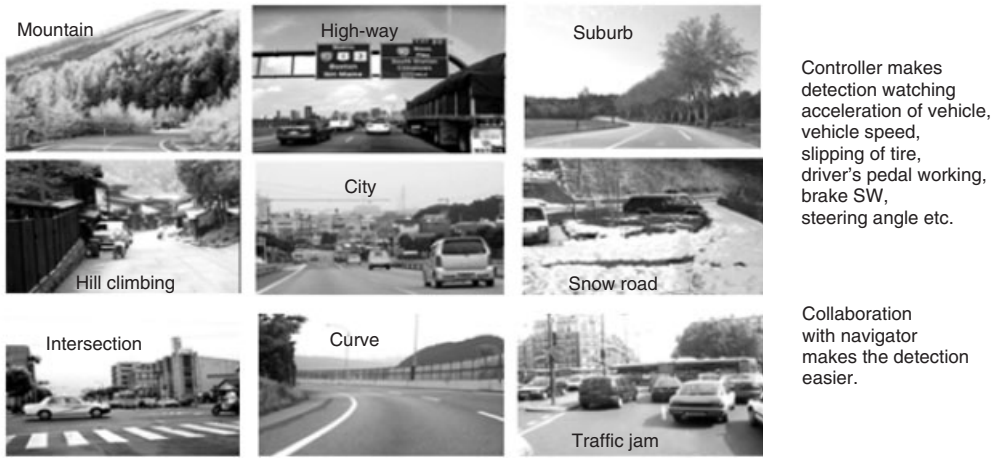
Category	Control device, control way	Control contents
Manual shift	Manual mode SW shift lever tipping Paddle switch	1,2,3,4,5,6,7 gear hold gear will be changes only by driver's operation. Engine overrunning and engine stall can be prevented automatically
Schedule change by manual	Economy, sporty made SW snow mode. winter mode SW OD INH SW	Shift schedule will be changes by driver's SW operation. Higher or lower shift line will be adopted. Some gears may be avoid for some selected purpose.
Schedule change automatically	Adaptive shift control Hill climbing mode auto engine braking before stop	Watching vehicle's condition and driver's in tension for acceleration or braking by detecting acceleration and driver's pedal working, braking SW etc. Shift schedule will be automatically changed, or gear will be automatically kept, for example on mountain road or by sporty pedal working.
	Collaborate with steering system collaborate with navigation system	Watching road condition curve, traffic signal etc. gear will be automatically changed or automatically be kept temporally

(a) Various shift schedule change controls



(b) Various gear selector system for manual mode

(c) Example of shift schedule change by mode switch



(d) Adaptive shift schedule control for various driving conditions

Figure 5. (a–d) Various additional shift schedule controls. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

information to the driver. Once the kick down area is entered, down shifts will be made as available as possible, and engine max power operation can be achieved. Figure 4d shows such a shift schedule with kick a down area.

As discussed earlier, basic concept of AT shift schedule design has three important points of view.

(i) Driving PowerAbility of obtaining good acceleration performance and keeping vehicle speed without busy shifting. In emergency conditions, max power can be derived.
(ii) Fuel EconomyPossibility to use optimal engine operation point.
(iii) Shifting SmoothnessTo consider adequate shift timing to avoid hesitation and shift shock.

5 VARIOUS ADDITIONAL CONTROL FOR AT SHIFT SCHEDULE

There are various additional shift schedule controls; Figure 5 shows variations (Yamaguchi *et al*., 1994; Petersmann, Seidel and Moellers, 1990; Stebar *et al*., 1990).

6 CONTROL FOR GOOD AT SHIFT FEELING

During gear shifting, especially, during upshifting, fluctuation on output torque will be caused as is shown in Figure 6. Just after shifting has been started, clutch capacity for high

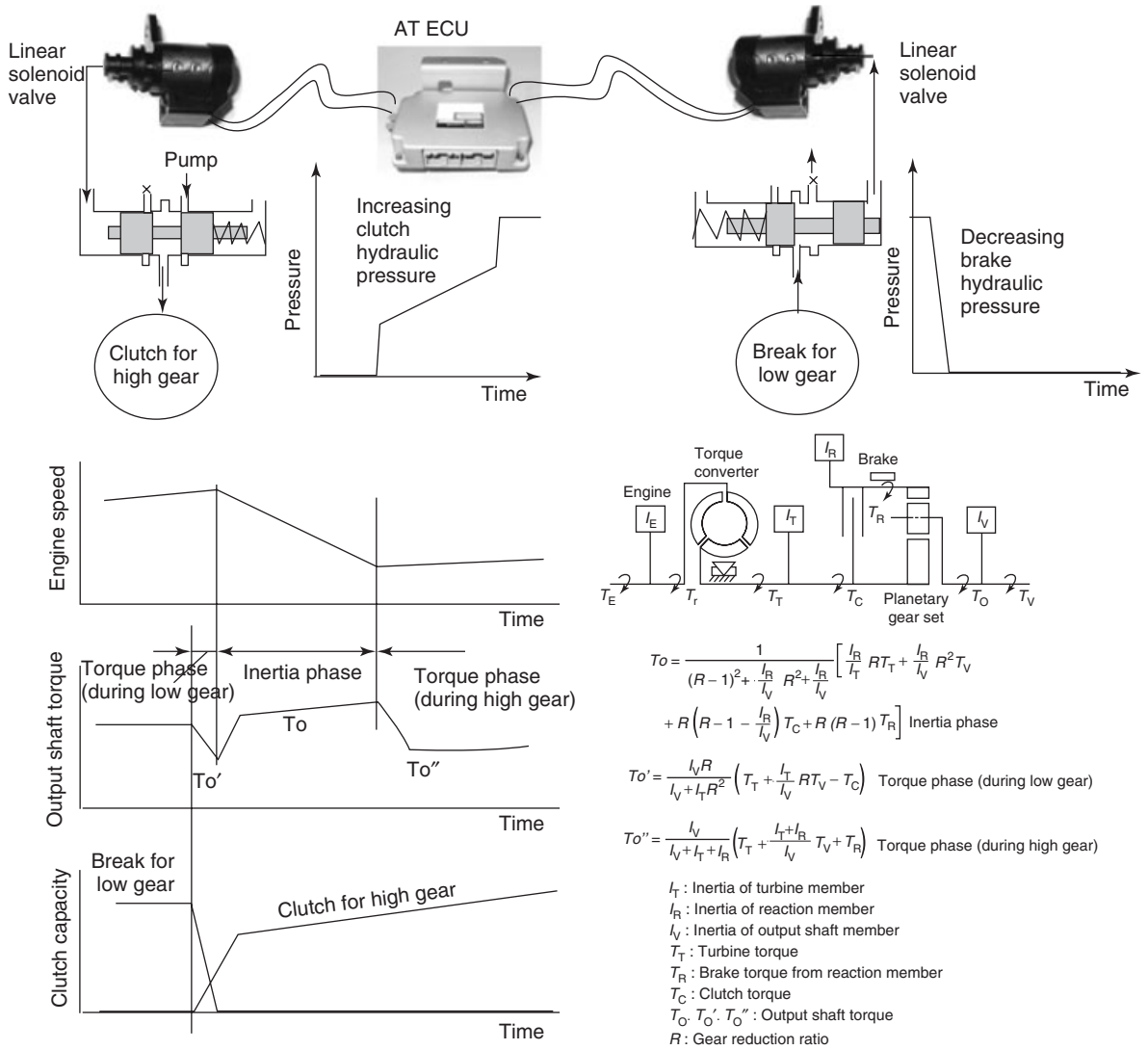


Figure 6. A typical upshift feeling control system. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

gear will increase, and by that increase, conflict (or "tie up") between the clutch for low gear and the brake for high gear will be caused. By the conflict, output shaft torque will be temporarily decreased, this period is called *first-torque phase*. Then, engine speed will go down and additional output torque, caused by an inertia torque generated because of the change in engine rotating speed, will be added. This period is called *inertia phase*. Finally, shifting will be completed and "second-torque phase" will come. For shift control, it is very important to control such torque fluctuations. If the fluctuation is too large, drivers feel it as harsh shift. Harsh shift can be prevented by precise control of clutch capacity.

The output shaft torque To, To′, To' can be calculated by theoretical calculations, as shown in Figure 6 (Society of Automotive Engineers, 1962). As is shown in the theoretical output torque formula, the output shaft torque is influenced by clutch torque capacity. In other words, output torque can be controlled by precise clutch capacity control. Referencing this theory, an AT shift control system is "controlling the engaging clutch hydraulic pressure, and the disengaging brake pressure precisely, to obtain expected torque fluctuation wave form." The solenoid valves and ECU provide this pressure control.

Proportional solenoid valves are often adopted for such precise clutch hydraulic pressure control. Output pressure is almost linear in relation to the electric current going through the solenoid.

During shifting, the AT ECU provides electric current control to the solenoids, in order to obtain adequate increasing hydraulic clutch pressure (Wilfinger and Thompson, 1988).

This increasing hydraulic pressure is very similar to the gradual clutch engagement of a manual transmission. A driver of a manual transmission is making similar control to obtain smooth shift feeling.

In actual calibration, optimal pressure control is derived empirically. Figure 7 shows the calibration methodology (Shinohara *et al*., 1989).

If clutch pressure increases suddenly like case 1, harsh shift will result. On the other hand, if clutch pressure increases too slowly or at too low a level like case 3, shift duration will become too long, thereby providing a slip feeling. The calibrator watches and evaluates shift feel and decides on optimal pressure shape.

Such calibration is executed at first on the test bench with a dynamometer using the transmission and engine only. After test bench calibration has been completed, final calibration using an actual vehicle will be executed.

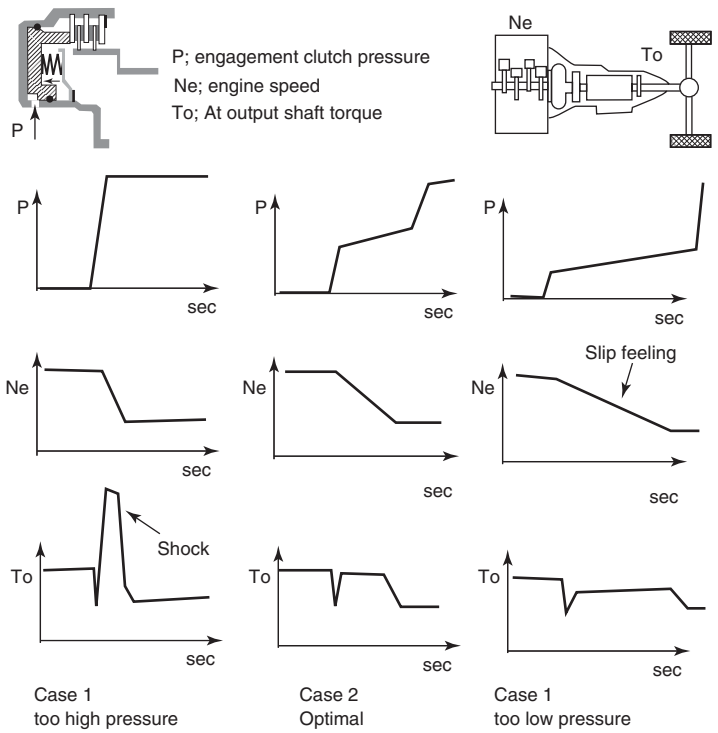


Figure 7. Clutch pressure calibration methodology. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

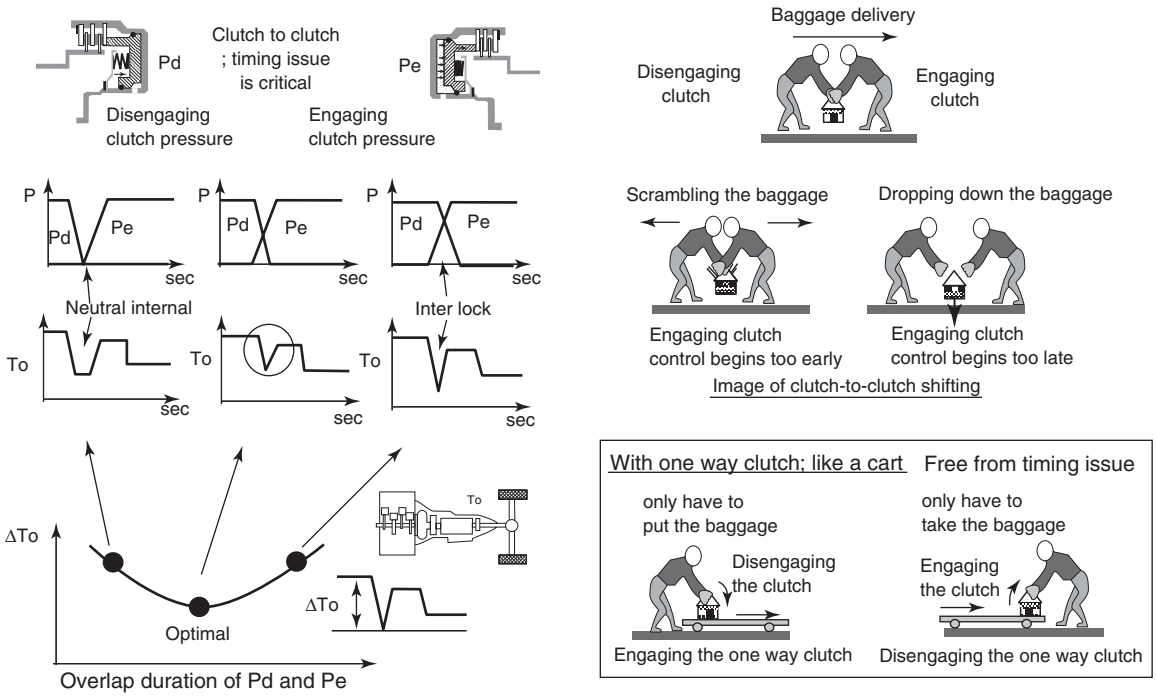


Figure 8. The usefulness of the one-way clutch for the shift feeling control. (Reproduced by permission of Jatco, Ltd.)

In vehicle calibration, calibrators evaluate not only harsh shift but also quickness, response, time lag, engine sound changing feeling, etc. All factors that drivers may notice will be checked.

If a one-way clutch is adopted, such calibration is very easy. Shift control has to be achieved using the engaging clutch. However, if a one-way clutch is not applied, shifting control becomes more difficult. Clutch-to-clutch shifting control has many difficulties.

Figure 8 shows the image of such difficulty of clutch-to-clutch shifting control and the usefulness of the one-way clutch. Clutch-to-clutch shifting is like baggage delivery from one person to another person (Society of Automotive Engineers, 1973). Timing is critical. If engaging person begins his work too early, conflict may happen. On the other hand, if engaging person begins his work too late, the baggage will drop down to the ground. This is the difficulty of clutch-to-clutch shifting control.

With a one-way clutch, the delivery is like using a cart. Disengaging person only has to put the baggage on the cart and engaging person only has to take the baggage from the cart, so, shifting control becomes free from timing issues. This is the reason why many ATs adopt one-way clutches. If a one-way clutch is eliminated, cost and space will be saved, but shifting control becomes much more difficult.

However, with recent remarkable improvements of hardware and software, clutch-to-clutch shifting control is becoming more widely used, the evolution of control technology has conquered the difficulty of clutch-to-clutch shifting control.

Cutch-to-clutch shifting control calibration is shown in Figure 8. Calibrators search for the best timing to obtain minimum fluctuation of transmission output torque. Beginning of disengaging clutch pressure decrease and beginning of engaging clutch pressure increase timing is critical. Both neutral interval and interlocking causes large output torque fluctuations. The fluctuation will be minimal at optimal timing control. With optimal control, the fluctuation of output torque will become almost the same as that of one-way clutch shifting.

There are many control approaches to make optimal pressure control.

Time Table Type. Making a timing table for various shift conditions

Sequence Control Type. Facilitating pressure sensor or pressure sensing valve that detects pressure movement.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

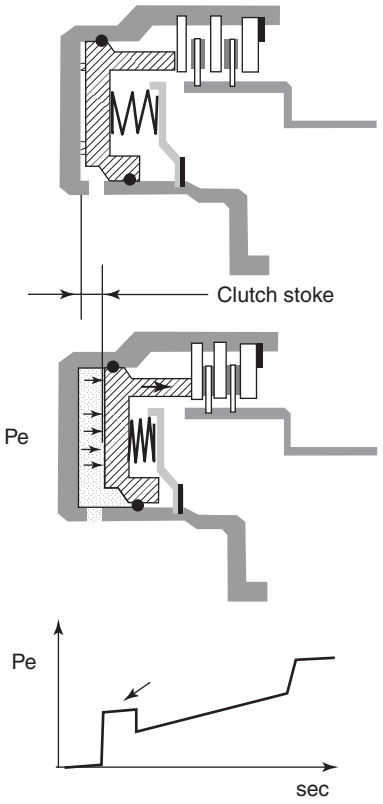


Figure 9. Precharge control to reduce time lag. (Reproduced by permission of Jatco, Ltd.)

Adaptive Learning Control Type (Narita, 1991). Watching engine speed carefully and making slight pressure adjustments for subsequent shifts.

Real-Time Feedback Type. Watching engine speed or pressure carefully and making real-time adjustment.

In actual control, the time lag caused by clutch stroking taking up free play is an important factor to consider. To reduce the time lag, precharge control is sometimes adopted. Figure 9 shows the precharge control. Until clutch piston stroke is over, the controller keeps clutch inlet pressure slightly high, and by this high pressure, large oil flow into clutch pressure chamber can be obtained.

The precharge control is especially important for garage shifting (N to D and N to R selector movement) control. Because the time lag is an important evaluation point, the controller has to reduce time lag by prestroking the clutch as much as possible.

However, the slightly high pressure becomes harmful if high pressure remains after clutch stroking. The remaining high pressure to the piston will result in harsh shift feeling.

Therefore, duration of precharge should be kept to an optimal time length by adopting some adjusting control: for example, constant map for various conditions or some adaptive learning control. Here, the influence of oil temperature is very large. At low temperatures, viscosity of the ATF is high, so precharge duration should be longer: for example, garage shifting control just after engine starting in winter.

7 ENGINE—AT INTEGRATED SHIFTING CONTROL

During AT shifting, the engine speed has to change to adjust its speed to the next gear position. During upshifting, the engine speed goes down, and during downshifting, engine speed rises up. Engine-AT integrated control can assist with this engine speed change. During upshifting, the engine ECU reduces engine torque to help engine speed going down, and during downshifting, the engine ECU increases engine torque to help the engine speed to increase. This is the basic concept (Kondo *et al*., 1990).

As discussed in Section 5, transmission output torque is related to engaging clutch torque capacity and turbine torque. Therefore, if engine torque makes a change, then transmission output torque will be affected. For the integration control, the detail design is investigated considering the theoretical formula described.

Figure 10 shows detail for engine-AT integrated control. By reducing engine torque during upshifting, shift duration can be shortened and the heat generated on the friction material in the engaging clutch can be reduced. Shift feeling is improved to have quick, sharp feeling, and durability of friction material can also be improved. Therefore, this integrated control has become very popular for AT shifting control.

Recently, throttle control of the engine has been improved greatly by introducing electronic throttle control. By this new system, the engine ECU can increase its output torque during downshifting. This is very effective in making the down shift feel quick and sharp. Therefore, many ATs with manual shift control for sports cars adopt this control to produce quick and sharp shifts.

Engine torque control is effective for not only helping the engine speed change but also preventing sudden torque rises during downshifting. The right side of Figure 10 shows integrated control during downshifting. By temporarily reducing engine torque, the sudden rise of driving force can be prevented and harsh shift can be avoided.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

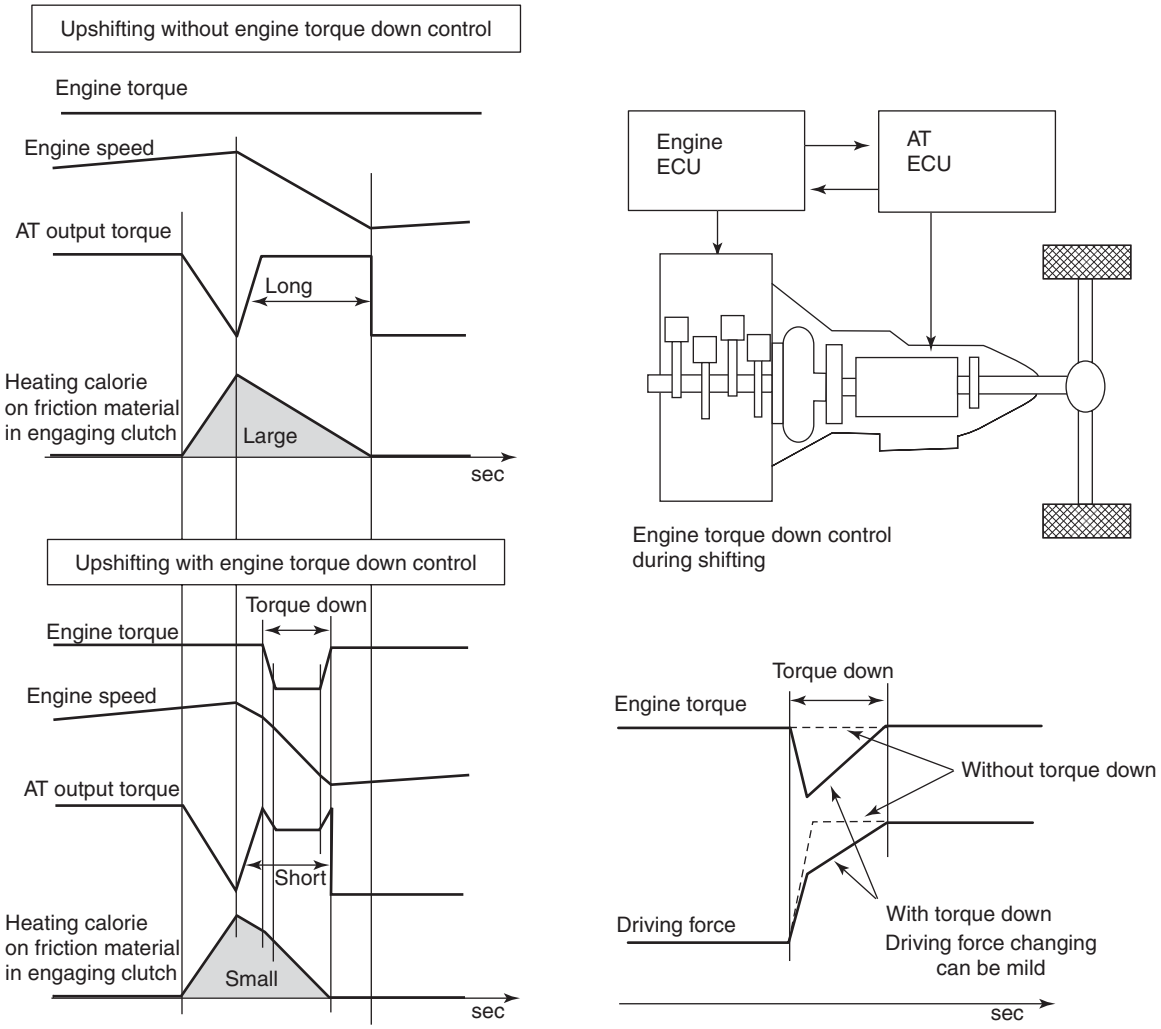


Figure 10. Engine-AT integrated control. (Reproduced by permission of Jatco, Ltd.)

8 EXAMPLE OF DETAIL TIME CHART FOR SHIFTING CONTROL

As for typical upshifting and downshifting, a detailed timing chart for shifting control is described in Figure 11. Examples are clutch-to-clutch shifting, second to third upshifting on the left, and third to second downshifting on the right.

9 OTHER AT CONTROL

As the final section of this chapter, other AT control functions are described.

Figure 12 shows various other AT controls

9.1 Torque converter lock-up control

In order to improve fuel economy, the torque converter is locked up during cruise driving. At low vehicle speeds, "controlled slip" control is often adopted. By this slip control, lockup clutch facing material is slowly slipped around 20–70 rpm. Vibration coming from the engine can be filtered by the slight slipping and that slip control produces a much lower loss than the normal "open" torque converter without lock-up slip control. This control is also done with solenoid valves and hydraulic valves as is shown in Figure 12a. The control device is similar to that of shifting control. For slip control, feedback control is adopted watching converter slip speed detecting engine speed and turbine speed.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

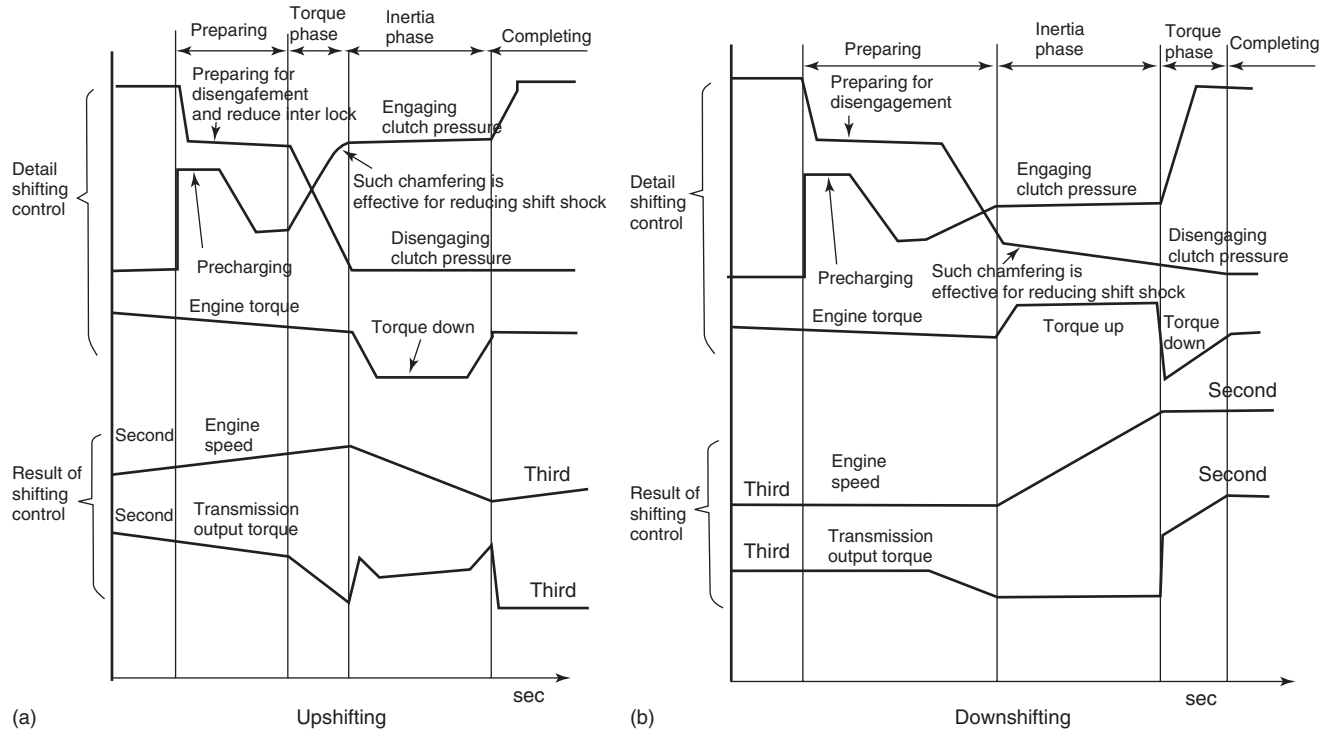


Figure 11. (a,b) A detailed timing chart for shifting control. (Reproduced by permission of Jatco, Ltd.)

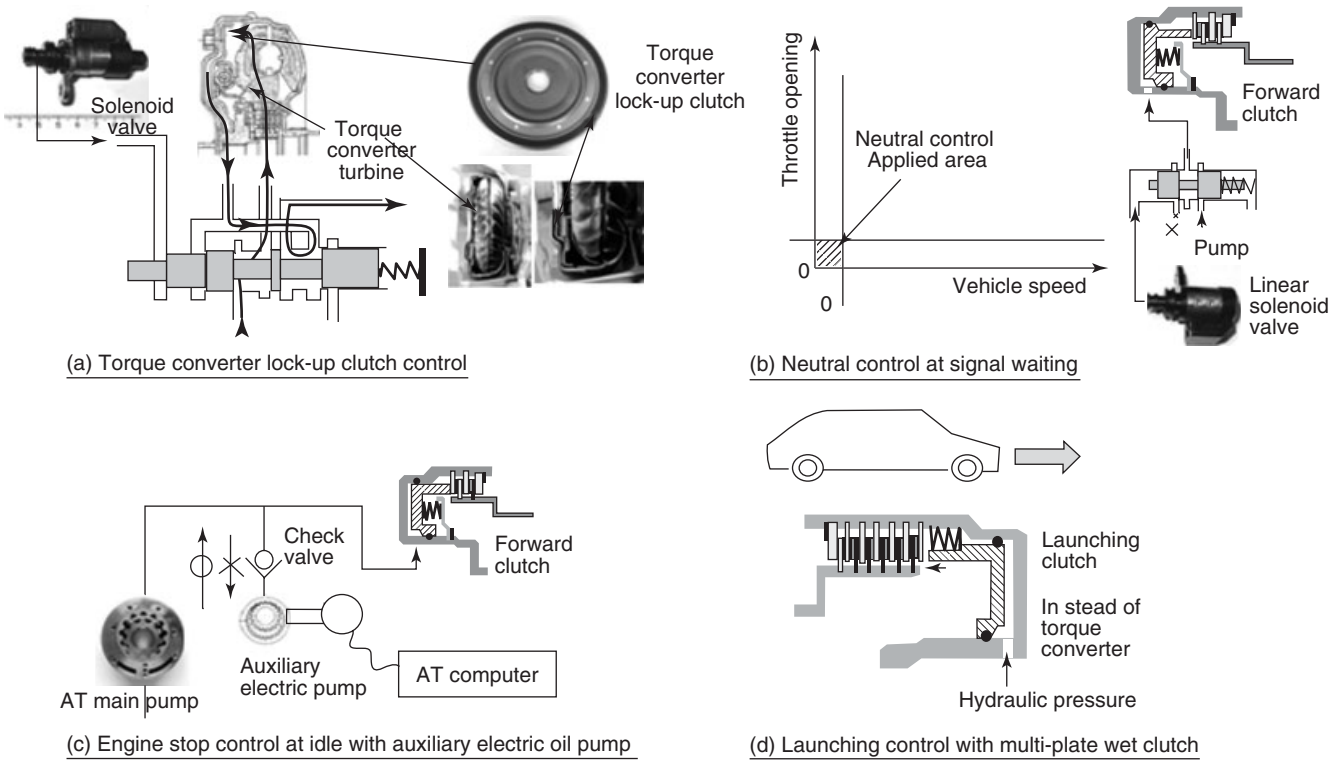


Figure 12. (a–d) Various other AT controls. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

9.2 Idle neutral and idle stop control

Recently, idle neutral control and idle stop control have become popular because they help to improve fuel economy significantly. In D range stop condition, the engine increases fuel consumption to prevent engine stall, owing to the resistance load of the torque converter imparting drag on the engine. To reduce such loss, idle neutral or idle stop control is adopted.

This control is applied in a very narrow area. Figure 12b shows applied area of idle neutral control. It is applied in the condition of vehicle almost stopping and throttle opening is almost zero. Clutch pressure for forward is released and the clutch is disengaged.

In preparation for after the traffic signal turns green, clutch pressure is set at a slight pressure level to keep the clutch piston prestroked. Subsequent pressure increase to the clutch provides drive torque capacity to enable no-delay drive-away. This is the key point of this control. The pressure increase needs to be carefully calibrated to minimize output torque disturbance, which can be noticed particularly in start/stop city traffic.

For idle stop control, an electrically driven auxiliary pump is sometimes adopted, because oil pressure becomes completely zero during engine shutoff. Therefore, to realize quick recovery of forward clutch engagement, auxiliary pump oil pressure is utilized.

Figure 12c shows such system with an auxiliary electrically driven pump. A check valve is installed into the oil circuit and this valve will open if main pump pressure goes close to zero and the auxiliary pump pressure is increased.

9.3 Launching control by multiplate wet clutch

AT used for large Displacement engine or AT for hybrid systems, the torque converter is sometimes eliminated. Instead of a torque converter, a multiplate clutch is used as a launching device as is shown in Figure 12d. Control devices and method are similar to normal shifting clutches. These launching clutches have to endure high speed slipping conditions, so heating energy into the friction clutch plates is a critical issue, especially during hill climbing or traffic jam conditions. The clutch has to provide a creeping force for sustained low speed driving up the hill. Large heat generation is caused by such long duration slipping. Thus, in this system, an effective lube cooling system for the clutch is a necessity to protect the friction clutches.

ACKNOWLEDGMENTS

The author wishes to thank Prof. Nicholas Vaughan for his very helpful advice to improve the contents of this subscription and also would like to thank all colleagues in JATCO and JATCO US.

REFERENCES

Kondo, T., Iwatsuki, K., Yutaka, T. (1990) Toyota "ECT-i" a new automatic transmission with intelligent electronic control system. SAE Paper 900550, or, Society of Automotive Engineers, 1994, Design Practices, Passenger Car Automatic Transmissions Third Edition, AE-18 (Advances in engineering, vol. 18), pp. 541–550.

Narita, Y. (1991) Improving automatic transmission shift quality by feedback control with a turbine speed sensor. SAE Paper 911938, or, Society of Automotive Engineers, 1994, Design Practices, Passenger Car Automatic Transmissions Third Edition, AE-18 (Advances in engineering, vol. 18), pp. 557–566.

Petersmann, J., Seidel, W. and Moellers, W. (1990) Porshe Carrera 2 tiiptronic transmission. SAE Technical Paper 901760, or, Society of Automotive Engineers, 1994, Design Practices, Passenger Car Automatic Transmissions Third Edition, AE-18 (Advances in engineering, vol. 18), pp. 603–616.

Shinohara, M., Shibayama, T., Ohtsuka, K. *et al*. (1989) Nissan electronically controlled four-speed automatic transmission. SAE Technical Paper 890530, or, Society of Automotive Engineers, 1994, Design Practices, Passenger Car Automatic Transmissions Third Edition, AE-18 (Advances in engineering, vol. 18), pp. 523–539.

Society of Automotive Engineers (1962) Design practices, passenger car automatic transmissions. AE-1 & AE-2 (Advances in engineering, vol. 1&2), pp. 57–80.

Society of Automotive Engineers. (1973) Design practices, passenger car automatic transmissions. AE-05 (Advances in engineering, vol. 5).

Stebar, R.F., Davison, E.D., and Linden, J.L., (1990) Determining frictional performance of automatic transmission fluids in a band clutch. SAE Technical Paper 902146, or, Society of Automotive Engineers, 1994, Design Practices, Passenger Car Automatic Transmissions Third Edition, AE-18 (Advances in engineering, vol. 18), pp. 827–840.

Wilfinger, E. and Thompson, J. (1988) Borg-Warner Australia model 85 automatic transmission. SAE Technical Paper 880480, or, Society of Automotive Engineers, 1994, Design Practices, Passenger Car Automatic Transmissions Third Edition, AE-18 (Advances in engineering, vol. 18), pp. 513–520.

Yamaguchi, H., Narita, Y., *et al*. (1993) Automatic transmission shift schedule control using fuzzy logic. SAE Technical Paper 930674, or, Society of Automotive Engineers, 1994, Design Practices, Passenger Car Automatic Transmissions Third Edition, AE-18 (Advances in engineering, vol. 18), pp. 617–627.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

FURTHER READING

Jatco (2000–2012) Jatco Technical Review, No.1, 2000, – No.11, 2012.

Society of Automotive Engineers (2012) Design Practices: Passenger Car Automatic Transmissions, Forth Edition, AE-29 (Advances in engineering, vol. 29).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto106
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Engine/Transmission Matching

Gabriele Virzi' Mariotti

Universita' di Palermo, Palermo, Italy

1 Vehicle Load Calculations	1
2 Gear Ratio Selection	15
References	33
Further Reading	34

1 VEHICLE LOAD CALCULATIONS

1.1 Introduction

A vehicle in motion with a possible trailer has to overcome forces that are named *motion resistances*. As motion can take place, the engine has to be able to generate an equal and opposite force to the resistances instant for instant.

Motion resistances vary with the vehicle kind and as a function of the road; moreover, they depend on hills or curves, in the instantaneous direction of the motion. To compare the motion resistance of several vehicles, they can be referred to the unitary weight (specific resistances r expressed in newtons per kilonewton). In the case of towing, they are distinguished as resistance to the wheel periphery and the hook resistance.

1.2 Road loads

Motion resistances are subdivided in two classes (Stagni, 1980; De Gregorio, 1970):

- accidental road loads and
- ordinary road loads.

In general, they are a function of the speed. Knowing the speed v and the weight P, and determining the single resistances, the total value of the resistance can be calculated as

$$R_{\text{tot}} = P \cdot \sum r \tag{1}$$

This value has to be equal to the traction load T generated by the engine and applied to the wheels:

$$T = R_{\text{tot}}\,[\text{N}] \tag{2}$$

The power furnished to the wheels is given by

$$N_{\text{w}} = \frac{T \cdot v}{3600}\,[\text{kW}] \tag{3}$$

with v the vehicle speed in kilometers per hour. The horse power is given by

$$N_{\text{w}} = \frac{1000}{750 \times 3600} \times T \cdot V = \frac{T \cdot v}{2700}\,[\text{hp}] \tag{4}$$

Power to weight ratio means the power of the vehicle for every kilonewton of weight.

1.2.1 Accidental road loads

1.2.1.1 Hills and towing. Indicating the angle of inclination of the road with the horizontal by α, the added load due to the hill is $P \cdot \sin\alpha$. The slope of the road is given by (Figure 1)

$$\tan\alpha = \frac{h}{l} = \frac{i}{100} \tag{5}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.

This article is © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

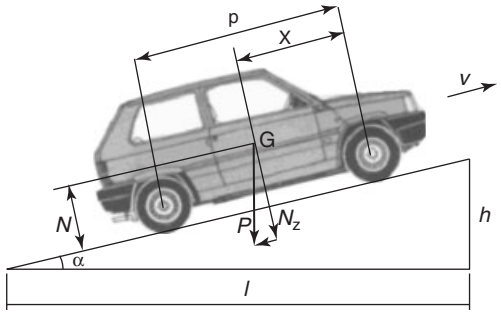


Figure 1. Loads in hill and vehicle size.

i being expressed in percent. One has

$$N_z = P \cdot \cos\alpha$$
$$T = P \cdot \sin\alpha \tag{6}$$

Because the road slope does not exceed 10–15%, $\sin\alpha \approx \tan\alpha$, so the hill resistance is

$$R_h = P \cdot \sin\alpha \cong P \cdot \tan\alpha \quad \text{or} \quad R_h = \frac{Ph}{l} \tag{7}$$

The specific resistance is

$$r_h = \frac{R_h}{P} = 1000 \cdot \tan\alpha \quad [\text{N/kN}] \approx 10i \tag{8}$$

With a steep road, the reaction of the ground becomes $N_z = P \cdot \cos\alpha$; this influences the evaluation of the rolling resistance (Section 1.2.2.1), which is also influenced by the tire stiffness. If the road is slopping down, gravity gives a favorable component to the motion and the resistance has a negative value.

A vehicle can overcome the maximum hill resistance at the adhesion limit under the following condition:

$$10i \cdot P \leq 1000f \cdot L \tag{9}$$

where f is the adhesion coefficient and L [kN] is the adhesion weight, that is, the vertical load on the driving wheels; 1000 is a factor to regulate the measurement unity. In general, the traction load $T = 1000f \cdot L$, but if L is equal to P [kN] like in a four-wheel-drive (4WD) vehicle, the traction load assumes its limiting value

$$T = 1000f \cdot P \quad [\text{N}] \tag{10}$$

The total specific resistance to motion in a hill is

$$r_{tot} = 10i + r_{ord} \tag{11}$$

where r_{ord} indicates the value of the ordinary (rolling and aerodynamic) specific resistances. Also,

$$R_{tot} = P \cdot r_{tot} = 1000f \cdot P \tag{12}$$

so that

$$r_{tot} = 1000f \tag{13}$$

That is a condition to the adhesion limit. Finally, the gradability is

$$i_{max} = \frac{1000f - r_{ord}}{10} \tag{14}$$

If $L < P$, the difference Q [kN] $= P - L$ is the weight on the supporting wheels. By putting

$$m = \frac{Q}{L} \tag{15}$$

one has

$$1000f \cdot L = (r_{ord} + 10i_{max})(L + Q) \equiv (r_{ord} + 10i_{max}) \cdot P \tag{16}$$

and

$$i_{max} = \frac{\dfrac{1000f \cdot L}{L+Q} - r_{ord}}{10} = \frac{1000f\dfrac{1}{1+m} - r_{ord}}{10} \tag{17}$$

In the case of a towing vehicle, the implicit assumption is made that r_{ord} of the driving vehicle and of the tow are equal (this is not strictly true); moreover, the ratio m is a function of the slope. In the case of 4WD, m assumes the value 0, and in towing m is equal to $L/(Q+M)$, with M [kN] the tow weight.

Indicating P_a as the weight on the front wheel and P_p as the weight on the rear wheel, z the height of the center of gravity, p the wheelbase, and x the distance of the center of gravity from the front axis, the equations of the rotation and translation equilibrium are (Figure 1) (Fessia, 1948)

$$-Pz\sin\alpha + P\cos\alpha(p - x) - P_a p = 0$$
$$P_p + P_a = P\cos\alpha \tag{18}$$

For a front-drive vehicle ($L = P_a$, $Q = P_p$), one obtains

$$m = \frac{\dfrac{x}{p} + \dfrac{z}{p}\dfrac{i}{100}}{1 - \dfrac{x}{p} - \dfrac{z}{p}\dfrac{i}{100}} \tag{19}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

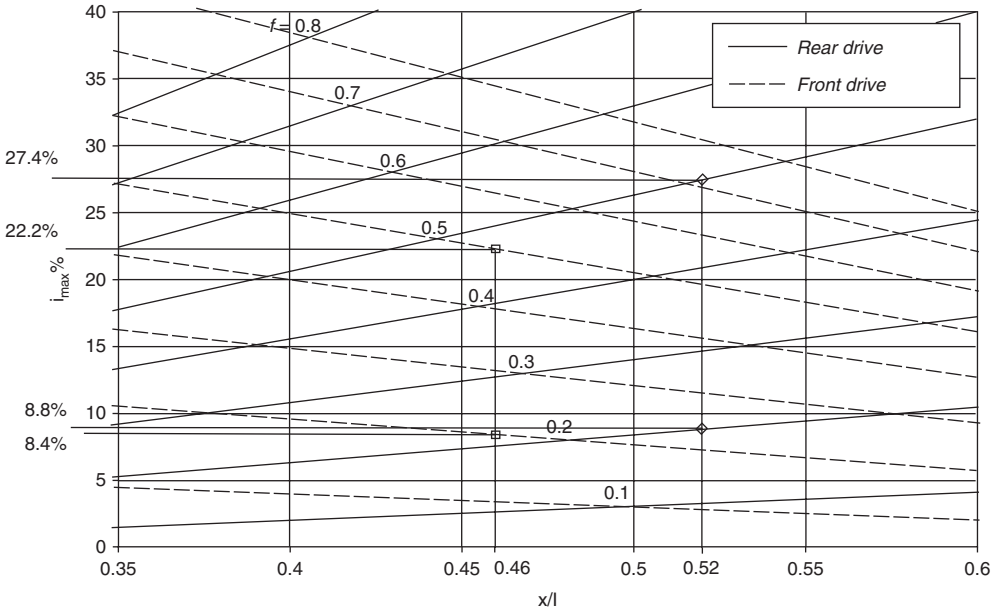


Figure 2. Graveability with ordinary resistances (20 N/kN) ($z/p = 0.25$).

For a rear-drive vehicle ($L = P_p$, $Q = P_a$), one obtains, instead

$$m = \frac{1 - \frac{x}{p} - \frac{z}{p}\frac{i}{100}}{\frac{x}{p} + \frac{z}{p}\frac{i}{100}} \tag{20}$$

Figure 2 shows the comparison of the gradability between vehicles ($y/p = 0.25$, $r_{ord} = 20$ N/kN) for several values of the adhesion coefficient. On low adhesion ($f = 0.2$), a front drive with $x/p = 0.45$ and a rear drive with $x/p = 0.52$ are equivalent; the surmountable limit hill is equal to ~8% for both vehicles. On mean adhesion ($f = 0.5$), the front drive climbs up the hill 22.2%, while the rear drive climbs up 27.4%. By varying the ordinary resistance more or less, the situation does not change much. Both the vehicles climb up the same hill in very different conditions: the front drive climbs in full understeering, at the limit of the adhesion, while the rear drive climbs in conditions of oversteering, with high utilization of the rear adhesion. Downhill has a contrary effect and the front drive is more stable.

1.2.2 Ordinary resistances

1.2.2.1 Rolling resistance. Rolling resistance consists of the resistance induced in the wheel bearings and that induced by the wheel–ground contact. The former is negligible and can be determined by the theory of bearings. The latter is due to the rolling friction between the wheel and the ground and to motion unevenness.

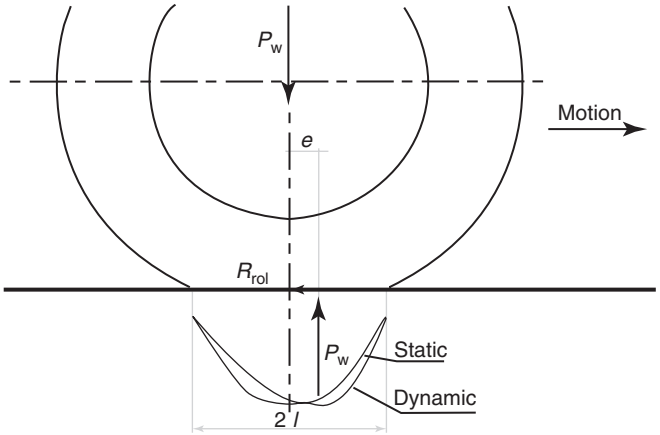


Figure 3. Pressure on wheel–ground contact.

When the wheel does not run, the diagram of the contact pressures is symmetrical: the reaction P_w is equal and opposite to the weight on the wheel (Figure 3). If the wheel is in motion, both elastic hysteresis and inertia distort the pressure diagram, so that the reaction is subjected to a displacement, indicated by e. The rolling resistance is given by

$$R_{rol} = \frac{P_w \cdot e}{R_0} \tag{21}$$

The eccentricity e (hence the specific resistance) increases with the speed v, and is proportional to the track length $2l$. R_0 is the rolling radius of the wheel. The specific

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

resistance for rolling friction increases with the weight and diminishes when the diameter increases. This resistance can be ascribed to several causes of energy loss:

1. Elastic hysteresis and mechanical inertia of the tire sidewalls subjected to deformation (contact with the ground);
2. Local sliding between the contact surfaces in the wheel track due to the slipping;
3. Motion of the compressed air inside the tire due to section reduction in correspondence to the contact.

Eccentricity cannot be determined by theory, so the resistance may be determined experimentally only. Experimental results show that starting from zero until about 50 km/h, the unitary resistance is quasi constant; the resistance increases very fast for greater values. The parameters influencing the rolling resistance are the rolling radius and the aspect ratio. In general, the increase of the radius and the decrease of the aspect ratio cause a decrease of the rolling resistance. Some empirical formulae used in the practice are as follows:

Andreau relationship:

$$r = \frac{1}{{p_0}^{0.64}}\left[20 + \frac{v^{3.7}}{1.29 \cdot 10^6 \cdot {p_0}^{1.44}}\right] \quad [\text{N/kN}] \qquad (22)$$

where
p_0 inflation pressure [bar]
v speed [km/h].

Kamm relationship: The Kamm formula presented by the Society of Automotive Engineers (SAE) is

$$r = k\left(5.1 + \frac{5.5 \times 10^5 + 90P_w}{p} + \frac{1100 + 0.0388P_w}{p}v^2\right) \times [\text{N/kN}] \qquad (23)$$

where k assumes the value 1 for conventional tires and 0.8 for radial tires; the normal force P_w is measured in newtons, the inflating pressure p in newtons per square meter or pascals, and v is the speed [m/s]. Both Equations 22 and 23 refer to rolling conditions on smooth and dry surfaces. The Andreau equation does not consider the weight on the wheel; both do not consider the size of the tire (Figure 4).

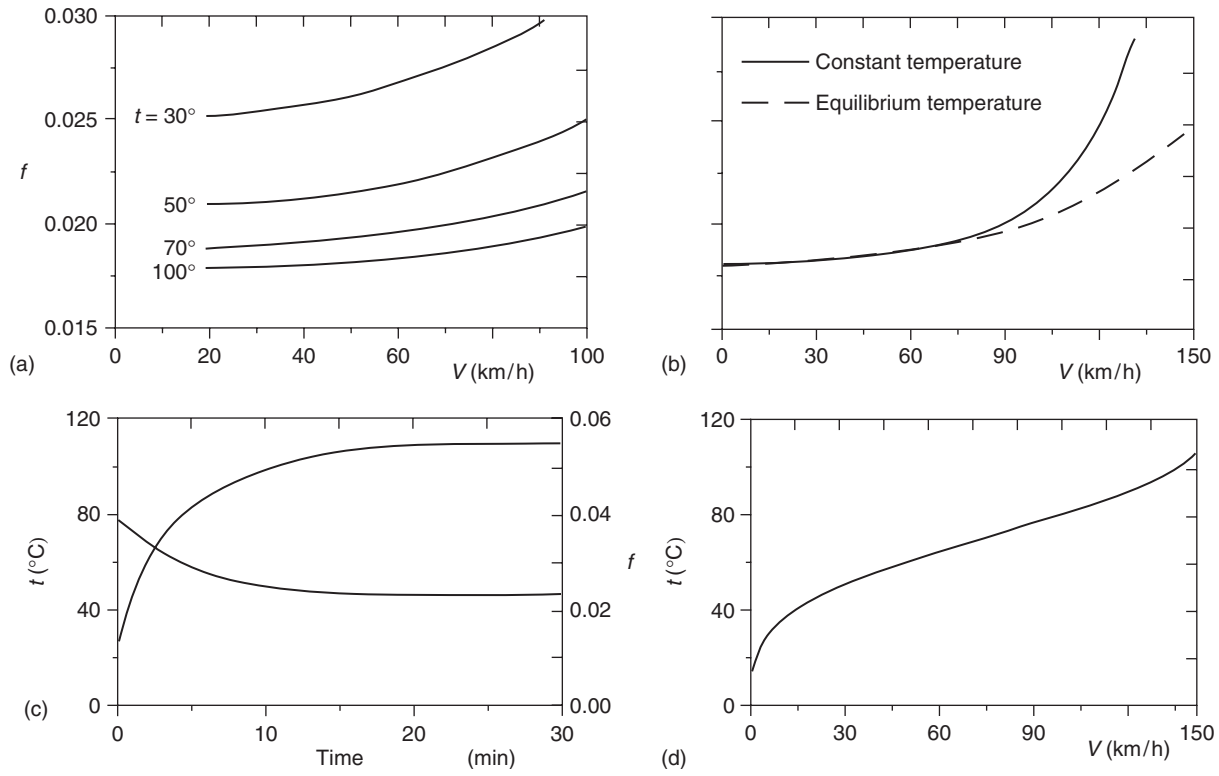


Figure 4. (a) Rolling coefficient as a function of the temperature at constant speed. (b) Comparison between the law $f(v)$ at constant temperature and at the equilibrium temperature at each speed. (c) Decrease in time of the rolling resistance and increase of the temperature in a tire rolling at 185 km/h. (d) Equilibrium temperature as a function of the speed. (From Figure 2.11, p. 48 in Genta, 2003. Reproduced by permission of World Scientific Publishing Co Pte Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Binomial Formula:

$$r = a + bv^2 \ \ [\mathrm{N/kN}] \tag{24}$$

where a [N/kN] assumes different values depending on the road surface:

$a = 8–10$ for smooth concrete
$a = 20$ for rough concrete
$a = 10–12$ for smooth asphalt
$a = 16–18$ for not much rough asphalt
$a = 22$ for very rough asphalt, and
$a = 15–25$ for sand.

b [$\mathrm{N\,s^2/kN/m^2}$] depends on the type of tire; for example, 5×10^{-3}.

v is the speed [m/s].

De Gregorio relationship:

$$r = 15 + 0.00003(v - 50)^3 \ \ [\mathrm{N/kN}] \tag{25}$$

v is expressed in kilometers per hour. For $v < 50$ km/h, the resistance is assumed as constant and is equal to 15 N/kN.

Figure 5 shows a comparison between the formulas reported earlier. All the relationships give good results until $v \leq 100$ km/h; only SAE and binomial formula are in accordance for higher speeds.

Rolling resistance is the cause of heat production. The thermal power [W] in the tire is

$$W = P \cdot w = P \cdot r \cdot \left(\frac{v}{3.6}\right) \tag{26}$$

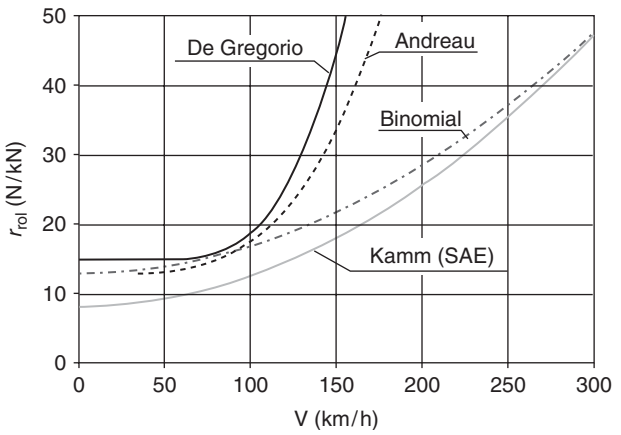


Figure 5. Comparison between the formulas for rolling resistance. SAE: $k = 1$, $P = 800$ N, $p = 2$ bar; Andreau $p_0 = 2$ bar; binomial: $a = 13$, $b = 5 \times 10^{-3}$.

where w is the unitary thermal power [W/kN] and is referred to the weight P [kN]. It is transmitted to the surroundings trough the walls of the tire. Also,

$$W = P \cdot w = K \cdot \Delta T \tag{27}$$

where K indicates the global coefficient of thermal conductivity and ΔT [°C] is the temperature difference between the tire and the surroundings (both w and K are increasing functions of the speed). The working temperature of the tire has to be limited both for preventing the risk of explosion and avoiding excessive wear.

SAE has developed test practices to measure the RRC (rolling resistance coefficient or specific resistance) of tires. These tests (SAE J1269 and SAE J2452) (Terziyski and Kennedy, 2009; Popio and Luchini, 2007) are usually performed on new tires. When measured using these standard test practices, most new passenger vehicle tires have reported RRCs ranging from 7 to 14 N/kN. In the case of bicycle tires, values from 5 to 25 N/kN are achieved. In the latter two cases, the effect of air resistance must be subtracted or the tests performed at very low speeds. ISO 28580 (2009) (International Standards Organization) is used to test rolling resistance in Europe.

1.2.2.2 Aerodynamics and wind. A body in motion in a fluid mass (air) with speed v with respect to the ground (the air is assumed motionless regard to the ground) is subjected to a resistance R_a, which is a function of

- the shape and size of the body,
- the density ρ and viscosity μ of the fluid medium, and
- the speed v.

Resistance can be determined by making use of the similitude theory (Stagni, 1980). The following relationship can be written:

$$R_a \equiv f(\mathrm{Le}, \rho, v, \mu) \tag{28}$$

where Le is the length of the vehicle. Expressing R_a in a power series, one has

$$R_a = \sum C(\mathrm{Le}^a \cdot \rho^b \cdot v^c \cdot \mu^d) \tag{29}$$

where C is a constant. Expressing the dimensional homogeneity of both terms

$$(\mathrm{kgm/s^2}) \equiv (\mathrm{m})^a (\mathrm{kg/m^3})^b (\mathrm{m/s})^c (\mathrm{kg/ms})^d \tag{30}$$

For kg one has $1 = b + d$
for m $1 = a - 3b + c - d$
and for s $2 = c + d$.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

giving

$$b = 1 - d$$
$$c = 2 - d$$
$$a = 3 - 3d - 2 + d + d + 1 = 2 - d \tag{31}$$

Then

$$R_a = \sum C \cdot (\mathrm{Le}^{2-d} \cdot \rho^{1-d} \cdot v^{2-d} \cdot \mu^d)$$
$$\equiv \sum C \cdot \rho \cdot \mathrm{Le}^2 \cdot v^2 \left(\frac{\mu}{\mathrm{Le} \cdot v \cdot \rho} \right)^d \tag{32}$$

with d being an unknown exponent. Equation 32 can be written as

$$R_a = C \cdot \rho \cdot \mathrm{Le}^2 \cdot v^2 \cdot \phi \left(\frac{\mathrm{Le} \cdot v \cdot \rho}{\mu} \right) \tag{33}$$

where φ is an unknown function of the Reynolds number (Re)

$$\mathrm{Re} = \frac{\mathrm{Le} \cdot v \cdot \rho}{\mu} \tag{34}$$

R_a can be written in the following way:

$$R_a = \frac{1}{2} C_d(\mathrm{Re}) \cdot \rho \cdot S \cdot v^2 \tag{35}$$

where S is the area of the main section of the body in a plane normal to the motion direction x, and C_d (or C_x) is the drag coefficient characterizing the external shape of the body and is a function of the Reynolds number.

Equation 35 is written considering the expression of the resistance of a disk having section S, lapping up a fluid stream having speed v, when turbulence effects downstream the disk (tail effect) are neglected (C_d (Re) = 1).

In reality, six aerodynamic coefficients can be attributed to a body in motion inside a fluid mass: three for the resistance calculation along the coordinate axes, and three for the moment calculation around the same axes (Genta, 2003; Hucho, 1998). In the case of a vehicle, the more important coefficients are the coefficient C_d along the translation direction and the coefficient along the normal axis outgoing from the ground C_z, (or C_l) which is named the *lift coefficient*; it has the effect of creating a vertical force that is added algebraically to the weight force. The coefficient along the y direction is important for sensitivity to the lateral wind. Among the three rotation coefficients, the pitching coefficient C_{My} around the y-axis is more important because it influences the trim of the vehicle; the equation to calculate the moment coefficient is analogous to Equation 35, considering a reference volume ($S \cdot \mathrm{Le}$, where Le is a generic length, for example, the wheelbase) instead of the main section. Aerodynamic coefficients can be determined numerically by computational fluid dynamics (CFD) or by wind tunnel experiments.

Wind tunnel: It consists (Figure 6) of a Venturi tube with a test room C. The honeycomb grid R has the purpose of avoiding turbulence in the tunnel. A fan placed at the end of the tube and set in action by an electric engine generates the necessary airflow. In general, the fan works in aspiration mode. The model M is placed at the center of the room C and is supported by one or more beams connected to the dynamometric balance B. Moreover, the wind tunnel is equipped with control devices permitting the measurement and the regulation of the air speed and of the characteristics influencing the viscosity and the density (temperature, pressure, and relative humidity). The capacity of the wind tunnel is determined also by the section of the test room; if this section is wider, the models can be larger and the measurements more reliable.

Scale effect: Measurement accuracy depends on the scale model. In the immediate surroundings of any body, the fluid flow is laminar rather than turbulent. If the model is of a very small scale, the small protuberances will not be reproduced faithfully, because they are not contained in the laminar layer, unlike a model in scale 1:1. An analogous effect gives the walls roughness, which is greater in the model for working reasons. Vehicle 1 and the model 2 have

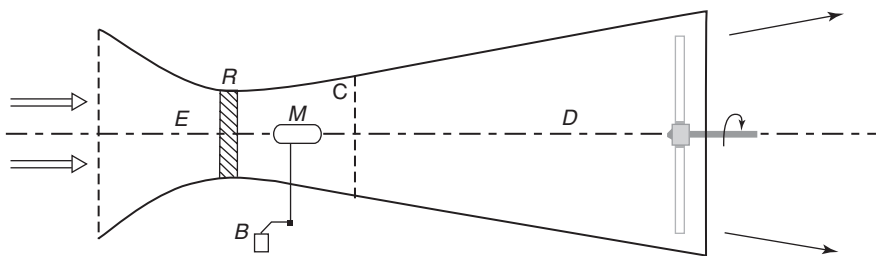


Figure 6. Scheme of wind tunnel.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the same drag coefficient if

$$\rho_1 v_1 \frac{l_1}{\mu_1} = \rho_2 v_2 \frac{l_2}{\mu_2} \tag{36}$$

If the environmental conditions outside and inside the tunnel are equal, from Equation 36 one has

$$v_1 = v_2 \frac{l_2}{l_1} \tag{37}$$

Determination of the drag coefficient is convenient at speed equal to $\frac{2}{3}v_{\max}$ and no more than 100 km/h; it is the speed suggested by standards (Comitato Unificazione Nazionale Autoveicoli) CUNA NC 003–01 (2004) and CUNA NC 003–02 (2004) (today they are European standards) for the determination of the fuel consumption on the road. If the test room is wide enough, the test on motorcycles and vehicles can be executed in 1 : 1 scale. In the case of other vehicles, or if the test room is not wide enough, the test needs a model in reduced scale, with the disadvantage that it is impossible to obtain the speed in the tunnel very close to the speed of sound, c. This situation does not hold the validity of Equation 28 because the aerodynamic coefficients are functions also of the Mach number (Ma):

$$\mathrm{Ma} = \frac{v}{c} \tag{38}$$

This is a disadvantage because the maintenance of the same Re and Ma is not possible in the case of test on small scale models. The speed in the tunnel can be reduced by making use of the fluid at a greater density, maintaining Re as constant, and following Equation 36. Studies conducted recently (D'Anca, Mancuso, and Virzì Mariotti, 2005) show that the drag coefficient varies very little with the speed for a road vehicle, at least starting from a certain value of the speed; this helps the solution of the problem. CFD tests are executed on a Maserati Biturbo model; only a half the vehicle is analyzed because of the longitudinal symmetry of the model and the symmetry of the motion field. The model is a flat-bottomed vehicle, has no rear view mirrors, has smooth rim covers, and does not present prominences because of handles, lights, and so on. CFD analysis gives the result $C_d = 0.472$. Comparison with the value obtained by the White method (Section 1.2.2.2), namely $C_d = 0.4735$, is rather satisfactory. On the same model, another analysis was carried out to confirm the validity of C_d value at $\frac{2}{3}v_{\max}$ by determining the variation versus the vehicle speed. The results are reported in (D'Anca, Mancuso, Virzi' Mariotti, 2005); they show only a qualitative validity and cannot be generalized because aerodynamic coefficients are functions of the shape. Moreover, the results are obtained by two-dimensional simulations because the three-dimensional simulation requires a computational effort that is beyond our ability at this time. For the purpose of forecasting the vehicle performance on the road, the aerodynamic coefficients are considered as constants by neglecting their variation with the Reynolds number and with the speed in particular. This established a negligible cause of error at low speed.

Ground effect (Cogotti, 1998; Stagni, 1980; Katz, 2006): It is assumed that airflow has a relative speed v with respect to the ground, which is contrary to the reality. This is another cause of inaccuracy of the wind tunnel tests. The fluid speed diagram between the ground and vehicle has a semi-parabolic trend, while the trend is parabolic in the model (Figure 7). This changes the pressure distribution.

A better way to eliminate or reduce ground effect is the use of a double-mirror-like model. The traveling band model is less valid but also efficient, but the solutions are very expensive and require a very wide wind tunnel.

Aerodynamic resistance of a vehicle can be considered as the sum of factors such as

- the head effect (wind pressures exerted on the front),
- the tail effect (wind pressures exerted on the rear),
- the wall effect (resistance diffused along the lateral walls, roughness, concentrated resistances).

Experience shows that both the ratio of the total length of the vehicle to the mean diameter of the main section and the head shape influence head and tail effects. The regularity of the lateral surfaces, in particular the lower surface, influences the wall effect.

The tail effect can be negative or positive. It is due to the presence of a vortex in the rear of the vehicle depending on the shape of the rear bonnet. In general, the vortex induces a negative pressure opposing a resultant force to the motion, increasing the drag coefficient. If the rear of the vehicle is finely shaped, the pressure can induce a favorable force to the motion, reducing the drag coefficient. Pressure distribution on the Maserati Biturbo model and the streamlines of the airflow are shown in (D'Anca, Mancuso, Virzi' Mariotti, 2005).

Style variations: They have great importance on the value of the aerodynamic coefficients. Effect of some particular variation of the bodywork are shown in (D'Anca, Mancuso, Virzi' Mariotti, 2005). Also in this case the vehicle is a Maserati Biturbo. The results are useful for manual optimization of the bodywork. However, there are other parts that influence the aerodynamic coefficient, such as wheels and cavities and tank. Another example is the manual optimization of the geometry of FIAT Bravo

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

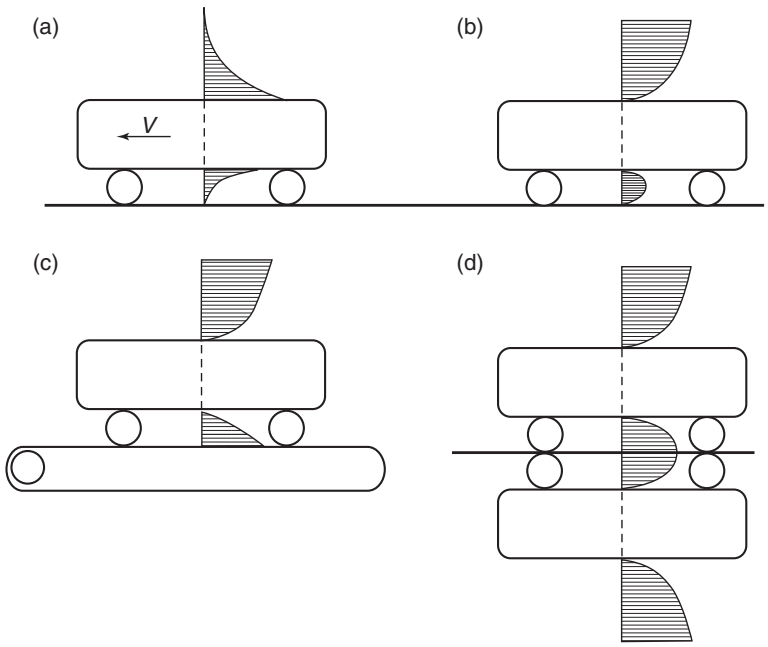


Figure 7. Ground effect and equipment for its reduction. (a) Reality; (b) fixed ground model; (c) traveling band model; and (d) mirror-like model.

Blueprint (Terranova, 2009; Milone, 2011). The purpose is the reduction of the motion resistance of the vehicle; then the drag and lift coefficients are determined.

Three different configurations for the front bumper are examined, eliminating the zones where the flow has quasi-zero speed and is stagnant.

First configuration: Modification of the inclination angle from 14.24° to 13.30° (Figure 8a). Drag is reduced by 1.7% and lift is increased by 28%.

Second configuration: Following the condition of optimum nose (Janssen and Hucho, 1975) (Figure 8b), drag is reduced by 2.7% and lift is reduced by 4.6%.

Third configuration: Two configurations (Buchheim, Deutenbach, and Luckoff, 1981) in Figure 8c, varying the height of the point stagnation, give the following results:

- *Configuration A*: height $z_s = 221\,\text{mm}$; $C_d = 0.33$; $C_l = 0.13$,
- *Configuration B*: height $z_s = 149\,\text{mm}$; $C_d = 0.33$; $C_l = 0.06$.

CFD analysis on configuration B shows that the zone at low speed in the front end (Figure 9) reduces its extension and the flow is not stagnant. The reduction of the drag coefficient is 3.6%, while lift reduction is 66%.

Study on a small rear spoiler over the rear window shows that it can give lower aerodynamic resistance and a better lift effect. Analyses are executed on a FIAT Bravo model by varying the length of the wing contour and the incidence angle with respect to the horizontal. Figure 10 shows the several lengths for the incidence angle -20°, and Figure 11 shows the results for the length 75 mm. The drag coefficient is optimum for a length 75 mm and incidence angle -25°. The lift coefficient reaches 0.33, while the corresponding apparent reduction of the weight is equal to 12.3% at the maximum speed.

CFD analyses are executed on the complete model, obtaining a drag coefficient reduction of 7% and a lift coefficient increase of 60%. Figure 12 shows the trend of the streamlines in the modified configuration. Instead, a negative C_l value induces an increase of the apparent weight with an increase of the force on the tire–ground contact (greater ability of acceleration and greater stability), but with an increase of the rolling resistance and an increase of the drag coefficient.

Portion of aerodynamic drag that is linked to the generation of lift is named the *induced drag* (Genta, 2003; Dabbene, 2009). In the case of vehicles with a high power to weight ratio, the use of negative aerodynamic lift allows the transfer of all the power on the wheel–ground contact. Aerodynamic pitching moment causes strong variation of the forces exerted by the wheels on the road (Figure 13a). The pitching moment is considered positive if it increases the load on the front wheels. The lift of the wing is due to a difference of pressure, which induces a lift force (Figure 13b). If the wing has an infinite length, there is no induced

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

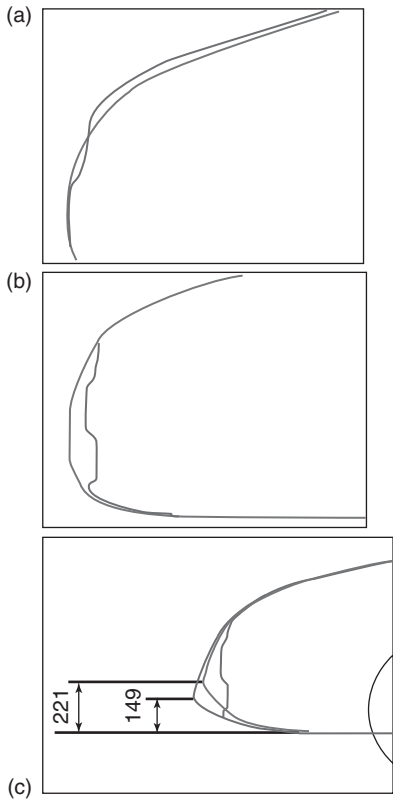


Figure 8. Modification of the high part of a FIAT Bravo front bumper (a), optimum nose condition (b), and stagnation point position (c).

drag, but in the case of finite length the vortex at the tips is deflected backwards (Figure 13c). An unfavorable situation is when the breakaway of the fluid is total on the entire superior surface: one has the stall, and the lift effect is reduced totally (Figure 13d). The study on the wing profile was executed in the past generating national advisory committee for aeronautics (NACA) profiles, which are optimized wing profiles to choose in accordance with the requested performances. In the automotive field, they are used for the design of spoilers with the aim to negate the lift coefficient (Schenkel, 1977; Katz, 2006). The minimum drag coefficient is due to the drop shape. It has to be hit by the fluid along its symmetry axis and thus it is ∼0.05 if there are no other bodies in the neighborhood. The ground has a negative effect given that drag coefficient increases by ∼150% because of the fact that the streamlines have to remain parallel to the ground. If the vehicle is supposed to be in motion in contact with the ground, the optimum shape is the semidrop (or semilune). However, the distance of the vehicle from the ground cannot vanish, and this solution gives high values of the drag coefficient.

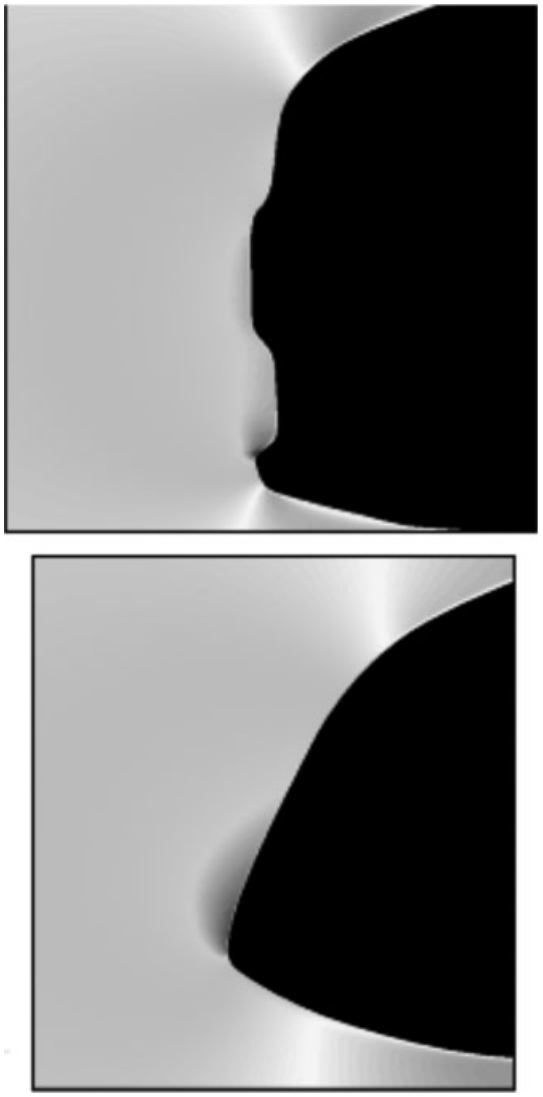


Figure 9. Comparison between the speeds (head effect) in the original model FIAT Bravo and configuration B (top).

If the vehicle is in a symmetrical position with regard to the flow, with roll and aerodynamic angles equal to zero, the side force, the rolling moment, and the yawing moment are equal to zero. In general, the coefficient C_y is a function of the roll and sideslip angles, while it can also be a function of the steering angle in race vehicles. Moreover, it is a function of the width to height ratio. For small variations of the above parameters, the aerodynamic coefficients C_y, C_{Mx}, and C_{Mz} can be approximated by a linear function. For a preliminary evaluation, the following empirical relationship is suggested:

$$C_y = \beta \frac{\text{lateral area}}{\text{frontal area}} (0.005 + 0.0019 n_f) \qquad (39)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

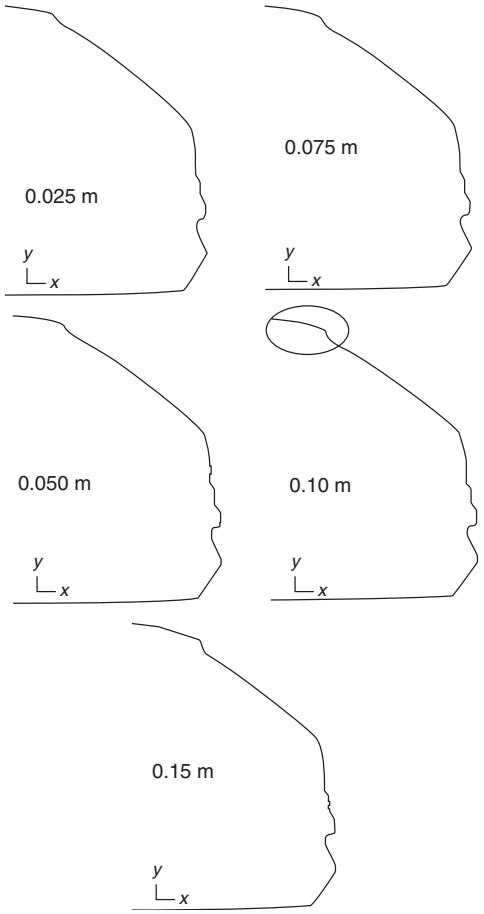


Figure 10. Configuration spoiler -20° (FIAT Bravo).

where β is the aerodynamic sideslip angle (the angle that the speed vector makes with the longitudinal axis of the vehicle) and n_f is a numerical factor that can be obtained experimentally or from data on similar vehicles.

White's method: The method proposed by White (1967–1969) is useful for the evaluation of the drag coefficient C_d. The vehicle is divided in nine parts, with as many different configurations, as shown in Table 1. A weight x_i is assigned to every part and is introduced in the empirical relationship

$$C_d = a_0 + \sum_{i=1}^{9} x_i \cdot a_i = a_0 + a_i \sum_{i=1}^{9} x_i$$
$$= 0.16 + 0.0095 \sum_{i=1}^{9} x_i \quad (40)$$

The coefficients a_0 and a_i are obtained by statistical analysis of the results in wind tunnel tests. Formula (40) gives C_d with an approximation of 7%. The nine parts are

- the plant contour
- the outline
- the crosswise section
- the windscreen plant
- the windscreen–roof intersection
- the plant
- the rear interior compartment–tail intersection
- the rear part
- the floor.

Today, the White method is replaced by modern CFD methods. All the same, it is useful for the validation of experimental or computational results.

1.3 Acceleration (inertia)

When the vehicle is subjected to a speed variation (positive or negative acceleration), it is subjected to an additional resistance because of inertia, given by

$$R_{acc} = \frac{a \cdot P}{g} \quad (41)$$

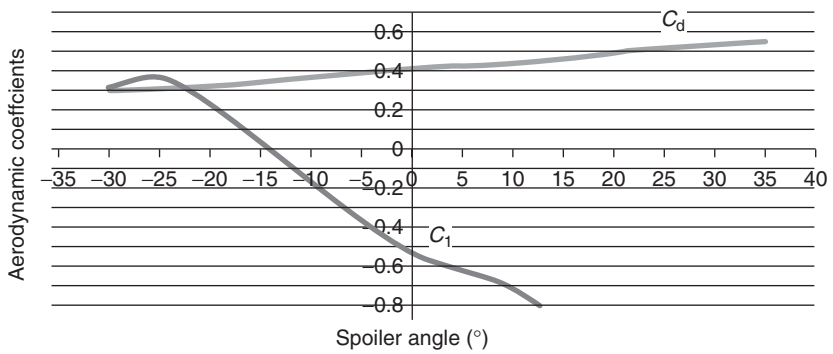


Figure 11. Trend of aerodynamic coefficients versus the spoiler angle; spoiler length 75 mm (FIAT Bravo).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

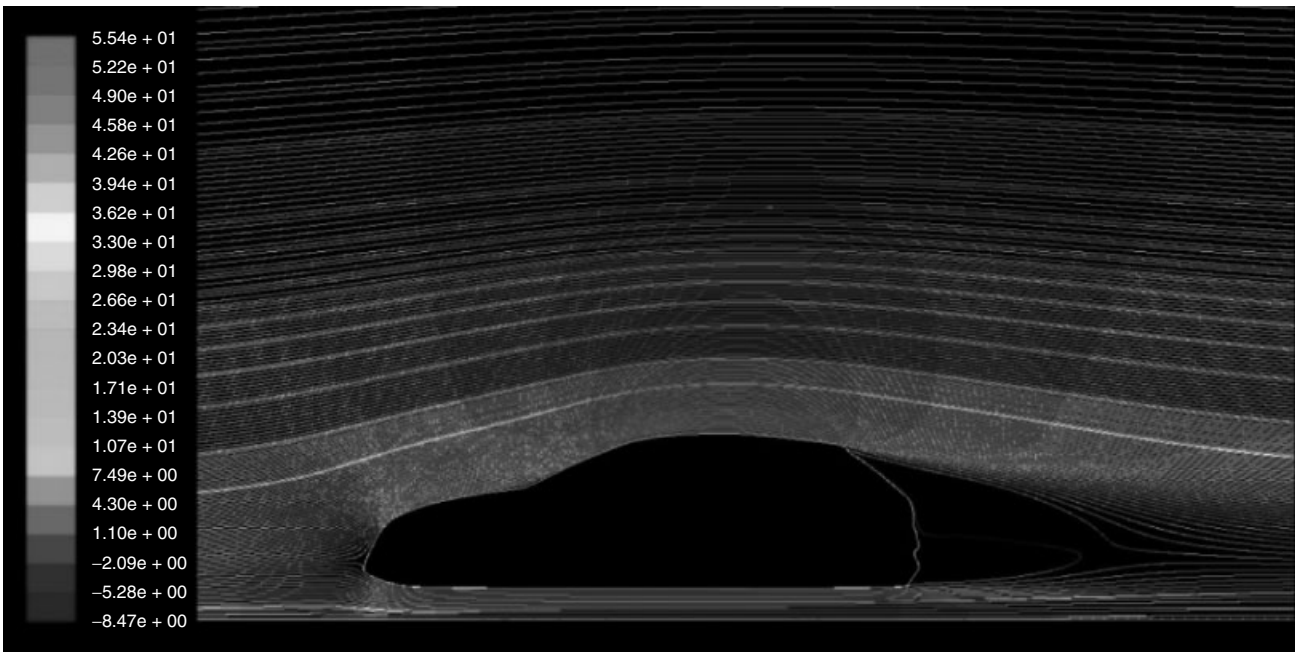


Figure 12. Streamlines *X*-velocity [m/s] (FIAT Bravo).

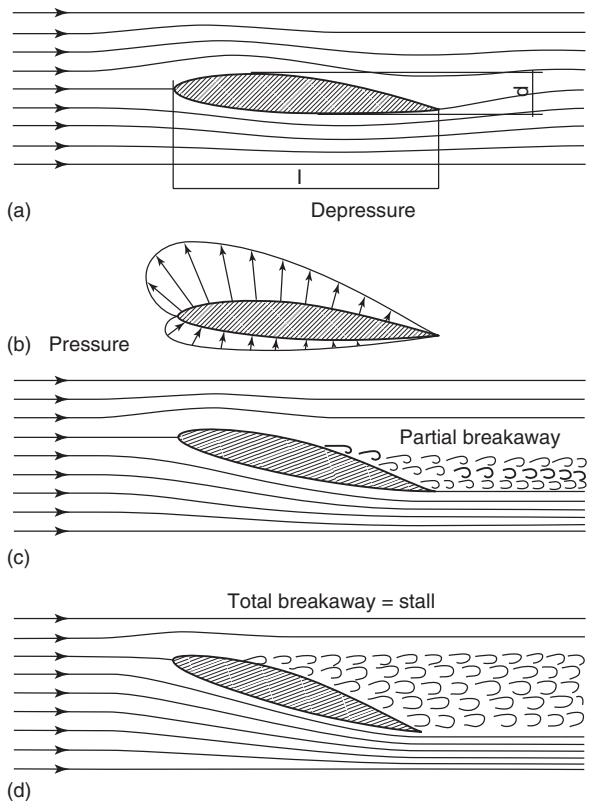


Figure 13. Vorticity in a wing.

where a is the vehicle acceleration, P is the vehicle weight, and g is the gravity. Dividing by P, the specific resistance is obtained as

$$r_{\text{acc}} = 1000 \cdot \frac{a}{g} \cong 102 \cdot a \quad [\text{N/kN}] \tag{42}$$

Considering the inertia of all the rotating mass (gears, shafts, wheels, etc.), the total kinetic energy of the vehicle is the sum of the kinetic energy of motion E_c of the entire vehicle and the kinetic energy of rotation of all the rotating masses, that is,

$$E_c = \frac{1}{2}\frac{P}{g}v^2 = \frac{1}{2}\frac{P}{g}(\omega_r R_0)^2 = \frac{1}{2}J_v\omega_r^2 \tag{43}$$

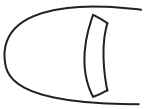
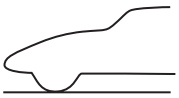
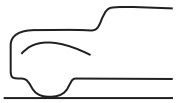
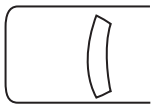
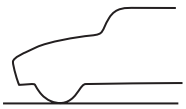
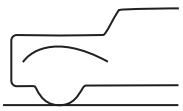
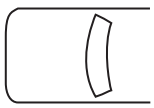
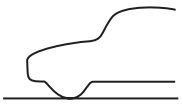
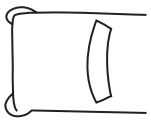
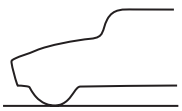
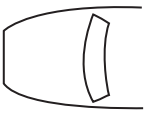
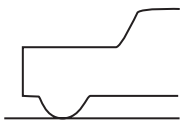
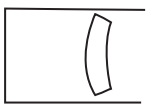
$$\frac{1}{2}J_v'\,\omega_r^2 = \frac{1}{2}J_m\omega_m^2 + \frac{1}{2}J_t\omega_t^2 + 4\frac{1}{2}J_r\omega_r^2 + \frac{1}{2}J_v\omega_r^2 \tag{44}$$

where J_v' is the equivalent inertia of the entire vehicle, J_v is its equivalent inertia without the rotating masses, J_m is the moment of inertia of the engine, J_r is the moment of inertia of one wheel, and J_t is the moment of inertia of the transmission shaft; ω_r is the angular speed [rad/s] of the wheels, ω_m that of the engine, and ω_t that of the transmission; R_0 [m] is the rolling radius of the wheels; and the ratio P/g is the vehicle mass [kg]. One has

$$\omega_m = \omega_r/(\tau_1\tau_p) \qquad \omega_t = \omega_r/\tau_p \tag{45}$$

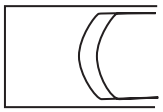
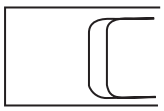
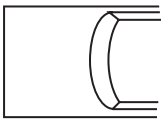
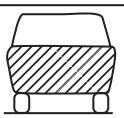
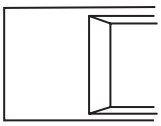
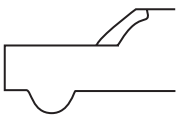
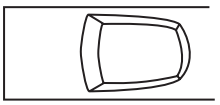
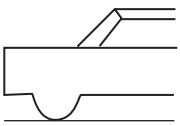
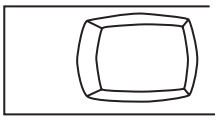
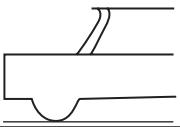
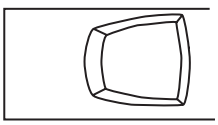
Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. White method.

Front part		
x_i	1 Plant contour	2 Outline
1	About semicircular 	(a) Rounded low grille  (b) High, rounded, tapered bonnet 
2	Frontal and sidewalls well filleted 	(a) Low, but squared grille  (b) High, tapered, unrounded bonnet 
3	Frontal and sidewalls filleted without protuberances 	(a) Few high, tapered grille 
4	Frontal and sidewalls filleted with protuberances (*) 	(a) A few high, squared grille  (b) High, rounded grille with horizontal bonnet 
5	Squared frontal with tapered sidewalls 	High, squared grille with horizontal bonnet 
6	Squared frontal, untapered sidewalls 	
	(*) Door mirrors: considered like protuberance, x as the weight $x_i = 4$. Otherwise add 1 to the assigned weight	Adding 3 for divided fenders 4 for distant fenders from the frontal 2 for retractable headlights 4 for small projecting headlights 7 for great projecting headlights

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

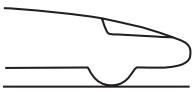
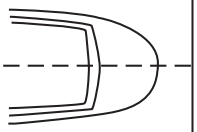
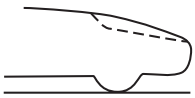
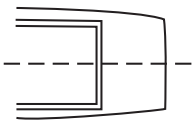
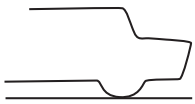
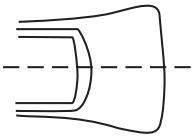
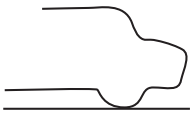
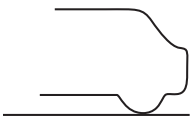
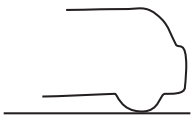
Table 1. (*Continued*).

Windscreen–interior compartment intersection		
x_i	3 - Crosswise section	4 - Windscreen plant
1	Smooth bonnet and fender, well filleted with the sidewalls 	large fillet at sidewalls (practically semicircular plant) 
2	High, bonnet, low fenders 	Lateral fillets 
3	(a) Bonnet at the same level of fenders, with upper fillet  (b) High bonnet with fenders filleted in the top 	Convex 
4	Bonnet at the same level of angular fenders 	Flat 
5	Lower bonnet than the angular fenders 	
	Adding: 1 for vertical windscreen 1 for prominent profilings or moldings	
Interior compartment		
x_i	5 - Roof – windscreen intersection	6 - Plant
1	Rounded 	Tapered towards the rear part 
2	Squared, also with moldings 	Tapered in front and behind or with about constant width 
3	With prominences in front 	Tapered in front 

(*continued overleaf*)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. (*Continued*).

Tail and floor			
x_i	Rear interior compartment–tail	8 Rear plant	9 - Floor (lower part)
1	Fast-back 	Well tapered 	Completely smooth or with small mechanical projections
2	Semifast-back (discontinuinty in the tail) 	Small or lacking tapering 	Intermediate
3	Squared interior compartment, wit squared rear bonnet 	Negatively tapered 	Not smooth, with structural and mechanical projections
4	(a) Rounded interior compartment and rounded rear bonnet  (b) Squared interior compartment and small or lacking rear bonnet 		Intermediate
5	Rounded interior compartment and small or lacking rear bonnet 		With projecting chassis
	Adding: 3 for longitudinal fins until the rear bonnet 2 for splitted fenders Note: All the x_i of this column are referred to rear bonnet rounded sideways		Note: Applying intermediate x_i for intermediate situations

(Reproduced from Morelli, 1970. © A. Morelli.)

In the above, τ_1 is the ratio of the first gear and τ_p is the axle ratio. Substituting into Equation 44, one has

$$J'_v = \frac{J_m}{(\tau_1 \tau_p)^2} + \frac{J_t}{\tau_p^2} + 4J_r + \frac{P}{g}R_0^2 \qquad (46)$$

The inertia coefficient in the first gear is

$$k_I = \frac{J'_v}{J_v} = \frac{gJ'_v}{PR_0^2} \qquad (47)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The inertia coefficients of the other gears are obtained by substituting τ_1 with the considered gear ratio in Equation 46, considering the variation of the moment of inertia as a function of the pair of gears in contact. The inertia coefficients are slightly greater than 1. The equivalent inertia of the vehicle in neutral position is

$$J_v^n = \frac{J_t}{\tau_p^2} + 4J_r + \frac{P}{g}R_0^2 \tag{48}$$

and the inertia coefficient is

$$k_n = \frac{gJ_v^n}{PR_0^2} \tag{49}$$

The quantities J_m and J_t have to be calculated considering the inertia of the rotating parts of all the linked masses.

The maximum available acceleration of the vehicle depends on the power to weight ratio. In general, the acceleration is higher at start, and diminishes with the increase of speed, both because the motion resistances increase with the speed and because the traction is reduced. Introducing the inertia coefficient, Equation 42 becomes

$$r_{acc} = k_m 1000 \cdot \frac{a}{g} = k_m 102 \cdot a \quad [\text{N/kN}] \tag{50}$$

The inertia coefficient k_m changes with the inserted gear.

2 GEAR RATIO SELECTION

2.1 Introduction

An engine giving constant power versus the rotation (i.e., electric one) is optimum for the traction, assuming that the kinematic connection between the engine shaft and wheels is fixed. When the continuous use of the maximum power of the engine is not possible, a gear drive system is necessary, and consequently both the running cost of the vehicle and the energy consumption increase. The *power characteristic* is the curve relating the power to the rotation number [rpm] (De Gregorio, 1970; Bencini, 1956). The power N is the product of the angular speed and the engine torque C [Nm]:

$$N = \frac{\omega \cdot C}{1000} = \frac{2 \cdot \pi \cdot n}{60} \cdot \frac{C}{1000} \quad [\text{kW}] \tag{51}$$

Here, ω is the engine angular speed [rad/s] and n [rpm] the rotation number (Figure 14).

An engine having a constant characteristic of the torque is defined *rigid* (so that the power has a linear trend); in contrast, the engine at constant power is defined *elastic* (so that the engine torque has a hyperbolic trend). In

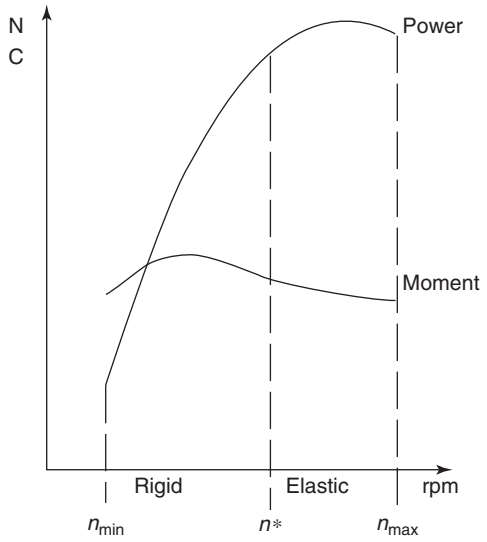


Figure 14. Characteristic of an internal combustion engine.

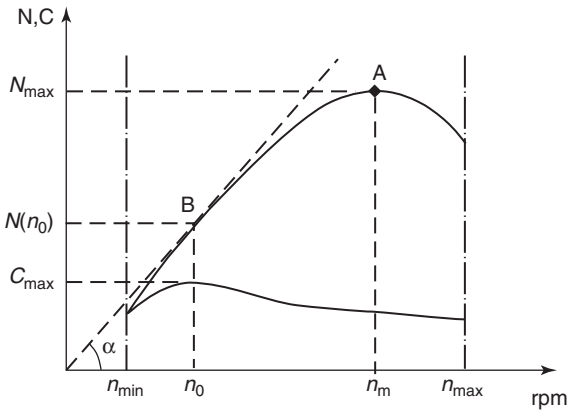


Figure 15. Maximum torque rate.

general, internal combustion engines have small elasticity because the developed power increases proportionally to the rotation number. The engine is rigid in the zone between n_{min} and n^* and is elastic between n^* and n_{max} because the characteristic is flat at the maximum power rate. The characteristic curves of Otto and Diesel engines have very similar trends (Figure 15).

2.2 Power characteristic

If the engine characteristic is not flat, the vehicle requires a gear drive in the rigid zone to increase the traction at low speed. The gear ratio has to be very low at the starting point. Moreover, the choice of a certain number of gears permits the correction of rigidity, so that the curve of the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

engine torque is very close to an equilateral hyperbola. An internal combustion engine presents the maximum torque at a certain intermediate rate n_0 (Figure 14). For this value, the engine torque curve has a horizontal tangent, while the power curve has a tangent through the origin. In fact,

$$\left(\frac{dC}{dn}\right)_{n=n_0} = 0 \tag{52}$$

Because $N = C \cdot n$, one has

$$\frac{dN}{dn} = n \cdot \frac{dC}{dn} + C \tag{53}$$

and

$$\left(\frac{dN}{dn}\right)_{n=n_0} = C(n_0) = \frac{N(n_0)}{n_0} = \tan\,\alpha \tag{54}$$

In correspondence of the rotation rate n_0, the excess of power is maximum, and then the maximum acceleration can be obtained. The gear change is convenient near n_0 for consumption reduction.

At rotation lower than n_{min}, the combustion is uneven or missing and the power is not sufficient to balance the internal friction of the engine. The characteristic of power presents a certain concavity toward the bottom because of the fact that the several efficacies of the engine are not constant. The maximum rotation number n_{max} is slightly higher than at the maximum power rate. It allows the implementation of a greater speed than that at maximum power rate, when possible (i.e., on light downhill). In general, the maximum rotation number and the maximum power rate can be evaluated by thermo-fluid dynamical analysis and by reducing the masses in the motion of the engine.

Engine torque and power curves are obtained experimentally by the *engine bench*. The engine works at full introduction, and engine torque and power are measured by a dynamometric brake varying the number of rotations. The curves are obtained by interpolating the obtained measurements. The consumption tests are executed by the same bench. Following the international rules, power curve obtained by the bench does not consider the absorption of numerous auxiliaries, such as the fan, pump of the coolant, and generator; they absorb 7–10% of the developed power.

In general, the power characteristic of an automotive engine can be represented by a cubic parabola:

$$N = an^3 + bn^2 + cn + d \tag{55}$$

The four constants a, b, c, and d can be determined by the data of maximum power and maximum engine torque. Both high power engines and motorcycle engines are exceptions. They need a more accurate analytical definition. When possible, the aspect of the curve may be improved by increasing the parabola order and adding other points with known values of torque or power.

Small losses of power in the gear drive and in the axle are considered by introducing the efficacy of the transmission, given by the product of the single efficacy of the parts in the kinematic chain, that is,

$$\eta_t = \eta_1 \cdot \eta_2 \cdot \cdots \cdot \eta_n \tag{56}$$

Indicative values are reported in Table 2. The efficacy of the transmission is greater if the vehicle rides in direct gear. Just as an indication, in direct gear $\eta_t = 0.93$; in reduced gears $\eta_t = 0.87$; and if the gear drive has no step gear, the efficacy increases and $\eta_t = \sim 0.91$.

The measured power has to be corrected as a function of the conditions of pressure and temperature. The correction has to be executed using relationships indicated by the standardization institutions. The more important ones are SAE method (SAE J1349, 2004)

$$N_{cor} = N \cdot \left[1.176 \cdot \left(\frac{990}{p_d} \cdot \sqrt{\frac{273+t}{298}}\right) - 0.176\right] \tag{57}$$

ISO method (ISO 1585, 1992)

$$N_{cor} = N \cdot \left(\frac{990}{p_d}\right)^{1.2} \left(\frac{273+t}{298}\right)^{0.6} \tag{58}$$

where N and N_{cor} are the measured and the corrected power, respectively, p_d is the pressure of dry air [hPa], which is found by subtracting the vapor pressure p_v from the actual air pressure; and t is the air temperature [°C]. Equation 57 is

Table 2. Efficacy of several parts of mechanical transmission (also for vehicles with two or three wheels).

	Type	Efficacy
Gears	Straight toothed	0.98
	Helical toothed	0.97
	Conic toothed	0.96
	Hypoid toothed (efficacy diminishes if the eccentricity increases)	0.90
Chain	Roller	0.95
	Silent	0.98
Belt	Timing	0.95
	V-belt	0.94
Joint or hydraulic converter	Unlocked hydraulic joint	0.92
	Unlocked hydraulic converter	0.92
Joint	Cardan or CV	0.98

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

valid in the case of spark ignition for temperatures between 15 and 35°C and dry air pressure between 900 and 1050 hPa (90–105 kPa). Instead, Equation 58 is valid only if the ratio $0.93 \leq N_{cor}/N \leq 1.07$ with a reference temperature of 25°C and a pressure 99 kPa. In the case of compression ignition, it becomes (Commission Directive 32002L0041, 2002; Commission Directive 31988L0195, 1988)

$$N_{cor} = N \cdot f_m \left(\frac{990}{p_d}\right)\left(\frac{273+t}{298}\right)^{0.7} \tag{59}$$

And, in presence of a turbo supercharger

$$N_{cor} = N \cdot f_m \left(\frac{990}{p_d}\right)^{0.7}\left(\frac{273+t}{298}\right)^{1.5} \tag{60}$$

The coefficient f_m is given by

$$f_m = 0.036\frac{q}{r} - 1.14 \tag{61}$$

where q is the correct flux of the fuel [mg/L/cycle] for liter of total volume, and r is the ratio between the pressures of the output and input of the supercharger ($r > 1$). Equation 60 is valid if $40\,\text{mg/L/cycle} < q/r < 65\,\text{mg/L/cycle}$. If the ratio is lower, f_m assumes the value 0.3; and if it is greater, f_m assumes the value of 1.2. The test is valid if $0.9 \leq N_{cor}/N \leq 1.1$. The above relationships replace the earlier CUNA, DIN (Deutsches Institut fur Normung), and other obsolete European rules.

2.3 Road loads versus speed

2.3.1 Diagram T–v

The performance of the vehicle can be foreseen by knowing the characteristic of power and engine torque (Heisler, 2004; De Gregorio, 1970; Bencini, 1956). To do it, one has to know

- the gear ratios $\tau_1, \tau_2, \ldots, \tau_v$,
- the axle ratio τ_p, and
- the rolling radius R_0 of the tires.

The axle ratio is established with the appropriate design criteria on the basis of the performances of the engine power. τ_p is implemented by

$$\tau_p = \frac{n_p}{n_c} \tag{62}$$

where n_p is the number of pinion teeth at the end of the transmission shaft and n_c is the number of gear teeth at the differential box ($\tau_p < 1$). Traction T is given by

$$T = \frac{1000N}{\frac{2\cdot\pi\cdot n}{60}\cdot R_0} \cdot \frac{1}{\tau_n \tau_p} \cdot \eta_t \cdot [\text{N}] \tag{63}$$

where N is expressed in kilowatts, R_0 is the rolling radius [m], η_t is the transmission efficacy, τ_n is the ratio of the inserted gear, and n [rpm] is the engine's angular speed. The traction characteristic $T = T(v)$ can be obtained from the characteristic $N(n)$ by Equation 63, changing the abscissa, as

$$v = \frac{2\pi n}{60}\tau_n \tau_p R_0 \times 3.6 \tag{64}$$

with v [km/h] and a 3.6 conversion factor. By making the hypothesis that there are four gears, indicating τ_3, τ_2, and τ_1 the ratios of the third, second, and first gear, respectively, and assuming the ratio of the fourth gear to be equal to 1 (it is an ancient solution), the traction characteristic in third, second, and first gear are obtained easily (Figure 16). P_3, the corresponding point in third of P_4, has abscissa

$$X_3 = X_4 \cdot \tau_3 \tag{65}$$

and ordinate

$$Y_3 = \frac{Y_4}{\tau_3} \cdot \frac{\eta_3}{\eta_4} \tag{66}$$

where η_3 and η_4 are the transmission efficacy in the third and fourth gears, respectively. Figure 16 is completed with the curve of the resistance power obtained considering ordinary resistance and hill resistance, so that

$$R_{tot} = \frac{1}{2}C_d \rho S v^2 + r_{rol} P + 10iP \quad [\text{N}] \tag{67}$$

where r_{rol} is the specific rolling resistance, i is the hill resistance in percentage (downhill is considered negative), and P is the vehicle weight [kN]. Lift effect can be considered by calculating P in the following way:

$$P = \frac{1}{1000}\left(P_s - \frac{1}{2}C_l \rho S v^2\right) \quad [\text{kN}] \tag{68}$$

where P_s [N] is the static weight of the vehicle, and C_l is the lift coefficient; if it is positive, an apparent reduction of the weight is obtained. If the lift effect is neglected, the weight is constant and the resistance curves are parabolas having axes coincident with the ordinates. The zone between the traction curve and abscissa axis represents the possible working field of the vehicle. Figure 16 shows that a hill resistance of 6% can be surmounted in fourth gear, and that the same hill can be surmounted in third gear at a greater speed. If the working field corresponding to two different gears is superimposed, the gear change is possible.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

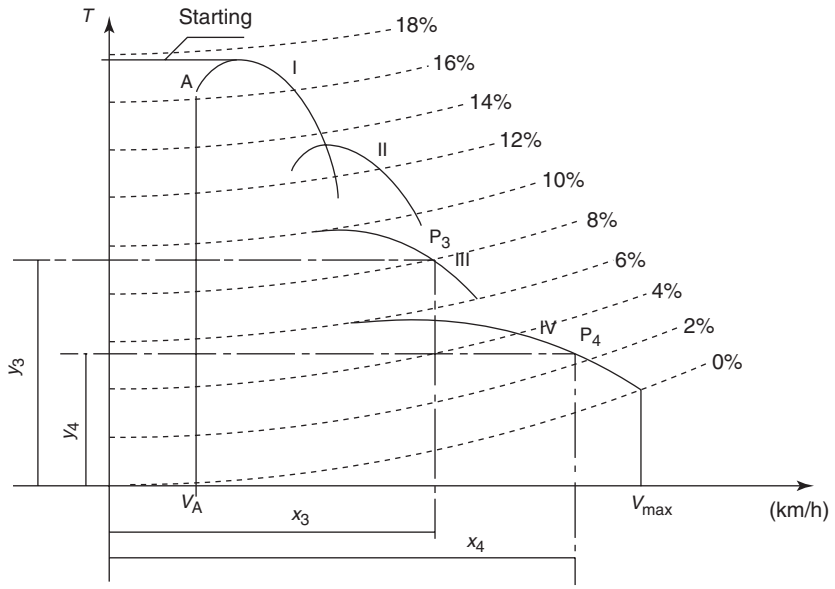


Figure 16. Traction versus speed.

Characteristic in first gear does not start from the ordinate axis because the engine cannot run under the minimum rotation number. The abscissa of the point A gives the minimum speed of the vehicle with the inserted engine; a lower speed than v_A requires the sliding of the clutch. For this reason, the acceleration at the starting point is considered constant, as is customary.

Traction characteristic of the vehicle can be obtained by Figure 16, knowing the point of gear change. Fixing the speed, the useful traction for the acceleration can be obtained by subtracting the resistance by the traction of the engine:

$$T_u = T\eta_t - R_{tot} \quad [N] \tag{69}$$

Because

$$N_u = T_u v = \frac{d}{dt}\left(\frac{1}{2}k_m \frac{P}{g}v^2\right) = k_m \frac{P}{g}va \tag{70}$$

the acceleration is

$$a = \frac{gN_u}{k_m P v} = \frac{gT_u}{k_m P} \quad [m/s^2] \tag{71}$$

with k_m is the inertia coefficient of the inserted gear. Diagram a–v in Figure 17 is useful to obtain the curve of speed and space versus time in the way described in the following paragraph.

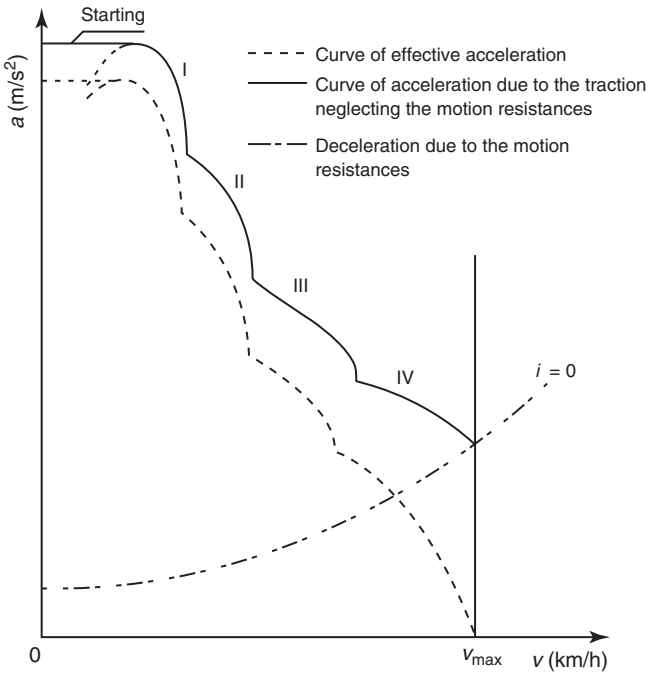


Figure 17. Acceleration versus speed curves; the vertical lines represent the gear changes.

2.3.2 *Diagrams N–v*

The estimate of the performance can be also executed by the construction of diagrams of power versus speed (De Gregorio, 1970; Heisler, 2004; Bencini, 1956). The speed

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

[km/h] is reported in the abscissa, while the power [kW] is reported in the ordinate. The curves of the resistance power are cubic parabolas approximately

$$N_r = R_{tot} \cdot \frac{v}{3.6} \cdot \frac{1}{1000} = \frac{R_{tot} \cdot v}{3600} \quad [kW] \qquad (72)$$

where R_{tot} is given by Equation 67 and v by Equation 64. Equation 64 converts the rotation of the engine into vehicle speed and then allows the construction of the used curves with several gears. The power characteristic of the engine has to be reduced by the efficacy of the transmission in the inserted gear, and the position of the point M (Figure 18) is in correspondence with the curve of the resistance power in plane if the axle ratio has the optimum value (Section 2.4.1). The speed corresponding to the point M is the maximum speed that the vehicle can reach in plane; slightly greater speed can be reached on a light downhill, without exceeding the maximum rotation rate. The diagram gives the following information:

1. Maximum hill resistance on the given gear;
2. Maximum speed with the given gear on a given hill;
3. The gear permitting the maximum speed on a given hill;
4. Cruise level, which is defined as the speed corresponding to the maximum engine torque rate.

Figure 18 shows also the resistance power on road with slope 2%. Until the speed v_n, the required power is negative for the gravity effect.

If the vehicle rides at speed v_n (at constant speed with disengaged engine), 2% is the harmful limit downhill. By increasing the number of gears, the extension of the dotted area is reduced, so that one has better utilization of the engine with better ability of acceleration and overcoming hills. In general, the maximum and minimum ratios are fixed with the performance of speed and hill; better ability of acceleration is obtained choosing the ratios by attempts, in order to make lower the dotted area.

Adaptation of the engine to the vehicle is the whole objective to obtain the requested performance compatible with the performances of the engine. Given a vehicle (S, C_d, P) and an engine with given characteristic of power and torque, the connection has to be implemented so that the performance is the more satisfactory, eventually for the optimum of one in detriment of others. The more reliable compromise constitutes the problem of the preliminary design of the vehicle (maximum speed, consumption on the road, acceleration, surmountable hill, towing, etc.).

2.4 Limits

2.4.1 Axle ratio calculation

The maximum achievable speed v_{max} can be obtained by solving Equation 72 versus the speed, putting

$$N_r = \eta_t \cdot N_{max} \qquad (73)$$

That is a limit condition. The axle ratio τ_p can be determined using Equation 63 by putting $v = v_{max}$ and $n = n_m$ [rpm], which is the maximum power rate (Figure 19), and putting τ_n equal to the highest gear ratio τ_υ. If the gear drive has a direct connection, τ_υ assumes the value 1, whereas if the highest ratio is implemented by a pair of gears, it assumes values very close to 1 because the two gears do not have the same number of teeth, to avoid the

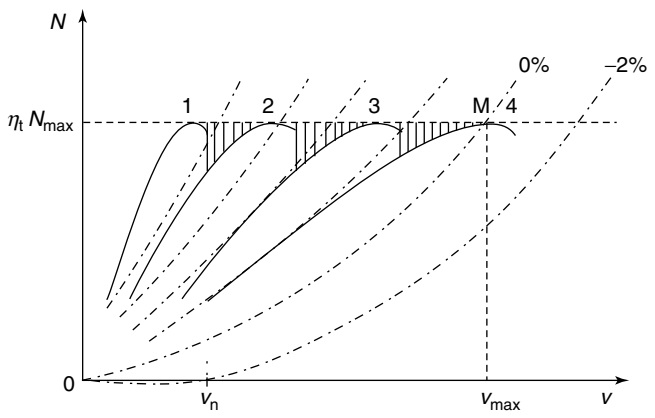


Figure 18. Four-gear N–v diagram.

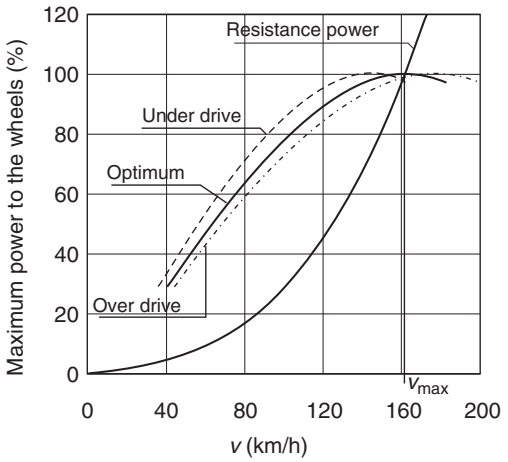


Figure 19. Determination of the optimum axle ratio.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

continuous contact between the same teeth. One has

$$\tau_{opt} = \frac{60}{2\pi n_m} \frac{1}{\tau_\upsilon R_0} \frac{v_{max}}{3.6} \tag{74}$$

The value calculated by Equation 74 is called the *optimum axle ratio*; the characteristic curve intersects the curve of resistance power corresponding to the maximum power rate. It is the value that allows the maximum speed with the selected engine.

If the intersection point corresponds to a greater or lower rate than the maximum power, the maximum speed of the vehicle will be lower than the possible one. For this reason, a decrease of the maximum speed is obtained if $\tau_p \neq \tau_{opt}$; a different ratio is chosen to obtain an improvement of other performances. In fact, greater ability of acceleration and greater gradability are obtained with $\tau_p < \tau_{opt}$ (underdrive), while the consumption on the road is lower if $\tau_p > \tau_{opt}$ (overdrive). In reality, small variations of τ_p give sensible improvements in some performances, while the consequent speed variation is only a percent of theoretical v_{max}. Moreover, the fact that the ratio τ_p is given by Equation 62, such as the ratio between two entire numbers, compels the designer to modify the optimum value. This is true also for all the other gears. By Figure 20, the following performances are obtained by increasing τ_p:

- Cruise level increases, with reduction of the consumption on the road (the fuel consumption is lowest at maximum engine torque).
- Minimum speed on the road increases. This is a negative aspect for urban running and for the urban drivability. A minimum speed of about 40 km/h should be ensured.
- Power reserve for the acceleration diminishes. This reduces the vehicle burst.

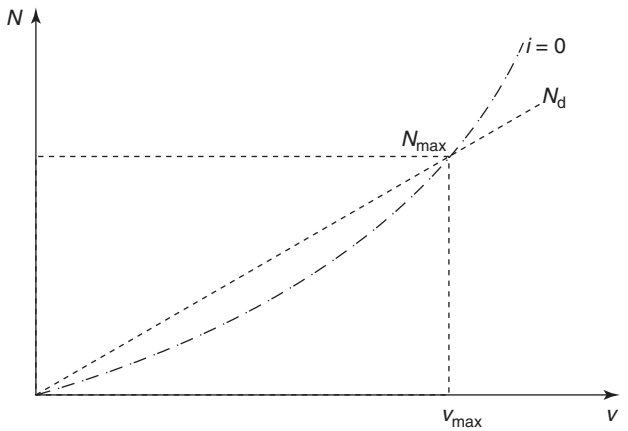


Figure 20. Maximum power on the vehicle and maximum speed.

2.4.2 *Top speed and power*

A limit exists to the power that may be implemented on the vehicle and transmitted to the wheels. If L is the adhesion weight and f the value of the longitudinal adhesion coefficient, the traction T_{max} is given by

$$T_{max} = f \cdot L \tag{75}$$

Equation 75 in term of power is

$$N_d = \frac{f \cdot L \cdot v}{3600} \quad [\text{kW}] \tag{76}$$

Fixing f and L for a given vehicle and for the wheel–ground coupling, the power is a linear function of the speed v, and then is

$$N_d = \frac{f \cdot L \cdot v}{3600} = R_{ord} v = A_0 \cdot v + B_0 \cdot v^3 \tag{77}$$

given that R_{ord} is a function of v^2. Neglecting the lift effect, A_0 and B_0 are constants and contain the conversion factors of the measurement units. Then

$$v_{max} = \sqrt{\frac{1}{B_0}\left(\frac{f \cdot L}{3600} - A_0\right)} \tag{78}$$

which expresses the maximum speed of the vehicle on $i = 0$ in absolute terms, as Figure 20 shows. The corresponding maximum power is

$$N_{max} = \frac{1}{\eta_t} \frac{f \cdot L}{3600} \cdot \sqrt{\frac{1}{B_0}\left(\frac{f \cdot L}{3600} - A_0\right)} \tag{79}$$

which is the admissible value of the maximum power on the vehicle.

2.4.3 *Minimum ratio*

A limit exists to the minimum ratio imposed by the maximum traction applied to the wheels. The traction is

$$T_m = \frac{\eta_t \cdot C_{max}}{R_0 \cdot \tau_{min} \tau_p} \tag{80}$$

As the wheels do not skid, one has

$$T_m \leq T_{max} \tag{81}$$

Then

$$\tau_{min} = \frac{\eta_t \cdot C_{max}}{R_0 \cdot f \cdot L \cdot \tau_p} = \frac{\eta_t \cdot C_{max}}{R_0 \cdot f \cdot \tau_p} \frac{1+m}{P} \tag{82}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where C_{max} is the maximum engine torque, η_t is the transmission efficacy, R_0 is the rolling radius of the wheels, P is the weight of the vehicle, τ_p is the axle ratio fixed in the previous way, and m is the ratio defined by Equation 15 and obtained by Equation 19 or 20. A lower value than given by Equation 82 is not useful because the greater part of the traction is not used, so that $\tau_1 \geq \tau_{min}$. Moreover, if the wheels begin to skid, f diminishes and the transmitted traction decreases further.

The maximum hill can be determined by Equation 17. In a first approximation, the quantity r_{ord} can be calculated by neglecting the aerodynamic resistance and the variation of the rolling resistance, as the vehicle speed is very low. This permits the calculation of m and of the ratio τ_{min} by Equation 82. Note that 4WD traction permits a lower gear ratio, with greater possibility of gradability. The calculated ratio has to be rounded up to consider the number of teeth of the gear drive that implements it. This permits the determination of the first gear ratio τ_1. The relative position of the curves is shown in Figure 21, which is a condition of tangency as the motion resistance can be approximated by a straight line. The calculated hill resistance can be overcome with a speed corresponding to the point M. The road irregularity, which introduces a diminishing of the speed, induces a deceleration of the vehicle until the stop. In other words, the gradability is not stable because the resistance power is greater than the wheel power if the speed does not correspond to the maximum engine torque rate. In reality, the run is stable if the residual acceleration has the value a_M. The corresponding power N_{acc} is given by

$$N_{acc} = k_I \cdot a_M \cdot \frac{P}{g} \cdot \frac{v_M}{3.6} \cdot \frac{1}{1000} \quad [kW] \tag{83}$$

where k_I is the inertia coefficient in the first gear. The vehicle is able to run on the hill if

$$i^* = i_{max} - k_I a_M \frac{100}{g} \tag{84}$$

The second term of the right-hand side is obtained by Equation 50. Curve i^* is drawn by putting $a_M = 0.3$ m/s with an inertia coefficient 1.4 and an adhesion coefficient 0.6. This permits the stable run of the vehicle between the speeds corresponding to the points A and A′.

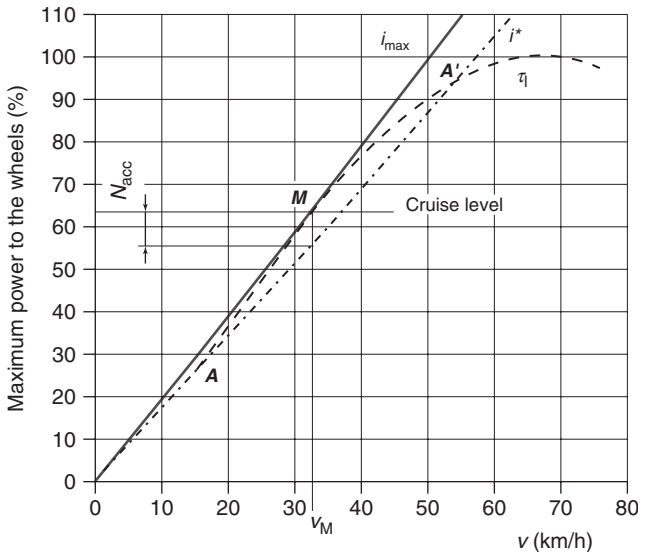


Figure 21. Minimum gear ratio and maximum hill.

2.5 Intermediate ratios selection

Transmission has a fixed number of gear ratios, so that the curves should be interrupted for loss of engine rotation between each gear change. For a saloon car or light van having a large power to weight ratio, a five or six gearbox is adequate to maintain the traction effect without high loss in engine rotation during the gear change. Heavy goods vehicles hauling large loads have a low power to weight ratio, and the fall-off of engine rotation is so high so that higher gear ratios are necessary. A comparison between Figures 22 and 23 shows that the diminishing of engine rotation is very small by doubling the gear ratios. To avoid too large and heavy gearboxes, a better way of extending the gear ratios is the use of a two-speed auxiliary gearbox in series with a conventional one. The combination of main and auxiliary gearboxes can be designed to be used as a splitter gear change or a range gear change (Figure 23). Also, in the case of a splitter gear change, the high ratio is kept nearly equal to 1. The second gear ratio is chosen so that it splits the main gearbox ratio steps in to half. The splitter indirect ratio normally is set 1.1–1.4 : 1, and it is set to 1.2 : 1 in Figure 23.

Ratios between the top and bottom gears have to be spaced in order to provide the traction effort as close to the ideal as possible. Intermediate ratios can be selected using a geometric progression as a first approximation. This requires the engine to operate within the same speed range in each gear to obtain the best fuel economy. The limits are the rotations n_H and n_L. Neglecting the change of the road speed, the speeds are v_1, v_2, v_3, v_4, and v_5 if the number of gears is equal to 5 (Figure 24). The gear ratios are τ_1, τ_2, τ_3, τ_4, and τ_5, and the overall gear ratios are obtained multiplying them by τ_p. The following

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

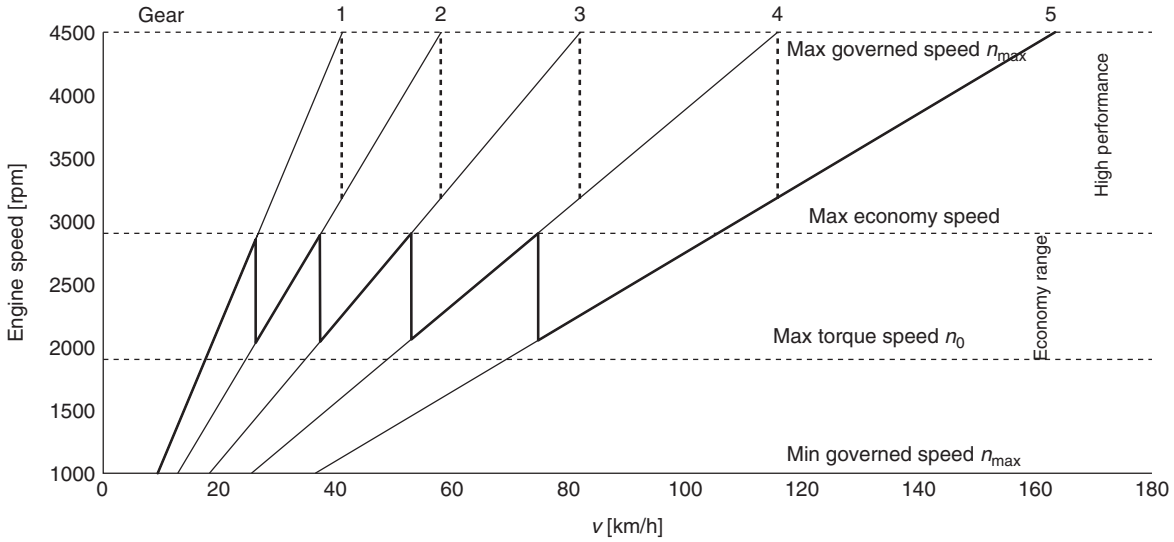


Figure 22. Engine speed ratio chart for a vehicle employing a five-speed gearbox.

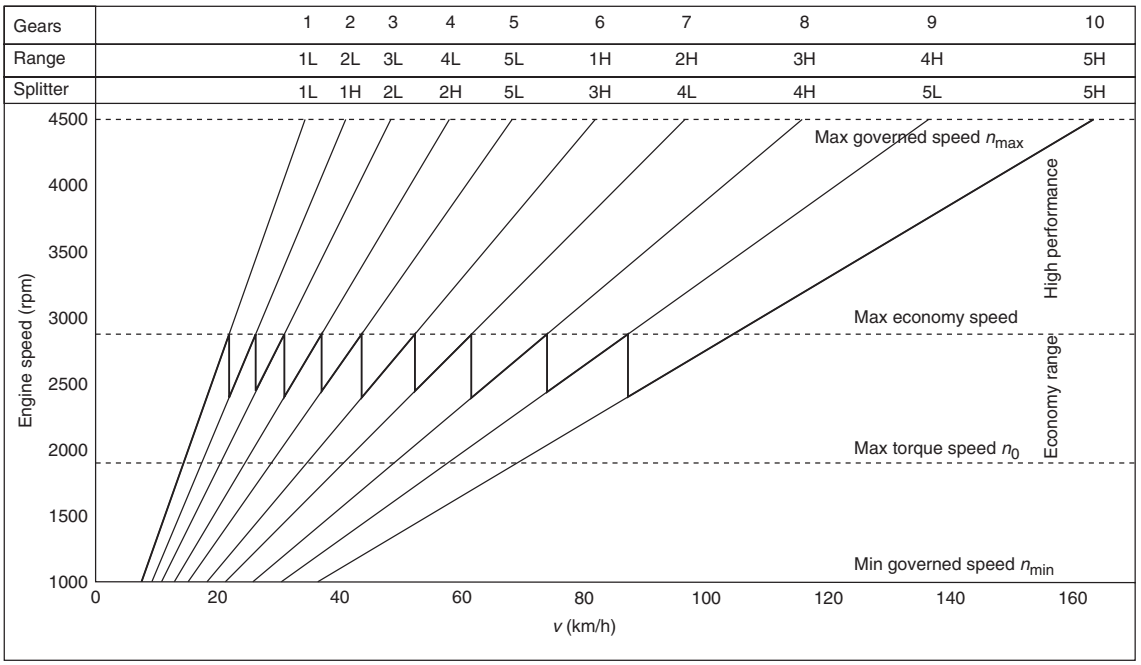


Figure 23. Engine speed ratio chart for a vehicle with a 10-speed range or a splitter change gearbox.

relationships are obtained:

$$\begin{aligned} v_5 &= n_H \tau_p \tau_5 & v_4 &= n_L \tau_p \tau_5 \\ v_4 &= n_H \tau_p \tau_4 & v_3 &= n_L \tau_p \tau_4 \\ v_3 &= n_H \tau_p \tau_3 & v_2 &= n_L \tau_p \tau_3 \\ v_2 &= n_H \tau_p \tau_2 & v_1 &= n_L \tau_p \tau_2 \\ v_1 &= n_H \tau_p \tau_1 \end{aligned} \tag{85}$$

Eliminating the speeds, one has

$$\tau_2 = \tau_1 \frac{n_H}{n_L} \qquad \tau_3 = \tau_1 \left(\frac{n_H}{n_L}\right)^2$$

$$\tau_4 = \tau_1 \left(\frac{n_H}{n_L}\right)^3 \qquad \tau_5 = \tau_1 \left(\frac{n_H}{n_L}\right)^4 \tag{86}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

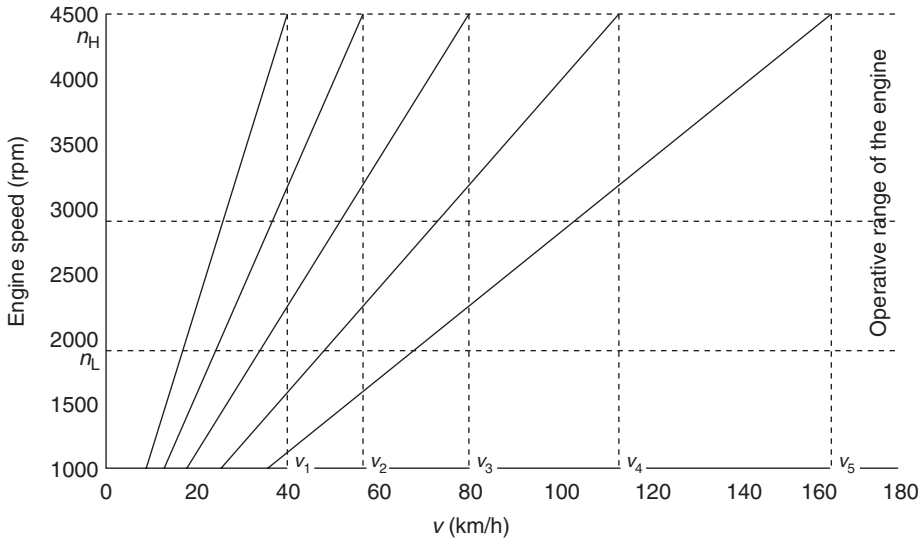


Figure 24. Gear ratio selected on geometric progression.

In general, the first gear ratio is known ($\tau_1 \geq \tau_{min}$), and the highest ratio τ_υ is equal to 1 or has a value very close to 1. By fixing the number υ of gears, the value of the generic ratio τ_n is given by the following relationship:

$$\tau_n = \tau_1 \left(\frac{\tau_\upsilon}{\tau_1} \right)^{\frac{n-1}{\upsilon-1}} \qquad n = 1, \ldots, \upsilon \tag{87}$$

The ratio k of the geometric progression is

$$k = \frac{n_H}{n_L} = \sqrt[\upsilon-1]{\frac{\tau_\upsilon}{\tau_1}} \tag{88}$$

The determination of n_H and n_L can give indications on the need of a splitter or range gear change. A correction can be executed choosing a few lower values if the performance in acceleration or ability to surmount hills has to be improved, and a few higher values if the consumption has to be diminished.

The overall number υ of the gears depends on the following characteristic:

1. n_{min} and n_{max} rates and ratio between the corresponding engine powers. If the characteristic is flat, the number of gears decreases;
2. Maximum engine power;
3. Vehicle acceleration. The characteristic of pickup improves by increasing υ.

Power–speed characteristic can be represented also in logarithmic coordinates. Their use was useful in the past, as the curves have to be drawn by hand. Today, the spread of computers does not make it convenient.

2.6 Vehicle acceleration

The starting, with the sliding of the clutch, takes a time T_a, so that the vehicle begins the acceleration phase by increasing the speed. In this phase, a human factor is present in the way the driver operates the clutch. It is customary to consider the acceleration to be constant during the entire starting time, for the purpose of this article. If the engine is at maximum torque rate and at full introduction

$$\frac{dv}{dt} = \left(\frac{C_{max}}{\tau_1 \tau_p R_0} - R_{rot} \right) \frac{g}{k_1 P} \tag{89}$$

2.6.1 Acceleration, speed, and space diagrams

To forecast the performances on the road, that is, in plane road, the diagrams N–v (or T–v) have to be completed by making some hypotheses on the gear changes. Also, this aspect is influenced by a human factor, because the driver's maneuver may be executed in several ways. Figure 25 shows a possibility. The drive change occurs in order to obtain the maximum torque in the successive gears and the maximum performance in terms of acceleration.

For the first condition, the starting is represented by a straight line linking the origin with the point of maximum torque in the first gear, like the tangent to the curve. All diagrams are drawn by making the hypothesis of the engine working at full introduction. Fixing a generic point A having speed v [km/h], the powers of interest are (i) the resistance power N_r [kW], (ii) the available power at wheels N_u [kW] at full introduction, and (iii) the available power

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

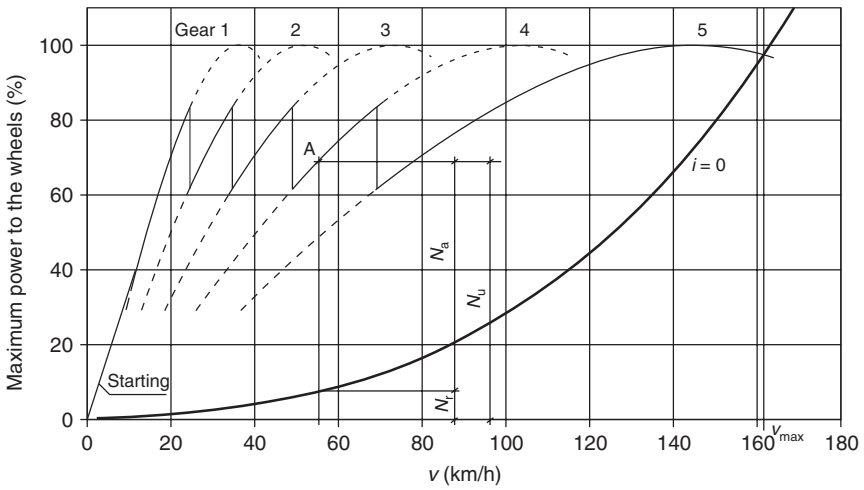


Figure 25. Hypothesis of maneuver of gear changes. The driver executes the change, represented by vertical lines, in order to obtain the maximum torque in the inserted gear.

for the acceleration $N_a = N_u - N_r$. The acceleration is given by

$$a = \frac{dv}{dt} = \frac{(N_u - N_r) \times 3.6}{v} \frac{g}{k_m P} \quad [m/s^2] \qquad (90)$$

where k_m is the inertia coefficient in the inserted gear, P [kN] is the vehicle weight, and g is the gravity. Acceleration versus the speed can be thus drawn, and it is similar in all respects to the graph of Figure 17 (in dotted line). To improve the performance forecast, the drawn curve may be corrected to consider the time t_l due to the change gears. During this time, the engine is disconnected from the wheels and the useful power is equal to zero, so that the vehicle suffers a deceleration d that is given by Equation 90, putting $N_u = 0$.

$$d = -\frac{N_r \times 3.6}{v} \frac{g}{k_n P} \qquad (91)$$

where k_n is the inertia coefficient in neutral position. This deceleration acts on the vehicle during the time of gear change depending both on a human factor and on the gearbox construction. It produces a diminishing of speed, which is given by

$$\Delta v = d \cdot t_l \qquad (92)$$

A more accurate definition is not necessary because the gear change time is small and assumes the value 0.5 s for an automated gearbox until 2.0 s, or more for a nonsynchronized gearbox. Then, the graph a–v is shown in Figure 26. Of course, the lines of gear change have to

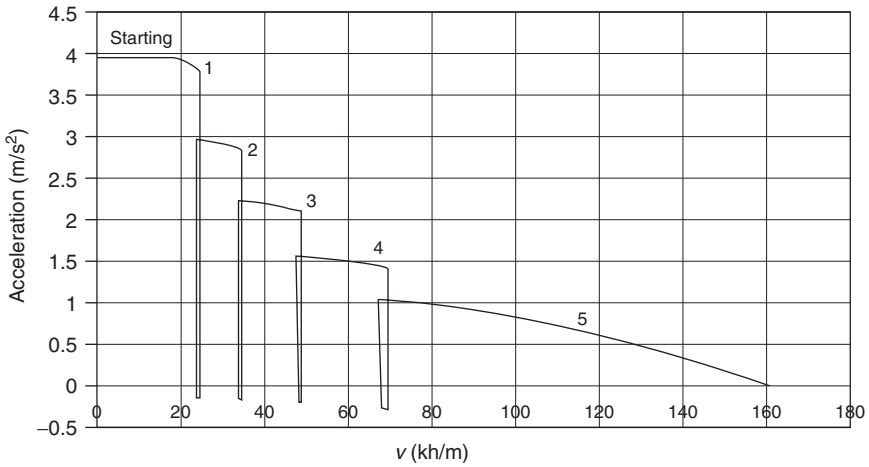


Figure 26. Acceleration versus speed considering the time of gear change, which is taken as 1.5 s.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

be considered as approximate. Now the graph v–t may be drawn because

$$a = \frac{\mathrm{d}v}{\mathrm{d}t} \tag{93}$$

One has

$$t = \int_0^v \frac{\mathrm{d}v}{a} \tag{94}$$

The diagram $1/a$–v can be integrated by a numerical method. For example, the use of a spreadsheet foil gives satisfactory results using Simpson's rule applied to the inverse of acceleration $1/a$. The optimum gear ratios make the time minimum by Equation 94; in this way, the graph of Figure 27 is obtained. It shows the times of gear change and the corresponding speed reductions. The curve is asymptotic to the value of maximum speed. The covered space versus time can be drawn (Figure 28) by a further integration. The curve of the space is asymptotic to a straight line having an angular coefficient equal to v_{max}. The time t_p is the waste of time to accelerate the vehicle and is the difference between the time to cover the distance s_0 with standing start and the time to cover the same distance with a constant speed equal to v_{max}.

To increase the ability of acceleration, the inertia moments of the rotating parts have to be reduced, in order to reduce the inertia coefficients. Very advantageous is the reduction of the time of gear change, which shows the utility of devices to make the insertion of the gears easy, reducing the mass in motion and automating.

2.7 Engine map and kilometric consumption performance

Equal consumption curves q [g/kW/h] are obtained experimentally by the engine bench and are drawn on diagram N–n or C–ω [rad/s] (Bencini, 1956; Heisler, 2004; Genta,

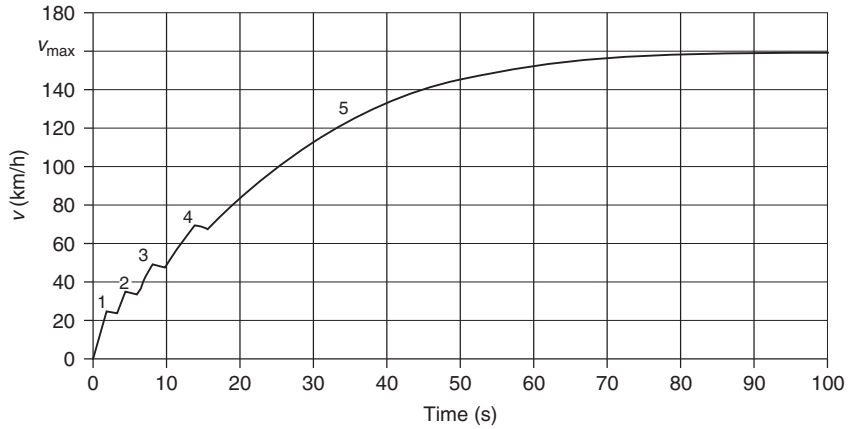


Figure 27. Speed versus time.

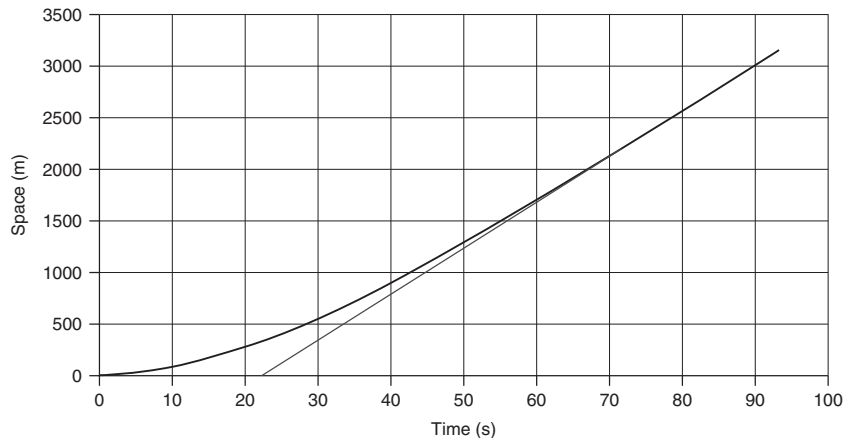


Figure 28. Space versus time

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

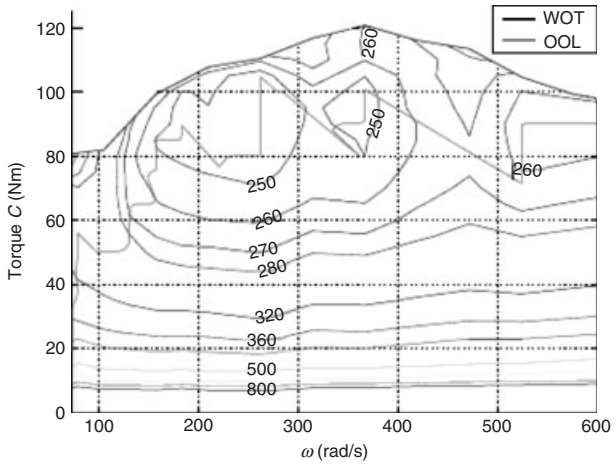


Figure 29. Static fuel map [g/kW/h] SI engine. $C_{max} = 120\,\text{Nm}$ at 3500 rpm, $N_{max} = 60\,\text{kW}$ at 5700 rpm.

2003). Figure 29 shows an engine map with the curves of equal consumption for all open throttle and optimized throttle. The line "WOT" represents the maximum obtainable torque by varying the rotation number of the engine; instead, the line "OOL" connects the points of minimum consumption of fuel.

Figure 30 shows both the power and torque characteristics and the corresponding specific consumption for the map of Figure 29. The power characteristic is obtained using a fourth-order polynomial, as the third-order one by Equation 55 is not satisfactory. Of course, the constraints of maximum power and maximum torque are respected. Specific consumption depends on the rotation rate n and on the introduction x, intended like a percentage of the full power, so that $0 < x < 1$. For example, the power corresponding to the point A in Figure 30 is 40 kW; the full power at the same rotation rate is 56.65 kW, so that the introduction corresponding is $x = 40/56.65 = 0.706$. The lines of specific consumption q are determined by interpolating the data in Figure 29. Both the lines at full introduction and at $x = 25\%$ of the power are shown in Figure 30. The interpolation versus n is executed with a polynomial of the second order, as a greater order does not give satisfactory results. The specific consumption increases, reducing the grade of introduction of the engine and has a minimum corresponding to the maximum torque rate. The economy range is a zone where the specific consumption has a minimum value and is the zone of normal working of the engine.

The interpolation of q versus x is not easy, as the curves at high percentage are close to the full introduction curve $x = 1$, so that both the linear and parabolic interpolation are not satisfactory. The work requires more advanced techniques of interpolation. After doing this, a definition of $q = q(x, n)$ at the entire field is obtained (Scamardi, 2007; Terruso, 2008; Terruso and Virzi' Mariotti, 2009). The precision is low because of the uncertainties of the extrapolation and in the experimental determination.

The minimum value of consumption for unity of run, in plane and in speed, is the maximum performance of kilometric consumption. The hourly consumption of the engine is defined by

$$C_t = \frac{q \cdot N'_w}{\eta_t \cdot \rho} \quad [\text{L/h}] \tag{95}$$

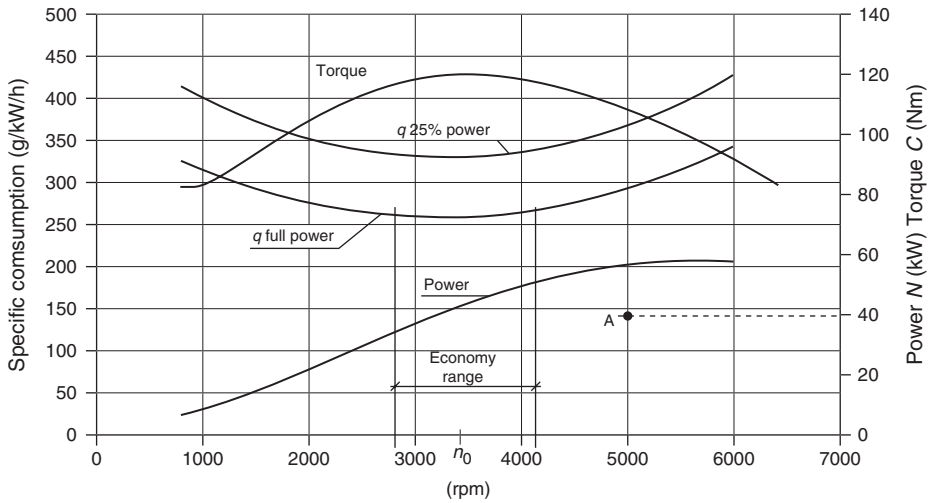


Figure 30. Characteristic of power, torque, and specific consumption of the engine.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where q is the specific consumption of the engine [g/kW/h], ρ is the density of the fuel [g/L], and η_t is the transmission efficacy. The available power N'_w to the wheels is a percentage of the full power N_w:

$$N'_w = xN_w = x\eta_t N \tag{96}$$

In the case of a hybrid vehicle, the power furnished by the electric motor has to be subtracted by Equation 96. Substituting Equation 96 into Equation 95 gives

$$C_t = \frac{x \cdot q \cdot N_w}{\eta_t \cdot \rho} = \frac{x \cdot q \cdot N}{\rho} \quad [\text{L/h}] \tag{97}$$

Fixing the C_t value, Equation 97 allows the calculation of x for a given value of n (or of the speed v), so that the curves at constant C_t may be drawn. Of course, the points with $x > 1$ have to be discarded. An example is shown in Figure 31, which is drawn for an overdrive, but the curves at constant C_t are associated with the curve of use, so that for the other gears they are drawn by changing the gear ratio to calculate the speed in the abscissa. The resistance power at $i = 0$ is also shown; the consumption in plane road at speed v_{max} is equal to ~21/L/h, while the consumption at speed 100 km/h ($\frac{2}{3}v_{max}$ is greater, following CUNA rule) is about 5.5 L/h. The graph of C_t versus time may be drawn after the construction of curves similar to Figure 25 and the successive figures, and then the necessary fuel C on a given stretch can be determined by the relationship

$$C = \int_{t_1}^{t_2} C_t \mathrm{d}t \tag{98}$$

Introducing a suitable conversion factor; t_1 and t_2 indicate the interval of the time.

The total consumption for 100 km is defined by

$$C_{100} = \frac{C_t}{v} \cdot 100 \quad [\text{L/100 km}] \tag{99}$$

By substituting Equation 97, one obtains

$$C_{100} = \frac{x \cdot q \cdot N_w}{\eta_t \cdot \rho \cdot v} 100 = \frac{x \cdot q \cdot N}{\rho \cdot v} 100 \ [\text{L/100 km}] \tag{100}$$

Curves at constant consumption corresponding to 100 km of run, for every use condition (v, i, τ), can be calculated by an analogous procedure knowing the vehicle speed v.

Figure 32 shows the curves at constant C_{100} in the case of an overdrive. If the vehicle rides on a plane road at speed v_{max}, the consumption is ~12 L/100 km, while riding at speed 100 km/h the consumption is about 5.5 L/100 km. Figure 33 shows the curves in the third gear; at constant speed 100 km/h, the consumption is ~6.2 L/100 km, while it increases to ~18 L/100 km at full acceleration. C_{100} curves depend on the gear ratio and have to be determined for every gear, unlike C_t. The family of curves can be reported in the N–v diagram to obtain an overall representation of both thermodynamic and economical performances of the vehicle. In general, a greater consumption corresponds to lower gear ratios. Also, the ratio

$$C_d = \frac{v}{C_t} = \frac{100}{C_{100}} \ [\text{km/L}] \tag{101}$$

is used in practice; it defines the number of kilometers covered by 1 L of fuel.

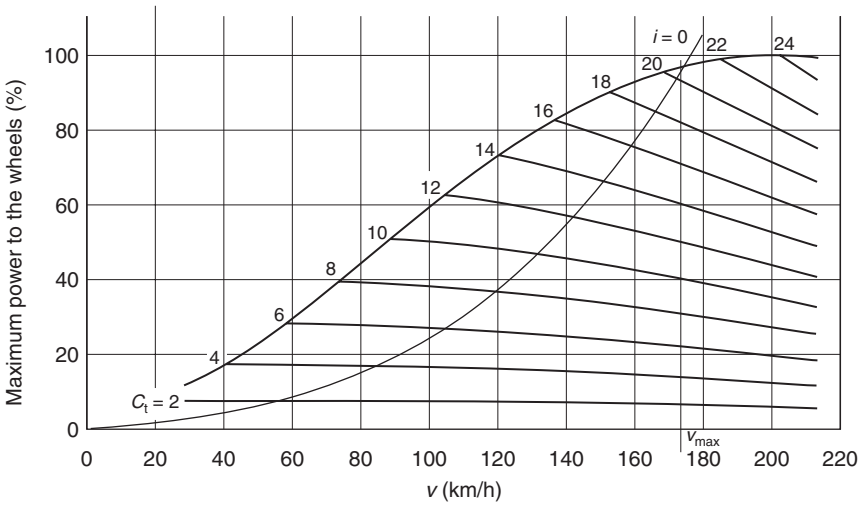


Figure 31. Curves at constant C_t [L/h] for an overdrive; $\eta_t = 0.95$; $\rho = 785$ g/L.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

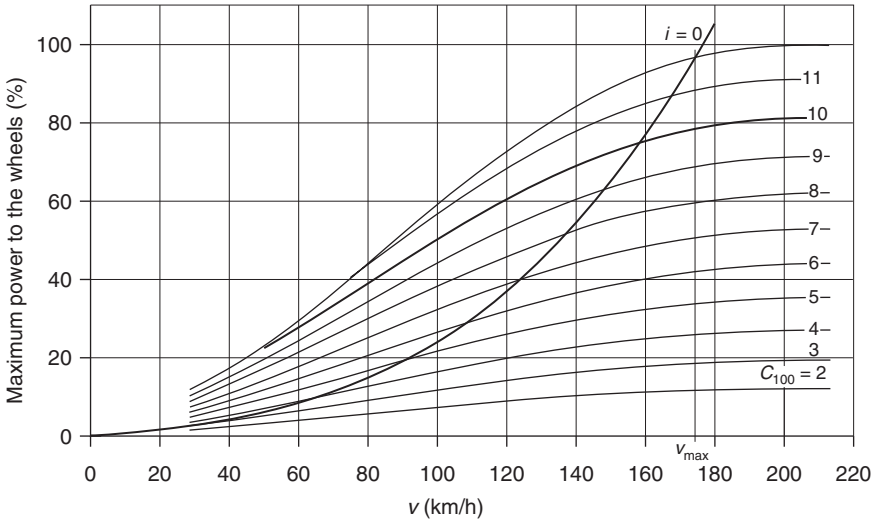


Figure 32. Kilometric consumption curves [L/100 km] for an overdrive.

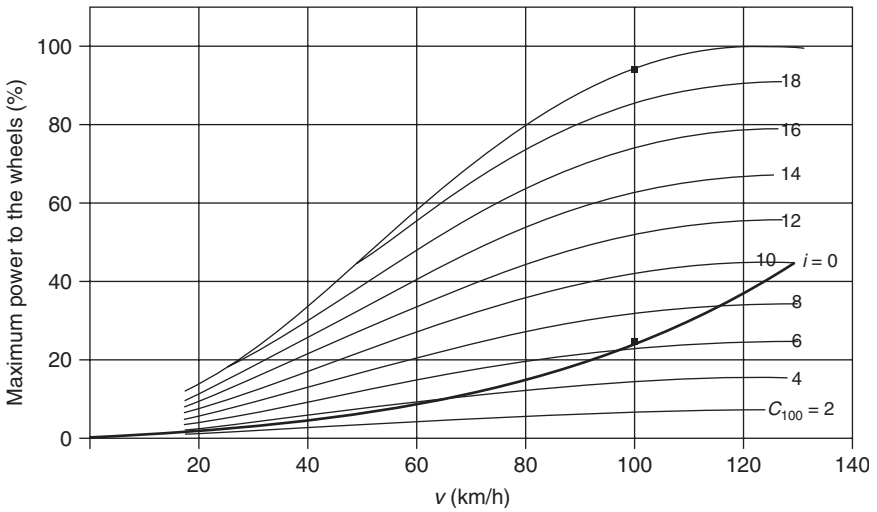


Figure 33. Kilometric consumption curves [L/100 km] in third gear.

A vehicle can be considered as a thermal machine, receiving energy (fuel) and furnishing work (transport), so that the overall efficacy of the change of thermochemical energy of the fuel into mechanical energy used by the vehicle during the motion can be calculated. The energy furnished to the vehicle in 1 h for a determinate mechanical performance (N, v, i) is

$$E_f = C_t \cdot \rho \cdot H_i \quad [\text{kJ/h}] \tag{102}$$

where ρ is the density of the fuel [kg/L] and H_i is the net heat value [kJ/kg]. Introducing C_{100}, one has

$$E_f = \frac{C_{100} \cdot v}{100} \cdot \rho \cdot H_i \quad [\text{kJ/h}] \tag{103}$$

The mechanical energy used in 1 h is

$$E_u = 3600 \cdot N \quad [\text{kJ/h}] \tag{104}$$

N [kW] being the power. The overall efficacy is

$$\eta_{cv} = \frac{E_u}{E_c} = \frac{3600 \cdot N}{C_t \cdot \rho \cdot H_i} = \frac{3600 \cdot N}{\frac{C_{100} \cdot v}{100} \cdot \rho \cdot H_i} = \frac{3600 \cdot N \cdot C_d}{v \cdot \rho \cdot H_i} \tag{105}$$

The efficacy varies with the transmission ratio and assumes values between 0.25 and 0.3 for a passenger car.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Consumption for kilonewton of transported load Q_u is

$$C_u = \frac{C_{100}}{Q_u} = \frac{C_t}{Q_u \cdot v} \cdot 100 \quad [\text{L/kN/100 km}] \qquad (106)$$

This parameter expresses the goodness of the engine (C_t) and of the vehicle (performance v and utilization Q_u) at the same time.

Kilometric consumption varies with v and i. The consumption assumes different values depending on the inserted gear at the same speed. The consumption at $\frac{2}{3}v_{max}$ is not indicative, as the consumption depends on

- the mean commercial speed v_{mc}, which in general is much lower than $\frac{2}{3}v_{max}$;
- the altimetry variation of the road (if the run is closed, the loss energy in the hill is recovered in part in the downhill motion);
- the acceleration and deceleration of the vehicle during the run;
- the time of thermal rate of the engine (at the starting, a part of the energy is lost in heating the coolant).

Today, other data are declared, such as the urban consumption and consumption on a mixed way.

2.8 Drivability

Handling is a dynamic property of the vehicle that is not to be mistaken with the maneuverability. Both these properties constitute the drivability of a vehicle. The maneuverability can be defined as the maximum obtainable performance, quantized by the necessary time to execute a maneuver, starting from an initial point until the arrival, satisfying the physical constraints (maximum forces supported by the tires and maximum engine torque) and geometry (the vehicle has to remain inside the route). Handling is, instead, defined in terms of the maximum obtainable performance considering also the driver limits, for example, the speed to apply the brakes, the gas, and the steering. In other words, handling is linked to the answer characteristic of the vehicle considering the physical and mental effort required by the driver. Very important is the facility of the driver to integrate with the vehicle.

Kinematic steering (Pollone, 1970; Stagni, 1980) is defined as the motion of a vehicle on a curved trajectory determined by the pure rolling of the tires, so that the sideslip angles are equal to zero. The center of instantaneous rotation O has to be the same for all the wheels (Figure 34). Referring α and β as the steering angle of the two front wheels, the following relationships are valid:

$$\frac{OC}{AC} = \frac{1}{\tan\beta} = \frac{R_2 + \frac{s}{2}}{p} \quad \frac{OD}{BD} = \frac{1}{\tan\alpha} = \frac{R_2 - \frac{s}{2}}{p} \qquad (107)$$

where p is the wheelbase, s is the front carriage, and R_2 is the radius of curvature at the rear carriage. Subtracting Equations 107, the following relationship is obtained:

$$\frac{1}{\tan\beta} - \frac{1}{\tan\alpha} = \frac{s}{p} \qquad (108)$$

That is known as the *Ackermann fundamental equation*. Today, steering systems are derived by the "Jeantaud quadrilateral" (Figure 35), but do not satisfy Ackermann relationship, so that a steering error is committed. One has

$$\tan\gamma = \frac{d}{p} \cong \frac{s}{2p} \qquad (109)$$

Moreover

$$a \cong \frac{s}{10} \qquad (110)$$

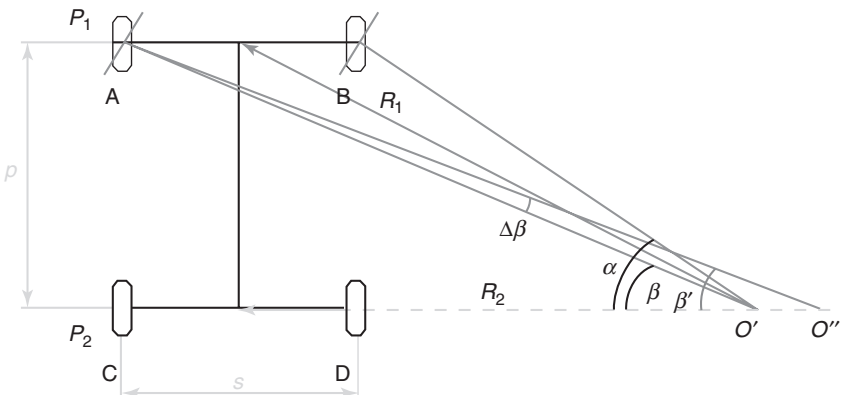


Figure 34. Ideal steering of the vehicle and steering error.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

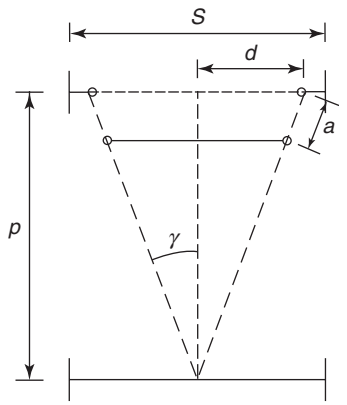


Figure 35. Jeantaud quadrilateral.

a and γ are two parameters to correct for the reduction of steering error. In particular, the designer has to reduce the steering error at low values of steering angle in order to improve the drivability during the runs with low steering angle (roadway or mixed way). This reduces the additional resistance in curves.

Figure 36 shows that the steering error can be strongly reduced increasing the γ value. α and β are the steering angles of wheels A and B, respectively (Figure 34). $\Delta\beta = \beta - \beta'$ is the steering error committed by Jeantaud quadrilateral with regard to the ideal condition of Ackermann.

The acronym 4WS indicates four-wheel steering. The working is divided in to *in contra phase* and *in phase*. The first is used when the vehicle has low speed on road; the rear wheels have steering angle of opposite sign than front wheels. The second is used when the vehicle runs at high speed; the rear wheels have steering of the same sign.

The elk test (Figure 37) (or moose test) (De Gaetano, Guerrera, and Virzì Mariotti, 2003; Lo Guasto, 2011) is executed to verify the stability of a vehicle during the execution of sudden steering to avoid unexpected obstacles on the road. The test is carried out on a dry surface; traffic cones are disposed at *S* along the way, the road, and its side

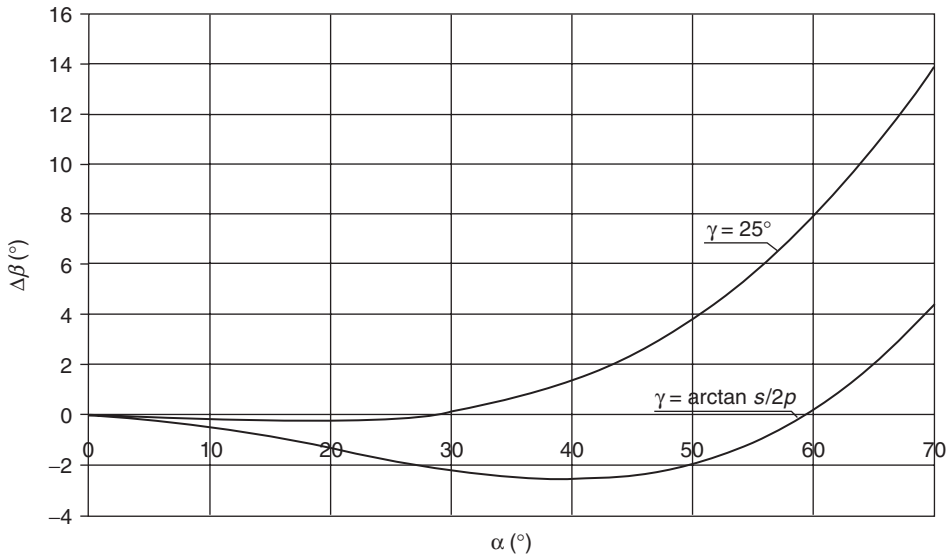


Figure 36. Steering error with increasing value of γ.

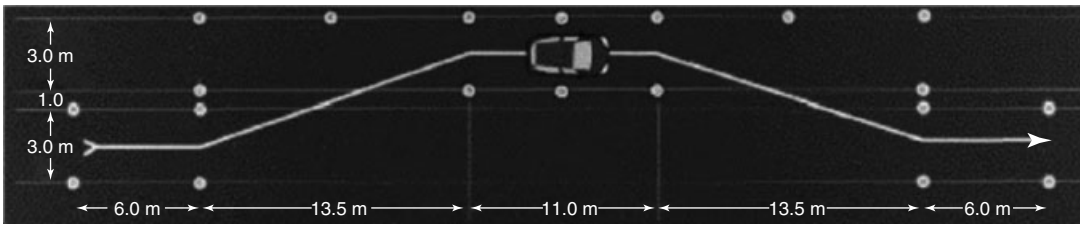


Figure 37. Elk test.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to simulate the obstacle; and the vehicle is at full load. In the neighborhood of the obstacle, the driver steers suddenly toward the opposite lane to return at once in the run lane. The test consists in the overcoming of the way at 120 km/h. This allows the comparison of the data of two-wheel drive (2WD) and 4WD, such as sideslip angles of the wheels, angle and jaw speed, roll, force in the ground–tire contact, and percentage error in the trajectory. The way the stability test is conducted consists of an initial part at constant speed long 70 m, a gap of 30 m, a part on the left lane of 20 m, and a successive return to the right lane. In the simulation, the control logic of the vehicle is of type "closed loop," given that the driver executes corrections on the trajectory. Typical data may be imposed in the following way: driver preview time = 0.73 s, driver time lag = 0.4 s. The first time corresponds to the prevision of the driver on the trajectory, and the second is the reaction time. They constitute a human factor with similar values to the real drivers. The vehicle is a saloon car of class D at three volumes, with long wheelbase (Table 3), and the overall duration of the handling test is 7 s. The simulator furnishes the data of both the vehicles; as an example, Figure 38 shows the front load F_x in the motion direction.

Results can be validated by comparing longitudinal force F_x, the side force F_y, and the aligning moment M_z with the values obtained by the Pacejka magic formulae. The necessary coefficients may are assumed from the international literature for a saloon car having a mean size (i.e., Genta, 2003), while the vertical load F_z ($=P_w$) is obtained by the same simulation: so, like the sideslip angle, the camber angle and the longitudinal slip. Another comparison may be executed between the vertical load obtained by the simulation and the vertical load obtained by the following theoretical relationship (Stagni, 1980):

Table 3. Vehicle data.

Total mass	1530 kg
Sprung mass	1370 kg
Unsprung mass	160 kg
Height	1.471 m
Length	4.620 m
Width	1.795 m
Carriage	1.550 m
Front camber	-2°
Front toe in	-1°
Rear camber	-4°
Rear toe in	0° 4WS vehicle 1° 2WS vehicle
Front king pin	10°
Front caster	7°
Rear kingpin	10°
Rear caster	5°
Front spring constant	153,000 N/m
Rear spring constant	82,000 N/m
Tires	215/55 R17
Rear wheels maximum steering	3.5°

$$\frac{\Delta F_{z1}}{F_{z1}} = \frac{2a_0}{g}\left[\frac{h'}{s} + \phi\frac{h''}{s}\frac{P}{F_{z1}}\left(\frac{m_1}{m_s - \phi \cdot P \cdot h''}\right)\right] \quad (111)$$

The symbols are listed in Table 4. An analogous formula can be written for the rear carriage, by changing the index 1 in 2. m_s [Nm] is called the *crosswise stiffness moment*; applying it around the longitudinal axis of the vehicle, a roll angle equal to 1 rad is obtained. It is the sum of the values on the front carriage m_1 and on the rear carriage m_2:

$$m_s = m_1 + m_2 = 2b_1^2k_1 + 2b_2^2k_2 \quad \text{[Nm]} \quad (112)$$

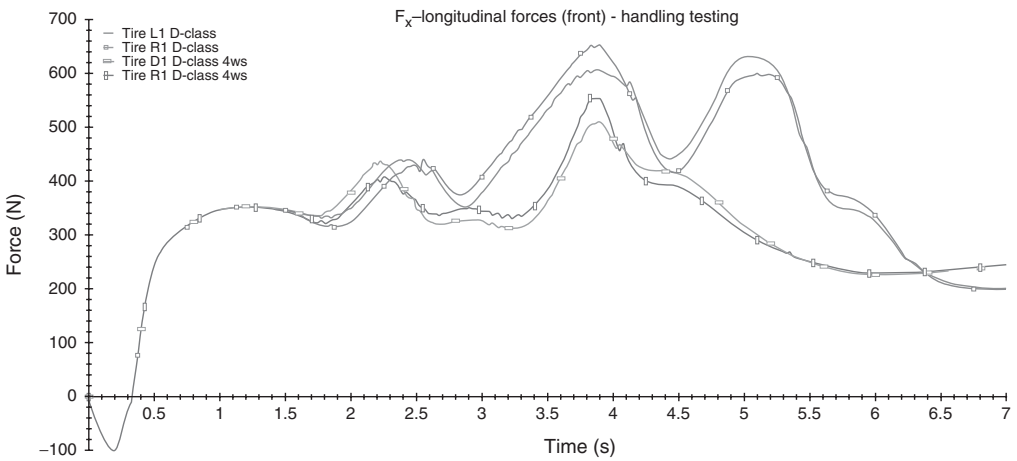


Figure 38. Front load F_x.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Data for drivability.

a_0	Centrifugal acceleration of the vehicle	[m/s^2]
φ	Ratio between sprung and total mass	—
F_{z1}	Weight acting on the front carriage	[N]
F_{z2}	Weight acting on the rear carriage	[N]
ΔF_{z1}	Weight difference between the wheels of the front carriage	[N]
P	Vehicle weight	$F_{z1}+F_{z2}$
s	Front carriage	[m]
G′	Center of gravity of the sprung mass	—
G″	Center of gravity of the unsprung mass	—
h'	Distance between G′ and G″	0.205 m
h''	Distance between G″ and the ground	0.335 m
h	Height of the center of gravity $= h'+h''$	0.540 m
$2b_1$	Distance between the constraints of the front springs	0.900 m
$2b_2$	Distance between the constraints of the rear springs	1.460 m

where k_1 and k_2 [N/m] are the front and rear spring stiffness, respectively. A 4WS vehicle has a better performance because the front and rear sideslip angles are lower than for a two-wheel steering (2WS) vehicle, resulting in greater vehicle stability. A 2WS vehicle has a greater yaw and can be maneuvered with only greater difficulty. Also, roll angle and roll speed improves much in the 4WS vehicle, as the displacements of the weight due to centrifugal acceleration are reduced. The force F_z is always lower in the 4WS vehicle.

Handling test is executed in a different way than the previous test. Because the 4WS system works contra phase under 60 km/h, the test speed varies between 10 and 50 km/h. The chosen way is similar to the elk test.

Figure 39 shows the handling map of a 4WS vehicle. It is obtained by simulating several tests in the way, for several values of speed. The map has the speed v of the vehicle as abscissa and the bend value of the trajectory as ordinate; moreover also the steering angle varies. Each maneuver is represented by a point. Knowing the speed and the bend, the optimum value of the steering angle can be determined. The same tests are executed on 2WS vehicle, obtaining the handling map of Figure 40.

For the same values of speed and bend, a 4WS vehicle needs a lower steering angle compared to a 2WS vehicle, so that is more handy; in fact, it is advantaged by the contra phase steering (3.5°) of the rear wheels, which increases the handling in notable way.

Another test of drivability is the *slalom test*. The vehicle is subjected to a repeated sequence of curves that reaches harsh stress. The centrifugal force compels the vehicle to a trajectory error, forcing the driver to make sudden corrections.

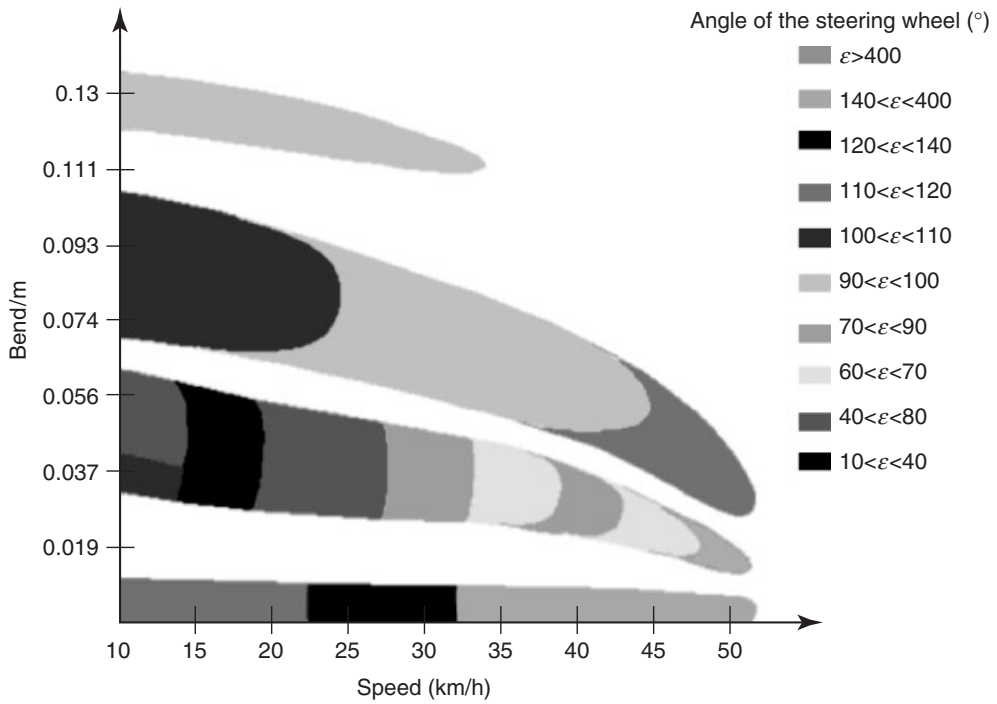


Figure 39. Handling map of a 4WS vehicle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

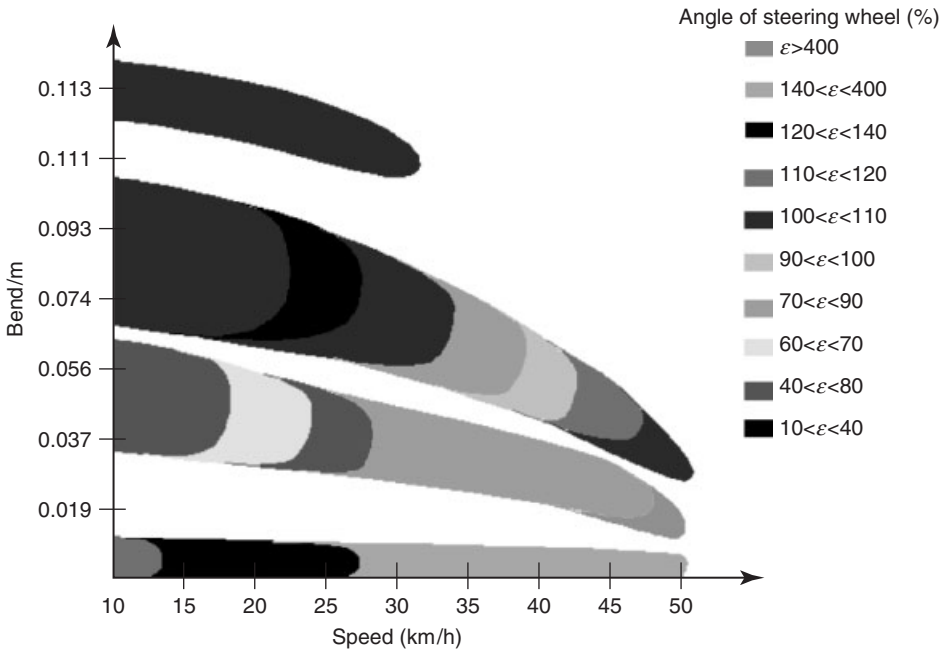


Figure 40. Handling map of a 2WS vehicle.

REFERENCES

Bencini, M. (1956) *Dinamica del veicolo considerato come punto*, Tamburini, Milano.

Buchheim, R., Deutenbach, K.-R. and Luckoff, H.-J. (1981) Necessity and Premises for Reducing the Aerodynamic Drag of Future Passengers Cars. SAE Paper No. 810185, Society of Automotive Engineers, Warrandale, PA.

Cogotti A. (1998). A parametric study on the ground effect of a simplified car model. SAE Technical Papers 980031, Warrendale, PA, USA, 02-01-1998.

Commission Directive 31988L0195 1988/195/EEC of 24 March 1988 adapting to technical progress Council Directive 80/1269/EEC on the approximation of the laws of the Member States relating to the engine power of motor vehicles.

Commission Directive 32002L0041 2002/41/EC of 17 May 2002 adapting to technical progress Directive 95/1/EC of the European Parliament and of the Council on the maximum design speed, maximum torque and maximum net engine power of two- or three-wheel motor vehicles.

CUNA NC 003–01 (2004) ed. 07.00 Veicoli a motore—Metodo per la determinazione del consumo di combustibile.

CUNA NC 003–02 (2004) ed. 07.00 Veicoli a motore—Metodo per la determinazione della velocità massima.

Dabbene F. (2009) Analisi CFD delle prestazioni di un catamarano a semiscafi asimmetrici dotato di sostentamento idrodinamico. Graduate Thesis. University of Palermo, Mechanics Department.

D'Anca, C., Mancuso, A. and Virzi' Mariotti, G. (2005) Optimization of a vehicle shape by CFD code *International Journal of Vehicle Design*, **38** (1), 26–41.

De Gaetano S., Guerrera G. and Virzi' Mariotti G. (2003) Dynamic Simulation of the Behaviour of a "Light Vehicle" by Disable Users, Tested on a Virtual Road. *XIX International Conference Science and Motor Vehicles 2003*, May 26–28, 2003, on CD, JUMV, Belgrade (Serbia).

De Gregorio, C. (1970) *Meccanica della locomozione terrestre, marittima ed aerea*, Denaro Editore, Palermo, Italy.

Fessia A. (1948) Sulle condizioni limite di aderenza longitudinale sui veicoli a trazione anteriore, *ATA*, February–March 1948.

Genta, G. (2003) *Motor Vehicle Dynamics Modeling and Simulation*, World Scientific, Singapore.

Heisler, H. (2004) *Advanced Vehicle Technology*, 2nd edn, Elsevier, Oxford.

Hucho, W.-H. (1998) *Aerodynamics of Road Vehicles, from Fluid Mechanics to Vehicle Engineering*, Society of Automotive Engineers 4th Revised edn (1 July 1998), US.

ISO 1585:1992 Road vehicles—engine test code—net power.

ISO 28580 (2009) Passenger car, truck and bus tyres—methods of measuring rolling resistance—single point test and correlation of measurement results Edition: 1 | Stage: 60.60 | TC 31 ICS: 83.160.01.

Janssen, L.J. and Hucho, W.H. (1975) Aerodynamic optimization of body details of the Volkswagen passenger cars Golf and Scirocco, Aerodynamische Entwicklung von VW Golf und Scirocco *ATZ*, **77** (11), 309–313.

Katz, J. (2006) Aerodynamics of race cars *Annual Review of Fluid Mechanics*, **38**, 27–63.

Lo Guasto S. (2011) Comportamento su strada di un autoveicolo a Quattro ruote sterzanti. Graduate thesis. University of Palermo, DICGIM Department, July 2011.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Milone S. (2011) Studio fluidodinamico di ottimizzazione di forma in campo automobilistico e navale, con implementazione dei codici CFD nei software VPP. PhD Thesis. University of Palermo, DICGIM, Palermo (Italy).

Morelli, A. (1970) in *Costruzioni Automobilistiche, Enciclopedia dell'ingegneria* (ed M. Lenti) Vol. III, parte 14, Isedi, Milano, Italy.

Pollone, G. (1970) *Costruzioni automobilistiche, il veicolo*, Levrotto e Bella, Torino, Italy.

Popio, J.A. and Luchini, J.R. (2007) Fidelity of J1269 and J2452 *Tire Science and Technology*, **35** (2), 94–117.

SAE J1269 (2000) Rolling Resistance Measurement Procedure for Passenger Car, Light Truck, and Highway Truck and Bus Tires. Date Published: 2000-09-12.

SAE J1349_200403: Engine Power Test Code-Spark Ignition and Compression Ignition-Net Power Rating, March 2004.

SAE J2452 (1999) Stepwise Coastdown Methodology for Measuring Tire Rolling Resistance. Date Published: 1999-06-01.

Scamardi M. (2007) Studio sugli autoveicoli a trazione ibrida ICEHV. Graduate Thesis. University of Palermo, Dip. Meccanica, Plaermo, Italy.

Schenkel, F.K. (1977) The origins of drag and lift reductions on automobiles with front and rear spoiler. SAE Paper No. 770389. Society of Automotive Engineers, Warrandale, PA.

Stagni, E. (1980) *Meccanica della Locomozione*, Patron, Bologna, Italy.

Terranova L. (2009). Determinazione del coefficiente di penetrazione aerodinamica di un veicolo da turismo mediante analisi CFD. Graduate Thesis. University of Palermo, Department of Mechanics, Italy.

Terruso M. (2008) Previsione del comportamento su strada di un autoveicolo ibrido con trasmissione CVT, Graduate Thesis, University of Palermo, Dip. Meccanica.

Terruso, M. and Virzi' Mariotti, G. (2009) Prediction of driving performance of a hybrid vehicle with CVT transmission *Mobility and Vehicle Mechanics* RS, **35** (3), 27–42. ISSN: 1450-5304

Terziyski, J. and Kennedy, R. (2009) Accuracy, sensitivity, and correlation of FEA-computed coastdown rolling resistance *Tire Science and Technology*, **37** (1), 4–31.

White G.R.S.A. (1967–1969) Rating method of assessing vehicle aerodynamic drag coefficient. Report MIRA (Motor Industry Research Ass.) Nuneaton, Warwickshire, England.

FURTHER READING

Cossalter, V. (1999) *Cinematica e dinamica della motocicletta*, Casa Progetto, Padova, Italy.

Deutsch, C. (1970) *Dynamique des Vehicules Routiers*, O.N.S.E.R., Arcueil, France.

Fenton, J. (1998) *Handbook of Automotive Powertrain and Chassis Design*, Professional Engineering Publishing Limited, London.

Giri, R.N.K. (1996) *Automotive Mechanics*, 7th edn, Khanna Publishers, New Delhi.

Gillespie, T.D. (1992) *Fundamentals of Vehicle Dynamics*, Society of Automotive Engineers, Warrendale, USA.

Heisler, H. (1985) *Vehicle and Engine Technology*, Edward Arnold, London.

Milliken, W.F. and Milliken, D.L. (1995) *Race Car Vehicle Dynamics*, SAE International, Warrendale, USA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Please note that the abstract and keywords will not be included in the printed book, but are required for the online presentation of this book which will be published on Wiley Online Library (http://onlinelibrary.wiley.com/). If the abstract and keywords are not present below, please take this opportunity to add them now.

The abstract should be a short paragraph of between 150–200 words in length and there should be 5 to 10 keywords

Abstract: This chapter deals with the forecasting of vehicle performance on the road, according to the installed engine. It begins with Section 1, which shows the calculation of the resistance to motion, that is, the forces that the vehicle must overcome to advance on flat roads and uphill. The calculation excludes accidental resistance in curves because it is not essential for predicting the behavior of the vehicle on the road. Section 2 begins with the power characteristic, the forecast by traction–speed and power–speed diagrams, calculation of gear ratios, and then passing to the limits to the estimated acceleration, feature consumption, and distance. This chapter concludes with a section on the drivability of the vehicle, and sets out the advantages of four-wheel-drive vehicles.

Keywords: motion resistances; inertia coefficient; performance forecast; power characteristic; torque characteristic; traction–speed diagram; power–speed diagram; consumption diagram; maneuverability; handling

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto107
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Emission Control Systems—Oxides of Nitrogen

Soonho Song

Yonsei University, Seoul, Korea

1 Introduction 1
2 NO_X Control Technologies 1
3 Applications 12
References 15

1 INTRODUCTION

Facing the reinforced emission regulations for oxides of nitrogen (NO_X), many approaches have been proposed and applied to control NO_X emissions from current engine systems. NO_X control technologies can be grouped into two categories: (i) in-cylinder control and (ii) aftertreatment systems.

In-cylinder control provides NO_X reduction before and during the combustion process by decreasing the gas temperature and controlling the chemical reactions of combustion. In-cylinder control technologies include the following:

- Combustion control by optimizing the spark timing for spark ignition (SI) engine and injection strategies for compression ignition (CI) engines.
- Applying exhaust gas recirculation (EGR) to control combustion temperature and chemical reactions.
- Intercooling under boosting systems to control intake air temperature.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Aftertreatment systems provide the means to reduce already formed NO_X after combustion process via chemical reactions between the exhaust gas and other chemical components such as catalysts. Aftertreatment technologies could include the following:

- three-way catalyst in the SI engine and
- selective catalytic reduction (SCR) and NO_X adsorber in the CI engine.

In the SI engine system representing the gasoline engine, the problems of NO_X emission are less important because most of the NO_X reduction can be achieved using a three-way catalyst. However, in CI engine system representing the diesel engine, there will much room for improvement to develop a suitable and effective de-NO_X technologies. (see NO_x Formation and Models).

2 NO_X CONTROL TECHNOLOGIES

2.1 In-cylinder control in SI engine

2.1.1 Spark timing and EGR

Spark timing significantly affects NO_X emission levels. However, retarding the SI to reduce NO_X has been limited in order to avoid performance degradation. Thus, EGR has been combined to control NO_X with spark timing. EGR is combustion technology to reduce NO_X by recirculating cooled exhaust gas back into the combustion chamber. A more detailed description of EGR will be discussed in Section 2.2.2.

As shown in Figure 1, both EGR and spark retard (delaying the spark timing and hence the start of combustion) effectively decrease the NO_X emissions by reducing the peak combustion pressure and flame temperature.

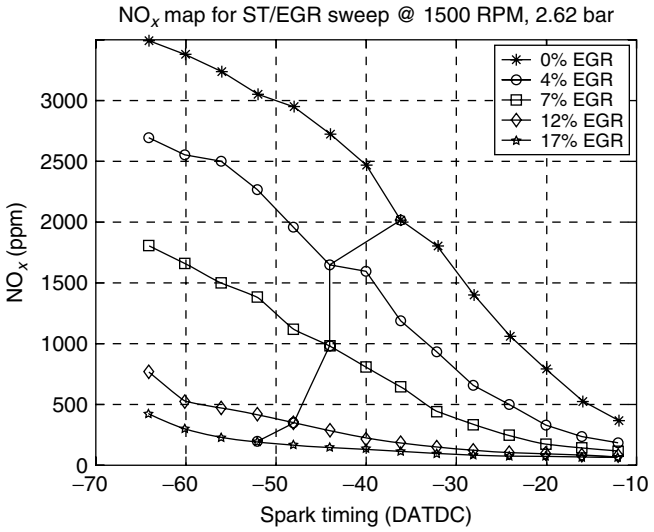


Figure 1. NO_X at different EGR/ST settings. (Reproduced by permission of American Automatic Control Council. (Haskara *et al.*, 2006).)

As maximum brake torque (MBT) spark is the optimal spark setting for fuel economy, EGR is left as the most suitable remaining control variable for NO_X emissions. (see Spark Ignition Combustion).

2.2 In-cylinder control in CI engine

2.2.1 Fuel injection system

2.2.1.1 Injection timing. Adjustment in injection timing is one of the fundamental means of reducing NO_X emissions. Mechanical fuel injection systems were the first to incorporate variable injection timing. However, as electronics become more prevalent in diesel engine control, electronically controlled injectors have become the preferred means for the achieving variable injection timing; offering unprecedented flexibility in injection timing settings.

While NO_X reduction via retarded injection timing retard can be effective, there would be significant trade-offs in terms of fuel consumption and particulate matter (PM) emissions. In many cases, these trade-offs must be dealt with through additional engine design enhancements. One early approach to reduce the fuel penalty associated with retarded injection timing was to reduce ignition delay using a high compression ratio and higher injection pressures. Additional measures such as reductions in oil consumption, increases in charge air pressure, increases in injection pressure, reductions in injector nozzle hole size, reductions in engine friction losses, and reductions in intake manifold temperature can also be taken to control fuel consumption and increases in PM emissions. Thus, injection timing itself has a clear limitation in its ability to reduce NO_X emissions. This places a practical lower limit of around 4 g/kWh NO_X that can be achieved with injection timing retardation. Further NO_X reductions have required additional measures such as injection rate shaping, pilot injections, as well as EGR and NO_X aftertreatment. While injection timing retardation is no longer the primary means of NO_X control, it is still an important tool that can be used in conjunction with other control measures to provide further NO_X reductions (Jääskeläinen, 2010).

2.2.1.2 Injection pressure. As demands for lower emissions, improved performance, and reduced fuel consumption have intensified, one prominent trend in fuel injection system design that has gained momentum is a steady increase in maximum injection pressure in new diesel engines. Figure 2 represents one of the dilemmas of increase fuel injection system pressure—while brake-specific fuel consumption (and PM) can be reduced, NO_X emissions

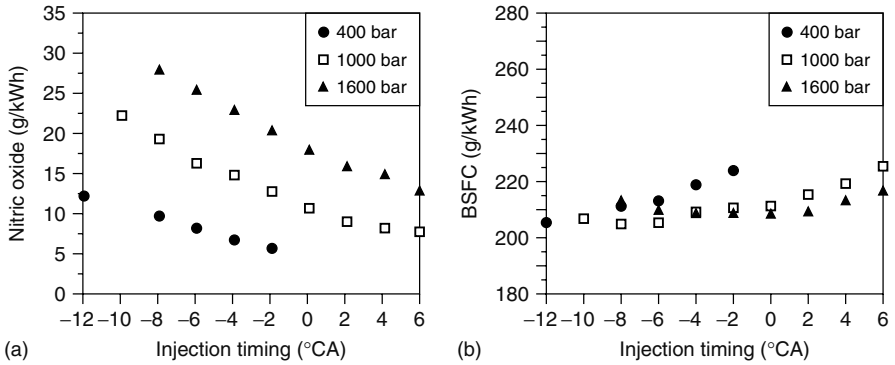


Figure 2. (a,b) Effect of injection pressure and injection timing on NOx and BSFC. (From Hountalas *et al.*, 2004. Copyright © 2004 SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

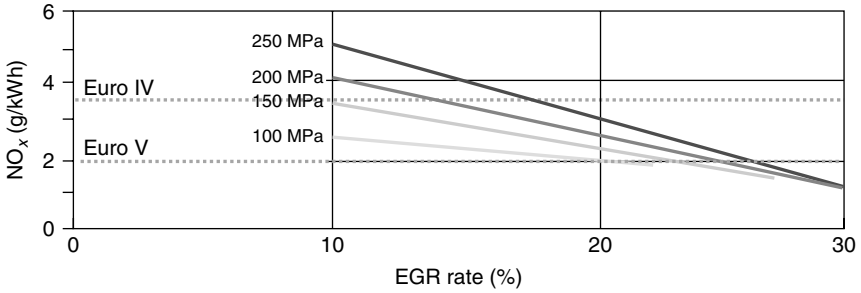


Figure 3. Relationship between EGR rate, injection pressure, and NO_X (Single cylinder, 2-L engine. Constant start of injection. BMEP 15 bar, $\lambda = 2.0$, 1500 rpm.)

have shown to increase. One reason for the increased NO_X emissions is the advanced combustion phasing with higher pressure—for *the same start of injection angle, fuel injection at a higher pressure shortens the combustion duration* and advances combustion phasing. One way of addressing this challenge is using cooled EGR because higher injection pressure can enable reductions in NO_X when combined with other control measures such as cooled EGR, which will be introduced in Section 3.2.1. EGR can provide the necessary NO_X reductions but will require higher fuel injection pressures to control the PM emissions and avoid significant BSFC penalties than NO_X abatement approaches using other technologies (Figure 3). As an example, it has been suggested that Euro VI heavy-duty standards can be achieved with 250 MPa pressure if NO_X aftertreatment is used—perhaps with lower levels of EGR—while 300 MPa would be required if only EGR is used to control NO_X (Jääskeläinen, 2010). (see Petrol Fuel Injection Systems and Diesel Fuel Injection Systems).

2.2.2 *Exhaust gas recirculation (EGR)*

EGR is a highly effective method for the internal combustion engines measure to lower NO_X emissions. A distinction is made between:

- *Internal EGR, which is determined by valve timing and residual gas, and*
- *External EGR, which is routed to the combustion chamber through additional lines and a control valve*.

2.2.2.1 Principle of operation. In general, the most important factor contributing to the NO_X reduction effect of EGR is a decrease in the peak temperature inside of the cylinder during combustion. The effects leading to the lower peak temperature can be broken down into (i) a dilution effect, (ii) a thermal effect, (iii) a chemical effect, and (iv) an added-mass effect (Ladommatos *et al*., 1996; Maiboom, Tauzia and Hétet, 2008).

- *Dilution Effect*. Decrease of inlet O_2 concentration, whose principal consequence is the deceleration of the mixing between O_2 and fuel resulting in the extension of flame region. Thus, the gas quantity that absorbs the heat release is increased, resulting in a lower flame temperature. As a result, one consequence of the dilution effect is the reduction of local temperatures that can also be considered as a thermal effect ("local" thermal effect). Another consequence of the dilution effect is the reduction of the oxygen partial pressure and its effect on kinetics of the elementary NO formation reactions.
- *Thermal Effect*. EGR contains significant amounts of water and carbon dioxides (CO_2), both of which have significantly higher specific heat capacities than air. The effect of increased specific heat capacity is the thermal effect. As already mentioned, intake air dilution with EGR can simultaneously introduce the dilution and thermal effects.
- *Chemical Effect*. The recirculated water vapor and CO_2 are dissociated during combustion, modifying the combustion process, and the NO_X formation. In particular, the endothermic dissociation of H_2O results in a decrease of the flame temperature.
- *Added-Mass Effect*. If adding a diluent to the intake charge results in an increased mass flow rate, an additional effect is introduced. This added flow has an additional heat capacity because of its mass. This is different from the thermal effect because of any specific heat capacity differences that may exist.

Figure 4 illustrates the effect of charge dilution on the oxygen mass fraction and the product of intake charge mass and heat capacity (mCp) as might occur in an engine with EGR. All effects may occur simultaneously making it difficult to ascertain which are most important. The dilution effect only accounts for the reduction in oxygen mass fraction; the thermal effect for differences

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

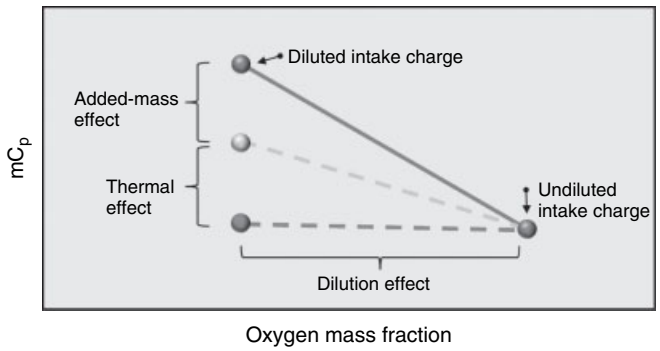


Figure 4. Conceptual view of EGR effect. (Based on data from Ladommatos *et al*., 1996.)

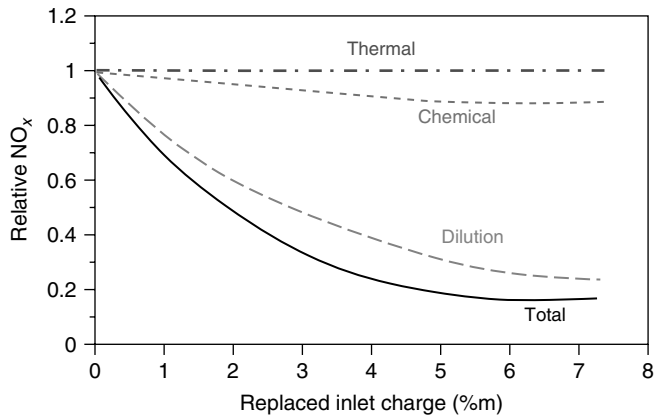


Figure 5. EGR effects with pure CO_2 charge dilution. (Based on data from Ladommatos *et al*., 1996.)

in average-specific heat and the added-mass effect for differences in intake charge mass. The chemical effect may also be present. Figure 5 shows the relative NO_X reductions for the thermal, chemical, and dilution effects in isolation and for the combined NO_X reduction for these three effects when charge dilution is made with pure CO_2. Intake charge and fuel mass flows were fixed, so there is no "added-mass" effect. It is apparent that most of the NO_X reduction from CO_2 is through the dilution effect with a small additional contribution due to a chemical effect. The thermal effect was found to be insignificant at dilution levels up to 7% even though CO_2 has a considerably higher specific heat capacity than air (1.24 and 1.16 kJ/kg at 1000 K, respectively). This is hardly surprising as adding 6% CO_2—the amount present with ∼50% EGR—to air increases the specific heat capacity by less than 0.5% (Khair, 2006).

2.2.2.2 External EGR. External EGR is usually distinguished as follows:

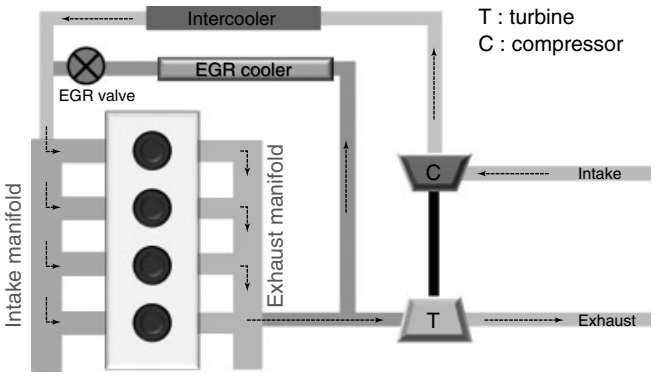


Figure 6. Schematic of HPL EGR.

– *High Pressure Loop EGR*. High pressure loop exhaust gas recirculation (HPL EGR) is generally used in light-duty vehicles. From Figure 6, it is implied that a pressure difference exists between the exhaust and intake manifold without which EGR could not flow from the former to the latter. In turbocharged heavy-duty diesel engines, it is often difficult to introduce EGR into the intake manifold. The problem is that intake manifold pressures are usually greater than exhaust system pressures. To circumvent this dilemma, exhaust is intercepted at a point upstream of the turbocharger where exhaust pressure is higher than that of the intake manifold. A portion of the exhaust flow is returned to the engine cylinders through an electronically controlled EGR valve after being cooled as shown in Figure 6. This approach is commonly referred to as *HPL EGR* (Khair, 2006).
– *Low Pressure Loop EGR*. Another type of EGR implementation in heavy-duty diesel engines is the low pressure loop system (LPL EGR). This system is often employed in conjunction with particulate filter-based aftertreatment systems, where several benefits may accrue. Rather than sourcing EGR from a preturbine location (as in the HPL EGR case), LPL EGR systems use exhaust that has been filtered through diesel particulate filters. Figure 7 shows a schematic representation of an LPL EGR system where EGR was recirculated from a point downstream from the DPF. This alternate configuration sought to preserve turbocharger performance by supplying exhaust gas from a point downstream of the trap (Figure 7), thus allowing all the exhaust to be utilized in the turbine. At this location, exhaust gas pressure is at a lower level than that of the intake manifold. To promote flow, EGR is introduced back in the engine just upstream of the turbocharger compressor. The pressure difference between points downstream of the trap and upstream of the turbocharger is generally adequate

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

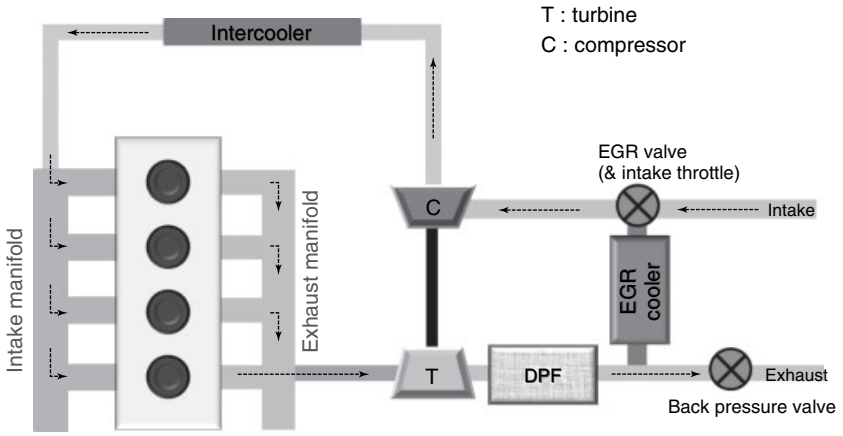


Figure 7. Schematic of LPL EGR.

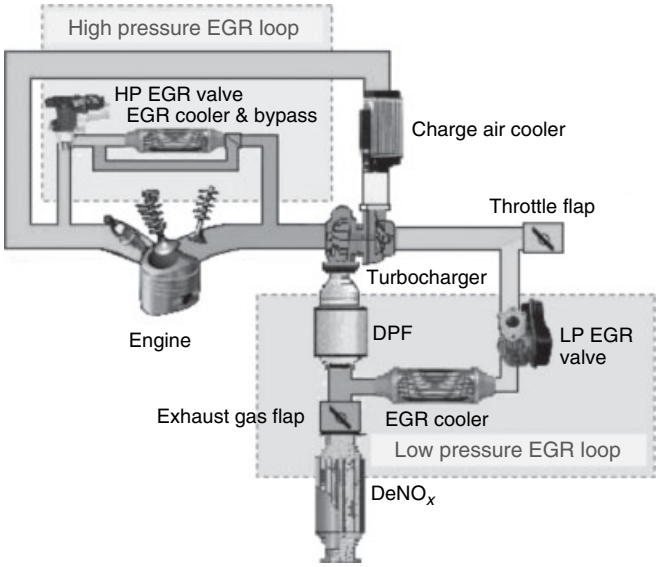


Figure 8. Schematic of dual-loop EGR.

for EGR flow rates needed to reduce US FTP NO_X to at least the 2.0 g/bhp-h level for heavy-duty diesel engines.

– *Dual-Loop EGR*. Some engines may be equipped with waste-gated turbochargers, where some of the exhaust gas bypasses the turbine at high engine speed conditions. Bypassing exhaust around turbochargers generally reduces exhaust manifold pressure as well as the pressure difference between exhaust and intake manifolds above peak torque speed. In such cases, the future types of EGR, a newly developed, dual-loop EGR system, which has combining features of HPL and LPL, have become a common option to consider as shown in Figure 8 (Park *et al*., 2010). In Figure 8, EGR is sourced from a point upstream of the turbine (preturbine) as in an HPL configuration and delivered precompressor as in an LPL configuration. While this system presents some of the same adverse features of the LPL EGR system, it provides an adequate pressure difference between the exhaust and intake manifolds. This allows EGR rates for substantial NO_X reduction without needing a pump or the application of excessive exhaust back pressure to drive EGR into the engine.

2.2.2.3 Internal EGR. An alternative way of achieving NO_X reduction through combustion products is performed through the residual exhaust gas. In theory, if it were possible to retain these combustion products in the cylinder until the next combustion cycle, then it would be possible to achieve NO_X reduction without the complication of an external EGR system with its many components and their control. The use of residual gas for NO_X reduction is commonly referred to as *internal EGR*, even though exhaust gas is not recirculated but generally retained in the cylinder. In some cases, exhaust gas products may be returned to the cylinder through the exhaust valve actuation. Therefore, the term *internal EGR* could still be correctly employed as the acronym for exhaust gas retained as well as exhaust gas returned. An example of the exhaust gas returned is pulse EGR. In the pulse EGR system, the exhaust valve is reopened during the intake stroke by means of a modified exhaust valve cam lobe design. This modified design features a second lobe on the valve cam referred to as a *sublift lobe* as shown in Figure 9. As the piston moves from top dead center to bottom dead center during the intake stroke, the sublift cam lobe lifts the exhaust valve off of its seat and allows higher pressure exhaust to return to the cylinder. Of course, the timing and the degree of valve lift as well as the pressure difference across the exhaust valve

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

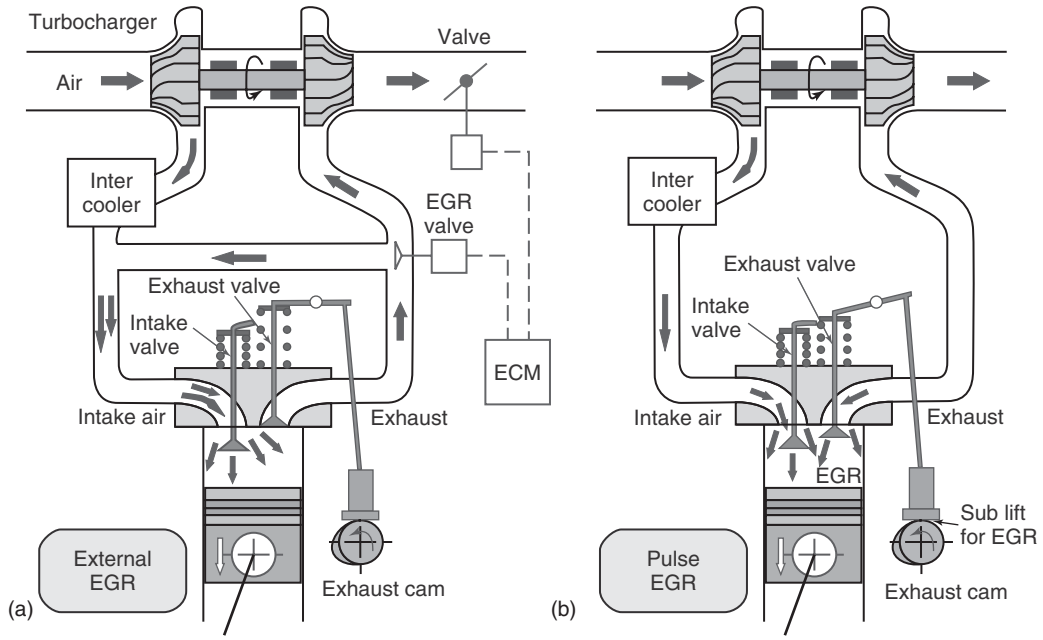


Figure 9. (a,b) External EGR(HPL) versus internal(pulse) EGR. (Reproduced with permission from Khair, 2006. Source: Diesel Progress. (c) EcoPoint.)

are extremely important parameters in controlling the rate of the exhaust gas returned.

An added modification over the pulse EGR system can be achieved if the function of the sublift lobe can be made to be more flexible and controllable. This flexibility can be provided through variable valve actuation (VVA). VVA mechanisms generally allow the control of the timing of the valve lift, the rate of the valve lift, the valve lift, and the number of valve openings. Control of these parameters may be limited at times and have limited impact on performance and emissions depending on a complex array of technical conditions. For instance, the closeness of the piston crown to the valve face may limit either the valve lift timing or the total valve lift itself. Therefore, system optimization is usually performed to obtain the best results within the overall physical and technical limitations of the system (Khair, 2006).

2.2.3 Charge air cooling

In modern engines, it is also important to ensure that the temperature of the charge does not become excessive. In modern boosted engines, this is a real possibility. Excessive temperatures can lead to reduced charge density and higher combustion temperatures, which can affect NO_X emissions. While turbochargers and superchargers increase charge air density, they also increase the temperature of the air in the intake manifold. As emission standards became increasingly stringent, additional increases in charge air density were needed. While this could be achieved through compression to higher pressures, this would require more expensive compression equipment and would further increase cycle temperatures. On the other hand, if intake manifold temperature could be reduced, the intake density could be further increased and more air could be supplied to the engine without necessarily increasing the intake manifold pressure. Cooling the air with a heat exchanger as it leaves the compressor is a common way to achieve this charge air cooling. Such a heat exchanger is referred to as a charge air cooler (CAC), intercooler, or aftercooler. Increasing demand for improvements in fuel economy and exhaust emissions has made the CAC an important component of most modern turbocharged engines. By controlling intake air temperature, up to 75% of NO_X reduction can be achieved by combining with other de-NO_X technologies (Jääskeläinen and Khair, 2011). (see Turbocharging and Diesel and Diesel LTC Combustion).

2.3 Aftertreatment systems in SI engine

2.3.1 Three-way catalytic converter (TWC)

A catalytic converter is a device used to convert toxic exhaust emissions into nontoxic substances. Inside a catalytic converter, a catalyst stimulates a chemical reaction in which noxious by-products of combustion

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

undergo a chemical reaction. The type of chemical reaction varies depending on the type of catalyst installed. In SI engine, three-way catalyst is applied to reduce not only carbon monoxide (CO), unburned hydrocarbons (HC), but also NO_X.

2.3.1.1 TWC construction. The three-way catalyst is created by coating the internal converter substrate with the following key materials:

- *Platinum/Palladium*: Oxidizing catalysts for HC and CO
- *Rhodium*: Reducing catalyst for NO_X
- *Cerium*: Promotes oxygen storage to improve oxidation efficiency.

2.3.1.2 TWC operation. As exhaust gases flow through the converter passageways, they come in contact with the coated surface that initiates the catalytic process. As exhaust and catalyst temperatures rise, the following reaction occurs:

- Oxides of nitrogen(NO_X) are reduced into simple nitrogen (N_2) and carbon dioxide (CO_2).
- Hydrocarbons (HC) and carbon monoxide (CO) are oxidized to create water (H_2O) and carbon dioxide (CO_2).

Three-way catalysts are effective when the engine is operated within a narrow band of air–fuel ratios near stoichiometry, such that the exhaust gas oscillates between rich (excess fuel) and lean (excess oxygen) conditions. However, conversion efficiency falls very rapidly when the engine is operated outside of that band of air–fuel ratios as shown in Figure 10. Under lean engine operation, there is excess oxygen and the reduction of NO_X is not favored. Under rich conditions, the excess fuel consumes all of the available oxygen before the catalyst, thus only stored oxygen is available for the oxidation function. Closed-loop control systems are necessary because of the conflicting requirements for effective NO_X reduction.

2.3.1.3 Oxygen storage. Three-way catalytic converters (TWCs) can store oxygen from the exhaust gas stream, usually when the air–fuel ratio goes lean. When insufficient oxygen is available from the exhaust stream, the stored oxygen is released and consumed. A lack of sufficient oxygen occurs either when oxygen derived from NO_X reduction is unavailable or when certain maneuvers such as hard acceleration enrich the mixture beyond the ability of the converter to supply oxygen. (see Stoichiometric Exhaust Emission Control).

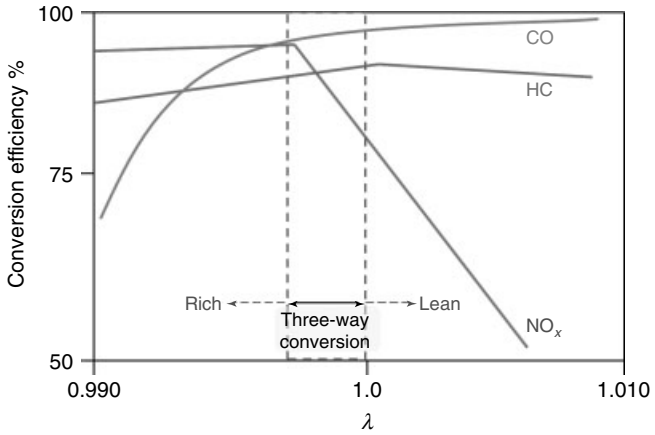


Figure 10. Lambda region for control of three-way catalyst.

2.3.2 *NO_X adsorber for lean burn spark ignition engine*

The concept of the NO_X adsorber/catalyst has been developed based on acid–base wash coat chemistry. It involves storage of NO_X on the catalyst wash coat during lean exhaust conditions and release during rich operation. The released NO_X is catalytically converted to nitrogen, in a process similar to that occurring over three-way catalysts (TWC) widely used in stoichiometric gasoline engines. Normally, three-way catalysts are inactive in converting NO_X under lean exhaust conditions, when oxygen is present in the exhaust gas. By alternating the lean storage and rich release-and-conversion phases, the applicability of the three-way catalyst has been extended to lean burn engines. The technology was first commercialized on gasoline direct injected (GDI) engines. More detailed explanations for NO_X absorber will be introduced in Section 2.4.2.

2.4 Aftertreatment systems in CI engine

2.4.1 *Selective catalytic reduction (SCR)*

SCR system is an aftertreatment system that reduces the emitted NO_X after combustion in CI engines. SI engines can be an adopted TWC system. SI engine is operated at all times with an air/fuel ratio at or close to stoichiometry in which there exist enough reducing gases for NO reduction. However, CI system is different from SI engines. CI engines operate in an air index range between $\lambda = 4$ at idle and $\lambda = 1.2$ at full load that makes NO_X reduction by TWC impossible. Hence, for controlling NO_X for diesel engines and lean-burn gasoline engines, special treatment or SCR of NO_X is necessary.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.4.1.1 Reductants. Several reductants are currently used in SCR applications including anhydrous ammonia, aqueous ammonia, urea, or hydrocarbon. All those reductants are widely available in large quantities. Pure anhydrous ammonia is extremely toxic and difficult to store safely but needs no further conversion to operate within an SCR. It is typically favored by large industrial SCR operators. Aqueous ammonia must be hydrolyzed in order to be used, but it is substantially safer to store and transport than anhydrous ammonia. Urea is the safest to store but requires conversion to ammonia through thermal decomposition in order to be used as an effective reductant. Hydrocarbon such as diesel fuel is used as reducing agent that can be added into exhaust stream via in-cylinder postinjection or in-exhaust secondary injection. This is the most attractive feature of HC-SCR because it can make exhaust systems simple and inexpensive.

2.4.1.2 Ammonia-SCR. Two forms of ammonia may be in SCR systems: (i) pure anhydrous ammonia and (ii) aqueous ammonia. Anhydrous ammonia is toxic, hazardous, and requires thick-shell, pressurized storage tanks and piping because of its high vapor pressure. Aqueous ammonia, NH_3-H_2O, is less hazardous and easier to handle. A typical industrial grade ammonia, containing about 27% ammonia and 73% water by weight, has nearly atmospheric vapor pressure at normal temperatures and can be safely transported on highways.

A number of chemical reactions occur in the ammonia-SCR system, as expressed by Equations 1–5. All of these processes represent desirable reactions that reduce NO_X to elemental nitrogen. Equation 2 represents the dominant reaction mechanism (Cho, 1994). Reactions given by Equation 3 through 5 involve nitrogen dioxide reactant. The reaction path described by Equation 5 is very fast. This reaction is responsible for the promotion of low temperature SCR by NO_X. Normally, NO_2 concentrations in most flue gases, including diesel exhaust, are low. In some diesel SCR systems, NO_2 levels are purposely increased to enhance NO_X conversion at low temperatures.

$$6NO + 4NH_3 \rightarrow 5N_2 + 6H_2O \quad (1)$$

$$4NO + 4NH_3 + O_2 \rightarrow 4N_2 + 6H_2O \quad (2)$$

$$6NO_2 + 8NH_3 \rightarrow 7N_2 + 12H_2O \quad (3)$$

$$2NO_2 + 4NH_3 + O_2 \rightarrow 3N_2 + 6H_2O \quad (4)$$

$$NO + NO_2 + 2NH_3 \rightarrow 2N_2 + 3H_2O \quad (5)$$

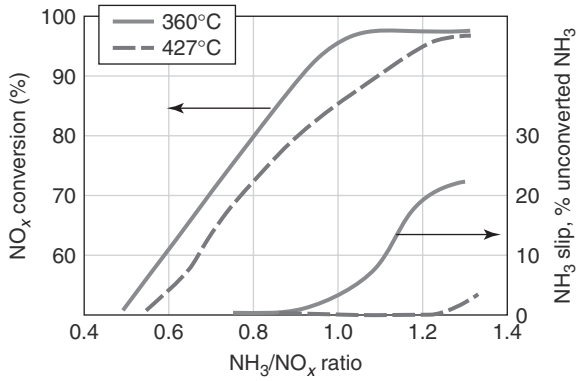


Figure 11. NO_X conversion and Ammonia slip for different NO_3/NO_X ratios. (Reproduced with permission from Majewski, 2005. Source: Dieselnet.com © Ecopoint Inc., 2005.)

The SCR process requires precise control of the ammonia injection rate. An insufficient injection may result in unacceptably low NO_X conversions. An injection rate that is too high results in release of undesirable ammonia to the atmosphere. These ammonia emissions from SCR systems are known as *ammonia slip*. The ammonia slip increases at higher NH_3/NO_X ratios. According to the dominant SCR reaction (Equation 2, the stoichiometric NH_3/NO_X ratio in the SCR system is about 1. Ratios higher than 1 significantly increase the ammonia slip. In practice, ratios between 0.9 and 1 are used, which minimize the ammonia slip while still providing satisfactory NO_X conversions. Figure 11 shows an example relationship between the NH_3/NO_X ratio, NO_X conversion, temperature, and ammonia slip (Heck, Farrauto and Gulati, 2009). The ammonia slip decreases with increasing temperature, whereas the NO_X conversion in an SCR catalyst may either increase or decrease with temperature, depending on the particular temperature range and catalyst system.

2.4.1.3 Urea-SCR. Owing to the toxicity and handling problems with ammonia, there has been a need for more convenient SCR reductants. From the technical point of view, the alternative reductant has to easily and completely decompose to ammonia, producing no harmful by-products, under the conditions in the SCR reactor. From the commercial perspective, the perfect reductant would be nontoxic, easy to transport and handle, inexpensive, and commonly available. Urea, $CO(NH_2)_2$, which meets the criteria of nontoxicity and safety and is commercially available, became the reductant of choice for use in mobile SCR applications. In the SCR process, water solutions of urea are injected into the process gas stream and evaporated, followed by decomposition of urea. The overall process of urea decomposition is often described by the following

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

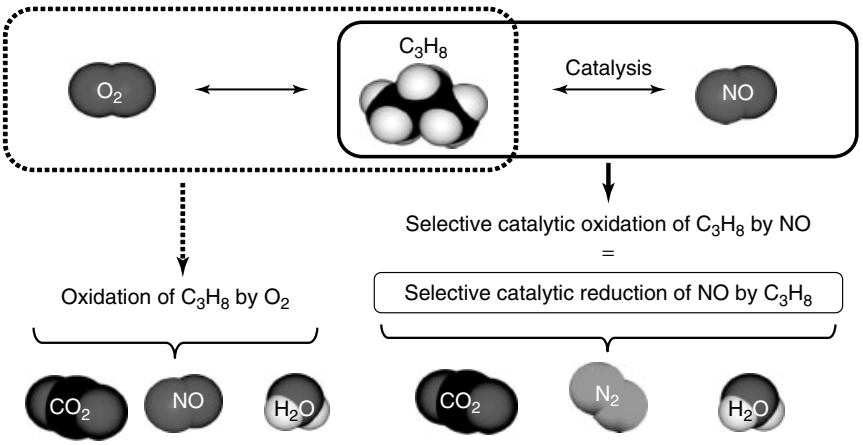


Figure 12. A schematic of selective catalytic reduction of NO with hydrocarbons. Hydrocarbons are selectively oxidized by NO.

hydrolysis reaction (Equation 6) (Forzatti, 2001):

$$CO(NH_2)_2 + H_2O \rightarrow CO_2 + 2NH_3 \quad (6)$$

In practice, however, the decomposition of urea proceeds through two separate reaction steps, involving an isocyanic acid (HNCO) intermediate. In the first step, HNCO and one molecule of ammonia are formed by thermolysis of urea, followed by hydrolysis of the HNCO with the formation of a second NH_3 molecule:

$$CO(NH_2)_2 \rightarrow NH_3 + HNCO \quad (7)$$

$$HNCO + H_2O \rightarrow NH_3 + CO_2 \quad (8)$$

While urea starts to decompose already at around 160°C, the decomposition cannot reach completion in the gas phase at temperatures typical for diesel exhaust and at the residence time in SCR systems. It is believed that only up to about 20% of the urea decomposes to HNCO and NH_3 in the gas phase at 330°C and only about 50% decomposes at 400°C (Kleemann *et al*., 2000). The remaining urea decomposes only after reaching the surface of the catalyst. It was also shown that HNCO is very stable in the gas phase, requiring an oxide surface to catalyze its decomposition to NH_3. In effect, most of the urea decomposition—especially at low temperature conditions—will occur on the catalyst surface, rather than in the gas phase. In laboratory bench tests, the effects of urea decomposition were visible in the inlet section of the SCR catalyst (Sluder, Storey and Lewis, 2005).

2.4.1.4 Hydrocarbon-selective catalytic reduction. In an SCR of NO_X with hydrocarbon reaction, small amount of hydrocarbon is added to the exhaust gas via in-cylinder postinjection or in-exhaust secondary injection. Injected hydrocarbons act as reducing agents for the NO_X reduction reaction. In a diesel exhaust, oxygen exists in a few percentage by volume, whereas NO_X in a few hundred parts per million. Therefore, supplied hydrocarbons may tend to react with oxygen rather than nitrogen oxides. Then, using the appropriate catalyst system, oxidation of hydrocarbons by NO_X can be achieved. This reaction is very selective, because extremely small amount of NO_X oxidizes hydrocarbons instead of oxygen that exists in a few percentages in a diesel exhaust. Next, Figure 12 illustrates the overall reaction of selective of unselective oxidation of propane with O_2 and NO (Lee, 2007).

2.4.1.5 Catalysts. SCR of NO_X with ammonia was first discovered over a platinum catalyst. The Pt technology can be used only at low temperatures (<250°C), because of its poor selectivity for NO_X reduction at higher temperatures. Two groups of base metal SCR catalysts—vanadia and zeolite based—were later developed, which can operate at higher temperatures and have wider temperature windows, as illustrated in Figure 13.

Platinum catalysts lose their NO_X reduction activity above approximately 250°C. A V_2O_5/Al_2O_3 catalyst was used first for higher temperature applications. However, its use was limited to sulfur-free exhaust gases because the alumina reacted with SO_3 to form $Al_2(SO_4)_3$, resulting in catalyst deactivation. To solve this problem, a nonsulfating TiO_2 carrier was used for the V_2O_5, which then became the catalyst of choice. These catalysts functioned at higher temperatures and over a broader temperature range than Pt. Other base metal oxides, such as tungsten trioxide (WO_3) and molybdenum trioxide (MoO_3), are often added to V_2O_5 as promoters to further decrease the SO_3 formation and to extend catalyst life. The upper temperature limit of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

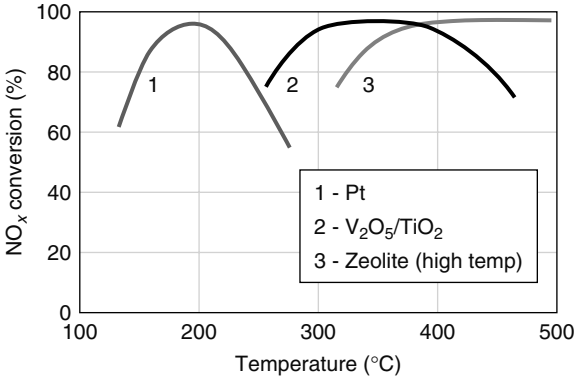


Figure 13. Operating temperature windows for different SCR catalysts. (Reproduced with permission from Majewski, 2005. Source: Dieselnet.com © Ecopoint Inc., 2005.)

Table 1. SCR catalyst technologies.

Catalyst	Temperature Range (°C)
Platinum (Pt)	175–250
Vanadium (V_2O_5)	300–450
Zeolite (high temperature)	350–600
Zeolite (low temperature)	150–450

vanadia catalysts—about 450°C—is still insufficient for certain hot gas applications, such as gas-fired cogeneration plants. Zeolite-based catalysts have been developed and commercialized in the 1990s that function at higher temperatures. Finally, ion-exchanged zeolites of greatly improved low temperature activity (at the expense of a reduced upper temperature limit) have been developed for mobile applications. The operating temperature ranges for different SCR catalyst technologies are shown in Table 1. These temperature ranges should be considered as being approximate. Catalysts are under development, especially for mobile SCR applications, which are characterized by increasingly wider temperature windows (Figure 14).

For HC-SCR, the first efficient catalyst was ion-exchanged Cu/ZSM-5, but the catalyst has the major problem reported as the hydrothermal stability and deactivation by a prolonged exposure to temperatures of the order of 800°C. Therefore, many researchers concentrated on metal oxide catalyst. The priority was given to PGM (platinum group metal) catalysts. In addition, base metal/metal oxide catalysts have been studied extensively with the purpose of making a durable and N_2 selective lean-DeNO_X catalyst. Typically, copper (Cu), cobalt (Co), nickel (Ni), iron (Fe), and silver (Ag) were supported on alumina (Al_2O_3), titania (TiO_2), or zirconia (ZrO_2) (LEE, 2007).

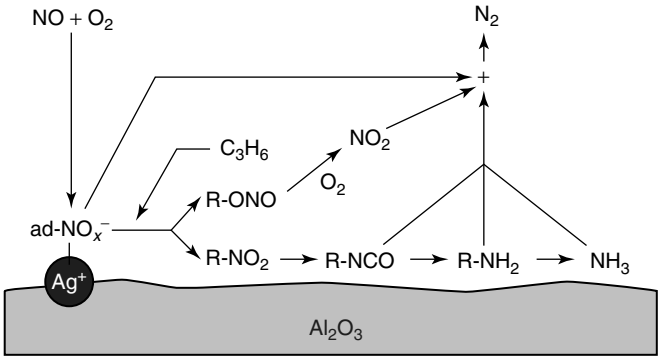


Figure 14. Proposed reaction mechanism of C_3H_6-SCR over low loading Ag/Al_2O_3 catalyst.

2.4.2 NO_X adsorber

2.4.2.1 Principle of operation. NO_X adsorber technology reduces NO_X via cyclic operation from which an example of one cycle is illustrated in Figure 15. The catalyst traps NO_X when the engine is run in a lean-burn mode. The nearly complete removal of NO_X for some time period (almost 2 min) as shown in Figure 15 is one of the primary characteristics that makes NO_X adsorber catalysts a leading candidate for lean-burn aftertreatment systems. The catalyst has a certain NO_X trapping capacity; and once this capacity approaches some level of saturation, unacceptable amounts of NO_X will slip through the catalyst. At or before this point, the catalyst is exposed to a rich environment or exhaust, which induces NO_X release from the surface and reduction to N_2. In the experiment described by Figure 15, this rich period begins at 10 min into the cycle. The rich event also results in the recovery of some or all of the original trapping capacity of the surface (Epling *et al*., 2004; Majewski, 2007).

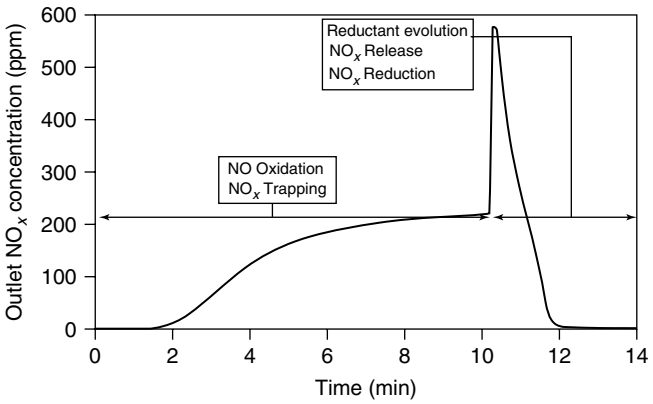


Figure 15. NO_X release profile in NO_X adsorber.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.4.2.2 Chemical reactions. The overall chemical reactions of catalyst operation during the cycle can be described by these five reaction steps:

- *Step 1*. NO oxidation to nitrogen dioxides (NO_2),
- *Step 2*. NO or NO_2 adsorption on the surface in the form of nitrites or nitrates,
- *Step 3*. Reductant evolution when the exhaust is switched to rich conditions,
- *Step 4*. NO_X release from the nitrite or nitrate sites, and
- *Step 5*. NO_X reduction to N_2.

Such operations of the catalyst require a combination of strong basic properties to accomplish storage of NO and NO_2 (step 2) and redox catalytic properties to facilitate all of the steps. Thus, NO_X adsorber catalysts are typically composed of at least one alkali- or alkaline-earth component and at least one precious-metal component, which are supported on a high surface area refractory oxide.

The NO_X adsorption/reduction mechanism is illustrated in Figure 16 (MECA 2000). The catalyst wash coat combines three active components: (i) an oxidation catalyst, for example, Pt, (ii) an adsorbent, for example, barium oxide (BaO), and (iii) a reduction catalyst, for example, Rh.

- *NO Oxidation*. NO_X emissions from the diesel engine are composed mostly of nitric oxide, NO, but most NO_X trapping materials more effectively adsorb NO_2 compared to NO, or NO_2 may even be the required intermediate compound for NO_X adsorption. In the first step, described by Equation 9, nitric oxide reacts with oxygen on active oxidation catalyst sites (Pt) to form NO_2.

$$NO + {}^1\!/_2O_2 \rightarrow NO_2 \tag{9}$$

- *NO_2 or NO Adsorption*. Extensive NO_X accumulation occurs on the catalyst surface, owing to NO_X adsorption in the form of nitrates or nitrites with the formation of ionic bonds. Equation 10 represents adsorption of NO_2 by the storage material in the form of barium nitrate.

$$BaO + NO_2 + {}^1\!/_2O_2 \rightarrow Ba(NO_3)_2 \tag{10}$$

- *Reductant Evolution*. Once exhaust is switched to the rich condition, oxygen is replaced by reducing species, including hydrocarbons, carbon monoxide, and hydrogen.
- *NO_X Release from the Nitrite or Nitrate Sites*. When the engine runs under excessive fuel conditions or at elevated temperatures, the nitrate species become thermodynamically unstable and decompose, producing NO or NO_2, according to Equations 11 and 12 (Brogan, Clark and Brisley,1998; Erkfeld *et al*., 1999).

$$Ba(NO_3)_2 \rightarrow BaO + 2NO + {}^1\!/_2O_2 \tag{11}$$

$$Ba(NO_3)_2 \rightarrow BaO + 2NO_2 + {}^1\!/_2O_2 \tag{12}$$

- *NO_X Reduction to Nitrogen*. Under rich conditions, the nitrogen oxides are reduced by HC/CO/H_2 to N_2 over the reduction catalyst, in a conventional three-way catalyst process. One of the possible reduction paths is described by Equation 13.

$$NO + CO \rightarrow {}^1\!/_2N_2 + CO_2 \tag{13}$$

The above set of reactions allows for an understanding of the basic NO_X adsorber chemistry, but the actual chemical and physical processes are more complex, and have yet to be fully explained. A more detailed analysis should also include other reaction paths and species, for example, barium carbonate and barium hydroxide that coexist with barium oxide on the catalyst surface.

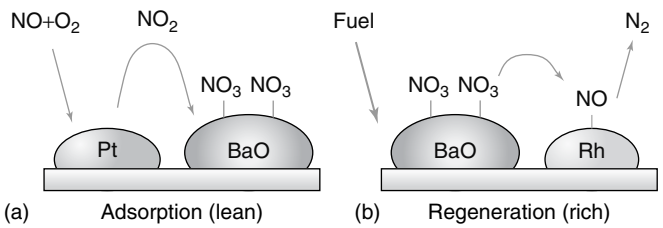


Figure 16. (a,b) NO_X adsorption and reduction mechanism. (Adapted from Manufacturers of Emission Controls Association, MECA 2000.)

2.4.2.3 Regeneration. During the adsorption cycle, the adsorber is gradually converted into its nitrate form (e.g., barium nitrate) and the adsorption capacity becomes saturated. At this time, the stored NO_X needs to be released and catalytically reduced in a process called the *regeneration*. At lean exhaust conditions, NO_X is released from barium sites at temperatures above 450–500°C. The regeneration occurs at much lower temperatures if a short pulse of fuel-rich mixture is provided. NO_X adsorbers can fully regenerate at 250°C, with the onset of a partial regeneration at temperatures as low as 150°C, if the air-to-fuel ratio is maintained at $\lambda < 1$. Therefore, the operation of the adsorber catalyst system involves continuous cycling through lean and rich fuel condition.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3 APPLICATIONS

3.1 SI engine

NO_X control in SI engine is affected by TWC control. Thus, closed-loop control has been proposed to maintain efficient operation of TWC.

3.1.1 Closed-loop control of TWC

The closed-loop control system for TWC was designed to rapidly alternate the A/F ratio slightly rich and then slightly lean of stoichiometry. Three-way catalysts operate in a closed-loop system including a lambda sensor (also called *oxygen sensor*) to regulate the A/F ratio. The closed-loop control works as follows:

- *When the A/F ratio is leaner than stoichiometry*, the oxygen concentration of the exhaust stream rises and the carbon monoxide concentration falls. This provides a highly efficient operating environment for the oxidizing catalysts (platinum and palladium). During this lean cycle, the catalyst (containing cerium) also stores excess oxygen that will be released to promote better oxidation during the rich cycle.
- *When the A/F ratio is richer than stoichiometry*, the carbon monoxide concentration of the exhaust gas increases and the oxygen concentration decreases. This results in very efficient operating condition for the reducing catalyst (such as rhodium). The oxidizing catalyst maintains its efficiency as stored oxygen is released.

3.1.2 Catalysts

While the original three-way catalyst used the combination of Pt and Rh, typically in a 5 : 1 ratio, further development has led to a multitude of formulas, including Pt/Pd/Rh, Pd/Rh, Pd only, etc.

3.2 CI engine

3.2.1 Exhaust gas recirculation

3.2.1.1 Detroit diesel corporation's heavy-duty engine. The typical cooled HPL EGR system was designed for the DDC Series 60. The engine was equipped with a variable geometry turbocharger that was used in the control of the EGR system. The EGR employed for this engine setup was a high pressure loop system where a portion of the exhaust was taken from upstream of the turbocharger. The EGR then flew through an EGR control valve responsible for controlling the EGR rate. Pressure taps upstream and downstream of the EGR control valve monitored the pressure difference across the EGR valve and were connected to a transducer providing a feedback signal monitoring the EGR rate. EGR proceeds through an EGR cooler with water from the engine jacket cooling water system and flows through an EGR pipe to the intake manifold where it mixes with cooled charge air before being inducted into the engine.

3.2.1.2 AUDI's light-duty engine. The discussion regarding EGR is not limited to heavy-duty engines but extends to engines powering light-duty vehicles as well. One application that is representative of other production-style applications is shown in Figure 17, which is a schematic representation of a passenger car type EGR system. The EGR system is a high pressure loop, cooled EGR configuration. A portion of the exhaust is channeled through an EGR control valve and proceeds to the EGR cooler. From the cooler EGR flows to a throttle valve assembly where it is mixed with filtered, high pressure, fresh combustion air that had been cooled through an intercooler to recover some of its density. The mixture of air and EGR is then inducted into the engine through

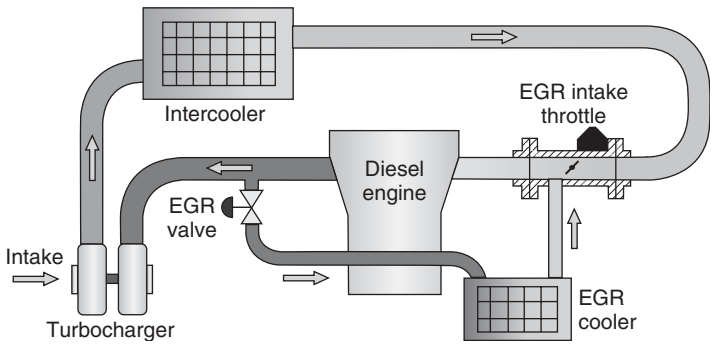


Figure 17. Schematic representation of a high-speed passenger car EGR/intake throttle system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

its intake manifold. The use of intake throttling is needed to increase EGR rate for purposes of achieving the target NO_X reduction.

3.2.1.3 STT Emtec's retrofit EGR system. The EGR technology has expanded its penetration beyond the light- and heavy-duty original equipment manufacturers into the retrofit market, mainly for heavy-duty applications. Retrofit kits—typically utilizing low pressure loop EGR—have been developed, which can provide up to about 40% NO_X reduction capability. In some applications, it is possible to achieve even greater NO_X reductions with a slight degradation in fuel economy.

A schematic of the STT Emtec DNO_X system—a retrofit EGR kit that has been used in many heavy-duty diesel engine applications in Europe and in the USA—is shown in Figure 18.

3.2.2 Selective catalytic reduction

3.2.2.1 BOSCH GmbH's Denoxtronic. Commercial deployment of urea-SCR systems depends on the development of not only the catalyst but also the urea injection system. The schematic of a urea system developed by Bosch is shown in Figure 19 known as *DENOXTRONIC 2.2*. The *DENOXTRONIC 2.2* has been designed for use in the heavy- and medium-duty segments. An integrated pump in the supply module of the DENOXTRONIC feeds the reducing agent AdBlue/DEF (a 32.5% solution of urea in water) from the tank to the dosing module. This injects the AdBlue/DEF directly into the flow of exhaust gas upstream of the SCR catalytic converter. The urea is then converted via hydrolysis into ammonia required for the further reaction. In the second step, inside the SCR catalytic converter, the ammonia reduces the exhaust gas nitrogen oxides into harmless water and nitrogen. The electronic control can be integrated into the engine control unit (ECU) or, alternatively, into a dosing control unit (DCU). The control unit matches the amount of AdBlue/DEF precisely to the relevant engine parameters. In this way, it is possible to reduce NO_X emissions by up to 95%.

3.2.2.2 Nissan's urea-SCR. A schematic of the commercial urea-SCR system launched in Japan by Nissan Diesel is shown in Figure 20. The system was installed on a 13-L, 6-cylinder and turbocharged diesel engine rated 280 kW@1800 rpm. The engine meets the JP 2005 NO_X limit of 2 g/kWh through the SCR aftertreatment. NO_X reduction efficiency in the SCR system was reported at about 70% (average from various conditions during field testing). The SCR catalyst system includes a preoxidation catalyst, a zeolite-based SCR catalyst, and an NH_3 slip catalyst. The urea dosing system was supplied by Bosch. This urea metering unit can provide minimum NH_3 slip and optimum time for NO_X reduction. To enable operation in cold climate conditions, system components incorporated electric heaters, controlled based on the temperature of the urea solution and ambient air. The urea tank was also

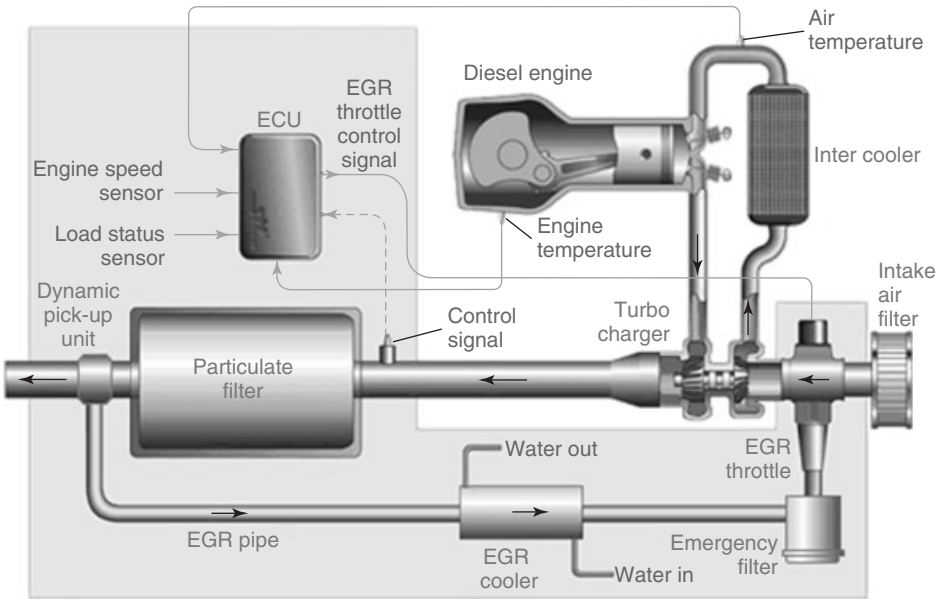


Figure 18. Schematic of STT Emtec EGR/throttle system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

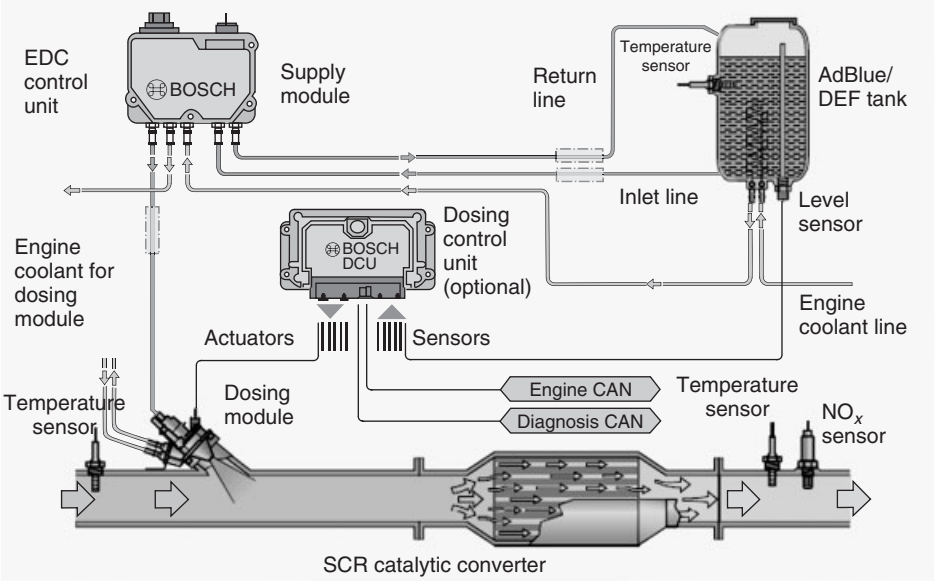


Figure 19. DENOXTRONIC: BOSCH GmbH.

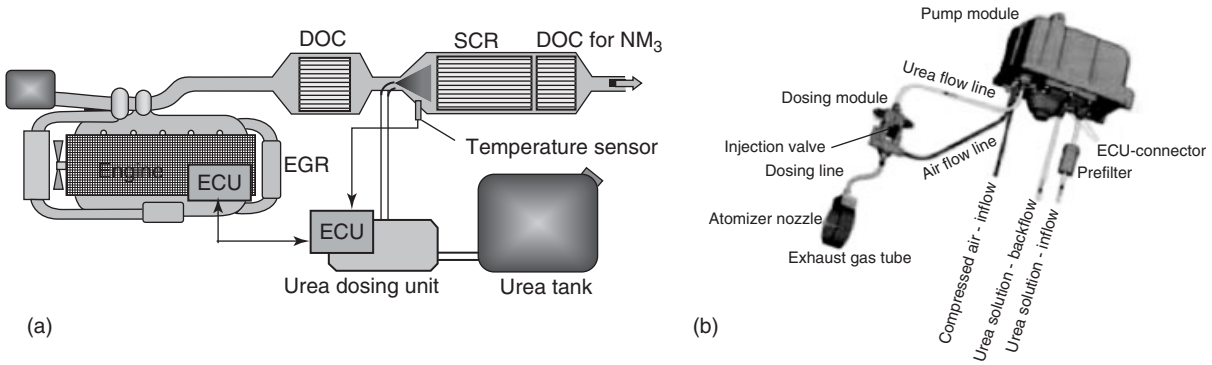


Figure 20. (a,b) Schematic of the urea-SCR system, NISSAN. (From Hirata et al., 2005. Copyright © 2005 SAE International. Reprinted with permission.)

equipped with an engine coolant heater (Hirata *et al*., 2005).

3.2.3 NO_X adsorber

3.2.3.1 US Environmental Protection Agency (EPA). One of the first NO_X adsorber demonstrations on a medium heavy-duty engine was presented by the US EPA with dual flow path exhaust emission control system as shown in Figure 21. NO_X reduction efficiency was 84% over the SET (supplemental emission test) at an equivalent 818,000 miles of engine operation with the test system. The data shows that improvements continue to be made in the area of NO_X adsorbers with respect to overall performance and thermal durability (Laroo and Schenk, 2007).

3.2.3.2 Johnson Mattey Inc. A NO_X adsorber catalysts and a Cu-based SCR catalyst were developed specifically for combined NAC+SCR applications by Johnson Mattey Inc. A 3.0L 2007 Mercedes E320 Bluetec passenger vehicle equipped with an NAC+SCR emission control system was used to evaluate the performance of the NAC and SCR components developed. The catalyst configuration and catalyst volume of the emission control system are containing a 0.6-L diesel oxidation catalyst (DOC), a 3.2-L NO_X NAC, a 3.5-L catalyzed soot filter (CSF), and a 2.5-L SCR. The newly developed Cu SCR catalyst exhibited exceptional stability against high temperature lean/rich aging and excellent low temperature NH_3-SCR activity. When tested as a combined system on a vehicle, the improved catalysts achieved high NO_X reduction efficiency

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

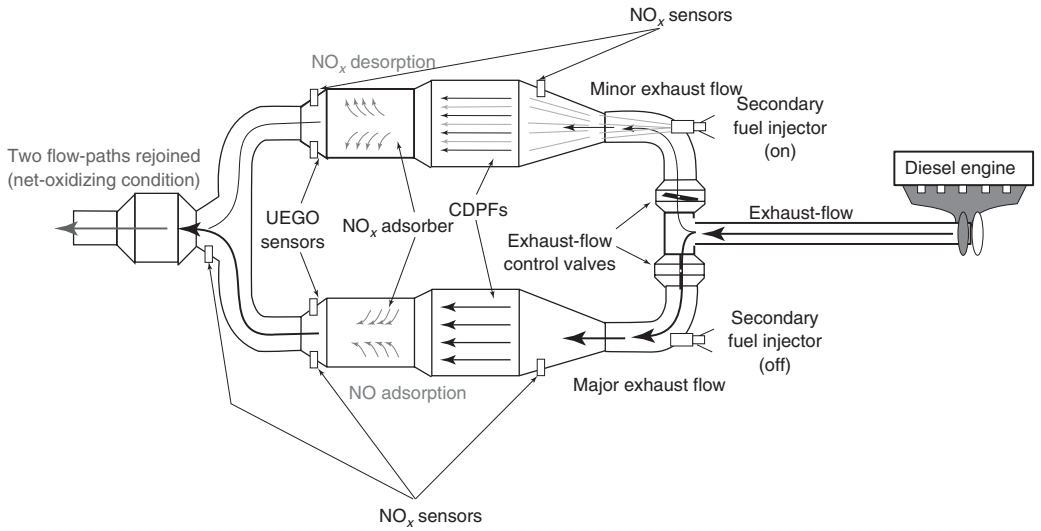


Figure 21. Schematic representation of the layout and functioning of the exhaust emission control system.

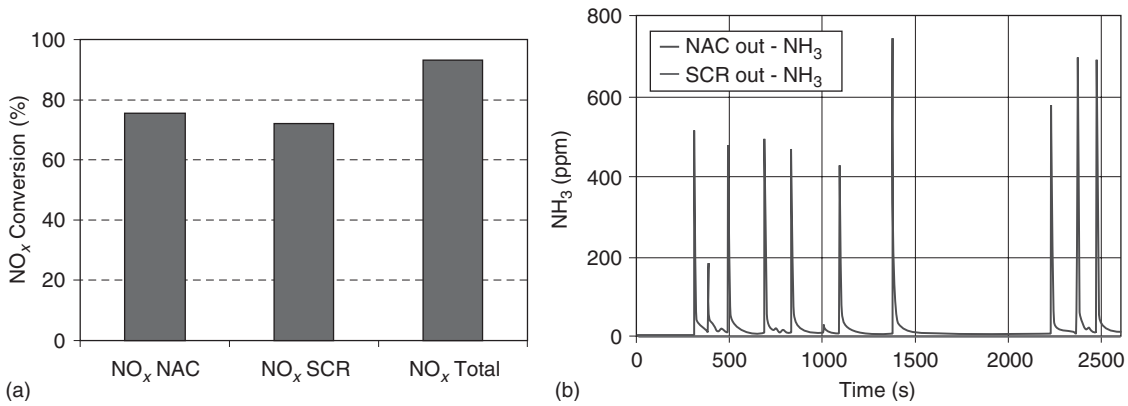


Figure 22. (a) NO_X conversion across the individual NAC and SCR components, and the conversion efficiency of the overall system during FTP. (Adapted from Chen *et al*., 2010. Copyright © 2010 SAE International. Reprinted with permission.) (b) The typical NH_3 concentration measured from the outlet of the NAC component or from the tailpipe during an FTP test. (Adapted from Chen *et al*., 2010. Copyright © 2010 SAE International. Reprinted with permission.)

up to 93% during an FTP cycle as shown in Figure 22a. Furthermore, significantly improved NH_3 generation was achieved on the new NAC component by the reduction of its oxygen storage capacity and the incorporation of a Pd component from Figure 22b. It was clear that the SCR component had sufficient NH_3 storage capacity to store this NH_3 and utilize it for further NO_X reduction (Chen *et al*., 2010).

REFERENCES

Brogan, M., Clark, A.D. and Brisley, R.J. (1998) Recent progress in NO_X trap technology. *SAE International Congress and Exposition*, Detroit, USA 26-26 February 1998. Warrendale, Pennsylvania: SAE International.

Chen, H. Y., Weigert, E. C., Fedeyko, J. M., *et al*. (2010) Advanced catalysts for combined (NAC + SCR) emission control systems, *SAE World Congress*, Detroit, USA 13-15 April 2010. Warrendale, Pennsylvania: SAE International.

Cho, S.M. (1994) Properly apply selective catalytic reduction for NO_X removal. *Chemical Engineering Progress*, **90**(1), 39–45.

Epling, W.S., Campbell, L.E., Yezerets, A., *et al*. (2004) Overview of the fundamental reactions and degradation mechanisms of NO_X storage/reduction catalysts. *Catalysis Reviews*, **46**(2), 163–245.

Erkfeld, S., Larsson, M., Hedblom, H. *et al*. (1999) Sulphur poisoning and regeneration of NO_X trap catalyst for direct

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

injected gasoline engines. S*AE International Fall Fuels and Lubricants Meeting and Exhibition*, Toronto, Canada 25-28 October 1999. Warrendale, Pennsylvania: SAE International.

Forzatti, P. (2001) Present status and perspectives in de-NO_X SCR catalysis. *Applied Catalysis A: General*, **222**, 221–236.

Haskara, I., Zhu, G. G., Winkelman, J. (2006) Multivariable EGR/Spark Timing Control for IC Engines via Extremum Seeking, *Proceedings of the 2006 American Control Conference*, Minneapolis 14-16 June 2006. Minneapolis, Minnesota: IEEE.

Heck, R.M., Rarrauto, R.J., and Gulati, S.T. (2009) *Catalytic Air Pollution Control: Commercial Technology*, 3th edn, John Wiley & Sons, Inc, New York.

Hirata, K., Masaki, N., Ueno, H. *et al*. (2005) Development of urea-SCR system for heavy-duty commercial vehicles, *SAE World Congress*, Detroit 11-14 April 2005. Warrendale, Pennsylvania: SAE International.

Jääskeläinen, H. (2010) *Fuel Injection for Clean Diesel Engines*, http://www.dieselnet.com/tech/engine_fi.html (accessed 8 October 2013).

Jääskeläinen, H. and Khair, M. K. (2011) *Charge Air Cooling*, http://www.dieselnet.com/tech/diesel_air_cool.html (accessed 8 October 2013).

Khair, M. K. (2006) *Exhaust Gas Recirculation*, http://www.dieselnet.com/tech/engine_egr.html (accessed 8 October 2013).

Kleemann, M., Elsener, M., Koebel, M., *et al*. (2000) Hydrolysis of isocyanic acid on SCR catalysts. *Industrial & Engineering Chemistry Research*, **39**(11), 4120–4126.

Ladommatos, N., Abdelhaim, S. M., Zhao, H. *et al*. (1996) The dilution, chemical, and thermal effects of exhaust gas recirculation on diesel engine emissions—Part 1: effect of reducing inlet charge oxygen. *SAE International Spring Fuels and Lubricants Meeting and Exposition*, Dearborn, USA 6-8 May 1996. Warrendale, Pennsylvania: SAE International.

Laroo, C. and Schenk, C. (2007) NO_X adsorber aging on a heavy-duty, on-highway diesel engine-part two, *SAE World Congress*, Detroit USA 16-19 April. Warrendale, Pennsylvania: SAE International.

Lee, J. (2007) An experimental study on HC–SCR and its activity enhancement by catalytic partial oxidation of hydrocarbons. PhD. Yonsei University.

Maiboom, A., Tauzia, X., and Hétet, J.F. (2008) Experimental study of various effects of exhaust gas recirculation (EGR) on combustion and emissions of an automotive direct injection diesel engine. *Energy*, **33**, 22–34.

Majewski, W.A. (2005) *Selective Catalytic Reduction*, http://www.dieselnet.com/tech/cat_scr.html (accessed 8 October 2012).

Majewski, W.A. (2007) *NO_X Adsorber*, http://www.dieselnet.com/tech/cat_nox-trap.html (accessed 8 October 2012).

Manufacturers of Emission Controls Association, MECA. (2000). *Catalyst-Based Diesel Particulate Filters and NO_X Adsorbers: A Summary of the Technologies and the Effects of Fuel Sulfur*, http://www.meca.org/galleries/default-file/cbdpf-noxadwp.pdf (accessed 8 October 2012).

Park, J., Lee, K.S., Song, S.H., *et al*. (2010) Numerical study of a light-duty diesel engine with a dual loop EGR system under frequent engine operating conditions using the DoE method. *International Journal of Automotive Technology*, **11**(5), 617–623.

Sluder, C.S., Storey, J.M.E. and Lewis, L.A. (2005) Low temperature urea decomposition and SCR performance. *SAE world congress*, Detroit, USA 11-14 April 2005. Warrendale, Pennsylvania: SAE International.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto108
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Pistons

Sanjeet Kanungo

Tolani Maritime Institute, Pune, India

1 Introduction 1
2 Piston and Ring Development 1
3 Materials and Coatings 6
4 Temperature Control and Piston Cooling 8
5 Ring Pack Design 10
6 Defects and Failures 13
References 14

1 INTRODUCTION

The only component of the engine that works against all the odds of high fluctuating temperatures, pressures, and more importantly varying velocity makes piston the most important element of the internal combustion engine. The connecting rod design encompasses all the varying forces arising due to change of reciprocating and gas forces to rotational forces. The crankshaft is designed purely on the effects of the rotational, reciprocating forces and torque. In contrast, piston is subjected to the adversaries of very high inertia forces arising due to the high piston speeds as is common in automobile engines, high temperature fluctuations as combustion is limited to one-fourth of the stroke and consequent high pressures. The fluctuations are more pronounced in automobile engines, as engine may not achieve steady state while underway because of traffic signals, bends, unequal road contours, and textures.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

The piston is a key engine part, which is used to compress gases within the combustion chamber or cylinder of an engine. The piston is sealed with a piston ring, often constructed under freezing temperatures, to allow expansion and therefore complete sealing, to seal the chamber to prevent leaks, or dangerous gases escaping. As the piston moves downward, a valve opens to allow fuel and air mixture into the chamber, this fuel charge is then compressed, combusted, and expelled as in SI engines. The motion is thus transferred through the propeller shaft, through a differential and gear box, into the axle and ultimately onto the road via the tire.

2 PISTON AND RING DEVELOPMENT

2.1 Design features and considerations

The main function of the piston itself is twofold:

1. It acts as a moving pressure transmitter, by means of which the force of combustion is impressed on the crankshaft through the medium of the connecting rod and its bearings.
2. By supporting a gudgeon pin, the piston provides a guiding function for the small end of the connecting rod. Figure 1 shows the general view of piston and piston rings.

The piston assumes a trunk form to present a sliding bearing surface against the cylinder wall, which thus reacts against the side thrust arising from the angular motion of the connecting rod. As the piston is a major reciprocating part, it must of necessity be light in weight to minimize the inertia forces created by its changing motion—bearing in mind that the piston momentarily stops at each end of

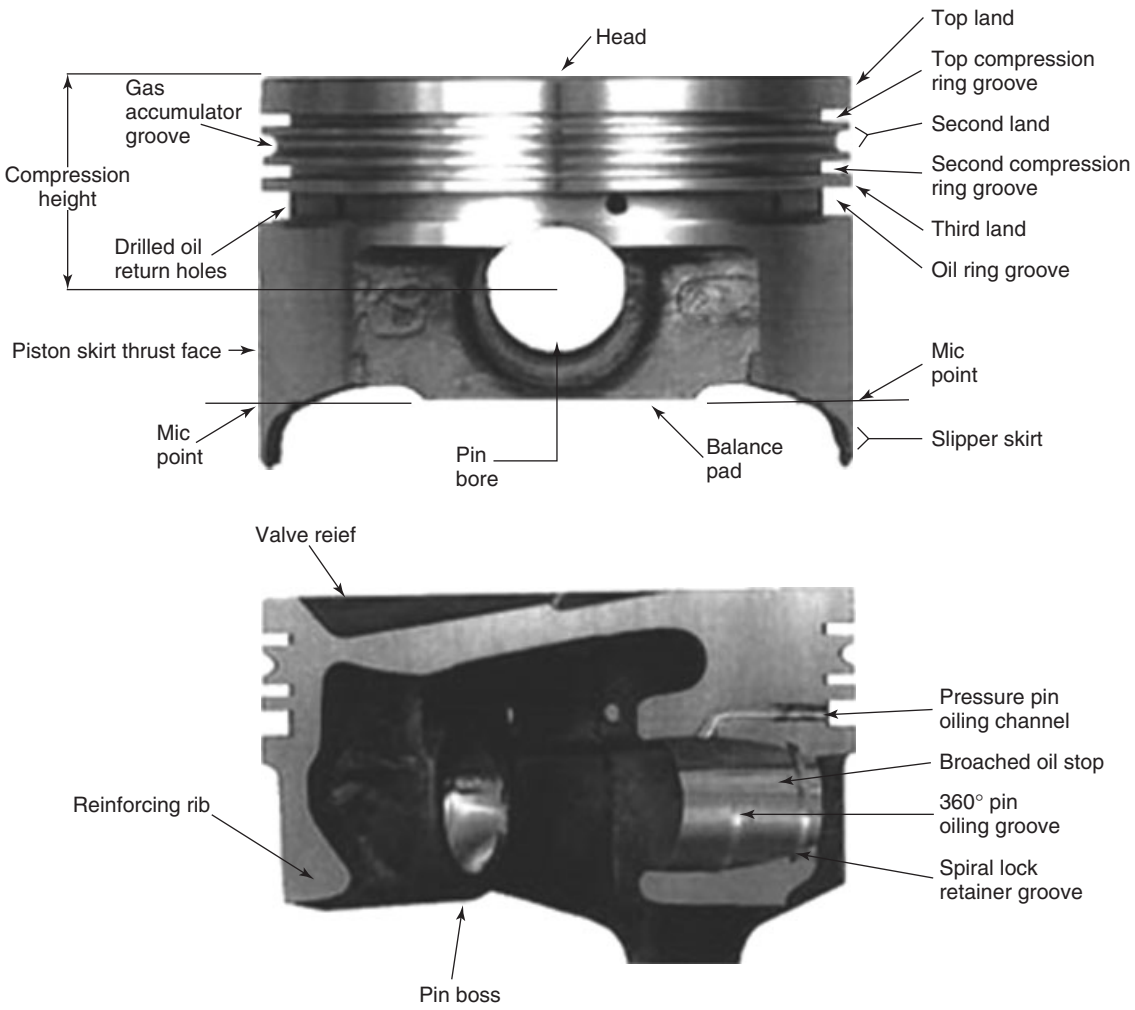


Figure 1. Cut section of an automobile piston. (Reproduced with permission from JE pistons.)

its stroke! Another, perhaps obvious, requirement is that the piston must be able to withstand the heat of combustion and should operate quietly in its cylinder, both during warm-up and at the normal running temperature of the engine.

To perform its sealing function efficiently, the upper part of the piston is encircled by flexible metal sealing rings known as the *piston rings*, of which there are typically three in number for petrol engines. In combination, the piston rings perform several important functions, as follows:

1. The upper compression rings must maintain an effective seal against combustion gases leaking past the pistons into the crankcase.
2. These rings also provide a means by which surplus heat is transmitted from the piston to the cylinder wall and thence to the cooling jacket.
3. The lower oil control ring serves to control and effectively distribute the lubricating oil thrown on to the cylinder walls, consistent with maintaining good lubrication and an acceptable oil consumption.

2.2 General description

The piston is a cylindrical piece of metal (always made from nice light aluminum alloy as far as we are concerned), which moves up and down in the cylinder. Automobile pistons are limited to single forgings, unlike large power medium speed engines with composite arrangement, such that the cast iron piston crown is capable of withstanding the high pressures and the light aluminum skirts not only aid in transferring the heat to the inner wall but more importantly contribute to reduction of inertia weights.

The piston is made of essentially seven parts. The piston's top or crown takes the brunt of combustion's forces and heat. Consequently, the crown is the hottest

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

part of the engine after the spark plug (for a SI engine). It must therefore be quite thick so as to not collapse, though it is not always the thickest part of the piston. Moving down the piston, the next thing is the ring groove. The closely manufactured groove accepts the third part, the precisely made piston ring. In the four-stroke engine, natural harmonics cause the ring to rotate as the piston goes up and down in the cylinder. This helps the groove stay clean of carbon. The solid pieces between the grooves are called *ring lands*. They support the shock loads the rings receive during combustion. The next part is the piston pin hole. This hole accepts a pin that connects the piston to the connecting rod. The hole is offset from the piston's center slightly so that when the piston and rod reach top dead center (TDC), they do so at slightly different times. This spreads the shock loads at high revolutions per minute, easing stresses on the connecting rod and eliminating a noise called *piston slap*. Surrounding the hole inside the piston are pin bosses, thick masses of metal that support the pin when it is inserted in the hole. The pin bosses are sometimes the thickest part of the piston. In some cases, they are not as thick as the crown. In either case, however, the thickness of these two parts is important, as it determines much about how the piston deals with heat. Last in the line is the piston skirt. The skirt is the bearing portion of the piston. It slides against the cylinder wall, bearing the force of combustion on the power stroke, and the loads of compression on the compression stroke. There are also stresses involved with revolutions per minute that the piston and cylinder are designed together to deal with. The skirt is the part of the piston most in need of lubrication. Thus, most lubrication problems show up on the piston skirt first.

The crown is considerably thicker as compared to the wall thickness of the skirt and is given shapes principally to accommodate the exhaust and inlet valves without compromising on the clearance volume and at the same time to effectively contribute to mixing of air and fuel and easy flame path.

Some of a diesel engine lubricant's most important basic functions are control of deposits, prevention of wear, and resistance to breakdown and thickening in service. Compression ignition, as utilized in the diesel engine, can lead to the production of unburnt fuel and carbon particles. These combustion by-products form combustion chamber and piston deposits and enter the crankcase along with blow-by gases. It is primarily deposits that may form on the top ring land, that portion of the piston wall above the top piston ring, and the susceptibility of the engine to these deposits, that determines the style of lubricant required. Hence, the lubricant for any particular diesel engine is determined by the design of that diesel engine, and largely by the design of the piston.

The volume between the piston and the cylinder wall, above the top ring and below the top of the piston crown, is called the *crevice volume*. Reducing the crevice volume increases a diesel engine's efficiency and reduces the amount of unburnt fuel and soot in the exhaust gases.

A piston with a larger clearance between the piston and cylinder wall, that is a piston with a "cutaway" crown land, will need a top ring located high on the piston to achieve an acceptably low crevice volume. The rings of a piston with a cutaway crown land will be exposed to higher forces from the combustion chamber gases through the wider clearance between piston and cylinder wall. These higher forces on the top rings will produce better ring action and minimize oil consumption. Control of hard carbonaceous deposits in the top ring land is the major requirement in diesel engines with reduced clearance between the piston and cylinder wall (Figure 2).

Four-cycle diesel engines from different areas of manufacture have different design features, such as the position of the top piston ring, power to capacity ratios, and sump

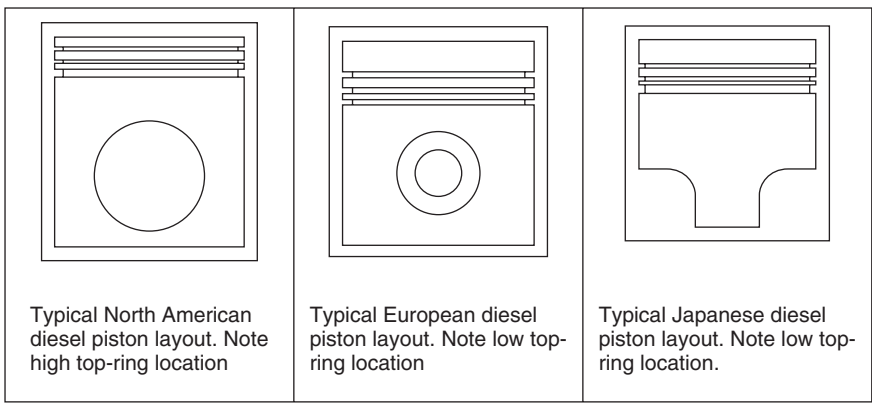


Figure 2. Some piston types with different crown lands.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 3. Antipolishing rings on a liner.

capacity. They are also subjected to differing operational expectations, such as extended oil drain intervals and the likelihood of their operation with high sulfur fuel. Deposits on land lead to bore polishing, increased oil consumption and reduced engine life.

This problem of bore polishing encountered due to the high crevice volume can be overcome by insertion of an antipolishing ring in the liner at the top part just above the maximum travel of the first piston ring. Figure 3 shows an antipolishing ring used in medium speed diesel engine.

2.2.1 *Piston shapes*

There are two important ways in which pistons are shaped. First, the piston is not round but elliptical in shape. The reason is the aforementioned pin bosses. The bosses' mass makes them absorb a lot of heat, which makes them expand more than any part of the piston. If the piston was instead made round, it would not be when fully warmed up. That would be a problem. Therefore, the width of the piston at the area of the bosses is narrower than it is elsewhere. The resulting shape (looking downward onto the piston crown) is an ellipse (an oval). Marine pistons are sometimes called *cam ground*, which refers to the same thing (however, it is not the shape that is being referred to in that case, but rather the machine that produces it). Secondly, all pistons have a taper. That is, the diameter of the piston at its crown is considerably smaller than its diameter at the skirt. The reason is the same as for the piston's ellipse. Only this time it is the crown, not the pin bosses, that necessitates the shape. The crown absorbs so much heat that it must be made smaller so that when fully heated, the piston will be straight.

The shape of the piston crown depends on the shape of its combustion chamber and its compression ratio. In diesel engines, the combustion chamber may be formed totally or in part in the piston crown, depending on the method of injection, so they use pistons with different shapes. Piston design has a great influence on the power output and efficiency of the motor. Pistons can be either round or oval, with a protruding, flat or concave crown. The piston head shape can also be made to induce swirling or vortices in the combustion gases to promote more even mixing and better combustion.

New technologies are being developed for gasoline direct injection engines as well. In direct injection engines, fuel is injected directly into the cylinder during the intake stroke (instead of being injected into the intake manifold near the intake valve outside of the combustion chamber), (U.S. Pat. 6, 3250, 40). This provides better atomization of the fuel and better combustion, but in order to maximize the benefits, the piston head must be carefully shaped. The piston head for direct injection has what is called a *bowl*. This bowl has two effects; the main objective is to deflect the incoming fuel spray toward the spark plug and secondly to direct it around the combustion chamber.

Direct gasoline injection also allows higher compression ratios of about 12 times versus the standard 10.5 (Singer, 1999). The result is a cleaner burning, more powerful and more efficient combustion process (Singer, 1999). Another side effect of spraying the fuel off of the piston head is to keep it away from the spark plug. If fuel hits the spark plug prematurely, it will combust and leave sooty deposits on the plug, reducing its effectiveness (Singer, 1999).

Piston heads can also simply have recesses that allow for better combustion (Haman and Craft, 1997). More complex designs cut away material around the circumference of the piston to achieve better mixing (Pouring, 1988). The yin-yang arrangement that is intended to induce swirl in the combustion gases (Yunick, 1986). With correct piston head shape, stress concentrations can be all but eliminated.

Efforts to enhance combustion by optimizing in-cylinder flows fall into two broad categories: (i) those that invigorate combustion by promoting tumble, swirl, and other in-cylinder flows throughout the engine's intake and compression strokes and (ii) those that use squish to create turbulence mainly in the vicinity of TDC on the compression stroke.

Three main piston-crown shapes and their respective merits and demerits are discussed. A flat piston has a small crown surface area and consequently has the merits of small heat losses and low weight. A horizontal-squish piston is effective at promoting combustion, but small

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

valve diameters are necessary to provide a squish area. During high speed operation, therefore, there is a tendency for air volumes to decrease and for performance to be concomitantly limited. A slanted squish piston is well suited to a pent roof combustion chamber in which ignition takes place at a central point between the four valves (the typical combustion-chamber configuration in recent engines), and it has the merit of providing a squish area without requiring any reduction in valve diameter. However, it has the demerits of a relatively large piston-crown surface area and relatively high weight. It has been reported that slanted-squish pistons suppress knock by accelerating combustion toward the end of the combustion process.

The skirt, an extended portion of the crown, appears like a hollow cylinder with no apparent use. However, skirts in four-stroke engines have the important role of transferring the side thrust generated because of the angularity of the connecting rod to the engine block. In so doing, they also act as medium of transfer of heat from piston crown to the cooled engine block. As skirts are intended to transfer the side thrust, they are fulcrum about the gudgeon pin, such that the horizontal component of the small end thrust of connecting rod is effectively transferred to the skirt and then on to the engine block as shown in Figure 4. However, the skirts also contribute to frictional loss because of its rubbing against liner. This leads to ovality of the liners. Developments are in the offing to simulate the piston movement as in large two-stroke slow speed crosshead engines, where skirts do not lean on the liner. Skirts are made oval to reduce the frictional contact and additionally coated with antifriction polymer coatings to reduce the abrasive wear.

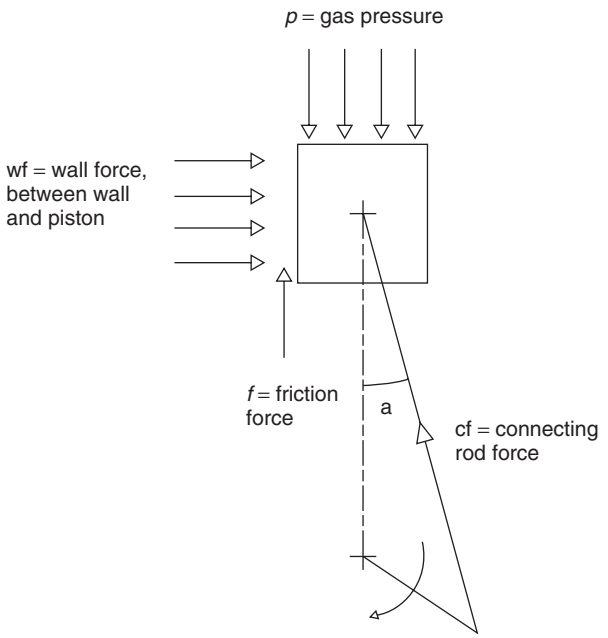


Figure 4. Forces acting on piston during combustion stroke.

As said earlier, it is important to keep the piston weight as light as possible to reduce the inertia effects. However, the prevailing pressures and temperatures influence the thickness and material. This in turn governs the weight of the piston.

In petrol engines, the combustion characteristics are largely different from the way diesel engines. Petrol engines do not self-ignite as a source of ignition always aids in combustion, thereby reducing the need of a high compression pressure. Moreover, external source of ignition helps flatter and lower peak pressure over a large crank angle for a given mean effective pressure. This reduction in peak pressure and the associated temperatures alongside a splash cooling from the crank throw aids in achieving a lighter piston. Therefore, petrol engines are generally made of aluminum, which are lighter and can with stand the temperature close to 400°C safely without loss of strength.

In contrast, diesel engines work under the principle of self-ignition, wherein air is compressed to the extent that the directly injected fuel into the compressed air can self-ignite. Thermodynamically, this calls for a lot of work done and the temperature of the air rises to ignite the initial part of injected fluid with the preset ignition delay thus providing spontaneous ignition to the incoming fuel, leading to a large uncontrolled combustion and thus high pressure rise. High pressure rises are also accompanied by high local temperature, although this temperature falls down rapidly as the piston moves down. This pressure rise calls for high strength and thus higher thickness across sections. It will thus be prudent to say the diesel engine pistons are heavier than petrol engine pistons.

Rings development: piston rings for automotive engines cater to the following functions:

1. Ensure sealing of the combustion space from the crankcase for better power output and prevent blow-by.
2. Spread the oil uniformly to prevent wear of liner.
3. Scrap the excess unregulated oil back to the crankcase.

Failure to achieve this leads to the following:

1. Engine not developing the required power.
2. Crankcase getting pressurized.
3. Crankcase oil becoming black.
4. Carryover of oil to the combustion space, thereby fouling the exhaust passage.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5. Thick gummy deposits due to the burning of lubricating oil.
6. Piston ring getting stuck.

3 MATERIALS AND COATINGS

The pistons in automobile engines as stated earlier are made of single metal, generally aluminum and in some rare cases cast iron, but are rarely composite. Most automotive engines use aluminum pistons that move in an iron cylinder. The average temperature of a piston crown in a gasoline engine during normal operation is typically about 300°C (600°F) and the coolant that runs through the engine block is usually regulated at approximately 90°C (190°F). Aluminum expands more than iron at this temperature range so for the piston to fit the cylinder properly when at a normal operating temperature, the piston must have a loose fit when cold. Expansion coefficient is a function of the silicon content of the aluminum alloy.

Forging requires alloy compositions lower in silicon content. Forged pistons have a finer microstructure than cast pistons with the same alloy composition. The production process results in greater strength in the lower temperature range. A further advantage is the possibility to produce lower wall thicknesses—and hence reducing the piston weight.

These have higher expansion coefficients than cast alloys. Therefore, it is not the forging, but the alloy selected that dominates the expansion coefficients and thus the piston to cylinder clearances required. More silicon means lower expansion coefficient. However, hardness and brittleness go up as silicon content goes up. Modern cast pistons have lots of silicon content. These alloys are termed *hypereutectic* (more than eutectic mixture, which is 12% silicon in aluminum). Because higher silicon contents make the alloy harder and more brittle, alloys used in forging are necessarily lower in silicon. The resulting alloy is tougher, sometimes stronger, more resistant to detonation damage and more suited to extreme use such as racing. However, it has a higher expansion coefficient requiring more piston to cylinder clearance to allow for the piston to grow when it heats. This is particularly true when using a steel or iron bore as steel and iron have lower thermal expansion coefficients than aluminum.

For aluminum pistons operating in steel cylinders, the key figure of merit is the difference in thermal expansion coefficients between steel and aluminum.

The majority of pistons are produced by gravity die casting. Optimized alloy compositions and properly controlled solidification conditions allow the production of pistons with low weight and high structural strength.

In addition, aluminum metal matrix composite materials are used in special cases. Pistons with Al_2O_3 fiber-reinforced bottoms are produced by squeeze casting and used mainly in direct injection diesel engines. The matrix holds the reinforcement together. Aluminum matrix composite refers to a class of material where aluminum is the metal matrix reinforced by materials such as SiC, Al_2O_3, TiC, TiB_2, graphite, and certain other ceramics.

Apart from the reinforced material, the morphology of the reinforcement too is of importance. The three major morphologies are continuous fiber, chipped fiber or whisker, and particulate. With further options such as reinforcement volume fraction and reinforcement orientation and aluminum alloy composition and heat treatment, a wide range of materials and resultant properties are feasible. The main advantage, apart from a general improvement of the mechanical properties, is an improvement of the thermal fatigue behavior. Squeeze casting process is a combination of gravity die casting and closed die forging. The technique in which metal solidifies under pressure within closed die halves. The applied pressure and the instantaneous contact of molten metal with the die surface produces rapid heat transfer that yields a pore-free casting with mechanical properties approaching the wrought product.

Squeeze casting offers high metal yield, nil or minimum gas or shrinkage porosity, excellent surface finish, and low operating costs. The patent on this process seems to be that of James Hollingrake in 1819 from Manchester. The steps involved in this process are (i) pouring of metered quantity of liquid metal with adequate super heat in to the die cavity, (ii) application of pressure on the liquid metal and maintaining the same until the solidification is complete, and (iii) removal of the casting and preparation of the die for the next cycle. The process is basically divided into two types: direct and indirect. The direct process is where the squeeze pressure is applied through the die-closing punch itself, whereas in the indirect process, the squeeze pressure is applied after closing of die, by a secondary ram.

The advantages of cast pistons with higher silicon alloy are

- lower expansion coefficient,
- less bore clearance,
- less noise on start up when pistons are cold, and
- better long-term wear in "normal" applications because the excess silicon hardens surfaces.

The advantages of forged pistons with lower silicon content are

- greater toughness,
- resistance to cracking,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- tolerance of abuse, and detonation
- but at the cost of greater required piston to cylinder clearance to accommodate the higher thermal expansion coefficient.

KS Kolbenschmidt uses two alternative steel materials for the production of the current monoblock steel pistons. The quenched and tempered $42CrMo_4V$ steel is considered a very good compromise between formability, material strength, resistance against scaling, machinability, and cost. Microalloyed or AFP steel grades offer cost advantages for medium-duty applications.

Schunk Kohlenstofftechnik offers pistons made of special carbon materials with significant advantages compared to conventional piston materials. The most important characteristics of these pistons are a low specific weight, high temperature and thermal shock resistance, a low coefficient of thermal expansion as well as excellent wear resistance, and low friction properties. The highlights of this piston are increased in mechanical strength with the increase in temperature unlike other materials. This enables the usage of such pistons in engine with very high thermal loads. A reduction in weight contributes a fuel saving of 5%. Low thermal expansion and shock resistance helps for dimensional stability and low crevice volume. The self-lubricating properties help for emergency running even when piston lubrication fails.

3.1 Coatings

Coatings of any kind should meet the following functional requirements:

- A thermal barrier coating on the piston crown.
- A friction reduction coating on the piston skirts.
- A wear-resistant coating on the piston grooves and piston rings.
- Oil shed coating.

These coatings, in turn, aim at the following:

- Increased power
- Reduced emissions
- Better response
- Increased overhaul periods
- More resistance to catastrophic failure.

The thermal barrier coating gives a more consistent finish to the top of the piston crown. It helps reflect heat into the combustion chamber, rather than dissipating it through the piston materials and thus improves combustion speed and the completeness of the combustion process. In other words, thermal barrier coatings increase the thermal impedance. However, this increase in combustion temperature and trapping the power rather than dissipating it does require reduced timing advance but will pay back dividends on higher revolutions per minute motors or on engines with short rod/stroke ratios where piston acceleration away from TDC becomes a problem for power transfer into the pistons at high revolutions per minute.

One thing to note here is that as the thermal barrier coating improves torque delivery by accelerating the burn rate inside the engine, it can be used to boost torque output on cars with centrifugal superchargers or large turbochargers to give better response before the boost builds.

Overall, it may seem like it's disadvantageous to trap more heat in the chamber and that it possibly reduces the octane, boost, or timing limits of the motor but this is not true. The increased heat can be counteracted with timing reduction without power loss (as the burn rate is maintained) and the added advantages are that the thermal coating helps spread the heat out over the entire crown area of the piston, thus protecting any thin or weak spots from being pummeled to failure. In addition, in the rare event that you do have some minor detonation in the engine because of high load, a bad fill of gas, or some other factors, the thermal coating prevents piston pitting due to minor occurrences of detonation and thus it prevents the creation of hot spots on the piston crown, which could have become hot beds for further detonation and a prevented runaway toward total piston failure.

The low friction coating on the piston skirt reduces frictional losses between the piston sides and the cylinder walls. This protects the pistons from damage and scuffing on cold starts, if the engine is overheated or overboosted (and the piston expands due to heat), and during oil starvation conditions.

Reducing friction in the engine delivers more horsepower to the crank, improves the engine's operation, and gives the engine crisper response. One study showed that using lower profile skirts, with proper friction coating, as well as lower friction wrist pins can reduce the total engine internal friction by 40%.

It is well known that the piston rings rotate in their respective grooves leading to what can be closely associated to fretting wear. Continuous wear of this kind will lead to less gas force acting behind the piston rings and hence loss of sealing.

Oil clinging to the underside of the crown adds to the weight of piston and so the coating should repel the oil.

No specific coatings are being highlighted for wide range of high temperature application coatings is available in the market. Figure 5 shows the coatings ranging polymers to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

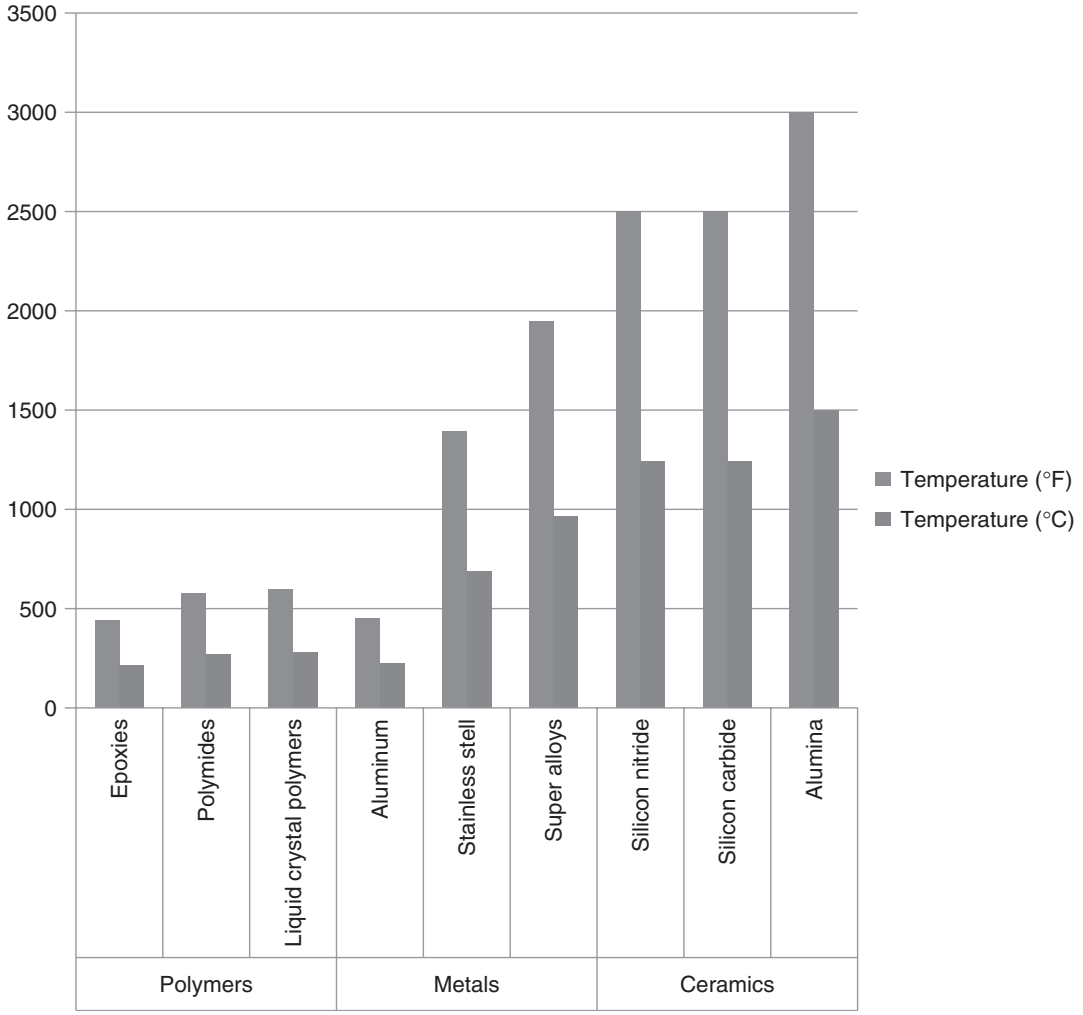


Figure 5. Coating types for different temperatures ranges.

metals and cermets. Choice of coatings also affects the cost. Hence, prudence has to be exercised for selection of materials. Aluminum pistons can be easily anodized to enhance its hardness and thus its wear resistance. Thermal barrier coatings are applied by either mechanical bonding or metallurgical bonding. Delamination is a possibility in mechanical bonding, commonly applied by high velocity oxy fuel (HVOF), plasma spray, and so on. Metallurgical bonding can be achieved by laser cladding, which is in vogue in bigger pistons.

4 TEMPERATURE CONTROL AND PISTON COOLING

Piston temperature has to be controlled for two primary reasons:

1. To prevent thermal weakening of the material.
2. To avoid knocking.

Materials become structurally weak when the operating at temperatures half the melting point. Thermal fatigue setsin at such temperatures and yielding is inevitable. Therefore, the surface temperatures have to be closely controlled. In case of aluminum, it is around 200°C.

A hot piston can additionally lead to early combustion and thus promote knocking. It is pertinent to note here that four-stroke engines (automobile engines generally are) have the inertia forces reversing every 90°, and the period when inertia forces are upward, the small end bearings get oil for lubrication. Knocking will affect the lubrication because of the predominating gas forces and the bearings will starve.

Therefore, effective cooling is of paramount importance. A large share of the heat absorbed by the piston top is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

transferred by the piston ring belt area. The remainder is essentially removed by the oil lubricant impinging on the underside of the piston. Automobile pistons get copious amount of lubricating oil, acting as coolant through drilled passage of the connecting rod and the piston gudgeon pin assembly. Connecting rod big end can also be provided with scoops for splashing the oil on the underside of the piston crown and inner face of skirt. Excess oil is scraped back to the crankcase by the scraper rings.

With increment in power, the piston temperatures are bound to rise. One method of reducing the temperature of the piston is by the use of a heat-pipe cooling system. This system allows for a channel inside the piston skirt that directs heat away from the piston itself. This greatly increases the heat transfer through the piston, which does not help the efficiency of the engine, but would allow for the use of special light alloys to form the piston (Wang *et al*., 2000). Magnesium and its alloys have much larger creep rates than other metals and therefore can usually not sustain the same load and temperatures as steel or aluminum. A heat-pipe system can drastically reduce the temperature of the piston crown from about 700°C to only 350°C (Wang *et al*., 2000). Therefore, using heat-pipe technology makes it easier to employ magnesium alloys in pistons.

The piston cooling configuration is of particular importance for the design of the monoblock steel piston. The decreased heat transfer capability of steel results in increased surface temperatures that require more supply of oil through the piston cooling nozzle. A stable oil jet at high pressures and reduced piston wall thicknesses between the cooling gallery and the combustion bowl, where the most extreme temperatures in the piston are observed, are the basis for high functionality and long life.

Piston cooling gallery is created for operation in high temperature environment such as in turbo engines, through oil circulation. The key factors for the design of the cooling gallery are the distance to the bowl rim and the thermal shielding of the top groove. The crown-cooling gallery piston has a curved cooling gallery, giving it a larger surface area that can remove more heat from the critical combustion chamber area and top-ring groove without allowing increased emissions. Additional height and surface of the cooling gallery enables increased heat transfer factors because of cocktail shaker effect and heat flow. The cocktail shaker effect is due to the inertia of the resident oil in the gallery at TDC. The splashing effect of the oil due to this enable oil reaching the difficult corners of the thrust side of the hot piston (Figure 6).

The monoblock steel piston design can be further improved by introducing drilled holes that connect the inner chamber with the outer cooling gallery. The added cooling effect reduces the surface temperature below the combustion bowl by approximately 25°C. This has the advantage of less carbon build and a reduction of oil evaporation from the piston surface, resulting in much less oil in the blow-by gas and lower emissions to the environment. Such cooling effect also results in smaller bore-to-piston clearance resulting in smoother and more fuel-efficient engines. Figures 7 and 8 show a general over view of the oil consumption and the operating temperatures of automotive engines.

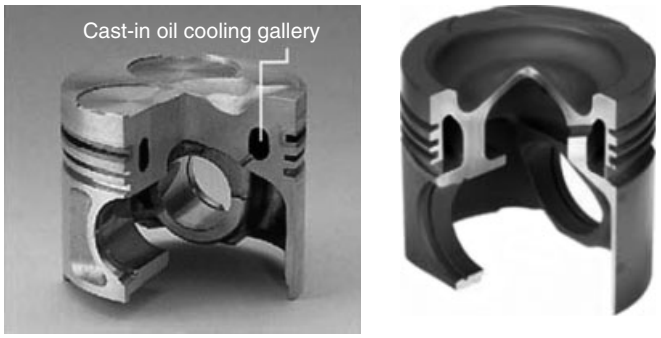


Figure 6. Pistons with cooling gallery. (Reproduced with permission from KS Kolbenschmidt. © KSPG AG.)

The role of the piston rings in maintaining the piston temperature is equally noteworthy. Piston rings make a physical contact with the bottom land of the piston groove and its side flank with the water-cooled or fin-cooled liner. This provides for a good conductance of thermal flux.

Heat gradient set up between the hot piston crown and the relatively cooler skirt aids in keeping the piston temperatures under control. Hence, aluminum skirts additionally help.

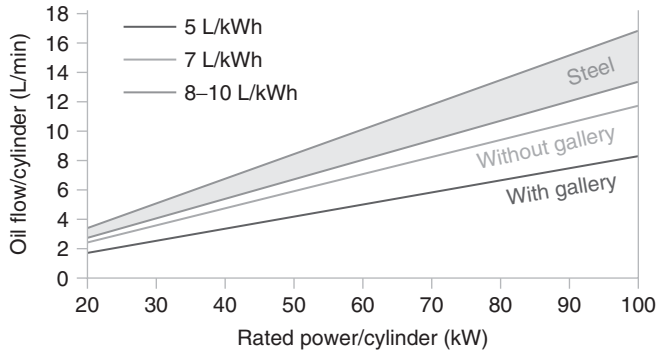


Figure 7. Oil consumption versus engine power. (Reproduced with permission from KS Kolbenschmidt. © KSPG AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

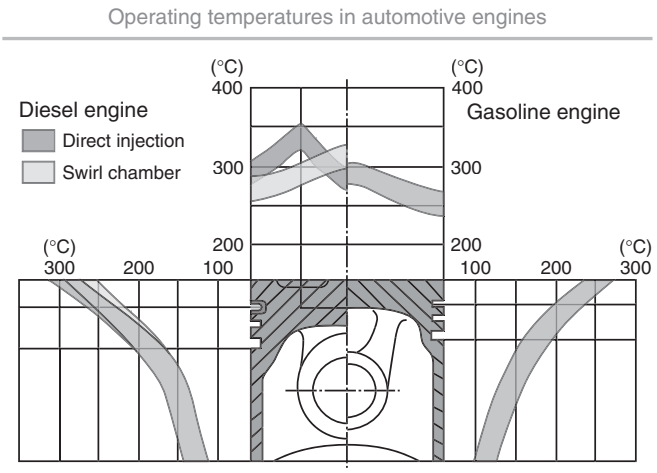


Figure 8. Operative chambers in automotive engines. (Reproduced from Röhrle (1995). © Verlag moderne industrie GmbH.)

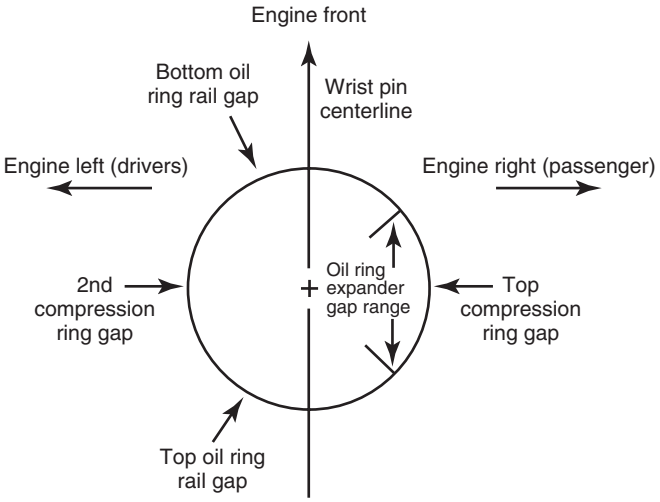


Figure 9. Ring orientation diagram.

5 RING PACK DESIGN

5.1 Materials and design features

The materials and design features of piston rings should meet the three functional requirements. They are as follows:

1. Prevent high pressure gas from leaking through the piston and liner interface, which would result in power loss.
2. Distribute the lubricating oil effectively to reduce friction between ring and liner and prevent lubricant oil migrating to the combustion chamber and getting burnt.
3. To serve as an effective conduit for heat dissipation from piston to the cooled liner.

The first ring commonly identified as the top ring does the sealing job. The third ring is an oil ring and regulates the oil flow. The second ring is primarily a scrapper ring and scrapes of the excess oil back to the crank case. Figure 9 shows the common ring assembly procedure to avoid blow past. Figure 10 shows the general arrangement of piston rings and their terminologies.

Three ring systems are common in automobile engines with the exception in racing engines with no second ring as height of the piston is a major restriction. In large engines, four or five ring system is adopted. Here, the second and third rings are essentially backup rings for the top ring. The top ring is that such cases are not fully gas tight and the sealing is thus achieved by labyrinth effect (Figure 13). This reduces the loading on the first ring and the material and design is thus different. With respect to the automobile engines, the top ring is designed to offer full gas tight seal.

Compression rings as stated earlier experience the toughest environment with respect to temperature and pressure. If the velocity and acceleration profile of the piston ring is plotted, it can be easily traced that it has maximum speed at mid-stroke and obviously very low at both dead centers. This proves to be beneficial at mid-stroke where hydrodynamic lubrication is possible with very less lubricating oil. Toward either end, the lubrication goes into mixed film and finally boundary regime. In addition at TDC, during the power stroke, the temperature and pressure both are high and create a high radial force, squeezing out the lubricating film. Effective lubrication is possible by barrel-shaped face profile. Failure to provide such profiles will result in high wear rates.

Thermal analysis of the piston rings shows large contact force at the either open ends, which is known as *butt*. Butt clearance of certain degree is provided to consider the thermal expansion. Failure to do this will lead to ineffective sealing and collapse of the ring. As stated earlier if the contact force at the butt end is more, this implies that there will be nonuniform contact and hence more chances of blow past. This can be reduced by negative oval casting (Figure 11).

Scrapper ring has a tapered face showing that it cannot accumulate oil on its upper edge to scrape it in the upward direction toward the combustion chamber. However, it can very effectively accumulate oil on it lower edge to scrape it down to the crankcase, thereby preventing excessive consumption of oil.

Oil control ring (Figure 12), wherever employed, has a two-land edge backed up by coiled spring to enhance

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

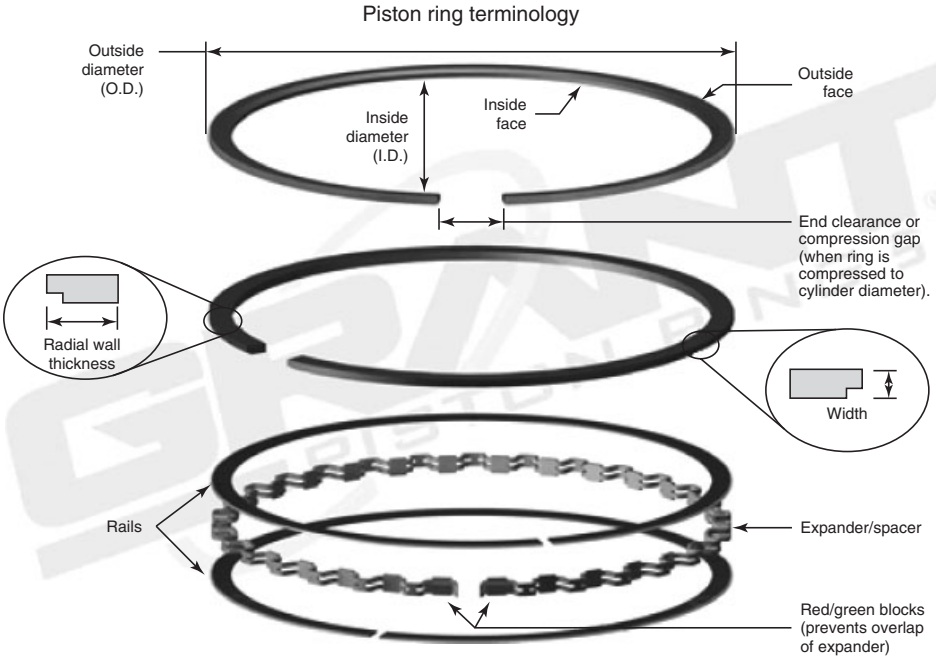


Figure 10. Piston ring terminologies. (Reproduced with permission from www.grantpistonrings.com. © Grant Piston Rings.)

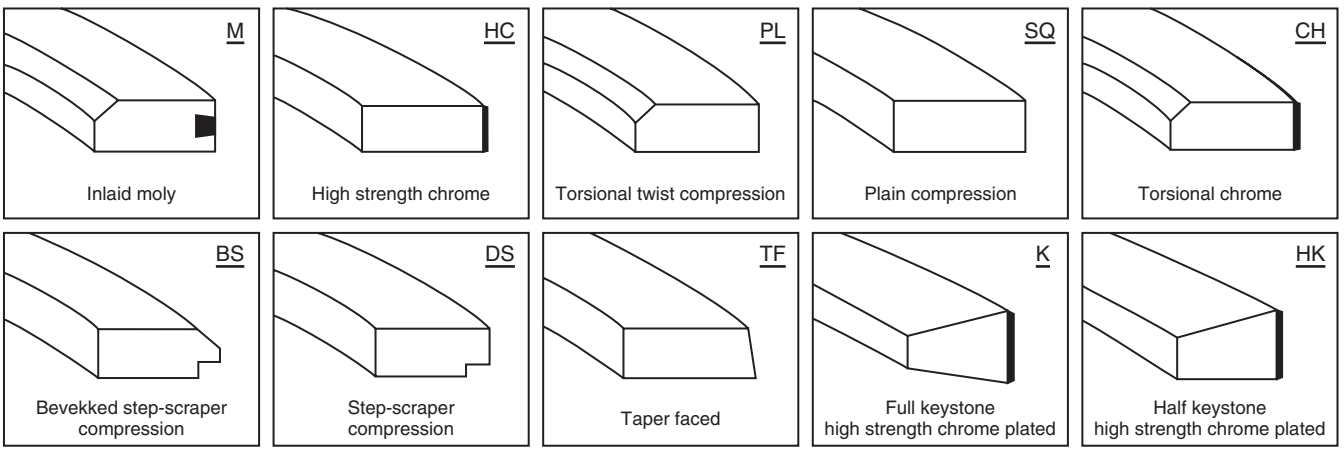


Figure 11. Types of butt opening for the top ring. (Reproduced from www.beco.in. © BECO Group.)

conformability with the liner. Two lands are used to ensure that at least one of the lands will control the oil during anytime of the engine cycle.

5.2 Ring dynamics

About 20/25% of mechanical friction losses comes from the piston rings rubbing against the liner. As a result, reduction in piston ring friction has the potential to improve fuel efficiency, fuel consumption, and emissions.

The piston ring secondary motions can be divided into piston ring motion in the transverse direction, piston ring rotation, ring lift, and ring twist. These types of motion result from different loads acting on the ring. Loads of these kinds are inertia loads arising from the piston acceleration and deceleration, oil film damping loads, loads owing to the pressure difference across the ring, and friction loads from the sliding contact between the ring and the cylinder liner. The forces acting on the ring are shown in Figure 13 (Ejakov, Diaz, and Chock, 1999).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

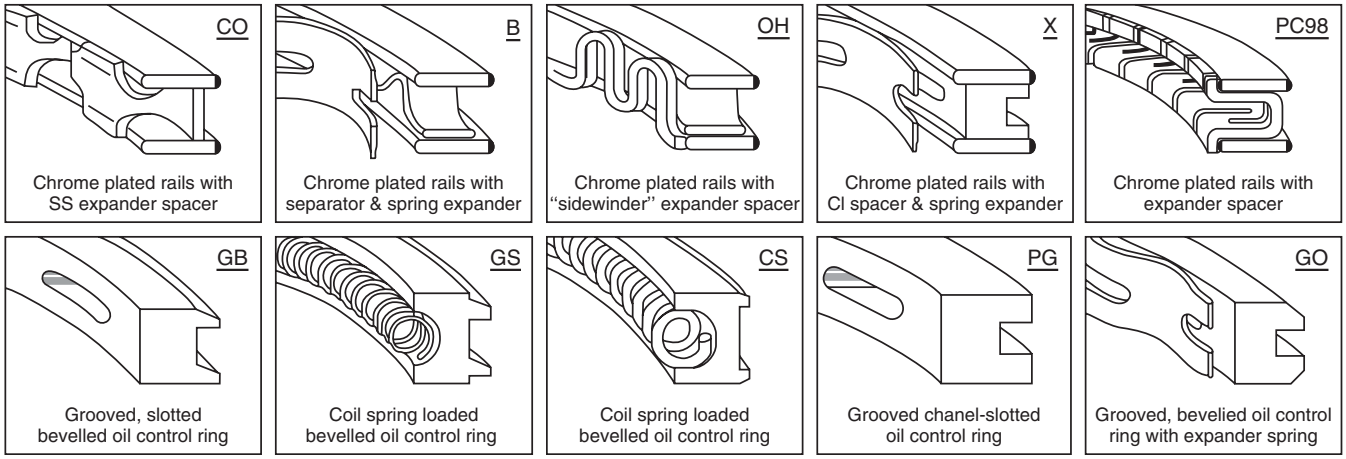


Figure 12. Types of oil scraper and control ring. (Reproduced from www.beco.in. © BECO Group.)

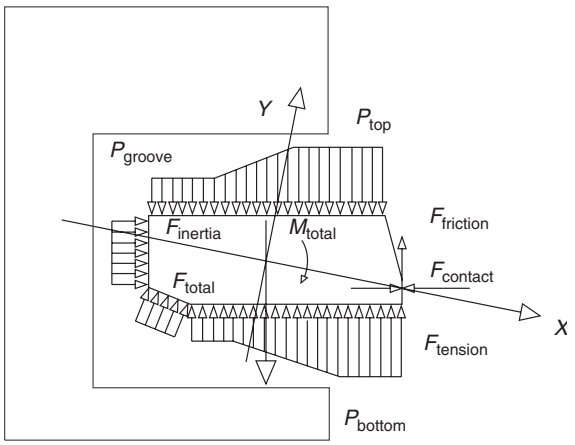


Figure 13. Force diagram of piston ring. (From Andersson, Tamminen and Sandström (2002) (After Ejakov, Diaz, and Chock, 1999).)

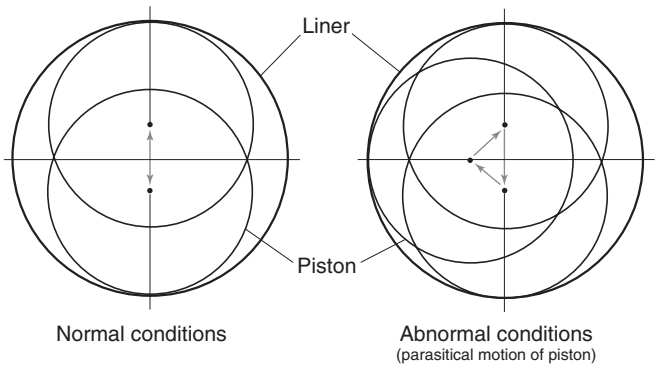


Figure 14. Parasitical motion of piston in a liner.

The gas pressure above, below, and behind the ring produces resultant forces on the ring section (Dowson, 1993). The inertia forces acting on the piston rings, as well as those acting on the other reciprocating crank mechanism components, change proportionally to the square of the engine speed (Röhrle, 1995). The side loading of the piston against the cylinder wall is a result of the articulated joint of the connecting rod (Röhrle, 1995). The shearing of the lubricating film, the sliding friction forces, and the contact pressure between the ring and the liner cause normal and tangential forces on the ring face.

The elastic distortion of the piston and liner can affect the effective geometry of the ring face and cylinder liner contact, which causes a nonuniform distribution of the contact pressure between the cylinder liner and the piston ring face and can thus lead to increased blow-by and oil consumption (Dowson, 1993).

As stated earlier, piston rings rotate in their grooves and exhibit what is called *fretting wear*. These can be observed by equally spaced patches both on the bottom groove land and on the rubbing surface of piston. These are unavoidable and are interestingly interpreted in four stroke engines. This is due to parasitical rotary motion of the piston (and accordingly piston rings) can be produced not only nonuniformly heated liner but also bend crankshaft, inaccuracy manufacture, and poor assembling (Figure 14).

The piston ring rotation is induced by the fact that the liner diameter varies over the stroke. The diameter variation is due to thermal deformation and wear. The ring will always be pressed toward the liner wall, thus the ring will expand and compress over the stroke. The continuous movement of the ring in combination with a variation in friction conditions and piston movements will then cause the ring to rotate in a stochastic way.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In addition to rotation, the piston rings also slam against the groove. These lead to wear of the grooves. The first ring that experiences the maximum contact pressure due to the primary sealing as well as secondary sealing also experiences wear. To mitigate the risk earlier failure and down time because of wear, the critical parts such as top piston ring and the groove bottom lands are coated with wear resistant coatings, the common among them being chromium coatings deposited by mostly electroplating. This coating meets the functional requirement of a hard coating with a suitable hardness, but the process of deposition is not environment friendly because of the release of chromium hexavalent mist (an ozone depleting layer). Suitable cermet coatings such as tungsten carbide cobalt matrix, nickel chromium, custom-blended chromium carbide, and nickel chromium-based coatings have proved their mettle with a suitable deposition technique. HVOF or plasma spray process is some of the common thermal spray process employed by industries. However, laser cladding or laser glazing methods are replacing the other thermal spray process because of the nature of metallurgical bonding without any distortion. This prevents delamination of the thermal coating and thus less risk of failure.

6 DEFECTS AND FAILURES

The study of pistons of their rings and the cylinder liner which together go into making of an effective sealing and thus influencing good combustion will be incomplete without understanding the defects and taking corrective actions thereafter. A lot of automobile manufacturers have been gracious enough to cite the defects on the mentioned components arising due to various operating reasons. It is urged that readers should have generous reading of "Piston damages" published by MSI Motor Service International GMBH. Figure 15 shows some photographs of the crown,

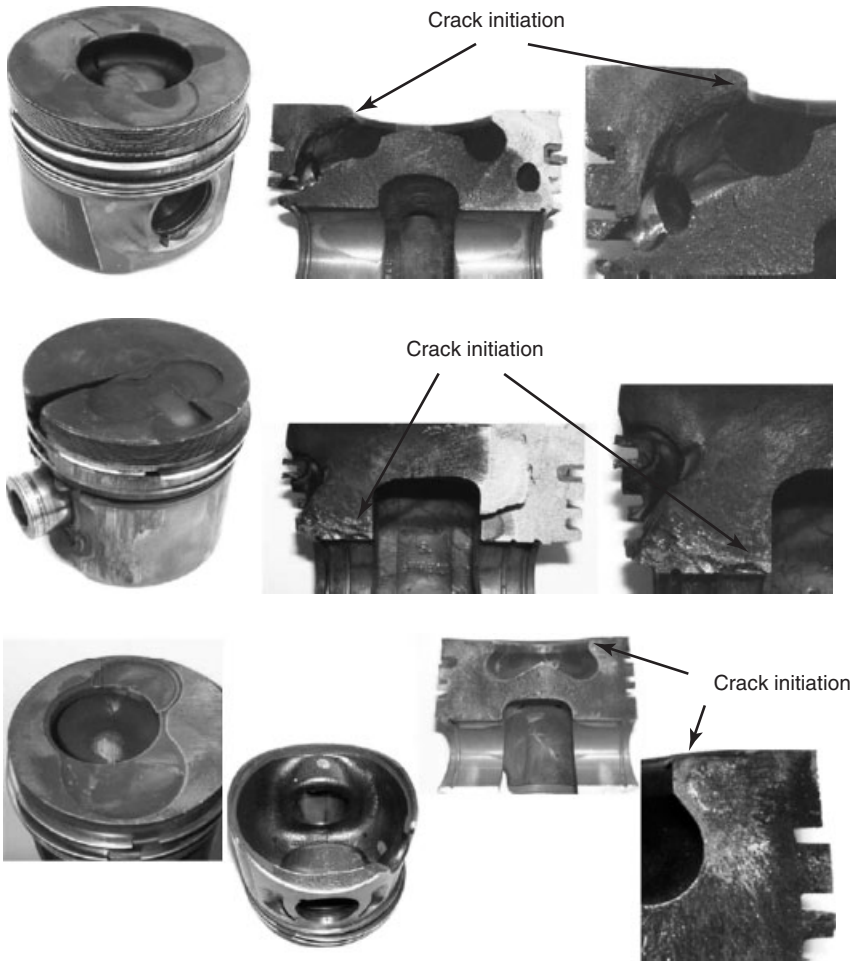


Figure 15. Fatigue failures on a typical piston. (Reproduced from Silva (2006). © Elsevier.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

gudgeon pin, and skirt of a piston due to fatigue (Silva, 2006).

REFERENCES

Andersson, P., Tamminen, J. and Sandström, C.-E. (2002) Piston Ring Tribology: A Literature Survey, http://vtt.fi/inf/pdf/tiedotteet/2002/T2178.pdf (accessed 29 October 2013).

Dowson, D. (1993) Piston assemblies; background and lubrication analysis in *Engine Tribology* (ed C.M. Taylor), Elsevier, Tribology series, 26, pp. 213–240. ISBN: 0-444-89755-0

Ejakov, M.A., Diaz, A.R. and Chock, H.J. (1999). Numerical optimization of ring-packbehavior, Society of Automotive Engineers, Inc. SAE Paper 1999-01-1521, 12, 10.4271/1999-01-1521, USA

Haman, D.F. and Craft T.D. (1997). Outboard Marine Corporation. Fuel-injected Internal Combustion Engine with Improved Combustion. Canadian Pat. CA 2177439, Canada.

Pouring, A. (1988). Sonex Research Inc. Piston for I.C. Engine. Canadian Pat. CA 1242120, Canada.

Röhrle, M.D. (1995) *Pistons for Internal Combustion Engines–Fundamentals of Pistontechnology*, MAHLE GmbH. Verlag Moderne Industrie, Landsberg/Lech, Germany, p. 70.

Silva, F.S. (2006) Fatigue on engine pistons–a compendium of case studies in *Engineering Failure Analysis*, vol. 13, Elsevier Ltd, UK, pp. 480–492.

Singer, S.. (1999). Combustion Engines Burn Super Clean in Machine Design. 71, 17, September 9, http://machinedesign.com/automotive/combustion-engines-burn-super-clean (accessed 29 October 2013).

Wang, Q., Cao, Y., Chen, G., *et al*. (2000) Studies of a heat-pipe cooled piston crown. *Journal of Engineering for Gas Turbines and Power*, **122**, 99–105.

Yunick, H.. (1986). Motortech, Inc. Internal Combustion Engine. Canadian Pat. CA1210285, Canada.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto111
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Axle Systems

Joseph Palazzolo
GKN Driveline, Auburn Hills, MI, USA

1 Introduction 1
2 Basic Functions 1
3 Axle Constructions 2
4 Independent Carrier Axles 5
5 Hypoid Gearing 5
6 Bearing Considerations 8
7 Manufacturing Methods 10
References 11

1 INTRODUCTION

With careful examination of a typical passenger car or light truck, it is clear that there is a single engine that can drive two or more wheels. In addition, it can be seen that in rear-wheel drive powertrain layouts, the axle assembly is connected to the powertrain with a long shaft called a *propshaft* or a *driveshaft*, whereas, on a front-wheel drive powertrain layout, the axle functionality is combined within the transaxle. For the purpose of this chapter, the main focus will be on traditional rear axle assemblies, but the theory and practice can be applied to front axles as well. The primary function of the axle assembly is to translate the axis of rotation from the longitudinal direction to the transverse direction in the rear-wheel drive vehicle. Simply put, the propshaft is configured to be running fore-aft in the vehicle and the axle outputs are perpendicular to that in order to divert the power to the wheels. There is a unique gear arrangement that provides this perpendicular powerflow and it is referred to as a hypoid gear set.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 BASIC FUNCTIONS

There are two generic categories of axle; *live* and *dead*. The *dead* axle is an axle that does not transfer power to the wheels and mainly acts to support the vehicle weight and suspension loads. These are typically found on the rear suspension of a front-wheel drive-based car. At times, they are referred to as *twist* beam axles. In contrast, the *live* axle transfers mechanical power to drive the wheels and is the main focus of this chapter. From here forward, the term axle will refer to a live axle arrangement.

2.1 Structural functions

As a beam element is located between the wheels and the chassis of the vehicle, the axle is required to support vehicle level loads. The first and the most obvious load is the actual rear vehicle weight. The axle structure must be able to carry the portion of the fully loaded vehicle that is distributed to the rear wheels from the chassis and any additional loads from passengers and cargo. The axle structure must also transmit and react both driving and braking torque reactions from the powertrain to the wheels and in turn from the wheels through the foundation brakes. The axle assembly also provides the mounting brackets and surfaces for the suspension components such as springs (leaf or coil), suspension arms, and the shock absorbers. The axle must also be able to resist the wheel loads that are applied to the structure when cornering events are encountered. The rear wheels do not articulate like the front wheels and are, therefore, dragged through a turn. This dragging actually produces significant loads on the wheel

hub flange portion of the axle assembly and the entire axle wants to travel straight as opposed to following the vehicle chassis in the turn. Some performance applications will make use of a panhard rod or track bars or even a Watt's linkage to counteract this transverse motion of the axle relative to the chassis during turn events. These are basically transverse-mounted suspension links to limit the range of side-to-side motion of the axle assembly relative to the chassis. The specific loads are very dependent on the type of suspension and spring arrangement. Two of the common types are leaf spring or Hotchkiss style and coil spring with a multilink (typically, three- or four-bar trailing arm) suspension arrangement. There are many variants of suspension arrangements with advantages and disadvantages to each. The specific detail of them is readily available in the public domain literature.

2.2 Mechanical functions

From a basic overview, the axle assembly needs to mechanically transmit torque at the required speed from the propshaft to the wheels. With mechanically transferring this power, the direction of the torque flow is translated 90° with a hypoid gear set. The hypoid gear set provides torque multiplication and speed reduction based on its gear tooth combination. In the traditional axle drivetrain system for automotive passenger vehicles, there is a need for a device that can divide torque as well to the driving wheels equal from side to side. This mechanical device is known as a *differential*, specifically referred to as an *open differential*. The open differential allows for a single input from the powertrain to be divided to the two outputs, typically the wheels. The open differential still accommodates the requirement of speed differences experienced in turn maneuvers while providing equal torque to the outputs. This equal torque balance of the conventional open differential can create a drive force issue in the vehicle if one wheel has lost tractive effort as experienced with one wheel is on a low coefficient of friction surface such as snow or ice. The details of open differential are covered in Basic Open Differentials, with the need and functionality of limited slip differentials, as reviewed in Passive and Active Limited Slip Differentials.

3 AXLE CONSTRUCTIONS

There are two main types of rear axles that cover the majority of passenger vehicle applications, which are beam- and independent carrier-style axles. The main difference is that the beam-style axle has rigid tubes that are part of the axle housing structure. These tubes typically have the suspension and brake mounting brackets rigidly attached to this structure. The beam axle also has the axle shaft supported inside this structure and is responsible for axle shaft and wheel bearing retentions.



Figure 1. The independent carrier-style axle.

In contrast, the independent carrier-style axle housing, as shown in Figure 1, supports the hypoid gearing and differential but does not transfer the vehicle weight, brake-reaction loads, or suspension motion through the axle structure. Typically, there are halfshafts (or side shafts) between the output of the differential and the wheel hubs to transfer the drive torque. These halfshafts allow for the wheel hubs to move through suspension travel without transferring the suspension loads to the axle housing. The brake forces are also supported within the wheel hub arrangement and not through this inboard-mounted axle housing. Basically, the suspension and brake loads are decoupled from the axle. This allows the chassis and suspension engineers much more flexibility to optimize the systems without the constraints of a beam axle system.

3.1 Beam axles

There are a few differential arrangements of beam axles, but the two most common are the cast center style with pressed-in tubes and the stamped and welded axle housing structures. The cast center style is commonly referred to as the *Salisbury style*, whereas the stamped and welded types are commonly referred to as the Banjo-style axle. The most common arrangement is the cast center type, but there are some original equipment manufacturers that still utilize the Banjo type.

3.2 Cast center section

The cast center section-style axle, as shown in Figure 2, uses a cast center section with the axle tubes pressed in place. The center portion is a single casting with an opening on the rear surface for assembly and service access. The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 2. Cast center-style beam axle.

tubes are retained via the press fit and redundant slug welds that are visible as circles near the top of the axle tube interface portion. The suspension and brake hardware mounting brackets are welded to the axle tubes accordingly. The hypoid gears, bearing, and differential are assembled from the rear-facing surface of the axle. There is a typically stamped steel or cast aluminum cover bolted to the rear that seals this large access port. One of the main advantages of this style axle structure is weight and cost to produce. Most of this style axles are designed and assembled with select fit shims for gear location and bearing preloads, which helps to reduce the manufacturing cost associated. While this arrangement aids in initial cost mitigation, service procedures can be more difficult. This style typically is configured with the axle shaft retention mechanism located inside the differential. This axle shaft retention is commonly referred to as semi-float and will be described later. The cast center arrangement normally does not utilize a separate pinion cartridge and this geometry is cast and machined directly as a part of the main housing. This arrangement requires that the bearing preload is disturbed to adjust pinion mounting distance shims.

3.3 Stamped center section

The stamped- and welded center-style axles will be referred to as *Banjo style* as shown in Figure 3. This arrangement of axle housing consists of a series of stamped steel sections. These stampings are arranged in fixtures and welded together to achieve the complete axle assembly. The rear portion of the axle housing is a part of this welded assembly and, therefore, the differential and hypoid gears must be installed from the forward-facing portion of the axle housing. There is a cast gear carrier that supports the gears, bearings, and differential. The gear carrier, when it is fully assembled, is commonly referred to as the third member. There are some advantages to this style axle arrangement from a service standpoint. Mainly, the end consumer can have multiple third members preassembled and change them relatively quickly if desired. Another

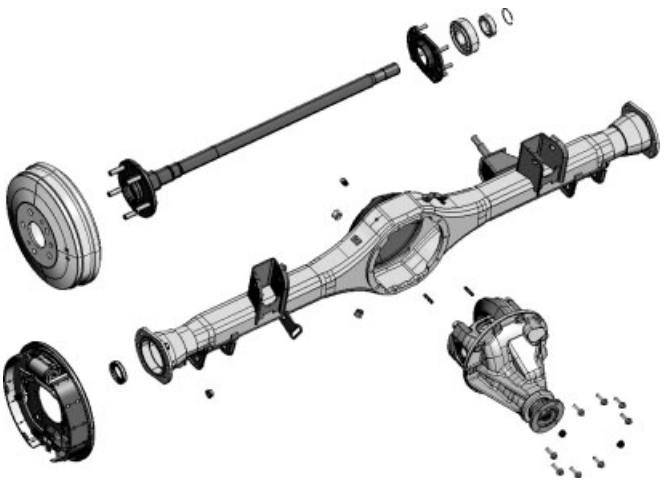


Figure 3. This is an exploded view of the typical stamped-style axle housing, commonly referred to as Banjo style. (Reproduced by permission of Joe Palazzolo.)

characteristic of this arrangement based on the assembly process and inability to access the differential is that the axle shaft retention mechanism must be outboard of the differential. This axle style employs what is known as three-quarter float wheel-end bearing arrangement, which will be reviewed later. One of the disadvantages or manufacturing concerns with this style axle structure, besides the complexity of assembly, is the correct alignment of all of the critical geometry for wheel bearings and gear case mounting throughout the welding and potential subsequent distortion process. In addition, there is a concern for minor leaks from the series of welds that are required to join all of the stamped pieces together. These welds provide the structure and act to seal the axle housing assembly. All of these items have been overcome with modern manufacturing processes and quality control systems. Some of the Banjo-style axle housings utilize a pinion cartridge arrangement that allows for the separation of bearing preload and pinion mounting distance. The pinion cartridge can be shimmed separately to achieve the desired mounting distance without disturbing the pinion bearings. This also requires additional parts and another potential leak path as a design trade-off.

3.4 Wheel-end bearing configurations

There are three main architectures of wheel-end bearing configurations and axle shaft retention schemes that are most common for automotive applications. The differences among the three different types is mainly how the axle shaft will be transferring torque to the wheel hub directly

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and whether it is carrying a portion of the vehicle mass and bending loads while rotating.

3.4.1 Full-float wheel end

The term *full-float wheel-end bearing* refers to the axle shaft retention and bearing strategy that totally isolates the axle shaft from the vehicle loading conditions. In this structure, the axle shaft is responsible solely for transferring the drive torque from the differential to the wheel hub. Referring to Figure 4, the tell-tale axle shaft bolts are at the left end of the diagram. Close examination will reveal the two tapered roller bearings that support the wheel hub to the axle housing. This load path isolates the axle shaft from the vehicle loads and is a very robust design solution. This is by far the most robust wheel-end bearing arrangement and, no surprise, the most expensive. In this architecture, the wheel hub is mounted directly to the axle housing and there is no endplay (or side-to-side motion) of the wheel assembly. The key differentiator is that the wheel hub is supported by separate bearings and transfers the vehicle loads directly back to the axle housing. The axle shaft is such that it can be removed from the axle and the vehicle wheels and tires will still support the vehicle. A quick visual indicator of a full-float axle shaft is that the axle shaft will have a series of bolts that attach it to the wheel hub that can be seen from the center of the wheel. On the basis of the separation of load and the ability of this arrangement to support heavy vehicle arrangements, this structure is often used for the heavier end light truck applications (under 8500 pound GVW) and medium and heavy truck applications.

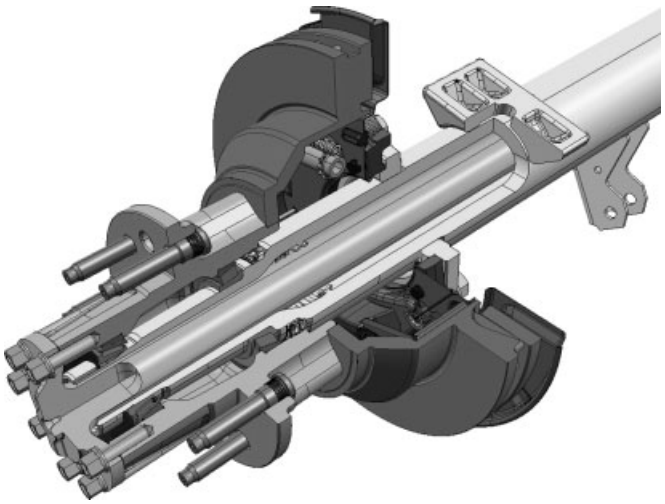


Figure 4. A partially cutaway view of a full-float wheel-end bearing arrangement. (Reproduced by permission of Joe Palazzolo.)

3.4.2 Three-quarter float wheel end

The three-quarter float wheel end, as shown in Figure 5, describes a wheel-end strategy that carries torque and partial vehicle load through the axle shaft. The axle shaft is still retained at the wheel end like the full-float style with the added difference in the bearing load path. The bearing load path for the vehicle mass is transferred through the axle in the axle housing. This wheel-end and axle shaft retention is commonly used on the Banjo- or third-member-style axle housings. There are a few different style bearing and seal arrangements that can be found on this style wheel end. The main function is the same and most applications use a greased and sealed bearing at the wheel end, whereas some of the higher loaded applications will use a tapered roller bearing that shares the lubricant with the main axle components.

3.4.3 Semi-float wheel end

The semi-float wheel end is the most common arrangement and is paired with the cast center-style axle housing structure (Figure 6). The axle shaft is actually retained with a special washer that is installed in a groove that is machined on the shaft end and located inside of the differential itself. In order to gain access to the washer, the axle must allow for access to the differential after it has been installed and the cast center with its rear access portion allows for this. The semi-float axle shaft uses a simple cylindrical roller bearing to support the axle shaft that used the axle shaft

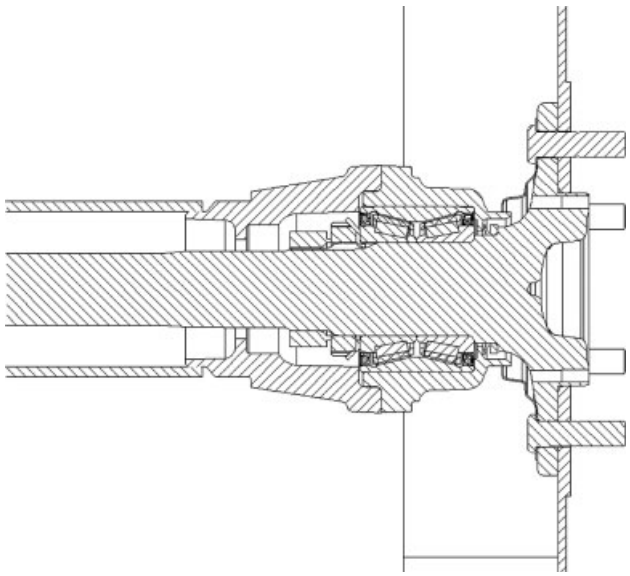


Figure 5. The three-quarter float wheel-end bearing arrangement. (Reproduced by permission of Joe Palazzolo.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

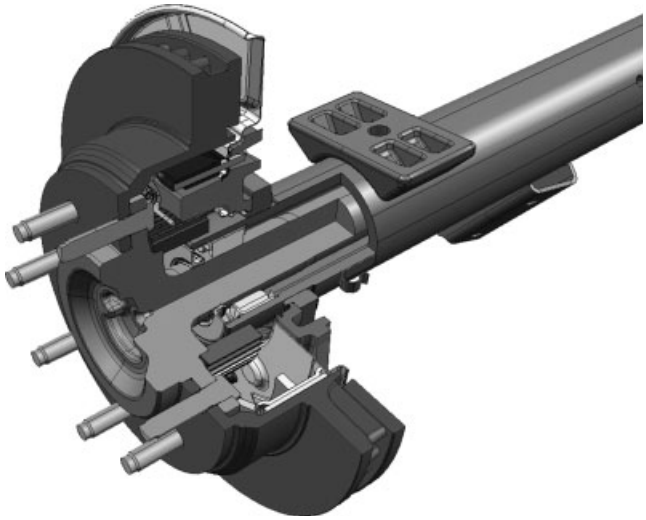


Figure 6. The semi-float wheel-end bearing arrangement. (Reproduced by permission of Joe Palazzolo.)

journal as the inner bearing race for a more economical solution. This bearing shares the lubricant with the main axle components. In operation, the axle shaft is subjected to not only torque to drive the wheels but also full vehicle loads and retains the wheel. On the basis of the geometry constraints of this style of axle, it exhibits the most endplay of the three types reviewed. This endplay needs to be considered when disk-style brakes are employed.

4 INDEPENDENT CARRIER AXLES

There is an additional axle arrangement that isolates the axle housing system completely from vehicle load. This type of axle is commonly called an independent carrier axle. The axle tubes and axle shafts from a typical beam-style axle are replaced with halfshafts and a suspension geometry that reacts to the vehicle loads. In this axle arrangement, the vehicle weight and suspension travel do not change the position or load the axle housing. This provides for additional freedom in suspension design along with packaging advantages and often found on luxury and sports cars.

5 HYPOID GEARING

5.1 Main functions

There are three main functions of the hypoid gearing; mainly, it provides a 90° direction change, provides torque multiplication, and acts as a mechanism to distribute lubricant to the gears and bearings. The fundamental geometry of a typical automotive hypoid is that the angle of the output gear, commonly called ring gear, is at angle, that is, 90° in relation to the input gear, commonly called the pinion gear. The angular relationship is often referred to with the Greek symbol alpha, α. With any two gears in mesh, the smaller of the two mating gears is referred to as the pinion and the larger as the gear. The fundamental function of the hypoid is to increase torque from the input to the output along with decreasing speed, or simply stated that torque and speed are inversely proportional. The torque ratio for the hypoid is the number of teeth on the ring gear divided by the number of teeth on the pinion. For example, a 41-tooth ring gear meshing with an 11-tooth pinion would yield a torque ratio of 3.7272, which is typically rounded to 3.73 : 1. The speed ratio is the inverse of that and 0.26829 : 1 for the example tooth combination. During the gear ratio selection process, it is typically desired to have a full hunting tooth combination to help with noise reduction and wear concerns. The term *hunting ratio* refers to the mathematical relationship between the tooth combinations of the gear set. A simple means to determine if a ratio is hunting or not is to examine the factors of the tooth combination and make certain that there are no common factors, excluding one, of course. For example, a 13-tooth pinion driving a 39-tooth ring has a common prime factor of 13 between the tooth pairs. This means that the pinion teeth will mesh with the same 13 teeth on gear, three times per revolution. For optimal noise and wear, it is desired to have the gear teeth always meshing with different teeth on the mating gear. The typical corrective action, for example, would be to increase the number of teeth on the ring gear to 40 and, therefore, there are no longer any common factors. The ratio changes slightly now to 3.08 : 1 compared to the 3.00 : 1. The last function that the hypoid gears perform is to adequately distribute the lubricant from the axle sump to the gears, bearings, and seals in all driving conditions. The gear system is supported with certain axle housing geometry features and ports to aid in channeling and retaining the lubricant. Figure 7 illustrates the oil flow from the axle housing sump being carried by the ring gear during its rotation to the pinion bearing oil port. The tapered roller bearings pump oil from their small diameter to the larger. There is a need for sufficient oil flow to the cavity between the pinion bearings to insure life of the bearings. There is also a return port for the pinion tail bearing to return oil back to the sump and not flood the pinion seal.

5.2 Design methodology

The term *hypoid* is derived from the fact that the pitch cones of the gears have a hyperbolic shape; more specifically, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

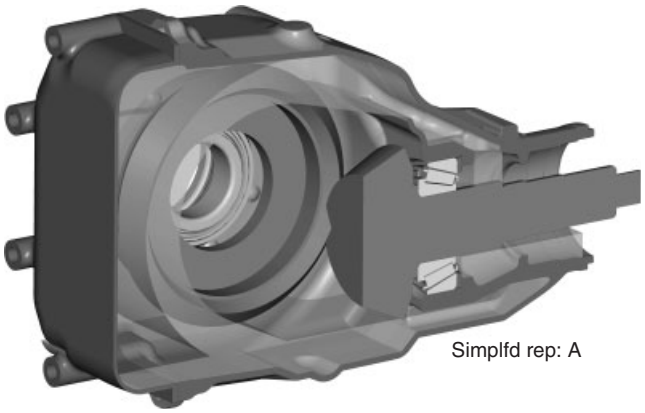


Figure 7. Axle oil flow illustration.

tooth form is created by a hypocycloidal curve. The hypoid gear set is very similar to a spiral bevel gear arrangement with the addition of an offset. The pinion axis of rotation does not intersect the gear axis of rotation, as shown in Figure 8. In this figure, two reference planes are displayed between the centerline of the ring gear and the axis of rotation of the pinion. The distance between these two planes is referred to as the hypoid offset. This offset factor also changes the paradigm of typical torque ratios equal to the ratio of the pitch diameters. As the hypoid gear teeth have different pressure angles and profiles of curvatures, it is possible to design an uncharacteristically large pinion even with a numerically large torque ratio. The offset allows such an arrangement to be possible and this allows for an overall stronger gear set.

The actual design equations and analysis are quite extensive with over 150 variables with approximately 30% of them requiring an experimental approach of trial and error with an iterative process until a suitable result is found. In order to have an appreciation of the amount of work that has been completed in this analysis area since the first commercial application of a hypoid axle in 1925 by Packard Automobiles, it is necessary to review some of the history behind hypoid gears. Even with the first hypoid available, the fundamental engineering behind the concept

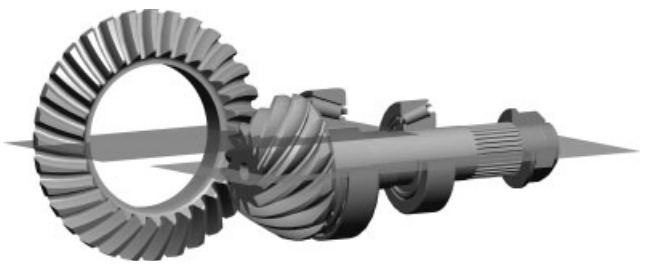


Figure 8. Hypoid offset illustration.

was not truly understood and no good design practices were in place for at least another 10 years. From the mid-1930s to -1970s, engineers applied physics and math to understand and develop kinematic relationships to design and manufacture these necessary gears for the automotive industry. To this day, these base equations, which are presented below, are at the core of the design optimization process for hypoid gears. This is a unique gear design problem, as besides the basic inputs of tooth combination, diameter of ring gear, pinion offset, and the like, the engineer needs to estimate the gear tooth cutter radius in the initial design process. Therefore, the engineer needs to be aware of the manufacturing tools and techniques upfront in the design process. Hypoid gears are the only gears that require this level of information in order to design the gears. From trial and error over the years, this is the only method that produces consistent results. Below are the five equations with five unknowns that need to be initially estimated and then iterated to find the exact solution for a given set of design inputs (Griffith, 1968).

$$A = \frac{1}{2}\left(\frac{D}{\sin\Gamma - F}\right) \tag{1}$$

$$r = A\left(\frac{n}{N}\right)\left(\frac{\sin\Gamma}{\cos\gamma}\right)\left(\frac{\cos(\psi - \varepsilon)}{\cos\psi}\right) \tag{2}$$

$$\sin\varepsilon = \frac{E}{(\cos\gamma\,(A\sin\Gamma + r\cos\Gamma))} \tag{3}$$

$$\begin{aligned}\tan\gamma = {} & \frac{r}{A}\left(\frac{\cos\psi}{\cos(\psi - \varepsilon)}\right) \\ & \times\left\{1 + \left[\frac{A}{r_c}\left(\frac{\cos\varepsilon}{\cos\psi}\right) - \tan\psi\right]\right. \\ & \left.\times\,[\tan\psi - \tan(\psi - \varepsilon)]\right\}\end{aligned} \tag{4}$$

$$\cot\Gamma = \frac{\tan\gamma}{\cos\varepsilon} \tag{5}$$

where the following five symbols are unknown:

$A =$ mean cone distance of gear (inch);
$r =$ distance from the mean point to the centerline of pinion, normal to the pitch cone (inch);
$\varepsilon =$ offset angle of pinion ($^\circ$);
$\gamma =$ pitch cone angle of pinion ($^\circ$);
$\Gamma =$ pitch cone angle of gear ($^\circ$).

where the following seven symbols are known:

$n =$ number of teeth of pinion;
$N =$ number of teeth of gear;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ψ = spiral angle of pinion (°);
r_c = mean radius of gear cutter (inch);
D = pitch diameter of gear (inch);
F = face of gear (inch);
E = offset of pinion (inch).

On the basis of the complexity of these calculations, they have been automated, to an extent, with computer-based tools. The most common of these is from the machine supplier, The Gleason Works. The other major tool and vendor is Oerlikon-Klingelnberg. Hypoid gears are so unique that the manufacturers of the equipment have also developed software tools to analyze the gear geometry along with the manufacturing tools. Some of these design tools do not produce stress values for ratios that are numerically below 1.8; therefore, if ratios lower than this are required, alternate stress analysis methods are required. These alternate methods can be finite element analysis or even altering offset to zero and temporarily modeling the gear set as a spiral bevel. These alternate methods are just estimates to help guide the design process and should never be used as absolutes.

On the basis of the complexity of the design process and the number of variables, this section will concentrate on the macro-level geometry and descriptions. There are four main geometry elements for an axle, of which two are adjusted during the assembly process. The first geometry feature is hypoid offset, which is the distance that the centerline of the pinion is away from the centerline of the ring gear. For typical automotive rear axle applications, the pinion is below the centerline of the ring gear. The offset dimension is typically 15–20% of the ring gear diameter. Hypoid offset is often given the symbol E and is a design parameter, but is not adjustable. The offset distance is machined into the cast center section of the axle. The next nonadjustable geometry feature is the shaft angle. For automotive applications, the ring and pinion axes are at 90° in relation to each other. This shaft angle is commonly referred to with the Greek symbol alpha, α, which is shown in Figure 9. This illustration shows references planes through the rotation axes of the gear and the pinion.

The next two geometry features are adjusted during the assembly process to achieve optimum durability and noise performance of the gears. These are pinion mounting distance, designated P, and ring gear mounting distance, designated G. The pinion mounting distance is the distance from the centerline of the ring gear to the pinion gear back face. This is illustrated in Figure 10 with two reference planes to depict the concept of pinion mounting distance. The ring gear mounting distance is the distance from the pinion centerline to the ring gear back face, as shown in Figure 11. Similar to Figure 10, except this time, the planes

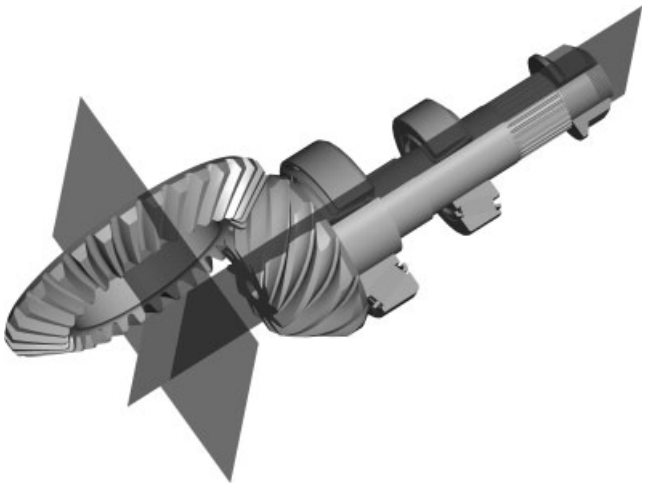


Figure 9. Shaft angle alpha illustration.

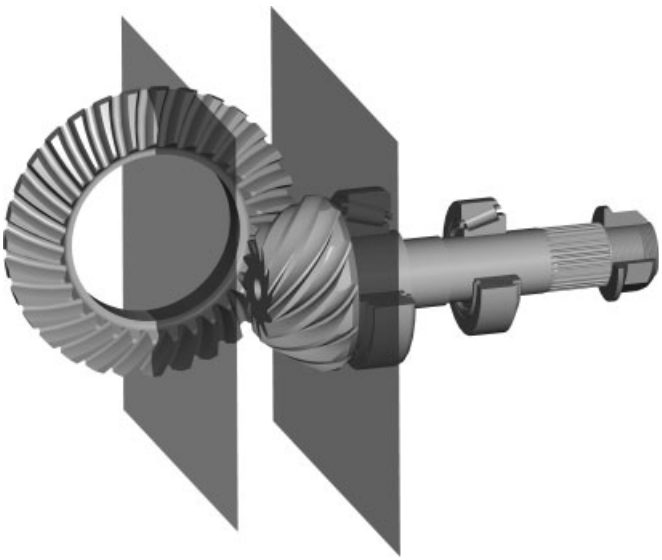


Figure 10. Pinion mounting distance illustration.

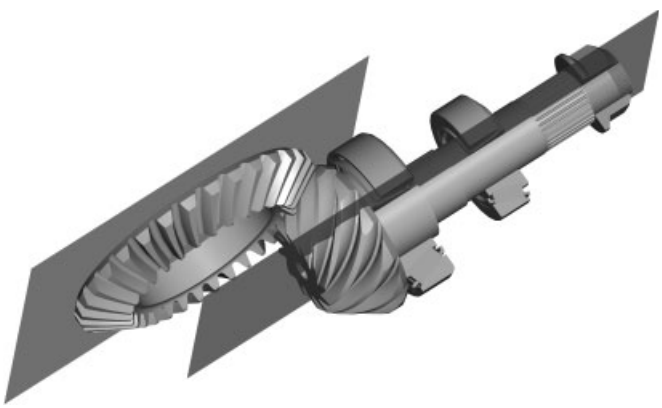


Figure 11. Ring gear mounting distance illustration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

are oriented to show the ring gear mounting distance. On the basis of difficulty of measuring at assembly, this dimension is measured and translated to an actual assembled gear mesh backlash value for assembly and service procedures. These two adjustments in gear mounting positions are so critical to correct operation of the gear pair that there is a special test performed on each gear set. This test is referred to as a *single-flank test* and determines the ideal location of these gears in relation to one another based on the manufacturing tolerances and final gear form. The pinion distance is recorded and select fit shimmed during assembly, whereas the ring gear mounting distance is translated to an actual backlash value for the gear to be installed. The main reason for translation gear mounting distance to backlash is to allow for ease of assembly and service. It is virtually impossible to measure ring gear mounting distance in the complete axle assembly, whereas ring gear backlash is relatively simple to measure.

The other unique geometry, in addition to offset, for a hypoid gear set is the spiral angle of the gear teeth. The spiral angle for the ring and pinion are opposite hand in order to operate correctly. Most automotive rear axle applications have the pinion located below the centerline of the ring gear with a left-hand spiral, whereas the mating ring gear is the opposite hand with a right-hand spiral angle. The typical spiral angle range is between 35° and 55° for the pinion.

6 BEARING CONSIDERATIONS

Any rotating members in a car and specifically axle assembly are supported by bearings. These bearings need to support and carry radial and axial loads as required. The main contributor for loads along the axis of rotation of the bearings is from the hypoid gear separation forces. The main contributor of radial loads is from the mass of the component's rotation and the weight of the vehicle. In this section, the loads associated from the hypoid gear mesh are reviewed. For the complete axle structure design, loads from the vehicle mass, wheels, suspension, cornering loads, and brakes must also be considered.

6.1 Hypoid load calculations

The hypoid gear arrangement creates three loads that are centered at the gear mesh point. These loads must be translated back to the bearing support structure based on the geometry and span of the bearings and axle layout. There are many different bearing support structures that have been used and range from straddle-mounted gears along with direct and indirect mounting strategies. These are all explained in the literature (Radzevich, 2012).

6.2 Preload

Most automotive axles utilize tapered roller bearings to support the thrust loads that are generated from the hypoid gear arrangement. There have been some recent axles that utilize tandem ball bearings; however, the main focus of this section is tapered roller bearings, and the theories are similar when tandem ball bearings are utilized. In order to properly locate and transfer load across the bearings, tapered roller bearings must be assembled with a fixed amount of preload, typically a value in the range 4000–6000 N (900–1350 pound force) is desired based on typical automotive size bearings. Of course, these values are just estimated for a typical light-duty automotive application. Another term to use instead of preload is interference fit, specifically that the bearings are initially installed in the axle in a situation where the bearings are forced into a smaller space than they would statically accommodate. The exact amount of interference to maintain the desired preload needs to be analyzed based on the characteristics of housing, shaft, and bearing deflection and fits, to name a few. There are many factors along with design and assembly methods that all contribute to the process for determining and setting bearing preload.

Preload can also be affected by the housing material selection, specifically when aluminum is utilized, as it has a greater rate of thermal expansion. For the pinion bearings, the higher coefficient of thermal expansion may be balanced, as the distance between the housing bearing seats increases (which increases the preload) and the bearing bores also increase in diameter, which decreases the bearing installed height (which decreases the preload). The net result is typically a balance back to the original preload but this is a function of the span between the bearing seats and the fit of the bearing cups in the housing. For the differential carrier bearings, the preload needs to be set higher to accommodate this thermal expansion in aluminum. As the axle temperature increases during operation, the housing will expand faster than the internal steel components, which results in a loss in bearing preload.

6.2.1 Pinion bearing preload

There are two common methods to achieve and control adequate pinion bearing preload. It is important to have the correct amount of preload without overloading the bearing. In order to properly space and preload the pinion bearings, there is a pinion spacer that is assembled in between the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

pinion bearing cones. At times, this pinion spacer is referred to as a preload limiter. The spacer also makes certain that there is a correct amount of pinion shaft elongation and subsequent clamp force on the drive flange. On the basis of the desire for ease of assembly for volume production, the pinion tail bearing cone is a slip fit to the pinion shaft. There must be enough clamp load of this bearing race between the spacer and the drive flange to guarantee that this bearing race does not rotate on the shaft.

The solid spacer-style preload limiter is one method for this clamp load and bearing race retention. This spacer is a select spacer that is selected during the assembly process. The exact length of the spacer is based on the exact measurements of all of the components involved, namely bearing heights, housing machining, pinion shaft machining, and press fit tolerance for shaft and housing. On the basis of these factors, the select fit spacer is chosen and assembled on the pinion. As a final check, the assembled pinion torque-to-rotate is measured and compared to a specific range based on the preload required, a typical range for this torque-to-rotate is 15–28 in-lbs (1.7–3.2 N-m). On the basis of the load that the spacer is subjected to, along with the pinion shaft bending while loaded, the spacer needs to be strong enough to not deform under these conditions. The solid spacer arrangement does allow the pinion nut to be tightened to large torque (i.e., 250 ft-lbs (340 N-m)) to make certain that the nut does not loosen over time.

The other method that is more common is to design and develop a spacer that can accommodate the range of manufacturing tolerances of all of the components while still maintaining proper bearing load under all conditions. The collapsible spacer is the typical method for this. This spacer is designed as a tube with a section that is prebent to allow the spacer to buckle in a controlled manner when loaded, as shown in Figure 12. This deflection allows the spacer to decrease in length while still providing a resistive load to the bearings. The spacer acts like a spring in the bearing mounting system to balance load versus deflection. The pinion shaft will deflect or bend under high loads and the collapsible spacer will maintain adequate bearing load throughout this deflection. This spacer is designed with a very specific and predictable compressive load versus deflection curve. This curve has an initial amount of deflection that achieves a high load and then the slope of the curve approaches zero. The flat portion of the curve allows the spacer to maintain the desired bearing load while still accommodating deflection in the system. Designed properly, the spacer is installed in a stable condition statically and maintains equilibrium throughout all dynamic operating conditions. On the basis of the physical arrangement of this spacer, during installation, the pinion nut is tightened until the desired torque-to-rotate on the bearings is achieved. Once achieved, the nut can no longer be tightened further or loosened without replacing the spacer; it is a single use spacer. As the nut cannot be tightened further, there must be a secondary mechanism in place to assure that the nut does not loosen or leak. The most common practice is to utilize a thread-locking sealant on threads and the nut flange to retain the nut and seal the interface.

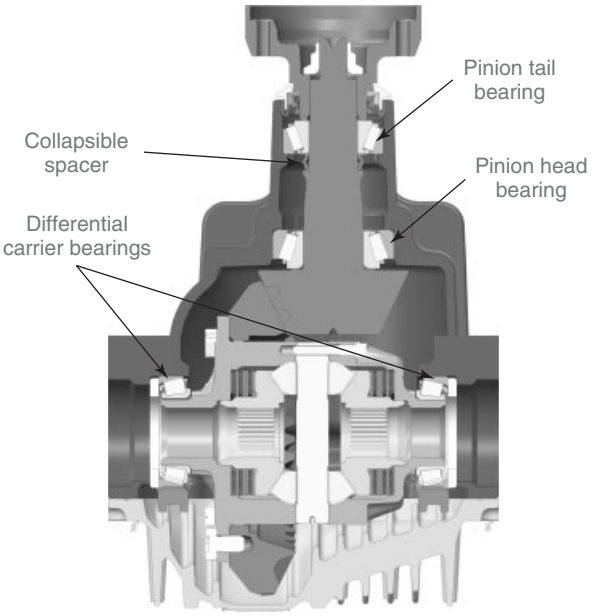


Figure 12. Axle center section cross section.

6.2.2 *Differential carrier bearings preload*

The differential carrier is supported by tapered roller bearings, and as stated earlier, these bearings need to have a predetermined amount of controlled preload in order to provide sufficient load-carrying capacity and durability. There are three common methods utilized to achieve carrier bearing preload. These are with large adjuster nuts, select fit snap rings and solid shims with bearing caps. There are some designs that even combine the snap ring fit for one side and bearing cap for the other. Along with developing the correct preload, the ring gear mounting distance is adjusted and verified with a backlash verification process. There is also a final visual check of the bench contact (low torque) of the gear in the final assembled position. If the contact pattern is incorrect, the gear set mounting and shims will be changed to obtain the correct contact pattern.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

7 MANUFACTURING METHODS

There are two mainstream manufacturing processes to cut hypoid gears. The traditional method of single indexing has been used for decades and was the most common until the late 1990s and early 2000s. The single indexing refers to the fact that a single gear tooth gap is completely cut at a time in the gear blank before the next tooth gap is cut. The second process is a continuous indexing process where both the gear cutter and the gear blank rotate while the tooth gaps are all partially cut or indexed. The gear teeth gaps are basically all completed at the same time. Independent of which gear manufacturing is employed, the typical process steps are generically as follows: gear forging, gear blank machining, tooth machining, heat treat, grind bearing journals, lapping, and finally, annealing of pinion threads. There may be additional processes required based on heat-treat distortion, length of pinion shaft, and other factors that may require pinion shaft straightening. For increased strength, there may be a shot-peening process and, of course, several part washing steps.

7.1 Face milling (five cut)

Traditional face milling is a single indexing process as described earlier and has been used to produce hypoid gears for quite some time. This process is often referred to as the *five-cut process*, as there are five discreet cutter heads and potentially five machines required to produce the ring and pinion gears for high volume production. The pinion requires three unique cutter heads; namely, one for rough cutting the tooth slot, another to finish cut the convex surface, and a third to finish cut the concave surface of the gear tooth form. The ring gear requires two cutter heads for finishing of the convex and concave surfaces of the gear tooth. This process is typically performed with a lubricant to increase tool life. The face-milling process produces a tapered tooth surface as seen in Figure 13. The tooth surface is outlined to highlight the trapezoidal nature of the surface. This process is adjusted by controlling and evaluating the tooth contact pattern during the manufacturing process. Specifically, the pinion tooth-form machining is adjusted to provide the desired position and shape of the contact pattern. These changes to the machining are not without problems, but the producers of these gears have learned how to handle these changes.

7.1.1 Face mill completing

There is a subset of face milling with a slight difference in that the entire tooth profile is cut with a single cutter head and is very similar to the face-hobbing process regarding material per cutter head. On the basis of this, only two cutter heads are required to cut the ring and pinion gears. However, the face mill completing process still removes material one tooth gap at a time. This process is usually the precursor for grinding operation.

7.2 Face hobbing (two cut)

Face hobbing is a gear-cutting process that is known as *continuous indexing* and became more popular in the late 1990s to modern day. There were a few changes to the manufacturing process that were all introduced along with face hobbing. The ability to cut the gear teeth without lubricant or dry-cutting along with a reduction in tooth-cutting time by approximately 30% was also enabled by face hobbing. With the introduction of high speed carbide cutter blades, the blade life was increased by almost a factor of 3. Lastly, the fact that only two-cutter heads and machines were required as compared to the five for face milling helped this process become the mainstream for the automotive industry. There are still legacy machines and components that are using the face-milling process; however, most new designs are face hobbed. The face-hobbing process produces a parallel tooth form as compared to the tapered tooth of the milling process as shown in Figure 14.



Figure 13. Face milled tooth surface.



Figure 14. Face-hobbed tooth surface.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

7.3 Lapping/grinding

After the gears are heat treated, there are two postheat-treat or hard-finishing processes that can be used to aid in final finishing and improving of the gear tooth surfaces. The first and the most common is a lapping process. For this process, a very abrasive material, specifically a silicon carbide material in oil, is put in the gear mesh while the ring and pinion gears are rotating under light torque. The gears are not only rotating but also being slid along the contact pattern to evenly remove any surface defects and *wear* the gears together. As the pinion acts as the tool to finish the gear and the gear as the tool to finish the pinion, the gears are a matched set after the lapping process and must stay together. It is no longer possible to mix and match the ring gear or pinion after lapping. The lapping process also created some desirable microstructure to aid in overall noise reduction of the gear set. As the pinion and ring gears are used together during the lapping process, a completed gear set is delivered and there is no subsequent machining required.

The second process is a grinding process, where the material that is removed is controlled and brings the gear tooth profiles back to design intent and corrects for manufacturing and heat-treat deviations. In order to grind the tooth profiles, they must be face milled or produced from a single-indexed process. It is not possible nowadays to grind gears that were produced from a continuous indexing process. A ground gear set is not a matched set and the ring and pinion gears can be intermingled. There are trade-offs with both of these processes. If the heat-treat process and subsequent distortion is controlled reliably and minimized, small amounts of material are only required to be removed and lapping is preferred. Contrarily, if large heat distortion and larger amounts of material are required to be removed, then grinding is preferred.

REFERENCES

General Motors Corporation (1951) *New Departure Handbook*, vol. II, 7th edn, General Motors Corporation, Bristol Connecticut, USA.

Griffith, B. (1968) Hypoid gear design for automotive axles. SAE International 680076, Warrendale, PA, USA

Radzevich, S.P. (2012) *DUDLEY'S Handbook of Practical Gear Design and Manufacture*, 2nd edn, CRC Press, Boca Raton, Florida, USA.

Stadtfeld, H.J. (2000) *Advanced Bevel Gear Technology*, The Gleason Works, Rochester, New York, USA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 GKN Driveline North America Inc.
DOI: 10.1002/9781118354179.auto112
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Supercharging

Scot Streeter[1] and Matt Swartzlander[2]

[1]*Eaton Corporation, Galesburg, MI, USA*
[2]*Eaton Corporation, Marshall, MI, USA*

1 Introduction 1
2 Compressor Design 1
3 Performance Measures and Mapping 5
4 Engine Matching 7
Reference 15
Further Reading 15

1 INTRODUCTION

By definition, the term *supercharging* as related to internal combustion engines is to increase the mass of charge air in the cylinder beyond the level capable by natural aspiration. In combination with the appropriate air/fuel ratio (AFR), the combustion pressures and useful specific work output are increased beyond the levels achievable for nonsupercharged applications. Typically, supercharged applications are discussed in terms of boost pressure referring to the peak pressure the device creates in the intake manifold. It is common to hear in North American engineering circles discussion of an application making numbers such as "12 pounds" of boost or "15 pounds" of boost when the actual unit of measure is PSI (pounds per square inch). These numbers are actually gaging pressure numbers above atmospheric pressure levels, hence, the term *boost* and why these performance levels cannot be achieved by naturally aspirated operation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

There are several ways to provide the work input to the supercharging device in applications that can be generally divided into two categories: mechanical and fluid drives. The term *supercharger* is generally associated with mechanical-drive devices, whereas the term *turbocharger* is the most common approach to a fluid-drive system. The scope of this chapter is to focus on the aspects of the device typically termed a *supercharger*.

2 COMPRESSOR DESIGN

The roots type compressor in its most basic form consists of an input shaft driving a pair of counterrotating rotors or impellors to move air from the inlet to the outlet of the device. As there is no internal compression in the roots device, a simple bypass loop can effectively eliminate unnecessary mass flow through the device when the operating conditions of the engine do not require the additional mass flow to meet the break mean effective pressure (BMEP) targets for the requested speed/load point demanded by the operator. When the bypass is fully open, the parasitic losses associated with the device are limited to the internal frictional losses and any throttling losses associated with the bypass flow restrictions. While these numbers are quite small relative to other parasitic losses is the system, the incorporation of a clutching device in the system can completely eliminate these losses when the supercharger is disengaged from the drive system (Figure 1).

2.1 Drive input system

The device is directly driven as an accessory on the engine most commonly via the Front Engine Accessory Drive

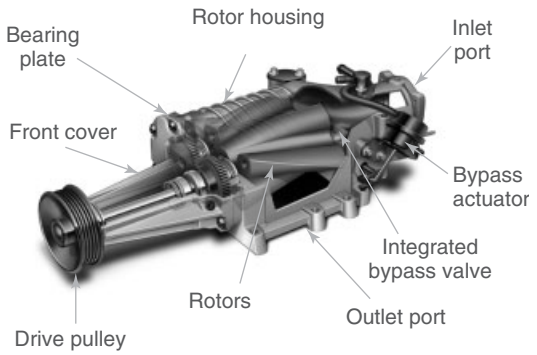


Figure 1. Supercharger component detail. (Reproduced by permission of Eaton Corporation.)

(FEAD) belt system. It is either in the general serpentine routing or potentially driven with a separate belt via a double pulley at some other location on the FEAD. For most engine applications, the supercharger belt ratio will generally increase the supercharger speed from 1.7 to 2.5 times the engine crankshaft speed. For gasoline, these ratios are typically sufficient when the supercharger is employed alone for the application. When a supercharger is used in conjunction with a turbocharger in a compounded configuration, the supercharger is typically clutched out at roughly half engine speed and, as such, a much higher overall drive ratio is required. For these types of applications, internal step-up gearing may be utilized within the supercharger or an additional belt ratio may be applied with a three-pulley system. This is also true for compression ignition where the peak engine speed is typically much lower than a gasoline spark ignition engine. Figure 2 shows an internal step-up gear configuration for a supercharger.

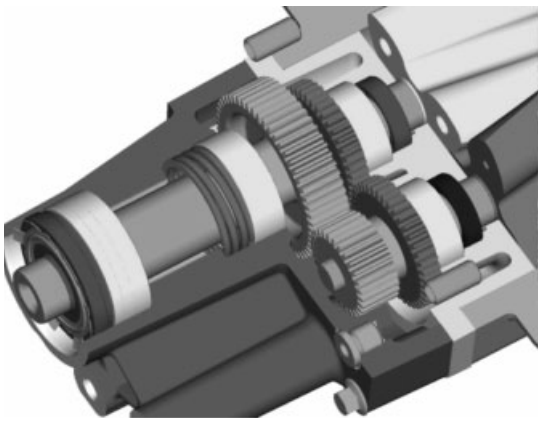


Figure 2. Supercharger step-up drive configuration. (Reproduced by permission of Eaton Corporation.)

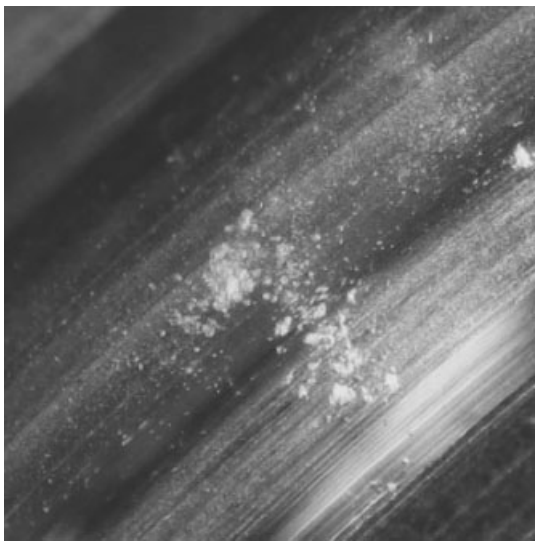


Figure 3. Supercharger bearing spalling. (Reproduced by permission of Eaton Corporation.)

The pulley on the supercharger is coupled with an input shaft supported on two bearings to provide the torque to drive the two rotors. The length of the input shaft, and subsequently the "snout" of the supercharger, can be varied in length to provide belt access while positioning the device in the optimal location of the engine. Care must be taken to assess the dynamic belt loads on the pulley to get the correct bearing design characteristics and assure that the application life requirements are met. If these loads are too high, given the overall duty cycle for the application and the bearing speed/load capacity, bearing raceway pitting can occur ultimately leading to premature bearing failures. Figure 3 shows the onset of microspalling in a supercharger pulley bearing raceway.

Additional areas of concern are that the cantilever loading does neither cause excessive stress in the front cover of the device nor cause a sealing issue in the joints between the bolted components of the device. The assessment of these conditions is typically assessed via finite element method (FEM) to ensure a robust design.

2.2 Input drive coupling

The input shaft system is connected to the rotor timing gears via a coupling that provides the necessary compliance in the appropriate degrees of freedom to couple the input shaft and rotor bearing systems. While a direct coupling between the input shaft and the timing gears is feasible, as shown in Figure 3, this is not desirable because the direct connection of the engine torsional excitation to the timing gears can cause gear rattle during unloaded

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

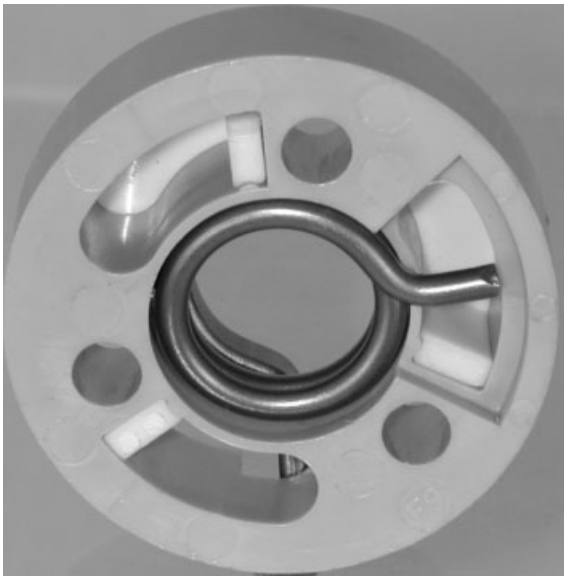


Figure 4. Supercharger torsional isolator. (Reproduced by permission of Eaton Corporation.)

operation. Most modern original equipment manufacturer (OEM) superchargers contain a torsional isolator between the input shaft and the timing gears to decouple firing torsional frequency from the gears (Figure 4).

These isolators are typically developed with dynamic systems models for the appropriate excitation frequency attenuation given the coupled inertia in the system.

2.3 Timing gears/lubrication

The timing gears are included in the system to prevent contact of the rotors during operation. As the rotors are designed to minimize the rotational inertia of the system, these types of materials do not make good load bearing candidates and will spall if in intimate contact. To prevent the contact of the rotors, the gear pair backlash is designed to be tighter than the rotor mesh backlash. The gear tolerances and the housing system tolerances that control gear and rotor center distances must be very tightly controlled during the manufacturing process. The gear design must account for the root stress, the contact stress, and the transmission error as they relate to gear acoustic performance (Figure 5).

Figure 5 shows a pair of supercharger timing gears on a vibration signature dynamometer.

Most modern OEM superchargers in production today use a very tightly toleranced profile quality spur timing gear. While helical gears can improve the contact ratio, timing the helical gears in concert with helical rotors is

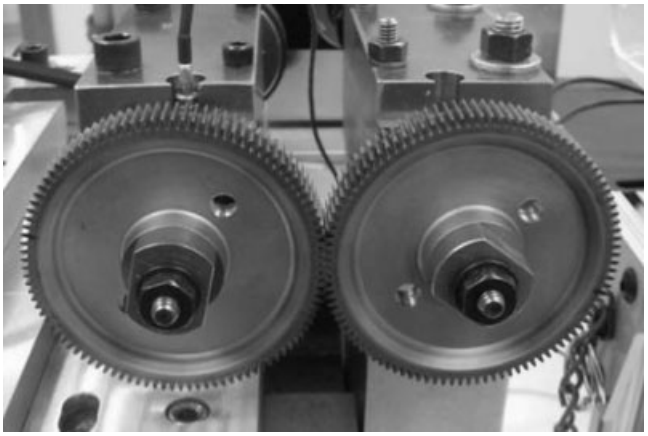


Figure 5. Gear torsional excitation assessment. (Reproduced by permission of Eaton Corporation.)

a very difficult process because of the thrust loads and deflection encountered during assembly. As the power transmission through the device is carried by the gears, the gear cavity is typically oil lubricated. While oil could be supplied via a feed and return line to the engine, it is very common for the supercharger to have a self-contained oil supply utilizing a high quality synthetic lubricant. These types of devices are considered "sealed for life" and require no maintenance for the life of the engine/vehicle. The gearbox must be sealed to retain the lubricant and the rotor shaft seals in the application have to survive with shaft speeds in excess of 20,000 rpm. In addition, the seals and churning losses in the oil are contributors to the parasitic losses and need to be designed specifically to minimize friction and viscous shear losses for the application to maximize efficiency.

2.4 Rotors

The timing gears then drive the rotors that actually pump the air from the inlet to the outlet of the device. Key design considerations to maximize efficiency are the rotor lobe geometry, rotor helix angle, number of lobes, rotor-to-rotor clearance, rotor-to-housing clearance, and the inlet/outlet port sizes and shapes. The system clearances directly impact the leakage between air transport volumes in the system and contribute negatively to low speed efficiency more so than high speed efficiency. Conversely, the porting of the device, as well as the inlet and outlet plumbing of the entire engine induction system, can cause choking because of either higher pressure drop or inefficient fill/expel events and generally have a greater negative impact on high speed efficiency. The air transfer process can be

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

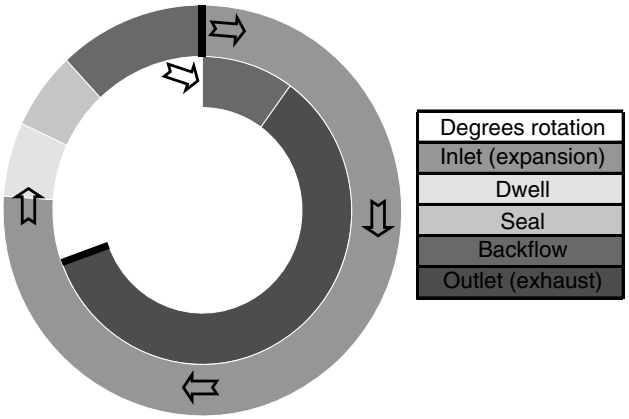


Figure 6. Supercharger air-handling timing diagram. (Reproduced by permission of Eaton Corporation.)

categorized into low pressure and high pressure events as follows.

1. Low pressure expansion—The time the transfer volume starts to expand until it comes completely out of rotor mesh.
2. Low pressure dwell—The time the transfer volume is completely expanded during the inlet process.
3. Low pressure seal—The time the transfer volume is sealed from the inlet and outlet.
4. High pressure backflow—The time the transfer volume is initially exposed to the high pressure manifold.
5. High pressure discharge—The time the transfer volume is completely exposed to the outlet port.

Transport time is a function of rotational speed and is typically referred to in degrees to take the time domain out of the equation. This helps with simplicity of explaining the geometric events but does negate the fluid dynamic impact on these events (Figure 6).

To optimize the rotor design, we want to control changes in the mass momentum of the charge air to the greatest extent possible. This can be achieved by matching the transport volume expansion rate with in inlet mass flow rate to avoid excessive changes in momentum (Figure 7).

Figure 7 demonstrates this goal of keeping velocity $V3 \leq$ velocity $V1$.

Figure 8 shows the dramatic improvement in these characteristics that Eaton Corporation achieved with its TVS® supercharger relative to their previous generation.

As mentioned earlier, the lower speed efficiency is governed by clearance in the device and leakage between air transport volumes. To optimize the efficiency at lower speeds, coating can be utilized to minimize the clearances in the device while still preventing metal-to-metal contact that would induce spalling and subsequent durability concerns. One innovation addressing this concern is the use of an abradeable coating on the supercharger rotors that wear into a match fit on the device. The rotors are coated before assembly and then assembled as interference fit. As the unit is rotated during a break-in operation, the coating then abrades and is expelled from the unit yielding an essentially zero clearance configuration. The volumetric efficiency improvement typical of this type of coating is shown in Figure 9.

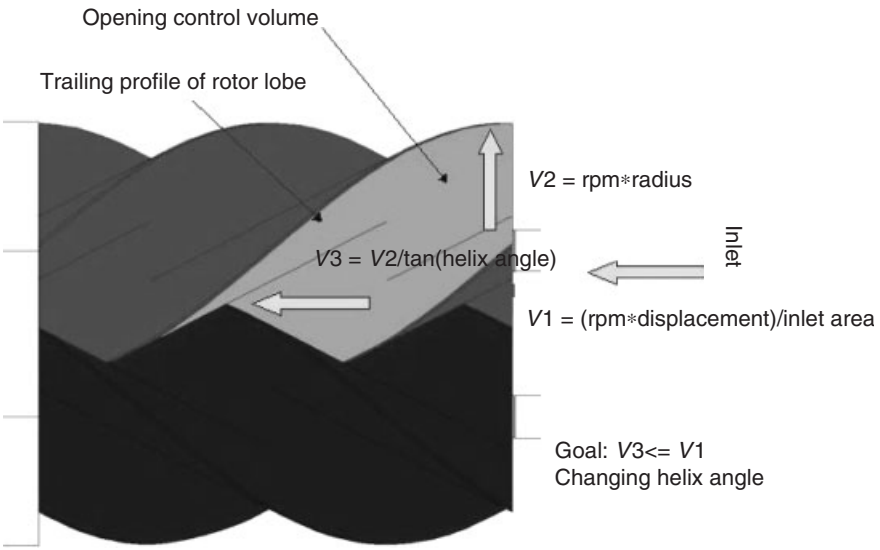


Figure 7. Mass air velocity optimization. (Reproduced by permission of Eaton Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

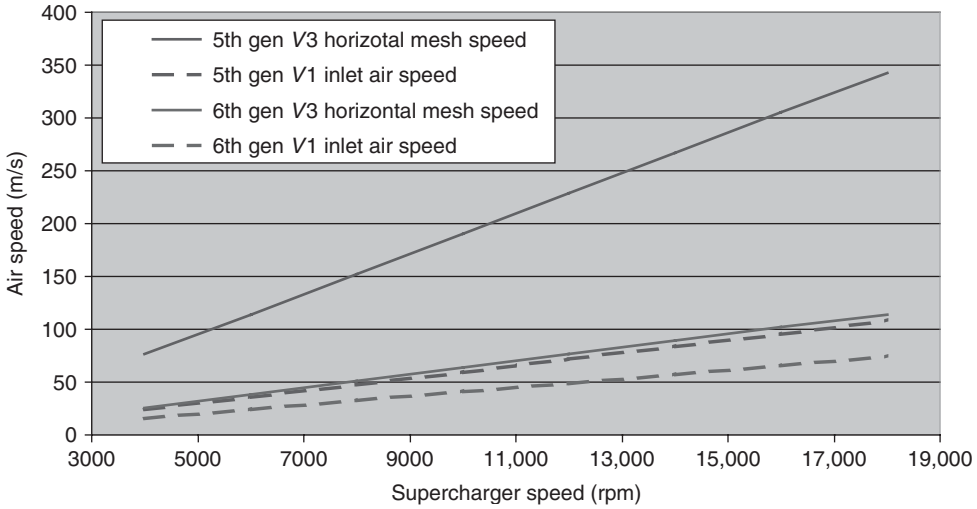


Figure 8. Inlet charge velocity to geometric volumetric expansion comparison. (Reproduced by permission of Eaton Corporation.)

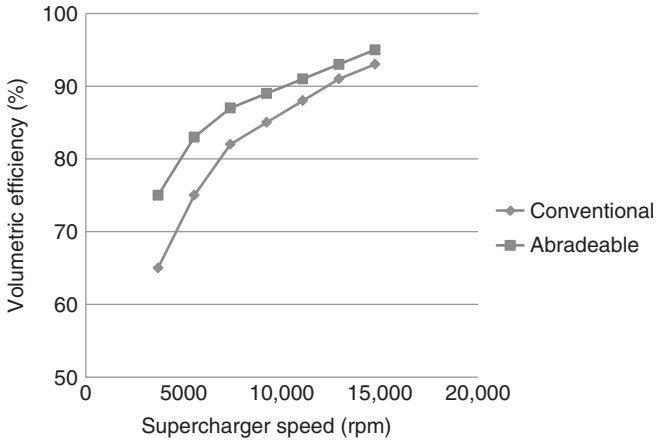


Figure 9. Volumetric efficiency comparison of coatings. (Reproduced by permission of Eaton Corporation.)

2.5 Ports

The inlet and outlet port shapes are primarily governed by the rotor configuration. If the ports are small compared to the optimum rotor timing, you will choke the flow or allow insufficient duration for the intake and exhaust events to occur. If the ports are excessively large compared to the rotor geometry and timing, you will generate leakage between the inlet and exhaust transport volumes. In Figure 10, the inlet ports show the difference between available port area for high and low twist rotors.

On the outlet port, the port edge is typically aligned with the rotor tip helix and exposes roughly half of the rotor length. With sufficient twist in the device, there is actually communication between transport volumes during the backflow portion of the exhaust event. This helps reduce the momentum of the high pressure in-rush of air to the exhaust volumes and minimizes the pulsation magnitude in the outlet. An exhaust (outlet) port is shown in Figure 11. The position of the outlet port can be used to tune the peak efficiency of the supercharger based on pressure ratio. The smaller the port, better efficiency is created at higher pressure ratios, the larger the port, better efficiency is present at lower pressure ratios. Setting the port mid rotor bores provides a good compromise for most applications (Figure 11).

The backflow geometry for pulsation control is shown in Figure 12.

3 PERFORMANCE MEASURES AND MAPPING

The performance of a supercharger is primarily gaged with two primary measures: volumetric efficiency and isentropic (thermal) efficiency. The measurement techniques are typically governed by industrial standards such as the Society of Automotive Engineers (SAE) standard J-1723. These standards provide consistency in test configurations to provide comparable results across the industry.

By design, roots type and screw compressors are positive displacement pumps and have a fixed geometric volume of displacement per revolution of the device. The typical descriptions of the devices incorporate this displacement to denote the size of the device. This may be listed in cubic inches, cubic centimeters, or liters of displacement per

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

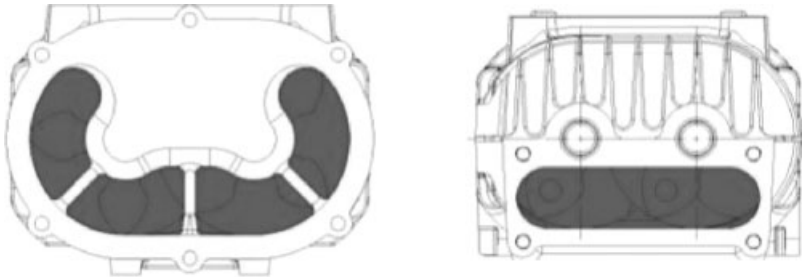


Figure 10. (a, b) Inlet port geometry optimization. (Reproduced by permission of Eaton Corporation.)

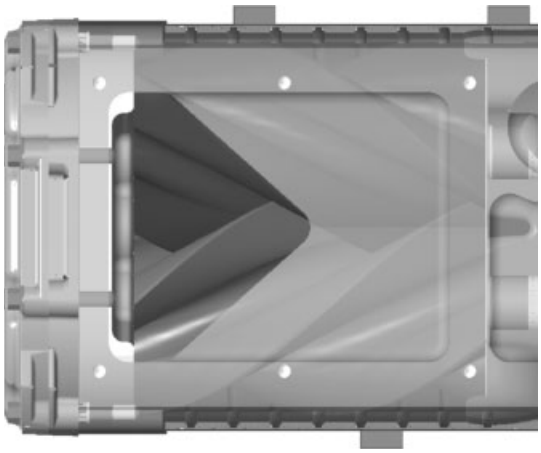


Figure 11. Outlet port geometry. (Reproduced by permission of Eaton Corporation.)

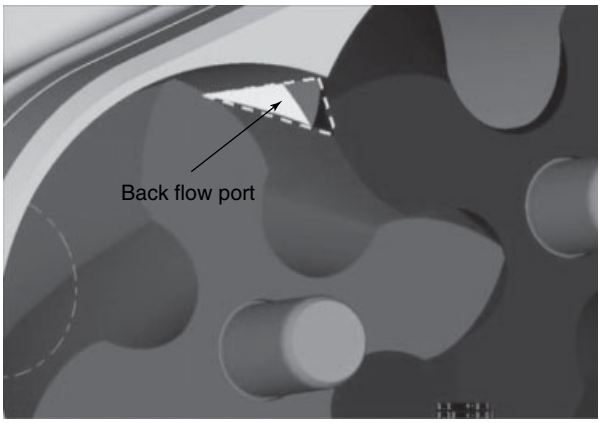


Figure 12. Internal backflow port geometry. (Reproduced by permission of Eaton Corporation.)

revolution. Before the introduction of Eaton's TVS supercharger, they would denote a model in terms of cubic inches of displacement per revolution (e.g., M90 = 90 in^3/rev). With the TVS supercharger, the nomenclature switched to cubic centimeters of displacement per revolution. An equivalently sized supercharger generating 90 in^3/rev displacement would be equal to 1475 cm^3/rev with the new nomenclature. The volumetric efficiency of the device is the ratio of the actual mass of ingested air into the device divided by the theoretical mass of air that would fill the fixed geometric volume of the device per revolution.

$$\eta_v = \frac{m_a}{\rho_{a,i} \cdot V_d}$$

where η_v is the volumetric efficiency, m_a is the actual air mass ingested, $\rho_{a,i}$ is the inlet air density, and V_d is the geometric volume of displacement per revolution in the positive displacement device (Equation from SAE J1723, 1995; SAE International, 1995).

The actual performance is affected by many things in the design including the ductwork feeding and exhausting the device, the inlet and outlet port design, and the rotor design features.

The isentropic efficiency of the device is quantified effectively by the thermal rise across the device compared to a completely isentropic compression process. The relationship is as follows:

$$\eta_T = \frac{T_{inlet} \cdot \left(\frac{P_{outlet}}{P_{inlet}}\right)^{\left(\frac{\gamma-1}{\gamma}\right)} - T_{inlet}}{T_{inlet} - T_{outlet}}$$

where η_T is the isentropic efficiency of the device, T_{inlet} and T_{outlet} are the inlet and outlet absolute temperatures, P_{inlet} and P_{outlet} are the inlet and outlet absolute pressures, and γ is the specific heat ratio of the air pumped through the device (Equation from SAE J1723, 1995; SAE International, 1995).

The efficiency varies with the speed and pressure ratio of the device and is typically assessed on a map similar to the following (Figure 13).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

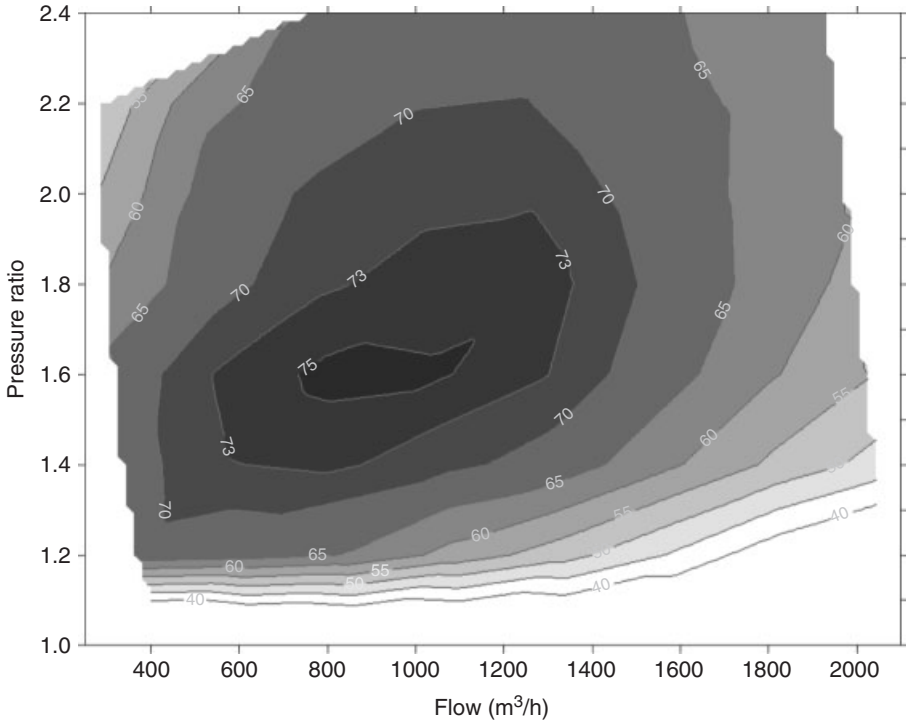


Figure 13. Isentropic compressor efficiency. (Reproduced by permission of Eaton Corporation.)

The actual isentropic efficiency in the final application also is affected by many variables, differing from the standardized in the test environment including duct geometry and atmospheric conditions. Within the device, the thermal efficiency is affected by inlet and outlet port geometry, rotor design, and particularly internal leakage within the device.

4 ENGINE MATCHING

Internal combustion engines are available in a wide range of sizes to fit many applications. Even considering the scope of this encyclopedia is limited to automotive engine design applications, once you consider the global market the variety is still vast. Engine displacements can range from a few hundred cubic centimeters for A-segment mini-cars dominant in Asia-Pacific markets to seven or more liters in J-segment sport utility vehicles (SUVs) and full-sized heavy-duty pickup trucks. This range of engine displacements coupled with the range of maximum engine speeds creates a spectrum of applications for potential supercharging. A supercharger needs to be matched to the application providing the desired mass air flow for the application. Figure 14 shows a typical relationship between engine displacement and the Eaton TVS supercharger displacement.

The range of the supercharger displacement for a variety of engine sizes is affected by primarily pulley ratio selection and bypass strategy.

4.1 Drive systems

The main drive system for a supercharger is either a belt drive or a gear drive. Belt drives are far more common than gear drives because of the ease of packaging and lower cost of integration. Gear drive systems would effectively require that the supercharger be integrated in either the front crank or cam system or the rear of the engine and require significant integration in the base engine design. While a gear drive system would be a more mechanically efficient solution, a belt drive approach allows the supercharger to be treated more like an accessory allowing the device to be mounted further away from the main engine rotating shaft systems, thus a majority of the OEM supercharger applications are driven via belt drive. Figure 15 shows a couple of types of belt routing.

Supercharged OEM engines are utilizing micro-V belts for supercharger drive systems for noise, efficiency, and durability. Depending on the application, the V belt/pulley system may range from a 3 V/groove system to a 12 V/groove system or even larger depending on the size

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

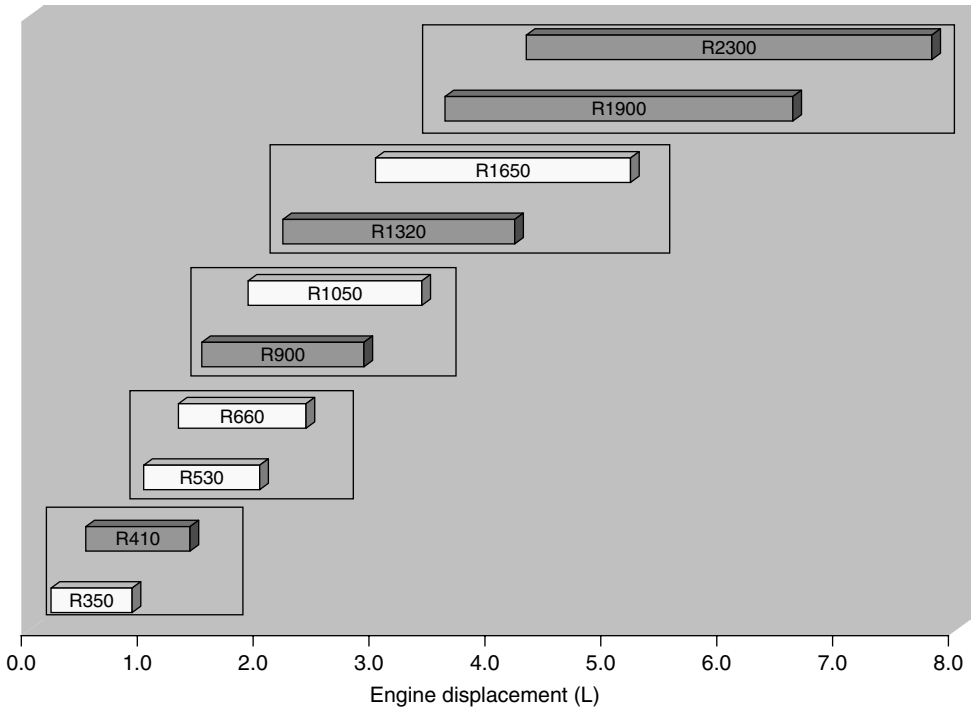


Figure 14. Supercharging sizing based on engine displacement. (Reproduced by permission of Eaton Corporation.)

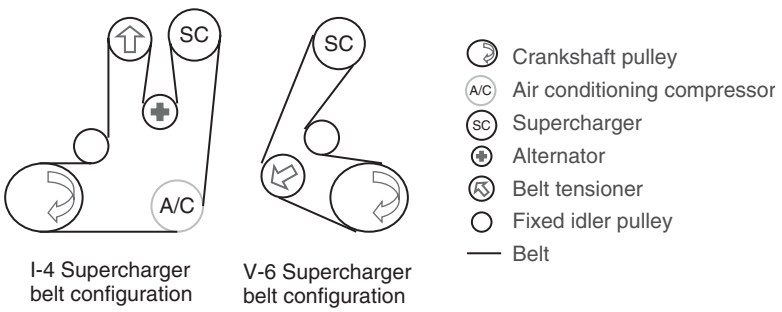


Figure 15. Front engine accessory drive configurations. (Reproduced by permission of Eaton Corporation.)

of the engine and supercharger. The peak power required to drive a supercharger can range from 5 to 150 kW for engines ranging from 0.5 to 7.0 L of displacement. Although this work is effectively recovered in positive pumping work during the intake stroke of the engine, the supercharger pulley and belt still need to be able to transmit this power. In addition to the width of the belt, the belt wrap angle is another key consideration in applying the supercharger to the application. A greater belt wrap angle increased the force transmission before the transition from static to dynamic friction in the belt/pulley interface. Most supercharger applications have belt wrap angles in excess of 180° to minimize the required belt width and handle the torque requirements to drive the device (Figure 16).

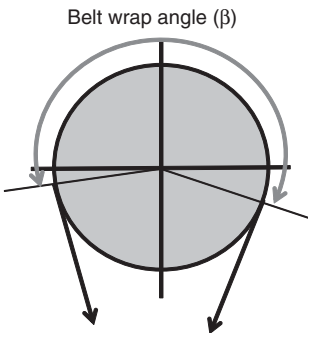


Figure 16. Belt wrap requirements. (Reproduced by permission of Eaton Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.2 Dual boosting configurations

1. While superchargers can be used as single boosting devices in a particular application, it is possible to use superchargers in conjunction with turbochargers to change the operating characteristics of the system and increase the power density of the application. Figure 17 shows the air routing for a super-turbo application.

The theory of operation is that the supercharger provides the benefit of quick transient response during low engine speed when the exhaust energy is low. At intermediate engine speeds, a controller starts to bypass the supercharger and both boosting devices work in concert. Then, at the highest engine speeds, the supercharger is completely bypassed and typically clutched out, whereas the turbocharger handles the boosting duties alone. This type of application could allow the use of a more cost-effective waste-gated turbocharger compared to the relatively more expensive variable geometry turbochargers (VGTs) on the market nowadays.

An alternative configuration is a turbo-super arrangement where the turbocharger is placed in advance of the supercharger. The schematic of this system is shown in Figure 18.

In this application, the supercharger multiplies the boost provided by the turbocharger by the supercharger's pressure ratio. As the supercharger is a positive displacement device, the pressure ratio across the device is constant regardless of the inlet pressure. This configuration also allows for quick transient manifold pressure increases, such as the super-turbo configuration.

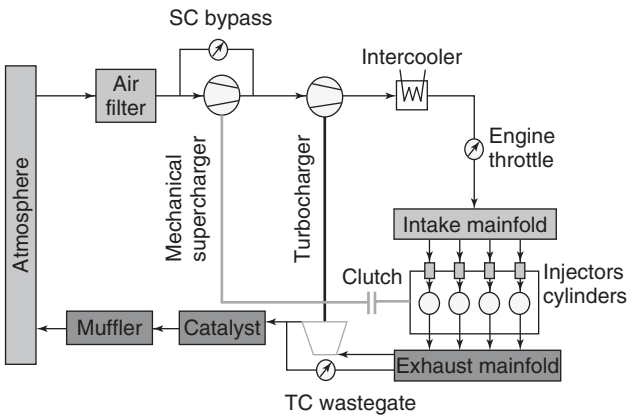


Figure 17. Supercharger–turbocharger configuration. (Reproduced by permission of Eaton Corporation.)

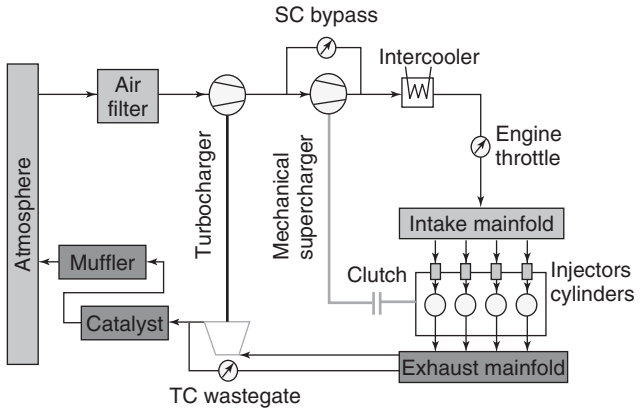


Figure 18. Turbocharger–supercharger configuration. (Reproduced by permission of Eaton Corporation.)

4.3 Bypass configurations

Bypass circuits are required to be used in supercharged induction systems when the engine does not require the mass flow rate that the supercharger is capable of generating. The bypass circuit can be used to control the manifold pressure to the appropriate pressure ratio level and the mass air flow sensor output allows the engine control unit (ECU) to determine the correct fueling rate. Typically, a supercharger is completely bypassed for BMEP load levels of 10 bar or less and the engine operates effectively naturally aspirated. The bypass circuit either can be an external loop outside of the supercharger itself or can be integrated directly in to the supercharger. Figure 19 shows a pneumatic bypass actuator and an integrated bypass butterfly valve on the left side of the model.

Bypass actuators are typically either pneumatic or electrically actuated throttles. The pneumatic actuators are proportionally balanced to close the bypass valve as the intake manifold pressure approaches atmospheric pressure,

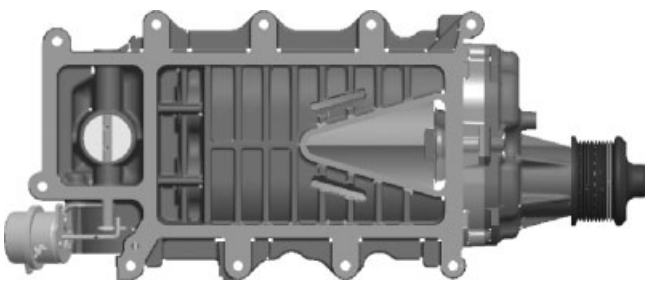


Figure 19. Integrated bypass configuration. (Reproduced by permission of Eaton Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

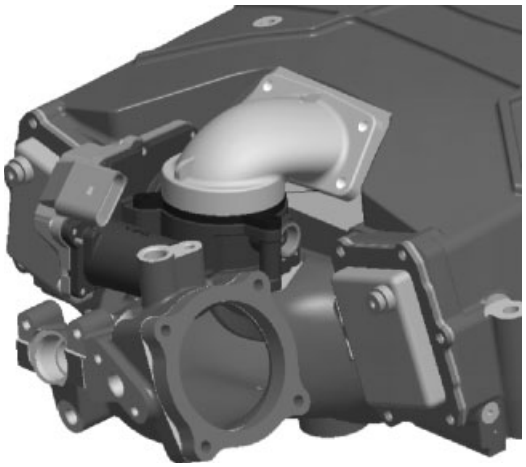


Figure 20. External bypass configuration. (Reproduced by permission of Eaton Corporation.)

which is in relationship to the main intake throttle's position. While the approach is very cost effective, it does not allow for manipulation of the boost profile in an active bypass control scenario. One way to use an active bypass control approach is to bleed excessive boost pressure to maintain a flat brake-torque level at higher engine speeds to control loads to drivetrain components. In addition, this also allows for a slightly larger supercharger than the application may require to further increase low speed torque levels and enable more aggressive down speeding for improved fuel economy. Figure 20 shows an integral bypass circuit coupled with an electric actuator in a complete upper manifold system.

Pressure drop in the bypass circuit is a key consideration for parasitic loss control during bypassed operation. Optimization techniques such as computational fluid dynamics (CFD) analysis are often utilized to minimize the losses associated with flow in bypass circuits (Figure 21).

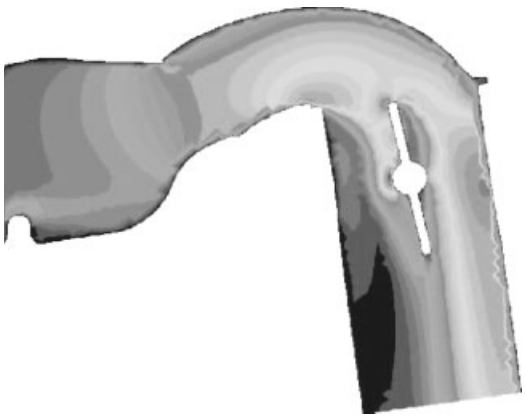


Figure 21. Bypass flow optimization. (Reproduced by permission of Eaton Corporation.)

4.4 Supercharger efficiency

The volumetric and isentropic efficiencies of a supercharger in application are controlled by the design characteristics of the supercharger itself and the way that it is applied to an engine. In this chapter, it is assumed that the engine manufacturers understand the effects of ductwork upstream and downstream of the device, as well as the charge air cooling strategy on overall engine performance. Focus is placed on the device itself in this discussion of supercharger efficiency.

The roots type supercharger has typically been characterized as a devise most suitable for low pressure ratio applications, but the resultant performance characteristics of the device can be drastically altered with careful design. This gives the ability to adjust the performance to the needs of the application. The characteristics that have the greatest effect on the isentropic and volumetric efficiency are the number of rotor lobes, lobe helix angle, outlet port geometry, rotor length, and rotor diameter. All these characteristics are interrelated. As shown in the following example, if you fix the number of lobes, the diameter, and the twist but vary the length of the rotor, then helix angle changes (Figure 22).

These parameters can be tuned to optimize the performance of the device for a particular application. For example, a super-turbo application will have significantly different optimal supercharger characteristics that a supercharger only application with this former preferring a greater volumetric efficiency at low engine speeds whereas the later prefers greater isentropic efficiency through a broad speed range. Another application, like supercharging a diesel engine, may desire higher pressure ratios that increased isentropic efficiency can enable. Tools such

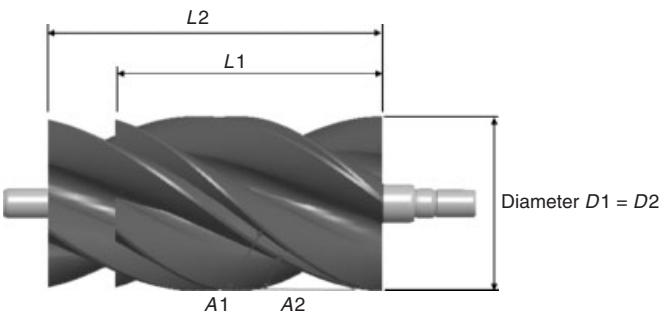


Figure 22. Rotor helix angle definition. (Reproduced by permission of Eaton Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

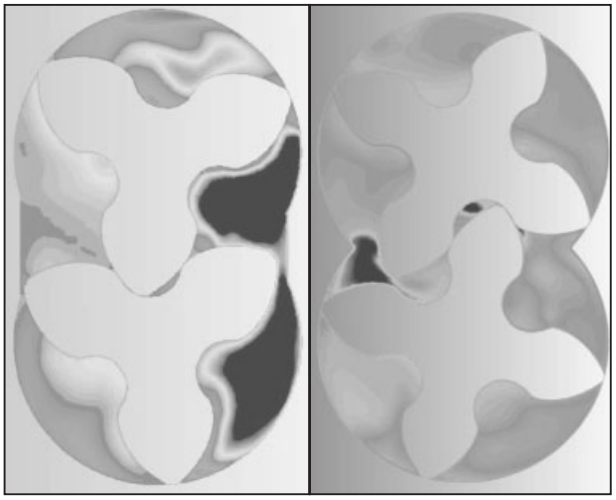


Figure 23. CFD device comparison. (Reproduced by permission of Eaton Corporation.)

as CFD can be utilized to virtually assess the effects of variable on the desired performance of the device (Figure 23).

The optimization of the design can have a dramatic effect on the performance characteristics of the device. Figure 24 shows a comparison of equal displacement superchargers with an optimization of volumetric efficiency at low speed.

The impact on performance varies with pressure ratio and speed, giving the designer the ability to push the high efficiency area of operation around the supercharger map. Figure 25 shows the effect of helix angle on isentropic efficiency at 1.4 and 1.8 pressure ratio operating conditions.

With optimization possibilities, we can target low speed volumetric efficiency for compound boosting applications, higher pressure ratio operation for compression ignition, and higher speed efficiency for supercharged only optimization (Figure 26).

The developments in this area of supercharger development have effectively changed the way that engine and vehicle manufacturers should think about traditional supercharging.

4.5 Supercharged performance

The ultimate goal of engine boosting is to efficiently increase the power/torque density of a specific engine application. With optimization of the engine, the use of charge air cooling, and appropriate selection of the boosting device, significant increases in power density can occur for a given engine displacement. Figure 27 compares the change in power and torque as the result of adding at 1900 cc/rev supercharger with an intercooler to a gasoline 6.2 L V-8 application. In this application example, there is a 50% increase in brake torque over most of the operating speed range compared to the naturally aspirated engine and a 46% increase in peak power (Figure 27).

Traditionally, power density levels of 75 kW/L have been considered well optimized for a boosted engine, but with proper design of the entire engine levels up to, or beyond, 150 kW/L can be feasible.

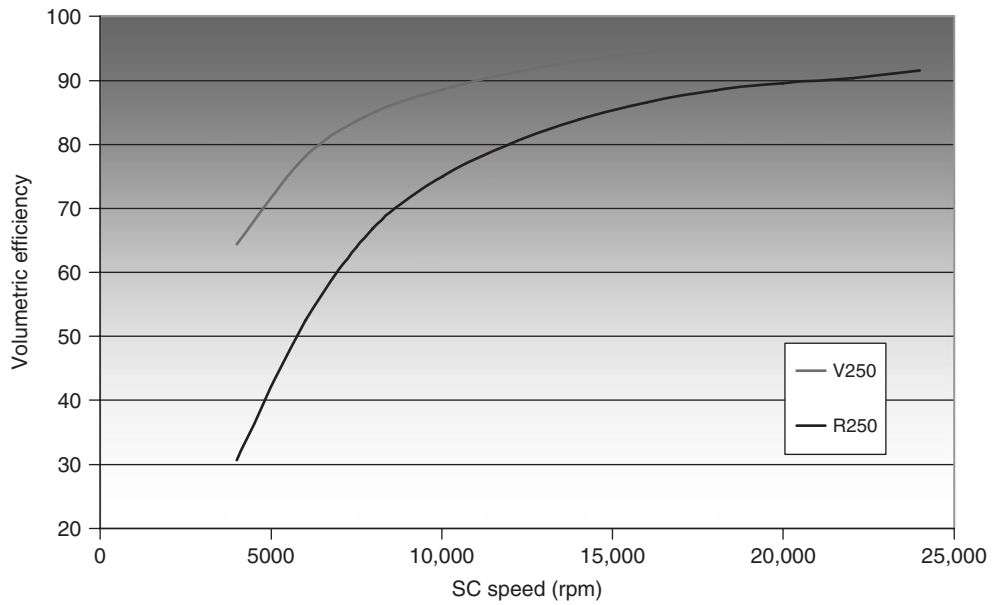


Figure 24. Volumetric efficiency comparison of devices. (Reproduced by permission of Eaton Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

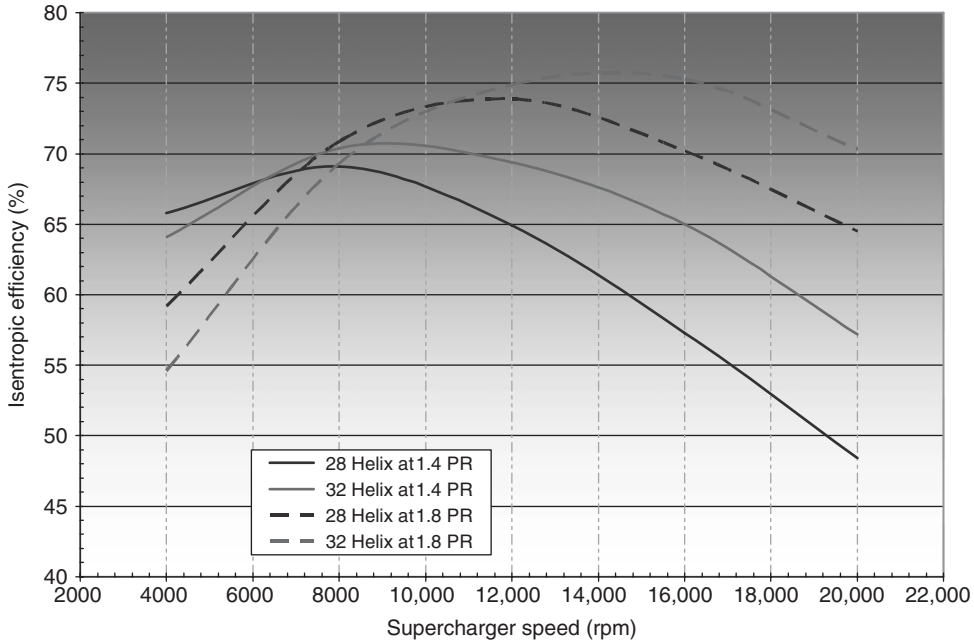


Figure 25. Isentropic efficiency versus helix angle. (Reproduced by permission of Eaton Corporation.)

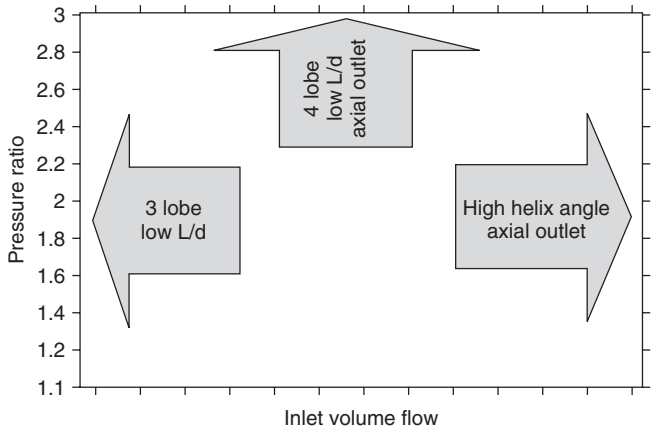


Figure 26. Device function optimization. (Reproduced by permission of Eaton Corporation.)

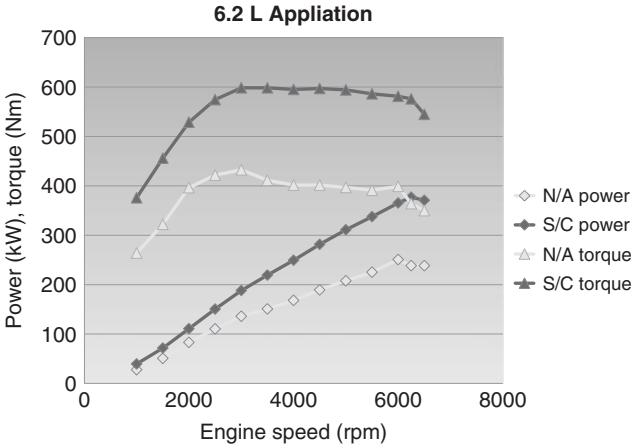


Figure 27. Power density optimization. (Reproduced by permission of Eaton Corporation.)

An additional benefit of mechanical supercharging is the transient impact of performance. Given the direct mechanical connection between the supercharger and engine, boost response is significantly faster than devices that are passively coupled. The following example shows manifold pressure response from 900 rpm with a wide open throttle (WOT) step input at 0 s in first gear (Figure 28).

The effect on vehicle response from a stop light or during the initiation of a passing maneuver is very noticeable. In addition, the transient torque rise rate allows for applications to be geared to run slower during typical cruise conditions to reduce frictional losses in the entire powertrain, yet provide acceptable performance for quick acceleration. This technique is called *Down-Speeding* in the industry.

The transient response of a supercharger can be exploited when coupled in a compound boosting system, such as a super-turbo configuration. This combination of devices gives quick transient response while continuing to maintain the highest pressure ratio capability and maintaining altitude compensation inherent with turbocharging. Figure 29 shows

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

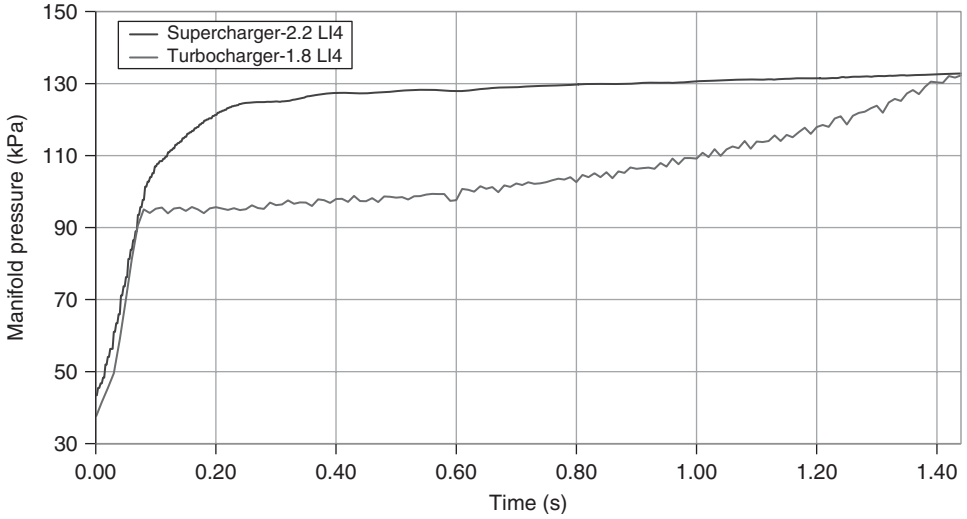


Figure 28. Transient boost response. (Reproduced by permission of Eaton Corporation.)

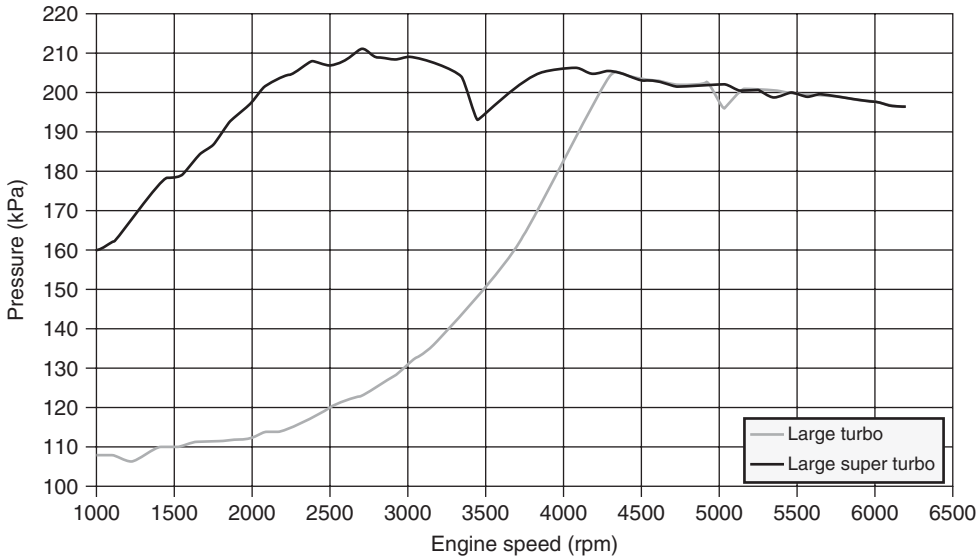


Figure 29. Compound boosting response. (Reproduced by permission of Eaton Corporation.)

the manifold pressure during a 1060 rpm/s WOT transient acceleration.

Finally, one performance area that may not initially come to mind when thinking about supercharging is emissions performance for diesel applications. As regulations around the world continue to tighten allowable emissions for automotive transportation, there may be a technology opportunity for supercharging that had not previously been required. With diesel applications using technology such as oxidation catalysts, the cold-start emissions start to have the same concerns that they have been for many years on gasoline engines. The following test data compares diesel engine boosting between superchargers and turbochargers. This particular test had both systems tuned to produce the same BMEP levels. The test cycle is the first 75 s on US FTP 75 (Figure 30).

The exhaust temperature entering the catalysts of the test systems is typically 50°C higher is the supercharged system compared to the turbocharged system with the same exhaust temperature in the manifold collector (Figure 31).

The key to cold-start emissions performance is the time it takes the catalysts to reach acceptable conversion

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

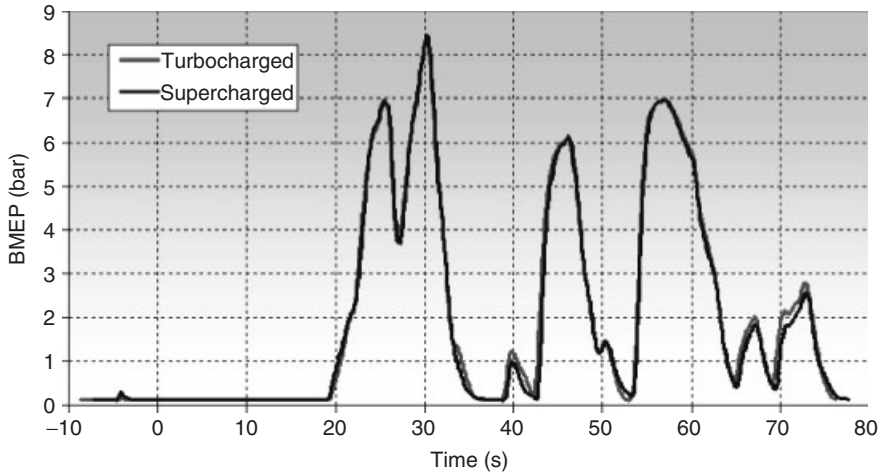


Figure 30. Matched BMEP response. (Reproduced by permission of Eaton Corporation.)

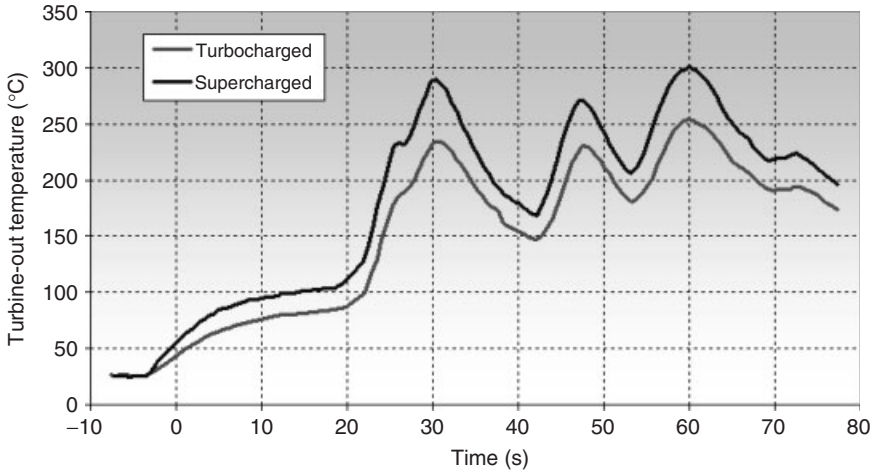


Figure 31. Exhaust temperature levels. (Reproduced by permission of Eaton Corporation.)

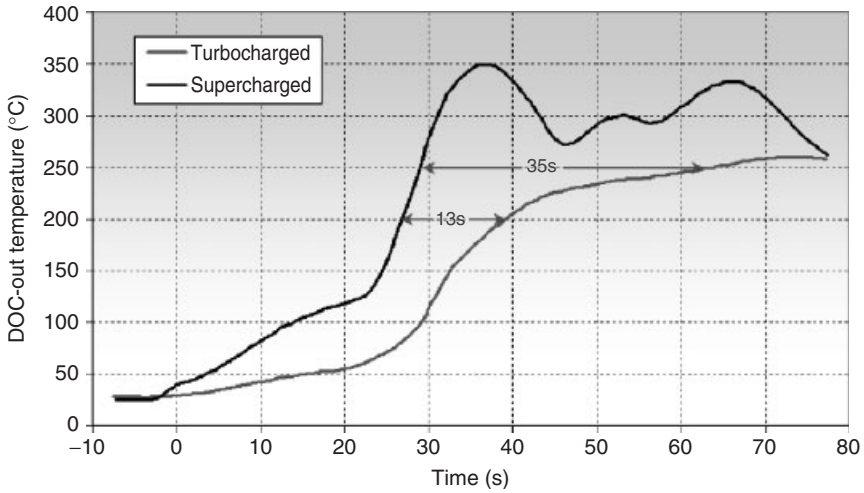


Figure 32. Catalyst light-off response. (Reproduced by permission of Eaton Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

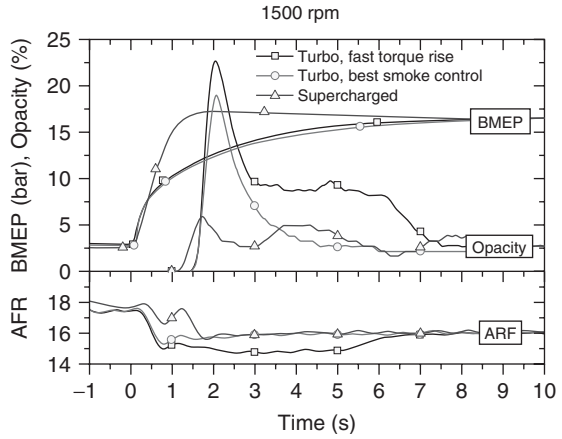


Figure 33. Transient particulate matter response. (Reproduced by permission of Eaton Corporation.)

efficiency, which is typically between 200 and 250°C operating temperature. With appropriate exhaust energy available, the catalyst "light-off" time can be substantially improved with a supercharged system (Figure 32).

Another interesting area of comparison is particulate matter levels during transient load steps leveraging the responsiveness of the air mass availability in a supercharged application. In the testing that follows, the control of the turbocharged system was optimized for either "Best Smoke Control" or "Best Torque Rise" system response. The port is for a 10–90% load step at 1500 rpm engine speed. The plot shows the BMEP, exhaust opacity, and the combustion AFR. The first thing to note is that the load, as expected, responds quickly allowing for leaner AFR levels during the transient. This promotes leaner operation in the supercharged configuration resulting in lower peak exhaust opacity during the event (Figure 33).

As technology and regulations move forward, the future could be a very interesting time for supercharged engines.

REFERENCE

SAE International (1995) *Supercharger Testing Standard*. SAE International, USA.

FURTHER READING

Heisler, H. (1995) *Advanced Engine Technology*. Butterwork-Heinemann, Oxford.

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*. McGraw-Hill, Inc., New York.

Hoag, K.L. (2006) *Vehicular Engine Design*. Springer-Verlag, Wien, Austria.

Rakopoulos, C.D. and Giakoumis, E.G. (2009) *Diesel Engine Transient Operation*. Springer-Verlag London Limited, London.

Richard, v.B. (2009) *Gasoline Engine with Direct Injection*. Vieweg-Teubner, Weisbaden.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto113
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fundamental Chemical Kinetics

Charles K. Westbrook and William J. Pitz
Lawrence Livermore National Laboratory, Livermore, CA, USA

1 Introduction 1
2 Basic Chemical Kinetics 1
3 Basic Classes of Reactions 2
4 Chemical Structure of Fuels 3
5 Reaction Pathways 3
6 Low Temperature Reaction Pathways 5
7 Primary Reference Fuels 7
8 Biofuels 9
9 Effects of C=C Double Bonds on Low Temperature Reactivity 10
10 Ignition of Biodiesel Fuel 11
11 Other Biodiesel Fuels 12
12 Conclusion 13
Acknowledgments 13
References 13

1 INTRODUCTION

Combustion in an internal combustion engine provides the energy to propel a vehicle. Combustion efficiency has economic implications such as miles per gallon of fuel consumed. Combustion in the engine also produces emissions of toxic chemicals and greenhouse gases that affect the environment. Ultimately, all of these factors are results of the specific chemical reactions as well as the rates and chemical pathways of those reactions that control the fuel combustion, all of which we call *fundamental chemical kinetics*.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Within a single engine cycle, chemical kinetics determines the rate and time of ignition, overall combustion rate, flame interaction with combustion chamber walls and piston rings, production of soot and other hydrocarbon emissions, and formation and emissions of other pollutants such as oxides of nitrogen (NO_x) and greenhouse gases, especially carbon dioxide (CO_2). Modeling of chemical kinetics of combustion has become an important tool (Westbrook *et al*., 2005) in engine research, and this chapter provides an overview of some of the fundamental aspects of chemical kinetics, with special attention on the types of fuel molecules that are characteristic of fuels used in modern vehicles. The same kinetic principles are important in combustion processes in spark-ignition (SI), diesel, and homogeneous charge, compression ignition (HCCI) engines as well as in aircraft gas turbine engines.

2 BASIC CHEMICAL KINETICS

Chemical kinetics of hydrocarbon combustion is complex, from initial fuels through stable and radical intermediate species to final products. The goal of combustion is to produce water (H_2O) and carbon dioxide (CO_2), as these products release the greatest possible amount of heat and do the most practical work.

It is convenient to write this combustion process as a single global reaction, using heptane for illustration

$$C_7H_{16} + 11\ O_2 \rightarrow 7\ CO_2 + 8\ H_2O \qquad (1)$$

The overall combustion rate of most hydrocarbons has a strong dependence on temperature, which can be simulated ideally as an Arrhenius expression:

$$\text{Rate} = A\ \exp\left(\frac{-E_a}{RT}\right) \qquad (2)$$

where T is the temperature, R is the universal gas constant, Ea is an activation energy, and A is a constant. The factors A and Ea would have different values for different fuels or reactants.

This type of expression has been used many times (Westbrook and Dryer, 1981) to provide approximate rates of overall reactions. However, the actual combustion of the hydrocarbon fuel does not take place via such a single, global reaction, and such simplified expressions are limited in their predictive capabilities. It is necessary to describe the reaction progress more accurately to gather insight into the details of the combustion process.

Combustion of a hydrocarbon fuel really takes place one atom at a time, removing atoms from the fuel, steadily taking apart the fuel, and building final products. Every step has its own rate, depending on the pressure and temperature of the environment in which the combustion is taking place. Depending on the size and structure of the fuel molecule, hundreds or even thousands of intermediate chemical species can be produced. Species interact via elementary chemical reactions, many of which proceed via complex excited states. For any specific fuel, it is possible to collect all the chemical species and elementary reactions involved in the combustion of that fuel into a chemical kinetic reaction mechanism that can be used as a computational tool to analyze combustion properties of that fuel.

Rates of elementary reactions are taken from many sources. Some particularly important reaction rates have been measured in laboratory experiments, and some reaction rates can be calculated from fundamental theoretical approaches (Wagner, 2002; Zador *et al*., 2011). Once the chemical species, elementary reactions, and reaction rates are assembled into a reaction mechanism, it is then "validated" or tested by comparing the computed results using the mechanism with results from laboratory experiments. Often, mechanism validation uses a variety of experimental problems such as laminar flame speeds, ignition delay measurements, and experiments in plug-flow reactors, stirred reactors, and engine tests to establish a "comprehensive" kinetic mechanism (Westbrook and Dryer, 1980, 1984) that has wide applicability and confidence that its numerical predictions will be as accurate as possible.

Recent kinetic modeling is focusing on large fuel molecules, as most transportation fuels consist mainly of such fuel molecules (Battin-Leclerc, 2008; Curran *et al*., 1998, 2002; Dagaut and Gail, 2007; Dagaut *et al*., 2007; Dagaut and HadjAli, 2009; Glaude *et al*., 2010; Herbinet *et al*., 2008, 2010; Oehlschlaeger *et al*., 2009; Ramirez *et al*., 2011; Ranzi *et al*., 2005, 2009; Westbrook *et al*., 2009, 2011a,b). Practical transportation fuels, especially gasoline, diesel fuel, and jet fuel, contain significant amounts of so many molecules that it is impossible to include kinetic models for all of them. The current solution for this dilemma is to include models for some of the most important components of the hydrocarbon fuel, expecting that simulations using such a selective "surrogate mechanism" (Farrell *et al*., 2007; Pitz *et al*., 2007a, 2007b; Pitz and Mueller, 2011; Colket *et al*., 2007; Violi *et al*., 2002) can produce results that will be approximately the same as the behavior of the real fuel.

3 BASIC CLASSES OF REACTIONS

Combustion is best characterized as a chain reaction, in which chain carriers are active radicals, highly reactive species such as H and O atoms, OH, HO_2, and CH_3. Small radical species have a collective identity as a "radical pool," which is a way of representing the overall reactivity of a hydrocarbon system. Radicals react primarily by removing H atoms from the fuel in "H-atom abstraction" reactions. The products of abstraction reactions react further, producing additional radical species that continue the overall combustion process. Some radicals participate in "chain-branching" reactions, which increase the total number of available radical species, further increasing the rate of H-atom abstraction reactions. Chain branching can produce an explosive growth in the number of radical species and the rate of heat release, leading to ignition. At the same time, other radical recombination reactions can reduce the number of radical species, mediating or reversing the effects of chain branching.

The most important chain-branching reaction (Westbrook and Dryer, 1980, 1984) in combustion is between H atoms and molecular oxygen:

$$H + O_2 = O + OH \tag{3}$$

consuming one radical H and producing two new radicals, O and OH. Other elementary chain-branching reactions include

$$O + H_2 = H + OH \tag{4}$$

and

$$CH_4 + O = CH_3 + OH \tag{5}$$

Another reaction of H atoms with O_2 that is important in converting H atoms to HO_2 radicals is

$$H + O_2 + M = HO_2 + M \tag{6}$$

where M represents the total concentration of all species in the reactive mixture and is needed to conserve energy and momentum in this type of reaction.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

It is impossible to overstate the importance of Reaction (3) in combustion chemistry. In models for flame propagation, high temperature ignition, or explosions, Reaction (3) contributes more to the observed rate of overall reaction than most of the other reactions combined. H atoms are highly diffusive and reactive and abstract H atoms from fuel molecules, and many other reactions produce new H atoms.

The importance of Reaction (3) has many implications. Any fuel that produces H atoms will generate high overall reaction rates, as the H atoms provide so much chain branching. For example, ethane as a fuel is unusually reactive, because every H-atom abstraction by a radical X from ethane leads to the ethyl radical (C_2H_5)

$$C_2H_6 + X = C_2H_5 + HX \tag{7}$$

$$C_2H_5 + M = C_2H_4 + H + M \tag{8}$$

and ethyl immediately produces H atoms by Reaction (8).

In contrast, methane (CH_4) produces low levels of H atoms, as H-atom abstraction from methane produces methyl radicals (CH_3), which react slowly without producing H atoms:

$$CH_4 + X = CH_3 + HX \tag{9}$$

As a result, methane combustion is unusually slow compared to combustion of other hydrocarbon fuels, because its main radical does not produce H atoms. Methane and ethane are special cases, as each leads to a single, unique product radical; larger hydrocarbon fuels lead to a mixture of H and CH_3 radicals, reacting at overall rates that are intermediate between those of methane and ethane (Westbrook, 1986).

It is easy to identify chain branching that occurs in a single reaction such as Reactions (3), (4), or (5), but sometimes chain branching occurs over a sequence of reactions. For example, during the intermediate temperature reaction of many hydrocarbons, a sequence of reactions occur for a fuel RH

$$RH + HO_2 = R + H_2O_2 \tag{10}$$

$$R + O_2 + M = Q + HO_2 \tag{11}$$

where Q is a stable intermediate species. These steps maintain the number of radicals and build up the concentration of hydrogen peroxide (H_2O_2). Once H_2O_2 decomposes in Reaction (12), the group of Reactions (10–12) collectively provides chain branching, with a net increase in the radical pool of two OH radicals.

$$H_2O_2 + M = OH + OH + M \tag{12}$$

However, the O–O bond is not easy to break, so it breaks into two OH radicals only when the temperature approaches 1000 K, when it decomposes rapidly. When Reactions (10) and (11) take place at temperatures *above* 1000 K, Reaction (12) proceeds rapidly and the chain branching is immediate. If these reactions take place at temperatures *below* 900 K, then the chain branching is delayed until the temperature reaches 1000 K, when multiple OH radicals are finally liberated, a result called *degenerate chain branching* (Griffiths and Barnard, 1995). Degenerate branching is a major component in so-called low temperature chemistry (LTC) and the negative temperature coefficient (NTC) of ignition. Diesel, SI, and HCCI engines all start to ignite at temperatures below 900 K, so degenerate chain branching is important in all three engine environments.

4 CHEMICAL STRUCTURE OF FUELS

Historically, most transportation fuels have been produced by refining petroleum liquids. In recent years, alternative fuels from other geological sources and biosources have begun to appear in automotive fuels, but their fundamental chemical kinetics are very similar to the kinetics of petroleum-based fuels. Gasoline, diesel fuel, and jet fuel contain measurable levels of hundreds or even thousands of different species. In order to characterize these fuels, it is useful to divide the fuel components into a small number of characteristic fuel classes, summarized in Figure 1.

The relative amounts of components in these classes of fuels are summarized in Figure 2. Gasoline is highest in amounts of branched alkanes and aromatics, fuel structures that resist engine knock; however, diesel fuels are high in *n*-alkanes and cycloalkanes, components that ignite rapidly. Diesel and jet fuels are low in aromatic compounds, as smoke and soot production are particularly troublesome in diesel and jet engines and aromatics are precursors for soot.

5 REACTION PATHWAYS

Overall reaction pathways that convert fuel to intermediates and products depend on the molecular structure of the fuel and the temperature and pressure of the reactive mixture. Two main routes are taken by nearly every hydrocarbon fuel, outlined in Figure 3.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

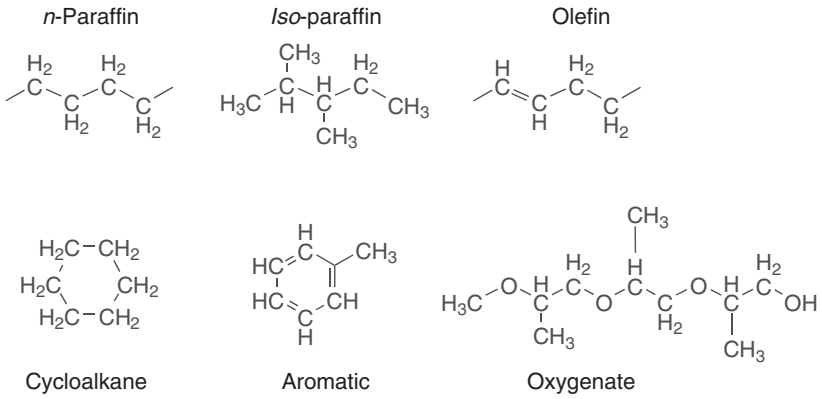


Figure 1. Structural classes of components in petrochemical fuels, with examples for each class.

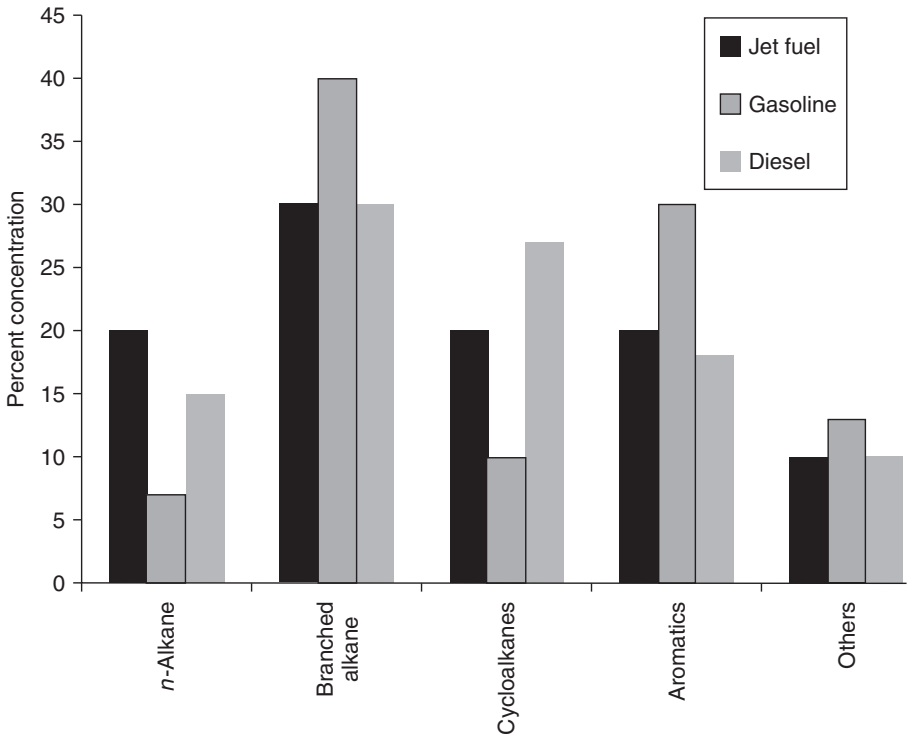
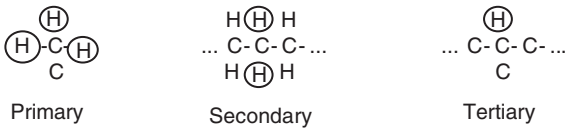


Figure 2. Fractions of fuel classes found in common gasoline, diesel fuel, and jet fuel.

Stable hydrocarbon molecules are consumed primarily by reactions with radical species. There are different types of C–H bonds in hydrocarbon molecules, differences that determine which H atoms are abstracted from each molecule. These distinctions also play a major role in establishing the effects of molecular structure on low temperature alkylperoxy radical isomerization reactions.

The major bond types are primary, secondary, and tertiary C–H bonds, and their bond strengths and rates of H-atom abstraction reactions depend on the number of other C atoms bonded to the C atom involved in the C–H bond. These bonds are shown below:



Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

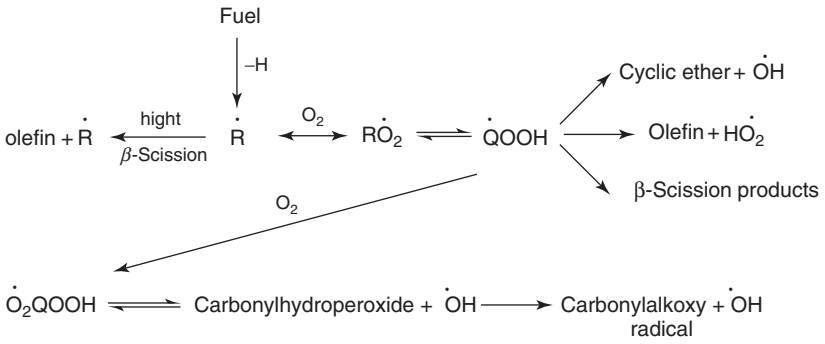
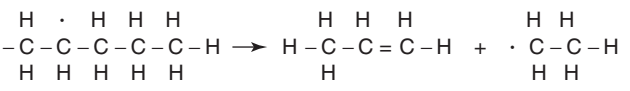


Figure 3. Reaction pathways, showing high temperature and low temperature routes.

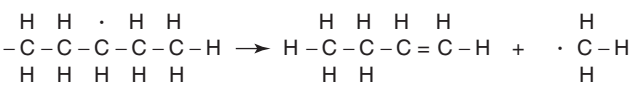
In the primary C–H bond, the C atom is bonded to only one other C atom, while in the secondary C–H bond the C atom is bonded to two other C atoms. Tertiary C–H bonds involve a C atom that has three other C atoms bonded to it. Primary C–H bonds are strongest at 101 kcal/mol, secondary C–H bonds next at 98.5 kcal/mol, and tertiary C–H bonds are weakest at 96.5 kcal/mol, so it is easiest to abstract H atoms from tertiary sites and most difficult to abstract them from primary C–H sites.

Virtually every fuel first reacts by losing an H atom to a radical, leaving a fuel radical R as a product. Subsequent reactions of that product depend on the reacting gas temperature and pressure, and on the structure of the radical species itself. At relatively high temperatures, the radical decomposes, producing smaller species. For example, decomposition of a 2-pentyl radical from *n*-pentane proceeds as



(13)

producing propene (C_3H_6) and an ethyl (C_2H_5) radical. The ethyl radical then produces ethene (C_2H_4) and H atoms (Reaction 8). Abstraction of the H atom in *n*-pentane at the "2" site therefore accelerates the overall rate of reaction since it leads to H atoms. In contrast, the 3-pentyl radical has a different impact because its products are different,

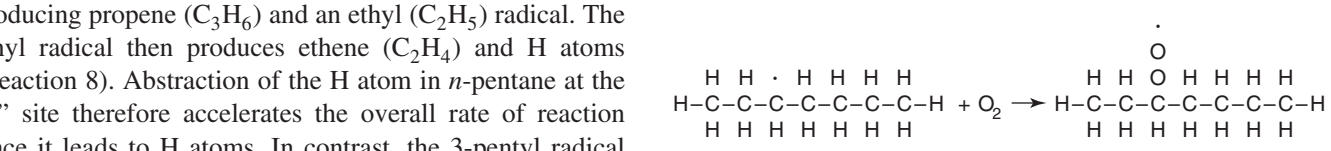


(14)

specifically 1-butene and methyl radicals, and production of methyl radicals decelerates the overall rate of reaction. Decomposition of the 1-pentyl radical is also decelerating, as it produces two stable C_2H_4 species and a methyl radical. For *n*-pentane, abstraction of the four H atoms at the 2-sites leads to H atom production, and abstraction of the eight H atoms at the 1-sites and 3-sites leads to methyl radicals. The fact that the three pentyl radicals decompose to different products demonstrates the necessity of accounting for "site-specific H-atom abstraction" from hydrocarbon fuels. Mixtures of accelerating and decelerating H-atom abstraction pathways are a common feature of most hydrocarbon fuels, and high temperature ignition delay times of most alkanes are very similar (Westbrook *et al*., 2001; Smith *et al*., 2005).

6 LOW TEMPERATURE REACTION PATHWAYS

Under low temperature conditions, the dominant reaction pathway for radicals R, shown in Figure 3, is addition of molecular oxygen to produce an alkylperoxy radical species RO_2, illustrated below for a 3-heptyl radical, producing a 3-heptylperoxy radical.



(15)

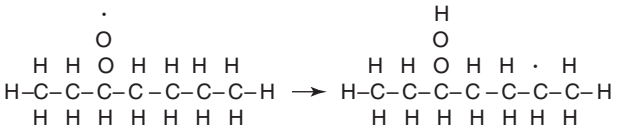
The equilibrium for this $R + O_2 = RO_2$ reaction is a strong function of temperature; the rate of the addition reaction producing RO_2 is effectively independent of temperature, while the decomposition reaction producing $R + O_2$ has large activation energy. The reaction proceeds rapidly toward RO_2 at lower temperatures (i.e., $T \leq 750$ K), but at higher temperatures this reaction becomes kinetically balanced, and, as the temperature increases further, RO_2 becomes very unstable and decomposes rapidly. The temperature at which the RO_2 becomes unstable is about

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

800 K and the equilibrium has shifted almost completely to $R + O_2$ by the time the temperature reaches 850–900 K.

Production of RO_2 species initiates a sequence of reactions that are collectively called *alkylperoxy radical isomerization reaction pathways*, *low temperature oxidation*, or *cool flame chemistry*. At temperatures where the RO_2 species is relatively stable, its primary reaction involves H-atom transfer from within the RO_2 species to the O–O• site in the radical.

This process can be shown, following Reaction (15):



(16)

This product is identified as QOOH in Figure 3, and it is named by specifying the location of the OOH group and the location of the remaining radical site, so the product in Reaction (16) is $C_7H_{14}OOH3\text{-}6$ with the OOH at the 3-site and the radical at the 6-site. Reaction (16) is commonly called an *RO_2 isomerization reaction* or *alkylperoxy radical isomerization*.

The example shown in Reaction (16) describes a process where the O–O• radical "reaches" to take an H atom located three carbon atoms away from the C atom hosting the O–O• group. In chemical terms, this reaction proceeds by forming a seven-membered transition state ring including the O–O group, its host carbon, the three additional C atoms down the molecule chain, and the H atom attached to that target C atom. Three major factors affect the rate of this H-atom transfer: the type of C–H bond being broken (i.e., primary, secondary, or tertiary bond), the number of atoms in the transition-state ring, and the number of equivalent H atoms available at similar sites. It is easy to form transition-state rings with six or seven atoms in the ring but difficult to form transition-state rings with four or five atoms or rings larger than seven atoms.

The QOOH species is a pivotal point, with QOOH having a choice of four or five possible consumption pathways. It can decompose back to the RO_2 reactant, and it can decompose by breaking the O–O bond, producing OH and a cyclic ether species.

$$C_7H_{14}OOH3\text{-}6 \rightarrow cycC_7H_{14}O3\text{-}6 + OH \qquad (17)$$

Reaction (16) can produce a different QOOH, specifically $C_7H_{14}OOH3\text{-}4$, by transferring the H atom from the adjacent C atom, leading to an olefin and HO_2.

$$C_7H_{14}OOH3\text{-}4 \rightarrow 3\,C_7H_{14} + HO_2 \qquad (18a)$$

Another important reaction pathway for fuel radicals R is a direct reaction with O_2:

$$3\,C_7H_{15} + O_2 \rightarrow 2\,C_7H_{14} + HO_2 \qquad (18b)$$

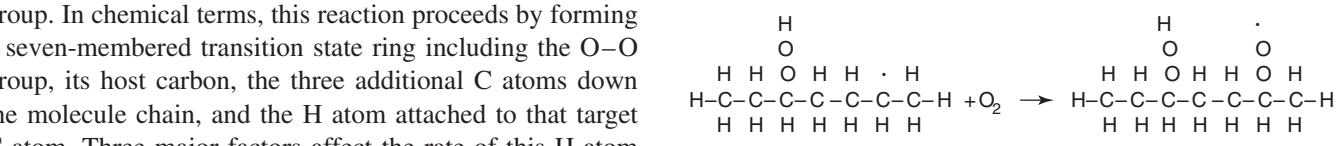
which is equivalent to Reactions (16) and (18a) in reactants and products but with a different rate because it proceeds by a different chemical process (Sheng *et al*., 2002; Zador *et al*., 2011).

Finally, if the QOOH abstracts the H atom two sites away, the result can be

$$C_7H_{14}OOH3\text{-}5 \rightarrow C_2H_5CHO + 1\,C_4H_8 + OH \qquad (19)$$

All these reaction sequences begin with one radical R in Reaction (15) and lead to stable smaller species plus one new radical, namely OH in Reactions (17) and (19) or HO_2 in Reactions (18a) and (18b). Therefore, all these pathways are chain-propagation pathways and cannot contribute to chain branching and ignition at low temperatures.

A final reaction pathway for QOOH does lead to low temperature chain branching and is the most important part of low temperature kinetics. Shown in Figure 3, QOOH can add a second O_2 at its radical site, in general terms a reaction $QOOH + O_2 = O_2QOOH$. In the example being used, this makes $C_7H_{14}OOH3\text{-}6O_2$:



(20)

The O_2QOOH species can then react by decomposing back to $QOOH + O_2$ reactants, or by internally transferring another H atom via a reaction analogous to Reaction (16). That product species then decomposes to smaller fragments and two OH radicals, providing low temperature chain branching. Both Reactions (15) and (20) are very sensitive to temperature, and increasing the temperature to about 850–900 K makes both reactions proceed back, toward the $R + O_2$ and $QOOH + O_2$ sides, terminating the low temperature alkylperoxy radical isomerization reaction pathway and ending the low temperature chain branching.

Fuel ignition at low temperatures can be illustrated in Figure 4 for ignition at 3 bar pressure of stoichiometric *n*-heptane/air at initial temperatures of 700 and 850 K, as well as stoichiometric 3-methyl hexane.

The computed results for *n*-heptane at 700 K show no appreciable heat release or temperature increase for the first 8 ms. During this 8 ms, radical species are produced from the fuel and O_2, but there are not enough radicals to consume a significant amount of fuel. Heptyl radicals

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

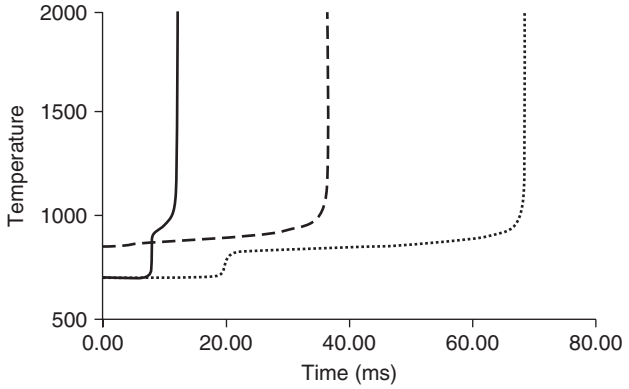


Figure 4. Computed ignition delay of stoichiometric *n*-heptane/air at 3 bar pressure, initial temperature 700 K (solid curve), 850 K (dashed curve), and 3-methyl hexane at 3 bar pressure and initial temperature 700 K (dotted line).

produced from the fuel react by addition to O_2 to produce $C_7H_{15}O_2$ radicals, which then isomerize to produce QOOH species. Enough of these QOOH species add a second O_2 to produce chain branching via Reaction (20), leading to low temperature ignition. This chain branching and ignition produces the rapid rise at 8 ms in temperature from 700 to about 900 K in Figure 4. At 900 K, the equilibrium in the heptane versions of Reactions (15) and (20) has shifted dynamically toward $R + O_2$ and $QOOH + O_2$, stopping the low temperature chain branching and the heat release. The temperature then goes through another slow growth stage for 4 ms longer until it reaches about 1000 K. At this point, the accumulated H_2O_2 decomposes rapidly into OH (i.e., Reaction 12), and this sudden flood of new OH radicals provides enough degenerate chain branching to drive the system to its final ignition phase at about 12 ms. The phenomenon shown for the mixture initially at 700 K in Figure 4 is called a *two-stage ignition*, with the first stage ignition at 8 ms and the second stage at 12 ms. The first-stage ignition is also called a *cool flame* or *low temperature* ignition.

The second case for *n*-heptane in Figure 4 has an initial temperature of 850 K. This temperature is at the high end of the low temperature regime, so this mixture experiences virtually no low temperature reactivity because the equilibrium for Reactions (15) and (20) is already favoring $R + O_2$ and $QOOH + O_2$. Low levels of radical species consume fuel slowly, while Reactions (10) and (11) build the levels of H_2O_2 (which is relatively inert because the temperature is still below 1000 K and H_2O_2 cannot decompose). When the temperature finally reaches 1000 K, decomposition of the accumulated H_2O_2 produces large amounts of OH and provides chain branching, producing the ignition at 36 ms seen in Figure 4.

It might seem counterintuitive that the mixture initially at 700 K ignites more rapidly (12 ms) than the mixture initially at 850 K (36 ms), but this is a very tangible demonstration of the impact of the low temperature kinetics regime on overall hydrocarbon ignition. The limited temperature interval over which this unusual chain branching pathway is active means that mixtures that pass through this temperature interval will have a "headstart" on the path to ignition. The mixture initially at 850 K does not experience any low temperature reactions because its initial temperature is above the low temperature reaction region. Not all fuels possess this reaction pathway; if we replace *n*-heptane by another heptane isomer, specifically 2,2,3-trimethyl butane, at an initial temperature of 700 K, which is certainly within the range of low temperature reactivity for many fuels, the latter shows no low temperature reactivity and no first-stage ignition, and the total ignition delay is very long, about 1 s rather than 12 ms. This different behavior is due to the specific molecular structure of 2,2,3-trimethyl butane, particularly the fact that all but one of its 16 H atoms are bound at primary sites in the fuel molecule, strongly inhibiting alkylperoxy radical isomerizations like Reaction (16), so it does not produce a first-stage ignition and its approach to 1000 K and the degenerate branching ignition remains very slow. The more reactive *n*-heptane has an octane number (ON) of 0 while 2,2,3-trimethyl butane has an ON of 112 (Lovell, 1948); *n*-heptane has a great deal of low temperature reactivity, while 2,2,3-trimethyl butane has virtually none, and the principal reason for the difference is that *n*-heptane has many H atoms located at secondary sites while 2,2,3-trimethyl butane has most of its H atoms located at primary sites.

One additional curve is plotted in Figure 4, showing the computed temperatures for another heptane isomer (3-methyl hexane) at an initial temperature of 700 K. This fuel shows a two-stage ignition, the first stage at about 20 ms and the final ignition at about 68 ms. The first stage is later and shows less temperature increase than the first stage for *n*-heptane, so 3-methyl hexane requires a longer time than *n*-heptane to reach the decomposition temperature of H_2O_2. The results for 3-methyl hexane are somewhat intermediate between those for *n*-heptane and 2,2,3-trimethyl butane, and the ON for 3-methyl butane is also intermediate at 52 (Westbrook *et al*., 2002; Silke *et al*., 2005).

7 PRIMARY REFERENCE FUELS

We noted earlier that the rate of RO_2 isomerization in Reaction (16) depends on the type of bond that must be broken to transfer an H atom from a bond with a C atom to a bond with the O–O• group. Three fuel molecules for which

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

nearly all of the H atoms are located at primary sites are *iso*-octane (2,2,4-trimethyl pentane), *iso*-cetane (2,2,4,4,6,8,8-heptamethyl nonane), and 2,2,3-trimethyl butane discussed earlier. *Iso*-octane has 18 H atoms, with 15 located at primary C–H sites, two at secondary sites, and one at a tertiary site. *Iso*-cetane has a total of 34 H atoms, with 1 at a tertiary site, 6 at secondary sites, and 27 at primary sites, and 2,2,3-trimethyl butane has 15 primary H atoms out of a total of 16 H atoms. Nearly all of the RO_2 isomerizations in these fuels require difficult internal abstraction of an H atom from primary sites, and all of them experience slow ignition delays. In Figure 5, the computed ignition delay times are plotted as functions of initial temperature for stoichiometric fuel/air mixtures at 13 bar pressure for *iso*-octane and *iso*-cetane, as well as results for *n*-heptane and *n*-hexadecane at the same conditions.

For the *n*-alkanes, the results in Figure 5 show that, from about 940 to 800 K, the ignition delay time actually becomes shorter (i.e., more reactive) as the temperature decreases, an NTC of reactivity that reflects the amount of low temperature combustion (LTC) that occurs as a result of the molecular structure of the fuel. The existence of an NTC region for a given fuel means the fuel will have two-stage ignitions for a range of initial temperatures. The strong NTC behavior for the *n*-alkanes is shown clearly, and the extents of NTC behavior for the *iso*-alkanes can be seen to be dramatically less than for the *n*-alkanes. Overall, the *n*-alkanes react much more rapidly than the *iso*-alkanes, especially in the NTC or low temperature oxidation region.

In a diesel engine, with no spark plug to ignite the fuel/air mixture, the liquid fuel must vaporize and mix with air quickly enough so that it can find a flammable condition for autoignition. The first mixtures of fuel and air that are able to ignite are quite fuel-rich (Dec, 1997) and are usually within the NTC region. To guarantee efficient operation, the fuel must encourage autoignition, and from Figure 5 it is clear that *n*-alkane fuels ignite rapidly. Diesel engines define the ignitability of their fuel by its cetane number (CN), which has an upper and a lower limit that serve as reference points. The CN of greatest ignitability is for *n*-hexadecane, with a defined CN of 100, and the low CN, slow-to-ignite reference fuel is *iso*-cetane, with a reference CN of 15. The correlation of the amount of low temperature kinetics with CN is very clear.

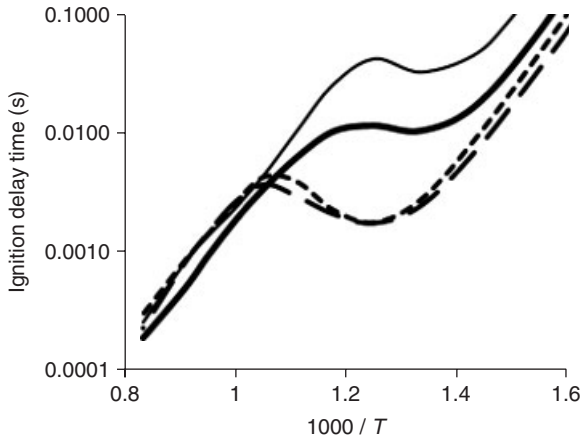


Figure 5. Computed ignition delay times at 13 bar pressure for stoichiometric fuel/air mixtures for *iso*-octane (thin solid line), *iso*-cetane (thick solid line), *n*-heptane (short dashed line), and *n*-hexadecane (long dashed line).

In the case of gasoline reference fuels in SI engines, the process that reflects relative autoignition fuel properties is the onset of knocking behavior. The problem of engine knock has a long history, and the relationship between ignition rate and fuel structure had been understood in phenomenological terms since the early 1920s (Midgley, 1923), making it possible to define reference fuels and a scale of knocking tendency (Lovell, 1948). This is the ON scale, with low values of ON indicating a high rate of autoignition and tendency to knock, while high values indicate low reactivity with low amounts of NTC behavior. *n*-Heptane has the defined low value of ON = 0 and *iso*-octane has the high value of ON = 100. These pairs of reference fuels, *iso*-cetane (CN = 15) and *n*-cetane (CN = 100) for diesel fuel, and *n*-heptane (ON = 0) and *iso*-octane (ON = 100) for gasoline, are the primary reference fuels (PRFs).

For both pairs of PRFs, assessing the ON or CN of an unknown fuel sample is done by finding the mixture of the reference fuels that shows the same autoignition behavior as that of the fuel in question. Experimentally, the comparisons are carried out under carefully specified conditions in a special test engine, a time-consuming process that is also quite expensive. Recent advances in chemical kinetic mechanisms for the PRFs (Westbrook *et al*., 1991, 2011b) have made it possible to do these comparisons between PRF mixtures and test fuels on the computer using detailed kinetic reaction mechanisms, which not only saves time and money but also provides a unique view into the fundamental chemistry that controls autoignition of these complex fuels.

Autoignition rates of the PRFs shown in Figure 5 are virtually identical at temperatures above 950 K, regardless of molecular structure, so the distinctions that affect both CN and ON are the differences in their amounts of low temperature alkylperoxy radical kinetics over the temperature range from 650 to 900 K. Both highly reactive PRFs, *n*-heptane and *n*-hexadecane, are long chains of secondary C–H bonds that are relatively easy to abstract and support extensive amounts of low temperature via alkylperoxy

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

radical kinetics and ignite easily. Both highly reactive PRFs, *n*-heptane and *n*-hexadecane, have long chains of C-H groups with H atoms that are relatively easy to abstract. As a result, both support extensive amounts of low temperature alkylperoxy kinetics and ignite readily.

8 BIOFUELS

Modern technology has been experimenting with biofuels for many years. The first diesel engine was demonstrated by its inventor, Rudolf Diesel, at the Paris Exposition of 1898; it burned peanut oil, making that the first biodiesel fuel. The motivation for using that fuel was that France had access to very large quantities of peanuts and peanut oil from its colonies in North Africa, so the opportunity of a country using whatever plant and plant oil resources it possesses has always been an important part of biofuel choice.

In recent years, liquid transportation biofuels have taken two pathways, one using small fuel molecules for use in SI engines as supplements or replacements for gasoline, and the other using larger fuels derived from plant oils, animal fats, and now including oils from algae to use in diesel engines. Significant amounts of research have examined a variety of energetic small molecules, including hydrogen and a number of small alcohols, for use in SI engines. Alcohols including methanol (CH_3OH), ethanol (C_2H_5OH), and *n*-butanol (*n*-C_4H_9OH) have been studied experimentally and in kinetic modeling as promising liquid fuels. Reliable kinetic reaction mechanisms have been developed for all of these potential small-molecule transportation fuels, and those mechanisms have been used to examine a wide range of combustion and engine properties. For the most part, these fuels are reliable, efficient supplements to gasoline and can be used in neat form without serious impacts on engine operations including overall combustion, and the detailed chemical kinetics of these small biofuels are quite well understood, although some material compatibility issues remain. Low temperature kinetics are unimportant for these small fuel molecules, and they have correspondingly high ONs.

Principal biodiesel fuels today (Graboski and McCormick, 1998) are produced from soybean oils and rapeseed (canola) oils, soy biodiesel fuel in the United States and rapeseed biodiesel fuel in Europe. Most vegetable oils do not burn conveniently in diesel engines, but a simple process called *transesterification* can convert the triglyceride vegetable oils into smaller, alkyl ester structures that burn much more efficiently. When the reagent for transesterification is methanol, the alkyl ester is a methyl ester, which is the case for most biodiesel fuels. This process is illustrated in Figure 6, showing that the initial triglyceride and three methanol molecules produce three methyl ester molecules and one glycerol molecule.

Both soy methyl ester (SME) and rapeseed methyl ester (RME) consist (>99%) of the same five distinct species, whose molecular structures are shown in Figure 7. Four of the species are in a single family, including methyl stearate, $C_{17}H_{35}(CO)OCH_3$, which is saturated with no C=C double bonds; methyl oleate, $C_{17}H_{33}(CO)OCH_3$, with one C=C double bond; methyl linoleate, $C_{17}H_{31}(CO)OCH_3$, with two C=C double bonds; and methyl linolenate, $C_{17}H_{29}(CO)OCH_3$, with three C=C double bonds. The fifth major component of biodiesel fuel is methyl palmitate, $C_{15}H_{31}(CO)OCH_3$, saturated like methyl stearate but with only 16 C atoms in the linear carbon chain. These components are also named as shown in the figure, with methyl stearate denoted as (C18:0), indicating a carbon chain with 18 C atoms in the continuous chain and zero double bonds in that chain. C18:1, C18:2, and C18:3 indicate carbon chains of the same length as methyl stearate, but with one, two, and three C=C double bonds in the chain. Methyl palmitate is denoted as C16:0 for its chain length of 16 C atoms and no C=C double bonds.

One important implication of the relative simplicity of the composition of biodiesel fuel is that, unlike petroleum-based fuels, it is possible to construct kinetic mechanisms that contain *all* of the components of the fuel. The amounts

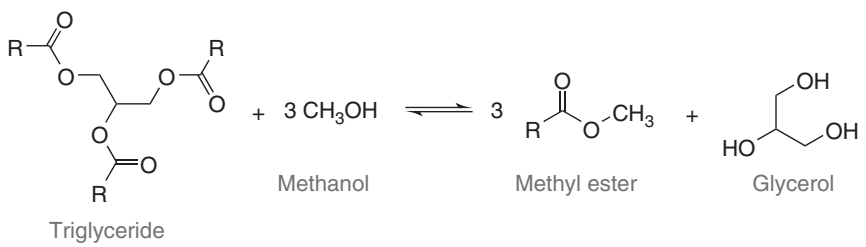


Figure 6. Schematic diagram for transesterification of a vegetable oil with methanol to produce methyl esters.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

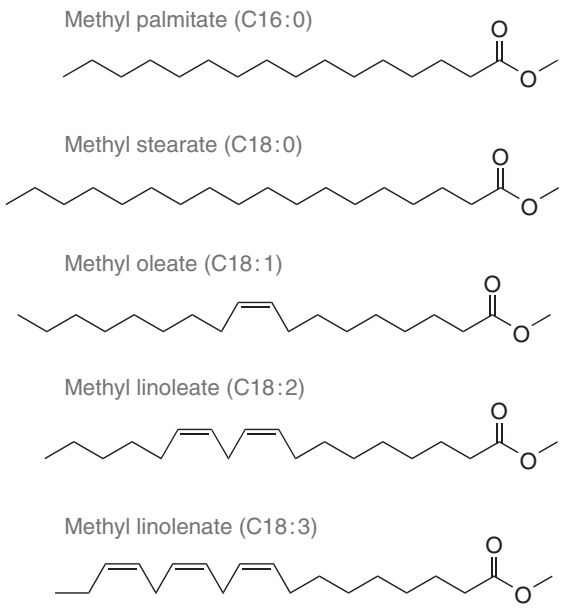


Figure 7. Structural diagram for the five primary components in most biodiesel fuels.

of each of the five components in SME can vary slightly, but the fractions are approximately 0.08 of C16:0, 0.04 of C18:0, 0.25 of C18:1, 0.55 of C18:2, and 0.08 of C18:3. The corresponding components in RME are 0.04 of C16:0, 0.01 of C18:0, 0.60 of C18:1, 0.21 of C18:2, and 0.14 of C18:3. The largest component in SME is C18:2 with two C=C double bonds, and the largest component in RME is C18:1 with one C=C double bond; these differences make a significant impact on the performance of these fuels in diesel engines.

Another interesting feature of biodiesel fuel is the components *not* present in the fuel. Most important, no aromatic species are present in biodiesel fuel, and aromatics are principal chemical precursors to soot production. The absence of aromatics in biodiesel fuels limits soot emissions, relative to emissions from conventional diesel fuel. Another benefit is that, unlike most hydrocarbon species in conventional diesel fuel, each component of soy biodiesel fuel contains two oxygen atoms from the ester group; it has been shown experimentally (Miyamoto *et al*., 1998) and in kinetic simulations (Westbrook *et al*., 2006) that oxygen atoms in the fuel reduce soot production in diesel combustion.

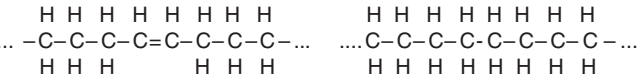
The most common use of biodiesel fuel is to blend it with conventional diesel fuel. Petroleum-based diesel components in these blended fuels will certainly include substantial amounts of aromatics, so blended diesel fuel will still produce soot, but the oxygen content in the biodiesel portion will help reduce the soot being produced by the blend.

9 EFFECTS OF C=C DOUBLE BONDS ON LOW TEMPERATURE REACTIVITY

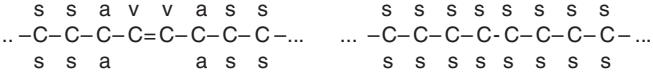
Most past kinetic analyses have focused on saturated hydrocarbon fuels without C=C double bonds and applies equally to conventional hydrocarbons and to the saturated components of biodiesel fuel, methyl stearate and methyl palmitate. However, the kinetic situation changes for the other biodiesel components with one, two, and three C=C double bonds. Computed results for ignition delay time for methyl linoleate and methyl stearate, at exactly the same conditions as those in Figure 5 for the conventional fuel PRF mixtures, show that the two double bonds delay the ignition of methyl linoleate by a factor of at least two, relative to the ignition time of methyl stearate, with almost all of the differences found in the low temperature, cool flame region.

As the only kinetic and molecular structure feature that is different for these two fuels is the inclusion of two C=C double bonds in methyl linoleate, we must look at the kinetic features of double bonds to understand how they affect the ignition rates.

The structure of a chain molecule in the immediate vicinity of a C=C double bond is illustrated on the left, with a saturated chain without a C=C double bond on the right,



or, labeled in a more instructive way



The "s" represents a conventional secondary C–H bond, which is relatively weak (98.5 kcal/mol) and is the main feature of the saturated carbon chains in methyl stearate, methyl palmitate, and *n*-alkane hydrocarbon molecules. When a C=C double bond is inserted into the chain, three changes occur. First, two H atoms are replaced by the double bond. Second, the two H atoms that remain bonded to the C atoms in the double bond become much more strongly bound to them by a "vinylic" C–H bond (indicated by the "v," 109 kcal/mol). Third, the four H atoms bound to atoms adjacent to the double bond become bound much less strongly than in secondary bonds at much weaker C–H bonds called *allylic* C–H bonds ("a" above, 88 kcal/mol).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

H atoms at allylic sites are much easier to abstract than those at secondary sites, and very much easier to abstract than H atoms at primary or vinylic sites, so H-atom abstraction occurs preferentially at these allylic sites. Intuition might then predict that O_2 addition should occur easily at these sites, leading to enhanced low temperature reactivity and faster ignition. However, this does not occur. The allylic radical site remains a weak location for any atom to remain bound for very long, so although O_2 is indeed added quickly to these allylic sites, the allylic RO_2 does not remain long enough to initiate the subsequent H-atom transfer. In chemical terms, the addition rate of Reaction (15) is fast at the allylic sites, but the dissociation rate of the reverse direction of Reaction (15) is extremely fast for allylic sites. Easy abstraction of allylic H atoms actually slows the overall rate of fuel oxidation in the low temperature regime, and the net result is a retarded rather than advanced ignition. Additional C=C double bonds lead to more nonproductive allylic sites in these methyl ester fuel molecules, further retarding ignition, and this is reflected by the steadily decreasing ignition rate for these molecules with more C=C double bonds.

We note that the same impact of adding C=C double bonds into a long carbon chain discussed earlier is a factor in conventional hydrocarbon fuel molecules, so the CNs of *n*-alkane fuels will be reduced when a double bond is inserted into that fuel. Most experimental studies of olefin CNs have examined 1-olefins, but 1-olefins provide only one C atom with two allyl H atom sites, which is only half the effect of the C=C double bonds in these biodiesel fuels. At the same time, on the basis of data from Murphy *et al*. (2004), CNs for 1-olefins are lower than values for *n*-alkanes with the same number of C atoms by significant amounts. For example, the CN values for *n*-octane, 1-octene, and 2-octene are 64, 41, and 43, respectively; for *n*-undecane and 1-undecene they are 79 and 65; for *n*-dodecane and 1-dodecane they are 80 and 71; for *n*-tetradecane and 1-tetradene they are 93 and 79; and for *n*-hexadecane and 1-hexadecene they are 100 and 84. The inhibiting effect of the C=C double bond and the resulting allyl radicals in the unsaturated species clearly has an effect on CN, which is consistent with the effects described here for the reduction in CN from methyl stearate to methyl oleate of 101 to 59, especially when we remember that, in methyl oleate, the C=C double bond is in the middle of the carbon chain and there are four H atoms at allylic sites, while the double bond is at the end of the chain for most available data in conventional hydrocarbon chains. Clearly, more information is needed on other olefins of hydrocarbons.

10 IGNITION OF BIODIESEL FUEL

The CN for soy biodiesel fuel is 47, which is acceptable for use as a "neat" fuel in the United States but too low for use in much of Europe, which requires a somewhat higher CN. Values of CN for individual components are 101 for methyl stearate, 59 for methyl oleate, 38 for methyl linoleate, and 23 for methyl linolenate, so the addition of each C=C double bond to methyl stearate reduces the CN by a considerable amount, with the largest impact seen for the addition of the first C=C double bond. The value of CN for methyl palmitate, with a carbon chain two atoms shorter than that of methyl stearate, is 86, showing the effect of a smaller chain length on CN.

A realistic combustion environment for simulations of diesel-like ignition is the jet-stirred reactor (JSR), commonly used in many research laboratories for basic chemical kinetics studies (Dagaut and Gail, 2007; Dagaut *et al*., 2007; Dagaut and HadjAli, 2009). The JSR provides a realistic simulation of diesel ignition, particularly if operated at elevated pressure (e.g., 50 bar) and fuel-rich conditions ($\phi = 2$ or 3), as diesel engines ignite under fuel-rich, high pressure conditions.

Reactivities of the components in biodiesel fuel are shown in Figure 8, with the fractional fuel consumptions (or conversions) plotted against the mixture temperature. A separate curve is shown for each component in biodiesel fuel, at conditions that are stoichiometric with 0.2% fuel, diluted by helium, at atmospheric pressure and a reactor residence time of 1.5 s. The two saturated components, namely methyl stearate and methyl palmitate, start reacting at 500 K. Both curves rise rapidly as the temperature of the reactor is increased to about 600 K, where these components show nearly equal conversions of about 0.7 or 70% consumption, due primarily to low temperature oxidation reactions. Both curves decrease significantly between 600 and 750 K to a minimum of about 25% conversion and then

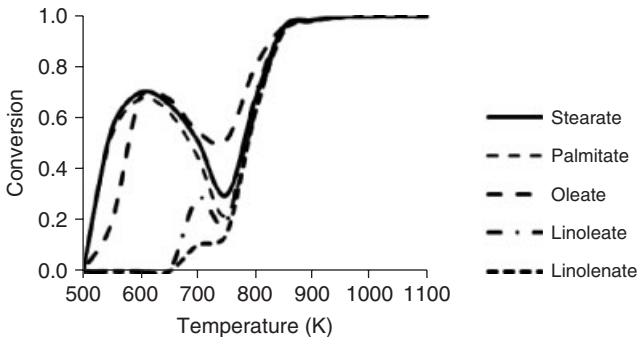


Figure 8. Simulated JSR conversion of each soy biodiesel component as individual fuel.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

increase to 100% conversion for temperatures above 875 K. Both saturated components show considerable amounts of NTC behavior from 600 to 750 K, and, above 750 K, their reactivity increases rapidly.

Methyl oleate, with one C=C double bond, is less reactive under JSR conditions than either of the saturated components. It shows significant NTC behavior, but not as much as the saturated components, and the oleate begins to react at temperatures slightly higher than the stearate or palmitate. Methyl linoleate is much different from the first three components; it does not begin to react until about 650 K and shows minor NTC behavior. Methyl linolenate is even less reactive at lower temperatures, shows no NTC behavior except for a slight inflection point in its conversion curve at 700 K, and is completely burned to products, like the other components, above 875 K. This group of fuels reacts in the same order at their CN values of 101, 86, 59, 38, and 23.

When the five components are combined into a single fuel with relative fractions characteristic of soy biodiesel, using the same conditions as those for Figure 8, the full SME biodiesel simulation results can be seen in Figure 9. Conversions of all five components are shown, and it is clear that, when combined into a fuel, they all burn at very much the same rates over the entire temperature range. The conversion curve for the mixture, with its CN of 47, falls neatly into line with the five separate component conversion curves in Figure 8, between methyl oleate (CN = 59) and methyl linoleate (CN = 38).

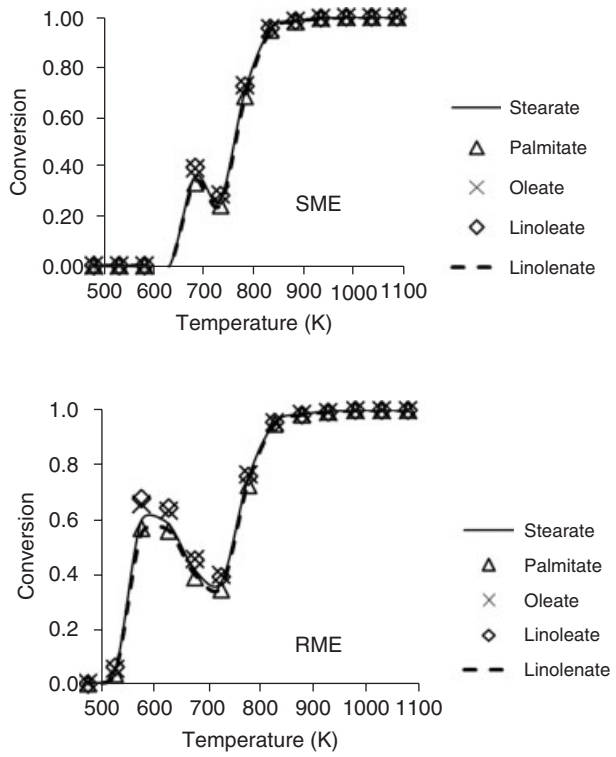


Figure 9. JSR simulations for soy methyl ester and rapeseed methyl ester fuels.

For comparison, a JSR simulation was carried out with the initial composition of RME fuel, with CN = 54. Its primary component is methyl oleate, with only one C=C double bond, and comparisons of the RME behavior in Figure 9 with the component curves in Figure 8 and the SME mixture in Figure 9 show that not only is the RME more reactive than SME but it also comes close to the computed behavior of the single-component methyl oleate, with CN = 59. These observations demonstrate that the JSR simulations are able to rank the single-component fuels and the composite SME and RME fuels in the same order as their known CN values and reproduce the effects of mixing their relative concentrations in a consistent way.

11 OTHER BIODIESEL FUELS

Many other biodiesel fuels have the same major components as RME and SME. The same chemical kinetic model used to simulate soy and rapeseed biodiesel kinetics can be used to study biodiesel fuels made by transesterification from sunflower oil, safflower oil, linseed oil, Jatropha oil, cottonseed oil, corn oil, olive oil, palm oil, and peanut oil, as well as oils from beef tallow. The CNs of these fuels range from 39 for linseed oil methyl ester to 62 for palm oil methyl ester. Chemical kinetic analysis easily explains these extreme values and reactivities in terms of their relative compositions of the five major components. That is, palm oil has 46% of the saturated component methyl palmitate and 40% of methyl oleate, which has one C=C bond, leading to its high CN value. In contrast, linseed oil methyl ester has virtually no saturated, high CN components, and includes 54% methyl linolenate, the component with CN = 23, leading to a CN value for the linseed oil methyl ester fuel of only 39. In fact, linseed oil methyl ester biodiesel fuel is the only common plant-derived biodiesel fuel that cannot be used in most diesel engines without CN-improving additives.

A final example shows that the hydrocarbon reference fuel scale and the biodiesel fuel CN ratings are internally consistent and calibrated to the same CN scale. This is illustrated in Figure 10, showing two diesel PRF mixtures of CN20 and CN60 and three different biodiesel fuels, namely beef tallow methyl ester, peanut oil methyl ester, and linseed oil methyl ester. This illustration uses a JSR environment at 50 bar pressure, with 200 ppm (parts per million) fuel at an equivalence ratio of 2, diluted in Helium,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

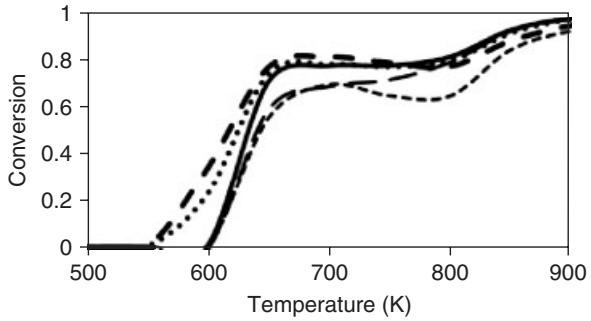


Figure 10. JSR simulations at 50 bar pressure, $\phi = 2200$ ppm fuel in O_2, diluted in He, 0.05 s residence time. Fuels are CN60 (dashed curve), beef tallow methyl ester (CN58, dotted curve), peanut oil methyl ester (CN54, solid curve), linseed oil (CN39, long dashes curve), and CN20 (short dashes curve).

and a residence time of 0.05 s. The CN60 and the beef tallow methyl ester curves are very close together over the entire temperature range and have CN values of 60 and 58, respectively. The least reactivity in Figure 10 is for CN20, and the biofuel closest to CN20 is the curve for linseed oil methyl ester, which has CN = 39. Peanut oil methyl ester, with CN = 54, is more reactive than linseed oil but less reactive than the beef tallow fuel with CN = 58, and all the differences between beef tallow methyl ester and peanut oil methyl ester are at the lowest temperatures of the calculations.

The differences between the biodiesel fuel reactivities, which are due to differences in the amounts of C=C double bonds in the carbon chains, can be related directly to the differences in reactivity of the PRF mixtures, which in turn are due to differences in the relative number of primary, secondary, and tertiary C–H bonds in the saturated hexadecane isomers of the PRF diesel fuels.

12 CONCLUSION

This chapter has described the major features of the chemical kinetics of combustion of transportation fuels and how the size and structures of those fuels control the rate of their oxidation reactions and heat release. Particular attention has been given to the kinetics of large hydrocarbon and biodiesel molecules that are common to SI and diesel engines, where the sizes of the fuel molecules present some unique challenges for kinetic analysis. The key to understanding these processes is to view them as chain reactions, so identification of the main chain-branching reactions leads naturally to a coherent picture of steady combustion in burners and flames, and quenching and autoignition in SI, HCCI, and diesel engines. Examples have been used to illustrate the overall principles involved in chemical kinetics, and special attention has been given to low temperature kinetics, which play important roles in many practical combustion problems.

ACKNOWLEDGMENTS

The authors acknowledge the support of the US Department of Energy, Office of Vehicle Technologies, Program Managers Gurpreet Singh and Kevin Stork. This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344.

REFERENCES

Battin-Leclerc, F. (2008) Detailed chemical kinetic models for the low temperature combustion of hydrocarbons with application to gasoline and diesel fuel. *Progress in Energy and Combustion Science*, **34**, 440–498.

Colket, M., Edwards, J.T., Williams, S., *et al*. (2007) *Development of an experimental database and kinetic models for surrogate Jet fuels*. 45th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, paper AIAA-2007-0770.

Curran, H.J., Gaffuri, P., Pitz, W.J., and Westbrook, C.K. (1998) A comprehensive modeling study of n-heptane oxidation. *Combustion and Flame*, **114**, 149–177.

Curran, H.J., Gaffuri, P., Pitz, W.J., and Westbrook, C.K. (2002) A comprehensive modeling study of iso-octane oxidation. *Combustion and Flame*, **129**, 253–280.

Dagaut, P. and Gail, S. (2007) Kinetic study of the effect of a biofuel additive on Jet-A1 combustion. *The Journal of Physical Chemistry A*, **111**, 3992–4000.

Dagaut, P., Gail, S., and Sahasrabudhe, M. (2007) Rapeseed oil methyl ester oxidation over extended ranges of pressure, temperature, and equivalence ration: experimental and modeling kinetic study. *Proceedings of the Combustion Institute*, **31**, 2955–2961.

Dagaut, P. and Hadj-Ali, K. (2009) Chemical kinetic study of the oxidation of isocetane (2,2,4,4,6,8,8-heptamethyl nonane) in a jet-stirred reactor: experimental and modeling. *Energy & Fuels*, **23**, 2389–2395.

Dec, J.E. (1997) A conceptual model of DI diesel combustion based on laser-sheet imaging. Society of Automotive Engineers Paper SAE-970873.

Farrell, J.T., Cernansky, N.P., Dryer, F.L., *et al*. (2007) Development of an experimental database and kinetic models for surrogate diesel fuels. Society of Automotive Engineers publication, SAE-2007-01-0201.

Glaude, P.W., Herbinet, O., Bax, S., *et al*. (2010) Modeling of the oxidation of methyl esters—validation for methyl hexanoate, methyl heptanoate, and methyl decanoate in a jet-stirred reactor. *Combustion and Flame*, **157**, 2035–2050.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Graboski, M.S., and McCormick, R.L. (1998) Combustion of fat and vegetable oil derived fuels in diesel engines. *Progress in Energy and Combustion Science*, **24**, 125–164.

Griffiths, J.F., and Barnard, J.A. (1995) *Flame and Combustion*, CRC Press, London.

Herbinet, O., Pitz, W.J., and Westbrook, C.K. (2008) Detailed chemical kinetic oxidation mechanism for a biodiesel surrogate. *Combustion and Flame*, **154**, 507–528.

Herbinet, O., Pitz, W.J., and Westbrook, C.K. (2010) Detailed chemical kinetic mechanism for the oxidation of biodiesel fuels blend surrogate. *Combustion and Flame*, **157**, 893–908.

Lovell, W.G. (1948) Knocking characteristics of hydrocarbons. *Industrial & Engineering Chemistry*, **40**, 2388–2438.

Midgley, T. (1923) Some fundamental relations among the elements and compounds as regards the suppression of gaseous detonation. *Industrial & Engineering Chemistry*, **15**, 421–423.

Miyamoto, N., Ogawa, H., Nurun, N.M., *et al*. (1998) Smokeless, low NOx, high thermal efficiency, and low noise diesel combustion with oxygenated agents as main fuel. Society of Automotive Engineers Paper SAE-980506.

Murphy, M.J., Taylor, J.D., and McCormick, R.L. (2004). Compendium of experimental cetane number data. NREL report NREL/SR-540-36805.

Oehlschlaeger, M.A., Steinberg, J., Westbrook, C.K., and Pitz, W.J. (2009) The autoignition of iso-cetane at high to moderate temperatures and elevated pressures: shock tube experiments and kinetic modeling. *Combustion and Flame*, **156**, 2165–2172.

Pitz, W.J., Naik, C.V., Mhaolduin, T.N., *et al*. (2007a) Modeling and experimental investigation of methylcyclohexane ignition in a rapid compression machine. *Proceedings of the Combustion Institute*, **31**, 267–275.

Pitz, W.J., Cernansky, N.P., Dryer, F.L., *et al*. (2007b) Development of an experimental database and kinetic models for surrogate gasoline fuels. Society of Automotive Engineers publication 2007-01-0175. SAE 200 Transactions of Journal of Passenger Cars: Mechanical Systems.

Pitz, W.J. and Mueller, C.J. (2011) Recent progress in the development of diesel surrogate fuels. *Progress in Energy and Combustion Science*, **37**, 330–350.

Ramirez, H.P., Hadj-Ali, K., Dievart, P., *et al*. (2011) Oxidation of commercial and surrogate bio-diesel fuels (B30) in a jet-stirred reactor at elevated pressure: experimental and modeling kinetic study. *Proceedings of the Combustion Institute*, **33**, 375–382.

Ranzi, E., Frassoldati, A., Granata, S., and Faravelli, T. (2005) Wide-range kinetic modeling of the pyrolysis, partial oxidation and combustion of heavy n-alkanes. *Industrial and Engineering Chemistry Research*, **44**, 5170–5183.

Ranzi, E., Frassoldati, T., Faravelli, T., and Cuoci, A. (2009) Lumped kinetic modeling of the oxidation of isocetane (2,2,4,4,6,8,8-heptamethyl nonane) in a jet-stirred reactor (JSR). *Energy & Fuels*, **23**, 5287–5289.

Sheng, C.Y., Bozzelli, J.W., Dean, A.M., and Chang, A.Y. (2002) Detailed kinetics and thermochemistry of $C_2H_5 + O_2$: reaction kinetics of the chemically-activated and stabilized CH_3CH_2OO adduct. *Journal of Physical Chemistry A*, **106**, 7276–7293.

Silke, E.J., Curran, H.J., and Simmie, J.M. (2005) The influence of fuel structure on combustion as demonstrated by the isomers of heptane: a rapid compression machine study. *Proceedings of the Combustion Institute*, **30**, 2639–2647.

Smith, J.M., Simmie, J.M., and Curran, H.J. (2005) Autoignition of heptanes; experiments and modeling. *International Journal of Chemical Kinetics*, **37**, 728–736.

Violi, A., Yan, S., Eddings, E.G., *et al*. (2002) Experimental formulation and kinetic model for JP-8 surrogate mixtures. *Combustion Science and Technology*, **174**, 399–417.

Wagner, A.F. (2002) The challenges of combustion for chemical theory. *Proceedings of the Combustion Institute*, **29**, 1173–1200.

Westbrook, C.K. (1986) Chemical kinetic modeling of higher hydrocarbon fuels. *AIAA Journal*, **24**, 2002–2009.

Westbrook, C.K. and Dryer, F.L. (1980) Chemical kinetics and modeling of combustion processes. *Proceedings of the Combustion Institute*, **18**, 749–767.

Westbrook, C.K. and Dryer, F.L. (1981) Simplified reaction mechanisms for the oxidation of hydrocarbon fuels in flames. *Combustion Science and Technology*, **27**, 31–43.

Westbrook, C.K. and Dryer, F.L. (1984) Chemical kinetics modeling of hydrocarbon combustion. *Progress in Energy and Combustion Science*, **10**, 1–57.

Westbrook, C.K., Pitz, W.J., and Leppard, W.R. (1991) The autoignition chemistry of paraffinic fuels and pro-knock and anti-knock additives: a detailed chemical kinetic study. SAE publication SAE-912314. *SAE Transactions*, Section 4, **100**, 605–622.

Westbrook, C.K., Pitz, W.J., Curran, H.C., *et al*. (2001) A detailed chemical kinetic modeling study of the shock tube ignition of isomers of heptane. *International Journal of Chemical Kinetics*, **33**, 868–877.

Westbrook, C.K., Pitz, W.J., Boercker, J.E., *et al*. (2002) Detailed chemical kinetic reaction mechanisms for autoignition of isomers of heptane under rapid compression. *Proceedings of the Combustion Institute*, **29**, 1311–1318.

Westbrook, C.K., Mizobuchi, Y., Poinsot, T., *et al*. (2005) Computational combustion. *Proceedings of the Combustion Institute*, **30**, 125–157.

Westbrook, C.K., Pitz, W.J., and Curran, H.J. (2006) Chemical kinetic modeling study of the effects of oxygenated hydrocarbons on soot emissions from diesel engines. *Journal of Physical Chemistry A*, **110**, 6912–6922.

Westbrook, C.K., Pitz, W.J., Herbinet, O., *et al*. (2009) A comprehensive detailed chemical kinetic reaction mechanism for combustion of n-alkane hydrocarbons from n-octane to n-hexadecane. *Combustion and Flame*, **156**, 181–199.

Westbrook, C.K., Naik, C.V., Herbinet, O., *et al*. (2011a) Detailed chemical kinetic reaction mechanisms for soy and rapeseed biodiesel fuels. *Combustion and Flame*, **158**, 742–755.

Westbrook, C.K., Pitz, W.J., Mehl, M., and Curran, H.J. (2011b) Detailed chemical kinetic reaction mechanisms for primary reference fuels for diesel cetane number and spark-ignition octane number. *Proceedings of the Combustion Institute*, **33**, 185–192.

Zador, J., Taatjes, C.A., and Fernandes, R.X. (2011) Kinetics of elementary reactions in low temperature autoignition chemistry. *Progress in Energy and Combustion Science*, **37**, 371–421.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto114
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fundamental Combustion Modes

Ronald D. Matthews
The University of Texas, Austin, TX, USA

1 Categories of Engines and Corresponding Modes of Combustion 1
2 Turbulent Premixed Flames 2
3 Mixing Controlled Combustion 11
4 Volumetric Energy Release 15
5 Mixed Combustion Modes 16
References 17

1 CATEGORIES OF ENGINES AND CORRESPONDING MODES OF COMBUSTION

In general, three types of engines are used in passenger cars and heavy-duty vehicles. The spark ignition (SI) engine is the most widely used engine in the world in passenger cars, primarily because it is the lowest cost option for simultaneously meeting performance expectations and standards imposed regarding fuel economy and emissions. The SI engine is by far the most dominant engine in the US passenger car and light-duty truck market. The diesel engine is used in some passenger cars and light-duty trucks, extensively so in Europe, and in almost all heavy-duty trucks. The gas turbine, primarily mentioned for completeness, was last used in a passenger car built by Chrysler in 1963, for the only consumer test of gas turbine-powered cars. However, the gas turbine is used in all of the tanks in the US military, and they are the dominant engine in air transportation applications. Although the gas turbine is not of much practical importance from the perspective of on-road automotive engineering, the idealization of the combustion process in a gas turbine will appear in the "combustion phase diagrams" to be discussed later in this chapter.

These three types of engines are characterized by three different modes of combustion. The homogeneous charge SI engine has a premixed flame. The combustion process in a diesel engine is dominated by a diffusion, or non-premixed, flame. The combustion process in a gas turbine can be idealized as a "perfectly stirred reactor." In a premixed flame, the fuel and air are uniformly mixed on the molecular scale but are separated from the products of combustion by a thin flame sheet. In a diffusion flame, the fuel and air are separated, must diffuse in opposite directions, and form a flame zone in the vicinity of the stoichiometric zone between the fuel and the air, with the rate of heat release controlled by the diffusion or mixing process. In a perfectly stirred reactor, the fuel, air, and products of combustion are homogeneously and instantaneously mixed on the molecular scale as reaction proceeds.

These three modes of combustion are not the complete set of the potential modes under which fuel and air can react under engine-type conditions. An important fourth mode is volumetric energy release. During volumetric energy release, the fuel and air are mixed on a molecular scale but combustion appears to be initiated throughout the volume by a rapidly developing kinetically controlled ignition process. Knock in SI engines is characterized by this mode of combustion. Volumetric combustion also occurs in a less premixed form during the "premixed burn" phase of diesel combustion and is a dominant mode of combustion in many forms of low temperature compression ignition combustion being explored worldwide for the fuel efficiency potential it offers

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

[e.g., homogeneous charge compression ignition (HCCI) combustion].

2 TURBULENT PREMIXED FLAMES

For the conditions within internal combustion engines, all of these fundamental modes of combustion are complicated by the presence of turbulence. The effects of turbulence on premixed flame propagation are discussed in this section. As already noted, turbulent premixed flames are the dominate mode of combustion within homogeneous charge SI engines. This section discusses a unified way to view and categorize turbulent premixed flames relevant to SI engines—a view that also shows how perfectly/well-stirred flames relevant to gas turbines fit into the picture. In addition, the effects of the critical parameters controlling the turbulent premixed flame front propagation speed are discussed. The overall discussions provide understanding of the nature and speed of turbulent premixed flames under various conditions, and which conditions are most relevant to SI engines.

As background to this topic, it is first useful to discuss the laminar premixed combustion process. Envision a premixed flame propagating in a well-insulated tube under laminar plug flow conditions in the absence of buoyancy. This laminar premixed flame can be held at a stationary position within the tube by flowing the reactants into the tube at the correct velocity: the "unstretched laminar flame speed," S_L, which depends upon the reactant composition, temperature, and pressure (as discussed in Section 2.3.1). If this premixed laminar flame is traversed with composition and temperature probes, where it must be noted that this laminar premixed flame is extremely thin, the resulting measurements will yield results similar to those presented in Figure 1. As the flame is approached from the reactant side, one notes that the fuel begins to disappear and intermediate hydrocarbons and molecular hydrogen are observed. Further into the flame, the intermediate hydrocarbons are oxidized to form CO and the molecular hydrogen is oxidized, eventually, to form water. Further through the flame, the CO is oxidized to form CO_2 in a relatively slow but highly exothermic reaction.

If the approach flow is turbulent rather than laminar, the central question is: how does the turbulence affect the laminar flame structure and chemistry? Understanding the interactions between turbulence and laminar combustion has been a central theme of combustion research for more than 30 years (e.g., Lancaster *et al*., 1976; Mattavi, Groff, and Matekunas, 1979). This need eventually led to methods for quantitatively categorizing various regimes within which turbulence would have specific effects on the combustion process. The Damkohler number versus turbulent Reynolds number diagram developed by Abraham, Williams, and Bracco (1985) is the subject of Section 2.1. Borghi's (1985) combustion phase diagram, and modifications thereof, is the subject of Section 2.2. Section 2.3 covers the elements of a physically based model of turbulent premixed flame propagation to provide the reader a physical understanding of flame propagation in SI engines, including stratified charge SI engines. In addition, the overall SI engine combustion system is discussed in Spark Ignition Combustion.

2.1 The Damkohler number versus turbulent Reynolds number diagram

Abraham, Williams, and Bracco (1985) made a breakthrough in understanding the interaction between turbulence and laminar flame-type combustion in homogeneous charge SI engines by constructing a diagram of a Damkohler number, Da_Λ, versus the turbulent Reynolds number, Re_Λ, as illustrated in Figure 2.

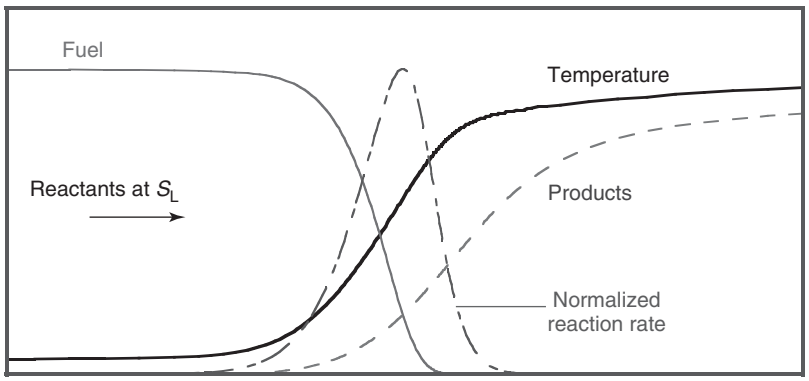


Figure 1. Temperature, composition, and reaction rate distributions through a premixed laminar flame.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

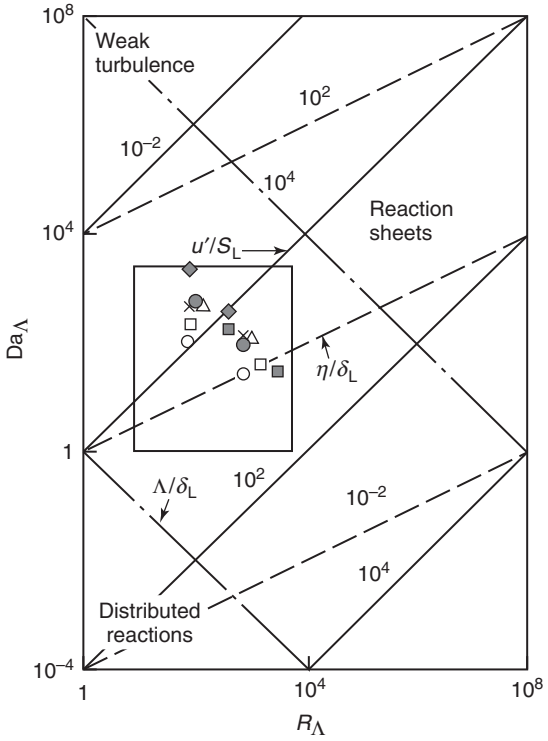


Figure 2. The $Da_\Lambda - Re_t$ diagram. (From Abraham, Williams, and Bracco, 1985. Copyright © SAE International. Reprinted with permission.)

Of the various Damkohler number definitions that exist, the Damkohler number chosen by Abraham, Williams, and Bracco (1985) for construction of Figure 2 is defined as the ratio of the characteristic turnover time of a turbulent eddy of the size of the integral length scale, $\tau_t = L_i/u'$, to the characteristic residence time in a laminar flame, $\tau_L = \delta_L/S_L$,

$$Da_{L_i} = Da_\Lambda = \frac{\tau_t}{\tau_L} = \frac{L_i}{u'}\frac{S_L}{\delta_L} \tag{1}$$

where u' is the root mean square (RMS) of the turbulent velocity fluctuations (the turbulence intensity), L_i or Λ is the integral length scale size of the spectrum of turbulent eddy sizes, S_L is the unstretched laminar flame speed, and δ_L is the laminar flame thickness. Abraham, Williams, and Bracco note that this Damkohler number can be perceived as an inverse measure of the influence of the turbulent processes on the chemical processes occurring in the flame. The Reynolds number they used is defined as

$$Re_t = R_\Lambda = \frac{u' L_i}{\nu} \tag{2}$$

where ν is the laminar kinematic viscosity. The integral length scale is the size of the largest turbulent eddies in the flow field.

Within Figure 2, Abraham, Williams, and Bracco identify two limiting regimes of turbulent combustion, reaction sheets, and distributed reactions (Williams, 1984, 1985), as originally suggested by Damkohler (1947). In the reaction sheet regime, the propagating flame front is wrinkled and convoluted by the turbulence but the turbulence does not affect the combustion chemistry; $\tau_L \ll \tau_t$. In the distributed reaction regime, pockets of fuel plus air and other pockets of combustion products are randomly mixed throughout the combustion zone.

Damkohler's original suggestion for the division separating these two limiting regimes of turbulent combustion was that $\Lambda/\delta_L = 1$, so this line is shown in Figure 2 (also shown is a parallel line for $\Lambda/\delta_L = 10^4$). A more recent suggestion, now known as the Klimov–Williams criterion (Klimov, 1963; Williams, 1976), $\eta/\delta_L = 1$, where η is the Kolmogorov scale of turbulence, is also shown in Figure 2 (included are parallel lines for the ratio with values of 10^{-2} and 10^2). The Kolmogorov scale of turbulence is the smallest size of turbulent eddies in the flow field.

Within the reaction sheet regime, turbulent flames are wrinkled laminar premixed flames. The structure of the wrinkled laminar flame depends upon the location within the reaction sheet regime on the $Da_\Lambda - Re_t$ diagram. For cases when $u'/S_L \ll 1$, the Damkohler number is very large (via Equation 1), turbulence is weak (as indicated in Figure 2), and wrinkling of the flame front by turbulence is gentle. This is a subregime that is not encountered under engine operating conditions. At higher turbulence intensities, flame front wrinkling is pronounced but a single continuous reaction sheet can still be identified. For even higher turbulence intensities, the turbulence wrinkles the flame so severely that adjacent areas of the wrinkled flame collide, cutting off pockets of unburned mixture, and forming "multiply connected" reaction sheets. In Figure 2, Abraham, Williams, and Bracco include a suggested boundary between the single sheet and multiple sheet subregimes, $u'/S_L = 1$ (Klimov, 1975, 1983), but they note that this boundary is quite uncertain.

By making some estimates, Abraham, Williams, and Bracco (1985) also imposed a rectangle that represented the operating envelope of homogeneous charge SI engines (the inset of Figure 2). Here, it should be noted that most of this rectangle lies within their multiply connected reaction sheet regime ($u'/S_L \geq 1$) although, again, they pointed out that this boundary criterion was quite uncertain.

The combustion phase diagram developed by Abraham, Williams, and Bracco (1985) only included one of the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

four modes of combustion, turbulent premixed flames, but was an important advance in our understanding of the interaction between turbulence and flame chemistry. They began this process by examining several prior relationships for the turbulent burning velocity (S_T) within the reaction sheet regime. In general, these prior relationships allowed solution for S_T as a function of the RMS of the turbulent velocity fluctuation (i.e., the turbulence intensity u') and the unstretched laminar flame speed (S_L). They concluded that the various models tended to agree that S_T/S_L depends on u'/S_L within the reaction sheet regime, but there was significant controversy over the form of the functional dependence. They further concluded that multidimensional models of premixed charge combustion within the reaction sheet regime needed to account for the turbulent flame structure "because the burning velocity and the flame surface area are the principle parameters that affect the pressure history" for homogeneous charge SI engines.

2.2 Borghi's combustion phase diagram

Our understanding of the interaction between turbulence and combustion was also significantly advanced by Borghi (1985), who developed the premixed combustion "phase diagram" that is provided in Figure 3a, following the modifications made by Peters (1986). This diagram includes

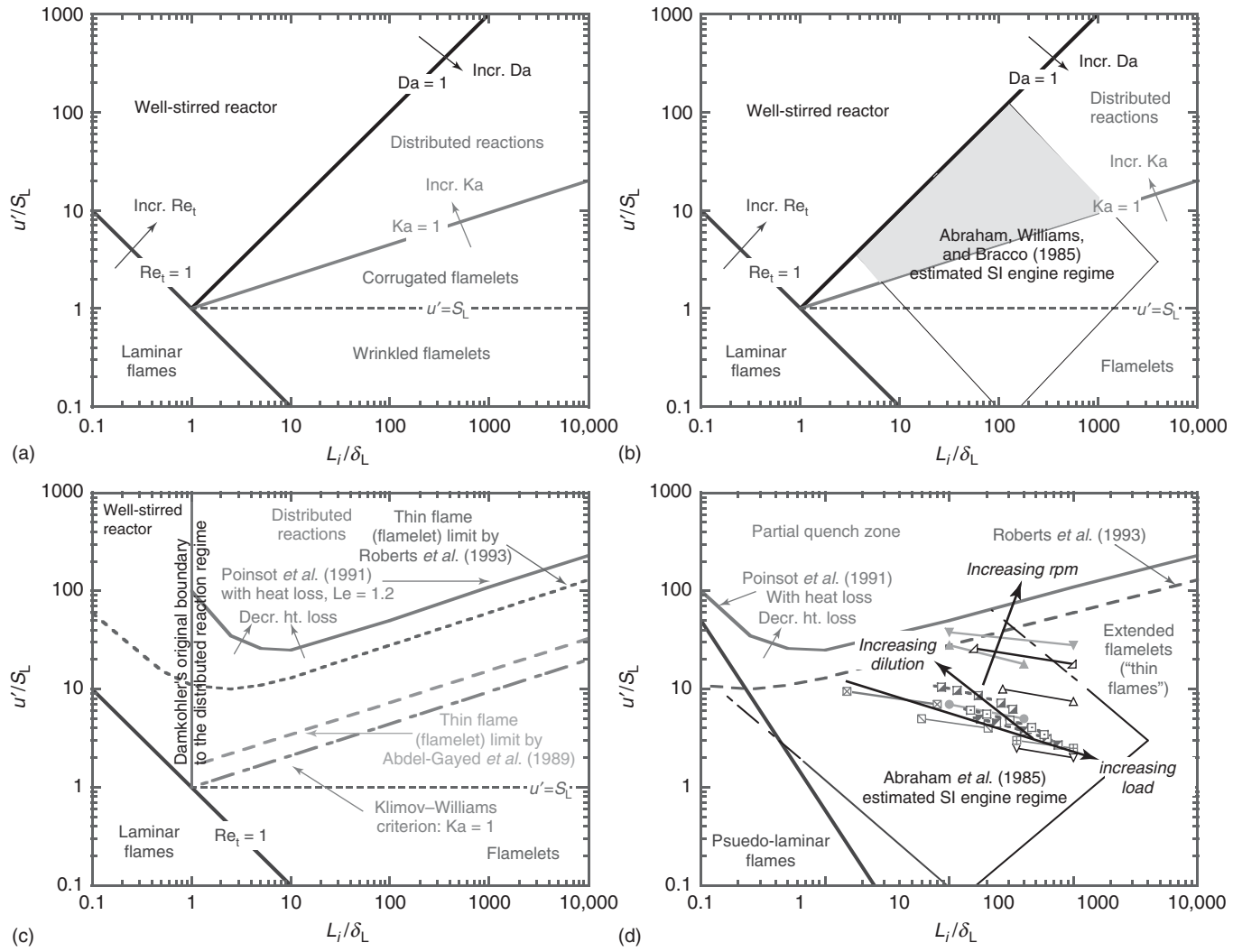


Figure 3. (a) Borghi's (1985) premixed combustion phase diagram with Peters' (1986) nomenclature. (b) Borghi's (1985) diagram with the estimated engine-operating regime from Abraham, Williams, and Bracco (1985) superimposed. (c) Borghi diagram showing shifting thoughts over the border between the flamelet and the distributed reaction regimes of turbulent premixed combustion. (d) Borghi diagram as modified by Poinsot, Veynante, and Candel (1991) and with the estimated engine operating regime from Abraham, Williams, and Bracco (1985) with the more recent engine operating results from Matthews *et al*. (1996) and Dai *et al*. (1998).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the turbulent Damkohler number and the turbulent Reynolds number that were the basis for the Abraham, Williams, and Bracco (1985) $Da-Re_t$ diagram discussed in Section 2.1. Here, the turbulent Damkohler number was defined as the ratio of a characteristic time for turbulence to a characteristic time for the reactions within the flame or, in other words, the ratio of the time scale for turbulent mixing to the time scale for reaction. Thus, this Da is essentially the equivalent of the Da used by Abraham, Williams, and Bracco (1985), discussed in Section 2.1. However, the Borghi diagram adds additional parameters that, in turn, allow division of turbulent premixed combustion into additional categories.

The Borghi combustion phase diagram is a rearrangement of the $Da_\Lambda - Re_t$ diagram, but uses the ratio of the turbulence intensity to the unstretched laminar flame speed, u'/S_L, as the abscissa rather than Re_t and the ratio of the turbulence integral length scale to the unstretched laminar flame thickness, L_i/δ_L, as the ordinate rather than Da_Λ. That is, Borghi identified two of the parameters appearing within the $Da_\Lambda - Re_t$ diagram as the most important parameters for dividing combustion into various regimes, with Da_Λ and Re_t becoming dividing lines between combustion regimes. Borghi also added the Karlovitz number, Ka (a measure of flame "stretch"), as a new parameter for dividing the regimes. The Karlovitz number that appears in Figure 3 is the ratio of the characteristic residence time within a laminar flame, t_F, to the time scale for dissipation of turbulence kinetic energy in the smallest (Kolmogorov) scale eddies, t_D:

$$\mathrm{Ka} = \frac{t_F}{t_D} = K\frac{\delta_L}{S_L} \tag{3}$$

where K is the flame stretch rate. Therefore, as used in Figure 3a, a large Ka ($\mathrm{Ka} > 1$) means that the flame is thick and/or the unstretched laminar flame speed is low (both of which could be caused by dilution with recirculated exhaust gas) and/or the flame stretch rate is high (Section 2.3.2). The flame falls within the laminar flamelet regime (corrugated and wrinkled flamelets; $\mathrm{Ka} < 1$) when the flame is thin and/or the unstretched laminar flame speed is high and/or the flame stretch rate is low.

Borghi's diagram in Figure 3 divides the regimes of turbulent premixed combustion into four categories based on the various parameters: laminar flames (whenever $Re_t < 1$), well-stirred reactors ($Re_t > 1$ and $Da < 1$), distributed reactions ($Re_t > 1$, $Da > 1$, and $Ka > 1$), and two types of laminar flamelets ($Re_t > 1$, $Da > 1$, and $Ka < 1$): (i) wrinkled flamelets ($u' < S_L$) and (ii) corrugated flamelets ($u' > S_L$). The well-stirred reactor in this diagram is similar to the perfectly stirred reactor mentioned in Section 1 except that mixing between the reactants and products is not perfect, but this regime is still an idealization of combustion in a gas turbine. In premixed combustion, laminar flamelets are flames for which the flame thickness is smaller than the smallest length scale of the turbulent flow (the Kolmogorov scale, η). In the wrinkled flamelet regime, the thin flame is only moderately wrinkled by the turbulent flow whereas the flame front is much more strongly wrinkled in the corrugated flamelet regime.

Figure 3b shows the Borghi diagram with the estimated engine-operating regime by Abraham, Williams, and Bracco (1985) superimposed (the thin-lined diamond with upper part lightly shaded in gray). According to this estimate of the engine operating regime, much of engine operation falls within the flamelet regime but much also falls within the distributed reaction regime. However, it must be noted that this estimate of the engine-operating regime relied on estimates of the turbulence intensity, integral length scale, and laminar flame thickness. More importantly, they note that "the sufficient condition for the presence of the distributed reaction regime is not satisfied in engine combustion" where they identified this condition as $L_i \ll \delta_L$ (the largest scale of the turbulent eddies must be much smaller than the laminar flame thickness). Thus, the addition of the Karlovitz number to the Borghi diagram shows that the lightly shaded gray part of the original engine-operating regime defined by Abraham, Williams, and Bracco (1985) can be eliminated by their conclusion that distributed reactions do not occur within SI engines.

Figure 3c illustrates how additional research resulted in further shifts and new definitions for the border between the flamelet and the distributed reaction regimes of turbulent premixed combustion. Several aspects of this figure are notable. First, Damkohler originally placed the border between well-stirred reactors and distributed reactions at $L_i/\delta_L = 1$ and $u'/S_L > 1$. More importantly, the border between the flamelet and the distributed reaction regime moved because of the results of additional research. In 1989, Abdel-Gayed, Bradley, and Lung argued that this border should occur at $\mathrm{Ka} = 1.64$ rather than at the Klimov–Williams criterion of $\mathrm{Ka} = 1.0$. Soon after that, Poinsot, Veynante, and Candel (1991) performed direct numerical simulations of flame/vortex interactions using the full Navier–Stokes equations and a simplified chemistry model to predict flame quenching by isolated vortices. Their formulation included non-unity Lewis number, non-constant viscosity, and heat losses so that the effect of flame stretch, flame curvature, transient dynamics, and viscous dissipation could be accounted for. Their results showed that flame fronts are much more resistant to quenching by vortices than previously expected. Their results also showed that strain is not the only important parameter

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

determining flame/vortex interaction. Heat losses, flame curvature, viscous dissipation, and transient dynamics had significant effects, especially for small scales, and they strongly influence the boundaries of the combustion regimes. It was found that the Klimov–Williams limit to the flamelet regime underestimates the size of the flamelet regime by more than an order of magnitude. Roberts *et al*. (1993) performed measurements, combined with concepts proposed by Poinsot, Veynante, and Candel (1991), in order to infer the "thin flame limit," which indicates when laminar flamelet theories become invalid, since quenching allows hot products and reactants to coexist (i.e., distributed reactions are possible). Their experiments included vortex core diameters as small as the flame thickness. Their main conclusion was that small vortices are less effective at quenching a flame than was previously believed. Therefore, the inferred regime within which thin flame (laminar flamelet) theories are valid extends to a turbulence intensity that is more than an order of magnitude larger than that which was previously predicted. Their results also indicated that micromixing models, which assume that the smallest eddies exert the largest strain on a flame, are not realistic. They found that the measured vortex Karlovitz number that is required to quench a flame is not constant but decreases by a factor of 4 as the vortex size increases from one to five flame thicknesses.

As noted previously, the well-stirred reactor regime identified in the upper left of Figure 3c is an idealization of the combustion process in a gas turbine. Although satisfactory agreement can be obtained between this type of model for combustion in a gas turbine and experimental data (Swithenbank, Poli, and Vincent, 1973), there are strong limitations to the well-stirred reactor approach (e.g., Kuo, 1986). However, because gas turbines are not used in the automotive market, gas turbine-type combustion will not be discussed any further. The interested reader is referred to Lefebvre (1983), Treager (1979), Mattingly (1996), and Turns (2000).

Figure 3c is the Borghi combustion phase diagram as modified by Poinsot, Veynante, and Candel (1991) via the direct numerical simulations discussed above. They defined an "extended flamelet" as a regime in which the reactants and products are separated by a thin reaction zone that is laminar-like but does not necessarily have the structure of a laminar flame. They chose to examine quenching because "If the local stretch induced by the turbulent flow … is sufficiently large and the flame is quenched at a given location, combustion stops in the vicinity of this point and fresh reactants will diffuse into the products without burning. Combustion ceases to take place in thin sheets and the flamelet concepts (laminar or extended) become less adequate. Therefore, quenching in a turbulent premixed flame determines the limit between two essentially different behaviors (i.e. flamelets or no flamelets)." Thus, in Figure 3, Poinsot, Veynante, and Candel replaced the distributed reaction regime (Figure 3c) with the "partial quench zone." Their DNS results also resulted in the boundary between the extended flamelet and partial quench zones illustrated in Figure 3, where the position of this boundary depends upon the heat loss from the flame.

Figure 3d includes an overlay of the engine-operating regime estimated by Abraham, Williams, and Bracco (1985) and the results of a more recent examination of engine-operating regimes for typical current technology homogeneous charge SI engines by Matthews *et al*. (1996) and the extension of this work by Dai *et al*. (1998). These latter results, represented by the data symbols in the plot, relied upon a well-calibrated quasi-dimensional engine model (including a calibrated turbulence model) to extract u'/S_L and L_i/δ_L throughout the combustion process of a 4.6 L V8 SI engine (Matthews *et al*., 1996) and both a 1.6 L and a 2.0-L four-valve I4 SI engine (Dai *et al*., 1998) over a range of operating conditions. The base operating conditions for these engines were 1500 rpm, 2.62 bar brake mean effective pressure (BMEP), stoichiometric, no exhaust gas recirculation (EGR), and minimum spark advance for best torque (MBT) spark timing. These investigations examined the effects of engine speed, load, and dilution with both excess air and recirculated exhaust gas. Only one parameter was varied at a time. In Figure 3d, the beginning of combustion is the far left point on each curve or line with the far right data point on that same curve or line as the end of combustion. The three curves (rather than straight lines) were generated by examining not just the beginning and end of combustion but a series of states between these two extremes. More importantly, as the engine speed increases the lines for each operating speed shift up and to the right; as the load increases the lines shift down and to the right, and as the mixture becomes more diluted the curves shift up and to the left. Only the last stages of combustion for the V8 at 5500 rpm, stoichiometric, and for the 1.6-L inline 4 at 9000 rpm, rich, appear to be outside of the engine-operating regime originally estimated by Abraham, Williams, and Bracco (1985). More details are available in the original papers (Matthews *et al*., 1996; Dai *et al*., 1998).

One conclusion that must be drawn from Figure 3d is that Abraham, Williams, and Bracco (1985) estimated the engine-operating regime fairly well, especially given that the "engineering tools" available for aiding such estimates have progressed significantly since they made their estimate. Their lower limit Reynolds number appears to have been a bit too conservative but their upper limit Reynolds number was quite accurate, except for some

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

engines operating at high engine speeds especially with a low load and/or with high dilution.

The most important conclusion to be drawn from Figure 3d is that flames within SI engines fall within the laminar flamelet regime of turbulent premixed combustion for most, or perhaps all, operating conditions. In this regime, the dominant effects of turbulence are to stretch and wrinkle the flame. Within the flamelet regime, the reactions are sufficiently fast that the burned gases and the unburned gases can be considered as separated by very thin flamelets, and the local structure of the flamelets is essentially that of a laminar flame. In other words, the dominant effects of turbulence on the premixed flame are to wrinkle and stretch the flame surface while the inner flame structure is not significantly altered by the turbulent flow field. The significance of the flamelet assumption is that it decouples the chemistry from the turbulence and, thereby, reduces the problem of modeling premixed turbulent combustion to the description of flame surfaces: the flame propagates locally at the instantaneous stretched laminar flame speed but the global turbulent flame propagates faster than the corresponding laminar flame because the flame surface area is increased by flame wrinkling. This means that the mass burning rate of the flame ($\dot{m}_b$) can be expressed as

$$\dot{m}_b = \rho_u u_L A_T = \rho_u I_o S_L A_T \tag{4}$$

where ρ_u is the unburned reactant (i.e., fuel–air-diluent mixture) density, u_L is the local instantaneous stretched laminar flame propagation speed, and A_T is the total (i.e., turbulent) surface area of the flame. Furthermore, u_L in Equation 4 is given by $I_o S_L$, where I_o is the flame stretch parameter, discussed more in Section 2.3.2, and S_L, as already defined, is the unstretched laminar flame speed discussed in Section 2.3.1.

Figure 4 illustrates the wrinkled laminar flame. Turbulent eddies over the full range of sizes from the integral length scale on the large end of the spectrum to the Kolmogorov scale on the small end, interact with the surface of the flame and, thereby, wrinkle and stretch the very thin flame. This variety of eddy sizes produces concave wrinkles in the flame surface of various sizes. Each local flamelet then propagates normal to the local flame surface at the local, instantaneous, stretched laminar flame speed, u_L. In turn, the small-scale concave wrinkles tend to become annihilated by the local flame propagation process while any large-scale convex wrinkles (portions of the flame not recently impacted by large eddies) tend to grow larger. Thus, the range of flame wrinkling scales differs somewhat from the range of turbulent eddy sizes. In addition, the flame wrinkling process continues as the flame propagates through the turbulent flow field. However, if one could freeze the flow and the reactions and probe any local flamelet to measure the local, instantaneous composition and temperature profiles of the very thin flamelet, this would yield results very similar to those shown in Figure 1 for the unstretched laminar premixed flame except that the local flamelet would be propagating normal to the flame surface at a somewhat different speed: u_L rather than S_L.

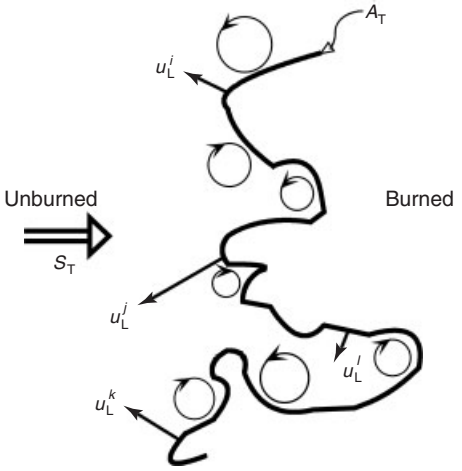


Figure 4. Snapshot illustration of turbulent premixed flame propagation. The turbulent approach flow has mean velocity S_T, the flame propagates normal to the local surface at velocity u_L (the superscript indicates that u_L is dependent upon the location along the instantaneous flame surface), and the total (turbulent) flame surface area is A_T.

2.3 Turbulent premixed flame speed

Several SI engine combustion models rapidly took advantage of the flamelet assumption, such as the coherent flamelet model (e.g., Cheng and Diringer, 1991; Boudier *et al*., 1993; Duclos, Veynante, and Poinsot, 1993) and the fractal flame model (e.g., Matthews and Chin, 1991; Chin *et al*., 1992; Zhao, Matthews, and Ellzey, 1993, 1994). Such models must (i) determine the instantaneous unstretched laminar flame speed S_L, (ii) quantitatively describe the effects of flame stretch on the local laminar flame speed (e.g., via I_o), and (iii) account for the increase in flame surface area caused by the turbulence. Each of these requirements is discussed in the following sections.

2.3.1 The unstretched laminar flame speed

As just noted, engine models that take advantage of the flamelet assumption must calculate the instantaneous

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

unstretched laminar flame speed. The unstretched laminar flame speed is a thermochemical property of a premixed reactive mixture. It depends upon the composition of the reactants (including the fuel, the air, and the amount of exhaust residual and/or EGR that may be present), the reactant temperature, and the system pressure.

The unstretched laminar flame speed, S_L, can be predicted using detailed kinetics models of the type discussed in Fundamental Chemical Kinetics (or appropriate simplified kinetics models if available) in conjunction with a suitable one-dimensional flame simulation code. Somewhat surprisingly, few codes are available for predicting this fundamental quantity. Perhaps the most widely used is PREMIX (Kee *et al*., 1985). PREMIX calls a very accurate code, as a subroutine, for calculation of the transport properties (TRANFIT, Kee, Warnatz, and Miller, 1983) since these transport properties have a significant influence on the flame propagation rate. Obviously, the results are also highly dependent upon the accuracy of the kinetics mechanism. Accurate high temperature kinetics mechanisms (i.e., schemes that are suitable for predicting flame propagation) are available for methane (which closely resembles natural gas), propane, and a variety of other hydrocarbons and alcohols, but not for gasoline. However, a kinetics mechanism is available for primary reference fuel (PRF) blends that can be used to predict the unstretched laminar flame speeds of various simulated gasolines. Unfortunately, such predictions are highly consumptive of computational time, since the time requirement increases roughly with the square of the number of species involved (Westbrook and Dryer, 1984). For methane, perhaps the simplest accurate detailed kinetics mechanism consists of 57 reactions and 25 species (Kee *et al*., 1985), while a recent detailed kinetics model for iso-octane (Curran *et al*., 1996) consists of 3445 elementary reactions and 805 species.

Although challenging, one method of determining S_L for engine simulation codes and accounting for the effects of pressure, unburned gas temperature, equivalence ratio, and residual gas fraction plus EGR on S_L is to use a laminar flame speed code to generate extensive tables of S_L for a broad range of conditions. These tables can then be called from the engine simulation code as a "laminar flame speed library."

An alternative approach is to predict the unstretched laminar flame speed using experimentally derived empirical correlations. Before introducing an example of an unstretched laminar flame speed correlation, it is first necessary to define a few symbols and terms. The unburned gas temperature, T_u, is the temperature of the reactants ahead of the flame in the unburned zone. The residual mass fraction, f, is the fraction of the mixture trapped within the cylinder at intake valve closing that consists of exhaust products remaining in the cylinder from the previous cycle. The equivalence ratio, ϕ, sometimes called the fuel/air equivalence ratio, is defined via

$$\phi \equiv \frac{F/A}{(F/A)_s} = \frac{(A/F)_s}{A/F} = \frac{1}{\lambda} \tag{5}$$

In the definition of ϕ, F/A is the actual fuel/air mass ratio, $(F/A)_s$ is the stoichiometric fuel/air mass ratio, A/F and $(A/F)_s$ are their respective reciprocals, and λ is the excess air ratio. For stoichiometric mixtures (no excess air, no excess fuel), $\phi = 1.0$, while for lean mixtures $\phi < 1.0$ and for rich mixtures $\phi > 1.0$.

Perhaps the most widely used empirical correlations for the unstretched laminar flame speed were developed at MIT (Metghalchi and Keck, 1980, 1982; Milton and Keck, 1984; Rhodes and Keck, 1985). This empirical correlation is valid for $0.7 < \phi < 1.6$, $0.4 < P < 12$ atm, $350 < T_u < 550$ K, and residual mass fractions (including EGR) up to 30% and is

$$S_L = S_L^o \left(\frac{T_u}{298}\right)^a \left(\frac{P}{1.0}\right)^b [1 - 2.06 f^{0.77}] \tag{6}$$

where the cylinder pressure P must be in atmospheres, the reference pressure is 1.0 atmosphere, and S_L^o is the unstretched laminar flame speed at the reference state of 298 K, 1 atm, with no dilution by residual or EGR ($f = 0$). S_L^o is empirically related to the equivalence ratio via

$$S_L^o = B_m - B_\phi (\phi - \phi_m)^2 \tag{7}$$

where the constants B_m, B_ϕ, and ϕ_m are given in Table 1 for four fuels.

For propane, iso-octane, and methanol, coefficients a and b in Equation 6 are empirically related to the equivalence ratio via

$$a = 2.18 - 0.8(\phi - 1) \tag{8}$$

Table 1. Coefficients for use in the empirical equations for the unstretched laminar flame speed.

Fuel	ϕ_m	B_m (cm/s)	B_ϕ (cm/s)
Methanol	1.11	36.9	140.5
Propane	1.08	34.2	138.7
Iso-octane	1.13	26.3	84.7
Reference gasoline	1.21	30.5	54.9

Metghalchi and Keck (1980, 1982), Milton and Keck (1984), and Rhodes and Keck (1985).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

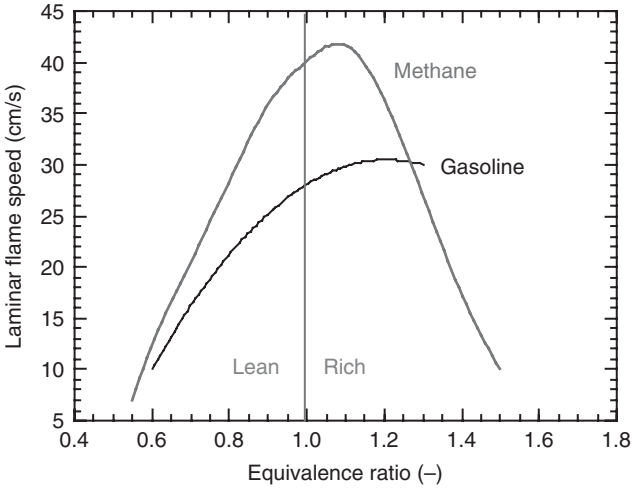


Figure 5. The effects of equivalence ratio on the unstretched laminar flame speeds of methane and gasoline at a pressure of 1 bar with a reactant temperature of 298 K. These results are from correlation equations obtained by fits to experimental data; gasoline from Metghalchi and Keck (1982); methane from Andrews and Bradley (1972).

$$b = -0.16 + 0.22(\phi - 1) \tag{9}$$

while for a reference gasoline, a and b are the following functions of equivalence ratio:

$$a = 2.4 - 0.271\phi^{3.51} \tag{10}$$

$$b = -0.357 + 0.11\phi^{2.77} \tag{11}$$

Using correlations such as these, the effects of equivalence ratio on the unstretched laminar flame speed are illustrated in Figure 5. The unstretched laminar flame speed is highest for a somewhat rich mixture not only for methane and the reference gasoline but also for all hydrocarbons. The same correlations can be used to show the effects of pressure, unburned gas temperature, and exhaust residual fraction plus EGR. An example is provided in Figure 6, which shows that the unstretched laminar flame speed decreases as the system pressure increases. This is caused by a falloff of chemical kinetic rates with increasing pressure. However, in spite of the pressure falloff of the combustion chemistry, the mass burning rate given by Equation 4 increases with increasing pressure because it is also a function of the density of the unburned mixture, and the effect of pressure on density increases faster with increasing system pressure (linearly) than the effect of the pressure falloff of the chemical kinetic rates.

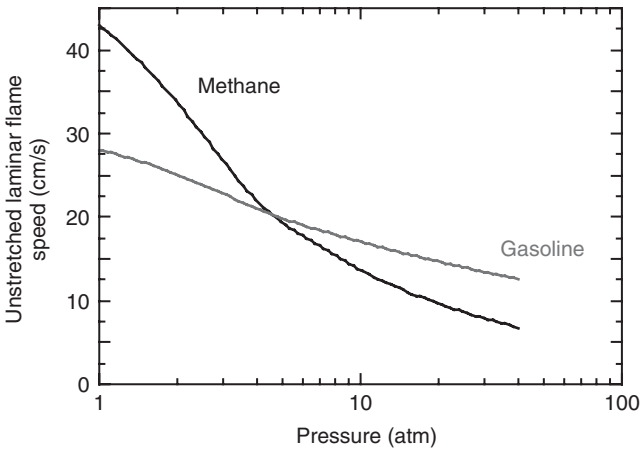


Figure 6. The effects of pressure on the unstretched laminar flame speeds of methane and gasoline for stoichiometric mixtures with a reactant temperature of 298 K. These results are from correlation equations obtained by fits to experimental data; gasoline from Metghalchi and Keck (1982); methane from Andrews and Bradley (1972).

2.3.2 *The effects of flame stretch on the laminar flame speed*

In general, flame expansion and turbulence strain slow down a laminar flame. These effects may be quantified via

$$u_L = I_o S_L \tag{12}$$

where u_L is the stretched laminar flame speed and I_o is the dimensionless flame stretch parameter.

Chung and Law (1988) showed that the flame stretch parameter is a function of the flame stretch rate (K), assuming a relatively small effect of stretch on the flame speed:

$$I_o = \frac{u_L}{S_L} = 1 - \left[\frac{1}{\mathrm{Le}} - \left(\frac{1}{\mathrm{Le}} - 1\right)\frac{E_A}{2\overline{R}T_{ad}}\right] \times \frac{D}{S_L^2}K - q_{rad}\frac{E_A}{2\overline{R}T_{ad}} \tag{13}$$

where $\overline{R}$ is the universal gas constant, Le is the Lewis number (the ratio of the thermal diffusivity to the mass diffusivity, α/D) of the deficient reactant (the fuel diffusing through the air), E_A is the global activation energy, T_{ad} is the adiabatic flame temperature, K is the flame stretch rate, and q_{rad} is the radiation heat loss from the flame. Assuming negligible heat loss from the flame via radiation (an excellent assumption for homogeneous charge SI engines), a Lewis number of approximately 1, and a Prandtl number

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

(Pr = ν/α) of approximately 1 yields (Chin *et al*., 1992):

$$I_o = \frac{u_L}{S_L} = 1 - \frac{\nu}{S_L^2} K \tag{14}$$

where ν is the laminar kinematic viscosity in the unburned gases. The flame stretch rate is composed of two additive components (Chung and Law, 1988):

$$K = K_E + K_S \tag{15}$$

where K_E is the flame stretch rate associated with flame expansion or large-scale curvature of the flame and K_S is the stretch rate associated with turbulence strain in the small scale wrinkles.

For a spherical flame, as is normally assumed for flame propagation in a homogeneous charge SI engine in quasi-dimensional engine simulations, Law (1988) showed that

$$K_E = \frac{2}{r_f}\frac{dr_f}{dt} = \frac{2}{r_f}\left[\frac{\rho_u}{\rho_b} S_L\right] \tag{16}$$

where r_f is the flame radius, the rate of change of the flame radius is equal to the terms within the square brackets if the flame thickness is much smaller than the flame radius, and ρ_u and ρ_b are the densities in the unburned and burned zones, respectively.

Turbulence strain can be expressed as the lifetime of an eddy of the size of the Taylor microscale, λ:

$$K_S = \frac{u'}{\lambda} \tag{17}$$

Various investigators (Herweg and Maly, 1992; Chin *et al*., 1992; Abdel-Gayed and Bradley, 1985, Abdel-Gayed, Bradley, and Lau, 1988) used this expression in combination with relationships for the Taylor microscale of turbulence, the integral length scale, the turbulence intensity, and the laminar kinematic viscosity to derive various, but similar, relationships for turbulence strain in their quasi-dimensional engine models with the form:

$$K_S = C_S \sqrt{\frac{\varepsilon}{\nu}} \tag{19}$$

where ε is the rate of dissipation of turbulence kinetic energy:

$$\varepsilon = \frac{(u')^3}{L_i} \tag{20}$$

and the constant, C_S, varied from 0.157 (Abdel-Gayed and Bradley, 1985) to 0.430 (Chin *et al*., 1992) depending upon the theory used to derive the constant. In spite of the factor of 3 difference in the flame stretch rates due to turbulence strain, it was found (Wu *et al*., 1993) that there was little effect on the burning rate predictions of a quasi-dimensional engine model for a stoichiometric propane/air mixture. More significant effects might be found for dilute mixtures and/or engine designs and operating conditions that produce high turbulence intensities.

No experimental data for engines is available for directly validating these flame stretch submodels, but all have a common form (when the Lewis and Prandtl numbers are both assumed to be 1.0) and differ only in magnitude.

2.3.3 *Turbulent premixed flame speed*

Under engine-like conditions, a turbulent flame will burn approximately 10 times faster than the corresponding laminar flame, primarily because of the increase in flame surface area caused by the turbulence. A relationship between the turbulent flame speed and the unstretched laminar flame speed is (e.g., Turns, 2000)

$$S_T = S_L \frac{A_T}{A_L} \tag{21a}$$

where S_T is the turbulent flame speed, S_L is the unstretched laminar flame speed, A_T is the total (turbulent) surface area of the wrinkled flame, and A_L is the surface area of a smooth laminar flame with the same mean radius as the turbulent flame. However, Equation 21a is more accurately written as

$$S_T = u_L \frac{A_T}{A_L} = I_o S_L \frac{A_T}{A_L} \tag{21b}$$

Equation 21b accounts for the additional effects of turbulence strain and flame curvature, both of which modify the unstretched laminar flame speed as discussed in the previous section, generally in a manner that decreases it compared to a laminar flame that is flat and not subjected to strain by a turbulent flow field. While these flame curvature and turbulence strain effects are important and decrease the turbulent flame speed, they do not overcome the turbulent flame speed increase induced by turbulent wrinkling of the flame. Thus, overall a turbulent flame burns faster than the corresponding laminar flame because of the dominant effect of the increase in surface area of the flame due to wrinkling of the flame front by the turbulent eddies.

The variety of turbulent premixed combustion models, such as those noted previously (e.g., Cheng and Diringer, 1991; Matthews and Chin, 1991; Chin *et al*., 1992; Herweg and Maly, 1992; Boudier *et al*., 1993; Duclos, Veynante, and Poinsot, 1993; Zhao, Matthews, and Ellzey, 1993,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

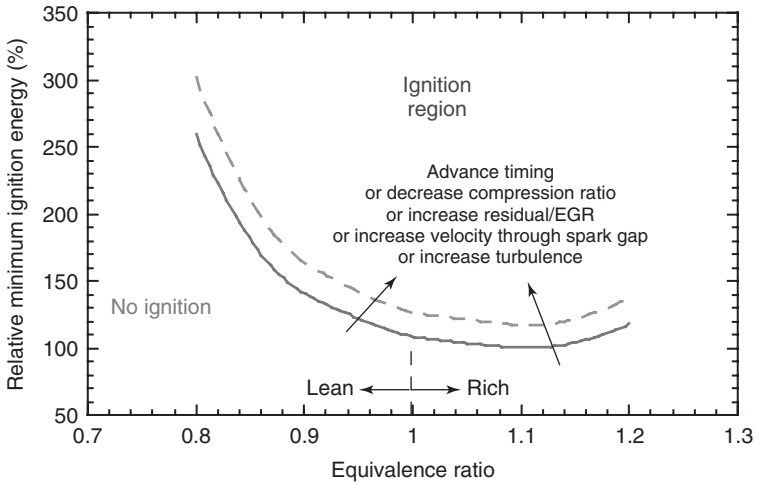


Figure 7. Effects of the equivalence ratio and other parameters on the energy required to achieve ignition. Figure generated from equations presented by Ballal and Lefebvre (1974).

1994), indicate that there are a wide variety of methods used to account for the increase in flame surface area due to wrinkling of the flame front by the turbulent flow field (the A_T/A_L term in Equation 21b).

2.3.4 Effects of stratification

An advanced, higher efficiency, spark-ignition engine concept involves using a fuel/air mixture with an equivalence ratio that is stratified (i.e., varies spatially) in the engine cylinder rather than a homogeneous premixed charge. Stratification is achieved by appropriate timing of fuel injection directly into the cylinder. The goal of this approach is to use the quantity of fuel injected to control load, rather than throttling the intake charge, and achieve overall lean operation at part load conditions. The net effect is an engine with significantly improved part load fuel efficiency. For optimal combustion, these stratified charge engines seek to place a stoichiometric or somewhat rich mixture near the spark plug at the time of spark with the mixture becoming progressively leaner away from the spark plug. The reason that it is important to produce a nearly stoichiometric mixture near the spark gap at the time of ignition is illustrated in Figure 7. This figure shows that stoichiometric and somewhat rich mixtures are much easier to ignite than lean mixtures, producing more stable ignition and combustion of a stratified and overall lean fuel–air mixture.

As the flame propagates through this mixture after ignition, it is evident from the laminar flame speeds in Figure 5 that the flame in the stratified environment will initially propagate relatively fast, and that the flame speed will decrease as the flame encounters increasingly lean mixtures away from the spark plug. In general, the turbulence intensity also decays throughout the combustion process, which further decreases the burning rate as the flame propagates through a stratified charge. This slowing of the combustion rate in the leaner mixtures can lead to combustion instabilities, one of the major design challenges for commercializing this approach for improving engine efficiency.

However, modeling such a turbulent, premixed, stratified charge is relatively straightforward. As noted above, for the homogeneous charge turbulent premixed flame, the modeler must account for the local instantaneous effects of flame stretch, flame wrinkling, and the unstretched laminar flame speed, which varies only because of the rapidly changing cylinder pressure and unburned gas temperature. For the turbulent, premixed, stratified charge engine, the modeler must also account for the local, instantaneous effects of the equivalence ratio on the unstretched laminar flame speed.

3 MIXING CONTROLLED COMBUSTION

Section 2 of this chapter pertained to the mode of combustion that is most relevant to SI engines. In diesel engines, the rate of heat release by the combustion process is dominated by mixing of the fuel with the air, both of which are separated by a flame front. This is classically called a diffusion, or non-premixed, flame, which is marked by separation of the fuel from the air, even under laminar conditions and with a gaseous fuel. In a diffusion flame, the fuel molecules diffuse toward the air, and the "air molecules"

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

diffuse toward the fuel, with both diffusion processes driven by concentration gradients. Thus, in a diffusion flame the mass burning rate is limited by the rate of mixing of the fuel and air. That is, diffusion flames are "mixing controlled."

The subject of mixing controlled combustion in a fuel jet is best introduced by first examining a nonreacting laminar jet of a gaseous fuel issuing into quiescent ambient oxidizer (such as air). The following assumptions will be used (Turns, 2000):

- The molecular weights of the fuel and oxidizer are equal. This assumption, combined with ideal gas behavior and constant pressure and temperature, yields uniform density throughout the flowfield.
- Transport of species is via binary diffusion governed by Fick's law.
- Momentum and species diffusivities are constant and equal (thus, the Schmidt number = Sc = $\nu/D = 1.0$, where ν is the momentum diffusivity (kinematic viscosity) and D is the species or mass diffusivity).
- Only the radial diffusion of momentum and species is important; axial diffusion is neglected (thus, the solution will only be valid some distance downstream from the nozzle exit).

Given these assumptions, the governing equations are as follows:

1. *Continuity (conservation of mass)*:

$$\frac{\partial v_x}{\partial x} + \frac{1}{r}\frac{\partial (v_r r)}{\partial r} = 0 \tag{25}$$

2. *Conservation of axial momentum*:

$$v_x \frac{\partial v_x}{\partial x} + v_r \frac{\partial v_x}{\partial r} = \nu \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_x}{\partial r}\right) \tag{26}$$

3. *Conservation of species*:

$$v_x \frac{\partial Y_\mathrm{f}}{\partial x} + v_r \frac{\partial Y_\mathrm{f}}{\partial r} = D \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial Y_\mathrm{f}}{\partial r}\right) \tag{27}$$

where x is the axis along the centerline of the jet, r is the radial direction, v_x is the velocity in the direction that is parallel to the jet axis, v_r is the velocity perpendicular to the jet axis, ν is the kinematic viscosity, D is the molecular diffusion coefficient, and Y_f is the mass fraction of the fuel. Furthermore, because there are only two species in this simplified analysis,

$$Y_\mathrm{f} + Y_\mathrm{ox} = 1 \tag{28}$$

where Y_ox is the mass fraction of the oxidizer.

The boundary conditions along the jet centerline ($r = 0$) are

$$v_r(x, 0) = 0 \tag{29a}$$

$$\frac{\partial v_x}{\partial r}(x, 0) = 0 \tag{29b}$$

$$\frac{\partial Y_\mathrm{f}}{\partial r}(x, 0) = 0 \tag{29c}$$

Moreover, at large radii,

$$v_x(x, \infty) = 0 \tag{30a}$$

$$Y_\mathrm{f}(x, \infty) = 0 \tag{30b}$$

Finally, at the jet exit it is assumed that the axial velocity, v_e, and fuel mass fraction, $Y_\mathrm{f,e}$, are uniform across the jet:

$$v_x(0, r \le R) = v_\mathrm{e} \tag{31a}$$
$$Y_\mathrm{f}(0, r \le R) = Y_\mathrm{f,e} = 1$$

where R is the radius of the jet nozzle. Elsewhere in the plane of the nozzle exit

$$v_x(0, r > R) = 0 \tag{31b}$$
$$Y_\mathrm{f}(0, r > R) = 0$$

The solution for the fuel mass fraction distribution is

$$Y_\mathrm{f} = \frac{3}{8\pi}\frac{v_\mathrm{e}\pi R^2}{D}\frac{1}{x}\left[1 + \frac{\left\{\left(\frac{3}{16}\right)(\rho_\mathrm{e} v_\mathrm{e} R)\frac{1}{\mu}\frac{r}{x}\right\}^2}{4}\right]^{-2} \tag{32a}$$

where ρ_e is the density at the jet exit and μ is the dynamic viscosity. Equation 32a may be rewritten as

$$Y_\mathrm{f} = \frac{3}{8\pi}\frac{\dot{V}_\mathrm{f,e}}{D}\frac{1}{x}\left[1 + \frac{\zeta^2}{4}\right]^{-2} \tag{32b}$$

where $\dot{V}_\mathrm{f,e}$ is the volumetric flow rate of fuel at the jet exit and

$$\zeta \equiv \left(\frac{3\rho_\mathrm{e} J_\mathrm{e}}{16\pi}\right)^{1/2}\frac{1}{\mu}\frac{r}{x}$$

where J_e is the momentum flow at the jet exit:

$$J_\mathrm{e} = \rho_\mathrm{e} v_\mathrm{e}^2 \pi R^2$$

A solution for the fuel and oxidizer mass fraction profiles resulting from Equations 28 and 32a as a function of radial distance from the jet centerline is illustrated in Figure 8. The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

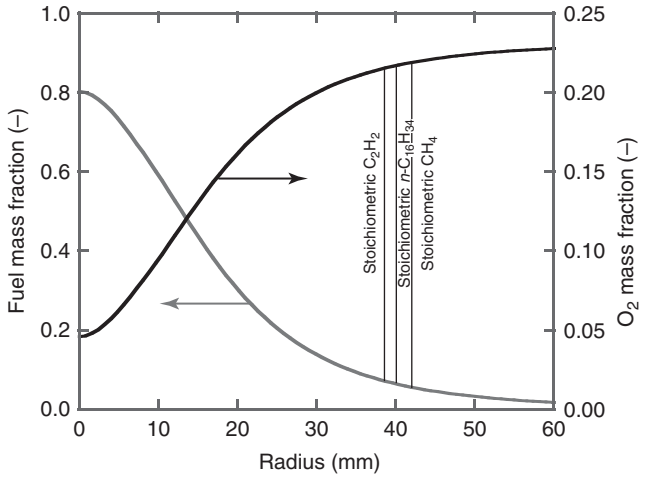


Figure 8. Fuel and molecular oxygen mass fraction profiles through a nonreacting laminar jet of fuel issuing into quiescent air through a 10-mm diameter nozzle at 10 cm/s. These profiles are for a plane that is 7.5 nozzle diameters downstream from the jet exit.

impacts of fuel and oxidizer diffusion are clearly evident. The radius at which the fuel/air mixture is stoichiometric at an axial distance of 75 mm is also illustrated for three different fuels by the vertical lines near methane (CH_4, H/C atom ratio = 4.0), *n*-hexadecane ($C_{16}H_{34}$, a primary reference fuel for Cetane number tests, H/C = 2.125), and ethyne (more commonly called acetylene, C_2H_2, H/C = 1.0). The results in Figure 8 illustrate that the stoichiometric contour is near the outer boundary of the jet and that fuel type will affect the radial location of this contour.

Figure 9 illustrates the locus of radii at which a stoichiometric mixture is located for the *n*-hexadecane and air example, with all other conditions being the same as for Figure 8. This stoichiometric surface is of interest because this is approximately where the flame would be located if the mixture were ignited. Fuel would diffuse from the inside and oxidizer from the outside to the flame at the stoichiometric surface.

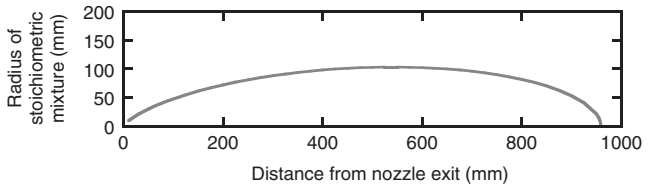


Figure 9. Radius at which a stoichiometric mixture of *n*-hexadecane and air is located for a nonreacting laminar jet of fuel (*n*-hexadecane) issuing into quiescent air; same conditions as Figure 8.

Taking the next step and extending this solution to the case of a reacting laminar jet is more difficult because of the following reasons:

- Constant density and viscosity should no longer be assumed.
- The species conservation equation will have a source/sink term.
- A species conservation equation will need to be employed for a minimum of three species (fuel, oxidizer, and product) and more than three species (e.g., fuel, O_2, N_2, CO_2, and H_2O) if increased accuracy is desired.
- Because the solution will be very sensitive to the transport properties, the Schmidt number should no longer be assumed equal to 1.0.
- An energy conservation equation must be included in the set of governing equations, and it will also include a source/sink term due to the chemical reactions.

However, a solution for a laminar jet diffusion flame has been developed (Turns, 2000) from which Figure 10 was derived. This solution is highly simplified because only three species were considered and various other simplifying assumptions were made. Among the assumptions is that chemical kinetics are infinitely fast, such that the flame is an infinitesimally thin surface at which the fuel and oxidant disappear and the products appear. This is the "flame sheet approximation," which is somewhat analogous to the thin flame (flamelet) assumption used for premixed turbulent flames. The solution shown in Figure 10 demonstrates that the fuel and oxidant disappear at the flame front (see Figure 10c) and that the "product" is formed at the flame front and diffuses away in both directions.

Figure 10 presents an illustration of this simple type of laminar diffusion flame: a laminar gaseous fuel jet issuing from a nozzle into quiescent ambient air. This illustrates a few essential features of mixing-controlled combustion. The type of diffusion flame shown in Figure 10 is highly idealized relative to actual diesel combustion, but is representative of the type of combustion that can occur in a diesel engine in a fuel jet issuing from an orifice of a fuel injector during the fuel injection phase of combustion. (See Diesel and Diesel LTC Combustion and Dec, 1997, for detailed discussion of combustion in diesels.) The flame will be located along the stoichiometric contour that is shown in Figure 9. Fuel diffuses radially away from the jet centerline to the flame front while the oxidizer diffuses in the opposite direction from the ambient gas to the flame front.

To proceed further and develop a simulation of a turbulent diffusion flame, the question again arises as to how

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

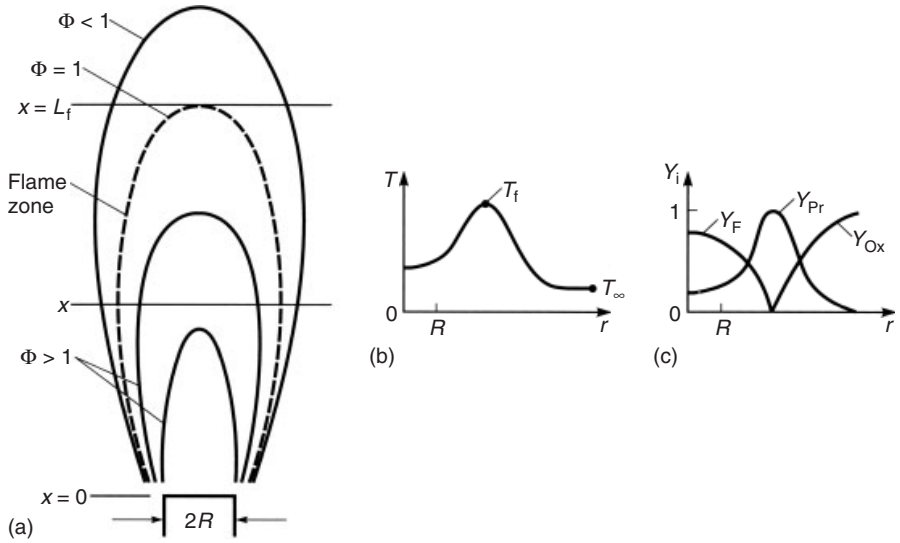


Figure 10. (a–c) Illustration of a simple laminar jet diffusion flame with a gaseous fuel. (Reproduced with permission from S.R. Turns, *An Introduction to Combustion* (Turns, 2000). © The McGraw-Hill Companies, Inc.)

the turbulence affects the combustion process. Turbulence increases the transport of momentum, species, and thermal energy and increases (wrinkles) the flame surface area (although for diffusion flames, unlike premixed flames, the flame surface area does not appear as an independent parameter). Again, various models have been developed to account for wrinkling of diffusion flames and the other effects of turbulence. An example of a gaseous jet turbulent diffusion flame is provided in Figure 11, which shows the diffusion flame location at various instants in time superimposed on top of each other and the minimum and maximum extent of the diffusion flame.

Here, it should be noted that the diesel-type diffusion combustion process is also complicated by several other factors that include the injected fuel being a liquid rather than a gas, the surrounding air being not only not quiescent but also a turbulent flowfield with potentially high velocities, and the process being transient rather than a steady state diffusion flame. Further discussions of the details of diesel combustion are provided in Diesel and Diesel LTC Combustion along with conceptual models for diesel combustion as it occurs in an engine.

It should also be noted that mixing-controlled combustion can play a role in hydrocarbon emissions from SI engines. These emissions are primarily due to fuel that emerges after peak pressure from places it collected during the engine cycle, such as deposits, oil layers, and—especially—crevices. This unburned fuel emerges when the cylinder pressure falls sufficiently low during the expansion stroke, and/or during "exhaust blowdown," and/or during the exhaust stroke. Whether this protected fuel eventually completely or partially oxidizes or does not react at all depends upon mixing with the hot products of combustion before they cool too much. Liquid fuel films on the cylinder walls, cylinder head, and top of the piston in port-injected SI engines, direct injected SI engines, and diesels can also produce mixing controlled combustion, leading to both hydrocarbon and soot emissions.

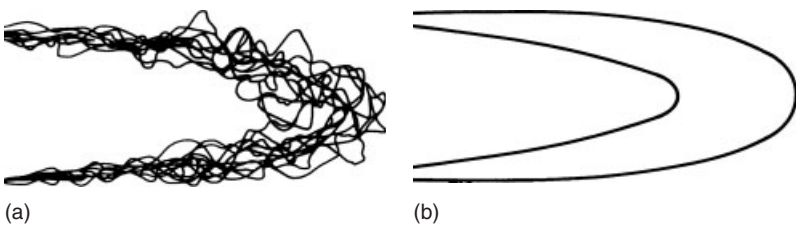


Figure 11. (a) Example time-resolved images of a turbulent diffusion flame (b) with the "flame brush" (the minimum and maximum extent of the diffusion flame). (Reproduced with permission from S.R. Turns, *An Introduction to Combustion* (Turns, 2000). © The McGraw-Hill Companies, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4 VOLUMETRIC ENERGY RELEASE

Volumetric energy release refers to a mode of combustion that differs significantly from the three modes introduced in the previous sections (premixed, diffusion, and well stirred). In this mode of combustion, reactants reach temperature, pressure, and mixture conditions that enable a spontaneous rapid oxidation process, basically an autoignition process. Examples of autoignition volumetric combustion include knock in SI engines, the ignition and premixed burn process in diesel engines, and the combustion process in high efficiency engines under development that employ various forms of low temperature combustion (LTC) [e.g., HCCI, controlled autoignition (CAI), and premixed charge compression ignition (PCCI)]. The volumetric energy release mode of combustion is dominated by chemical kinetic reaction rates. For high reactivity fuels (high Cetane number or low Octane number), the autoignition typically occurs in two stages. The first stage is initiated and controlled by low-to-intermediate temperature ($<$1000 K) hydrocarbon oxidation chemistry. During this stage, fuel is partially broken down and energy is released. This first stage helps raise the temperature toward approximately 1000 K, at which point "high temperature kinetics" kicks in and dominates the second stage of ignition. The ignition process rapidly progresses from here to a high heat release phase. For low reactivity fuels (high Octane number and low Cetane number), the low-to-intermediate temperature oxidation chemistry first stage of ignition is either not present or not nearly as significant for many relevant conditions. In this case, the ignition process takes longer for the same conditions and progresses directly into the high temperature kinetics phase when temperatures reach approximately 1000 K. Ignition processes and relevant chemical kinetics are discussed in detail in Fundamental Chemical Kinetics.

Knock in SI engines can be used as an example of this combustion mode. As the flame in a SI engine propagates away from the spark plug, the burned mixture behind the flame has a much higher temperature than the unburned mixture in front of the flame. Since both the burned and unburned mixtures can be treated as mixtures of ideal gases, they can be described by the ideal gas equation of state. Thus, the high temperature in the burned volume increases the pressure in the burned zone. This produces a pressure wave that propagates at the speed of sound, which is much faster than the turbulent flame speed. Therefore, the pressure increases throughout the combustion chamber essentially uniformly. This combustion-induced increase in pressure of the unburned "end gases" (the reactive mixture ahead of the flame front), together with the increase in pressure resulting from the upward motion of the piston, increases the temperature of the end gases. Because the rate of chemical reaction of the end gases generally increases strongly with increasing temperature, if the ignition delay time of the end gases (the time required for the reactions to reach an energy liberating phase, which is a function of the fuel type, the mixture composition, and the temperature history of the end gases) is less than the time required for the flame to reach the farthest point in the combustion chamber (which is primarily a function of engine design and engine speed), autoignition will occur in the unburned gases. Owing to inhomogeneities, all of the end gas does not autoignite simultaneously, but nearly so. This results in more-or-less a volumetric release and an energy release rate that can be much higher than for the normal flame propagation process. If sufficient energy is released during the autoignition reactions, a strong acoustic wave is developed, which reverberates in the combustion chamber yielding the audible sound that is called *knock*. Because knock can rapidly damage an SI engine, avoidance of knock is one of the primary factors limiting the design of SI engines.

In short, knock is the result of a "horse race" between the residence time of the unburned mixture ahead of the flame and the autoignition delay time. Knock is most likely to occur for conditions that either promote the reactivity of the unburned mixture (i.e., decrease the autoignition delay time) or decrease the flame propagation rate (i.e., increase the end gas residence time).

It has been shown that knock in SI engines is a result of low-to-intermediate temperature reactions involving alkanes, especially long straight chain alkanes (Cernansky *et al*., 1986; Westbrook, Pitz, and Leppard, 1991; Westbrook and Pitz, 1991, 1993; Leppard, 1992; Chevalier *et al*., 1992; Curran *et al*., 1995, 1996; Roberts, Matthews, and Leppard, 1996), with other hydrocarbon species essentially "soaking up" energy and/or active radicals.

The same kinds of reactions and chemical species are responsible for the autoignition process in diesel engines. In the case of diesels, the autoignition process is a more readily apparent two-stage ignition process, as previously described, that results from the autoignition characteristics of diesel fuel—a low reactivity fuel compared to gasoline.

For LTC approaches, much if not all of the cylinder undergoes volumetric energy release. To successfully achieve HCCI one must get the fuel vaporized and partially premixed with air prior to the autoignition chemistry reaching the point of autoignition. Gasoline vaporizes easily but does not autoignite easily whereas diesel fuel vaporizes slowly but autoignites easily. Based on the very different autoignition and vaporization characteristics of gasoline and diesel fuel, two very different overall approaches to LTC have arisen and are often referred to as HCCI [or gasoline compression ignition (GCI)] and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

diesel LTC. Ideally, for GCI a premixed charge of gasoline and air is inducted into an engine with a sufficiently high compression ratio to allow the entire cylinder contents to autoignite in rapid succession. However, to prevent the rate of pressure rise from being too excessive (leading to engine damage, similar to engine knock), the mixture is rarely truly homogeneous as the name HCCI implies. Stratification of the fuel–air mixture and of the temperature can be used to help slow the rate of progression of the combustion and broaden the load range capability of HCCI. In addition, charge dilution with excess air or EGR is used to slow down and spread out the heat release rate. Development of practical means of controlling the heat release rates and time of autoignition are primary development challenges for HCCI engines.

In the case of diesel LTC, there are two typical pathways (e.g., Dec, 2009) used. One is to inject the diesel fuel very early to allow time for vaporization and mixing of this low volatility fuel, but the ignition delay—which is inherently short for high Cetane fuels—must be stretched out via the relatively low oxygen concentration associated with high EGR rates. The EGR must also be aggressively cooled to further increase the delay in autoignition. Decreased compression ratio and late intake valve closing (which decreases the effective compression ratio while maintaining the expansion ratio) have also been used to increase the autoignition delay. High emissions of unburned hydrocarbons due to spray overpenetration/wall wetting have been reported, and control of combustion phasing (e.g., the crank angle at which peak pressure occurs) as the engine speed and load change remains a challenge. The second pathway used to attain diesel LTC involves fuel injection near or after top-dead-center compression. This pathway to diesel LTC provides better control over combustion phasing but the autoignition delay must still be "stretched out" to allow sufficient time for fuel evaporation and mixing. This is again accomplished via high levels of EGR. However, in this case, some of the heat release at the end of combustion often still occurs via mixing controlled combustion. As a result, care must be taken to ensure that NO*x* and PM emissions are controlled to low levels without inducing excessive unburned hydrocarbon and CO emissions. These remain as research challenges.

In summary, LTC is, in essence, controlled knock (Foster, 2012). For all LTC approaches, some inhomogeneity or stratification is essential so that the entire mixture within the combustion chamber does not all autoignite simultaneously resulting in excessive heat release rates and, thereby, excessive pressure rise rates. Instead, conditions must be generated such that a few locations autoignite, releasing sufficient energy to induce other regions to autoignite, which then induce other regions to autoignite. Thus, rather than flame propagation, LTC seems to involve propagation of autoignition, but this is not fully understood as yet. The desired stratification may be due to inhomogeneity in mixture composition or temperature, which will control the degree to which the autoignition reactions progress locally toward the energy liberation phase. Control of ignition timing and heat release rates, extending the load range capability, and avoidance of conditions that generate high levels of various emissions remain challenges for achieving the brake-specific fuel consumption potential of LTC (Dec, 2009).

The interested reader is referred to Spark Ignition Combustion for further details on spark-ignition combustion and knock, to Diesel and Diesel LTC Combustion for details on diesel combustion and diesel LTC, and Advanced Compression-Ignition Combustion for Ultra-Low NO_x and Soot for details on advanced gasoline compression ignition forms of LTC (e.g., HCCI).

5 MIXED COMBUSTION MODES

Engines have mixed modes of combustion more often than not. A few brief examples may be illustrative. Combustion in a SI engine is a premixed flame propagation process in which preflame reactions will occur in the end gases as already discussed, possibly progressing to the energy-liberating stage called *knock* (volumetric energy release). In addition, as already discussed, under the right conditions the fuel/air mixture that is hidden from the combustion process in crevices and deposits can emerge after peak pressure during the expansion stroke where it can become involved in a diffusion oxidation process—related to but probably not a diffusion flame. This is because the chain of reactions will not have sufficient time to reach an energy-liberating stage before exhaust valve opening with the consequent sudden decrease in pressure and temperature that will freeze the reactions. Another mixed combustion mode that can occur in both port-injected and direct-injected gasoline engines is due to "wall films"—films of liquid gasoline on the cylinder walls, piston top, and/or cylinder head. The flame can pass over such wall films without consuming them because the rate of evaporation is controlled by the surface temperature rather than by the gas temperature. As the wall film evaporates, it can take part in a diffusion-controlled combustion process that resembles a pool fire. The pool fires lead to emissions of particulate matter, and evaporation of the remaining wall film too late in the cycle produces hydrocarbon emission problems (e.g., Stanglmaier, Li, and Matthews, 1999; Li *et al*., 1999, 2000, 2001; Huang, Matthews, and Ellzey, 2001a; Huang

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

et al., 2001b, Alger *et al*., 2001; Steeper and Stevens, 2000; Stevens and Steeper, 2001; Warey *et al*., 2002).

Combustion in diesel engines occurs in multiple phases after autoignition. The first is dominated by a rapid volumetric energy release referred to as the *premixed combustion phase*. This is followed next by a quasi-steady very rich premixed flame embedded within a turbulent diffusion flame, which in turn is located in the vicinity of the mean stoichiometric mixture region zone surrounding the fuel jet for the remainder of the injection event (see Diesel and Diesel LTC Combustion and Dec 1997). After the end of fuel injection, mixing controlled combustion continues as pockets of fuel and air mix and burnout in the final stages of diesel combustion. If fuel during injection impinges and collects on walls, pool fires similar to those already discussed for gasoline engines can occur.

The diesel-fueled LTC combustion process briefly introduced in the previous section is dominated by a staged volumetric energy release process followed, in the last stage of combustion, by mixing controlled combustion of rich regions. Another recent approach to LTC is reactivity-controlled compression ignition (RCCI), in which a low reactivity fuel, such as gasoline, is port-injected to yield a nearly homogeneous charge while a more reactive fuel, such as diesel, is direct-injected to provide control over the combustion "phasing" (e.g., Inagaki *et al*., 2006; Kokjohn *et al*., 2009). In RCCI, the ignition process, which is controlled by the more reactive fuel, starts a volumetric energy release process, which is followed by a currently unknown but likely combination of volumetric energy release and premixed flame propagation through the mixture of the less reactive fuel with air (Kokjohn *et al*., 2011).

REFERENCES

Abdel-Gayed, R.G. and Bradley, D. (1985) Criteria for turbulent propagation limits of premixed flames. *Combustion and Flame*, **62** (1), 61–68.

Abdel-Gayed, R.G., Bradley, D., and Lau, K.C. (1988) *The Straining of Premixed Turbulent Flames*. Twenty-Second Symposium (International) on Combustion, The Combustion Institute, Pittsburgh, pp. 731–738.

Abdel-Gayed, R.G., Bradley, D., and Lung, F.K.-K. (1989) Combustion regimes and the straining of turbulent premixed flames. *Combustion and Flame*, **76**, 213–218.

Abraham, J., Williams, F.A., and Bracco, F.V. (1985) A discussion of turbulent flame structure in premixed charges. SAE Paper 850345.

Alger, T., Huang, Y., Hall, M.J., and Matthews, R.D. (2001) Liquid film evaporation off the piston of a direct injection gasoline engine. SAE Paper 2001-01-1204; also in *Journal of Engines*, **110** (3), 1295–1306.

Andrews, G.E. and Bradley, D. (1972) The burning velocity of methane-air mixtures. *Combustion and Flame*, **19**, 275–288.

Ballal, D.R. and Lefebvre, A.H. (1974) *Influence of flow parameters on minimum ignition energy and quenching distance*. 15th Symposium (International) on Combustion. The Combustion Institute, Pittsburgh, pp. 1473–1481.

Borghi, R. (1985) *Recent Advances in the Aerospace Sciences*. Plenum Press, New York.

Boudier, P., Henriot, S., Poinsot, T., and Baritaud, T. (1993) *A model for turbulent flame ignition and propagation in spark ignition engines*. 24th Symposium (International) on Combustion. The Combustion Institute, Pittsburgh, pp. 503–510.

Cernansky, N.P., Green, R.M., Pitz, W.J., and Westbrook, C.K. (1986) Chemistry of fuel oxidation preceding end-gas autoignition. *Combustion Science and Technology*, **50**, 3–25.

Cheng, W.K. and Diringer, J.A. (1991) Numerical modeling of SI engine combustion with a flame sheet model. SAE Paper 910268.

Chevalier, C., Pitz, W.J., Westbrook, C.K., and Melenk, H. (1992) *Hydrocarbon ignition: automatic generation of reaction mechanisms and applications to modeling engine knock*. Twenty-Fourth Symposium (International) on Combustion. The Combustion Institute, Pittsburgh, pp. 93–101.

Chin, Y.W., Matthews, R.D., Nichols, S.P., and Kiehne, T.M. (1992) Use of fractal geometry to model turbulent combustion in SI engines. *Combustion Science and Technology*, **8** (1–6), 1–30.

Chung, S.H. and Law, C.K. (1988) An integral analysis of the structure and propagation of stretched premixed flames. *Combustion and Flame*, **72**, 325–336.

Curran, H.J., Gaffuri, P., Pitz, W.J., *et al.* (1995) Autoignition chemistry of the hexane isomers: an experimental and kinetic modeling study. SAE Paper 952406.

Curran, H.J., Gaffuri, P., Pitz, W.J., *et al.* (1996) *Autoignition chemistry in a motored engine: an experimental and kinetic modeling study*. 26th Symposium (International) on Combustion. The Combustion Institute, Pittsburgh, pp. 2669–2678.

Dai, W., Russ, S.G., Trigui, N., and Tallio, K.V. (1998) Regimes of turbulent combustion and misfire modeling in SI engines. SAE Paper 982611; also in *SAE International Journal of Fuels and Lubricants*, **107** (2), 1738–1747.

Damkohler, G. (1947) The effect of turbulence on the flame velocity in gas mixtures. NACA Tech. Memo No. 112.

Dec, J.E. (1997) A conceptual model of DI diesel combustion based on laser-sheet imaging. SAE Paper 970873; also in *SAE International Journal of Engines*, **106** (3), 1319–1348.

Dec, J.E. (2009) Advanced compression-ignition engines—understanding the in-cylinder processes. *Proceedings of the Combustion Institute*, **32**, 2727–2742.

Duclos, J.M., Veynante, D., and Poinsot, T. (1993) A comparison of flamelet models for premixed turbulent combustion. *Combustion and Flame*, **95**, 101–117.

Foster, D.E. (2012) Low temperature combustion—a thermodynamic pathway to high efficiency engines. Prepared for the National Petroleum Council Fuels Study.

Herweg, R. and Maly, R.R. (1992) A fundamental model for flame kernel formation in S.I. engines. SAE Paper 922243.

Huang, Y., Matthews, R.D., and Ellzey, J.E. (2001a) The effects of fuel volatility and structure on HC emissions from piston wetting

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

in DISI engines. SAE Paper 2001-01-1205; also in *Journal of Fuels Lubricants*, **110** (4), 912–929.

Huang, Y., Matthews, R.D., Ellzey, J.E., and Dai, W. (2001b) The effects of fuel volatility, load, and speed on HC emissions due to piston wetting. SAE Paper 2001-01-2024; also in *Journal of Engines*, **110** (3), 1878–1889.

Inagaki, K., Fuyuto, T., Nishikawa, K., *et al.* (2006) Dual-fuel PCI combustion controlled by in-cylinder stratification of ignitability. SAE Paper 2006-01-0028.

Kee, R.J., Grcar, J.F., Smooke, M.D., and Miller, J.A. (1985) A Fortran program for modeling steady laminar one-dimensional premixed flames. Sandia National Laboratories Report SAND85-8240.

Kee, R.J., Warnatz, J., and Miller, J.A. (1983) A FORTRAN computer code package for the evaluation of gas-phase viscosities, conductivities, and diffusion coefficients. Sandia National Laboratories Report SAND83-8209.

Klimov, A.M. (1963) Laminar flame in a turbulent flow. *Prikladnaya Mekhanika I Tekhnicheskaya Fizika*, **3**, 49–58.

Klimov, A.M. (1975) Flame propagation under conditions of strong turbulence. *Doklady Akademii Nauk SSSR*, **221**, 56–59.

Klimov, A.M. (1983) Premixed turbulent flames—interplay of hydrodynamic and chemical phenomena, in *Flames, Lasers, and Reactive Systems* (eds J.R. Bowen, N. Manson, A.M. Oppenheim and R.I. Soloukhin), vol. 88, *Progress in Astronautics and Aeronautics*, Institute of Aeronautics and Astronautics, New York, pp. 133–146.

Kokjohn, S.L., Hanson, R.M., Splitter, D.A., and Reitz, R.D. (2009) Experiments and modeling of dual-fuel HCCI and PCCI combustion in a heavy-duty engine. SAE Paper 2009-01-2647; also in *SAE International Journal of Engines*, **2** (2), 24–39, 2010.

Kokjohn, S., Reitz, R.D., Splitter, D., and Musculus, M. (2011) Investigation of fuel reactivity stratification for controlling PCI heat-release rates using high-speed chemiluminescence imaging and fuel tracer fluorescence. SAE Paper 2012-01-0375; also in *SAE International Journal of Engines*, **5** (248–269), 2012.

Kuo, K.K. (1986) Principles of Combustion, John Wiley & Sons, New York.

Lancaster, D.R., Kreiger, R.B., Sorenson, S.C., and Hull, W.B. (1976) Effects of turbulence on spark-ignition engine combustion. SAE Paper 760160; also in *SAE Transactions*, **85**, 689–710.

Law, C.K. (1988) *Dynamics of stretched flames*. 22nd Symposium (International) on Combustion. The Combustion Institute, Pittsburgh, pp. 1381–1402.

Leppard, W.R. (1992) The autoignition chemistries of primary reference fuels, olefin/paraffin binary mixtures, and non-linear octane blending. SAE Paper 922325.

Lefebvre, A.H. (1983) *Gas Turbine Combustion*, McGraw-Hill, New York.

Li, J.W., Matthews, R.D., Stanglmaier, R.H., *et al.* (1999) Further experiments on in-cylinder wall wetting in direct injected gasoline engines. SAE Paper 1999-01-3661; also in *Journal of Fuels and Lubricants*, **108** (4), 2213–2224.

Li, J.W., Huang, Y., Alger, T.F., *et al.* (2000) Liquid fuel impingement on in-cylinder surfaces as a source of hydrocarbon emissions from direct injection gasoline engines, ASME Paper 2000-ICE-270, in *Fuel Injection, Combustion, and Engine Emissions*, ICE vol. 34–2, pp. 17–26.

Li, J.W., Huang, Y., Alger, T.F., *et al*. (2001) Liquid fuel impingement on in-cylinder surfaces as a source of hydrocarbon emissions from direct injection gasoline engines. *Journal of Gas Turbines and Power*, **123**, 659–668.

Mattavi, J.N., Groff, E.G., and Matekunas, F.V. (1979) Turbulence, Flame Motion, and Combustion Chamber Geometry—Their Interactions in a Lean Combustion Engine. *Proceedings of the IMechE Conference on Fuel Economy and Emissions of Lean Burn Engines*. Paper C100/79.

Matthews, R.D. and Chin, Y.-W. (1991) A fractal-based SI engine model: comparisons of predictions with experimental data. SAE Paper 910079; also in *Journal of Engines*, **100**, 99–117.

Matthews, R.D., Hall, M.J., Dai, W.-G., and Davis, G.C. (1996), Combustion modeling in SI engines with a peninsula-fractal combustion model. SAE Paper 960072.

Mattingly, J.D. (1996) *Elements of Gas Turbine Propulsion*, McGraw-Hill, New York.

Metghalchi, M. and Keck, J.C. (1980) Laminar burning velocity of propane-air mixtures at high temperature and pressure. *Combustion and Flame*, **38**, 143–154.

Metghalchi, M. and Keck, J.C. (1982) Burning velocities of mixtures of air with methanol, iso-octane, and Indolene at high pressure and temperature. *Combustion and Flame*, **48**, 191–210.

Milton, B.E. and Keck, J.C. (1984) Laminar burning velocities in stoichiometric hydrogen and hydrogen-hydrocarbon gas mixtures. *Combustion and Flame*, **58**, 13–22.

Peters, N. (1986) *Laminar flamelet concepts in turbulent combustion*, Invited Lecture, 21st Symposium (International) on Combustion. The Combustion Institute, Pittsburgh, pp. 1231–1250.

Poinsot, T., Veynante, D., and Candel, S. (1991) Quenching processes and premixed turbulent combustion diagrams. *Journal of Fluid Mechanics*, **228**, 561–606.

Rhodes, D.B., and Keck, J.C. (1985) Laminar burning speed measurements of Indolene-air-diluent mixtures at high temperature and pressure. SAE Paper 850047.

Roberts, W.L., Driscoll, J.F., Drake, M.C., and Goss, L.P. (1993) Images of the quenching of a flame by a vortex - to quantify regimes of turbulent combustion. *Combustion and Flame*, **94** (1/2), 58–69.

Roberts, C.E., Matthews, R.D., and Leppard, W.R. (1996) Development of a semi-detailed kinetics mechanism for the autoignition of iso-octane. SAE Paper 962107; also in *SAE International Journal of Fuels and Lubricants*, **105** (2), 2238–2266.

Stanglmaier, R.H., Li, J.W., and Matthews, R.D. (1999) The effect of in-cylinder wall wetting on HC emissions from SI engines. SAE Paper 1999-01-0502; also in *SAE International Journal of Engines*, **108** (3), 533–542.

Steeper, R.R., and Stevens, E.J. (2000) Characterization of combustion, piston temperatures, fuel sprays, and fuel-air mixing in a DISI optical engine. SAE Paper 2000-01-2900.

Stevens, E.J. and Steeper, R.R. (2001) Piston wetting in an optical DISI engine: fuel films, pool fires, and soot generation. SAE Paper 2001-01-1203.

Swithenbank, J., Poli, I., and Vincent, M.W. (1973) *Combustion design fundamentals*. Fourteenth Symposium (International) on Combustion. The Combustion Institute, Pittsburgh, pp. 627–638.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Treager, I.E. (1979) *Aircraft Gas Turbine Technology*, McGraw-Hill, New York.

Turns, S.R. (2000) *An Introduction to Combustion: Concepts and Applications*, McGraw-Hill, New York.

Warey, A., Huang, Y., Matthews, R.D., *et al*. (2002) Effects of piston wetting on size and mass of particulate matter emissions in a DISI engine. SAE Paper 2002-01-1140; also in *SAE International Journal of Engines*, **111** (3), 1977–1984.

Westbrook, C.K. and Dryer, F.L. (1984) Chemical kinetic modeling of hydrocarbon combustion. *Progress in Energy and Combustion Science*, **10**, 1–58.

Westbrook, C.K. and Pitz, W.J. (1991) The chemical kinetics of engine knock. Lawrence Livermore National Labs E&TR, February–March, pp. 1–13.

Westbrook, C.K. and Pitz, W.J. (1993) A chemical kinetic mechanism for the oxidation of paraffinic hydrocarbons needed for Primary Reference Fuels. Paper UCRL-JC-112696, presented at the Spring Meeting of the Western States Section of the Combustion Institute, Salt Lake City, Utah.

Westbrook, C.K., Pitz, W.J., and Leppard, W.R. (1991) The autoignition chemistry of paraffinic fuels and pro-knock and anti-knock additives: a detailed chemical kinetic study. SAE Paper 912314.

Williams, F.A. (1976) Criteria for existence of wrinkled laminar flame structure of turbulent premixed flames. *Combustion and Flame*, **26**, 269–270.

Williams, F.A. (1984) Asymptotic methods in turbulent combustion. Paper No. 84–0475, AIAA 22nd Aerospace Sciences Meeting, January.

Williams, F.A. (1985) Turbulent combustion, in *Mathematics of Combustion* (ed J. Buckmaster), pp. 97–131.

Wu, C.-M., Roberts, C. Matthews, R.D., and Hall, M.J. (1993) Effects of engine speed on combustion in SI engines: comparisons of predictions of a fractal burning model with experimental data. SAE Paper 932714; also in *Journal of Engines*, **102**(3), 2277–2291.

Zhao, X.-W., Matthews, R.D., and Ellzey, J.L. (1993) Three-dimensional numerical simulation of flame propagation in spark ignition engines. SAE Paper 932713.

Zhao, X., Matthews, R.D., and Ellzey, J.L. (1994) Numerical Simulations of Combustion in SI Engines Comparison of the Fractal Flame Model to the Coherent Flame Model. *Proceedings of the Third International Symposium on Diagnostics and Modeling of Combustion in Internal Combustion Engines*, JSME/JSAE, pp. 157–162.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto115
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Solving Combustion Chemistry in Engine Simulations

Norbert Peters and Heinz Pitsch
RWTH Aachen University, Aachen, Germany

1 Introduction 1
2 One-Step Global Kinetics: Auto-Ignition 2
3 Detailed Chemical Kinetics, their Reduction and Tabulation 4
4 Chemical Effects in SI Engines 5
5 Chemical Effects in CI Engines 6
6 Summary 8
References 8

1 INTRODUCTION

Combustion in engines depends critically on the appropriate choice of the fuel to be used—its chemical and physical properties as well as its availability, sustainability, and cost. Under these limiting conditions, liquid hydrocarbons derived from fossil fuels will continue to play a major role for the foreseeable future. However, readily available biofuels such as ethanol must also be considered as an alternative. Especially, blends of different fuels will gain importance for fuel-flexible engine operation in the future. The tendency to auto-ignite is an inherent property of all long-chain hydrocarbons or oxygenated fuels under high pressure and temperature conditions, as they occur in engines. When it occurs as an undesired event in the end gas of spark-ignition (SI) engines, it may lead to engine knock and may damage the engine. In diesel (or compression-ignition, CI) engines, auto-ignition is desired, as it occurs at a well-controlled time delay after injection to initiate the combustion process. SI engines, in order to avoid engine knock, are operated with high octane number (ON) fuels. The ON (defined as the percentage of isooctane in a mixture of isooctane and *n*-heptane that has in a reference experiment the same auto-ignition characteristics as the considered fuel) characterizes the auto-ignition behavior of gasoline fuels, such that low ON fuels are readily ignitable and high ON fuels are difficult to ignite. For CI engines, the desired auto-ignition properties are quite different. Ignition properties of diesel fuels are characterized by the cetane number (CN). A high CN (defined as the percentage of normal cetane in a mixture of normal cetane and isocetane that has in a reference experiment the same auto-ignition characteristics as the considered fuel) stands for a fuel that is easily ignitable (short ignition delay time), whereas a low CN characterizes a fuel with a long ignition delay time. *N*-Heptane, for instance, has a CN of 56, which is close to that of diesel fuel.

As it is impossible to represent the chemical composition of commercial fuels in their full complexity, surrogate fuels containing only a few chemical components are often used in numerical simulations. These components are chosen by their ability to reproduce the performance of the commercial fuel with respect to ignition, combustion, and pollutant formation, provided that a detailed chemical kinetics are available for them. A simple example for SI engines is primary reference fuels (PRF), which are a binary blend of *n*-heptane and isooctane. On the basis of such a surrogate fuel, for example, the laminar burning velocity—needed for the prediction of turbulent combustion in gasoline engines—can be calculated. Another example to be shown later is the use of *n*-heptane as a diesel surrogate for

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto116
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

the prediction of auto-ignition, combustion, and pollutant formation in a CI engine. The short ignition delay time of *n*-heptane is due to a particular group of reactions active at low temperatures. Those reactions will be discussed in detail later.

The chapter is outlined as follows: at first, one-step global kinetics are introduced and the resulting ignition delay time and thermal runaway is illustrated. This serves as a basis for a critical discussion of the Livengood–Wu integral, which is often used for the prediction of auto-ignition in engines. Then, as an example for elementary kinetics, the low temperature part of an *n*-heptane mechanism is presented. It is reduced to a four-step mechanism, from which the low temperature ignition delay time can be calculated analytically.

Chemical effects that appear in models for SI-engines are discussed next. While detailed chemistry has been used in models for the sparking process only recently, it enters into combustion models for turbulent flame propagation, such as the extended coherent flame model (ECFM) and the G-equation model, through the way it determines the laminar burning velocity. Chemical effects are also responsible for engine knock and preignition events leading to "super-knock" in super-charged SI engines.

In diesel or CI engines, chemistry plays a crucial role at all stages: auto-ignition at the premixed stage, combustion at the nonpremixed, diffusion-controlled stage, and finally, for pollutant formation. As an example of combustion and pollutant formation in a CI engine, an early application of the flamelet model is presented.

2 ONE-STEP GLOBAL KINETICS: AUTO-IGNITION

As mentioned in the previous section, auto-ignition is a desired and necessary step in CI engines, where it initiates combustion after the fuel has been injected into the compressed high temperature air (with or without exhaust gas recirculation, EGR), and it is an undesired side effect in SI engines, where it may occur in the yet unreacted mixture ahead of the propagating premixed flame front potentially leading to engine knock. Auto-ignition before spark timing is called *preignition* and may even lead to the so-called super-knock events that may damage the engine and must be avoided.

The chemical process of auto-ignition in engines may best be understood by considering a homogeneous charge in an adiabatic system at variable volume. Using a one-step reaction

$$\text{Fuel} + \text{Oxygen} \rightarrow \text{Products} \qquad (1)$$

one may assume a first-order reaction with respect to the fuel to one obtain the reaction rate

$$\omega = B\left(\frac{\rho Y_{\mathrm{F}}}{W_{\mathrm{F}}}\right)\left(\frac{\rho Y_{\mathrm{O_2}}}{W_{\mathrm{O_2}}}\right)\exp\left(-\frac{E}{\mathcal{R}T}\right) \qquad (2)$$

Here, B is the frequency factor, ρ the density, and Y_i and W_i the mass fractions and the molecular weights of fuel and oxygen, respectively. The temperature (in degrees Kelvin) is denoted by T, $\mathcal{R}$ is the universal gas constant, and E the activation energy. Assuming this rate expression, the frequency factor B and the activation energy E are adjustable constants for a given fuel. For instance, using auto-ignition measurements in a counterflow configuration over a liquid surface, Seshadri, Humer, and Seiser, 2008 have deduced activation temperatures $T_{\mathrm{a}} = E/\mathcal{R}$ for pure fuels, but also for jet and diesel fuels (Table 1). As auto-ignition under these conditions is governed by high temperature kinetics, these numbers are valid only if the initial temperature exceeds a value of typically 1000 K.

For illustration, we will first discuss the case of homogeneous combustion in a constant volume combustion chamber assuming adiabatic conditions, for simplicity. For the case of a homogeneous mixture having at $t = 0$ the temperature T_0, the temperature will increase at first slowly, then rapidly, if the assumed one-step reaction is exothermic. This is called *thermal runaway*. As the activation temperature is much larger than the initial temperature T_0, the exponential term is dominating in the rate expression in Equation 2 showing that the rate is very sensitive to temperature changes. To simplify the analysis, one may, therefore, only retain the effect of temperature changes and neglect the consumption of the reactants during the time of thermal

Table 1. Activation temperatures and frequency factors obtained from auto-ignition measurements in stagnation point flows.

Fuel	Chemical symbol	$T_{\mathrm{a}}[K]$	$B\,[\mathrm{m^3/(mol \cdot s)}]$
n-Heptane	C_7H_{16}	25 800	4.65×10^{10}
n-Octane	C_8H_{18}	19 500	1.66×10^{8}
n-Decane	$C_{10}H_{22}$	15 300	3.08×10^{6}
n-Dodecane	$C_{12}H_{26}$	15 200	2.68×10^{6}
n-Hexadecane	$C_{16}H_{36}$	14 800	1.77×10^{6}
isooctane	C_8H_{18}	30 600	2.18×10^{12}
Cyclohexane	C_6H_{12}	28 500	7.48×10^{11}
Methylcyclohexane	C_7H_{14}	29 000	5.79×10^{11}
o-Xylene	C_8H_{10}	28 000	8.95×10^{10}
JP-10	$C_{10}H_{16}$	26 000	4.63×10^{10}
JP-08	$C_{11}H_{21}$	29 600	6.08×10^{11}
Diesel	$C_{14.7}H_{26.8}$	25 000	6.57×10^{9}

Reproduced from Seshadri, Humer, and Seiser, 2008. © Taylor & Francis.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto116
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

runaway. For a variable volume chamber, setting the mass fractions of the fuel and the oxidizer equal to their initial values $Y_{F,0}$ and $Y_{O_2,0}$, one obtains the temperature equation as

$$\rho\, c_v \frac{dT}{dt} = \frac{p}{V}\frac{dV}{dt} + Q_v B \frac{\rho Y_{F,0}}{W_F}\frac{\rho Y_{O_2,0}}{W_{O_2}} \exp\left(-\frac{E}{\mathcal{R}T}\right) \quad (3)$$

where p is the pressure. The volume of the vessel changes with time as $V(t)$. The heat capacity c_v and the heat of reaction Q_v at constant volume are assumed constant and will be evaluated at the initial condition $T = T_0$. In the case of a constant volume explosion, the density is also constant and the first term on the r.h.s. vanishes. This case will be considered first.

Using the method of large activation energy asymptotics (Peters, 1992), p. 55), Equation 3 can be solved analytically for the case $dV/dt = 0$ to obtain the temperature as a function of time

$$T = T_0 - \frac{\mathcal{R}T_0^2}{E}\ln\left(1 - \frac{t}{t_I}\right) \quad (4)$$

This equation shows that T tends to infinity when t approaches t_I, thus illustrating the thermal runaway. The ignition delay time t_I, defined as the time at which thermal runaway occurs in the adiabatic constant volume vessel, is then determined as

$$t_I = \frac{\mathcal{R}T_0^2}{E}\frac{c_v}{\rho_0 Q_v B}\frac{W_F W_{O_2}}{Y_{F,0}Y_{O_2,0}} \exp\left(\frac{E}{\mathcal{R}T_0}\right) \quad (5)$$

It is noteworthy that in a plot of the logarithm of t_I over $1/T_0$, the so-called Arrhenius diagram, one obtains a nearly straight line with a positive slope equal to the activation temperature $T_a = E/\mathcal{R}$. Values for T_a and frequency factors B are given in Table 1 for different fuels.

During the compression stroke in an engine, the background temperature changes continuously. Thermal runaway is then an effect superimposed on this varying background temperature. A popular procedure to estimate the ignition time t_{ign} resulting from both the background temperature change and the thermal runaway is the use of the Livengood–Wu integral (Livengood and Wu, 1955) defined by

$$\int_0^{t_{ign}} \frac{1}{\tau(T(t), p(t))}\, dt = 1 \quad (6)$$

where $\tau(T(t), p(t))$ is the homogeneous ignition delay time evaluated at temperature $T(t)$ and pressure $p(t)$, and $T(t)$ and $p(t)$ are the in-cylinder temperature and pressure at time t considering the background compression. The auto-ignition delay time at different initial temperatures can be evaluated from detailed chemical kinetics, but will in the following, for illustration purposes, be taken from the analytic solution, Equation 5.

A correct prediction of homogeneous auto-ignition with changes of the background temperature because of compression would have to include the unsteady change of volume dV/dt in the temperature Equation 3. To illustrate this, we have performed such a simulation using a simple kinematic piston displacement model for a supercharged SI engine with a compression ratio of 12, a stroke of 94.6 mm, a bore of 86 mm, and a connecting rod of 152.2 mm. The intake temperature was 107 °C and the intake pressure was 2 bar. Global low temperature kinetics for *n*-heptane were used with $B = 1.86 \cdot 10^{11}$ m^3/(mol s), $E/\mathcal{R} = 22\,732$ K, and

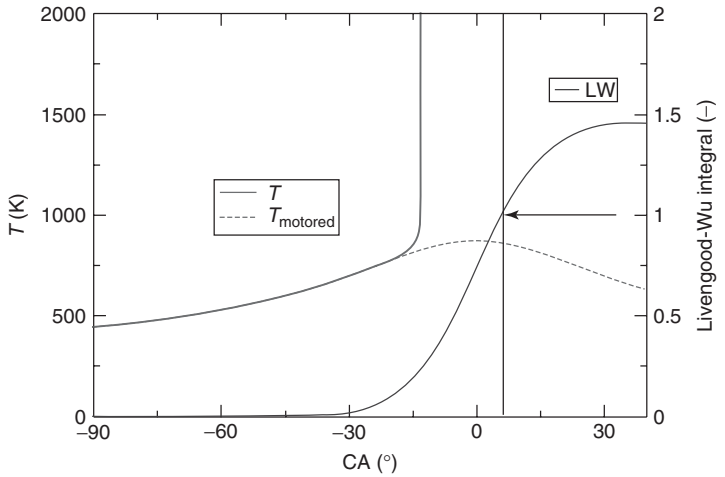


Figure 1. Illustration of auto-ignition calculated from Equation 3 and from the Livengood–Wu integral and the solution of Equation 3.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto116
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

$Q_v/c_v = 4824$ K kg/kmol. The thermal equation of state with the molecular weight of air was used.

The temperature of the simulation is shown in Figure 1 for motored conditions and for the case of homogeneous auto-ignition calculated numerically from Equation 3 at 3000 rpm. As the source term in Equation 3 is not limited at equilibrium conditions, the temperature increases without limit, just as in Equation 4. In addition, the Livengood–Wu integral evaluated with t_I from Equation 5 is shown, thereby considering only the background temperature changes. The time (in terms of deg. CA) when it crosses unity corresponds then to this prediction of auto-ignition. There is a noticeable difference between the two calculations illustrating the limitations in using the Livengood–Wu integral without considering thermal runaway.

3 DETAILED CHEMICAL KINETICS, THEIR REDUCTION AND TABULATION

Detailed chemical mechanisms for transportation fuel surrogates have been developed in the past. However, these mechanisms often consist of many thousands of elementary reactions and species, and are, therefore, impractical for engine simulations. Starting from an early account (Peters and Rogg, 1993) on how to reduce detailed chemical mechanisms in combustion by steady-state analysis, many groups have proposed reduction procedures (Pepiot-Desjardins and Pitsch, 2008; Lu and Law, 2006; Zeuch *et al.*, 2008; Løvås, 2009; Sun *et al.*, 2010; Lu, Ju, and Law, 2001; Lu and Law, 2006). As even reduced mechanisms often still contain too many chemical species and reactions, a further step to reduce the computational effort is to tabulate the resulting chemical source terms. An algorithm called in situ adaptive tabulation (ISAT) was developed in the context of transported probability density function (PDF) methods (Pope, 1997). As those methods are in general too expensive to be used in engine calculations, more specific tabulations have been proposed (Jay and Colin, 2011; Mittal and Pitsch, 2013).

For the example of *n*-heptane (C_7H_{16}, denoted as F), which often serves as a surrogate fuel for auto-ignition and combustion in diesel engines, we show in Table 2, taken from Ref. (Peters *et al.*, 2002), those reactions that are rate determining for the low temperature and the high temperature branch of auto-ignition. Calculated ignition delay times are shown in Figure 2 for several pressures in an Arrhenius diagram. The low temperature branch (on the right) and the high temperature branch (on the left) have a positive slope in this diagram. The intermediate branch is the so-called negative temperature coefficient (NTC) branch typical for diesel fuels. In an engine, the background pressure changes because of compression, as low temperature auto-ignition chemistry is occurring. This will impact the prediction of the auto-ignition time as discussed earlier. However, the general principles of mechanism reduction, as demonstrated in the following, will remain valid.

Assuming steady state (thus, canceling the species when adding two reactions) of the heptyl radical C_7H_{15} (denoted as R), we obtain by adding reaction 1 to reaction 4

$$\text{(I)}\quad \mathrm{F} + 2\mathrm{O_2} = \mathrm{RO_2} + \mathrm{HO_2} \tag{7}$$

where the heptyl peroxide radical $C_7H_{15}O_2$ is denoted by RO_2. Similarly, adding reaction 3 to reaction 4, we obtain

$$\text{(II)}\quad \mathrm{F} + \mathrm{OH} + \mathrm{O_2} = \mathrm{RO_2} + \mathrm{H_2O} \tag{8}$$

Table 2. The most important reactions for low temperature auto-ignition of *n*-heptane (C_7H_{16}).

Number	Reaction						
1	C_7H_{16}	+	O_2	→	C_7H_{15}	+	HO_2
2	C_7H_{16}	+	HO_2	→	C_7H_{15}	+	H_2O_2
3	C_7H_{16}	+	OH	→	C_7H_{15}	+	H_2O
4	C_7H_{15}	+	O_2	→	$C_7H_{15}O_2$		
5	$C_7H_{15}O_2$			→	$C_7H_{14}O_2H$		
6	$C_7H_{14}O_2H$	+	O_2	→	$O_2C_7H_{14}O_2H$		
7	$O_2C_7H_{14}O_2H$			→	$HO_2C_7H_{13}O_2H$		
8	$HO_2C_7H_{13}O_2H$			→	$OC_7H_{13}O_2H$	+	OH
9	$OC_7H_{13}O_2H$			→	$OC_7H_{13}O$	+	OH
10	$OC_7H_{13}O$			→	CH_2O	+	$C_5H_{11} + CO$
11	C_5H_{11}			→	C_2H_4	+	C_3H_7
12	C_3H_7	+	O_2	→	C_3H_6	+	HO_2
13	$2HO_2$			→	H_2O_2	+	O_2
14	H_2O_2	+	(M)	→	2OH	+	(M)

Reproduced from Peters *et al.*, 2002. © Elsevier.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto116
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

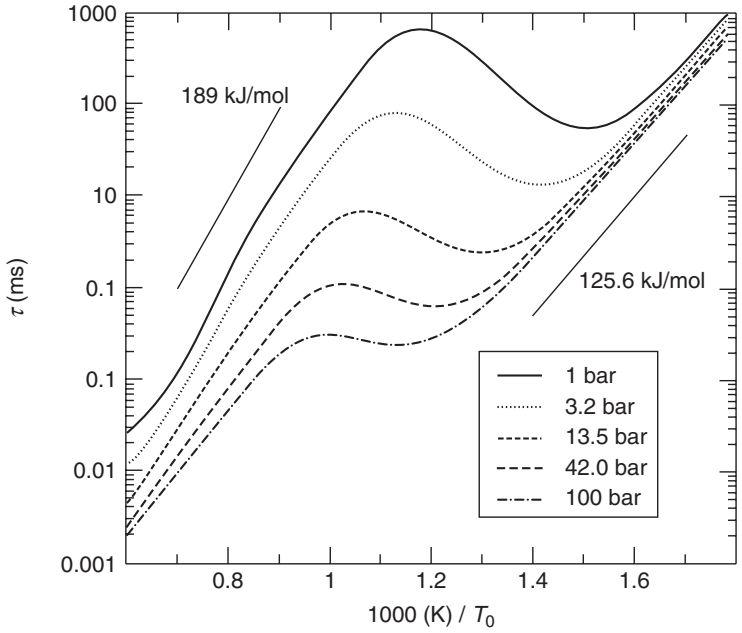


Figure 2. Ignition delay time of *n*-heptane as a function of the initial temperature T_0.

If we, furthermore, assume steady state for all intermediates up to ketohydroperoxide $OC_7H_{13}O_2H$ (denoted as KET), we obtain by adding reactions 5–8

$$\text{(III)}\quad RO_2 + O_2 = \text{KET} + OH \tag{9}$$

Finally, by adding reactions 9–12, we obtain by the same procedure

$$\text{(IV)}\quad \text{KET} + O_2 = P + OH + HO_2 \tag{10}$$

where P is the sum of the stable species CH_2O, CO, C_2H_4, and C_3H_6.

The four-step global mechanism I–IV illustrates the low temperature chain branching mechanism: by oxygen addition to the fuel, two OH radicals are formed in reactions III and IV, whereas only one OH radical is consumed in reaction II. The mechanism may be used to determine the low temperature ignition delay time, t_1, analytically (Peters *et al*., 2002)

$$t_1 = \frac{\ln\,(Y_F/\varepsilon)}{(k_5(T)k_9(T))^{1/2}} \tag{11}$$

Here, Y_F is the mass fraction of the fuel and $k_5(T)$ and $k_9(T)$ the reaction rate coefficients (in s^{-1}) of reactions 5 and 9 as a function of temperature, respectively, and ε is proportional to the ratio of the reaction rate of reaction 1 to the rate coefficient of reaction 9

$$\varepsilon = \frac{k_1(T)Y_{O_2}}{k_9(T)} \tag{12}$$

In a similar way, the ignition delay time of the high temperature branch shown in Figure 2 may be determined analytically. Here, only the reactions 2, 13, and 14 are rate determining, as shown in Peters *et al*., (2002). From these analytic solutions, the activation energies of the two branches are determined and are shown in Figure 2. The corresponding activation energies 125.6 and 189 kJ/kmol may be converted into activation temperatures as 15 179 and 22 732 K. The latter is to be compared to the value of 25 800 K (valid for the high temperature branch) given in Table 1.

4 CHEMICAL EFFECTS IN SI ENGINES

As a simple model for SI, heat deposition by the electrical spark may be modeled as a localized region of very high temperature. As the heat diffuses out, it forces the formation of radicals, which then initiate a self-sustained laminar, later turbulent premixed flame. Early examples for modeling SI are the Arc-and-Kernel tracking ignition model (AKTIM) (Duclos and Colin, 2001; Colin, Benkenida, and Angelberger, 2003) and the discrete particle ignition kernel (DPIK) model (Tan and Reitz, 2006). Chemistry enters into

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto116
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

these models through an expression for the laminar burning velocity.

A more detailed model that uses elementary chemistry is the spark channel ignition monitoring (Spark CIMM) model (Dahms *et al*., 2009). Here, the stratified mixture close to the spray in a direct-injection gasoline engine is monitored during the spark duration by tracing ignition kernels arising from the spark using fictitious particles that can lead to the formation of quasi-laminar flame kernels. Auto-ignition is modeled by adding a spark energy deposition source term to the unsteady flamelet equations, which consider the detailed ignition and combustion chemistry, as will be discussed later.

The early flame kernel development occurring after SI is typically modeled by spherical flame propagation. The turbulent flame that develops thereafter can be modeled, for instance, by the G-equation model (Peters, 1999; Dekana and Peters, 1999; Tan and Reitz, 2003; Tan, Kong, and Reitz, 2003) or the ECFM (Colin, Benkenida, and Angelberger, 2003; Candel and Poinsot, 1990; Duclos and Zolver, 1998). From the G-equation model, the turbulent burning velocity s_T was derived by Peters (1999) and can be written as Peters (2000)

$$s_T = s_L + v' \cdot [-\alpha\,\mathrm{Da} + \sqrt{(\alpha\,\mathrm{Da})^2 + 4\alpha\,\mathrm{Da}}\,] \qquad (13)$$

where $\alpha = 0.195$ and $\mathrm{Da} = s_L l/(v' l_F)$. This equation shows that in addition to the turbulence intensity v' and the integral length scale l, also the laminar burning velocity s_L and the flame thickness l_F enter into the expression. Therefore, approximations or tabulations of the laminar burning velocity and of the laminar flame thickness are needed as a function of initial temperature, pressure, and equivalence ratio. They can be constructed for gasoline fuels fairly easily on the basis of elementary reaction mechanisms, which are available for surrogate fuels.

A more controversial subject in SI-engine simulations is the prediction of engine knock. While Colin, Benkenida, and Angelberger (2003) considered the time evolution of a fictitious species with a source term based on a correlation with the ignition delay time τ_I, Bradley and Kalghatgi (2009), Kalghatgi and Bradley (2012), and Peters, Kerschgens, and Paczko (2013) based their prediction on the Livengood–Wu integral (1955) discussed earlier. Here, approximations of ignition delay times of surrogate fuels for gasoline were used. Furthermore, in Peters, Kerschgens, and Paczko (2013), a framework was developed for the prediction of the probability for a localized detonation, which leads to "super-knock" (Kalghatgi and Bradley, 2012). The prediction is based on a combination of the detonation diagram proposed in Bradley et al., (2002) and Gu, Emerson, and Bradley, 2003) with a refined theory of turbulence (Wang and Peters, 2006, 2008). Here, approximated ignition delay times for PRF surrogates are used, but also their derivative with respect to the temperature is needed.

5 CHEMICAL EFFECTS IN CI ENGINES

Owing to the different combustion process in CI engines as compared to SI engines, the influence of chemical kinetics becomes even more important. In CI engines, the diesel fuel is injected directly as a liquid spray into a hot environment composed of compressed air with or without recirculated exhaust gas. Atomization and evaporation occurs very rapidly resulting in a highly stratified mixture of gaseous fuel and oxidizer with local fuel–air equivalence ratios ranging from very lean to very rich mixtures. The part of that mixture, which is within the flammability limits, then auto-ignites first. This is called the *premixed burn stage* and is associated with a large heat release rate. The remaining part of the mixture burns in the following in a nonpremixed, diffusion-controlled stage. During this stage, soot is formed in the rich part of the mixture. As mixing with the surrounding oxidizer continues, most of the soot is oxidized, whereas NO_x is formed in the lean to stoichiometric part of the mixture as the temperature increases.

There are several models used nowadays for the computational fluid dynamics (CFD) modeling of ignition, combustion, and pollutant formation in CI engines based on detailed chemistry. Direct integration of the chemistry at every grid point of the mesh based on the mean values of all scalar variables may appear the most straightforward, although it ignores all scalar fluctuations and turbulence closure. Other models are the characteristic time (CTC) and the representative interactive flamelet (RIF) model. The CTC model (Kong, Han, and Reitz, 1995) is a local exchange model where the change of the mass fraction Y_i is modeled as

$$\frac{\mathrm{d}Y_i}{\mathrm{d}t} = -\frac{Y_i - Y_i^*}{\tau_L} \qquad (14)$$

Here, Y_i^* is the local and instantaneous thermodynamic equilibrium value of mass fraction Y_i and τ_L the CTC to achieve such equilibrium (Kong, Han, and Reitz, 1995). The CTC τ_L is the sum of a laminar timescale derived from a global reaction rate and a turbulent timescale that is proportional to the integral timescale of turbulence. The RIF model will be presented in more detail later.

Singh, Reitz, and Musculus (2006) have compared the direct integration, the CTC, and the RIF models. Depending

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto116
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

on the engine operation point, the heat release rates predicted by the three models differ in time (in terms of deg. CA) and their peak value, but no specific trend could be deduced. While the implementation of detailed chemistry by direct integration is probably the most expensive, the CTC and the RIF models solve chemistry "on the fly."

The RIF model uses a flamelet code (cf. Figure 3), where the mixture fraction Z is the independent variable. This is defined as m_1, the mass of the fuel (subscript 1), divided by the mass of fuel and oxidizer (subscript 2)

$$Z = \frac{m_1}{m_1 + m_2} \tag{15}$$

Denoting stoichiometric mixture fraction as Z_{st}, fuel is deficient and the mixture is called *fuel lean* for $Z < Z_{st}$. Similarly, if $Z > Z_{st}$, oxygen is deficient and the mixture is called *fuel rich*. There is a unique relation between Z and the local fuel-to-air equivalence ratio ϕ

$$\phi = \frac{Z}{1-Z}\frac{(1-Z_{st})}{Z_{st}} \tag{16}$$

The flamelet model is based on the assumption of a thin reaction zone. Then, the flamelet structure is to leading order described by the one-dimensional time-dependent equations

$$\rho\frac{\partial T}{\partial t} - \rho\frac{\chi_{st}}{2}\frac{\partial^2 T}{\partial Z^2} = \sum_{i=1}^{n}\frac{h_i}{c_p}\omega_i + \frac{\dot{q}_R}{c_p} + \frac{1}{c_p}\frac{\partial p}{\partial t}$$

$$\rho\frac{\partial Y_i}{\partial t} - \rho\frac{\chi_{st}}{2}\frac{\partial Y_i}{\partial Z^2} = \dot{\omega}_i \quad i = 1, 2, \ldots, k \tag{17}$$

Here, h_i is the specific enthalpy, ω_i the chemical source term of species i, c_p the heat capacity at constant pressure, and $\dot{q}_R$ the radiative heat loss. Furthermore,

$$\chi_{st} = 2D\left(\frac{\partial Z}{\partial x_\alpha}\right)^2_{st} \tag{18}$$

is the instantaneous scalar dissipation rate at stoichiometric conditions, where D is the thermal diffusivity and x_α are the spatial coordinates. It has the dimension $1/s$ and may be interpreted as the inverse of a CTC. It acts as a prescribed parameter in Equation 17, representing the influence of convection and diffusion normal to the surface of the stoichiometric mixture. In the limit $\chi_{st} \to 0$, the equations for a homogeneous reactor are obtained.

Models in nonpremixed turbulent combustion are often based on the presumed shape PDF approach. This requires the knowledge of the Favre mean mixture fraction $\tilde{Z}$ and its variance $\tilde{Z}''^2$ at each position $\mathbf{x}$ and time t. Then, a suitable

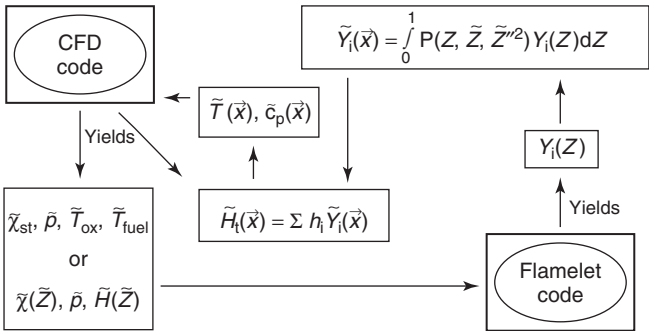


Figure 3. Code structure of the representative interactive flamelet concept.

two-parameter probability density function is "presumed" in advance, thereby fixing the functional form of the PDF by relating the two parameters in terms of the known values of $\tilde{Z}$ and $\tilde{Z}''^2$ at each point of the flow field. As in a two-feed system, the mixture fraction Z varies between $Z = 0$ and $Z = 1$; the beta function PDF is widely used for the function $P(Z, \tilde{Z}, \tilde{Z}''^2)$ referred to in Figure 3. More details about the beta function PDF are given in Peters (2000).

The solution of the flamelet equations provides the local flame structure in terms of species mass fractions as function of mixture fraction. Most importantly, this model allows for the use of detailed chemical kinetics, which is a prerequisite for accurate simulations of auto-ignition and pollutant formation. However, the solution of the flamelet equations requires further consideration. For stationary problems, the time dependence in these equations is often neglected. Then, the flamelet equations can be solved and the solutions can be tabulated as function of the governing parameters, in the standard formulation only the scalar dissipation rate. However, the time dependence of the flamelet equations becomes important in diesel engines, because ignition is a temporal process and because of the time-varying pressure and scalar dissipation rate. These parameters, therefore, have to be specified in the solution of the flamelet equations in a time-accurate manner from a CFD solution, which in itself requires the solution of the flamelet equations. In the RIF model, both CFD and flamelet equations are, therefore, solved simultaneously and interactively. The CFD solver then provides at each time step domain-averaged values of the flamelet parameters, stoichiometric scalar dissipation rate, and pressure; the solution of the flamelet equations together with the presumed PDF of mixture fraction provides the Favre-averaged density and mass fraction fields in the CFD domain (Peters, 2000; Barths, Pitsch, and Peters, 1999; Barths, Hasse, and Peters, 2000). The calculation procedure is schematically shown in Figure 3.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto116
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

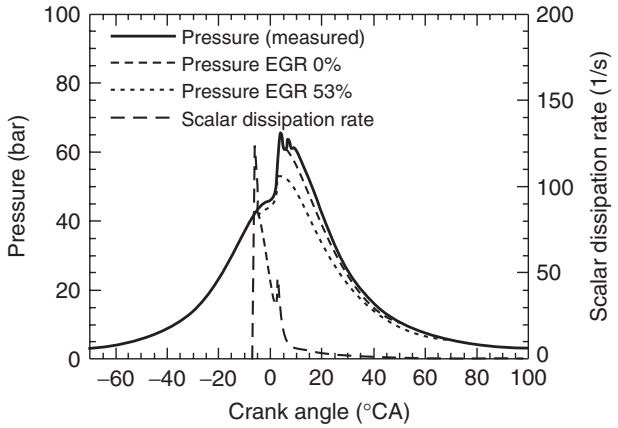


Figure 4. Calculated pressure compared with experiments for 0% EGR (From Pitsch, Barths, and Peters, 1996. Copyright © 1996 SAE International. Reprinted with permission.)

An early example of the application of the RIF model was presented by Barths, Pitsch, and Peters (1999), Pitsch, Barths, and Peters (1996), where it was used to predict ignition, soot, and NO_x, all with a detailed chemical kinetic model, for a Volkswagen DI 1.9l engine that was fueled with *n*-heptane and operated at varying levels of EGR. Exemplary results from the simulations are given in Figures 4 and 5 and show that ignition delay, pressure, NO_x, and soot are predicted with good accuracy. More recently, in an application by Cummins Inc., the RIF model has served as the central ingredient for the design process of modern CI engines (Eckerle and Stanton, 2006).

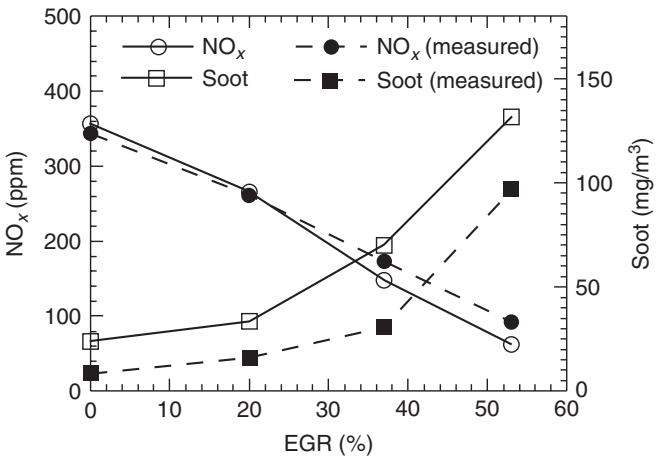


Figure 5. Calculated and measured exhaust gas concentrations of NO_x and soot mass for varying levels of EGR (Pitsch, Barths, and Peters (From Pitsch, Barths, and Peters, 1996. Copyright © 1996 SAE International. Reprinted with permission.)

6 SUMMARY

In this chapter, we have given a few examples that show how global and elementary kinetics can be used in numerical combustion simulations of SI and CI engines. There are, of course, many details to be considered in order to make their implementation successful. However, by replacing commercial fuels by appropriate surrogates and by reducing the detailed kinetic models for these surrogates appropriately, CFD calculations can be performed within reasonable compute times. A further reduction comes from tabulating either source terms or properties such as ignition delay times and burning velocities. Thereby, one now is able to capture many of the effects for which chemistry is responsible in engine combustion.

REFERENCES

Barths, H., Hasse, C., and Peters, N. (2000) Computational fluid dynamics modelling of non-premixed combustion in direct injection diesel engines. *International Journal of Engine Research*, **1**, 149–267.

Barths, H., Pitsch, H., and Peters, N. (1999) 3D-simulation of DI diesel combustion and pollutant formation using a two-component reference fuel. *Journal of Fluid Mechanics*, **54**, 233–244.

Bradley, D. and Kalghatgi, G.T. (2009) Influence of autoignition delay time characteristics of different fuels on pressure waves and knock in reciprocating engines. *Combustion and Flame*, **156**, 2307–2318.

Bradley, D., Morley, C., Gu, X.J., and Emerson, D.R. (2002) Amplified pressure waves during autoignition: relevance to CAI engines. SAE Technical Paper 2002-01-2868.

Candel, S. and Poinsot, T. (1990) Flame stretch and the balance equation for the flame area. *Combustion Science and Technology*, **70**, 1–15.

Colin, O., Benkenida, A., and Angelberger, C. (2003) 3D modeling of mixing, ignition and combustion phenomena in highly stratified gasoline engines. *Oil & Gas Science and Technology*, **58**, 47–62

Dahms, R., Fansler, T.D., Drake, M.C., Kuo, T.W., Lippert, A.M., and Peters, N. (2009) Modeling ignition phenomena in spray-guided spark-ignition engines. *Proceedings of the Combustion Institute*, **32**, 2743–2750.

Dekana, M. and Peters, N. (1999) Combustion modelling with the G-euquation. *Oil & Gas Science and Technology*, **54**, 265–270.

Duclos, J.M. and Colin, O. (2001) *Arc and kernel tracking ignition model for 3D spark-ignition engines calculations*. 5th International Symposium on Diagnostics and Modeling of Combustion in Internal Combustion Engines (COMODIA 2001), July, 2001, pp. 343–350.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto116
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Duclos, J.M. and Zolver, M. (1998) *3D modeling of intake, injection and combustion in a DI-SI engine under homogeneous and stratified operating conditions*. 4th International Symposium on Diagnostics and Modeling of Combustion in Internal Combustion Engines, July 20–23, 1998, Kyoto International Conference Hall, Kyoto, Japan, pp. 335–340.

Eckerle, W.A. and Stanton, D.W. (2006) Analysis-led design process for cumminas engine development. Thiesel Conference on Thermo- and Fluid Dynamic Processes in Diesel Engines.

Gu, X.J., Emerson, D.R., and Bradley, D. (2003) Modes of reaction front propagation from hot spots. *Combustion and Flame*, **133**, 63–74.

Jay, S. and Colin, O. (2011) A variable volume approach of tabulated detailed chemistry and its applications to multidimensional engine simulations. *Proceedings of the Combustion Institute*, **33**, 3065–3072.

Kalghatgi, G.T. and Bradley, D. (2012) Pre-ignition and 'Super-knock' in turbo-charged spark-ignition engines. *International Journal of Engine Research*, **13**, 399–414.

Kong, S., Han, Z., and Reitz, R.D. (1995) The development and application of a diesel ignition and combustion model for multidimensional engine simulation. SAE Technical Paper 950278.

Livengood, J.C. and Wu, P.C. (1955) *Correlation of autoignition phenomena in internal combustion engines and rapid compression machines*. 5th Symposium (International) on Combustion, The Combustion Institute, pp. 347–356.

Løvås, T. (2009) Automatic generation of skeletal mechanisms for ignition combustion based on level of importance analysis. *Combustion and Flame*, **156**, 1348–1358.

Lu, T. and Law, C.K. (2006) On the applicability of directed relation graphs to the reduction of reaction mechanisms. *Combustion and Flame*, **154**, 153–163.

Lu, T.F. and Law, C.K. (2008) Strategies for mechanism reduction for large hydrocarbons: n-heptane. *Journal of Fluid Mechanics*, **652**, 45–64.

Lu, T., Ju, Y., and Law, C.K. (2001) Complex csp for chemistry reduction and analysis. *Combustion and Flame*, **126**, 1445–1455.

Mittal, V. and Pitsch., H. (2013) A flamelet model for premixed combustion under variable pressure conditions. *Proceedings of the Combustion Institute*, **34**, 2995–3003.

Pepiot-Desjardins, P. and Pitsch, H. (2008) An efficient error-propagation-based reduction method for large chemical kinetic mechanisms. *Combustion and Flame*, **154** (1–2), 67–81.

Peters, N. (1992) *Fifteen Lectures On Laminar and Turbulent Combustion*, http://www.itv.rwth-aachen.de/fileadmin/Lehre-Seminar/Combustion/SummerSchool (accessed 19 September 2013).

Peters, N. (1999) The turbulent burning velocity for large scale and small scale turbulence. *Journal of Fluid Mechanics*, **384**, 107–132.

Peters, N. (2000) Turbulent Combustion, Cambridge University Press, Cambridge.

Peters, N. and Rogg, B. (1993) *Reduced Kinetic Mechanisms for Applications in Combustion Systems*, Lecture Notes in Physics, Monograph 15, Springer, Berlin.

Peters, N., Kerschgens, B., and Paczko, G. (2013) Super-Knock prediction using a refined theory of turbulence. SAE Technical Paper 2013-01-1109.

Peters, N., Paczko, G., Seiser, R., and Seshadri, K. (2002) Temperature cross-over and non-thermal runaway at two-slage ignition at n-heplane. *Combustion and Flame*, **128**, 38–59.

Pitsch, H., Barths, H., and Peters, N. (1996) Three-dimensional modeling of NOx and soot formation in Dl-diesel engines using detailed chemistry based on the interactive flamelet approach. SAE Technical Paper 962057.

Pope, S.B. (1997) Computational efficient implementation of combustion chemistry using in situ adaptive tabulation. *Combustion Theory and Modelling*, **1** (1), 41–63.

Seshadri, K., Humer, S., and Seiser, R. (2008) Activation-energy asymptotic theory of auto-ignition of condensed hydrocarbon fuels in non-premixed flows with comparison to experiment. *Combustion Theory and Modelling*, **12** (5), 831–855.

Singh, S., Reitz, R.D., and Musculus, M.P.B. (2006) Comparison of characteristic time (CTC), representative interactive flamelet (RIF) and direct integration with detailed chemistry combustion models against optical diagnostic data for multi-mode combustion in a heavy-duty DI diesel engine. SAE Technical Paper 2006-01-0055, pp. 129–141.

Sun, Y., Chen, Z., Gou, X., and Ju, Y. (2010) A path flux analysis method for the reduction of detailed chemical kinetic mechanisms. *Combustion and Flame*, **157**, 1298–1307.

Tan, Z., Kong, S.C., and Reitz, R.D. (2003) Modeling premixed and direct injection SI engines combustion using the G-equation model. SAE Technical Paper 2003-01-1843.

Tan, Z. and Reitz, R.D. (2003) Modeling ignition and combustion in spark-ignition engines using a level-set method. SAE Technical Paper 2003-01-0722.

Tan, Z. and Reitz, R.D. (2006) An ignition and combustion model based on the level-set method for spark ignition engine multidimensional modeling. *Combustion and Flame*, **145**, 1–15.

Wang, L. and Peters, N. (2006) The length scale distribution function of the distance between extremal points in passive scalar turbulence. *Journal of Fluid Mechanics*, **554**, 457–475.

Wang, L. and Peters, N. (2008) Length scale distribution fuctions and conditional means for various fields in turbulence. *Journal of Fluid Mechanics*, **608**, 113–138.

Zeuch, T., Moéac, G., Ahmed, S.S., and Mauss, F. (2008) A comprehensive skeletal mechanism for the oxidation of n-heptane generated by chemistryguided reduction. *Combustion and Flame*, **155**, 651–674.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto116
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

NO_x Formation and Models

Craig T. Bowman

Stanford University, Stanford, CA, USA

1 Introduction 1
2 Reaction Mechanisms for Nitric Oxide Formation 1
3 Reaction Mechanism for NO Formation from Nitrogen in the Fuel 4
4 Reaction Mechanisms for NO_2 and N_2O 4
5 Reburn Mechanism for NO Removal from Combustion Products 5
6 Contributions of Mechanisms to NO Emissions from Combustion Processes 6
7 Concluding Remarks 8
References 8

1 INTRODUCTION

The principal nitrogen oxides emitted from automotive engines are nitric oxide (NO) and nitrogen dioxide, collectively referred to as NO_x. A variety of strategies for reducing nitrogen oxide emissions from these engines have been implemented in modern engines and include both in-cylinder strategies and after-treatment strategies. The focus of this chapter is on in-cylinder NO_x reduction; after-treatment strategies are the subject of Gas Aftertreatment Systems. The concepts underlying in-cylinder NO_x emissions by reduction may be understood in terms of the reaction mechanisms for the formation and removal of NO_x in the engine combustion process.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto117
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

In the following sections, the current state of understanding of the gas-phase reaction mechanisms for nitrogen oxide chemistry in combustion processes is discussed. The discussion is separated into two parts—mechanisms for nitrogen oxide formation and mechanisms for nitrogen oxide removal—recognizing that in engine combustion processes these mechanisms may be coupled.

2 REACTION MECHANISMS FOR NITRIC OXIDE FORMATION

NO is the primary nitrogen oxide emitted from automotive engines. The principal route to NO formation in automotive engine combustion is oxidation of molecular nitrogen present in the combustion air or in the fuel (some natural gas contains up to 10% by volume of N_2). In addition, nitrogen-containing additives in the fuel can contribute to NO formation in the engine. NO may be formed from N_2 by three reaction mechanisms: (i) the thermal NO mechanism, (ii), the prompt NO mechanism, and (iii) the N_2O mechanism. In spark-ignition and compression ignition engines, the dominant NO formation process is the thermal mechanism. All three mechanisms are described in Sections 2.1–2.3. The mechanism for the formation of NO from nitrogen additives in the fuel is discussed in Section 4.

2.1 Thermal NO formation mechanism

Three reversible reactions comprise the thermal NO formation mechanism:

$$O + N_2 \rightleftharpoons NO + N \qquad \text{(R1)}$$

$$N + O_2 \rightleftharpoons NO + O \qquad \text{(R2)}$$

Table 1. Rate coefficient expressions for reactions (R1)–(R3)[a].

Reaction	A	B	C(K)
(R1) $O + N_2 \rightarrow NO + N$	1.81×10^{14}	0	−38,400
(R2) $N + O_2 \rightarrow NO + O$	9.0×10^{9}	1	−3,270
(R3) $N + OH \rightarrow NO + H$	1.08×10^{14}	−0.2	0

Units: cm^3, mol, s.
[a]Data compiled from Baulch *et al*. (2005).

$$N + OH \rightleftharpoons NO + H \qquad (R3)$$

The temperature-dependent rate parameters for reactions (R1)–(R3) have been measured over a wide temperature range. The rate-controlling reaction for thermal NO formation is the slow reaction $O + N_2 \rightarrow NO + N$ (R1). The rate coefficient expressions for (R1)–(R3), in the form $AT^B \exp[-C/T(K)]$, are available in Baulch *et al*. (2005) and are tabulated in Table 1.

The primary parameters governing thermal NO formation can be determined using an equation for the maximum NO formation rate. This equation may be derived from (R1)–(R3) by invoking as quasi-steady-state approximation for N-atoms neglecting back reactions,

$$\frac{d(N)}{dt} = k_1(O)(N_2) - k_2(N)(O_2) - k_3(N)(OH) \approx 0$$

In molar concentration units, the maximum NO formation rate may then be expressed,

$$\left[\frac{d(NO)}{dt}\right]_{max} = 2k_1(O)(N_2) \times \left\{\frac{1 - (NO)^2/K(O_2)(N_2)}{1 + k_{-1}(NO)/[k_2(O_2)k_3(OH)]}\right\} \qquad (1)$$

where $K = (k_1/k_{-1})(k_2/k_{-2})$ is an equilibrium constant for the reaction $N_2 + O_2 \rightleftharpoons 2NO$.

At engine combustion conditions, the characteristic time for NO formation is very much greater than the characteristic time for fuel oxidation. Hence, to a reasonable approximation, the values for (O), (O_2), (N_2), and (OH) and the temperature used to evaluate the rate parameters, k_1–k_3, in Equation 1 may be assumed to be their respective equilibrium values in the burned gas. For combustion of lean and stoichiometric fuel–air mixtures, $k_3(OH)_{eq} \ll k_2(O_2)_{eq}$, so that Equation 1 may be approximated by,

$$\left[\frac{d(NO)}{dt}\right]_{max} = 2k_1(O)_{eq}(N_2)_{eq} \qquad (2)$$

When $(O)_{eq}$ is related to $(O_2)_{eq}$ by the equilibrated reaction, $\frac{1}{2}O_2 \rightleftharpoons O$, then Equation 2 may be expressed,

$$\left[\frac{d(NO)}{dt}\right]_{max} = 1.70 \times 10^{17} T^{-\frac{1}{2}} \exp\left[-\frac{69750}{T(K)}\right] \times (O_2)_{eq}^{\frac{1}{2}}(N_2)_{eq} mol/cm^3/s^1 \qquad (3)$$

The strong dependence of the thermal NO formation rate on the post-combustion gas temperature and the somewhat weaker dependence on the O_2 concentration in the burned gas, $(O_2)_{eq}$, are evident in Equation 3. Methods for reducing NO_x emissions from automotive engines, where NO is produced mainly by the thermal mechanism, are directed toward reduction of the NO formation rate rather than by enhancing the NO removal rate, which is very slow at typical engine combustion temperatures. From Equation 3, it can be seen that modification of the combustion process, such as exhaust gas recirculation or controlling spark or fuel injection timing, to reduce the maximum burned gas temperatures, the availability of oxygen, or both will reduce the NO formation rate and, hence, NO_x emissions.

Also evident from Equation 3 is that the amount of NO formed is dependent on time, especially the time at which the burned gas temperature is high. A characteristic time for NO formation in lean and stoichiometric fuel–air mixtures may be defined as

$$\tau_{NO} \equiv \frac{1}{\frac{d(NO)}{dt} \cdot \frac{1}{(NO)_{eq}}} \qquad (4)$$

Combining Equations 3 and 4, together with an expression for the equilibrium constant of $N_2 + O_2 \rightleftharpoons 2NO$, yields

$$\tau_{NO} = \frac{8 \times 10^{-16} T \exp[58300/T(K)]}{\sqrt{P}} \qquad (5)$$

2.2 Prompt NO formation mechanisms

Measurements of NO concentrations near the flame zones of hydrocarbon–air and hydrogen–air premixed flames and on the air side of hydrocarbon diffusion flames show that maximum NO formation rates exceed those given by Equation 1. Three mechanisms for this accelerated NO formation have been identified: (i) superequilibrium O and OH concentrations that accelerate the rates of the thermal NO mechanism; (ii) a reaction sequence, shown in Figure 1, that is initiated by reactions of hydrocarbon radicals, present in and near the reaction zone, with molecular nitrogen

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto117
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

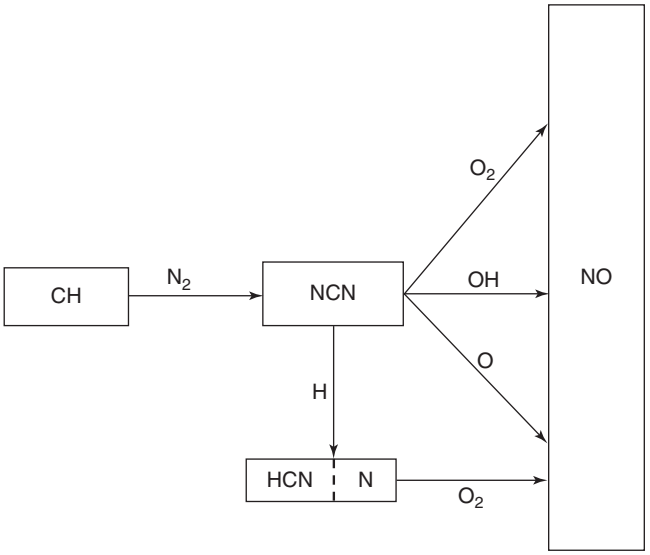


Figure 1. Reaction path diagram showing the principal pathways leading to prompt NO formation.

(Fenimore, 1970); and (3) reaction between the nitrogen-containing radical NNH and O atoms (Bozzelli and Dean, 1995; Harrington *et al*., 1996), present in the burned gas.

The relative importance of the three mechanisms of prompt NO formation depends on the combustion conditions. Acceleration of NO formation via the thermal mechanism by superequilibrium O and OH concentrations can be important in non-premixed flames, in stirred reactors for lean fuel–air mixtures, and in low pressure premixed flames. For near stoichiometric and rich premixed hydrocarbon–air flames and for hydrocarbon–air diffusion flames, prompt NO formation occurs primarily by the reaction sequence shown in Figure 1. Mechanism 3 has been proposed as an important source of NO in lower temperature flames (Bozzelli and Dean, 1995).

Rate coefficient expressions for the prompt NO reaction pathways shown in Figure 1 are available from Vasudevan *et al*. (2007) and Lin and coworkers (2005, 2007, 2009) and are tabulated in Table 2. Of particular importance are the rates of the prompt NO initiation reaction,

$$CH + N_2 \longrightarrow NCN + H \qquad (R4)$$

and the subsequent reactions of NCN to form NO via (R5)–(R7).

$$NCN + O_2 \longrightarrow NCO + NO \qquad (R5)$$

$$NCN + OH \longrightarrow HCN + NO \qquad (R6)$$

$$NCN + O \longrightarrow CN + NO \qquad (R7)$$

Table 2. Rate coefficient expressions for reactions (R4)–(R8).

Reaction	A	B	C (K)	References
(R4)	6.03×10^{12}	0	−11150	Vasudevan *et al*. (2007)
(R5)	3.80×10^{9}	0.51	−12377	Zhu and Lin (2005)
(R6)	4.71×10^{10}	0.44	−2013	Zhu, Nguyen, and Lin (2009)
(R7)	2.55×10^{13}	0.15	+17	Zhu and Lin (2007)

Units for A: cm^3, mol, s.

An alternative precursor species for NCN formation, C_2O, via the following reaction pathway has been proposed by Williams and Fleming (2007) and Konnov (2009).

$$C_2O + N_2 \longrightarrow NCN + CO \qquad (R8)$$

At this time, the importance of this pathway for prompt NO formation has not been established. Furthermore, reaction pathways for C_2O formation in flames are uncertain and accurate expressions for the rate coefficient of (R8) are not currently available.

2.3 Mechanism for NO formation from N_2O

Wolfrum (1972) suggested that NO can also be formed from N_2 by a reaction sequence involving the formation of N_2O by the three-body recombination reaction (M = any collision partner),

$$O + N_2 + M \longrightarrow N_2O + M \qquad (R9)$$

and subsequent reaction of the N_2O to form NO via,

$$N_2O + O \longrightarrow NO + NO \qquad (R10)$$

The formation of NO by the N_2O mechanism increases in importance as the fuel–air ratio decreases, as the burned gas temperature decreases, or as pressure increases. The N_2O mechanism would be the most important under conditions where the NO emissions are relatively low, as for example, in lean premixed gas turbine combustors (Malte and Pratt, 1974) or in homogeneous charge compression ignition (HCCI) engines.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto117
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3 REACTION MECHANISM FOR NO FORMATION FROM NITROGEN IN THE FUEL

Conventional distillate fuels used in spark-ignition and compression ignition engines contain very little nitrogen, although organic nitrogen-containing compounds are used as additives. These compounds typically are high molecular weight amines that are added in amounts ranging from a few tens to a few hundred parts per million. While it is likely that organic nitrogen additives at these levels will not be significant contributors to NO_x emissions from these engines, for the purpose of completeness, the chemical pathways for amine conversion to NO will be presented.

Figure 2 shows these reaction pathways. In the figure, ammonia is used as a surrogate amine compound. Ammonia will react with small radicals present in the hot combustion products in a sequential process leading, eventually to N atoms. The N atoms will react with oxygen-containing species (O_2 and OH) to form NO. Some N atoms will react with NO present in the burned gas to reduce it to N_2. An intermediate nitrogen species, HNO, also can contribute to NO formation, via the pathway shown in Figure 2. The competition between the NO-forming and NO-reducing reaction pathways determines the net amount of NO formed by this process. In oxygen-containing combustion products and for low levels of the nitrogen additive, the conversion of the fuel–N to NO is nearly quantitative (Bowman, 1991).

4 REACTION MECHANISMS FOR NO_2 AND N_2O

4.1 Mechanism for formation and removal of NO_2

Measured NO_2 emissions from spark-ignition engines are relatively small; however, in diesel engines, NO_2 can be 10–30% of the total NO_x emissions (Heywood, 1988). Measurements of NO and NO_2 in premixed and diffusion flames show that there are relatively large NO_2/NO ratios near the flame zone (Merryman and Levy, 1975; Drake *et al*., 1987). The gas-phase reaction mechanism for the formation and removal of NO_2 in combustion is well known. The principal NO_2 formation pathway is

$$NO + RO_2 \longrightarrow NO_2 + RO \qquad \text{(R11)}$$

In this reaction, RO_2 is a peroxy radical and the R is an H atom or an alkyl group. In low temperature regions of

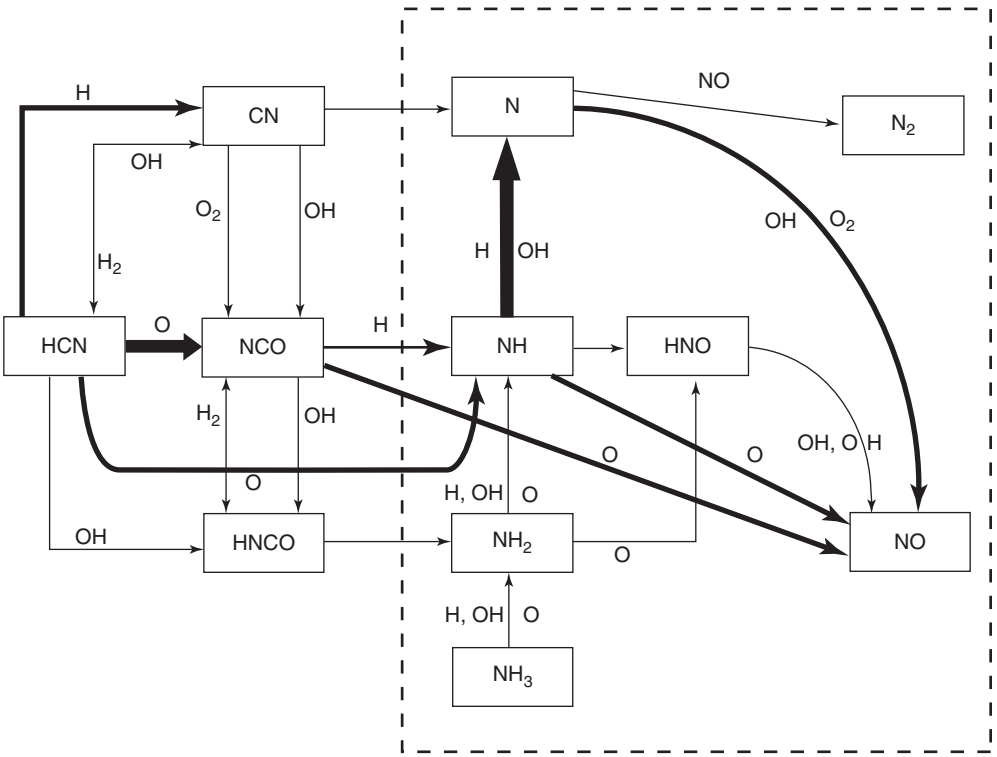


Figure 2. Reaction path diagram for oxidation of NH_3 and HCN in flames.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto117
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

flames, RO_2 is relatively stable and it can react with NO formed in higher temperature regions and be transported by diffusion to lower temperature regions. The NO_2 formed in (R11) is converted back to NO via

$$NO_2 + H \longrightarrow NO + OH \quad \text{(R12)}$$

$$NO_2 + O \longrightarrow NO + O_2 \quad \text{(R13)}$$

At elevated temperatures ($T > 1500$ K), NO_2 removal by (R12) and (R13) is rapid because of the presence of high radical concentrations in the combustion products. Typical lifetimes of NO_2 in combustion products at 1500 K are less than 10 ms. Persistence of NO_2 in the engine exhaust is likely due to quenching by mixing with cooler fluids.

4.2 Mechanism for formation and removal of N_2O

N_2O emissions from automotive engines are not currently regulated, however N_2O is a greenhouse gas and, as such, N_2O emissions are of interest. In current light-duty gasoline engines with three-way catalysts, N_2O is formed primarily in the catalyst and N_2O emissions are dependent on the sulfur content of the fuel. N_2O can also be formed as a by-product in diesel engines equipped with SCR NO reduction systems. Direct in-cylinder production of N_2O in current engine designs should be very small; however, as noted in Section 2.3, in-cylinder N_2O formation will increase at the elevated pressures and lean-burn conditions found in some HCCI engines.

The dominant gas-phase N_2O formation reaction in non-nitrogen containing fuels is (R9). The primary N_2O removal steps in combustion products are

$$N_2O + H \longrightarrow N_2 + OH \quad \text{(R14)}$$

$$N_2O + O \longrightarrow N_2 + O_2 \quad \text{(R15)}$$

$$\longrightarrow NO + NO \quad \text{(R10)}$$

The rate parameters for these reactions are known fairly accurately at combustion temperatures (Baulch *et al.*, 2005). For temperatures above 1500 K, the lifetime of N_2O in combustion products is less than 10 ms.

5 REBURN MECHANISM FOR NO REMOVAL FROM COMBUSTION PRODUCTS

Reburning is the name given to a reaction process for NO removal in combustion products by reaction with small hydrocarbon radicals to form cyano species. The reburn reaction pathways are active in fuel-rich regions, such as found in diffusion flames, where these cyano species react preferentially to form N_2. The final levels of NO emitted from hydrocarbon–air flames are the balance of NO formation via the reaction pathways described in Section 2 and NO removal via reburn.

A reaction path diagram for NO reburn in methane combustion, due to Miller, Durant, and Glarborg (1998), is shown in Figure 3. The important reaction pathways, in so far as reburn is concerned, are the reactions of the radicals CH_3 and HCCO with NO to produce cyano species. Figure 3 also shows the reaction pathways for these radicals from hydrocarbon fuel species.

The specific elementary reaction steps that participate in the reburn process include

$$CH_3 + NO \longrightarrow HCN + H_2O \quad \text{(R16a)}$$

$$CH_3 + NO \longrightarrow H_2CN + OH \quad \text{(R16b)}$$

and

$$HCCO + NO \longrightarrow HCN + CO_2 \quad \text{(R17a)}$$

$$HCCO + NO \longrightarrow HCNO + CO \quad \text{(R17b)}$$

At elevated temperatures, the reaction of NO with CH will contribute to NO removal via

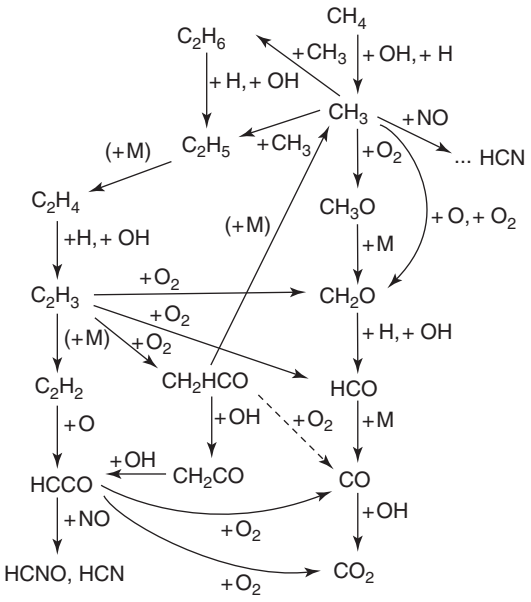


Figure 3. Reaction path diagram for NO reburn in CH_4 combustion. (Reproduced with permission from Miller, Durant, and Glarborg (1998). © The Combustion Institute.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto117
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

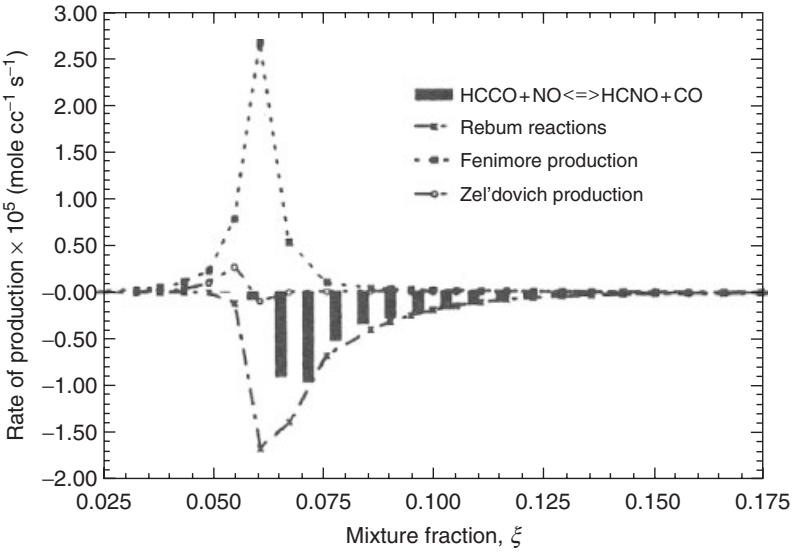


Figure 4. Contributions of reactions to NO reburn in an atmospheric pressure diffusion flame. (Reproduced with permission from Marro, Pivovarov and Miller (1997). © The Combustion Institute.)

$$CH + NO \longrightarrow HCN + O \qquad (R18)$$

Subsequent reaction of the HCN produced in (R16a), (R17a), and (R18) occurs via the reaction pathways shown in Figure 2. Under fuel-rich conditions, the pathways leading to N_2 is favored. The H_2CN formed in (R16b) dissociates to produce HCN via

$$H_2CN \longrightarrow HCN + H \qquad (R19)$$

and, therefore, contributes to the $NO \rightarrow HCN \rightarrow N_2$ process.

The HCNO formed in reaction (R17b) reacts primarily via

$$HCNO + O \longrightarrow HCO + NO \qquad (R20)$$

$$HCNO + OH \longrightarrow HCOH + NO \qquad (R21)$$

$$HCNO + H \longrightarrow HNCO + H \qquad (R22)$$

Reactions (R20) and (R22) effectively regenerate the NO removed in reaction (R17b) and, hence, are ineffective in removing NO. The HNCO produced in reaction (R22) will react with O, H, or OH to produce amine species (NH, NH_2), HNO or NCO, that may react to form NO, N_2O, or N_2.

The rate parameters for key reactions in the reburn processes, (R16a), (R16b), (R17a), and (R17b), from Miller, Durant, and Glarborg (1998), are

$$k_{16} = 0.3T^{3.52} \exp[-1990\ \mathrm{K/T}]\ \mathrm{cm^3/mol/s}$$

$$k_{17a} = 1.45 \times 10^{16} T^{-0.968} \exp[-326\ \mathrm{K/T}]\ \mathrm{cm^3/mol/s}$$

$$k_{17b} = 1.17 \times 10^{11} T^{0.65}\ \mathrm{cm^3/mol/s}$$

Figure 4 shows calculated contributions of reburn and NO formation reactions in an atmospheric pressure CH_4–air diffusion flame (Marro, Pivovarov, and Miller, 1997). The vertical axis shows the NO formation (>0) and removal (<0) rates. The horizontal axis is mixture fraction, ξ defined as $\xi \equiv$ initial CH_4 mass flow rate/total (CH_4 + air) mass flow rate, the Fenimore production is due to the prompt NO mechanism and Zeldovich production is due to the thermal NO mechanism, described in Sections 2.2 and 2.1, respectively. NO removal by reburn occurs on the fuel-rich side of the flame. The HCCO + NO reaction plays a major role in the reburn process in this flame, with the remainder of the reburn occurring through CH_x + NO reactions.

6 CONTRIBUTIONS OF MECHANISMS TO NO EMISSIONS FROM COMBUSTION PROCESSES

In this section, representative contributions to NO production of the NO formation and removal mechanisms described earlier are presented for combustion of methane and air. Figure 5 shows the calculated contributions of the thermal, prompt, and N_2O mechanisms to NO formation in a one-dimensional, premixed, laminar CH_4–air flame at

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto117
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

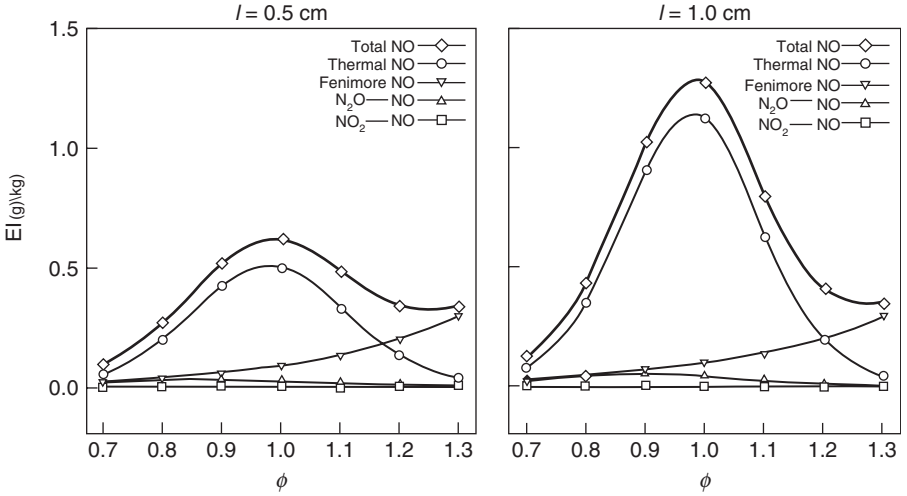


Figure 5. Contributions of mechanisms to NO formation in a premixed laminar CH_4–air flame at atmospheric pressure for various fuel–air equivalence ratios. Two positions above the burner surface, *l*, are shown. (Reproduced with permission from Nishioka *et al*. (1994). © The Combustion Institute.)

atmospheric pressure as a function of fuel–air equivalence ratio, ϕ, defined as

$$\phi \equiv \frac{n_{\text{fuel}}/n_{O_2}}{(n_{\text{fuel}}/n_{O_2})_{\text{stoichiometric}}}$$

at two axial locations, *l*, in the flame, measured with respect to the burner surface (Nishioka *et al*., 1994). The NO levels are reported in terms of an emission index, EI_{NO}, (g NO/kg CH_4).

From Figure 5, it can be seen that for lean and stoichiometric mixtures ($\phi \leq 1$), the thermal NO mechanism is the dominant source of the NO produced in the flame. The prompt NO mechanism, labeled here as Fenimore NO, increases in importance as ϕ increases. For values of $\phi > 1.2$, the prompt NO mechanism is the dominant source of NO. For these particular flame conditions, the N_2O mechanism contributes only a small amount of NO production and, then, only for the leanest mixtures.

Figure 6 shows the calculated contributions of the thermal, prompt, and N_2O formation mechanism and the NO reburn mechanism to total NO production in laminar CH_4–air diffusion flames at pressures of 1 and 40 bar (Hewson and Bollig, 1996). The vertical axis is the NO production rate (kmol NO/m^3/s) and the horizontal axis is the mixture fraction, *z* (which is equivalent to ξ). Mixture fraction may be viewed as a normalized fuel–air ratio, ranging from 0 for pure air to 1 for pure fuel. The *z*-value corresponding to a stoichiometric CH_4–air mixture

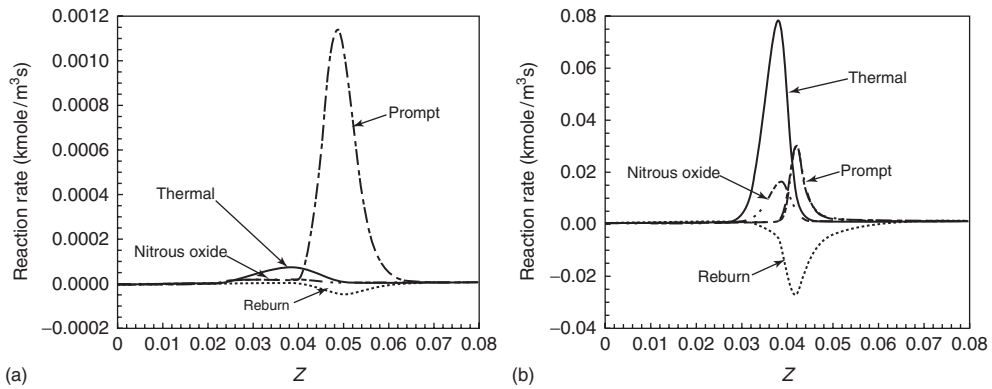


Figure 6. Contributions of mechanisms to NO formation and removal in a laminar CH_4–air diffusion flame at pressures of 1 bar and 40 bar. (Reproduced with permission from Hewson and Bollig (1996). © The Combustion Institute.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto117
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is $z = 0.055$. Values of $z < 0.055$ correspond to the air side of the flame, whereas values of $z > 0.055$ correspond to the fuel side.

From Figure 6, it can be seen that the relative contributions for the different NO formation and removal mechanisms depend on pressure and on the location in the flame. At a pressure of 1 bar, the prompt NO mechanism is the dominant source of NO, with the maximum rate of NO formation occurring near the stoichiometric mixture fraction, $z = 0.055$, that is, near the location of the flame. At this pressure, there are minor contributions to NO formation from the thermal and N_2O mechanisms, with the maximum NO formation rates occurring on the air side of the flame. At the elevated pressure, the relative contribution of the N_2O mechanism to NO formation increases, due mainly to the fact that the rate of (R9) increases as P^3, whereas the rates of (R1) (thermal NO) and (R4) (prompt NO) increase as P^2. NO removal by reburn reaction pathways increases with increasing pressure. Hence, accurate predictions of NO emissions in high pressure diffusion flames, such as occur in diesel engines, require consideration of all three NO formation mechanisms and the reburn NO removal mechanism.

7 CONCLUDING REMARKS

This article has summarized the current understanding of the gas-phase reaction mechanisms and rate parameters for nitrogen oxide formation and removal in combustion. At this time, detailed reaction mechanisms for predicting NO, NO_2, and N_2O emissions are available in the literature (e.g., Miller and Bowman, 1989; Konnov, 2009) and also on the internet (e.g., http://www.me.berkeley.edu/gri_mech/).

REFERENCES

Baulch, D.L., Bowman, C.T., Cobos, C.J., *et al*. (2005) Evaluated rate data for combustion modelling: supplement II *Journal of Physical and Chemical Reference Data*, **34**, 757–1397.

Bowman, C.T. (1991) Chemistry of gaseous pollutant formation and destruction in *Fossil Fuel Combustion: A Source Book* (eds W. Bartok and A.F. Sarofim), John Wiley & Sons, pp. 215–260.

Bozzelli, J.W. and Dean, A.M. (1995) O + NNH: a possible new route for NO production in flames *International Journal of Chemical Kinetics*, **27**, 1097–1110.

Drake, M.C., Correa, S.M., Pitz, R.W., *et al*. (1987) Superequilibrium and thermal nitric oxide formation in turbulent diffusion flames *Combustion and Flame*, **69**, 347–365.

Fenimore, C.P. (1970) Formation of nitric oxide in premixed hydrocarbon flames *Proceedings of the Combustion Institute*, **13**, 373–380.

Harrington, J.E., Smith, G.P., Berg, P.A., *et al*. (1996) Evidence for a new NO production mechanism in flame *Proceedings of the Combustion Institute*, **26**, 2133–2138.

Hewson, J.C. and Bollig, M. (1996) Reduced mechanisms for NO_x emissions from hydrocarbon diffusion flames *Proceedings of the Combustion Institute*, **26**, 2171–2179.

Heywood, J.B. (1988) Internal Combustion Engine Fundamentals, McGraw-Hill, New York.

Konnov, A.A. (2009) Implementation of the NCN pathway of prompt-NO formation in the detailed reaction mechanism *Combustion and Flame*, **156**, 2093–2105.

Malte, P.C. and Pratt, D.T. (1974) The role of energy-releasing kinetics in NO_x formation: fuel lean, jet-stirred CO-air combustion *Combustion Science and Technology*, **9**, 221–231.

Marro, M.A., Pivovarov, M.A., and Miller, J.H. (1997) Strategy for simplification of nitrogen oxide chemistry in a laminar methane/air diffusion flame *Combustion and Flame*, **111**, 208–221.

Merryman, E.L. and Levy, A. (1975) Nitrogen oxide formation in flames: the role of NO_2 and fuel nitrogen *Proceedings of the Combustion Institute*, **15**, 347–365.

Miller, J.A. and Bowman, C.T. (1989) Mechanism and modeling of nitrogen chemistry in combustion *Progress in Energy and Combustion Science*, **4**, 287–338.

Miller, J.A., Durant, J.L., and Glarborg, P. (1998) Some chemical kinetic issues in reburning: the branching ratio of the HCCO + NO reaction *Proceedings of the Symposium on Combustion*, **27**, 235–243.

Nishioka, N., Nakagawa, S., Ishikawa, Y., and Takeno, T. (1994) NO emission characteristics of methane-air double flame *Combustion and Flame*, **98**, 127–138.

Vasudevan, V., Hanson, R.K., Bowman, C.T., *et al*. (2007) Shock tube study of the reaction of CH with N_2: overall rate and branching ratio *The Journal of Physical Chemistry A*, **111**, 11818–11830.

Williams, B.A. and Fleming, J.W. (2007) Experimental and modeling study of NO formation in 10 torr methane and propane flames: evidence for additional prompt-NO precursors *Proceedings of the Combustion Institute*, **31**, 1109–1117.

Wolfrum, J. (1972) Formation of nitric oxide in combustion *Chemie Ingenieur Technik*, **44**, 656–659.

Zhu, R.S. and Lin, M.C. (2005) Ab initio study of the oxidation of NCN by O_2 *International Journal of Chemical Kinetics*, **37**, 593–598.

Zhu, R.S., Nguyen, H.M.T., Lin, M.C. (2009) Ab initio study on the oxidation of NCN by OH: prediction of the individual and total rate constants *The Journal of Physical Chemistry A*, **113**, 298–304.

Zhu, R.S. and Lin, M.C. (2007) Ab initio study on the oxidation of NCN by $O(^3P)$: prediction of the total rate constant and product branching ratios *The Journal of Physical Chemistry A*, **111**, 6766–6771.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto117
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fuel Introduction

Rolf Reitz[1], Lyle Pickett[2], and Mario Trujillo[1]

[1] *University of Wisconsin–Madison, Madison, WI, USA*
[2] *Sandia National Laboratories, Albuquerque, NM, USA*

1 Introduction 1
2 Fuel Delivery in Automotive Engines 3
3 Theories of Atomization, Drop Breakup and Spray Processes 7
4 Direct Numerical Simulations of Atomization 15
References 23

1 INTRODUCTION

Fuel is the lifeblood of the internal combustion engine and various fuel introduction methods are used to mix the fuel with air in the correct proportions for optimal combustion. A key consideration is to be able to vaporize the liquid fuel in the short time available in each engine cycle, and this generally requires that the fuel be finely atomized during the mixing process. Owing to the importance of fuel injection and spray processes in IC engines, many review articles and books have been written on these topics and are available for further study (Reitz and Bracco, 1986; Stiesch, 2003; Reitz, 2006; Jiang *et al*., 2010). Excellent reference books include Challen and Baranescu (1998) for diesel engine fuel injection and Heywood (1988) for carbureted and manifold injection gasoline engines.

This section provides a review of the fundamental processes involved in the introduction of fuels into the combustion chamber. We begin with a brief discussion of various types of spray nozzles and those commonly used in engine applications. Fuel introduction strategies are discussed in Section 2, ranging from classical intake manifold fuel introduction using carburetors to modern port fuel injection (PFI) fuel injectors, with discussion of fuel film and deposit effects. These are followed by a review of direct in-cylinder fuel introduction, with consideration of issues relevant for spark-ignition and diesel engines. Experimental findings on fundamental spray processes, such as the effect of air entrainment on spray penetration and vaporization, injector nozzle, injection timing and fuel effects on combustion, oil dilution, and engine wear are then discussed. In addition, the effect of fuel sprays interactions with the in-cylinder airflows is discussed.

An in-depth understanding of fundamental spray processes is required for successful modeling of engines. Thus, Section 3 reviews current theories of atomization, drop breakup, collision and coalescence, vaporization, drag and deformation, turbulent diffusion, and spray/wall impingement processes. The final section of this chapter is devoted to discussion of new computational tools that are currently under development for improved understanding of spray processes.

1.1 Types of sprays

Sprays are used in many applications, including the process industries (e.g., spray drying, spray cooling, and powdered metals), treatment applications (e.g., humidification and gas scrubbing), coating applications (e.g., surface treatment, spray painting, and crop spraying), medicinal and printing applications, combustors (e.g., burners, furnaces, rockets, and gas turbines), and automotive diesel and gasoline engines. Generally, the liquid to be sprayed is injected

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

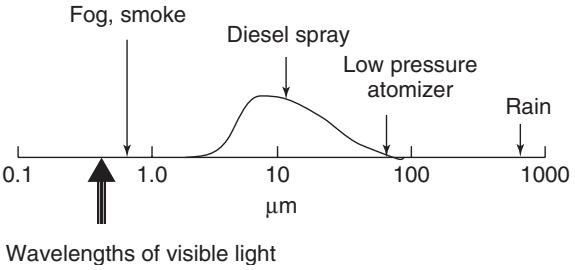


Figure 1. Drop sizes in various applications.

at a high velocity through a small orifice, and atomization is the process whereby the liquid is broken up into droplets (Lefebvre, 1989). Atomization influences spray vaporization rates because, in the form of small droplets, the total surface area of the injected liquid is increased. In automotive engine applications, fast vaporization is desirable for good fuel–air mixing (Heywood, 1988).

The trajectories of the injected spray drops formed by atomization are governed by the momentum of the drops, drag forces, and interactions between the drops and the surrounding gas, as discussed later. Practical atomizers generate sprays with a distribution of drop sizes, with average drop diameters ranging from a few micrometers to as large as 0.5 mm (Reitz, 2005). The smaller drops in the spray vaporize fast, if there is adequate mixing with surrounding air, to form combustible mixtures. Therefore, the rate of atomization and mixing ultimately controls the ignition process in combustion systems. On the other hand, large drops carry most of the mass and momentum of the injected liquid and these drops are able to penetrate into the high pressure gases in engine combustion chambers. Typical average drop sizes for broad classes of sprays are shown in Figure 1.

1.1.1 Pressure atomizers

Pressure atomizers are used to produce solid cone or hollow cone sprays, which can be classified as low or high pressure sprays. In solid cone (or full cone) sprays, the injected liquid is concentrated along the spray axis and, with high injection pressures, can provide high spray penetration, such as is required in direct-injection diesel engines where injection takes place near top dead center (TDC) into a high gas density environment (Figure 2a, b). Low pressure solid cone sprays may be used for early injections when chamber gas densities are low and where low spray penetration is favored to prevent spray wall impingement. In pressure atomizers, the chamber gas density, injection pressure, and nozzle hole diameter control the momentum exchange between the spray and the ambient gases and the resulting spray drop size (Hiroyasu, Kadota, and Arai 1980).

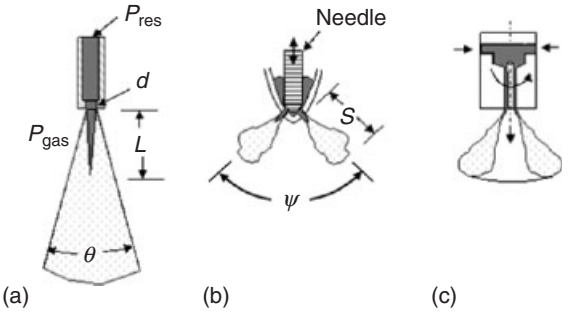


Figure 2. Pressure atomizers (a) plain orifice, solid cone (b) multihole, diesel, or asoline direct injection, and (c) swirl nozzle, hollow cone.

In hollow cone sprays, a swirling flow is setup within the nozzle passage, yielding a spray whose central axis region can be relatively free of drops, giving wide spray dispersal (Figure 2c). The air core vortex within the swirl chamber plays an important role in determining the thickness of the liquid sheet or film at the nozzle exit. This type of nozzle produces relatively coarse sprays and has been proposed for some gasoline direct-injection engines. However, precautions are needed for use in transient applications because the atomizer can dribble at start-up and shut down when the air core is not fully formed.

In the plain orifice nozzle (Figure 2a), the liquid emerges at a velocity related to the theoretical limit

$$U_f = \frac{C_d}{C_a} \cdot \sqrt{2 \cdot \frac{P_f - P_a}{\rho_f}} \tag{1}$$

where P is the pressure, ρ is the density, and the subscripts f and a signify fuel or charge-gas (ambient) density ("res" and "gas" in Figure 2), respectively. In addition, C_a is the orifice area contraction coefficient and C_d is the orifice discharge coefficient, the ratio of which lowers the actual injection velocity from the ideal Bernoulli velocity by accounting for flow losses due to friction and cavitation effects within the nozzle. An intact liquid core of length L may exist somewhat downstream of the nozzle exit (Figure 2a), depending on upstream cavitation and turbulence. As the spray emerges, it develops a conical shape with spray angle, θ. High pressure spray injection is intermittent and is required to start and stop quickly without dribble between injections. This is accomplished by means of a needle valve that is actuated by a cam and spring system in older mechanical "jerk" pump systems or by electromagnetic solenoids or piezo stacks in modern electronic injectors. These latter approaches permit the duration and injection pressure to be varied independently of each other and engine speed.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

1.1.2 Other atomizers

Other atomizer designs are found in combustion applications, as described by Lefebvre (1989). The most common atomizer types include the pressure atomizers described in Figure 2 and rotary, twin-fluid (air assist, air blast, and effervescent), flashing, electrostatic, vibratory, and ultrasonic atomizers. Group hole nozzles have also been proposed for internal combustion engine applications (Kim *et al*., 2009). In this case, each nozzle hole of the showerhead nozzle (Figure 2b) is replaced by a pair of smaller converging or diverging holes. In the converging nozzle case, the impinging jets create unstable liquid sheets that help break up the liquid to produce the sprays.

Air assist and air blast are twin-fluid atomizers that use airflows to aid the liquid breakup. However, they require a source of compressed air, and this introduces significant complexity and cost in automotive applications.

A common feature of the above-mentioned nozzle designs is that atomization is enhanced by reducing the liquid jet diameter or liquid sheet thickness and by arranging for a high relative velocity between the liquid and gas flows. For the common pressure atomizer, this can be achieved by increasing the injection pressure, and indeed, injection pressures exceeding 2000 bar are being introduced in modern diesel engines.

2 FUEL DELIVERY IN AUTOMOTIVE ENGINES

For automobile applications, specific fuel introduction technologies are most common. Figure 3 summarizes these different methods for injecting fuel, dependent on the type of engine. Fuel may be introduced and mixed with air before entering the engine cylinder, or it may be injected directly into the cylinder. The characteristics of fuel sprays produced by each of these technologies are reviewed in this section.

2.1 Fuel introduction systems

2.1.1 Carburetor

A carburetor blends fuel and air by creating a subatmospheric pressure environment to draw liquid fuel from the tank into the engine's intake system. As shown in the upper left schematic, intake air is directed through a venturi contraction, which increases the velocity and reduces local pressure, thereby creating a vacuum to pull fuel into the air stream. Typically, a single carburetor is used to mix fuel and air before dividing the prepared charge toward multiple cylinders, but multiple carburetors with separate intake manifolds are also utilized. Modern passenger cars

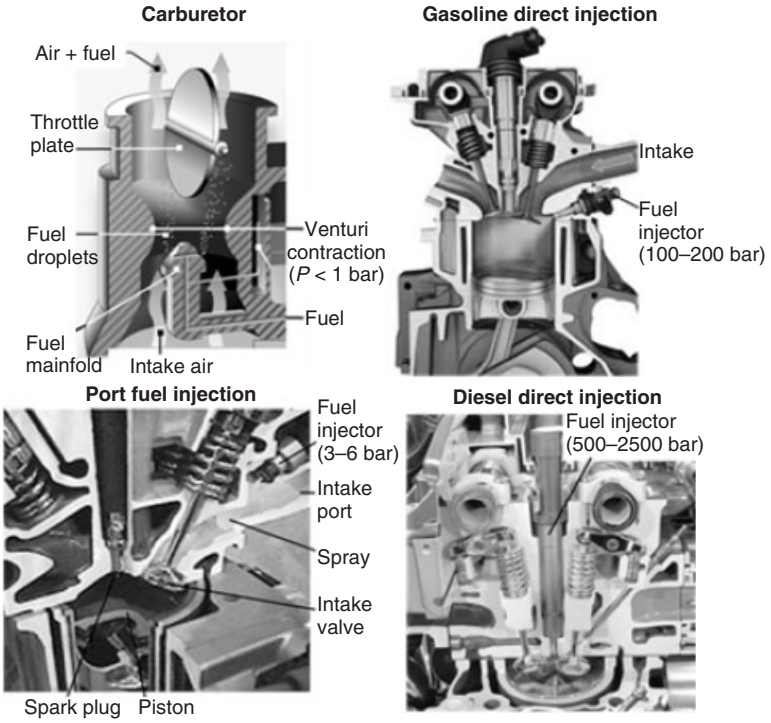


Figure 3. Schematics and cutaways of various fuel introduction systems representative of common automobile technology. Fuel injector location and other engine components are labeled, along with typical fuel injection absolute pressure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

no longer use carburetor fueling but carburetors are still very common in small utility engines.

When fuel is ejected from the fuel manifold, only a small portion of the fuel vaporizes upstream of the throttle plate, as typical for large, poorly atomized droplets. Large droplets with diameters in the range 25–100 μm or larger are typical (Heywood, 1988). Significant liquid fuel impacts the throttle plate and throttle-body walls, forming a liquid film (Heywood, 1988). The liquid film vaporizes as air continuously flows into the engine. Because the fueling system and conditions within the carburetor are poor for vaporization, high volatility fuels such as gasoline are required for this type of engine.

Fueling rates within the carburetor are adjusted automatically when the airflow rate increases or decreases as dictated by the throttle plate position. As the airflow rate increases, the velocity at the venturi contraction increases and the pressure decreases, thereby increasing the fueling rate. While this type of fueling adjustment is functional and simple because it requires no fuel pressurization (i.e., a fuel pump), it is imprecise over the large operating range of engine. Consequently, carburetors were replaced with pressurized, electronically controlled fuel injectors of type shown in Figure 2b or c. A single fuel injector positioned at the throttle body came to be known as a single-point, central-fueling, or throttle-body injector, as engine fueling is performed at a single input location. While the throttle-body injection system is not shown schematically in Figure 3, the PFI system, discussed in Section 2.1.2, bears similarities.

2.1.2 Port injection

PFI systems deliver fuel very close to the intake of each cylinder of an engine, as shown in Figure 3. The injector is modestly pressurized and controlled by an electronically activated needle valve to open and close at the proper timing for the cycle of that particular cylinder. The injection duration is varied depending on the engine power requirement, with typical injection durations ranging from 1.5 to 10 ms (Heywood, 1988). As one fuel injector is required for each cylinder, this system is also known as *multipoint fuel injection*.

Similar to carburetor fueling, the port injection fuel spray mixes with intake air before being ingested into the cylinder. The fuel spray is directed toward the valve of the intake port. Droplet atomization and vaporization is improved compared to a carburetor for several reasons. First, the spray shape is actively designed to mix with intake air and minimize wetting of the intake port. Droplets may even be ingested directly into the cylinder during the intake cycle, particularly when the intake valve is open, with in-cylinder droplets amounting to 10–20% of the fueling at some conditions (Peters, 1982). Second, higher fuel injection pressure forms smaller droplets that are easier to atomize. Third, the intake port and valve are warmer because of the proximity to the cylinder combustion chamber, which naturally vaporizes more fuel. Vaporization is also enhanced by the backflow of hot residuals into the intake (Heywood, 1988). Despite the improved vaporization, substantial fuel films are still formed along the intake port and valve and must be vaporized by the flow of incoming air.

2.1.3 Direct cylinder injection

Two different variants of injection technology to introduce fuel directly into the engine cylinder and combustion chamber are shown in Figure 3. Gasoline direct injection from the side of the cylinder is shown at the upper right. Diesel direct injection from the cylinder center is shown at the lower right. As with other technologies, variants for gasoline and diesel direct injection exist. For example, gasoline injection with centrally mounted orientation has been implemented; as has diesel injection from the side. Other variants may include indirect, prechamber injection, but these technologies have now generally been superseded by direct injection.

The design for a direct-injection gasoline nozzle is typically a 5–8 hole pattern with a somewhat narrow "umbrella" or included angle between sprays (ψ in Figure 2b). Gasoline injectors with an annular swirl pattern such as that shown in Figure 2c are also common. The optimum hole number and included angle is dependent on the injector location relative to the cylinder and to the timing of injection. As depicted in Figure 3, a gasoline spray may be injected early in the cycle with the piston near bottom dead center (BDC) and the intake valve open. The spray fills the combustion chamber as it penetrates across the chamber and mixes with in-cylinder gases. Although there is some distance between the injector and opposing cylinder wall, allowing vaporization, liquid droplets may impact the cylinder liner or the advancing piston. Wall wetting may be undesirable because of oil dilution and other concerns, demonstrating the care that must be taken when designing a direct-injection fueling system. The location of the fuel cloud relative to the spark plug is another consideration.

Gasoline direct-injection fuel pressures are higher than that of PFI systems. One reason for this is that the fuel pressure must exceed in-cylinder pressures that become elevated during the compression cycle. Higher injection pressure also promotes atomization and allows shorter

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

fueling intervals to deliver the same mass. In general, gasoline direct-injection systems will experience significant variation of the in-cylinder charge-gas properties because of the huge pressure and temperature variations experienced during intake, compression, exhaust, and so forth, as well as mixing with different exhaust gas residuals. The spray development and vaporization will strongly depend on the in-cylinder conditions and flow patterns during injection.

As opposed to gasoline injection systems where fuel is injected early in the cycle and later ignited with a spark plug, diesel injection timing typically occurs slightly before full compression (TDC). Because the piston is in close proximity to the fuel injector at TDC, diesel injectors are designed with an open included angle ($\psi \approx 140-150^{\circ}$) to direct each plume mainly in the radial direction. The number of spray plumes/nozzle holes varies, but 5–8 is typical. As the charge-gas pressure and temperature are quite high at TDC compression, evaporation of liquid fuel is enhanced and low volatility (high boiling point temperature) fuels such as diesel or biodiesel mix and evaporate without serious wall wetting and fuel film formation. Rapidly mixing fuel with dense charge gases is enhanced by very high fuel injection pressure, which Figure 3 shows may exceed 2000 bar. As depicted in Figure 1, diesel sprays are finely atomized with smaller droplet sizes. Mixing, vaporization, ignition, and completion of combustion occur in only a few crank angle degrees near TDC, also aided by the high injection pressure and high rate of fuel delivery. Injection durations may be as short as 0.2 ms, extending to approximately 3 ms. Multiple injection events during a single cycle are also widely used.

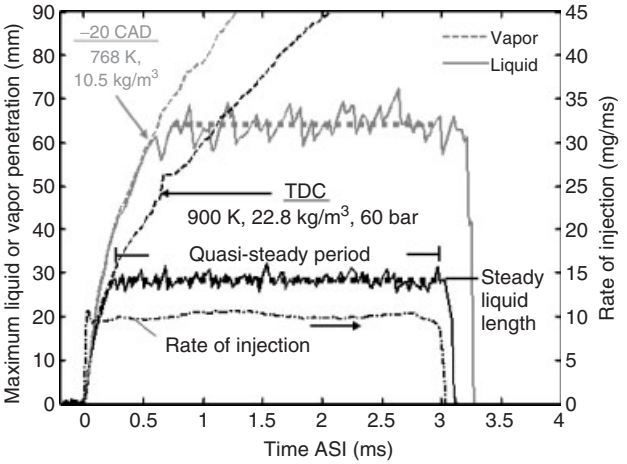


Figure 4. Measured vapor- and liquid-phase axial penetrations at near-TDC (black) and early-injection (gray) conditions. Rate of injection at right axis. Ambient: 0% O_2 at temperature and density given. Injector: 0.181 mm orifice diameter, 110 MPa pressure, and #2 diesel fuel. (Reproduced with permission from Pickett *et al.*, 2009. © L.M. Pickett, S. Kook and T.C. Williams, Sandia National Laboratories.)

2.2 Spray penetration fundamentals

Given that different injection systems deliver fuel into much different charge gas environments, either within the intake system or directly into the cylinder, it is helpful to understand how various injector and ambient conditions will affect spray penetration, mixing, and vaporization. A simple example of measurements depicting the effect of different injection timings, wherein the spray would mix with quiescent gases at different density and temperature, is given in Figure 4 for single-orifice sprays. The spray liquid- and vapor-phase spray tip penetrations (S, in Figure 2b) are shown at conditions typical for near-TDC diesel injection, as well as injection at 20° before TDC. For both conditions, the rate of injection has a top-hat shape (bottom curve) and the injection duration is the same. With increasing time after the start of injection (ASI), the liquid and vapor phases of the spray penetrate together. The spray penetrates more slowly into a high density environment. As hot gases are entrained into the spray, the liquid phase eventually reaches an axial position where the liquid is vaporized completely. The vapor phase continues to penetrate downstream, but the liquid phase fluctuates about a mean "quasi-steady" position throughout the duration of injection, as indicated in the figure. This maximum liquid penetration distance is defined as the quasi-steady liquid length (LL) or simply the LL. Note that the LL more than doubles, comparing TDC to −20 CAD injection timing.

When describing spray penetration, note the importance of distinguishing between liquid- and vapor-phase penetrations. The overall penetration will always be described by the vapor-phase penetration, and if upstream of the LL, the liquid- and vapor-phase penetrations are the same.

2.2.1 *Overall penetration and mixing*

With guidance from fundamental measurements such as that shown in Figure 4, empirical formulas and analytical solutions have been developed to predict overall penetration. Naber and Siebers (1996) derived an analytical solution, relating a nondimensional penetration time ($\widetilde{t}$) to a nondimensional axial penetration distance ($\widetilde{S} = \widetilde{x}$) as:

$$\widetilde{t} = \frac{\widetilde{S}}{2} + \frac{\widetilde{S}}{4} \cdot \sqrt{1 + 16 \cdot \widetilde{S}^2} + \frac{1}{16} \cdot \ln\left(4 \cdot \widetilde{S} + \sqrt{1 + 16 \cdot \widetilde{S}^2}\right) \quad (2)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

$$\tilde{t} = t \cdot \sqrt{\frac{\rho_a}{\rho_f}} \frac{a \cdot \tan(\theta/2)}{\sqrt{C_a} \cdot d_0} \cdot U_f \quad (3)$$

$$\tilde{S} = \tilde{x} = x \cdot \sqrt{\frac{\rho_a}{\rho_f}} \frac{a \cdot \tan(\theta/2)}{\sqrt{C_a} \cdot d_0} \quad (4)$$

where t is the time, x is the axial distance from the injector, d_0 is the orifice diameter, θ is the full spreading angle, and the constant a has a value of 0.66–0.75. In the near-field limit of a spray, Equation 2 has an $\tilde{S} = \tilde{t}$ dependency, and in the far-field limit, an $\tilde{S} = \tilde{t}^{0.5}$ dependency. Expectations for the effect of injector and charge-gas conditions follow from Equations 1–4. For example, penetration rate increases with the increasing fuel injection pressure or nozzle diameter and decreases with the increasing spreading angle or charge-gas density.

2.2.2 *Liquid-phase penetration*

Without considering atomization and breakup processes directly, an understanding of the penetration and mixing of the spray also can be used to explain the maximum liquid penetration. For example, Siebers (1999) showed that LL predictions were in good agreement with measurements when assuming mixing-limited vaporization. Estimates for mixing, or an average fuel–ambient mass ratio at a particular axial distance, come from Equations 2–4 as:

$$\frac{F}{A} = \frac{2}{\sqrt{1 + 16 \cdot \tilde{x}^2} - 1} \quad (5)$$

Then, assuming perfect mixing and liquid–vapor equilibrium, the fuel–ambient ratio $(F/A)_{liq}$ by mass where complete vaporization is expected is determined to be that when the enthalpy change in the ambient gas entrained into the spray matches the energy required to heat and vaporize the liquid fuel at the local equilibrium temperature. $(F/A)_{liq}$ depends on the charge-gas temperature and density, as well as the particular fuel properties and fuel temperature. Combining terms from Equations 4 and 5, and recognizing the need for a constant to account for local conditions at the center of the spray, the equation to predict the steady LL is (Siebers, 1999):

$$LL = 0.62 \cdot \sqrt{\frac{\rho_f}{\rho_a}} \cdot \frac{\sqrt{C_a} \cdot d_0}{\tan(\theta/2)} \sqrt{\left(\frac{2}{(F/A)_{liq}} + 1\right)^2 - 1} \quad (6)$$

Exploration of the dependency of Equation 6 on charge-gas conditions and fuel type provides insight about the different injection technologies utilized in automobile applications. Figure 5 shows the predicted maximum liquid penetration length if fuel from a single-orifice nozzle is injected into gas at certain crank angles before TDC. This calculation assumes that there is no wall impingement, that is, that the gas medium is infinite. As such, there is no interaction between neighboring spray plumes. The predictions shown are for single-orifice nozzle of 0.181-mm diameter, typical of production fuel injectors. The charge-gas conditions are calculated based on polytropic compression of gases initially at 340 K and 1 bar at BDC. The top axis shows the charge temperature and density at a given crank angle. Charge density is indicated because overall (vapor) penetration depends on density, rather than pressure (Naber and Siebers, 1996).

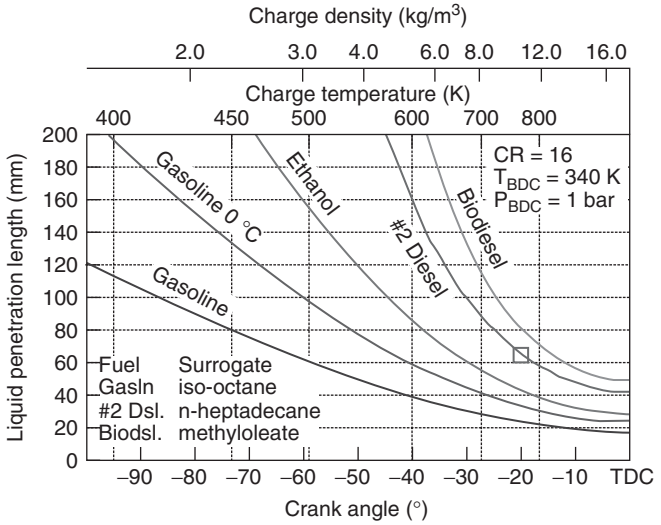


Figure 5. Predicted liquid penetration length when injecting various types of fuel at different timings before top dead center based on Equation 6 (Siebers, 1999) for single-orifice 0.181 mm nozzle at 95°C, unless indicated. Single-component surrogate fuels (lower left) are used to represent realistic fuels. Charge density and temperature at the time of compression are given on the top axis, resulting from bottom center conditions and compression ratio (CR) indicated at upper right.

Figure 5 also shows that the LL depends strongly on fuel type. Fuels with a high boiling point temperature, such as #2 diesel and biodiesel, have a longer liquid penetration than fuels with a low boiling point temperature, such as gasoline. In this calculation, the thermodynamic properties of single-component surrogate fuels are used to provide representative $(F/A)_{liq}$ for realistic fuels. For instance, *n*-heptadecane has a boiling point temperature that corresponds to the 90% distillation temperature point for #2 diesel, and methyl oleate is one of the major components for soy methyl ester biodiesel. The use of these surrogates has been justified by past comparison with experimental

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

data (Siebers, 1999). For example, the measured LL for the same conditions of Figure 4 is shown as a square symbol, falling right in line with the #2 diesel predictions in Figure 5.

The very strong effect of charge-gas temperature and density is also evident in Figure 5. Injecting #2 diesel and biodiesel before approximately −40 CAD results in LLs that exceed 100 mm, which is much larger than the bore radius of a typical heavy-duty diesel engine. Wall wetting, fuel film formation, and oil contamination are expected, particularly for biodiesel, as the predictions show that the LL of biodiesel is 15–50% higher than that of diesel with more significant differences at earlier injection timing. As such, one can appreciate why diesel fuels are typically injected only near TDC. On the other hand, the high volatility gasoline produces much shorter LLs, remaining well below 100 mm for injections as early as −90 CAD. At −90 CAD, where the expected charge temperature and density are much lower, approximately 410 K and 1.7 kg/m^3, respectively. Clearly, gasoline evaporates quite effectively, consistent with the practice for injection of gasoline even at BDC.

With reference to the gasoline results, we also show the effect of fuel temperature or the use of ethanol as an alternative to gasoline. Simulating a cold engine start at winter conditions, the gasoline temperature was changed from 95°C, a warm engine coolant temperature, to 0°C. The LL increases by 50–70%, suggesting a far greater likelihood for wall wetting at cold-start conditions. If significant fuel remains as a film along the wall and does not vaporize during compression and reach the spark plug, the engine will misfire during cold start. The predicted LLs for ethanol are also much higher than that of gasoline, by more than a factor of two. Curiously, the boiling point temperature is lower for ethanol compared to iso-octane, the gasoline surrogate fuel. However, the heat of vaporization is factor of 2.7 higher for ethanol than for iso-octane. Therefore, ethanol requires more entrainment of hot charge gases to reach the point of complete evaporation. As a result, ethanol has a longer LL than that of gasoline. Fuel and fuel temperature dependencies on LL obviously deserve attention for the injector and engine design.

While the perspective of overall and maximum liquid-phase penetration for a single spray is helpful because of the basic understanding it affords, actual injection systems coupled with gas flows that are highly variable and turbulent require deeper analysis. In the following sections, we describe spray modeling as applied to computational fluid dynamics (CFD) simulations.

3 THEORIES OF ATOMIZATION, DROP BREAKUP AND SPRAY PROCESSES

Engine sprays are multidimensional, multiphase, multicomponent, turbulent, compressible, reactive flows. The length scales of the smallest drops in the spray are of the order of a few micrometers (Figure 1) and their lifetimes are tens of microseconds. However, IC engine physical dimensions and timescales are of the order of centimeters and tens of milliseconds. This three to four orders of magnitude range disparity of scales necessitates that submodels be formulated for processes that are too small to be resolved in practical CFD. These processes include atomization and drop breakup, drop collision and coalescence, vaporization, drag and deformation, turbulent diffusion, and wall impingement, as discussed in this section.

The basic governing equations are the continuity (mass conservation), momentum (Navier–Stokes equations), energy, turbulence [normally the Reynolds averaged Navier–Stokes (RANS) $k-\epsilon$ equations] equations, and equations of state (ideal gas law). Assumptions include the use of the ideal gas law, Fick's law of mass diffusion, and others, depending on the details of the model. The discrete droplet model (DDM) (Dukowicz, 1980) method is widely used by researchers and also in commercial engine CFD codes to model the gas–liquid flow interactions. The gas phase is treated as continuous and described by the conventional Eulerian formulation. The liquid phase is treated as consisting of discrete particles and is described using the Lagrangian approach, in which liquid droplets are represented as assemblages of particles using the DDM method. The Lagrangian description of the particles avoids numerical diffusion and allows individual attributes, such as particle size, velocity, and position, to be statistically assigned to each particle. Mass, momentum, and heat transfer between the gas and liquid phases is accounted for with source terms that describe unresolved physical processes using submodels. Some important spray models are discussed next.

3.1 Governing equations

3.1.1 Gas phase

The motion of a gas-phase flow is described using the Navier–Stokes equations. The gas mixture consists of multiple species, and the continuity equation for species k is

$$\frac{\partial(\rho Y_k)}{\partial t} + \frac{\partial(\rho Y_k U_j)}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\rho D \frac{\partial Y_k}{\partial x_j}\right] + \dot{\omega}_k + \dot{\rho}_k^{\mathrm{s}} \quad (7)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where Y_k is the mass fraction of species k, ρ is the total mass density of the mixture, t is the time, x_j is the Cartesian position, and U_j is the fluid velocity. D is the single diffusion coefficient with the assumption of Fick's law diffusion. $\dot{\omega}_k$ and $\dot{\rho}_i^{\mathrm{s}}$ are source terms because of chemical reaction and spray evaporation/condensation, respectively. Summing Equation 7 over all species yields the continuity equation for the gas phase:

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho U_j)}{\partial x_j} = \dot{\rho}^{\mathrm{s}} \tag{8}$$

The momentum equations for the gas-phase are

$$\frac{\partial (\rho U_i)}{\partial t} + \frac{\partial (\rho U_i U_j)}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial \tau_{ij}}{\partial x_j} + F_i^{\mathrm{s}} + F_i^{\mathrm{b}} \tag{9}$$

where p is the pressure and τ is the viscous stress tensor:

$$\tau_{ij} = \mu \left(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i} - \frac{2}{3} \frac{\partial U_k}{\partial x_k} \delta_{ij} \right) \tag{10}$$

where μ is the dynamic viscosity, δ is the Kronecker delta, and F_i^{s} is a source term due to spray drop drag forces. F_i^{b} is the gravitation body force equal to ρg.

The energy conservation equation can be expressed in terms of sensible energy, e, which is the specific internal energy exclusive of chemical energy, as

$$\frac{\partial (\rho e)}{\partial t} + \frac{\partial (\rho e U_j)}{\partial x_j} = -p \frac{\partial U_j}{\partial x_j} + \frac{\partial J_j}{\partial x_j} + \dot{Q}^{\mathrm{s}} + \dot{Q}^{\mathrm{c}} \tag{11}$$

The heat flux vector J is the sum of contributions because of heat conduction and enthalpy diffusion:

$$J_j = -K \frac{\partial T}{\partial x_j} - \rho D \sum_{k=1}^{N_s} h_k \frac{\partial Y_k}{\partial x_j} \tag{12}$$

$\dot{Q}^{\mathrm{s}}$ and $\dot{Q}^{\mathrm{c}}$ are source terms because of spray and chemical reaction, respectively, and N_{s} is the total number of species. Assuming an ideal gas, the state equation is used to relate pressure and density:

$$p = \rho R T \sum_{k=1}^{N_s} \frac{Y_k}{W_k} \tag{13}$$

where W_k is the molecular mass of the kth species.

3.1.2 Liquid phase

The liquid-phase interactions with the gas-phase flow are complicated two-way couplings. For example, turbulence modifies the dispersed drops behavior, which, in turn, modifies turbulence produced by the drops. When the particle number density is sufficiently large, the effect of the particle–particle interactions (e.g., collisions and coalescences) must be considered.

The locally homogeneous flow (LHF) model neglects the slip effect between the liquid and gas phases, which are assumed to be in dynamic and thermodynamic equilibrium (Crowe, Sommerfeld, and Tsuji, 1998). This is the limiting case with infinitely small droplets. To consider the effects of the finite rate transport between the phases, various separated flow (SF) models have been proposed, including the DDM. In the DDM, the spray is represented by a finite number of parcels of identical drops. The motion and transport of these droplet groups are tracked and the effects of the liquid phase on the gas phase are considered by introducing appropriate spray source terms into the governing equations of the gas phase. This approach assumes that after primary breakup the droplets are small enough to be viewed as point sources. Thus, the spray dynamics can be described by the spray equation (Williams, 1958), where the spray is represented by a droplet distribution function, f:

$$f = f(\mathbf{V}_{\mathrm{d}}, r_{\mathrm{d}}, T_{\mathrm{d}}, y, \dot{y}; \mathbf{x}, t) \tag{14}$$

where $\mathbf{x}$, $\mathbf{V}_{\mathrm{d}}$, r_{d}, and T_{d} are the spatial location, velocity, equilibrium radius (the radius that the droplet would have if it was spherical), and the temperature of the droplet, respectively. y and $\dot{y}$ monitor the distortion of the drop from the spherical shape and its time rate of change. The droplet distribution function is defined such that $f \mathrm{d}\mathbf{V}_{\mathrm{d}} \mathrm{d}r_{\mathrm{d}} \mathrm{d}T_{\mathrm{d}} \mathrm{d}y \mathrm{d}\dot{y}$ is the probable number of droplets per unit volume at position $\mathbf{x}$ and time t with velocity in the interval $(\mathbf{V}, \mathbf{V} + \mathrm{d}\mathbf{V})$, radii in the interval $(r_{\mathrm{d}}, r_{\mathrm{d}} + \mathrm{d}r_{\mathrm{d}})$, temperature in the interval $(T_{\mathrm{d}}, T_{\mathrm{d}} + \mathrm{d}T_{\mathrm{d}})$, and distortion parameters in the intervals $(y, y + \mathrm{d}y)$ and $(\dot{y}, \dot{y} + \mathrm{d}\dot{y})$. The first moment of f is the number density of the droplets:

$$n = \int f \mathrm{d}\mathbf{V}_{\mathrm{d}} \mathrm{d}r_{\mathrm{d}} \mathrm{d}T_{\mathrm{d}} \mathrm{d}y \mathrm{d}\dot{y} \tag{15}$$

The second moment about radius r_{d} is the liquid volume fraction θ or liquid macroscopic density ρ'_1:

$$\theta = \int \frac{4}{3} \pi r^3 f \mathrm{d}\mathbf{V}_{\mathrm{d}} \mathrm{d}r_{\mathrm{d}} \mathrm{d}T_{\mathrm{d}} \mathrm{d}y \mathrm{d}\dot{y}$$

$$\rho'_1 = \int \frac{4}{3} \pi r^3 \rho_{\mathrm{d}} f \mathrm{d}\mathbf{V}_{\mathrm{d}} \mathrm{d}r_{\mathrm{d}} \mathrm{d}T_{\mathrm{d}} \mathrm{d}y \mathrm{d}\dot{y} \tag{16}$$

The time evolution of f is obtained by solving the spray equation:

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

$$\frac{\partial f}{\partial t} + \nabla_{\mathbf{x}} \cdot (f\mathbf{V}_{\mathrm{d}}) + \nabla_{\mathbf{V}} \cdot (f\mathbf{F}) + \frac{\partial}{\partial r_{\mathrm{d}}}(f\dot{r}_{\mathrm{d}}) + \frac{\partial}{\partial T_{\mathrm{d}}}(f\dot{T}_{\mathrm{d}}) + \frac{\partial}{\partial y}(f\dot{y}) + \frac{\partial}{\partial \dot{y}}(f\ddot{y}) = \dot{f}_{\mathrm{coll}} + \dot{f}_{\mathrm{bu}} \tag{17}$$

where $\mathbf{F}$, $\dot{r}_{\mathrm{d}}$, $\dot{T}_{\mathrm{d}}$, and $\ddot{y}$ are the time rates of change of the drop's velocity, radius, temperature, and oscillation velocity $\dot{y}$, respectively. The terms on the right-hand side account for collision and coalescence and drop breakup effects on the distribution function. With f solved for, the effects of the liquid phase on the gas phase are modeled with the source terms in the conservation equations (Equations 8, 9, and 11), namely,

$$\dot{\rho}^{\mathrm{s}} = -\int 4\pi r_{\mathrm{d}}^2 \dot{r}_{\mathrm{d}} \rho_{\mathrm{d}} f \, \mathrm{d}\mathbf{V}_{\mathrm{d}} \mathrm{d}r_{\mathrm{d}} \mathrm{d}T_{\mathrm{d}} \mathrm{d}y \mathrm{d}\ddot{y} \tag{18}$$

$$\mathbf{F}^{\mathrm{s}} = -\int \left(\frac{4}{3}\pi r_{\mathrm{d}}^3 \mathbf{F}' + 4\pi r_{\mathrm{d}}^2 \dot{r}_{\mathrm{d}} \mathbf{V}_{\mathrm{d}} \right) \rho_{\mathrm{d}} f \, \mathrm{d}\mathbf{V}_{\mathrm{d}} \mathrm{d}r_{\mathrm{d}} \mathrm{d}T_{\mathrm{d}} \mathrm{d}y \mathrm{d}\ddot{y} \tag{19}$$

$$\dot{Q}^{\mathrm{s}} = -\int \left\{ \frac{4}{3}\pi r_{\mathrm{d}}^3 \left[C_{\mathrm{p,l}} \dot{T}_{\mathrm{d}} + \mathbf{F}' \cdot \left(\mathbf{V}_{\mathrm{d}} - \widetilde{\mathbf{U}} - \mathbf{u}' \right) \right] + 4\pi r_{\mathrm{d}}^2 \dot{r}_{\mathrm{d}} \left[e_{\mathrm{l}}\left(T_{\mathrm{d}}\right) + \frac{1}{2}(\mathbf{V}_{\mathrm{d}} - \widetilde{\mathbf{U}})^2 \right] \right\} \times \rho_{\mathrm{d}} f \, \mathrm{d}\mathbf{V}_{\mathrm{d}} \mathrm{d}r_{\mathrm{d}} \mathrm{d}T_{\mathrm{d}} \mathrm{d}y \mathrm{d}\ddot{y} \tag{20}$$

where $\mathbf{F}' = \mathbf{F} - \mathbf{g}$.

3.2 Atomization models

In the primary breakup process, the liquid exiting the injector nozzle disintegrates into filaments and drops because of interaction with the surrounding gas, as depicted in Figure 6. In spite of decades of research, the details of the breakup process are not well understood, partially due to the fact that the optically dense region near the nozzle exit cannot be resolved easily experimentally. Instabilities on the liquid–gas interface are thought to be a major driving force for the breakup process. Recent high resolution X-ray spray visualization experiments and direct numerical simulations (DNS)s (Section 4) provide some support for the existence of instabilities that lead to surface waves and their role in the breakup process.

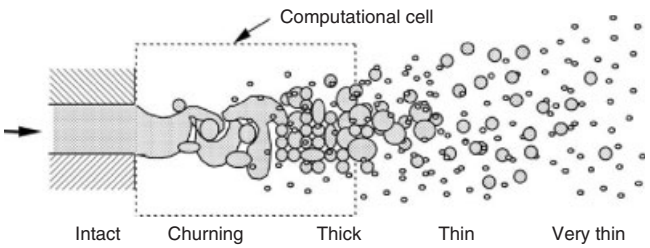


Figure 6. Schematic diagram of jet breakup regimes.

In practical engine CFD simulations, the atomization process occurs on length scales that are below the scale of resolution of the computational mesh, as depicted in Figure 6. In this case, initial conditions for the drop distribution function parameters in the DDM at the injector nozzle exit must be supplied using submodels. The necessary parameters include the injection velocity and initial drop sizes and trajectories. The injection velocity or injection rate shape can be determined from detailed injector flow dynamic simulations that account for needle movement and cavitation processes (Lee and Reitz, 2010). However, in engine codes, simpler phenomenological nozzle flow models such as that of Sarre, Kong, and Reitz (1999) are often used with a measured injection rate shape. In this model, the nozzle discharge coefficient ratio in Equation 1 is modeled as

$$C_{\mathrm{d}}/C_{\alpha} = C_{\mathrm{V}} = \left(K + \mathrm{frict}\ \frac{L}{D} + 1 \right)^{1/2} \tag{21}$$

where the friction factor frict and inlet loss coefficient K are determined from tabulated data.

3.2.1 Primary breakup

The breakup process has been modeled using wave instability models, such as the hybrid Kelvin–Helmholtz (KH) and Rayleigh–Taylor (RT) models of Beale and Reitz (1999). The KH model is based on a linearized analysis of the stability of the surface of a cylindrical liquid jet or sheet to perturbations, η, as depicted in Figure 7.

These linear stability analyses lead to dispersion equations that include the physical and dynamical parameters of the liquid jet or sheet and the surrounding gas, and which relate the growth rate of unstable waves to their wavelength. Curve fits of numerical solutions have been generated by Reitz (1988) for the maximum growth rate, Λ_{KH} (Λ in Figure 7), and the corresponding wavelength, Ω_{KH} (Ω in Figure 7), for jets as

$$\Lambda_{\mathrm{KH}} = \frac{9.02 r_{\mathrm{d}} (1 + 0.45\mathrm{Z}^{0.5})(1 + 0.4\mathrm{T}^{0.7})}{(1 + 0.865\mathrm{We}_{\mathrm{air}}^{1.67})^{0.6}} \tag{22}$$

$$\Omega_{\mathrm{KH}} = \frac{0.34 + 0.38\mathrm{We}_{\mathrm{air}}^{1.5}}{(1 + \mathrm{Z})(1 + 1.4\mathrm{T}^{0.6})} \sqrt{\frac{\sigma}{\rho_{\mathrm{l}} r_{\mathrm{d}}^3}} \tag{23}$$

where the gas Weber number is $\mathrm{We}_{\mathrm{air}} = \rho U_{\mathrm{rel}}^2 r_{\mathrm{d}}/\sigma$, and the Ohnesorge number is $\mathrm{Z} = \mathrm{We}_{\mathrm{d}}^{1/2}\mathrm{Re}_{\mathrm{l}}^{-1}$. U_{rel} is the relative

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

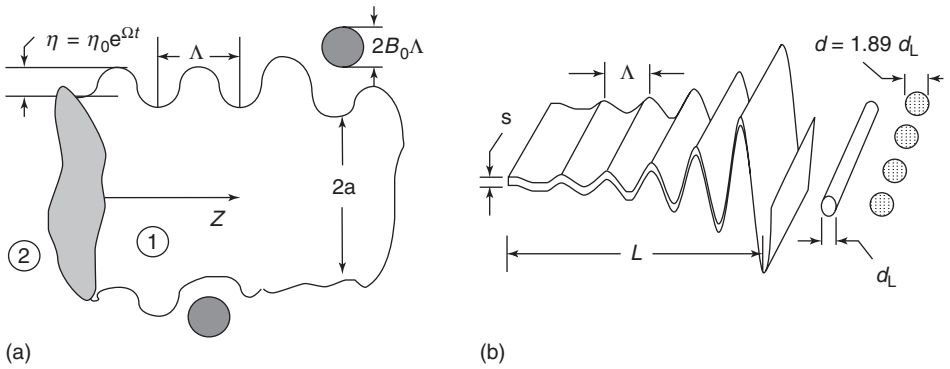


Figure 7. Schematic diagrams depicting surface wave-induced breakup of (a) liquid jets or "blobs" and (b) liquid sheets.

velocity between the gas and the droplet and σ is the surface tension. The droplet Weber number is $We_d = \rho_l U_{rel}^2 r_d/\sigma$, the liquid Reynolds number is $Re_l = \rho_l |\mathbf{U}_{rel}|^2 r_d/\mu_{l,}$ and T is the Taylor number $T = We_{air}^{1/2} Z$.

The KH model is implemented by postulating that a parent drop with radius r_d breaks up from a liquid jet of diameter $d_L = 2a$ to form new droplets with radius $r_{d,c}$ determined by (Reitz, 1988):

$$r_{d,c} = B_0 \Lambda_{KH} \tag{24}$$

as depicted in Figure 7a, where B_0 is a model constant. The rate of change in the radius of the parent droplet or liquid blob is described by

$$\frac{dr_d}{dt} = \frac{r_d - r_{d,c}}{\tau_{KH}} \tag{25}$$

where the breakup timescale τ_{KH} is

$$\tau_{KH} = \frac{3.726 B_1 r_d}{\Omega_{KH} \Lambda_{KH}} \tag{26}$$

where B_1 is a second model constant.

The RT model describes the growth of instabilities associated with deceleration of the injected drops. The frequency of the fastest growing wave, Ω_{RT} (Ω in Figure 7), and its corresponding wavelength, Λ_{RT} (Λ in Figure 7), are given by (Beale and Reitz, 1999)

$$\Omega_{RT} = \sqrt{\frac{2}{3\sqrt{3\sigma}} \frac{[-a_d(\rho_l - \rho)]^{1.5}}{\rho_l + \rho}} \tag{27}$$

$$\Lambda_{RT} = 2\pi \Omega_{RT} \sqrt{\frac{3\sigma}{-a_d(\rho_l - \rho)}} \tag{28}$$

where a_d is the droplet acceleration. When the predicted wavelength is less than the droplet diameter, the RT waves are assumed to grow on the surface of the droplet and the wave growth time is tracked. When it reaches the RT breakup timescale $\tau_{RT} = C_\tau/\Omega_{RT}$, the drop is assumed to break up. The radii of the new droplets are computed using

$$r_{d,c} = 2 C_{RT} \Lambda_{RT} \tag{29}$$

where C_τ and C_{RT} are additional model constants. Beale and Reitz (1999) applied the RT model to the decelerating drops beyond a liquid breakup length that is proportional to the distance, $L \sim t_{BK} U_{mean}$ (Figure 2).

Other liquid breakup and atomization models have been proposed that consider drop deformation processes such as the Taylor analogy breakup (TAB) model of O'Rourke and Amsden (1987), as described in Section 3.2.2.

3.2.2 Droplet breakup and drag

When the relative velocity between the drop and the surrounding gas is high, the drop deforms, which also influences its drag coefficient. The drop drag coefficient was related empirically by Liu, Mather, and Reitz (1993) to the magnitude of the drop deformation using

$$C_d = C_{d,sphere}(1 + 2.632y) \tag{30}$$

where $C_{d,sphere}$ is the drag coefficient of a spherical drop and y is the drop distortion parameter that is proportional to the displacement of the drop's surface from its equilibrium position divided by the drop's radius. $y = 0$ represents the lower limit of a sphere and $y = 1$ the upper limit of a flat disk, as shown in Figure 8a.

y can be computed using the TAB model of O'Rourke and Amsden (1987) or the similar droplet deformation and breakup (DDB) model of Ibrahim, Yang, and Przekwas (1993). The TAB model assumes that a liquid drop is analogous to a spring–mass system (analogy of Taylor,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

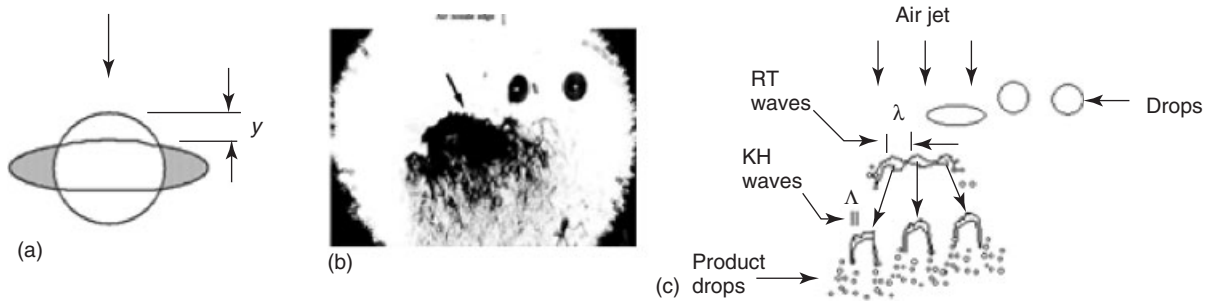


Figure 8. Schematic diagram of (a) drop deformation leading to breakup and (b) photograph of 170 mm diesel fuel drops deflected and broken up by 250 m/s air jet. Monodisperse drop stream enters from right, and the air jet is oriented vertically downward. RT waves are indicated by arrow. (c) Proposed KH–RT drop breakup mechanisms. (Reprinted from Atomization and Sprays, Vol. 6, Hwang, S.S., Liu, Z., and Reitz, R.D., Breakup Mechanisms and Drag Coefficients of High Speed Vaporizing Liquid Drops, pp. 353–376, Copyright (1996), with permission from Begell House, Inc.)

1963), and drop breakup occurs when the amplitude of the drop oscillation $y = 1$. The drop distortion parameter is given as

$$\ddot{y} = -5\frac{\mu_l}{\rho_l}\frac{\dot{y}}{r_d^2} - \frac{8\sigma y}{\rho_l r_d^3} + \frac{2}{3}\frac{\rho}{\rho_l}\frac{U_{rel}^2}{r_d^2} \tag{31}$$

where U_{rel} is the relative velocity between the gas and the drop. As can be seen in Figure 8b, in the case of high speed drop breakup, the flattened drop disintegrates into a shower of drops. Hwang, Liu, and Reitz (1996) proposed that the breakup is due to KH and RT instabilities (Figure 8c), as discussed for jet breakup earlier.

3.3 Models of other spray processes

3.3.1 Turbulent dispersion models

In most DDMs, the drop moves according to its velocity, $\frac{d\mathbf{x}_d}{dt} = \mathbf{V}_d$, and its acceleration is

$$\frac{d\mathbf{V}_d}{dt} = \frac{\mathbf{F}_d}{m_d} = D_d(\mathbf{U} - \mathbf{V}_d) + \mathbf{g} \tag{32}$$

where the drag function $D_d = \frac{3}{8}\frac{\rho|\mathbf{U}-\mathbf{V}_d|}{\rho_d r_d}C_d(\mathrm{Re}_d)$ and the drag coefficient are given by Amsden *et al*. (1985) as

$$C_d = \begin{cases} 24\mathrm{Re}_d^{-1}\left(1 + \mathrm{Re}_d^{2/3}/6\right), & \mathrm{Re}_d < 1000 \\ 0.424, & \mathrm{Re}_d \geq 1000 \end{cases} \tag{33}$$

The drop Reynolds number is $\mathrm{Re}_d = \frac{2\rho r_d|\mathbf{U}-\mathbf{V}_d|}{\mu_{air}(\widehat{T})}$, where the average temperature $\widehat{T} = (T + 2T_d)/3$, and the gas instantaneous velocity $\mathbf{U}$ seen by the droplet is $\mathbf{U} = \widetilde{\mathbf{U}} + \mathbf{u}''_d$, where $\widetilde{\mathbf{U}}$ is the mean gas velocity.

Turbulence effects are modeled by a turbulent dispersion model, which provides the gas turbulent velocity, $\mathbf{u}''_d$. Each component of $\mathbf{u}''_d$ is randomly chosen from a Gaussian distribution with standard deviation, $\sigma = \sqrt{2k/3}$. $\mathbf{u}''_d$ is chosen once every turbulence correlation time t_{turb}, which is the minimum of an eddy breakup time and a time for the droplet to penetrate an eddy (Gosman and Ioannides, 1981).

3.3.2 Drop collision and coalescence models

The seminal drop collision model of O'Rourke (1981) considers collisions between drop parcels located in the same computational cell. The collision frequency of a pair of particles is

$$\nu = \frac{N_{mp}}{V_{cell}}\pi(r_{lp} + r_{mp})^2|\mathbf{V}_{lp} - \mathbf{V}_{mp}| \tag{34}$$

where the subscripts mp and lp denote the more populous and less populous droplet parcels, respectively, N is the number of drops in the parcel, and V_{cell} is the volume of the numerical cell. The probability that the more numerous particles undergo n collisions with the smaller number of particles is assumed to follow a Poisson distribution, and the probability of no collision is $P_0 = e^{-\nu\Delta t}$, where Δt is the computational time step. A random variable is selected to determine whether the collision event will occur or not. Another independent random number YY in the interval (0,1) is used to calculate the collision impact parameter $b = YY^{1/2}(r_{lp} + r_{mp})$, which determines the collision outcome by comparing it with the critical impact parameter b_{cr}

$$b_{cr} = (r_{lp} + r_{mp})\{\min[1, 2.4(\gamma^3 - 2.4\gamma^2 + 2.7\gamma)/\mathrm{We}_{lp}]\}^{1/2} \tag{35}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

with $\gamma = r_S/r_L$, and the Weber number of the larger drop parcel $We_{lp} = \rho_d |\mathbf{V}_{lp} - \mathbf{V}_{mp}| r_{lp}/\alpha(\overline{T}_d)$. For the Weber number, the mean temperature $\overline{T}_d = \frac{r_{mp}^3 T_{d,lp} + r_{mp}^3 T_{d,mp}}{r_{lp}^3 + r_{mp}^3}$ and the liquid surface tension coefficient $\alpha(\overline{T}_d) = \frac{\overline{T}_d - T_{cr}}{T_0 - T_{cr}}\alpha_0$.

When $b < b_{cr}$, coalescence occurs, otherwise each collision is assumed to be a grazing collision. In the case of a coalescence collision, the droplet number of the new parcel is adjusted and the properties of the new parcel are determined from mass and momentum conservation laws. Improvements have been made to the O'Rourke model to account for other collisions outcomes, such as satellite drop formation and drop shattering by Munnannur and Reitz (2007).

As the O'Rourke model only considers collision events between groups of identical droplets (parcel) that are located in the same computational cell, the outcome of the model depends on mesh size. To remove mesh dependency, Munnannur and Reitz (2009) proposed a radius-of-influence (ROI) collision model. The ROI model considers potential collision between every pair of droplet parcels whose distance $D_{lp,mp}$ is smaller than the maximum of their influence radii, R_{lp} and R_{mp}, where R_p is the distance of the colliding drops, and l and m can travel at their current velocities in one numerical time step:

$$D_{lp,mp} \leq \max(R_{lp}, R_{mp}) \tag{36}$$

The collision frequency is then computed as:

$$\nu = \frac{N_{mp}}{V_{col}} \pi (r_{lp} + r_{mp})^2 |\mathbf{V}_{lp} - \mathbf{V}_{mp}| \tag{37}$$

The collision volume V_{col} is based on the radii of influence:

$$V_{col} = \frac{4}{3}\pi (R_{lp} + R_{mp})^3 \tag{38}$$

Thus, the influence of mesh topology is removed and the collision events only depend on the droplet parcel distribution in space. Other elements of the ROI model are similar to those of the O'Rourke model.

3.3.3 *Drop vaporization*

Assuming a single composition and homogeneous distribution of temperature inside the droplet, the rate of droplet radius change due to evaporation is given by the Frossling correlation (Faeth, 1983):

$$r_d = \frac{dr_d}{dt} = -\frac{\rho D}{2\rho_d r_d} B_m Sh_d \tag{39}$$

where Sh_d is the Sherwood number for mass transfer:

$$Sh_d = (2.0 + 0.6\ Re_d^{1/2} Sc_d^{1/3})\ \frac{\ln(1 + B_m)}{B_m} \tag{40}$$

with Schmidt number $Sc_d = \frac{\mu_{air}(T)}{\rho D}$. D is the fuel vapor diffusivity and $B_m = \frac{Y_{F,s} - Y_{F,\infty}}{1 - Y_{F,s}}$ is the Spalding mass transfer number, with $Y_{F,s}$ and $Y_{F,\infty}$ the fuel vapor mass fraction at the droplet's surface and at the outer boundary of the film surrounding the droplet. The surface mass fraction $Y_{F,s}$ is computed from the Clausius–Clapeyron equation:

$$Y_{F,s}(T_d) = \frac{1}{1 + \frac{W_{air}}{W_F}\left[\frac{p}{p_v(T_d)} - 1\right]} \tag{41}$$

where W_F is the molecular weight of fuel vapor, and W_{air} is the local average molecular weight of all species except fuel vapor. p is the local pressure and p_v is the equilibrium fuel vapor pressure.

The rate of droplet temperature change is determined from the energy equation:

$$\frac{4}{3}\rho_d \pi r_d^3 C_l \dot{T}_d - 4\rho_d \pi r_d^2 \dot{r}_d L(T_d) = 4\pi r_d^2 \dot{Q}_d \tag{42}$$

where C_l is the liquid specific heat, $L(T_d)$ is the latent heat of vaporization, and $\dot{Q}_d$ is the rate of heat conduction to the droplet surface per unit area and is computed using the Ranz–Marshall correlation (Faeth, 1983):

$$\dot{Q}_d = \frac{k_{air}\ (\widehat{T})\ (T - T_d)}{2r_d}\ Nu_d \tag{43}$$

where the Nusselt number Nu_d is

$$Nu_d = \left(2.0 + 0.6 Re_d^{1/2} Pr_d^{1/3}\right) \frac{\ln(1 + B_m)}{B_m} \tag{44}$$

and the Prandtl number $Pr_d = \frac{\mu_{air} C_p}{k_{air}}$, C_p is the local specific heat at constant pressure, and k_{air} is the thermal conductivity coefficient of air.

Ra and Reitz (2009) have developed a more accurate model that considers transient heat transfer inside the droplet. Instead of assuming a uniform droplet temperature, the droplet surface temperature $T_{d,s}$ is computed from a heat and mass transfer balance at the interface between the droplet and surrounded gas when the droplet size is larger than a preset critical value. The energy balance at the interface is written as

$$4\rho_d \pi r_d^2 \dot{r}_d L(T_{d,s}) = 4\pi r_d^2 (Q_i + Q_d) \tag{45}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where Q_i is the energy flux from inside the droplet to the surface. It is modeled as a convective heat transfer process with internal circulation considered and written as

$$Q_i = \frac{k_l}{\delta_e}(T_d - T_{d,s}) \quad (46)$$

where k_l is the liquid thermal conductivity. δ_e is the unsteady equivalent thickness of the thermal boundary layer and calculated from the effective thermal diffusivity:

$$\delta_e = \sqrt{\pi \alpha_{eff} t} = \sqrt{\pi \chi \alpha_l t} \quad (47)$$

where $\chi = 1.86 + 0.86 \tan h[2.225 \log_{10}(Pe_l/30)]$ and Pe_l is the Peclet number of the droplet. Q_d is also computed from the droplet surface temperature $T_{d,s}$ from

$$Q_d = \frac{k_{air}\ (\widehat{T})\ (T - T_{d,s})}{2r_d}\ Nu_d \quad (48)$$

As the effective heat transfer coefficient for the outer heat flux is coupled with the vaporization rate, the surface temperature of the droplet must be determined by solving the two equations iteratively.

Other evaporation models have been proposed that consider multicomponent fuel effects. These include continuous multicomponent models (Lippert and Reitz, 1997), discrete multicomponent models (Ra and Reitz, 2009), and discrete/continuous multicomponent models (Yang and Reitz, 2010), which have been developed and applied to IC engine simulations.

3.3.4 Spray/wall impingement

Spray wall impingement is important for both PFI and DI engines. Fuel wall films in the intake ports can cause an undesirable fuel delivery delay and an associated fuel metering error in PFI engines. In wall-guided *direct-injection-spark-ignition* (DISI) engines, the fuel is directly injected into a specially designed piston bowl and the spray impingement is used to generate an optimal stratified mixture. In diesel engines, spray wall impingement may lead to unacceptable unburned hydrocarbon (UHC) and/or soot emission, as well as low fuel efficiency. Therefore, an accurate spray/wall interaction model is important in engine simulations.

The outcome of a drop impingement event depends on the properties of the drop, the wall surface roughness conditions, and the details of the gas boundary layer in the near-wall region. If the wall temperature T_w is less than the liquid boiling temperature T_B, a collision of a drop on the solid surface may result in sticking, bouncing, spreading, or splashing, as described by Bai and Gosman (1995). The stick regime occurs when the impact energy is low and the wall temperature is below the pure adhesion temperature T_{pa}. The impinging drop adheres to the solid surface or coalesces with any liquid film existing on the surface. As the Weber number is increased, drop rebound can occur from the wall or liquid film, and the velocity and direction of the rebounding drop is often determined from experimental and analytical correlations (Naber and Reitz, 1988).

The spreading regime occurs at higher incident drop Weber numbers. Here, the drop spreads on the wall surface for a dry wall or merges with the liquid wall film on impact for a wet wall. When a train of drops impacts the wall surface, the time between multidrop impacts, or the impact frequency, must be considered. A splashing regime occurs at higher incident Weber numbers because of the development of a crown instability, which leads to secondary atomization of the impinging drop and/or wall film. The criterion for splash is

$$E^2_{splash} = \frac{We}{\min\ (H_f/d_d, 1) + Re_d^{-1/2}} > E^2_{splask,\ crit} \quad (49)$$

where H_f is the thickness of the liquid film and the drop Reynolds number is based on the velocity normal to the surface and $E^2_{splash,crit} = 3330.0$ (O'Rourke and Amsden, 2000). Liquid films can also be modeled using a two-dimensional thin film assumption that includes consideration of fuel vaporization (Stanton and Rutland, 1998).

3.4 Modeling applications

Two applications are briefly discussed as examples of current CFD spray modeling capabilities based on the spray models discussed in the previous sections. The first considers steady vaporizing, noncombusting sprays injected in a constant-volume chamber. Simulations that use coarse meshes and large time steps are of interest in applications to reduce the computer time needed for engine simulations. However, coarse meshes can result in inaccurate predictions of mass, momentum, and energy transfer between the spray drops and the combustion chamber gas, or poor prediction of droplet breakup and collision and coalescence processes. Accordingly, the spray models discussed earlier have been proposed to address these deficiencies, including use of an unsteady gas jet model to improve momentum transfer predictions in underresolved regions of the spray near the injector nozzle (Figure 6), a vapor particle model to minimize numerical diffusion effects, and a radius of influence drop collision model to ensure consistent collision computations on different meshes (Shi, Ge, and Reitz 2011;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

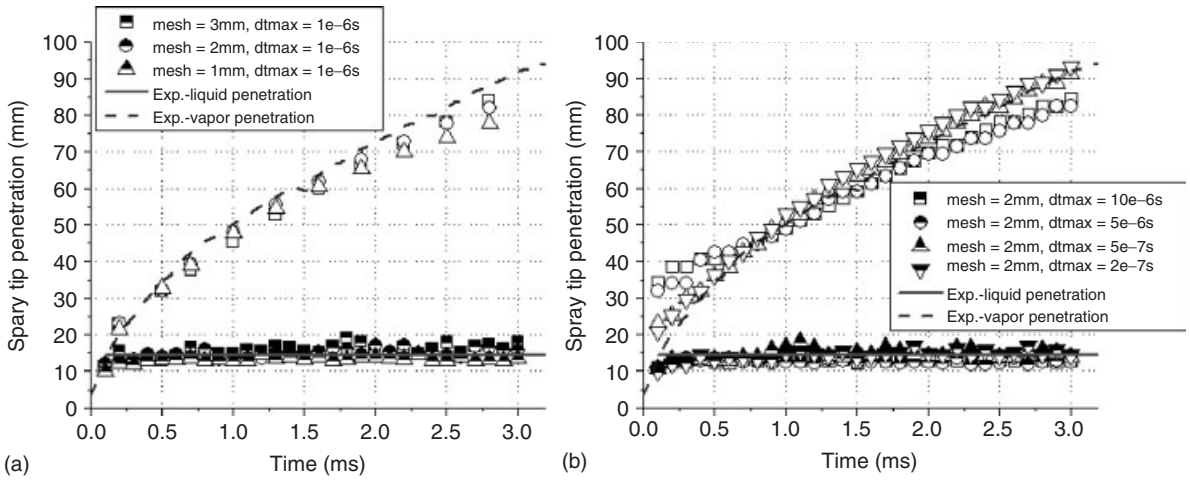


Figure 9. (a) Mesh and (b) time step independency of predicted vapor and liquid penetrations. Experimental data of Naber and Siebers (1996) and Pickett (2007). Diesel fuel injection, nozzle diameter 257 μm, injection pressure 1370 bar, gas temperature 1000 K, and gas density 58.6 kg/m^3.

http://www.springer.com/engineering/mechanical+eng/book/978-0-85729-618-4).

3.4.1 Spray penetration predictions

As discussed in Figure 4, spray penetration is an important process in direct injection engines, and thus accurate predictions are critical for CFD modeling. Wang, Ge, and Reitz (2010) combined the above models with KH–RT breakup models to improve the consistency of drop breakup predictions. A modified mean collision time model was also proposed to reduce the time step dependency of droplet collision predictions. The models were implemented into the KIVA CFD code (Amsden, 1997), and the results were found to demonstrate independence with respect to both mesh sizes and time steps, as shown in Figure 9. The vapor penetration and LL predictions agree well with the experiments of Naber and Siebers (1996) and Pickett (2007), even when various numerical grids are used whose mesh sizes vary by a factor of 3 (order of magnitude in cell volume), and with almost two orders of magnitude variations in time steps.

3.4.2 Engine fuel–air mixing predictions

A second example application of spray modeling in CFD simulations considers the single-cylinder optical light-duty diesel engine experiments simulated by Dempsey *et al*. (2012). The engine and injector geometry are summarized in Table 1, and a schematic of the engine is shown in Figure 10. A production, high pressure, solenoid-driven, common-rail Bosch CRIP2.2 fuel injector

Table 1. Engine and injector specifications for Figure 12.

Engine geometry

Bore/stroke (mm)	82.0/90.4
Displacement (cc)	4.77
Compression ratio	16.7
Squish height (mm)	0.88

Injector geometry
Bosch common rail CRIP2.2

Sac volume (mm^3)	0.23
Number of holes	7
Hole diameter (mm)	0.14
Included angle (deg)	149
Tip protrusion (mm)	2.1
Hole protrusion (mm)	0.3

	Baseline	*Current*
Engine speed (rpm)	1500	1500
Intake charge (Mole fraction)	O_2: 0.10 N_2: 0.81 CO_2: 0.09	N_2: 1.0
Intake temperature (K)	372	300
Intake pressure (bar)	1.5	1.5
Injection fuel (mg)	8.8	8.8
Fuel rail pressure (bar)	860	860
Global equivalence ratio	0.3	N.A.
Start of injection (° ATDC)*	–23.6	–23.6
Injection duration (° CA)	4.5	4.5
Swirl ratio	2.2	2.2
Motored TDC density (kg/m^3)	20.9	21.1
Motored TDC Temperature (K)	909	908

*Actual start of fuel injection, which was optically confirmed in the engine.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

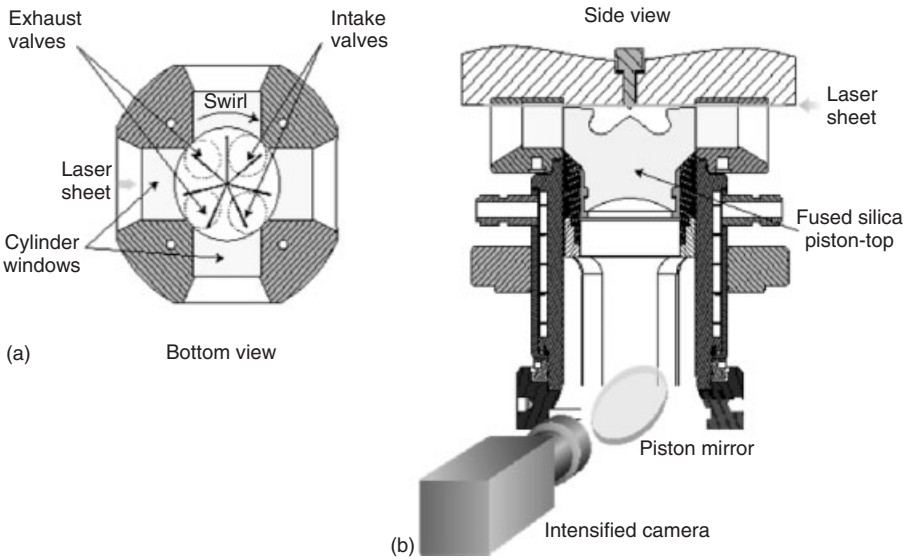


Figure 10. (a, b) Schematic of light-duty optical diesel engine showing laser sheet location and camera setup. (a) The left-hand image shows the view seen through fused silica piston top by the camera using the 45 degree piston mirror. (Reprinted with permission. Copyright © 2012 SAE International (Dempsey *et al*., 2012).)

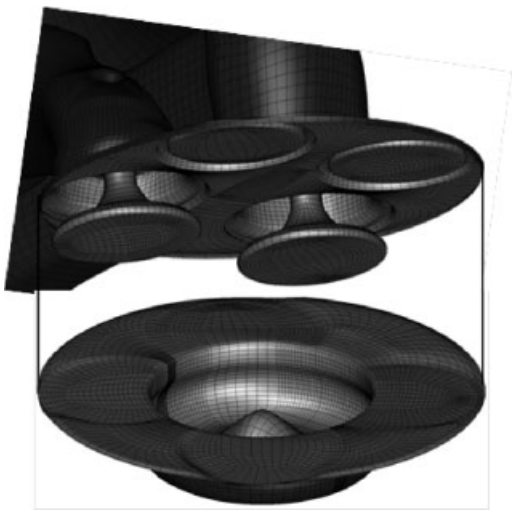


Figure 11. Three-dimensional intake simulation computational grid, which includes valve recesses in the cylinder head and valve cutouts in the piston. (Reprinted with permission. Copyright © 2012 SAE International (Dempsey *et al*., 2012).)

was centrally mounted in the cylinder head and features seven equally spaced holes with nominal nozzle hole diameter of 0.14 mm. The piston top is constructed from fused silica but retains the production piston geometry, featuring a reentrant bowl design and piston crown valve cutouts, as seen in the numerical mesh shown in Figure 11. Full 3D simulation results were used to initialize sector mesh simulations.

The engine was operated with pure nitrogen intake and thus under noncombusting conditions to allow equivalence ratio measurements using a toluene fluorescence diagnostic. The near-TDC temperature and density of a baseline combusting operating condition were closely matched by maintaining the same intake charge mass flow rate and lowering the intake temperature to account for the lower specific heat of the pure N_2 charge. The operating conditions for the baseline combusting case and the nonreacting case are shown in Table 1.

A comparison of the measured (Figure 12a) and predicted (Figure 12b) equivalence ratio fields is shown in Figure 12. The predictions show only one-seventh of the chamber as symmetry is assumed about each of the seven fuel jets from the nozzle. It can be seen that rapid mixing is observed in the head of the injected spray jets in the region above the bowl, with peak equivalence ratios near 2.0. This is well captured by the simulation where the peak equivalence ratio in the head of the jet is ∼2.1. It is also seen that the simulations adequately predict the bulk effects of swirl motion for this engine on the fuel spray.

4 DIRECT NUMERICAL SIMULATIONS OF ATOMIZATION

As described in Section 3.2, the physics of the liquid breakup process are still poorly understood. Accordingly, recent efforts to predict the atomization process, not through

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

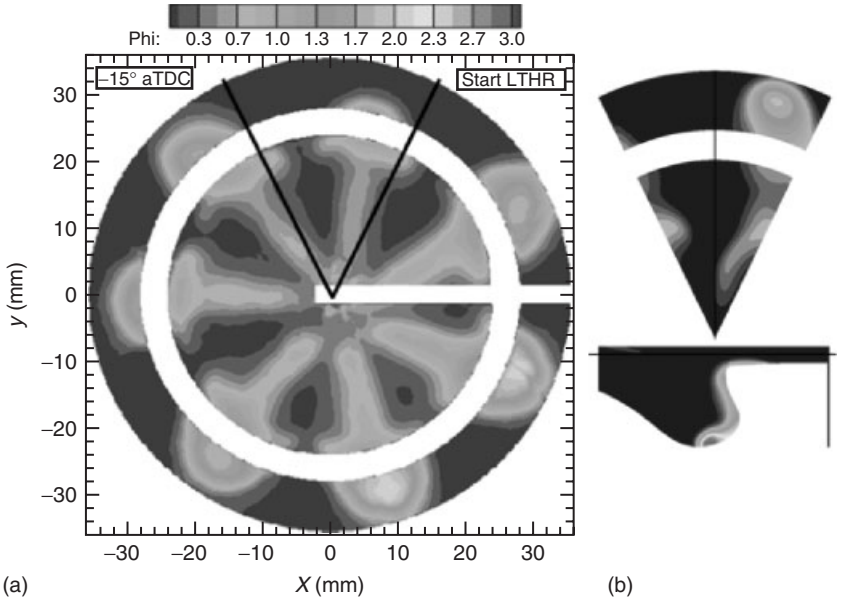


Figure 12. Comparison of measured (a) and predicted (b, top) equivalence ratio distributions in a horizontal plane 1.4 mm below the firedeck, 8.6° after the start of injection (4.1° after the end of injection), which is the time of the start of low temperature heat release (LTHR). The measurement plane location is also shown in the end elevation view through the plane of the spray (b, bottom. Symmetry about fuel jets is assumed in the computation, so only one-seventh of the computational plan is shown in (b). (Reprinted with permission. Copyright © 2012 SAE International (Dempsey *et al*., 2012).)

modeling, but directly from a fully resolved simulation of the underlying equations or a DNS have commenced (Tryggvason, Scardovelli, and Zaleski 2011; Shinjo and Umemura, 2010, 2011a,2011b). DNS is enabled through a combination of improvements in numerical methods coupled with increases in computational power and the proliferation of massively parallel computing. The field is relatively young, with the first review been reported by Gorokhovski and Herrman (2008), which summarized work that appeared in the literature as far back as 2004. Owing to its recent inception and its potential impact into a variety of industrial applications, it is attracting an increasing number of research efforts and progress is being made at an accelerated pace.

An important aspect to note from recent DNS studies of atomization, for instance Shinjo and Umemura (2011b), is the significant level of computational power required. Under the traditional Lagrangian–Eulerian model approach, the number of cells (n) often used nowadays is around $O(10^5)$ and $O(10^6)$, whereas the DNS calculations are $O(10^9)$. Hence, the computational expense, which at best scales as $n\log(n)$, is approximately three to four orders of magnitude larger. The cost is also reflected in both memory usage and disk space. In addition, owing to the numerical constraints on time step size, for example, the Courant–Fredricks–Lewy (CFL) condition, the computational burden is further exacerbated. Nevertheless, the level of physical detail offered by DNS is well beyond any conventional spray treatment through simulation or experiments. It provides a complete inside view of the dense region of the spray, which, at this point, has not been possible. Equally as important, DNS offers a means of model testing that is significantly superior to all experimental validation exercises to date (provided the equations solved are an accurate reflection of the physics). All quantities calculated in a DNS can be interrogated and postprocessed for detailed comparison against models. Issues regarding the density of the spray plume, which have proved difficult to overcome in experimental diagnostic techniques, are simply nonexistent and irrelevant in DNS. Moreover, quantities such as local Weber numbers that cannot be obtained directly from experiments are available from DNS data sets.

Owing to the potential benefits arising from the use of DNS data in model validation, as well as in studies of the physics of atomization, it is important to recognize the challenges faced by the numerical solution of the governing equations. As opposed to single-phase flows, the presence of a gas–liquid interface poses significant numerical difficulties and continues to be an active area of research. In what follows, a brief review is given of the main numerical methods used in atomization simulations followed by a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

description of the latest simulation results. Most of the work that has been performed thus far consists of the advection of a gas–liquid interface under incompressible and isothermal flow conditions (both gas and liquid), where phase change is absent.

4.1 Numerical methods

Owing to the numerous topological changes undergoing in a typical atomization scenario, all numerical methods presently employ an implicit interface representation strategy. This essentially implies that the underlying computational mesh remains stationary, and the treatment avoids numerical stability issues (e.g., negative Jacobians) inherent in interface tracking methods, where the mesh points follow the interface. Under this strategy, the more common methods are the volume of fluid (VoF) (Hirt and Nichols, 1981), level set (Osher and Sethian, 1988), and hybrid schemes (Sethian and Smereka, 2003). These are summarized in the following subsections. More extensive reviews can be found in Tryggvason, Scardovelli, and Zaleski (2011) for the VoF approach and Osher and Fedkiw (2003) for the level set method.

4.1.1 Volume-of-fluid

The central part of the method is the transport of a volumetric liquid fraction, α, often referred to as the *color function*. If we define an indicator function as

$$I(\mathbf{x}, t) = \begin{cases} 1 & x \in \text{ the liquid phase} \\ 0 & x \in \text{ the gas phase} \end{cases} \tag{50}$$

the liquid fraction corresponding to a given computational cell, Ω_k, becomes

$$\alpha_k = \frac{1}{|\Omega_k|}\int_{\Omega_k} I(\mathbf{x}, t)\mathrm{d}\forall \tag{51}$$

The VoF method is essentially composed of two steps. In the first step, the gas–liquid interface or free surface is reconstructed from the liquid fraction field. In the second step, this geometric construction is used to solved a transport equation for α, for instance

$$\frac{\partial \alpha}{\partial t} + \nabla \cdot (\alpha \mathbf{u}) = 0 \tag{52}$$

where $\mathbf{u}$ is the local liquid velocity. To date, a multitude of VoF methods have been presented with the first ones being reported by Debar (1974) and Hirt and Nichols (1981). In these works, interfacial cells, that is, cells where $0 < \alpha < 1$, were treated such that the interface was formed from line segments oriented parallel to either coordinate axis (SLIC, simple line interface). This low order geometric reconstruction was then used in the advection of fluid across computational cells. Owing to problems arising from erroneous generation of liquid bodies ejected from the liquid domain, floatsam or jetsam, improved reconstruction methods have been presented. For instance, the SLIC algorithm allowed piecewise linear segments to be arbitrarily oriented with respect to the coordinate axes (PLIC, piecewise linear interface) (Youngs, 1982) and provided an improvement over previous methods. A representation of both PLIC and SLIC methods is shown in Figure 13 adapted from the work of Rider and Kothe (1998).

A key parameter in the correct reconstruction of the interface arises from the accurate calculation of the interface normal, which can be directly linked to the slope of the interface. A number of methods have arisen to compute this quantity including the method of Youngs (1982), ELVIRA (efficient least-squares volume-of-fluid interface reconstruction algorithm) by Pilliod and Puckett (2004), and the least-squares fit method by Scardovelli and Zaleski (2003), as well as a number of other alternatives and combinations as discussed within these articles and in the text by Tryggvason, Scardovelli, and Zaleski (2011). Once the interface is reconstructed, it is employed in the solution of Equation 52, written in integral form as

$$\left(\frac{\alpha_k^{n+1} - \alpha_k^n}{\Delta t}\right)\mathrm{d}V = \int_{\partial\Omega_k} \alpha \mathbf{u} \cdot \mathbf{n}\mathrm{d}s \tag{53}$$

The key difficulty is ensuring that the geometrical information employed in the interface reconstruction is appropriately used in the calculation of volumetric mass fluxes. Due to this limitation, it is challenging to employ a broader range of mesh types, for instance, unstructured meshes.

Owing to the geometrical complexity of these previous schemes, alternatives to the entire interface reconstruction have been proposed within the VoF framework. Some of these are discussed by Gopala and van Wachem (2008). Among these alternatives, one approach, which has received increasing attention in recent years (Kissling *et al.*, 2010; Rusche, 2002), is based on adding a compressive term to the transport equation for α. This is the approach that characterizes the numerical treatment in the open source VoF code interFOAM, which forms part of the openFOAM libraries. In integral form, Equation 52 can be modified according to

$$\int_{\Omega_k} \frac{\partial \alpha}{\partial t}\mathrm{d}V + \sum_{f \in \partial\Omega_k} (\alpha_\mathrm{f}\varphi_\mathrm{f}) = -\sum_{f \in \partial\Omega_k} (\varphi_{rf})[\alpha(1-\alpha)]_\mathrm{rf} \tag{54}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

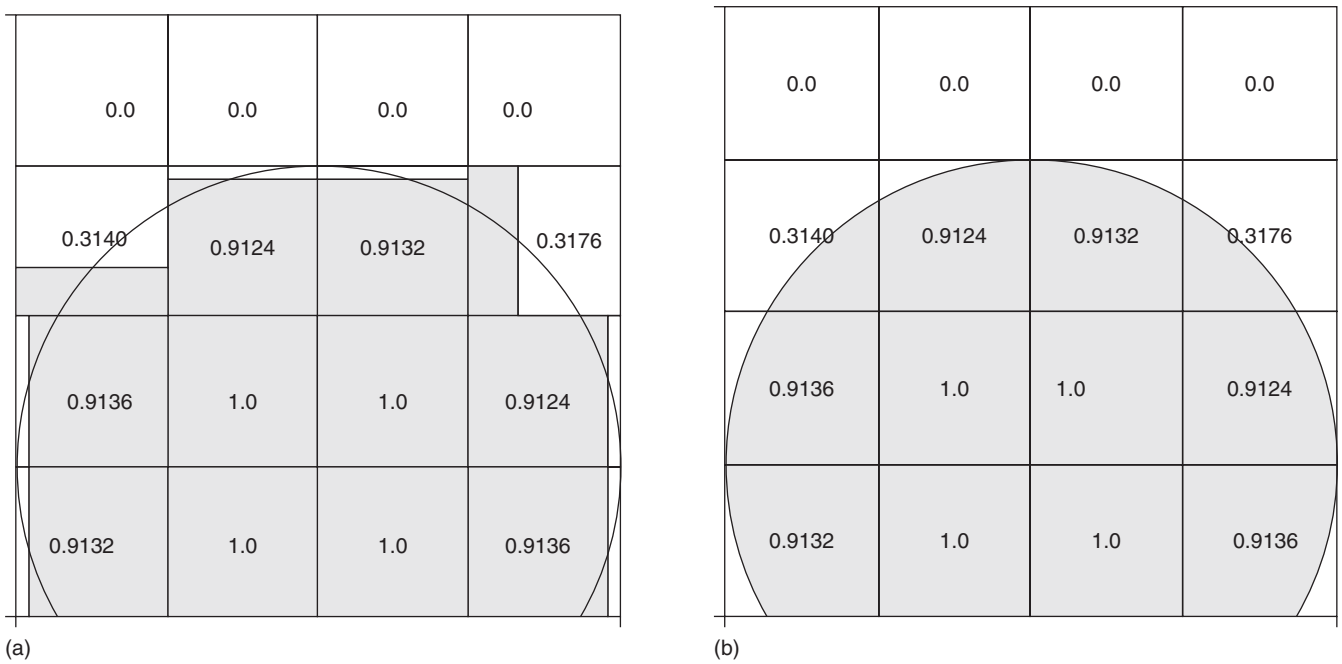


Figure 13. Shaded regions correspond to reconstructed liquid region for both the (a) SLIC and PLIC (b) methods. The PLIC method allows for a more realistic and accurate representation of the interface. Numbers in cells denote volume fractions. (Adapted from Rider and Kothe, 1998. © Elsevier.)

where the subscript f denotes each face associated with cell Ω_k and the flux φ is given by $u_f \cdot S_f$, where S_f is the vector area of a given face and u_f is the cell-centered projected velocity on the face. The compressive flux term appears on the right-hand side of this expression. A limiter is used in the evolution for α, such that the compressive flux is only active in the neighborhood of the interface. Its effect is to maintain the interface sharp by mitigating numerical diffusion. This solver has been tested against a variety of canonical advection test cases, for example, Zalesak notched disk and dynamic problems (e.g., sloshing and single droplet splashing) leading to results that range between first and second order in the L_1 error norm. The accuracy of the calculations are comparable to the coupled level set and volume-of-fluid (CLSVoF) of Sussman and Puckett (2000), the VoF method (SOVoF) of Pilliod and Puckett (2004), and the various VoF schemes studied in Gopala and van Wachem (2008).

4.1.2 Level set methods

A perhaps more popular approach in the direct simulation of multiphase flows with severe topological changes is the use of the level set method. Under this approach (Osher and Fedkiw, 2003), a level set function, ϕ, is solved as a Lagrangian invariant, namely

$$\frac{\partial \phi}{\partial t} + \mathbf{u} \cdot \nabla \phi = 0 \tag{55}$$

This function is then directly used to track the gas–liquid interface, which is generally denoted as the iso-surface corresponding to $\phi = 0$, that is, the zero level set. Hence, regions where ϕ is less than zero are typically interpreted as corresponding to a liquid phase, and similarly regions where ϕ is greater than zero pertain to the gas phase. In addition, the level set function is initialized and periodically maintained to be equal to a distance function to the interface, that is, $\phi(x) = |\mathbf{x} - \mathbf{x}_f|$, where $\mathbf{x}_f$ is the closest interfacial point to $\mathbf{x}$. This distance function property has been shown to help in the preservation of accuracy (Sussman, Smereka, and Osher 1994) as the level set field is evolved.

A variety of methods have been presented on the most efficient and accurate ways to calculate this equation. In fact, much of the literature on this topic does not even address our atomization problem but emphasizes the numerical issues in solving Equation 55. This continues to be an active area of research, which will eventually translate to better and more efficient methods for simulating the liquid breakup problem. The latest methods using this approach (Nave, Rosales, and Seibold 2010; Czajkowski and Desjardins, 2011; Yan and Osher, 2011) report an impressive order of accuracy with second-order behavior in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the calculation of interfacial curvature. This last parameter is particularly important in the calculation of surface tension forces, which govern the breakup behavior for sufficiently small droplets or ligaments.

4.1.3 *Momentum calculation in atomization calculations*

Due to the presence of both gas and liquid phases and the significant changes in material properties, special care has to be given to the solution of momentum, which is derived in Chang *et al*. (1996), yielding

$$\frac{\partial \rho \mathbf{u}}{\partial t} + \nabla \cdot (\rho \mathbf{u}\mathbf{u}) = -\nabla P + \nabla \cdot [\mu(\nabla \mathbf{u} + \nabla \mathbf{u}^T)] + \rho \mathbf{g} + \int_\Gamma \sigma \delta(\mathbf{x} - \mathbf{x}_s)\kappa \mathbf{n} \qquad (56)$$

where the conservative form is presented here and Γ is the gas–liquid interface. For atomization calculations, the surface tension force is particularly important in governing the formation of the resulting droplet population. This force is represented by the last term in Equation (56), where the Dirac delta function, δ, is used to isolate the surface tension effects to the interface location $\mathbf{x}_s$. The local curvature and interface normal are, respectively, given by κ and $\mathbf{n}$. Two common approaches exist in the literature for implementing the surface tension into a numerical discretization of Equation (56), namely the continuous surface force method (Brackbill, Kothe, and Zemach 1992) and the continuous surface stress method. Both of these are elaborated in detail in Tryggvason, Scardovelli, and Zaleski (2011).

For VoF methods, the viscosity and density are generally expressed as a weighted function of the liquid fraction α,

$$\mu(\mathbf{x}_\kappa) = \mu_l \alpha(\mathbf{x}_\kappa) + \mu_g(1 - \alpha(\mathbf{x}_\kappa)) \quad \text{and}$$
$$\rho(\mathbf{x}_\kappa) = \rho_l \alpha(\mathbf{x}_\kappa) + \rho_g(1 - \alpha(\mathbf{x}_\kappa)) \qquad (57)$$

where $\mathbf{x}_k$ corresponds to the spatial location for a given cell. Similarly, in level set methods, these material properties are smoothed over the interface using hyperbolic functions that essentially spread the interface over one or two grid cells. Recent methods have advocated the use of sharp interface strategies, which implies that the above relatively smooth transition in material properties, for instance Equation 57, occurs abruptly. A closer look at these strategies generally reveals that such an exact approach is often voiced but rarely strictly enforced. The final solution of Equation (56) for atomization simulations is often exercised using the classical projection methods introduced by Chorin (1968) or the PISO algorithm (Issa, 1985), where a predictor step is followed by a solution of a pressure Poisson equation to ensure the velocity field at the new time level is divergence free. The stability of the solution is restricted by CFL and capillary timescale constraints (Galusinski and Vigneaux, 2008).

4.2 Simulating atomization

As discussed earlier, this is a fast changing field that has received an increasing amount of attention over recent few years. Gorokhovski and Herrmann (2008) presented a review of the state of the art at that time. Since then, the size of the calculations and the degree of detail afforded by them has notably increased. One of the recurring themes in their reviews is that many of the calculations presented have not reported a systematic grid study in which key properties, such as drop size distribution, could be analyzed. This measure will show directly whether adequate grid resolution has been achieved by demonstrating convergence of the calculated drop size distribution with increasing grid resolution. Many of the calculations were categorized as underresolved DNS, and it was noted that most of the calculations did not provide quantitative comparisons against experimental findings. Similar arguments can be made nowadays with respect to the newer studies reported. In fact, grid studies and numerical convergence on the prediction of interfacial disturbances and subsequent ligament/droplet formation continue to be an area of research need.

In a series of articles, Shinjo and Umemura (2010, 2011a,2011b) have reported results on some of the largest atomization calculations to date. Their study focused on the qualitative and quantitative characteristics of a liquid jet undergoing breakup after being injected into a quiescent domain. Three cases were reported in terms of increasing bulk Weber number, $\mathrm{We} = (\rho_l U^2 D/2)/\sigma$, namely $\mathrm{We} = 1270$ with 400×10^6 cells, $\mathrm{We} = 3530$ with 1.16×10^9 cells, and $\mathrm{We} = 14{,}100$ with 6.0×10^9 cells. A snapshot of the gas–liquid interface colored by axial velocity is shown in Figure 14 corresponding to these three cases. The results clearly indicate an increase in the population of small droplets with increasing Weber number. From the role of the Weber number, which is a measure of the stability of a liquid body to atomization, this is an expected result. Moreover, the authors provide substantial evidence for the manner in which atomization takes place. It begins with instabilities in both the liquid jet tip and the liquid core, which grow and breakup into ligaments, which subsequently break into smaller droplets. This is in agreement with observed experimental findings and provides a reassuring notion used in the development of atomization models.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

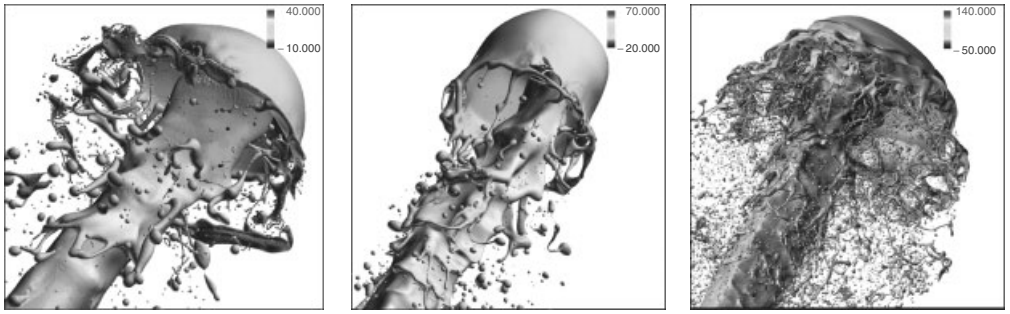


Figure 14. Gas–liquid interface of a liquid jet undergoing atomization corresponding to increasing values of (a) We = 1270, (b) We = 3530, and (c) We = 14, 100. (Adapted from Shinjo and Umemura, 2010. © Elsevier.)

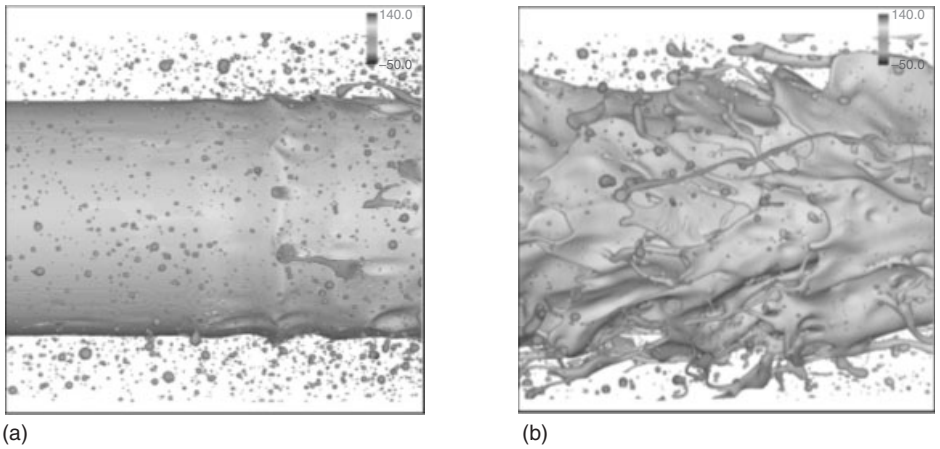


Figure 15. Gas–liquid interface of the jet core displayed in Figure 14 showing the growth of instabilities and subsequent formation of ligaments and droplets. Early time shown in (a) and later time in (b). Colored by axial velocity. (Adapted from Shinjo and Umemura, 2011b. © Elsevier.)

In their latest article, Shinjo and Umemura (2011b) report a study on the breakup characteristics of the liquid core of the jet. These are shown in Figure 15, at two different times. Through a detailed viscous boundary layer analysis of this core region, the authors conclude that the Tollmien–Schlichting (TS) mode is responsible for breakup. Moreover, the authors provide a convincing picture of the physical mechanisms leading to atomization. Initially, the liquid jet tip is deformed because of aerodynamic forcing into the shape of a mushroom head (Figure 14). The deformation proceeds leading to growth of instabilities and formation of ligaments that are periodically stripped from the head. These ligaments subsequently undergo stretching and droplet formation. Downstream of the mushroom head region, a recirculation system is established in the flow field, which entraps small atomized drops. The TS instabilities take some time to develop in the core; however, once they are formed, they are convected into the jet tip, further destabilizing the liquid head. The recirculation flow region also aids in accelerating the growth of the TS waves.

The effect of global Weber number on the resulting droplet population in an atomization calculation has also been reported in the work of Trontin *et al*. (2010), where a liquid sheet having a thickness of δ is exposed to a classical homogenous isotropic turbulent flow field, which is allowed to decay. The sheet is surrounded above and below by gas regions. The gas-to-liquid density and viscosity ratio is one. The authors report that with small Weber number $\mathrm{We} = (\rho_1 U^2 D/2)/\sigma$ of 0.2, the initial disturbance does not result in the atomization of the liquid sheet. With increasing Weber number (We = 2), the sheet atomizes into larger droplets accompanied by small satellite droplets. With even larger Weber number, the atomization intensifies producing a larger population of small droplets. Time histories of the turbulent energy indicate that the maximum deviation from

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

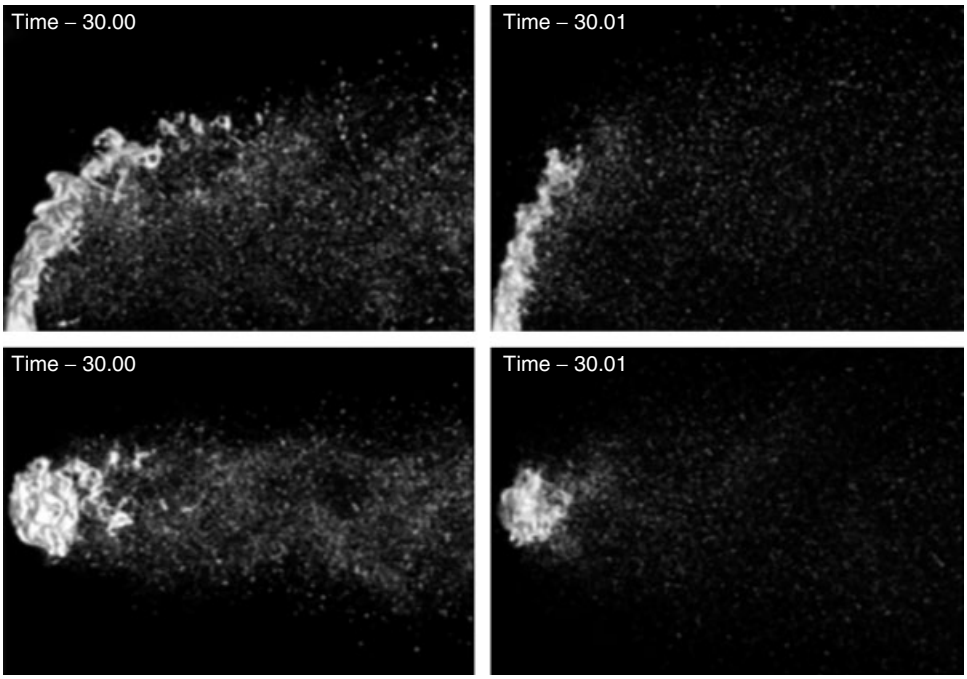


Figure 16. Atomization of a liquid jet in crossflow. Figures on the left and right correspond to $\rho_l/\rho_g = 10$ and 100, respectively. (Adapted from Herrmann, 2011. © Elsevier.)

the corresponding single-phase configuration occurs for the $We = 2$ case, where the oscillations of the larger droplets grow and persist for relatively long times.

The effect of density ratio, which is particularly important in pressurized combustion applications, has been studied in the simulations reported by Herrmann (2011). The effect of varying this ratio is analyzed by considering two cases having $\rho_l/\rho_g = 10$ and 100, respectively, while keeping the momentum flux ratio, crossflow and jet Weber numbers, and crossflow and jet Reynolds number constant. A snapshot of these two cases is shown in Figure 16, where the higher liquid density case exhibits more penetration and appears to bend less in the direction of the crossflow. The results also indicate a decrease in mean drop size with increasing density ratio and a change in the location where atomization takes place. For the higher liquid density case, the formation of droplets occurs closer to the injector. This also impacts the droplet velocity distribution in the crossflow direction, where the higher density case has a significantly lower mean component.

To ease the burden of computation, Gorokhovski and Herrmann (2008) and Herrmann, 2011 introduce Lagrangian elements to represent small-scale liquid structures that are nearly spherical. It is the resolution of these structures that requires the greatest computational power, and hence their translation into a Lagrangian–Eulerian representation provides significant relief on the computational burden. Moreover, the atomization process for these elements is practically complete; hence, full resolution of their dynamics is not completely necessary. A snapshot of this hybrid treatment is shown in Figure 17. Similar approaches have also been reported by Tomar *et al*. (2010), where in their case the underlying simulation is based on the VoF method. They also provide a mechanism in which the Lagrangian droplets are transferred to a VoF description based on the proximity to the gas–liquid interface. This effectively allows for both Lagrangian droplet formation produced by the atomization process and coalescence, in cases where the droplets merge back into the VoF resolved liquid domain.

4.3 DNS summary and future challenges

As pointed out by Shinjo and Umemura (2010) and earlier by Gorokhovski and Herrmann (2008), level set formulations can result in the formation of fictitious droplets on the order of the grid spacing. This imposes a constraint on the grid density to ensure that the spectrum of breakup processes occurring in atomization calculations is fully resolved. With increasing Weber number and the associated shift to smaller droplet populations as evidenced in Figure 14, it becomes increasingly difficult to obey this constraint. It is primarily this fact that establishes a ceiling

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

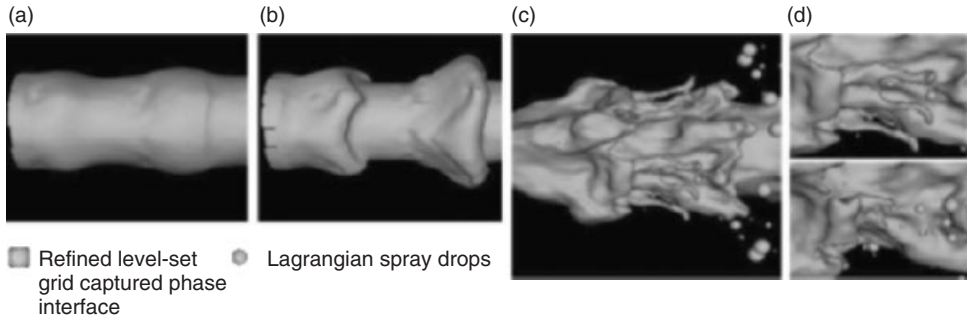


Figure 17. (a–d) Use of Lagragian elements to represent near spherical objects in the level set calculations of atomization. (Adapted with permission from Gorokhovski and Herrmann, 2008. © Annual Reviews.)

on the global Weber number that can be solved in DNS manner. Up to this point, grid studies consider mostly the final droplet population as a measure of convergence. While this is definitely an important point, it leaves other important characteristics such as the growth of instabilities and the process leading to the breakup of fine ligaments undetermined. Much work remains to be done in this aspect. A particularly sensitive topic is the breakup characteristics under various degrees of grid resolution using numerical schemes other than the level set method of Shinjo and Umemura (2010) or Gorokhovski and Herrmann (2008) and Herrmann, (2011).

On the basis of experience involving canonical advection tests using both level set and VoF methods, we found that the errors increase with increasing interfacial curvature. This implies that as the gas–liquid interface becomes increasingly corrugated during the atomization process, the level of error attributed to advection increases substantially. This is illustrated exercising the augmented level set method (Nave, Rosales, and Seibold 2010; Anumolu and Trujillo, 2011), where both the level set function and its gradient are transported in semi-Lagrangian manner. The order of the method is reported as being fourth order locally and to perform better than fifth order weighted essentially nonoscillatory (WENO) schemes. We have confirmed this convergence rate, but only found it to be the case for objects that were relatively smooth. Objects having levels of curvature above a specified threshold show a worsening level of convergence rate. Results are shown in Figure 18 corresponding to the advection of a spherically harmonic shape defined as

$$\phi = \sqrt{(x - x_o)^2 + (y - y_o)^2} - (r + A\cos(4H\theta));$$
$$\theta = \tan^{-1}\left(\frac{y - y_o}{x - x_o}\right) \text{ with } \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \tag{58}$$

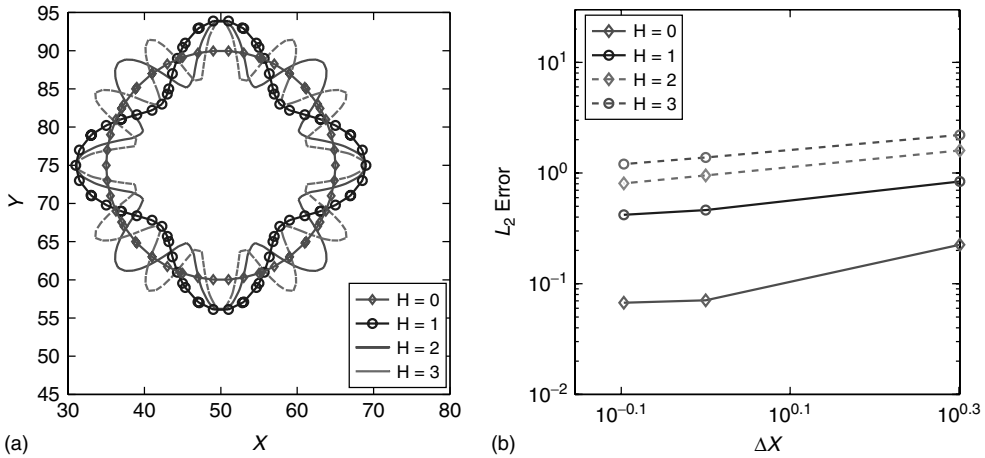


Figure 18. Advection test corresponding to the transport of a spherically harmonic object (a) with increasing levels of interfacial curvature. L_2 error norms are shown on the accompanying plot (b).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

under the advection of a circular flow field

$$\mathbf{u} = (u, v); u = \frac{\pi}{314}(50 - y); v = \frac{\pi}{314}(x - 50) \quad (59)$$

The results clearly indicate that increasing local curvature leads to an increase in the L^2 norm of the error by approximately an order of magnitude (convergence rates are approximately the same). Again, the implications are that the fine structures in an atomization calculation are especially vulnerable to this type of error. In addition, we believe that they are also prone to errors in the prediction of the dynamics of breakup, that is, the two-phase momentum calculations. Although at this time, evidence for this idea has not yet been provided.

Compared to the state of the art of atomization a decade ago, significant improvements in computational power and methods over recent 4–6 years have given birth to a new field, namely DNS of atomization. With the level of detail available in these large calculations, it has become possible to interrogate the physical processes taking place in the dense part of the spray region—an area that has been largely unexplored. These calculations remain challenging, not only in terms of computational demands but also more importantly in controlling and understanding the propagation of errors that can easily overwhelm the entire atomization prediction. Hence, a significant aspect of the research over the next few years will undoubtedly be targeted at minimizing the error. In addition, it is expected that computational models based on Lagrangian–Eulerian or partially resolved Eulerian–Eulerian approaches, as well as hybrid schemes, will be presented as alternative strategies for calculating atomization. These models should have the benefit of having DNS data available for detailed comparisons. This is an unprecedented step and marks a turning point in the development of such models, where before only experiments could be used.

Another route in the DNS approach is the implementation of phase change, as well as other future improvements, to make the simulation approach capable of representing all of the relevant physics occurring in realistic fuel injection configurations, for example, diesel sprays. While the phase change and cavitation capability has been reported within the context of implicit interface capturing methodologies limited to small domains (Lee and Son, 2011; Huang and Zhang, 2007), future improvements including a comprehensive and generalized treatment of the thermodynamics, in particular the interfacial conditions at high pressures remain as challenges. These generalizations are required in the adequate treatment of the two-phase flow processes occurring in various engine configurations. It is reasonable to expect that another decade will have to pass, before the full integration of these effects will have become mature within the framework of implicit interface capturing technologies.

REFERENCES

Amsden, A.A. (1997) KIVA-3 V: A block-structured KIVA program for engines with vertical or canted valves. Los Alamos National Laboratory Report No. LA-13313-MS.

Amsden, A.A., Ramshaw, J.D., O'Rourke, P.J. and Dukowicz, J.K. (1985) KIVA: a computer program for two- and three-dimensional fluid flows with chemical reactions and fuel sprays. Los Alamos Report No. LA-10245-MS.

Anumolu, L. and Trujillo, M.F. (2011) Re-initialization for the Augmented-Level-Set Scheme. *23rd Annual Conference on Liquid Atomization and Spray Systems Conference*, Ventura, CA.

Bai, C., and Gosman, A.D. (1995) Development of methodology for spray impingement simulation. SAE Paper 950283.

Beale, J.C. and Reitz, R.D. (1999) Modeling spray atomization with the Kelvin-Helmholtz/Rayleigh-Taylor hybrid model. *Atomization and Sprays*, **9**, 623–650.

Brackbill, J.U., Kothe, D.B., and Zemach, C. (1992) A continuum method for modeling surface tension. *Journal of Computational Physics*, **100**, 335–354.

Challen, B. and Baranescu, R. (eds) (1998) *Diesel Engine Reference Book*, 2nd edn, Society of Automotive Engineers, Warrendale, PA.

Chang, Y.C., Hou, T.Y., Merriman, B., and Osher, S. (1996) A level set formulation of Eulerian interface capturing methods for incompressible fluid flows. *Journal of Computational Physics*, **124**, 449–464.

Chorin, A.J. (1968) Numerical solution of the Navier–Stokes equations. *Mathematics of Computation*, **22**, 745–762.

Crowe, C., Sommerfeld, M., and Tsuji, Y. (1998) *Multiphase Flows with Droplets and Particles*, CRC Press, Taylor & Francis Group, Boca Raton.

Czajkowski, M.F. and Desjardins, O. (2011) A Discontinuous Galerkin Conservative Level Set Scheme for Simulating Turbulent Primary Atomization, *23rd Annual Conference, Institute of Liquid Atomization and Spray Systems*, Ventura, CA.

DeBar, R. (1974) Fundamentals of the KRAKEN code. Technical Report UCIR-760, LLNL.

Dempsey, A.B., Wang, B., Reitz, R.D., *et al*. (2012) Comparison of quantitative in-cylinder equivalence ratio measurements with CFD predictions for varying injection timings, pressures and swirl ratios in a light duty diesel engine. SAE Paper 2012-01-0143.

Dukowicz, J.K. (1980) A particle-fluid numerical model for liquid sprays. *Journal of Computational Physics*, **35**, 229–253.

Faeth, G.M. (1983) Evaporation and combustion of sprays in *Progress in Energy and Combustion Science*, vol. 9, Pergamon Press, New York, pp. 1–76.

Galusinski, C. and Vigneaux, P. (2008) On stability condition for bifluid flows with surface tension: application to microfluidics. *Journal of Computational Physics*, **227**, 6140–6164.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Gopala, V.R. and van Wachem, B.G.M. (2008) Volume of fluid methods for immiscible-fluid and free-surface flows. *Chemical Engineering Journal*, **141**, 204–221.

Gorokhovski, M. and Herrmann, M. (2008) Modeling primary atomization. *Annual Review of Fluid Mechanics*, **40**, 343–366.

Gosman, A.D., and Ioannides, E. (1981) Aspects of computer simulation of liquid-fueled combustors. AIAA Paper No. 81–0323.

Herrmann, M. (2011) The influence of density ratio on the primary atomization of a turbulent liquid jet in crossflow. *Proceedings of the Combustion Institute*, **33**, 2079–2088.

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*, McGraw Hill, New York.

Hiroyasu, H., Kadota, T., and Arai, M. (1980) Supplementary comments: fuel spray characterization in diesel engines in *Combustion Modeling in Reciprocating Engines*, Plenum Press, New York, pp. 349–405.

Hirt, C.W. and Nichols, B.D. (1981) Volume of fluid (VOF) method for the dynamics of free boundaries. *Journal of Computational Physics*, **39**, 201–225.

Huang, J. and Zhang, H. (2007) Level set method for numerical simulation of a cavitation bubble, its growth, collapse, and rebound near a rigid wall. *Acta Mechanica Solida Sinica*, **23**, 645–653.

Hwang, S.S., Liu, Z., and Reitz, R.D. (1996) Breakup mechanisms and drag coefficients of high speed vaporizing liquid drops. *Atomization and Sprays*, **6**, 353–376.

Ibrahim, E.A., Yang, H.Q., and Przekwas, A.J. (1993) Modeling of spray droplets deformation and breakup. *AIAA Journal of Propulsion and Power*, **9**, 651–654.

Issa, R.I. (1985) Solution of the implicitly discretised fluid flow equations by operator-splitting. *Journal of Computational Physics*, **62**, 40–65.

Jiang, X., Siamas, G.A., Jagus, K., and Karayiannis, T.G. (2010) Physical modelling and advanced simulations of gas liquid two-phase jet flows in atomization and sprays. *Progress in Energy and Combustion Science*, **36**, 131–167.

Kim, D., Desjardins, O., Herrmann, M., and Moin, P. (2007) The Primary Breakup of a Round Liquid Jet by a Coaxial Flow of Gas, *20th Annual Conference, Institute of Liquid Atomization and Spray Systems*, Toronto, CA.

Kim, J., Park, S.W., Sung, K., and Reitz, R.D. (2009) Experimental investigation of intake condition and group-hole nozzle effects on fuel economy and combustion noise for stoichiometric diesel combustion in an HSDI diesel engine. *SAE International Journal of Engines, V118-3*, **2** (1), 1054–1067.

Kissling, K., Springer J., Jasak H., *et al*. (2010) A Coupled Pressure Based Solution Algorithm Based on the Volume-of-Fluid Approach for Two or More Immiscible Fluids. *V European Conference on Computational Fluid Dynamics ECCOMAS CFD*, Lisbon, Portugal.

Lee, W.G. and Reitz, R.D. (2010) A numerical investigation of transient flow and cavitation within minisac and VCO diesel injector nozzles. *ASME Journal of Gas Turbines and Power*, **132**, 052802-1-8.

Lee, W. and Son, G. (2011) Numerical simulation of boiling enhancement on a microstructured surface. *International Communications in Heat and Mass Transfer*, **38**, 168–173.

Lefebvre, A.H. (1989) *Atomization and Sprays*, Hemisphere Publishing Corp., New York.

Lippert, A.M. and Reitz, R.D. (1997) Modeling of multicomponent fuels using continuous distributions with application to droplet evaporation and sprays. SAE Paper 972882.

Liu, A.B., Mather, D., and Reitz, R.D. (1993) Effects of drop drag and breakup on fuel sprays. *SAE Transactions, Journal of Engines*, **102** (3), 63–95.

Munnannur, A. and Reitz, R.D. (2007) A predictive model for fragmenting and non-fragmenting binary droplet collisions for use in multi-dimensional CFD codes. *International Journal of Multiphase Flow*, **33**, 873–896.

Munnannur, A. and Reitz, R.D. (2009) A comprehensive collision model for multi-dimensional engine spray computations. *Atomization and Sprays*, **19** (7), 597–619.

Naber, J.D. and Reitz, R.D. (1988) Modeling engine spray/wall impingement. SAE Paper 880107.

Naber, J. D. and Siebers, D. L. (1996) Effects of gas density and vaporization on penetration and dispersion of diesel sprays. SAE Paper 960034.

Nave, J.-C., Rosales, R.R., and Seibold, B. (2010) A gradient-augmented level set with an optimally local, coherent advection scheme. *Journal of Computational Physics*, **229**, 3802–3827.

O'Rourke, P.J. (1981) Collective drop effects on vaporizing liquid sprays. PhD Thesis, Princeton University.

O'Rourke, P.J. and Amsden, A.A. (1987) The TAB method for numerical calculation of spray droplet breakup. SAE Paper 872089.

O' Rourke, P.J., Amsden, A.A. (2000) A spray/wall interaction submodel for the KIVA-3 wall film model. SAE Paper 2000-01-0271.

Osher, S. and Fedkiw, R. (2003) *Level Set Methods and Dynamic Implicit Surfaces*, Springer, New York.

Osher, S. and Sethian, J.A. (1988) Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, **79**, 12–49.

Peters, B., (1982). Fuel droplets inside the cylinder of a spark ignition engine with axial stratification. SAE Technical Paper 820132, doi:10.4271/820132.

Pickett, L.M. (2007) Engine Combustion Network Data Archive, http://www.sandia.gov/ecn/ (accessed 13 August 2013).

Pickett, L.M., Kook, S., and Williams, T.C. (2009) Transient liquid penetration of early-injection diesel sprays. *SAE International Journal of Engines*, **2**, 785–804.

Pilliod, J.E. and Puckett, E.G. (2004) Second order accurate volume-of-fluid algorithms for tracking material interfaces. *Journal of Computational Physics*, **199**, 465–502.

Ra, Y. and Reitz, R.D. (2009) A vaporization model for discrete multi-component fuel sprays. *International Journal of Multiphase Flow*, **35** (2), 101–117.

Reitz, R.D. (2006) Computer Modeling of Sprays http://www.erc.wisc.edu/modeling/spray_course/modeling_spray_technology.htm (accessed 13 August 2013).

Reitz, R.D. (2005) Liquid atomization and spraying in *CRC Handbook of Mechanical Engineering*, vol. 3, 2nd edn (eds F. Kreith and D.Y. Goswami), CRC Press, Taylor & Francis Group, LLC, Boca Raton, pp. 182–190.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Reitz, R.D. (1988) Modeling atomization processes in high-pressure vaporizing sprays. *Atomisation and Spray Technology*, **3**, 309–337.

Reitz, R.D. and Bracco, F.V. (1986) Mechanisms of breakup of round liquid jets in *Encyclopedia of Fluid Mechanics*, vol. 3 Chapter 10 (ed N. Cheremisnoff), Gulf Publishing, Houston, TX, pp. 233–249.

Rider, W.J. and Kothe, D.B. (1998) Reconstructing volume tracking. *Journal of Computational Physics*, **141**, 112–152.

Rusche, H. (2002) Computational fluid dynamics of dispersed two phase flows at high phase fractions, PhD Thesis, Imperial College of Science, Technology, and Medicine.

Sarre C.V.K, Kong S.C., and Reitz R.D. (1999) Modeling the effects of injector nozzle geometry on diesel sprays. SAE Paper 1999-01-0912.

Scardovelli, R. and Zaleski, S. (2003) Interface reconstruction with least-square fit and split Eulerian–Lagrangian advection. *International Journal for Numerical Methods in Fluids*, **41**, 251–274.

Sethian, J.A. and Smereka, P. (2003) Level set methods for fluid interfaces. *Annual Review of Fluid Mechanics*, **35**, 341–372.

Shi, Y., Ge, H.-W., and Reitz, R.D. (2011) *Computational Optimization of Internal Combustion Engines*, Springer, ISBN: 978-0-85729-618-4.

Shinjo, J. and Umemura, A. (2010) Simulation of liquid jet primary breakup: dynamics of ligament and droplet formation. *International Journal of Multiphase Flow*, **35**, 513–532.

Shinjo, J. and Umemura, A. (2011a) Detailed simulation of primary atomization mechanisms in diesel jet sprays (isolated identification of liquid jet tip effects). *Proceedings of the Combustion Institute*, **33**, 2089–2097.

Shinjo, J. and Umemura, A. (2011b) Surface instability and primary atomization characteristics of straight liquid jet sprays **37**, 1294–1304.

Siebers, D.L. (1999) Scaling liquid-phase fuel penetration in diesel sprays based on mixing-limited vaporization. *SAE Transactions*, **108**, 703–728. 1999-01-0528

Stanton, D.W. and Rutland, C.J. (1998) Multi-dimensional modeling of thin liquid films and spray-wall interactions resulting from impinging sprays. *International Journal of Heat and Mass Transfer*, **41** (20), 3037–3054.

Stiesch, G. (2003) *Modeling Engine Spray and Combustion Processes*, Springer, New York.

Sussman, M. and Puckett, E.G. (2000) A coupled level set and volume-of-fluid method for computing 3D and axisymmetric incompressible two-phase flows. *Journal of Computational Physics*, **162**, 301–337.

Sussman, M., Smereka, P., and Osher, S. (1994) A level set method for computing solutions to incompressible two-phase flow. *Journal of Computational Physics*, **114**, 146–159.

Taylor, G.I. (1963) The shape and acceleration of a drop in a high speed air stream in *Scientific Papers*, vol. 3 (eds G.I. Taylor and G.K. Batchelor), University Press, Cambridge, pp. 457–464.

Tomar, G., Fuster, D., Zaleski, S., and Popinet, S. (2010) Multiscale simulations of primary atomization. *Computers and Fluids*, **39**, 1864–1874.

Trontin, P., Vincent, S., Estivalezes, J.L., and Caltagirone, J.P. (2010) Direct numerical simulation of a freely decaying turbulent interfacial flow. *International Journal of Multiphase Flow*, **36**, 891–907.

Tryggvason, G., Scardovelli, R., and Zaleski, S. (2011) *Direct Numerical Simulation of Gas–Liquid Multiphase Flows*, Cambridge University Press, Cambridge.

Wang, Y., Ge, H.-W., and Reitz, R.D. (2010) Validation of mesh- and timestep-independent spray models for multidimensional engine CFD simulations. *SAE International Journal of Fuels and Lubricants*, **3** (1), 277–302.

Williams, F.A. (1958) Spray combustion and atomization. *Physics of Fluids*, **1**, 541–555.

Yan, J. and Osher, S. (2011) A local discontinuous Galerkin method for directly solving Hamilton-Jacobi equations. *Journal of Computational Physics*, **230**, 232–244.

Yang, S.Y. and Reitz, R.D. (2010) A continuous multi-component fuel flame propagation and chemical 802 kinetics model. *ASME Journal of Engineering, Gas Turbines and Power*, **132**, 072802-1-7.

Youngs, D.L. (1982) Time dependent multi-material flow with large fluid distortion in *Numerical Methods for Fluid Dynamics* (eds K.W. Morton and M.J. Baines), Academic Press, London, pp. 273–285.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Please note that the abstract and keywords will not be included in the printed book, but are required for the online presentation of this book which will be published on Wiley Online Library (http://onlinelibrary.wiley.com/). If the abstract and keywords are not present below, please take this opportunity to add them now.

The abstract should be a short paragraph of between 150–200 words in length and there should be 5 to 10 keywords

Abstract: This chapter reviews the fundamental processes involved in the introduction of fuels into the engine. Spray nozzles commonly used in engine applications are summarized, together with fuel introduction strategies ranging from classical carburetor intake manifold fuel introduction to modern port fuel injection (PFI) fuel injectors. Issues relevant for direct in-cylinder fuel introduction spark-ignition and diesel engines are also reviewed. The effects of air entrainment on spray penetration and vaporization, together with injector nozzle, injection timing and fuel effects on combustion, oil dilution, and engine wear are also discussed. In addition, the effect of fuel spray interaction with in-cylinder airflows is discussed. Current models of fundamental spray processes in engines are reviewed, including models for atomization, drop breakup, drop collision and coalescence, vaporization, drag and deformation, turbulent diffusion, and spray/wall impingement processes. This chapter concludes with a discussion of new computational tools that are currently under development for improved understanding of spray processes.

Keywords: sprays; fuel injection; atomization; breakup; penetration; vaporization; DNS modeling

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto118
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In-Cylinder Flow

Jacques Borée[1] and Paul C. Miles[2]

[1] *ISAE-ENSMA, Futuroscope Chasseneuil, France*
[2] *Sandia National Laboratories, Livermore, CA, USA*

1 Introduction 1
2 General Concepts 2
3 Fundamental Properties of In-Cylinder Flows 5
4 In-Cylinder Macro Flows 11
5 Summary and Concluding Remarks 27
Acknowledgments 28
Endnotes 28
References 28

1 INTRODUCTION

> "An automobile engine, even at present, represents (at least from a fluid mechanical point of view) an accumulation of a century of practice and just a little science"
>
> John L. Lumley

As pointed out by Gosman (1986) in his treatise "Flow Processes in Cylinders", in-cylinder flows in automotive engines have far greater importance than merely the efficient filling and emptying of the cylinder. These flows are essential to the fuel–air mixing process, both through bulk transport by the mean flow and through the small-scale mixing by the turbulent eddies, which subsequently allows molecular diffusion to rapidly complete the intermingling of fuel and air on a molecular scale. The scaling of the turbulent flow structure—an expression we will use to encompass the mean, bulk flow structures as well as the turbulence—with engine speed is the fundamental reason that engines are able to run efficiently even as engine speed changes by an order of magnitude. In spark ignition (SI), predominantly premixed engines, the near proportionality of the turbulent flame speed to the turbulent fluctuating velocity is what allows the oxidation process to be completed in shorter times at high engine speeds, while in non-premixed or partially premixed compression ignition (CI) engines burn duration is shortened by the analogous increase in turbulent mixing rates. Turbulent mixing is also important in homogeneous charge CI combustion regimes, where the temperature inhomogeneities created by turbulent transport from the boundary layers act to smooth the heat release, preventing excessive pressure rise rates.

Other areas where in-cylinder flows impact engine operation include misfire and cycle-to-cycle variability (CCV), post-combustion mixing (important for completion of the final oxidation of CO and particulate matter as well as reduction of NO_x formation), heat transfer, and mixing of crevice hydrocarbons with hot bulk gases during the expansion stroke. An excellent recent example of the importance of these flows to engine design is provided by Yamakawa *et al.* (2011), who describe their efforts to control mean flow structures, turbulence and burn duration, heat transfer, and engine knock in a direct-injection SI engine through modification of piston geometry and manipulation of the flow structure by the injected fuel sprays.

As indicated by the quote opening this chapter, both the authors have been heavily influenced by the ideas of John Lumley, who has been a stalwart of both the fluid mechanics and the automotive engineering communities. In keeping with the spirit of his book *Engines: An Introduction*, our primary objective in this chapter is to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

provide the automotive engineer with a basic physical understanding of the underlying flow processes that occur in engines. We follow this with a description of the various flow structures and turbulence generation mechanisms observed in engines, and the parameters that the engine designer can vary to modify or control these structures and mechanisms. Our hope is that physical understanding, coupled with practical examples of the embodiment and control of flow structures, will give designers the fundamental insight and ideas they need to accomplish their goals. With this insight to guide them, the predictive modeling tools described in Zero- and One-Dimensional Methodologies and Tools and Multidimensional Simulation can then be effectively and efficiently employed. Our additional objective, however, is to provide the fluid mechanics community a synopsis of how basic turbulent flow structures are manifested in engines, and the gaps in our understanding of these structures, with the objective of guiding their work to improve the understanding and tools needed by the engine design community.

2 GENERAL CONCEPTS

2.1 Tumble and swirl and their generation

Engine induction port and valve geometries are carefully designed to produce large-scale organized motions known as *tumble* and *swirl* (Figure 1). Tumble, which is also called *barrel swirl*, is a rotational motion with an axis that is perpendicular to the cylinder axis and to the diametral plane bisecting the intake ports of a four-valve engine. Tumble motion, with an axis parallel to the bisecting plane, is usually referred to as *cross-tumble*; for our purposes, there will be little need to distinguish between the two. Swirl is defined as expected intuitively: rotational motion with an axis aligned with the cylinder axis. This terminology is a convenient way to define globally the three components of fluid rotation in an engine. If both swirl and tumble are present, the flow is often described as a tumble flow which is spinning about the cylinder axis or as a swirling flow with an axis that is tilted with respect to and that precesses about the cylinder axis. Although tumble and swirl are useful and widely used concepts, the detailed, instantaneous flow structure and its associated physics cannot generally be described by such simple motions or their superposition, as will be seen in subsequent text.

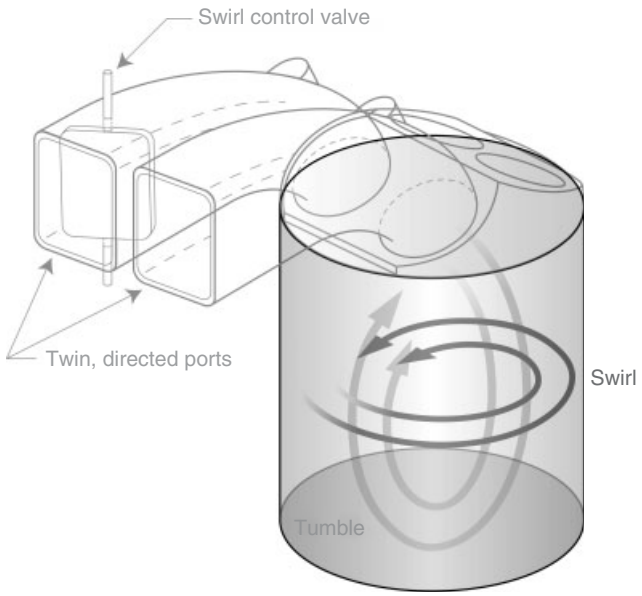


Figure 1. Definition of swirl and tumble, and illustration of the twin ports and optional swirl control valve typically employed in modern SI engines.

Engine designers are concerned with generating and quantifying the amount of in-cylinder swirl or tumble because momentum and energy stored in these rotational motions provides a means by which both mean flow structures and turbulence can be manipulated or generated later in the cycle, when these motions are needed to assist in the fuel–air mixture preparation process, the main combustion process, or the late-cycle, post-combustion oxidation processes. Inducing and manipulating internal aerodynamics to best aid these processes is the art of combustion system design, although very little has been written to explicitly describe how this is accomplished. However, many examples of flow manipulation are based on one common theme: the angular momentum stored within the tumble or swirl structures is manipulated to either force the breakdown of these structures, thereby releasing their energy into turbulence, or to drive the generation of new structures that can effectively transport fuel or combustion products within the cylinder or that can generate turbulence where it is needed locally. This manipulation is accomplished through either changes in geometry that occur with piston motion or through the action of an external stimulus—such as the fuel injection process.

The strength of the swirl and tumble motions are estimated from data acquired at several valve lifts on a steady-state "flow bench" (e.g., Xu, 2001). Through weighting the angular momentum of the tumble or swirl components of the steady flow by the mass flow rate through the valve at each lift, an estimate of the average angular momentum per unit mass inducted into the cylinder can be obtained (e.g., Heywood, 1988). The angular velocity ω_s or ω_t of the solid-body rotating flow with the same average angular

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

momentum is used to define swirl and tumble ratios

$$R_s = \frac{\omega_s}{2\pi N}$$

and

$$R_t = \frac{\omega_t}{2\pi N} \quad (1)$$

N represents the rotational speed of the crankshaft and the subscripts s and t denote swirl and tumble, respectively. Typical values of tumble ratios in engines designed for tumble are in the range of 1–2, while typical swirl ratios in engines designed for swirl range from 1 to 5. Generally, as will be discussed further in subsequent text, tumble is associated with SI engine technologies (although gasoline direct-injection engines can employ significant swirl), and swirl is associated with CI engine technologies (although large ~2 L per cylinder displacement CI engines often use very little swirl). In-cylinder measurements of the swirl velocity field generally show remarkably close agreement between the measured angular momentum and that estimated from flow bench measurements (Reuss *et al*., 1995; Petersen and Miles, 2011).

Tumble is also generated naturally in loop-scavenged, two-stroke engines—although two-stroke designers are often concerned primarily with scavenging efficiency and short-circuiting of fresh charge rather than flow manipulation and turbulence generation later in the cycle. Examples of measured tumble structures and engine design to enhance these structures can be found in the work of Fansler and Drake (2009) and of Tsui and Cheng (2007).

Notice that the swirl or tumble ratios discussed earlier are defined in terms of engine speed, but speed is not referenced when we speak of an engine as having a swirl ratio of 2, for example. The reason for this has already been alluded to in the introduction: the turbulent flow velocities in engines scale very closely with engine speed provided the effects of wave dynamics in the manifolds are not too strong. In research engines that have been specifically designed to minimize the effects of manifold dynamics (Margary, Nino, and Vafidis, 1990; Miles *et al*., 2004), detailed in-cylinder velocity measurements—when normalized by the mean piston speed $\overline{V}_p = 2SN$—are independent of engine speed. Here, S denotes the engine stroke. This scaling of the flow structure with piston speed applies to both the mean flow velocities as well as the statistics of the turbulent velocity fluctuations. A more thorough discussion and review of studies of engine speed scaling has been provided by Miles (2009).

Ports are generally classified as directed or helical, as shown in Figure 2. Classic work investigating the flows created by these ports in engines with a single intake valve includes that of Monaghan and Pettifer (1981), Tindal, Williams, and Aldoory (1982), and Brandstätter, Johns, and Wigley (1985). In four-valve engines, the outflows from the two intake ports interact, and the details of the in-cylinder flow structure depend strongly on the specific port geometry being investigated. However, four-valve engines designed for tumble usually have a similar port layout, which is sketched in Figure 1. The induction-generated flow in engines with this geometry has been characterized both experimentally and computationally and is described in detail in subsequent text. The nearly identical geometry of the twin, directed intake ports typically provides very low swirl. By installing a throttle plate (often called a *swirl control valve*) in one port, however, significant levels of flow swirl can be achieved. Tumble can also be controlled using throttle plates (tumble flaps) in the intake ports that direct the port flow toward the upper portion of the intake valve, where it enters the cylinder along the roof of the combustion chamber (e.g., Wurms *et al*., 2011). This can also be accomplished by designing the port to induce flow separation at the bottom of the port, just before the valve, thereby decreasing the flow through the lower portion of the intake valves. In engines with near-vertical intake ports, the flow velocities are highest near the lower portion of the valves, and a counter-rotating, "reverse-tumble" motion can be generated (e.g., Kume *et al*., 1996).

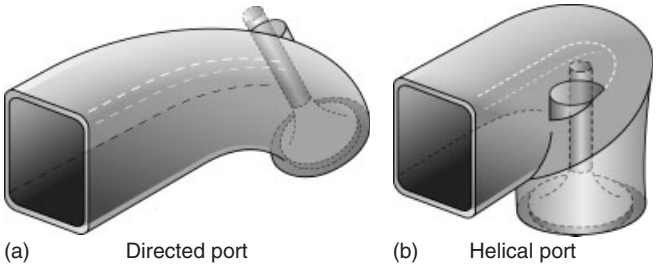


Figure 2. (a,b) Typical port geometries. Directed ports are employed in both SI and CI engines, while helical ports are typically found on CI engines.

In contrast, the two ports designed to generate swirl in modern four-valve engines are often very different. A common design employed in diesel engines incorporates one directed port, positioned such that the net linear momentum leaving the valve is oriented tangential to the cylinder wall and generates significant levels of swirl. The second port is often of helical design, with a lower level of swirl generation. Helical ports typically have higher discharge coefficients, and this port is often called the *fill* port. CI (diesel) engines are frequently found to require a higher swirl ratio at low speed and load for best performance. Under these conditions, the consequences of increased flow restriction are relatively unimportant, and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a swirl control valve is often placed in the fill port. By throttling the fill port, a higher average level of angular momentum is introduced into the cylinder, raising the swirl ratio. At high speeds and loads, both ports are open. Typical swirl ratios achieved with both ports open are in the range of 1–3, and in the range of 3–5 with the fill port throttled. As expected, these ports will also generate some in-cylinder tumble, which may contribute to the considerable tilt in the swirl axis that has been measured in typical diesel engines (Petersen, 2010).

While engine designers are generally concerned with only a single, global swirl ratio or tumble ratio, this is only because there is a lack of knowledge as to the true, detailed impact of in-cylinder flows on the combustion process. Often, the performance and emissions of an engine can vary dramatically at the same swirl ratio, depending on the particular geometry of the ports. A clear example of this can be found in the work of Krieger *et al*. (1997), in which the smoke emissions from an automotive diesel engine are found to differ significantly, at the same swirl ratio, when the engine is fitted with heads having different port geometries. In some instances, it is not only the level of smoke emissions that differs, but an opposite trend in smoke emissions is observed as swirl is varied. The impact of swirl on combustion and emissions can clearly not be correlated by the single, global parameter R_s. To develop and optimize engine combustion chambers, it is thus very important to consider the local, spatio-temporal properties of the turbulent, in-cylinder flows (ideally within each cycle).

2.2 Cycle-to-cycle variations and turbulent flow decomposition

A key challenge facing engine development engineers is the understanding, modeling, and control of cycle-to-cycle variation (CCV) in engine performance, which can contribute to unevenness in the running of the engine, excessive engine noise and emissions, and potentially damaging engine knock. The consequences of CCV are particularly important for direct-injection SI engines with stratified combustion.

Cyclic variability is caused by large-scale variations in flow structures that affect the gas motion and composition throughout the cylinder, although we are especially concerned with changes in the vicinity of the spark plug and at the time of ignition (Ozdor, Dulger, and Sher, 1994) that influence the early flame kernel development in SI engines. These large-scale variations are viewed as distinct from random fluctuations due to turbulent flow motions. Indeed, obtaining a high level of turbulent kinetic energy (TKE) near top dead center (TDC), carried predominantly by a range of energetic eddies of size $\sim l$, is the primary goal of the engine designer who seeks to minimize the duration of the combustion process. While these turbulent fluctuations are essential for combustion and flame propagation, they coexist with the generally lower frequency, larger scale coherent structures in the in-cylinder flow that contribute to CCV.

Turbulence is defined by fluctuations about an average velocity:

$$U_i(\mathbf{x},\theta,n) = \widetilde{U}_i(\mathbf{x},\theta,n) + u_i(\mathbf{x},\theta,n) \qquad (2)$$

U_i is the ith component of the velocity at the location $\mathbf{x}$, crank-angle or phase θ, and during the cycle number n. The definition of $\widetilde{U}_i$ (and hence u_i) is a matter of considerable debate. The simpler definition is the phase average $\widetilde{U}_i(\mathbf{x},\theta,n) = \langle U_i\rangle(\mathbf{x},\theta) = (1/N)\sum_{n=1}^{N} U_i(\mathbf{x},\theta,n)$ computed, at the phase θ, from a large number (N) of independent cycles. Such a definition obviously lumps together all of the fluctuations from the mean, including coherent fluctuations associated with CCV, into the TKE and its associated integral length scale l. This definition is adopted in all Reynolds-averaged Navier–Stokes (RANS) modeling approaches, wherein Equation 2 is substituted into the conservation equations for mass and momentum to derive equations that govern the evolution of the mean flow field as well as the turbulence. RANS-based approaches to turbulence modeling or data analysis are very efficient and useful tools to study engine flow fields. However, this approach is not appropriate for modeling, understanding, and ultimately controlling cyclic variability.

To separate random contributions from turbulence to u_i from the coherent contributions associated with CCV, moving averages in the time or in the spatial domain, or equivalently high pass filtering, have been proposed: see, for example, Fansler and French (1988) and Marc *et al*. (1997). More sophisticated time–frequency approaches, such as wavelet transforms (Sullivan, Ancimer, and Wallace, 1999; Söderberg, Johansson, and Lindoff, 1998) have also been explored. Such approaches are valid only if there is a distinct temporal (frequency) or spatial scale separation between the random and coherent motions (Lumley, 1999) at each phase. A careful analysis, however, shows that such a scale separation is not observed and that "*high pass filtering of velocity data in order to extract turbulence intensity estimates based on time scale discrimination is subject to arbitrary selection of filtering criteria*" (Enotiadis, Vafidis, and Whitelaw, 1990). Moreover, removal of large-scale, anisotropic turbulent motions by inappropriate filtering will significantly impact the ability of the remaining "turbulence" to transport mass and momentum. Finally, it is important to remark that these approaches are not statistically equivalent to crank-angle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

averaging. Although various modeling approaches are dealt with in detail in Multidimensional Simulation, it is relevant to note here that while large-eddy simulation (LES) techniques (Haworth, 1999; Vermorel *et al*., 2009) are based on spatial filtering, the filtering occurs on much smaller length scales that contain little energy and are approximately isotropic. Temporal filtering has also been recently proposed—see Gatski, Rumsey, and Manceau (2007) for a recent review.

A promising alternative approach to the problem of separating random turbulent fluctuations from cyclic variability associated with coherent structures is to extract these structures directly from the measured or computed velocity fields. Identification of large-scale coherent structures can also help isolate the distinct physical processes that are influencing flow development. Specific tools have been developed to accomplish this (Bonnet *et al*., 1998; Bonnet and Delville, 2001), including proper orthogonal decomposition (POD) (Lumley, 1967), a technique that expresses the spatial structure of the flow in terms of the set of orthogonal basis functions (modes) that capture the highest fraction of the flow kinetic energy using the fewest modes. POD has been applied to engine flows to identify flow structures in the valve jet region (Kapitsa *et al*., 2010) and in the central region of a tumbling flow (Maurel, Borée, and Lumley, 2001; Borée, Maurel, and Bazile, 2002). However, POD filtering, which involves projection of the velocity field on a limited number of modes at a given phase, can provide no clear definition of the structures contributing to CCV. As with the spatial or temporal filtering discussed earlier, this is because no clear threshold separating turbulence from the large-scale motions associated with CCV can be objectively defined.

Nevertheless, POD can be employed to define "families" of cycles according to their structure near TDC and more refined statistics can be computed, with $\widetilde{U}_i$ being computed within each family of cycles. Contributions to u_i from differences in $\widetilde{U}_i$ among the different families—that is, contributions from cycle-to-cycle differences in coherent structures—can then be readily computed. This technique was applied to tumble breakdown measured by particle image velocimetry (PIV) in a pent-roof chamber during the second half of the compression phase (Voisine *et al*., 2011). POD also shows promise for constructing reduced-order models built on a phase-invariant POD basis (Fogleman *et al*., 2004).

Additional techniques built to educe the spatial coherence of flows have also been recently applied in engines. When focusing on an unsteady saddle point (e.g., propagation of a jet, fluid ejection away from a solid boundary), Lagrangian finite time techniques (Haller, 2002; Shadden, Dabiri, and Marsden, 2006) can be successfully applied to engine flows (Voisine *et al*., 2011) or even used to define a local triple decomposition based on the location of separation (Ruiz *et al*., 2010). Such advanced approaches are believed to be very useful to the engine community in order to analyze CCV. However, the route to any quantitative and objective definition of CCV is still long, if indeed such a goal is possible in a cylinder that gathers flow regions of different and complex physics, promoting energy exchange at a wide range of scales.

3 FUNDAMENTAL PROPERTIES OF IN-CYLINDER FLOWS

In-cylinder flows are compressed, rotating, and turbulent. This section provides useful estimates of the order of magnitude of various flow quantities and physical insights associated with these three fundamental keywords. A note concerning near-wall turbulence is provided at the end of this section.

3.1 One-dimensional compression

The thermodynamic state of the gas varies significantly during the compression and expansion phases. These variations are, however, imposed by the varying boundary conditions and not by the flow itself. Indeed, except for the flow through valves (not considered in detail here), the Mach number of in-cylinder flows is always very low—the speed of sound is much larger than the flow velocity—see Mansour and Lundgren (1990) and Leblanc and Le Penven (1999). In-cylinder flows are therefore called *compressed flows* where the density varies in time in a deterministic way but not in space. During the compression and expansion phases, the mass balance then easily shows that the flow divergence is nonzero with $\nabla \cdot \mathbf{U} = -\dot{\rho}/\rho$, where $\dot{\rho}$ is the rate of variation of the density. The mean and fluctuating velocity components then satisfy $\nabla \cdot \langle \mathbf{U} \rangle = -\dot{\rho}/\rho$ and $\nabla \cdot \mathbf{u} = 0$. The fluctuating motion is therefore solenoidal for a deterministic variation of the density.

In the case of one-dimensional (1D) compression of a uniform-density, initially quiescent fluid (Gosman, Johns, and Watkins, 1980), the solution to the continuity equation yields a linear variation in the axial velocity $U(z) = V_p(z/z_p)$. Here, z is the axial distance from the cylinder head and z_p denotes the piston position. This expression is commonly used to initialize multidimensional simulations of engine flows when the induction stroke is not simulated. In the general case, the mean velocity field can be uniquely decomposed (Batchelor, 1967) as the sum of a deterministic velocity field induced by the known dilatation rate, in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the given cylinder geometry, and a solenoidal part. This property and the simplified expression $U(z) = V_p(z/z_p)$ was used by Cosadia *et al*. (2006) to analyze spatial properties of a swirling flow during compression.

Recognizing that the cylinder volume V is proportional to z_p, the given expression for $U(z)$ can be differentiated to obtain $\nabla \cdot \mathbf{U} \cong (1/V)(\mathrm{d}V/\mathrm{d}t)$. This expression is generally valid in engine flows, even those that are not 1D, provided we are considering locations outside the thermal boundary layers and provided mass loss rates due to blow-by and global heat transfer rates are small. The divergence can also be estimated directly from the measured cylinder pressure P as

$$\nabla \cdot \mathbf{U} \cong -\frac{1}{P^{1/\gamma}} \frac{\mathrm{d}P^{1/\gamma}}{\mathrm{d}t} \tag{3}$$

In this case, the estimate is expected to be accurate even if mass loss and heat transfer rates are large, provided local density gradients are small and an accurate, local value of the specific heat ratio γ is available.

As will be seen in subsequent text, a significant source (or sink) of turbulence in engines is associated with bulk compression. Likewise, the heating or cooling rates of the in-cylinder gases—which impact ignition, combustion, and emission formation processes—are affected by $\nabla \cdot \mathbf{U}$. Accordingly, the magnitude and timing of $\nabla \cdot \mathbf{U}$, and its dependency on basic engine design parameters, particularly compression ratio, are of interest. Figure 3 shows how $\nabla \cdot \mathbf{U}$ varies with engine crank angle. Note also that the magnitude of $\nabla \cdot \mathbf{U}$ typically peaks near 30° before or after TDC, and that the location of this maximum approaches TDC more closely as compression ratio is increased. The ratio of the crankshaft throw ($S/2$) to the connecting rod length can also impact $\nabla \cdot \mathbf{U}$, but its effect is minor.

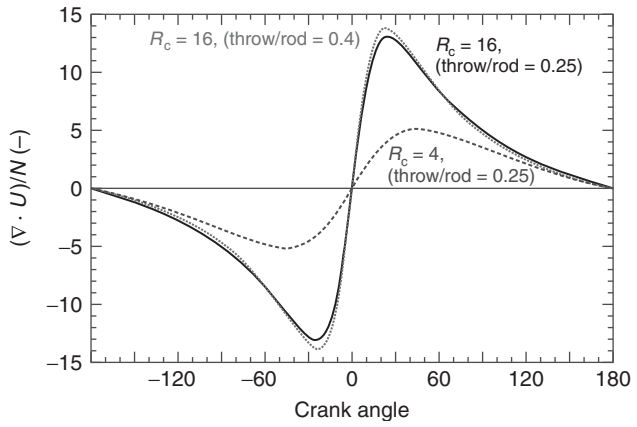


Figure 3. The impact of compression ratio and connecting rod length on the velocity field divergence normalized by the engine rotational rate.

3.2 Large-scale rotation

Engine flows are confined rotating flows, characterized by the swirl ratio or tumble ratio introduced earlier. We introduce here a few simple concepts in order to provide some insights into the physics of these flows. The equations governing the mean radial and azimuthal momentum of a fluid element in a spatially uniform density flow are, respectively,

$$\frac{D\langle U_r \rangle}{Dt} - \frac{\langle U_\theta^2 \rangle}{r} = -\frac{1}{\rho}\frac{\partial \langle p \rangle}{\partial r} + [\text{Reynolds stress terms}] \tag{4}$$

and

$$\frac{D(r\langle U_\theta \rangle)}{Dt} = -\frac{1}{\rho}\frac{\partial \langle p \rangle}{\partial \theta} + [\text{Reynolds stress terms}] \tag{5}$$

These equations are written in cylindrical coordinates. U_r is the radial velocity component, while U_θ is the azimuthal velocity component. Viscous terms can be neglected because the turbulent Reynolds number is large. The mean radial momentum equation (Equation 4) simply states that the acceleration responsible for a departure of a fluid element from a circular streamline is due to a competition between the fictitious centrifugal force and the radial pressure gradient. In engine flows, which are often nearly axisymmetric, the interpretation of the azimuthal momentum equation (Equation 5) is particularly simple: the mean angular momentum $r\langle U_\theta \rangle$ of the fluid element is conserved when the action of turbulent stresses can be neglected.

Conservation of angular momentum is also a very important concept if one wants to understand the evolution of rotational flows submitted to compression. Kelvin's circulation theorem (Batchelor, 1967) can be employed to show that in a flow field where the density is spatially uniform, the vorticity (twice the rotation rate) of a vortex with initial vorticity ω_0, length ℓ_0, and density ρ_0, after being stretched or compressed to a final length ℓ_f and density ρ_f, will be given by

$$\omega_f = \omega_0 \left(\frac{\ell_f}{\ell_0}\right)\left(\frac{\rho_f}{\rho_0}\right) \tag{6}$$

For a large-scale tumble vortex, the length of the vortex does not change during compression; hence, its rotation rate should increase as the vortex is compressed until it eventually breaks up because of the confinement in the reduced combustion chamber volume or internal instability. On the contrary, in a swirling flow, the axial compression of the swirl motion does not impact the rotation rate, as the change in the density is counteracted by the change in the vortex length.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.3 Turbulence: basic properties and production

To illustrate the various sources of turbulence and the physics governing its evolution, in this section we briefly review the fundamental physical principles underlying turbulent flows, discuss the RANS equations governing the evolution of the turbulence, and estimate the order of magnitude of the various terms in the equations. These ideas and equations are very important in order to understand the role of in-cylinder coherent flows, dominant sources of turbulence, and modeling issues further discussed in the chapter devoted to multidimensional modeling (see Multidimensional Simulation). Additional introductory material describing the fundamentals of turbulent flow is provided by Tennekes and Lumley (1972).

Although the specific details of the turbulent flow structure are as varied as the number of possible flow geometries that could exist, the turbulence itself has a remarkably similar structure in each case. Turbulence consists of random, 3D vortices that exist over a broad range of scales. The turbulence energy is extracted from the mean flow by large-scale turbulent vortices of size $\sim l$, which efficiently interact with the velocity gradients in the mean flow. The TKE k is defined by $k = \langle u_i u_i \rangle / 2$[2] with $u_i(\mathbf{x}, \theta, n) = U_i(\mathbf{x}, \theta, n) - \langle U_i \rangle(\mathbf{x}, \theta)$, and a characteristic velocity scale can be defined as $u = \sqrt{2k/3}$. For a large turbulent Reynolds number, $\mathrm{Re}_l = ul/\nu$, the viscous forces on these eddies are small, and very little dissipation of the turbulence energy occurs at scales of size l. Instead, energy is passed from the larger, most energetic scales to the smaller scales through a cascade process, wherein each successively smaller scale draws its energy from larger, but comparable size, scales that are able to interact efficiently. Viscous dissipation occurs predominantly at the smallest scales, which are approximately bounded by the Kolmogorov scale η, at which $\mathrm{Re}_\eta = u_\eta \eta / \nu = 1$. The relationships between the Kolmogorov length scale η, velocity scale u_η, and time scale τ_η and the corresponding large scales are:

$$\frac{\eta}{l} \sim \mathrm{Re}_l^{-3/4}, \quad \frac{u_\eta}{u} \sim \mathrm{Re}_l^{-1/4}, \quad \frac{\tau_\eta u}{l} \sim \mathrm{Re}_l^{-1/2} \tag{7}$$

In a statistically stationary flow, the energy cascade process described reaches a steady state, and the turbulence is said to be in equilibrium. If the flow is nonstationary, and the turbulence is being perturbed (by compression or by changing mean flow structure) on a time scale not significantly greater than l/u, then a steady energy cascade does not have time to develop.

As described, two key features of any turbulent flow are (i) that it receives kinetic energy from the mean motion and (ii) that it is strongly dissipative. The RANS equation governing the local turbulent energy balance reads:

$$\frac{\partial k}{\partial t} + \langle U_j \rangle \frac{\partial k}{\partial x_j} = \mathcal{P} + T - \varepsilon \tag{8}$$

Here, $\varepsilon = 2\nu \langle s_{ij} s_{ij} \rangle$ is the dissipation rate of k per unit mass and $s_{ij} = 1/2(\partial u_i/\partial x_j + \partial u_j/\partial x_i)$ is the fluctuating deformation rate tensor. The smallest scales have the smallest characteristic time scale ($1/s_{ij}$), and hence the largest deformation rate. T is often called *turbulent transport* because, neglecting the average power of the fluctuating pressure and the viscous stresses in the fluctuating motion (Tennekes and Lumley, 1972), the dominant effect of this term is the average transport of the fluctuating kinetic energy by the turbulence itself. The mathematical form of this term will not be useful here, but we note that it strongly contributes to the homogeneity of the turbulence in the chamber. $\mathcal{P}$ is the production term, responsible for kinetic energy exchanges with the mean flow, and reads (δ_{ij} is the Kronecker delta symbol):

$$\mathcal{P} = -\langle u_i u_j \rangle \frac{\partial \langle U_i \rangle}{\partial x_j} = -\left(\langle u_i u_j \rangle - \frac{2k}{3} \delta_{ij} \right) \frac{\partial \langle U_i \rangle}{\partial x_j} - \frac{2k}{3} (\nabla \cdot \langle \mathbf{U} \rangle) = \mathcal{P}_\mathrm{f} + \mathcal{P}_\mathrm{comp} \tag{9}$$

Notice that we have written Equation 9 to emphasize that the production of turbulent energy can be viewed as having two distinct components. The first, $\mathcal{P}_\mathrm{f}$, is highly dependent on the flow structure, as indicated by its dependence on the mean velocity gradient tensor and the anisotropic part of the turbulent stress tensor. The second, $\mathcal{P}_\mathrm{comp}$, is dependent only on scalar quantities: the existing TKE k and the rate of compression.

Considering first that turbulence is sharply suppressed during the expansion stroke and that very little turbulence is generated during the exhaust stroke, the TKE at the end of the compression phase is determined primarily by the intake and the compression phases. During the intake stroke, $\mathcal{P}_\mathrm{comp}$ is zero because the flow is essentially divergence-free (low Mach number). Turbulence is then mainly produced in the high mean shear regions of the inlet jets and by their impingement on the solid surfaces (responsible for turbulent boundary layer formation and secondary flow separation). At the end of the intake phase, the role of the intake jets, of course, vanishes, but the turbulence continues to interact with large-scale mean flow structures generated during intake. During the compression (resp. expansion) phase, the second term $\mathcal{P}_\mathrm{comp}$ provides a positive (resp. negative) contribution to k.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.4 Turbulence: estimation and evolution of turbulence energy, length, and time scales

Engine turbulence is fascinating and inherently complex. Despite this complexity, a great deal of understanding regarding the magnitude of the turbulence scales, how they relate to various engine parameters, their effectiveness in promoting mixing processes, and how the turbulence evolves throughout the cycle can be derived from simple order-of-magnitude estimates. We begin with the following general observations regarding the evolution of turbulent engine flows:

1. They are confined in a complex geometry that is strongly varying in time (say, by one order of magnitude for a typical compression ratio of $r \approx 10$). For order-of-magnitude estimation purposes, we will consider the characteristic length scale (l) of the most energetic eddies to scale with the most confined direction. Hence, $l \approx \alpha b$ at BDC and $l \approx \alpha h$ as the piston approaches TDC. Here, b is the cylinder bore and h the clearance height. A reasonable choice for α is $\alpha \approx 1/6$ (Lumley, 1999).
2. The time available for flow evolution is constrained by the engine rotation rate (N). Indeed, any portion of the cycle between two crank angles is proportional to $\tau_e = S/\overline{V}_p = 1/2N$, where S is the piston stroke and $\overline{V}_p = 2SN$ is the mean piston speed. Note that $\langle|\nabla \cdot \mathbf{U}|\rangle \cong \langle|(1/V)(\mathrm{d}V/\mathrm{d}t)|\rangle \cong (1/\tau_e)$.
3. The thermodynamic state of the gas varies significantly during the compression and expansion phases. Hence, the Reynolds number can vary significantly because of large changes in density and viscosity, even if changes in the flow velocities are small.

Beginning with the intake phase, we now estimate the order of magnitude of the various turbulence quantities introduced earlier. The order of magnitude of the turbulent velocity scale u ($\sim k^{1/2}$) can be obtained at mid-intake by assuming the flux of kinetic energy of the mean flow through the valves to be entirely converted to turbulence energy. Such a global estimation (Lumley, 1999) provides an estimate of $u \approx 10\overline{V}_p$ that approximately agrees with experimental values (Heywood, 1988) and will be retained here. We also adopt the following typical values of $\overline{V}_p \approx 5\,\mathrm{m/s}$, $l \approx 10$ mm and $\nu \approx 10^{-5}\,\mathrm{m^2/s}$, giving $R_l = 50{,}000$. Several observations can be made with these values:

- First, the turbulent Reynolds number $R_l = ul/\nu$ is large and scales linearly with the engine rotation rate. Consequently, the smallest, dissipative scales of the turbulent flow, bounded by the Kolmogorov scale (η), are very small. With the numerical values provided, $\eta \approx 3\ \mu\mathrm{m}$ and is characterized by a molecular diffusion time of roughly $\eta^2/\nu \sim 1\,\mu\mathrm{s}$. This is approximately two orders of magnitude smaller than the time required by the turbulence to mix l-sized inhomogeneities down to the smallest scales ($\approx l/u$).
- Second, such intense turbulence will transport any quantity very efficiently throughout the cylinder. Indeed, from a simple eddy viscosity hypothesis $\delta^2 = \nu_T \Delta t$, where δ is the distance covered by turbulent diffusion $\nu_T = ul$ during the time Δt, it is straightforward to show that with $u \approx 10\overline{V}_p$, $l \approx b/6$, and $b \approx S$, turbulent transport over a distance equal to the radius $\delta = b/2$ of the cylinder requires a time $\Delta t = 0.15S/\overline{V}_p$—corresponding to only 15% of the piston stroke.
- Third, of utmost importance although it has been neglected so far, the turbulence field is responsible for an intense dissipation rate ε. A classical experimental fact (Tennekes and Lumley, 1972; Pope, 2000) is that the amount of fluctuating kinetic energy ($\approx u^2$) and the time needed to dissipate it ($\approx l/u$) are governed by the energy-containing scales. This implies $\varepsilon \approx u^2/\tau_1 \approx u^3/l$, a relationship that is embedded in all classical RANS models. This relationship is only valid, however, if the turbulent energy cascade is in equilibrium and if no other physical phenomena perturb the turbulent energy cascade at a comparable rate. We will come back to this point in what follows. Comparing the decay time scale τ_1 with the engine period leads to $N\tau_1 = 1/120 \ll 1$ (with $u \approx 10\overline{V}_p$, $l \approx b/6$, and $b \approx S$). This means that turbulence has plenty of time to decay during the intake phase. Experimental data indicate that usually $u \approx \overline{V}_p$ at BDC, which means that 99% of the TKE generated during intake has been dissipated [compare $\overline{V}_p{}^2$ to $(10\overline{V}_p)^2$!].

From a global point of view, the objective of engine engineers when inducing large-scale, phase-averaged motions such as swirl or tumble is to store the kinetic energy of the intake flow into large-scale coherent motions that are less dissipative than the turbulent field. Tumble, in particular, is designed to give back this kinetic energy—through the production term $\mathcal{P}$ in Equation 8—in the second half of the compression stroke. It is interesting to compare the order of magnitude of the mean flow kinetic energy E stored in a coherent motion of radius $\approx b/2$, length $\approx b$, rotating in solid-body rotation at $\omega = R_s\omega_e$ (R_s is the swirl ratio but the tumble ratio could be adopted also and $\omega_e = 2\pi N$) to the estimate $E_{tot} = Mu^2/2$ (M is the total mass, $u \approx 10\overline{V}_p$) obtained by Lumley (1999) for the total energy flux into the cylinder. With $E \approx Mb^2\omega^2/16$, one gets $E/E_{tot} \approx 0.01R_t{}^2$. About 5% of the total kinetic energy is therefore stored in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

large-scale motions at BDC for the usual tumble or swirl ratios. This is significant when compared with the results obtained regarding the dissipation of TKE.

The intake valves close shortly after BDC and the in-cylinder flow is then compressed. The further development of the flow through TDC is influenced by the in-cylinder flow properties at valve closure. By dividing the turbulent energy equation by k, we see that the competing factors determining the relative rate of variation of k when following a mean fluid particle are

$$\frac{1}{k}\frac{Dk}{Dt} = \frac{\mathcal{P}_f}{k} + \frac{\mathcal{P}_{comp}}{k} + \frac{T}{k} - \frac{\varepsilon}{k} \tag{10}$$

The first term on the right-hand side ($\mathcal{P}_f/k$) is the rate of production of turbulence energy from the organized in-cylinder motion (tumble, swirl, squish). This term will be considered in detail in the next section. The second term is very specific to the engine context and can be written $\mathcal{P}_{comp}/k = -2/3(\nabla \cdot \mathbf{U})$. A compression time scale can be defined as $\tau_c = -(\nabla \cdot \mathbf{U})^{-1} = (\rho/\dot{\rho})$. Adopting from Figure 3 a typical value during the second half of the compression stroke of $-(\nabla \cdot \mathbf{U})/N = 6$, we obtain $\tau_c \approx 1/3(S/\overline{V}_p)$. Comparing the time scale of compression $\mathcal{P}_{comp}/k$ to the decay rate due to dissipation $\varepsilon/k \approx u/l = 1/\tau_t$ (with $u \approx \overline{V}_p$ and $l \approx b/6$), one obtains $\mathcal{P}_{comp}/\varepsilon = 2/3(\tau_t/\tau_c) \approx 1/3$.

This estimate has important consequences. It shows first that compression (or expansion) has a non-negligible contribution to the TKE balance. Furthermore, $\tau_t/\tau_c \approx 1/2$ means that the compression time scale is on the order of the turbulence time scale. This implies that compressed turbulence in an engine is neither quasi-steady (that would require $\tau_t/\tau_c \ll 1$) nor rapidly distorted (valid if $\tau_t/\tau_c \gg 1$). Therefore, the hypothesis of spectral equilibrium—assumed by nearly all current RANS models—is not strictly applicable (Schiestel, 1987; Lumley, 1990). Equally, simplified linear computations such as rapid distortion theory (RDT)—see, for example, Pope (2000) and references therein—cannot be pursued. To the authors' knowledge, a time-dependent scaling, valid for isotropic compression of an initially isotropic and homogeneous turbulence of very large Reynolds number, is the only strict theoretical frame proposed so far that allows a rational modification of the model equation for ε in RANS models (Cambon, Mao, and Jeandel, 1992). Although modeling is not the focus of this contribution, let us stress that LES techniques solve the equations of motion directly for all compression-affected scales, where $\tau_t \approx \tau_c$, and limit the modeling to the filtered scales where $\tau_t \ll \tau_c$.

Beyond this time scale comparison, one should of course have in mind that compression in an engine is unidirectional. Computation of the compression-induced production of each component of the Reynolds stress tensor for an initially homogeneous and isotropic turbulence shows that strictly 1D compression only "feeds" directly the component aligned with the piston motion. This is a direct consequence of Equation 6, which we considered in the discussion of large-scale structures. For smaller, turbulent structures as well, the length of a vortex aligned with the compression axis is shortened in proportion to the amount of compression, while it is unchanged if it is perpendicular to the compression axis. The angular velocity is therefore conserved for the former and enhanced for the latter, contributing to the anisotropy of the normal stresses and associated length scales. This effect is strongly opposed by the intercomponent energy exchange by the pressure-strain-rate correlations that try to equilibrate the energy in the three components.

We end this section by describing typical orders of magnitude of the state of the turbulent field near TDC. One difficulty, which we have ignored till now, arises when we try to define the length scale. Even in isotropic turbulence, there are two different length scales that can be defined on the basis of spatial velocity correlations. For incompressible, isotropic turbulence, the longitudinal scale, based on the correlation of velocities parallel to the separation vector, is demonstrated to be twice the magnitude of the lateral scale based on the correlation of the velocities perpendicular to the separation vector. Data obtained in near-isotropic grid turbulence (Comte-Bellot and Corrsin, 1971) indicate that the relation $\varepsilon \sim u^3/l$ is most closely satisfied when l is chosen to be approximately equal to the longitudinal scale.

During intake and early compression we have adopted Lumley's (1999) suggestion that $l \approx 1/6$ of the limiting dimension. Near TDC, experimental evidence in both model compression machines (Borée, Maurel, and Bazile, 2002) and engines (see the review in Miles, 2009) suggests that a larger estimate may be more appropriate. In a cylindrical combustion chamber, the limiting dimension (the clearance height) h is $h \approx S/(r-1)$. Here, we adopt $l \approx h/2$ as an estimate of the near-TDC length scale. Assuming that production due to organized motions and compression have been able to maintain $u \approx \overline{V}_p$, one sees that the TKE dissipates very quickly with $N\tau_t \approx 1/4(r-1) \ll 1$. For a compression ratio of 10, we expect the turbulence energy to be dissipated with a time scale characterized by just 10° of crankshaft rotation.

Even if the turbulence energy could be maintained near TDC, it is important to keep in mind that the turbulent transport due to near-TDC turbulence is much less efficient because scales are smaller and the available time for combustion is short. Assuming an available time

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of $\Delta t = S/10\overline{V}_p$, the distance $\delta \approx \sqrt{u\,l\,\Delta t}$ covered by turbulent transport during one-tenth of the piston stroke ($\Delta\theta = 18$ CAD) is just $\delta/b \approx 1/\sqrt{20(r-1)} \approx 0.07 \ll 1$. In contrast, large-scale, coherent motions can be expected to have a much more significant impact on bulk transport and subsequent combustion processes.

These estimates have addressed only the large-scale turbulent motions near TDC. Near TDC, $\nu \approx 4 \times 10^{-5}$ m/s^2 and with $u \approx \overline{V}_p \approx 5$ m/s and $h/2 \approx 5$ mm, then $\mathrm{Re}_l \approx 6000$. Using Equation 7, we obtain for the Kolmogorov scales: $\eta \approx 7\,\mu$m, $u_\eta \approx 0.6$ m/s, and $\tau_\eta \approx 12\,\mu$s. To put these numbers into perspective, the thermal thickness of a flame at a pressure of 30 bar is roughly 30 μm, corresponding to a chemical reaction time scale of $\sim$50 μs. Turbulent and chemical time scales thus overlap. Large scales of turbulence can be expected to transport reacting fluid (flames) without influencing the progression of chemical reaction, while the smaller scales will interact with and alter the progress of combustion. More complete discussions of the interaction between turbulence and chemistry can be found elsewhere, for example, Abraham, Williams, and Bracco (1985).

Finally, we close with the observation that this discussion is derived from studies of the evolution of in-cylinder flows in which the turbulent flow structure is generated by the intake process. Modern, direct-injection SI and CI engines present an additional level of complexity, in that the fuel injection event results in significant turbulence generation and modification of flow structures at times closely coupled to the combustion event. Recent results describing the impact of fuel injection on the in-cylinder flows will be described in greater detail in subsequent text. Here, we limit our remarks to noting that, similar to the intake-generated turbulence, the dissipation of turbulence generated by the fuel injection event will be very rapid. Modern SI fuel injection systems generate fuel jet velocities exceeding 100 m/s, while high pressure diesel injection systems can generate velocities of $\sim$500 m/s. Assuming a length scale of $\sim$1 mm, and a turbulent velocity scale of 30% of the jet centerline velocity, the turbulent time scale characterizing the decay of this turbulence is on the order of tens of microseconds or a fraction of a degree of crankshaft rotation. It will clearly be advantageous to employ the fuel injection process to generate or alter mean flow structures such that turbulence generation continues after the fuel injection event has ended.

3.5 Turbulence: boundary layer structure

Understanding and modeling the physics of near-wall flows is very important because of their role in convective heat and momentum transfer. Here, we briefly review the structure of turbulent boundary layers and sketch their salient properties through an order-of-magnitude analysis similar to that given earlier.

In a near-wall flow, the total shear stress, driven by the turbulent flow, is the sum of a Reynolds stress—dominant away from the wall—and a viscous stress. At the wall, the no-slip boundary condition dictates that all the Reynolds stresses are zero, and the wall shear stress is entirely due to the viscous contribution. Appropriate "friction velocity" and "viscous length" scales are thus defined, with the physical understanding that viscous diffusion of momentum has to adapt itself to the total stress τ_p imposed by turbulence. They read:

$$u_\tau = \sqrt{\frac{\tau_p}{\rho}}$$

and

$$\delta_\nu = \frac{\nu}{u_\tau} \tag{11}$$

The classical description of a turbulent boundary layer results from a two-scale problem (Tennekes and Lumley, 1972; Pope, 2000). The viscous length $\delta_\nu = \nu/u_\tau$ has to be compared with an external scale δ—thickness of a boundary layer, half height of a channel—with $\delta/\delta_\nu = \delta u_\tau/\nu = \mathrm{Re}_\tau$. Re_τ is the friction Reynolds number and is much larger than one. The distance from the wall can be naturally measured in terms of the viscous length $y^+ \equiv y/\delta_\nu$, and the local mean velocity is likewise represented in terms of $u^+ \equiv <U>/u_\tau$. Viscous diffusion of momentum dominates near the wall ($y^+ < 5$) and the mean velocity profile is linear with $u^+ = y^+$, as shown in Figure 4. At larger values in terms of y^+ ($y^+ \gtrsim 30$), but small enough values in terms of y/δ ($y/\delta < 0.3$), the velocity profile is logarithmic because the dependence on ν (or δ_ν) and δ vanishes. Beyond $y/\delta \approx 0.3$, the mean velocity profile is no longer universal and becomes flow specific.

In what follows, we examine if typical boundary layers in engine cylinders can be considered to possess this universal structure, despite the highly transient nature of the flow. To estimate the relevant viscous scales, we will use $u_\tau \approx \langle U_\theta \rangle/30$, an estimate expected to be valid over a wide range of Reynolds numbers (Lumley, 1999). Computing $\langle U_\theta \rangle$ for solid body rotation at a radius $r = b/2$, one obtains $\langle U_\theta \rangle \approx (\pi R_s/2)\,\overline{V}_p$ and $u_\tau \approx (R_s/20)\,\overline{V}_p$.

Our first remark is that the near-wall viscous length scale δ_ν is very small. Comparing δ_ν with the cylinder bore, one obtains $\delta_\nu/b \approx (20/R_s)1/\mathrm{Re}$, where the Reynolds number $Re = b\overline{V}_p/\nu$ is very large. Typically, for $R_s = 2$, $\overline{V}_p \approx 5$ m/s, $b = 80$ mm, and $\nu \approx 10^{-5}$ m^2/s, $\delta_\nu/b \approx 2.5 \times 10^{-4} \ll 1$. The thickness of the viscous

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

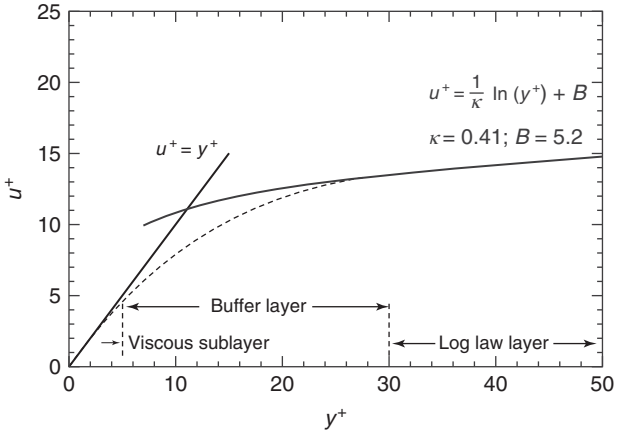


Figure 4. The structure of a turbulent boundary layer.

sublayer is consequently ∼0.1 mm, and a very fine near-wall grid resolution is required if the near-wall boundary layer is to be resolved by a computational model. It is also easy to show that the viscous time scale $\tau_\nu = {\delta_\nu}^2/\nu$ is much smaller than the swirl, tumble, or even compression time scale. This means that viscous diffusion dominates momentum transfer in the viscous sublayer (as expected) and that the local momentum balance is quasi-steady in this very thin layer.

We now move further away from the wall to, say, $y_1 = b/100$ (≈1 mm)—consistent with a typical engine computational grid resolution. In viscous lengths, this corresponds to ${y_1}^+ = 40$ for the numerical values chosen earlier. We, therefore, consider that this location is within the "log-law" region shown in Figure 4. Note that this value of ${y_1}^+$ increases linearly with $\overline{V}_\mathrm{p}$. For canonical near-wall turbulent flows, the expected universal logarithmic velocity profile in this region plus the fact that production and dissipation are in equilibrium can be used to build "wall functions," which constitute a mixed boundary condition specifying the mean velocity $\langle U \rangle$ and the shear stress τ_p at y_1. The great simplification and savings provided by such wall functions is very attractive. Therefore, they are used in most commercial computational fluid dynamics (CFD) codes. However, engines provide many flow conditions—, for example, unsteadiness, separated and impinging flow, regions with strong pressure gradients, compressibility—in which their physical basis and particular implementation may not be well founded.

A simple order of magnitude analysis can again help us here. From Tennekes and Lumley (1972), we learn that a good approximation of the turbulence time scale in a log-law region is $\tau_1 = \kappa y_1/u_\tau$, where the Von Karman constant $\kappa = 0.41$ and y_1 is held fixed at $b/100$. This near-wall timescale can then be compared with the turnover time scale $(1/\omega)$ of the swirl (or tumble) and to the bulk turbulence time scale $\tau_\mathrm{t} \approx l/u \approx b/6\overline{V}_\mathrm{p}$. The first comparison leads to $\omega\tau_1 \approx 0.25$; it is interesting to note in this straightforward computation that this ratio is of order one independent of the specific engine parameters because the dependencies on swirl ratio, engine rotation rate, and geometry cancel out. The second comparison leads to $\tau_1/\tau_\mathrm{t} = 0.5/R_\mathrm{s}$. If a log-law region is supposed to exist, we thus find that the turbulence time scale τ_1 in this region is of the same order of magnitude as the large-scale flow and bulk turbulence (and therefore of the compression time scale). Consequently, we expect near-wall turbulence to be significantly modified by bulk turbulence (having much larger length scales but comparable time scales). Bulk compression will also affect near-wall turbulence in a nonisotropic manner and any evolution of the large-scale rotating flow that occurs over its own turnover time scale—such as tumble breakdown—will lead to unsteady behavior in the near-wall region. Moreover, even steady swirling flows have been noted to cause significant departures from log-law behavior (Jakirlic, Hanjalic, and Tropea, 2002). These observations mean that neither the usual CFD RANS simulations using "wall functions," nor the more sophisticated LES simulations relying on wall functions are well adapted to engine conditions. Experimental data bases are needed on which to test and build models (e.g., Alharbi and Sick, 2010) but are very challenging to obtain.

4 IN-CYLINDER MACRO FLOWS

In this section we describe the generic mean flow structures that are observed in engines, as well as how they can be manipulated by the engine designer. Through this admittedly incomplete description, we hope to provide the reader with the background needed to foster new ideas of flow manipulation. We describe not only the flow structures observed in the cylinder but also available analytical solutions that describe certain simple flow features. Such descriptions are useful even in the age of multidimensional modeling, as they provide the engine designer with useful information on the order of magnitude of various flow features, as well as their approximate phasing within the engine cycle.

4.1 Tumbling flow

The need to induce a tumbling motion in a cylinder has been introduced in Section 2, and in Section 3 it was shown that such a structure can store a significant amount of kinetic

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

energy. Our goal in this section is to give more insights about the structure of the flow and its CCV during the intake and early compression phase. The important tumble breakdown phase will then be discussed and we will end with a note concerning spray–tumble interactions.

4.1.1 Intake flow: the impinging intake jet

Flow structure early in the intake stroke, say before 90 CAD, is generally not discussed in detail in the literature. The confinement is very high and the valve jets impact on the piston. Using horizontal PIV planes a few millimeters away from a flat piston surface, and analyzing the flow topology as proposed by Depardon *et al*. (2006), Voisine (2010) showed that the jet–piston interaction is very strong and results in complex patterns that exhibit noticeable CCV. A system of two mean ringlike vortices generated beneath the intake valves is detected by Dannemann *et al*. (2010), and constitutes the major flow structure at this time. However, in a study involving tumble control using moving flaps located inside the intake port, Keromnes *et al*. (2010) show that a mean tumbling flow can be induced as early as 90 CAD when the flaps are positioned to maximize the tumble. Finally, in a numerical study, Hasse, Sohm, and Durst (2010) identify key flow patterns responsible for cyclic variations in modern engines. These patterns include "*the flow separation at machined edges in the intake port, the separation at the inlet valve as well as the jet deflection at the moving piston*." CCV arising from the combination of these key flow patterns is expected to be amplified during early intake, thereby influencing the rest of the cycle.

Beyond 90 CAD, a tumbling jet structure develops along the wall of the cylinder, opposite to the intake valves, and impinges on the moving piston. The jet is deflected by the piston, spreads along the flat surface and interacts with the curved cylinder wall—generating an up-wash flow along the cylinder walls. Velocities measured in a horizontal plane located at mid-height show that the up-wash flows are recirculated and merge with the descending valve jets. A schematic of the mean flow structure at BDC is proposed in Figure 5 (Voisine *et al*., 2011). The sketch of Figure 5b demonstrates that the true structure of the flow is 3D, and that flow patterns measured in all planes are intimately coupled. Concentrating only on the description of a "tumble center," particularly if measurements are only taken in a symmetry plane, is an oversimplification and can be misleading. Figure 5b also emphasizes the fact that nonsymmetrical jet–piston interactions will significantly impact the formation of the so-called tumble vortex.

Several flow regions are marked in Figure 5a and were used by Voisine *et al*. (2011) to study the CCV of the flow structure. The results are believed to be of general value:

(i) The location of region A (the "tumble center") is strongly correlated with the momentum flux of the descending valve jets. The global sensitivity of the 3D structure at BDC to the development and merging of the valve jets is thus very clear.

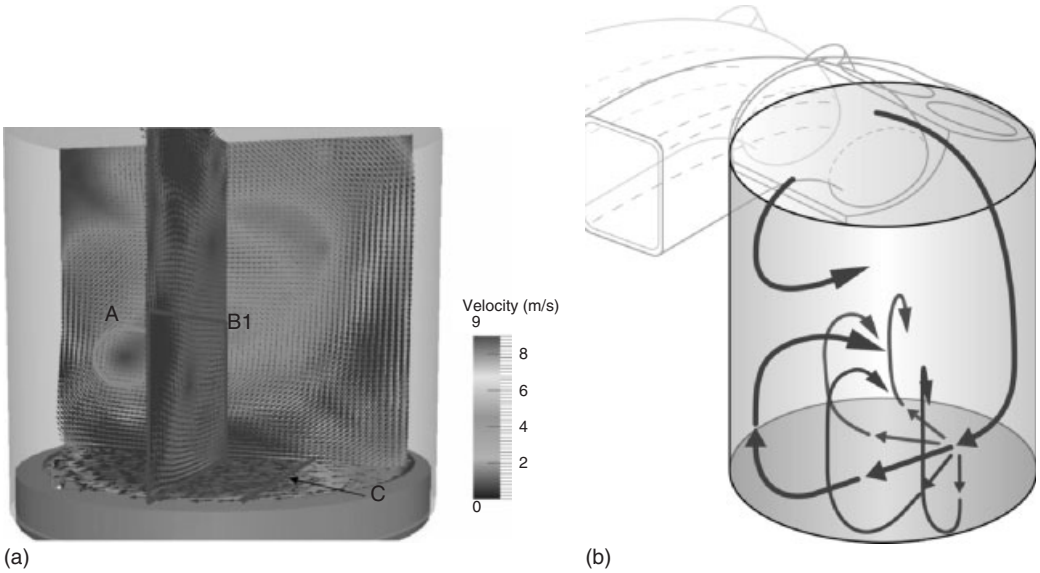


Figure 5. (a) Superposition of orthogonal, two-component PIV planes inside the cylinder at BDC (Reproduced from *Experiments in Fluids*, 2011, p. 1393–1407, Spatio-temporal structure and cycle to cycle variations of an in-cylinder tumbling flow, Voisine *et al*. With kind permission of Springer Science+Business Media); (b) sketch of the three-dimensional mean flow at BDC. (Color Figures are available in the electronic version.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

(ii) The circulation computed from the velocity along contour B1 in the transverse plane is of the order of $\overline{V}_p b$, and is comparable to the circulation of the main tumble structure in the symmetry plane. The coefficient of variation of the fluctuating circulation along B1 is high (about 30%) and fluctuations in circulation on both sides of the symmetry plane are uncorrelated. This means that the large-scale CCV associated with jet–piston interactions is significant, but that it induces no large-scale flow symmetry or antisymmetry relative to the engine symmetry plane.

(iii) RMS variations of about 15% in the integrated momentum flux across line C (line C is located in a horizontal plane 2 mm above the piston surface) show that there are large, coherent fluctuations of the momentum flux.

Each of these observations is the consequence of significant variation in the development, strength, and spatial extent of the valve jet and of the valve jet–piston interactions. Such high levels of fluctuations in integral quantities like circulation are significant and show that CCVs have a direct effect on the global 3D, recirculating flow structure at BDC.

4.1.2 Early compression phase

The early compression phase is characterized by the closure of the intake valves and a less intense jet–piston interaction. The tumbling structure becomes organized and its structure is influenced by the earlier cycle-to-cycle variations in the properties of the tumbling jet. The 3D representation of Figure 6b displays the footprint of the mean 3D motion in horizontal planes 2 mm above the piston top and in the upper region of the cylinder—locations shown by the horizontal lines in Figure 6a. We see that the flow along the piston top diverges away from the jet impact region, rolls up not only in the symmetry plane but also along the cylinder wall, recombines in the central region of the upper region of the cylinder, and impacts the piston in a down-wash motion—the downward motion is apparent when the velocity field in the vertical plane is considered in comparison with the approximate piston-induced velocity $U(z) = V_p(z/z_p)$. Consequently, the mean tumbling motion is not at all symmetric with respect to a "cross-tumble" plane. Moreover, we see that the two "vortical structures" detected in the upper horizontal plane are the signature of this 3D flow structure evolution and not the signature of the vortex breakdown.

A classical way of reasoning, adopted in the previous sections, is to consider the tumble as a simple transverse vortex. As explained in the context of Equation 6, if the length of this vortex does not change during compression, its rotation rate should increase as the compression stroke proceeds. While this spin-up process occurs, angular momentum is approximately conserved, but rotational kinetic energy is not. Conserving mass and assuming a circular vortex cross-section, we find that the ratio of the final-to-initial rotational moment of inertia of

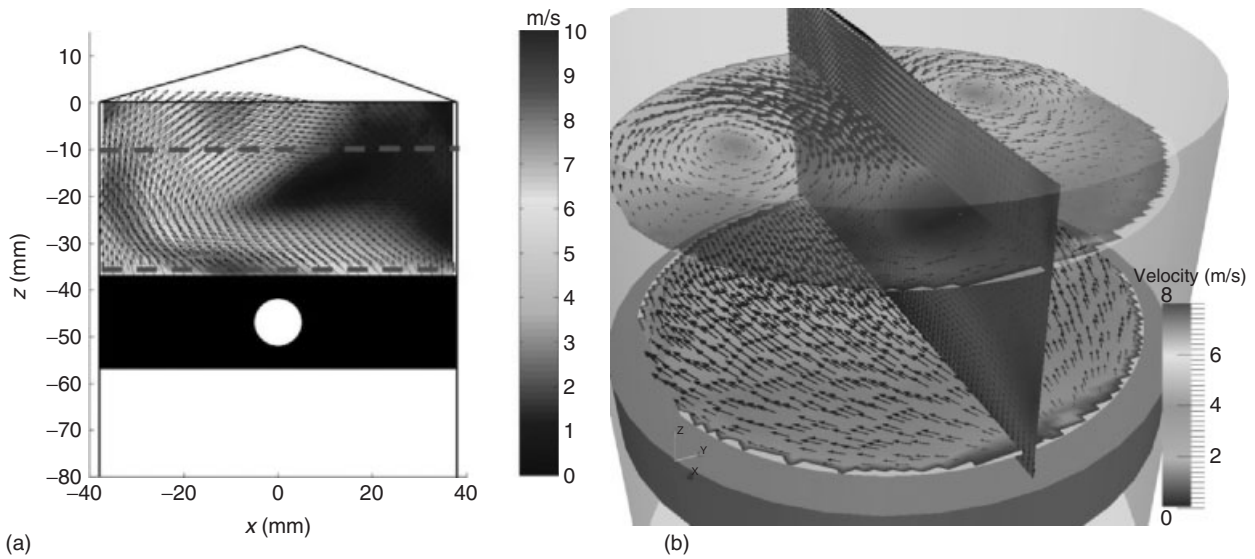


Figure 6. (a) Phase-averaged mean velocity field at $\theta = 270$ CAD, (b) superposition of 2D PIV planes inside the cylinder at half compression stroke. The vertical planes in Figure 6a and b are at the same location. (Reproduced from *Experiments in Fluids*, 2011, p. 1393–1407, Spatio-temporal structure and cycle to cycle variations of an in-cylinder tumbling flow, Voisine *et al*. With kind permission of Springer Science+Business Media.) (Color Figures are available in the electronic version.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the vortex I_f/I_0 is inversely proportional to the vorticity ratio, and hence the rotational kinetic energy $I\omega^2$ increases as ρ_f/ρ_0. A tumble vortex is thus not only a structure that allows the storage of mean flow momentum and energy but it also theoretically allows us to increase the mean flow energy during compression, making a greater amount of energy available to generate turbulence near TDC. The source of this energy is work done by the piston. Qualitatively, we expect the rate of energy transfer to the vortex to be similar to $-\nabla \cdot \mathbf{U}$, and hence the increase in kinetic energy should be strongly affected by compression ratio. The available data only partially support this expectation: at low compression ratio (~4) the mean kinetic energy associated with the tumble motion decreases throughout the compression process (Borée, Maurel, and Bazile, 2002), although a 30% increase is seen at higher compression ratio (Müller, Arndt, and Dreizler, 2011). In the latter case, however, the mean flow kinetic energy peaks at approximately 65° before TDC—well before the peak in $\nabla \cdot \mathbf{U}$. To our understanding, these results are not unexpected: as shown, the real mean flow structure is far more complex and is geometry dependent. Moreover, in each geometry, flow separations, for example, at the corners formed by the cylinder wall and head or piston surfaces, lead to shear flow regions and significant transfer of mean kinetic energy to turbulence.

Characterization of the single-cycle properties of the tumbling jet structure has led to significant progress in increasing our understanding of cycle-to-cycle, large-scale fluctuations of the tumbling motion. Lagrangian techniques for identifying coherent flow structures (Shadden, Dabiri, and Marsden, 2006) have been used to detect the unsteady flow separation away from the cylinder wall (left hand side, $x \approx -5$ mm in Figure 6a) induced by the roll up of the tumbling motion. In the symmetry plane, the vertical location of the jet front—which approximates the separation location—is accurately detected between BDC and 220 CAD and exhibits a very strong cycle-to-cycle variation. For the combustion chamber considered in Voisine *et al*. (2011), the RMS variation in the position of the jet front is approximately 10% of the piston stroke and is strongly correlated with the value of the instantaneous tumble ratio in the chamber. These Lagrangian techniques are believed to be promising ways to explore the in-cycle spatial and temporal coherence of engine flows and to link this coherence with the integral properties relevant for engine development (such as tumble ratio) and their cycle-to-cycle variations.

Finally, we close this section with a brief discussion on the potential for reduction of cycle-to-cycle variations in the tumbling motion. In the tumble control study referenced earlier, by calculating the spatial correlation coefficients for the two velocity components in the symmetry plane at 270 CAD, Keromnes *et al*. (2010) show very clearly that CCVs are considerably reduced when the tumble control flaps are positioned to maximize tumble. In particular, variability in the longitudinal (horizontal) velocity component correlation coefficient is reduced to less than 50% of its value without a flap. Moreover, a distinct tumble structure in the symmetry plane is observed later in the compression stroke than when the tumble is lower—behavior that will likely impact near-TDC turbulence levels, as discussed in subsequent text. More work is needed to determine the mechanisms by which this reduction in CCV and increase in stability of the tumble vortex occurs.

4.1.3 The breakdown phase

Near the end of the compression phase, a transfer of kinetic energy from the large-scale tumbling motion to small-scale turbulence occurs. Even if, as stated earlier, the 3D structure of tumble is geometry dependent, a literature survey of phase-averaged statistics (Arcoumanis, Hu, and Whitelaw, 1990; Hill and Zhang, 1994; Borée, Maurel, and Bazile, 2002; Müller *et al*., 2010; Voisine *et al*., 2011) shows that the transfer of energy from the mean flow to turbulence always occurs beyond $\theta \approx 300$ CAD. Moreover, in very different geometries—square-piston research compression chambers or cylindrical chambers—a semi-log plot of the CA evolution of the mean flow kinetic energy shows that the energy decay is exponential beyond $\theta \approx 300$ CAD and is characterized by the turnover time scale of the large-scale motion. The confined large-scale flow is thus believed to be intrinsically unstable near the end of the compression.

Many fundamental studies have addressed the problem of 3D instability of large 2D vortices in unbounded domains or in bounded geometries with free slip boundaries—see, for example, Lundgren and Mansour (1996). Likewise, a direct numerical simulation of a compressed Taylor vortex has been performed by Le Roy and Le Penven (1997). Lumley (2001) also points out the role of this instability in engine-related flows. The instability of the compressed tumble-vortex may be due to the elliptical nature of the 2D vortex streamlines. The plane strain associated with ellipticity is responsible for an amplification of perturbations carried by the rotating flow. A typical effect is the production of small-scale fluctuations directly from a smooth initial state. An evaluation of the time scale characterizing the growth of the elliptical instability in a model compression chamber was obtained by Borée *et al*. (1999) using the work of Lundgren and Mansour (1996). It shows that there is ample time for this instability to develop, and it is therefore a good candidate for causing tumble breakdown. It is, however, very important to note that the theoretical

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

studies always assume a free-slip boundary. In a real engine, near-wall flow separation has a strong impact on the topology of the flow in the chamber and the transfer of kinetic energy to turbulence. The route to turbulence in the compression chamber beyond 300 CAD might therefore be driven by the confined, naturally unstable, elliptical vortex core interacting both with the walls and with the resulting separated regions.

It is important to understand that the measured increase of TKE due to tumble breakdown is due to a combination of the transfer of energy to small scale turbulence and to large scale cycle-to-cycle variability. To clarify and separate the contributions from these two sources, Voisine *et al*. (2011) analyzed high speed PIV measurements in the pent-roof region during the second part of the compression phase. Distinct families of cycles were identified according to the spatial structure of the velocity field in the symmetry plane near TDC. More precisely, a globally coherent tumbling flow structure was still detected for some cycles near TDC, while other cycles experienced a loss of large-scale coherence. The conditional probabilities associated with these families were then used in order to compute conditional statistics. Phase invariant POD (Fogleman *et al*., 2004; Fogleman, 2005) was used as a rigorous and objective statistical approach to classify and quantify the single-cycle structures. The main conclusions of this analysis are as follows:

(i) In a statistical sense, flows having the largest kinetic energy in the symmetry plane before the breakdown phase are likely to transfer this kinetic energy to smaller scale turbulent fluctuations during the breakdown phase. This indicates that cyclic variability in tumble flow formation will impact near-TDC turbulence levels on a cycle-by-cycle basis.

(ii) The increase in the fluctuating kinetic energy is the largest for cycles that experience a loss of large-scale coherence, whereas no maximum is detected for cycles keeping their coherence. In conjunction with the previous conclusion, this suggests that both large-scale coherent flow structure variations and differences in small-scale turbulence occur on a cycle-to-cycle basis.

(iii) Conditional statistics can be used to decompose the phase-averaged Reynolds stresses into contributions from families of cycles with similar single-cycle coherent structures and contributions from the turbulence. In the experiments of Voisine *et al*. (2011), approximately 30% of the fluctuating kinetic energy is due to cycle-to-cycle coherent structure variations near TDC. Moreover, in the symmetry plane within the pent-roof chamber, this large-scale cyclic variability is mostly due to strong variations in the longitudinal component of velocity. Such variability, especially along the chamber head near the spark location, is expected to strongly influence the ignition process.

4.1.4 Spray–tumble interaction

A short note is made here to emphasize the fact that engine designers may employ the fuel injection process in order to influence and control in-cylinder flows. Supposing that the momentum flux of the liquid fuel spray is efficiently transferred to the gas motion (this is only partially true and depends on spray characteristics), an order of magnitude analysis indeed shows that the flux of angular momentum due to the spray can significantly perturb the tumble vortex for a typical tumble ratio. Yamakawa *et al*. (2011) specifically attempt to tailor the injection process to strengthen the tumble vortex. A similar strategy is described by Hyundai (Han *et al*., 2011). If not properly designed, however, injection–vortex interactions can interfere with vortex breakdown, even when the momentum introduced by fuel injection reinforces the tumble motion. Indeed, measurements made in a research engine (Moreau *et al*., 2004), designed to be dominated by a single tumble vortex, have shown that injection of a single jet can promote breakdown of the vortex structure and lead to turbulence energies three to four times lower at TDC than are seen in the undisrupted vortex because dissipation of TKE has more time to act. Simulation results also suggest that, even when fuel injection occurs during the intake stroke, spray-flow interactions can significantly impact the tumble development and lead to substantial differences in flow turbulence throughout much of the compression stroke (Chen *et al*., 2011).

4.2 Swirling and squish flows

Swirling flows have traditionally been most closely associated with CI engine technologies; accordingly, most of the examples given here have been drawn from the CI engine literature. However, as noted in Section 2, direct-injection SI engines are now often employing significant levels of swirl. An important feature of swirl-supported SI combustion systems is that they are generally less susceptible to CCVs (Zhao, Lai, and Harrington, 1999)—undoubtedly due in part to the fact that the swirl structure does not break down during compression.

So-called squish flows, created when the piston closely approaches a stationary combustion surface, were more important in older engines with L- or wedge-shaped combustion chambers. While modern pent-roof SI engines

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

may generate squish flows over a limited area, squish flows are more dominant in bowl-in-piston CI engine designs. Moreover, the impact of squish flows on turbulence generation is greatly enhanced when they are coupled with flow swirl. Hence, we treat them both in this section.

4.2.1 *One-dimensional analytical representation of a swirl vortex*

In parallel with the simple analytical representation of the axial velocity distribution, there is a simple representation for the radial profile of the swirl velocity, first noted in the context of engine flows by Johnston *et al*. (1979), which arises from a direct solution of the momentum equations. When U_θ is the only nonzero velocity component, the flow is axially and azimuthally uniform, and thermodynamic properties are uniform, the radial velocity profile and its decay rate can be expressed as

$$U_\theta(r,t) = \frac{(2\pi N R_s)\alpha B}{2J_2(\alpha)} \cdot J_1\left(\alpha\frac{2r}{B}\right) \cdot \exp\left(-\frac{4\alpha^2}{B^2}\nu t\right) \tag{12}$$

In Equation 12, B represents the cylinder bore diameter, ν is the viscosity, and α is a nondimensional parameter. J_1 and J_2 are Bessel functions of the first and second kind, respectively. If the no-slip condition at the cylinder wall is enforced, a value of $\alpha = 3.83$ is obtained, which provides an unrealistic shape for the radial profile; selecting $\alpha = 0$ results in a linear profile (solid body rotation). Velocity profiles for various representative values of α are shown in Figure 7.

This simple solution is of importance as it allows the tangential velocity profile to be initialized for multidimensional simulations by specifying only the engine speed, swirl ratio, and a value of α. In conjunction with $U(z) = V_p(z/z_p)$ (a convention that is inconsistent with the derivation of Equation 12), it is employed by several commercial CFD codes. A value of $\alpha = 3.11$ is often recommended (Amsden, O'Rourke, and Butler, 1989); however, a value of $\alpha = 2.2$ generally provides a better match to in-cylinder measurements (Petersen and Miles, 2011). These measurements have been made at $\sim 50^\circ$ before top center, in engines with various port geometries. Hence, well into the compression stroke, the value of $\alpha = 2.2$ can be assumed to have some generality.

If an appropriate turbulent viscosity were known, the decay rate of the swirl could be estimated from Equation 12. However, a more accurate estimate can likely be obtained from empirical formulae for the skin friction at the wall, and calculating the resulting torque acting on the cylinder contents. Employing the correlations in Heywood (1988), we find that it will take approximately one engine revolution to halve the angular velocity of the flow. Evidently, swirl is a very stable flow structure that is well able to survive the compression process.

4.2.2 *Cylindrical compression of a swirling flow*

As discussed earlier, in a swirling flow the axial compression of the swirl motion does not impact the rotation rate, as the change in the density is counteracted by the change in the vortex length. However, when the vortex is compressed into a reduced diameter combustion bowl, as illustrated in the simplified schematic shown in Figure 8, we expect an increase in rotation rate and in rotational kinetic energy. For a hypothetical bore-to-bowl diameter ratio of two, the volume occupied by the vortex is reduced by a factor of 4, providing a proportional increase in density. Hence, Equation 6 predicts a fourfold increase in rotational rate ω. The rotational kinetic energy $I\omega^2$ will also increase by a factor of 4, reflecting the reduction in the rotational moment of inertia which also occurs.

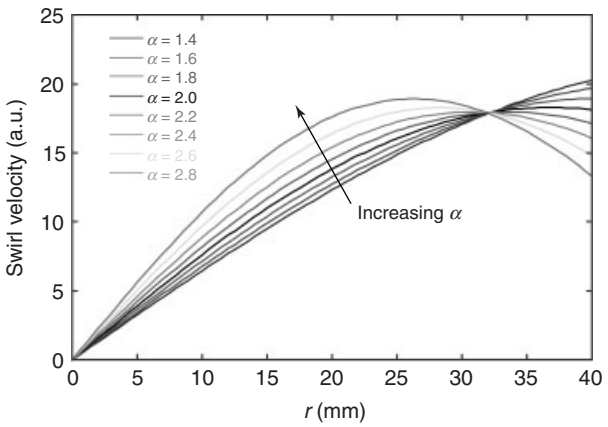


Figure 7. The impact of the parameter α on the shape of the radial profile of U_θ.

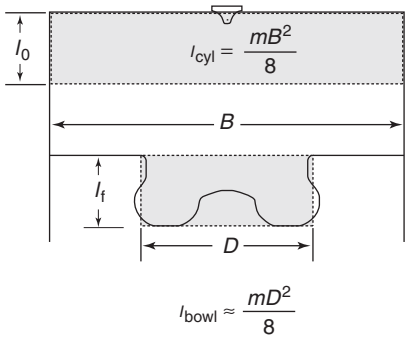


Figure 8. Simplified schematic illustrating the compression of a swirl vortex into a piston bowl.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Measurements (Arcoumanis, Hadjiapostolou, and Whitelaw, 1991) indicate that in a typical, bowl-in-piston automotive diesel engine geometry, approximately 60% of the rotational angular momentum is conserved as the swirl vortex is compressed into the bowl. With a bore-bowl diameter ratio of 2, the rotational kinetic energy in the bowl is thus increased by 40–50%. Similar to the spin-up of a tumble vortex during compression, the source of this energy is piston work.

4.2.3 Magnitude and phasing of the squish flow

To gain an appreciation for the strength of the squish flow, and how it varies with changing piston position and combustion chamber geometry, it is useful to examine the analytical expression derived by Fitzgeorge and Allison (1962) for a cylindrical bowl-in-piston geometry[3]

$$\frac{U_{sq}}{V_p} = \frac{D}{4z}\left[\left(\frac{B}{D}\right)^2 - 1\right]\frac{4V_{bowl}}{4V_{bowl} + \pi B^2 z} \qquad (13)$$

Equation 13 for the squish velocity, U_{sq}, assumes an axially and azimuthally uniform squish velocity through the surface defined by the bowl rim radius $D/2$ and the clearance height z. B is defined as in Figure 8, V_p represents the instantaneous piston speed, and V_{bowl} represents the volume of the piston bowl. The variation in squish velocity, normalized by the mean piston speed, is shown in Figure 9. Similar to the velocity field divergence shown in Figure 3, the peak magnitude of the squish flow occurs quite near to TDC, and turbulence generated by this jet will tend to peak at this time. As will be seen in subsequent text, both the mean flow structures and the turbulence generated by the squish flow are highly dependent on the swirl level.

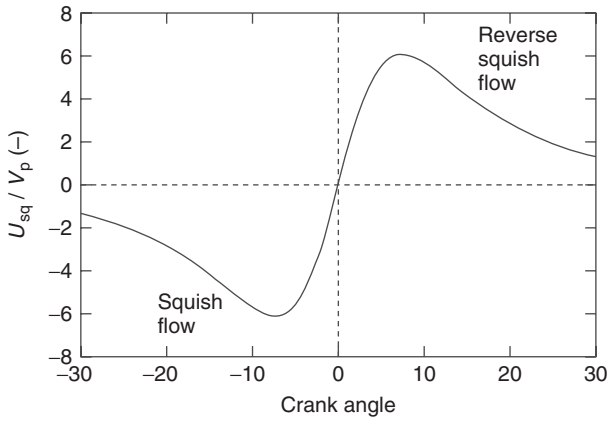


Figure 9. The variation in squish flow strength for a typical bowl-in-piston geometry.

4.2.4 Squish–swirl interaction

During the process of compressing the swirling flow into the bowl, different flow structures within the bowl are generated because of the rearrangement of the in-cylinder angular momentum by the "squish" flow (Gosman, 1986; El Tahry, 1982). The process is illustrated in Figure 10, where we consider the application of conservation of angular momentum to an individual fluid element. As the element is displaced inwardly, its tangential velocity increases in accordance with Equation 5. Owing to the increased tangential velocity, the centrifugal forces acting on the fluid element are increased. Eventually, as the fluid moves inward, the radial momentum imparted by the squish flow, in addition to the radial pressure force, is overcome by increased centrifugal forces and further inward motion is impeded.

Because the squish flow is dominated by the changing combustion chamber geometry, the radial momentum imparted to fluid elements during the squish process is independent of the flow swirl level. In contrast, the centrifugal forces acting on a fluid element will increase approximately quadratically with the flow swirl ratio. The inward penetration of the squish flow, therefore, is strongly influenced by the swirl, with the greatest penetration occurring at the lowest swirl levels. The consequence of this changing balance between squish-imparted momentum and centrifugal forces is shown in Figure 11. For low levels of flow swirl ($R_s \sim 0$), the squish flow penetrates to nearly the cylinder centerline before it turns down into the bowl, as required by symmetry constraints. As the flow swirl is increased, the inward penetration is reduced and the flow turns down into the bowl at larger r, when the increasing centrifugal forces have overcome the initial radial momentum imparted by the squish process. For high swirl levels, the centrifugal forces are sufficiently great that the squish flow turns down into the bowl as soon as the combustion chamber geometry permits. Note that the direction of rotation of the vortex structure that dominates

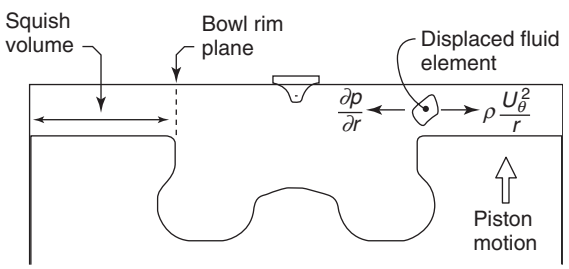


Figure 10. Illustration of the inward displacement of a fluid element by the squish flow.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

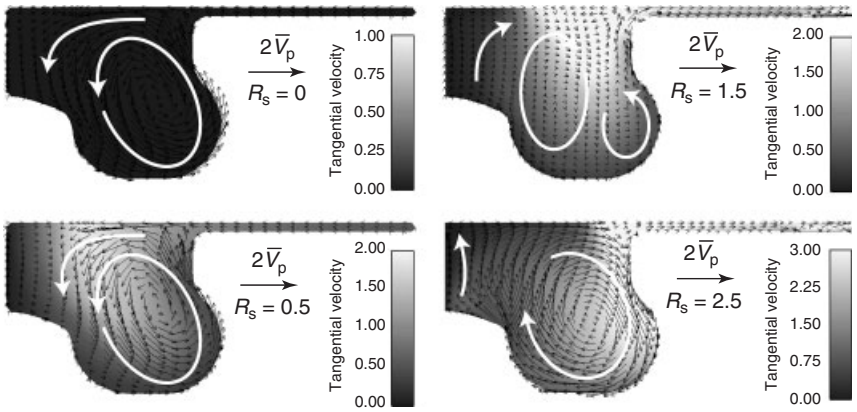


Figure 11. Vertical plane flow velocities induced by the squish swirl interaction at various levels of swirl. The vectors illustrate the motion in the axisymmetric vertical plane, while the color scale provides the magnitude of the out-of-plane tangential velocity.

the bowl at $R_s \sim 0$ and at $R_s \sim 2.5$ has been reversed in Figure 11.

4.2.5 Spray–swirl interaction

In-cylinder measurements coupled with multidimensional simulations (Miles, 2009 and references therein) have shown that the engine designer can influence the development of mean flow structures in a bowl-in-piston combustion chamber by altering the in-cylinder spatial distribution of angular momentum through the fuel injection process. Changes in the spatial distribution of angular momentum drive the formation of flow structures by the same mechanism as the squish–swirl interaction discussed—by balancing the momentum added to the flow by injection against the changing centrifugal forces on a fluid element as it is displaced radially. The redistribution of angular momentum is principally accomplished through two mechanisms: first, the fuel jets entrain nearly quiescent, low angular momentum fluid from the central portion of the cylinder and deposit it near the bowl periphery. Second, the fuel jets are a large source of linear momentum. Deflection of the fuel jets and their accompanying, entrained fluid by the walls of the combustion chamber results in energetic rotating flow structures in the vertical (r–z) plane. Both displacement by the low-momentum fluid transported from the center of the cylinder to the bowl periphery and convection by the rotating vertical plane structures generally act to transport high angular momentum fluid inward in a manner similar to the inward transport of high momentum fluid by the squish flow.

As depicted in Figure 12, by varying the injection pressure P_{inj} or the swirl ratio, the balance between linear momentum added by the fuel jets and the angular momentum associated with the commingled air can be changed. At low swirl ratios, depicted in the upper row of the figure, the deflected momentum of the fuel jet creates a single, clockwise-rotating structure. This structure transports high angular momentum fluid from near the bowl rim down into the bowl, where it is conveyed inward to the central bowl protrusion. Although the tangential velocity of this fluid increases as it moves inward, the centrifugal forces are sufficiently small that they do not strongly alter the vortex structure: the angular momentum is transported much like a passive scalar. As the swirl ratio is increased, larger centrifugal forces associated with the higher angular momentum compel the inwardly displaced fluid to return to the outer regions of the bowl after it turns upward from the bottom of the bowl. As a result, the clockwise-rotating vortex becomes progressively smaller and is centered lower and at larger radii within the bowl, and a complementary counter-rotating vortex forms in the upper central region of the combustion chamber. The relative sizes of these vortical structures can also be impacted by injection pressure: increasing injection pressure increases the momentum in the fuel-induced vortex structure, thereby allowing greater inward and upward displacement of a fluid element before the vortex momentum is overcome by centrifugal forces. Hence, the size of the lower vortex increases.

Generalizing, it is apparent that the geometry of the piston bowl lip, where the linear momentum of the fuel jet is deflected, will also impact the relative size of these vortices. Likewise, an increase in load can be expected to have an effect similar to an increase in injection pressure, as a greater amount of linear momentum is injected with the fuel. Finally, an increase in engine speed will have an effect on the structure of the flow similar to that of an increase in swirl ratio. This latter observation may help explain the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

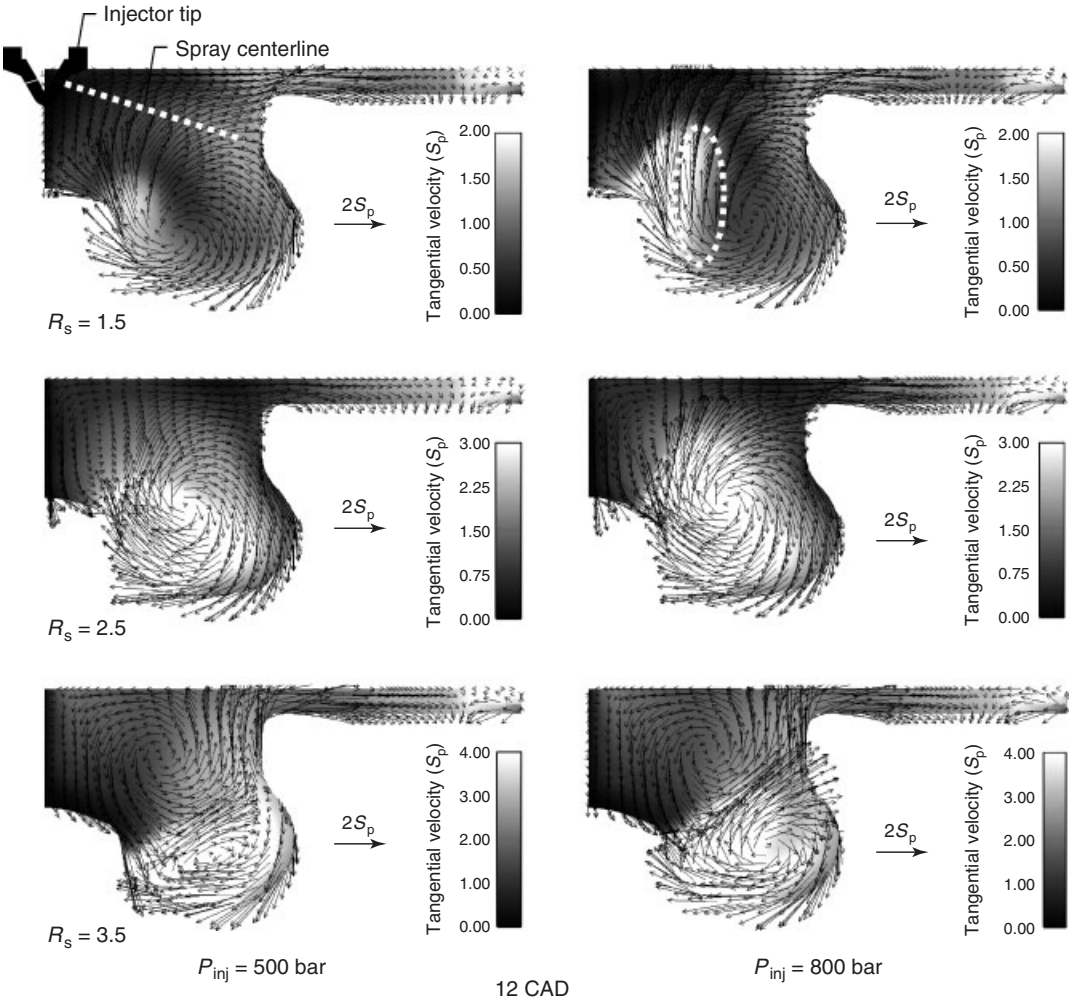


Figure 12. Flow structures in the vertical plane containing the spray centerline illustrating the interaction between the fuel spray and the swirl in bowl-in-piston engines. (Reproduced from *Flow and Combustion*, 2009, Ch. 2, Flow, mixture preparation, and combustion in direct-injection two-stroke gasoline engines, Fansler and Drake. With kind permission of Springer Science+Business Media.)

need for increased swirl ratio to obtain optimum low speed performance in many diesel engines.

Three additional observations related to the spray–swirl interaction are relevant. First, when high angular momentum fluid is transported inward the angular momentum of the fluid is conserved, but not the kinetic energy—which is theoretically expected to increase. Conceptually, the kinetic energy of the spray is employed to perform work on the fluid in transporting it inward. Thus, the spray–swirl interaction provides a mechanism by which the energy of the injection event can be stored within the mean, rotational energy of the fluid for later release as TKE. Second, the process of rearranging the in-cylinder angular momentum can lead to negative radial gradients in angular momentum; an example is shown by the dashed ellipsoid in the upper-right portion of Figure 12. Such flow patterns are inherently unstable—any radial perturbation of the position of a fluid element will be amplified (e.g., Bradshaw, 1973; Tritton, 1977). Consequently, the flow can be expected to very rapidly and efficiently produce turbulence, as will be discussed further in subsequent text. Third, at the interface between the two vortical structures, there are sharp gradients in the swirl velocity, as well as significant flow deformation in the vertical (r–z) plane. These attributes are likewise expected to result in high levels of turbulence production.

4.2.6 Toroidal vortex structures

A final example of strongly rotating flows in engines is the toroidal vortex, or vortex ring, which is a very

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

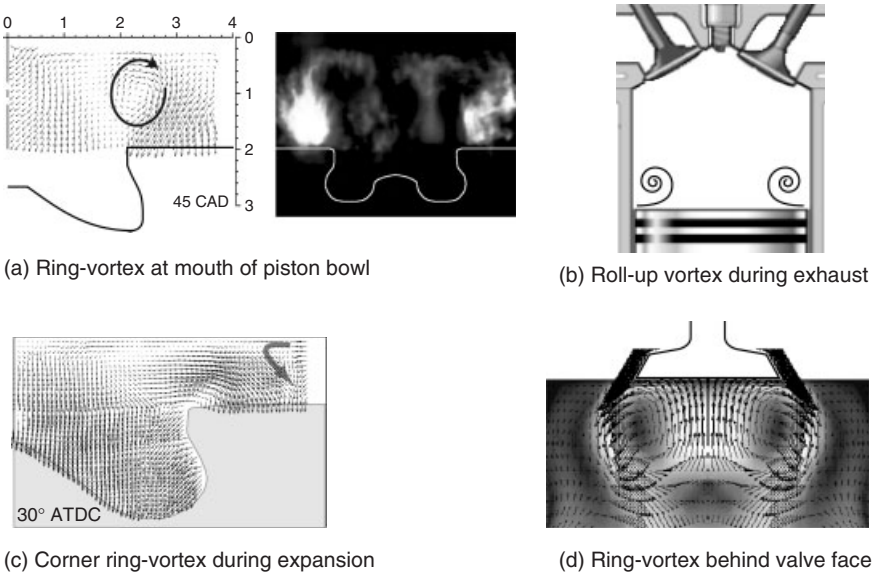
stable flow structure that can be produced in a variety of ways; the essential requirement for its production is that linear momentum should be imparted to a fluid in an axisymmetric manner (Batchelor, 1967). This structure is a very common feature of engine flows, in large part due to the axisymmetric geometry of cylinders and piston bowls and the approximately cylindrical geometry of valves and ports. One of the striking characteristics of a vortex ring is its persistence and stability. An example is shown in Figure 13a, which depicts a vortex ring formed at the piston bowl mouth by hot oxidation products leaving the bowl. It forms by 25° after TDC, and clearly persists until at least 50–55° after TC. The vortex forms a flow barrier separating and impeding mixing between fluid leaving the bowl and the remaining charge within the squish volume, causing the expanding squish volume to be filled from the top of the cylinder. Images of soot luminosity further suggest that this vortex may directly impact the mixing and further oxidation of soot in the squish volume of diesel engines.

Additional examples of toroidal vortices are also shown in Figure 13. The "roll-up" vortex (b), formed as the rising piston scrapes stagnant boundary layer fluid off the cylinder wall during the exhaust stroke, can grow to a diameter of approximately 15% of the bore. It is thought to contain a sizable fraction of the exhaust hydrocarbons emitted from SI engines. Unburned hydrocarbons have also been observed in a similar corner vortex in the upper portion of the cylinder in an optically accessible engine with a large ring-land crevice (c). CFD simulations, however, indicate that this vortex will be much less pronounced in engines with smaller crevice volumes. Finally, very distinct ring-vortex structures are predicted to form behind the intake valve face during induction (d). The experimental detection of these structures has been noted in the discussion of the intake jet flows in SI engines.

Notice that these ring-vortex structures can have rotation rates that significantly exceed the rotation rates of typical swirl or tumble structures. For example, ring vortices at the mouth of the piston bowl rotating at over 10 times the crankshaft speed have been measured in light-duty engines. As will be shown in the next section, rotation of this magnitude will clearly impact the structure of the turbulence.

4.2.7 *The impact of swirl on turbulence: theoretical considerations*

The effects of rotation on turbulence are very profound and complex (Sagaut and Cambon, 2008; Mathieu and Scott, 2000). Although analysis in a rotating coordinate system shows that for homogeneous turbulence, the equation governing the local TKE balance (Equation 8) is unchanged, rotation can still impact the turbulence energy through changes in both the production $\mathcal{P}$ and the dissipation ε. The production is impacted through the Reynolds stresses, while the dissipation is impacted by the tendency of rotation to inhibit the transfer of energy in the turbulent energy cascade, thereby reducing the dissipation rate.



(a) Ring-vortex at mouth of piston bowl

(b) Roll-up vortex during exhaust

(c) Corner ring-vortex during expansion

(d) Ring-vortex behind valve face

Figure 13. (a–d) Toroidal vortices observed in engine flows. The vortex behind the valve face (d) is a CFD simulation result. (Reproduced by permission of Christopher J. Rutland (University of Wisconsin).)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

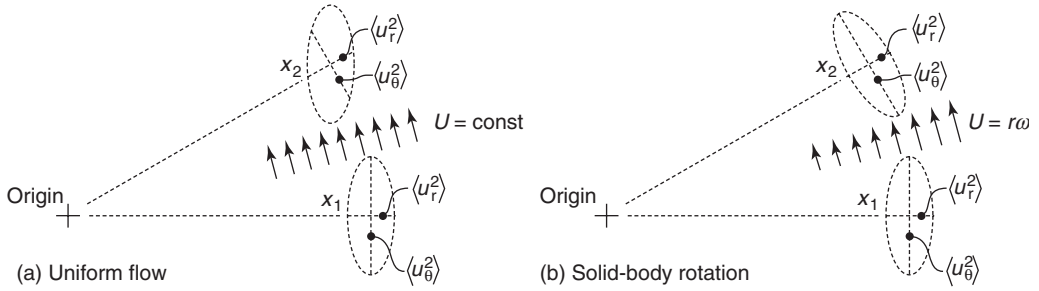


Figure 14. Examples of the impact of basis vector rotation on the individual components of the turbulent stress tensor. In (a) the production is zero and the stresses vary solely because of basis vector rotation. In (b) the effects of basis vector rotation and production cancel to leave the r–θ components of the stress tensor unchanged.

Here we pursue a simplified analysis of the production term, employing a cylindrical coordinate system within a stationary (inertial) reference frame. Before proceeding with the analysis, it will be useful to review briefly how the various Reynolds stress components vary as a result of mean flow convection in a cylindrical coordinate system. Consider first a steady, 2D flow in which the turbulence is transported from location $\mathbf{x}_1$ to location $\mathbf{x}_2$ by a uniform mean velocity field. We represent the turbulent stress tensor as an elliptical object shown on the left-hand side of Figure 14, where the minor axis of the ellipse at location $\mathbf{x}_1$ represents the magnitude of the radial stress $\langle u_r^2 \rangle$ and the major axis represents the tangential stress $\langle u_\theta^2 \rangle$. Because the velocity gradients are zero, the production of the individual Reynolds stresses is zero. Nevertheless, when expressed in cylindrical coordinates, the magnitude of both the normal stresses and the shear stress has clearly changed from location $\mathbf{x}_1$ to $\mathbf{x}_2$—an artifact of the changing orientation of the coordinate system basis vectors. Next, consider the right-hand-side of Figure 14, which depicts turbulence in equilibrium in a cylindrical coordinate system, where the mean flow is solid-body like and the radial and tangential stresses are unchanging. In this case, the production of $\langle u_r^2 \rangle$ and $\langle u_\theta^2 \rangle$ is nonzero, but the production-like terms that arise as a result of coordinate system rotation exactly cancel the production of the individual stress components such that the stresses at $\mathbf{x}_1$ and $\mathbf{x}_2$ are equal. These two examples clearly illustrate that both the true stress production and the production-like terms associated with the changing basis vector orientation must be considered jointly when we examine the variation in the turbulent stresses in an r–θ coordinate system. Note, however, that the terms arising from the changing basis vectors do not contribute to the production of TKE, they simply redistribute it among the various components.

As we will see in subsequent text, in swirling flows a dominant source of turbulence is often associated with production associated with r–θ plane velocity gradients, which are dominated by radial gradients in the tangential velocity component (the swirl velocity). Under these conditions, the production of k is

$$\mathcal{P}_f = -\langle u_r u_\theta \rangle r \frac{\partial}{\partial r}\left(\frac{\langle U_\theta \rangle}{r}\right) \tag{14}$$

Production of k can be further separated into terms that feed energy into the tangential and the radial fluctuations, and combined with the terms arising from coordinate system rotation, to yield

$$\mathcal{P}_{f,\theta} = -\langle u_r u_\theta \rangle \frac{\partial \langle U_\theta \rangle}{\partial r} - \langle u_r u_\theta \rangle \frac{\langle U_\theta \rangle}{r} = -\langle u_r u_\theta \rangle \frac{1}{r}\frac{\partial (r\langle U_\theta \rangle)}{\partial r} \tag{15}$$

and

$$\mathcal{P}_{f,r} = \langle u_r u_\theta \rangle \frac{\langle U_\theta \rangle}{r} + \langle u_r u_\theta \rangle \frac{\langle U_\theta \rangle}{r} = \langle u_r u_\theta \rangle \frac{2\langle U_\theta \rangle}{r} \tag{16}$$

For most engine flows, $\partial(\langle U_\theta \rangle / r)/\partial r < 0$, and a positive production of turbulence energy requires that $\langle u_r u_\theta \rangle$ be positive. Lumley (1999) argues that, provided the radial gradient in the mean flow angular momentum is positive, $\langle u_r u_\theta \rangle$ will be negative—indicating that swirl will suppress turbulence. The reasoning supporting this argument is that a positive radial excursion of a fluid particle, during which angular momentum is conserved, will lead to a local deficit in angular momentum at the new radius and hence, on average, a negative u_θ. This physical reasoning finds support in the RANS equation governing the evolution of $\langle u_r u_\theta \rangle$, which under the same assumptions leading to Equation 14 and accounting for changing basis vector orientation, shows that the production of $\langle u_r u_\theta \rangle$ is

$$\mathcal{P}_{r\theta} = -\langle u_r^2 \rangle \frac{1}{r}\frac{\partial (r\langle U_\theta \rangle)}{\partial r} + \langle u_\theta^2 \rangle \frac{2\langle U_\theta \rangle}{r} \tag{17}$$

Lumley's reasoning is embodied in the first term, where we recognize $\partial(r\langle U_\theta \rangle)/\partial r$ as the angular momentum gradient. However, the production of negative $\langle u_r u_\theta \rangle$ indicated

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

by the first term is counterbalanced by the second term, which is always positive. The form of the second term, by comparison with Equation 4, suggests that it is associated with centrifugal forces. Indeed, a positive fluctuation in u_θ would be expected to lead to a positive fluctuation in u_r, due to the increased centrifugal forces acting on a fluid element—as has been argued previously by Holloway and Tavoularis (1992). Equation 17 further shows that if the mean swirl velocity profile is solid body like, and the normal stresses are equal, then both terms are of equal magnitude. There is thus no compelling reason to expect $\langle u_r u_\theta \rangle < 0$. In fact, if $U_\theta \propto r^n$ and the normal stresses are equal, we expect from Equation 17 that $\langle u_r u_\theta \rangle > 0$ for $n < 1$, which is typical of local swirl velocity profiles seen in engines. If perchance $n > 1$, then $\partial(\langle U_\theta \rangle / r)/\partial r > 0$, and $\mathcal{P}$ is again positive. As we will see in subsequent text, direct measurements of $\langle u_r u_\theta \rangle$ in an automotive diesel show that it is often positive. Accordingly, swirl might more typically be expected to augment turbulence, a conclusion consistent with many of the measurements reported in the literature (Hill and Zhang, 1994).

Finally, note from Equation 17 that when the angular momentum gradient is negative, we expect $\langle u_r u_\theta \rangle$ to be positive and large, as both terms reinforce each other. Likewise, $\partial(\langle U_\theta \rangle / r)/\partial r$ will be large and negative, and Equation 14 therefore indicates that turbulent energy production will be large and positive. As we have seen earlier, negative angular momentum gradients can be readily produced in engine flows through interaction of the fuel injection process with the rotating motions.

The potential for the mean flow rotation to affect the turbulence by inhibiting the transfer of energy in the turbulent energy cascade can be evaluated by computing a nondimensional number, the Rossby number Ro, which compares the turbulence time scale to a rotation time scale ($1/2\omega$). With $l \approx b/6$, $u \approx \overline{V}_p$, and $b \approx S$, one gets: $\mathrm{Ro} = \frac{(1/2\omega)}{(l/u)} = \frac{u}{2\omega l} = \frac{3}{\pi R_s}$ or $\frac{3}{\pi R_t}$. For a swirl or tumble ratio of order 2, we therefore obtain $\mathrm{Ro} \approx 1/2$. Direct numerical simulations, reviewed by Sagaut and Cambon (2008), indicate that when $\mathrm{Ro} \lesssim 1$ rotation can significantly impact the structure of the turbulence. Consequently, modeling approaches that account for the impact of rotation (see e.g., Bardina, Ferziger, and Rogallo, 1985 for a RANS-based example) should be considered.

4.2.8 *The impact of swirl on turbulence: measurements*

The majority of the measurements made in swirling flows have been obtained in swirl-supported CI engine geometries, and the spatial and temporal properties of turbulence in these engines have been reviewed recently by Miles (2009). Here, we provide only a brief summary of the most important features and processes. Because swirling flows and flows driven by the rapidly changing combustion chamber geometry of bowl-in-piston engines are less susceptible to CCVs, in this section we treat measured velocity fluctuations as predominantly associated with turbulence. There are three principal observations supporting this practice. First, the results of frequency analysis of velocity measurements made in a typical firing automotive diesel engine (Miles *et al*., 2002) show that small-scale, high frequency flow fluctuations follow proportionally the evolution of the total flow fluctuations. Second, as will be seen in subsequent text, the evolution, anisotropy, and production of the measured fluctuations are remarkably consistent with the behavior that is predicted by the RANS equations governing the evolution of the turbulence energy and the individual component stresses. Third, the measured turbulent stresses compare very well with the predictions of models derived in canonical, ergodic flows (Miles, RempelEwert, and Reitz, 2009).

At low swirl ratios ($R_s \lesssim 1.5$), the turbulence is approximately homogeneous in the latter portion of the compression stroke, and remains homogeneous throughout the remainder of compression and early expansion. This is a direct consequence of the dominance of $\mathcal{P}_{comp}$, which depends on the existing turbulence energy and the rate of compression—both of which are spatially homogeneous. Measurements and simulations (Miles *et al*., 2004) show that near the end of compression (25° to 5° bTDC), turbulence production associated with r–θ plane production (Equations 14–16) and production associated with axial gradients in the swirl velocity become more important but do not significantly impact the spatial homogeneity. Surprisingly, the turbulence production by shear at the edges of the squish flow jets (i.e., production associated with $\langle u_r u_z \rangle \partial \langle U_r \rangle / \partial z$) is predicted to be small.

As the swirl ratio is increased ($R_s \sim 2.5$), significant inhomogeneity in the turbulence energy develops, with the highest levels near the piston bowl lip (Figure 15). Measurement and/or simulation of the various production terms near the lip shows that while $\mathcal{P}_{comp}$ continues to dominate before $\sim -25°$, r–θ plane production is a significant source of turbulence just before TDC. Note that the rate of change of TKE correlates closely with the measured r–θ plane net production rate. At this time, the turbulence in the near lip region is highly anisotropic, with the tangential fluctuations exceeding the radial fluctuations. Similar peaks in turbulence and anisotropy near the bowl lip have been reported in other studies (e.g., Fansler and French, 1987; Auriemma *et al*., 1998). The observed anisotropy is fully consistent with the behavior expected from Equations 15–17. Inward transport and "spin up" of high angular

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

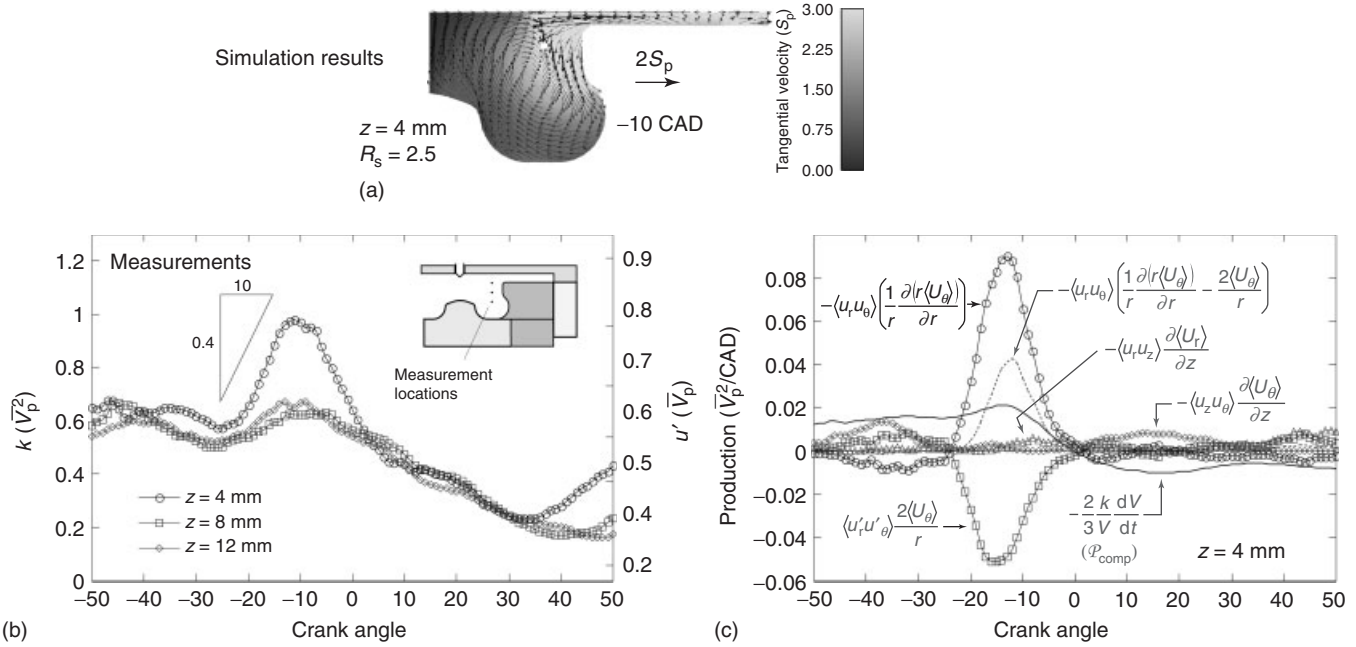


Figure 15. (a) Simulation results depicting the spatial distribution of the swirl velocity and the vertical plane flow structure; (b) the measured evolution of the turbulent kinetic energy; (c) turbulence production terms near the bowl lip ($z = 4$ mm, the upper measurement location designated in "(b)". $\mathcal{P}_{comp}$ and the r–θ plane production terms are estimated directly from the measurements; the eddy viscosity hypothesis is used to estimate $\langle u_z u_\theta \rangle$ from the measured $\partial \langle U_\theta \rangle / \partial z$, and $-\langle u_r u_z \rangle \partial \langle U_r \rangle / \partial z$ is obtained from simulations.

momentum fluid from the squish volume creates a large positive angular momentum gradient near the bowl lip. The angular momentum gradients force $\langle u_r u_\theta \rangle$ to become negative, leading to significant positive production of the tangential velocity fluctuations. In contrast, the component of the r–θ plane production that feeds the radial fluctuations is negative.

As with the results at the lower swirl ratio, turbulence production *directly* associated with the squish flow ($\langle u_r u_z \rangle \partial \langle U_r \rangle / \partial z$) is predicted to be small. However, the squish flow *indirectly* impacts the production of turbulence through its redistribution of angular momentum, leading to high rates of turbulence production associated with r–θ plane velocity gradients and with $\partial \langle U_\theta \rangle / \partial z$.

With further increases in swirl ratio (to $R_s \sim 3.5$), turbulence levels and anisotropy near the bowl lip continue to increase. However, near-TDC increases in turbulence are also observed lower in the bowl, where the radial fluctuations increase rapidly and eventually exceed the tangential fluctuations. These fluctuations are again being produced by r–θ plane flow deformation, but in this case angular momentum gradients are small, $\langle u_r u_\theta \rangle$ is positive (cf. Equation 17), and the production of radial component fluctuations is positive while the production of tangential component fluctuations is negative (cf. Equations 15 and 16).

To close this subsection, we mention a curious but potentially important phenomenon that has been measured in both SI research engines with a "pancake" combustion chamber at TDC (e.g., Reuss *et al*., 1995) and in CI engine configurations above the piston top during compression (Figure 16). With moderate-to-high swirl ratios ($R_s \gtrsim 2.5$), fluctuating velocities are observed to decrease strongly at intermediate radii, before again increasing as the high shear regions near the wall are approached. As discussed earlier, there is no clear reason to expect negative turbulence production in these flows and further work is needed to ascertain the cause of this behavior. However, considering that swirl is now a common feature in direct-injection SI engines, and that flame speeds and burn times are strongly dependent on the fluctuating velocity, understanding and correctly modeling this behavior is clearly important.

4.2.9 *Turbulence generation through spray–swirl interactions*

In the discussion of Figure 12, we have illustrated the ability of the fuel injection process to impact the angular momentum distributions within the combustion chamber, and in the previous section have shown that the inward transport of angular momentum by the squish flow can lead to significant turbulence production. It is reasonable

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

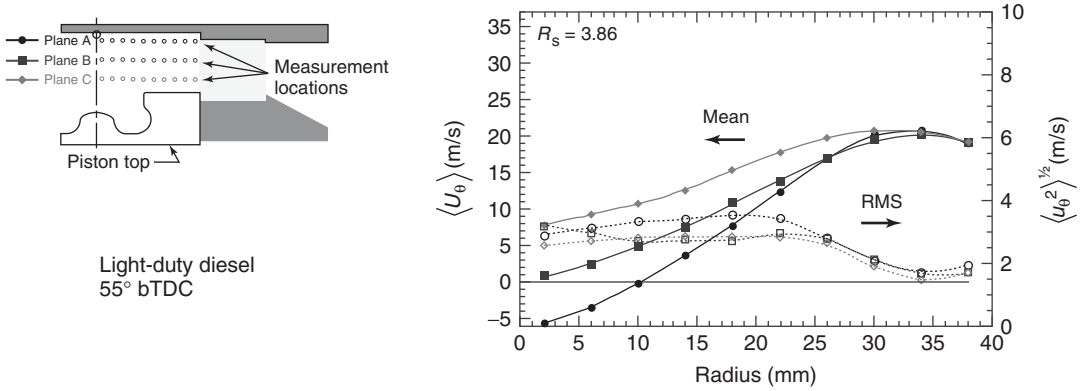


Figure 16. Radial profiles of the mean and RMS tangential velocities, demonstrating the attenuation of the fluctuations as the radius increases.

to expect, therefore, that the fuel injection process can be fruitfully employed to create mean flow structures that will generate additional turbulence. To isolate the impact of the spray on the turbulence field, the studies described in this section were performed in an engine operating with an N_2 intake charge. Consequently, complicating factors associated with cyclic variability in the ignition and combustion process, or in turbulence generation associated with combustion-induced density variations are avoided.

One mechanism by which turbulence can be generated efficiently is by formation of negative radial gradients in mean flow angular momentum, as was discussed in connection with Figure 12. Figure 17 shows experimental estimates of turbulence energy derived from two-component laser velocimetry measurements made with the engine operating at a simulated idle condition. As the measurement location approaches regions where simulations indicate that a negative angular momentum gradient is expected, the turbulence energy increases by an order of magnitude.

To further illustrate the impact of r–θ plane turbulence production, and angular momentum gradients in particular, the left-hand column of Figure 18 shows the evolution of the normal stresses and the components of the r–θ plane production given by Equations 15 and 16 ($\mathcal{P}_{f,\theta}$ and $\mathcal{P}_{f,r}$). The production terms were calculated from measurements of $\langle u_r u_\theta \rangle$ and of r–θ plane velocity gradients within the piston bowl of an engine operating with a typical swirl ratio of $R_s = 2.5$. These measurements are shown in the right-hand column of Figure 18. Our first observation is that in contrast to the production rates shown in Figure 15, during the period of maximum turbulence production, both the production of $\langle u_r^2 \rangle$ and of $\langle u_\theta^2 \rangle$ are positive. As required by Equations 15 and 16, this is due to the formation of a negative angular momentum gradient at this time, which can be seen in both the measured temporal profiles and the simulated velocity field shown in the right-hand column of Figure 18. A second observation is that the radial fluctuations are increased more than the tangential

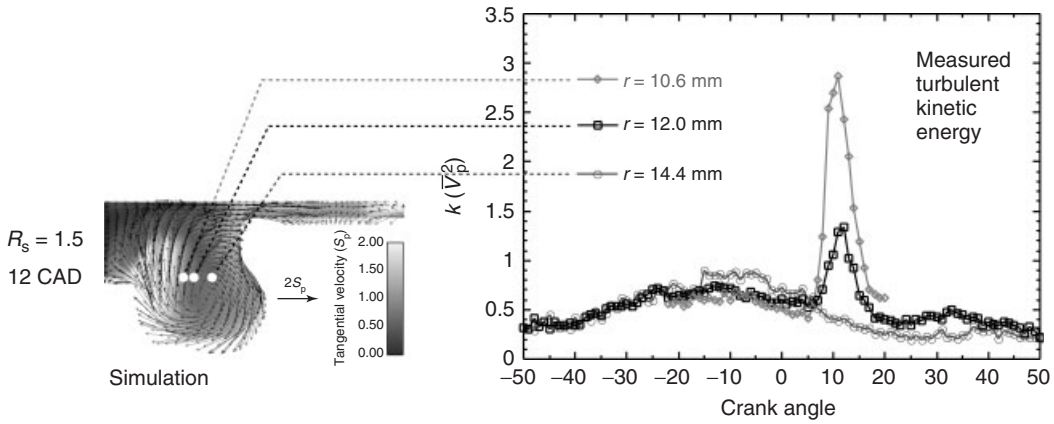


Figure 17. Variation in measured turbulent kinetic energy as the measurement location approaches regions with negative radial gradients in the mean flow angular momentum.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

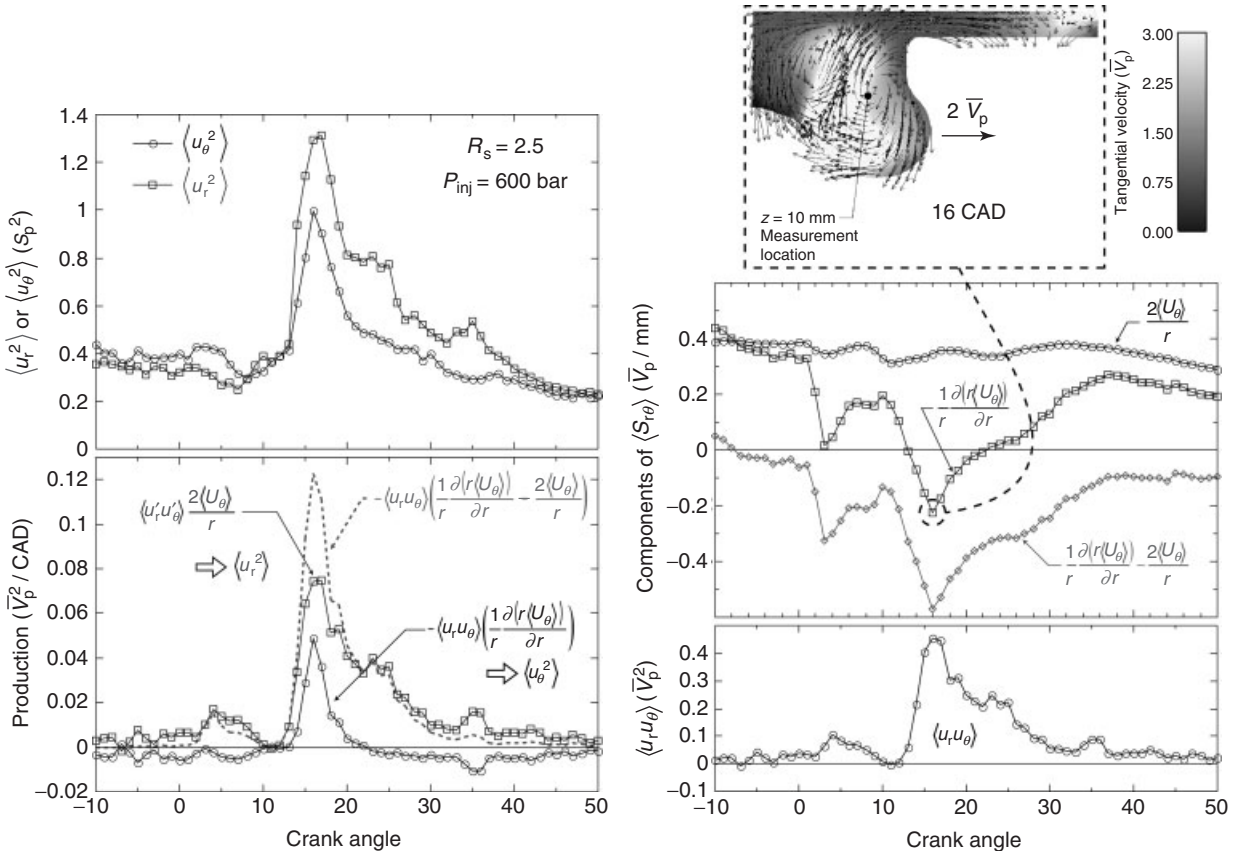


Figure 18. Variation in measured turbulent stresses and in the measured production terms and $r-\theta$ plane deformation rates at $R_s = 2.5$.

fluctuations, and remain elevated longer, in accordance with the measured difference in the production terms. Lastly, we also observe that $\langle u_r u_\theta \rangle$ shows a sharp, positive peak, while the angular momentum gradient is negative, as expected from Equation 17.

Although the correlation between the turbulent stresses and the turbulence production shown in Figure 18 is very good, we stress that the instantaneous, local value of the stresses will also be influenced by factors other than the $r-\theta$ plane production, especially in these highly inhomogeneous flows. Consider, for example, that the peak rate-of-change of turbulence energy that can be deduced from Figure 18 is $\sim 0.4-0.5\overline{V}_p^2/^\circ$CA, while the peak production is only 25% of this value. Moreover, the peak rate-of-change in energy is phased slightly earlier than the peak in production. The measurements indicate, however, that steep radial gradients in turbulence production are present, and that the $r-\theta$ plane production at a location just 1 mm inward from that shown in Figure 18 is over twice as great. In addition, unmeasured production terms, mean flow convection, and turbulent transport (cf. Equation 8) will also impact the overall turbulent energy, whereas redistributive terms associated with the fluctuating pressure–strain-rate correlation will further influence the relative magnitude of the individual normal stresses.

No direct measurements exist to illustrate the mechanisms of turbulence production associated with $r-z$ and $z-\theta$ flow deformation. However, simulation results indicate that significant sources of turbulence production, exceeding $0.1\overline{V}_p^2/^\circ$CA, are found near the interface between the two vertical plane vortices illustrated in Figure 12. These sources are associated with both $r-z$ plane deformation and with strong axial gradients in the swirl velocity, and are especially pronounced at higher swirl ratios. Examples are depicted in Figure 19, which also identifies sources of $r-\theta$ plane deformation, including those discussed earlier. As is discussed further in Diesel and Diesel LTC Combustion, the interfacial region between these two vortices can often correspond closely with the interface between partially oxidized fuel and additional air, and turbulence generated in this region can therefore directly impact fuel–air mixing rates and subsequent combustion.

We close this section by noting that in addition to employing the fuel injection event to enhance turbulence

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

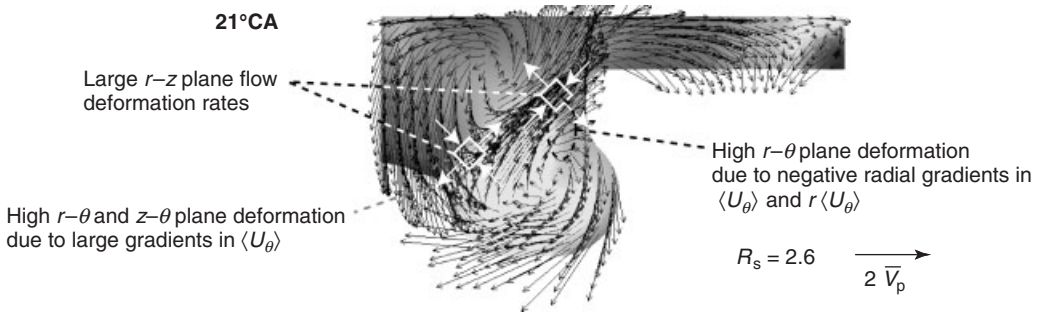


Figure 19. Flow features in the bowl of a light-duty diesel engine having significant turbulence generation potential. The square boxes/arrows depict $\langle S_{rz} \rangle$ in principal axes.

generation and mixing through the portion of the production that depends on flow structure $\mathcal{P}_f$, the fuel injection event can also be timed to take advantage of $\mathcal{P}_{comp}$. Although the intense turbulence generated by the fuel injection event will normally dissipate rapidly, compression can amplify this turbulence and extend the time over which it will impact fuel–air mixing. A simple calculation applying Equation 8 to homogeneous turbulence, considering $\mathcal{P}_{comp}$ to be the only source of production and modeling ε as proportional to $k^{3/2}$, indicates that "injecting" turbulence into the engine roughly 30–40°CA before TDC will maximize the turbulence energy when integrated over the 30°CA following the injection event.

4.2.10 The impact of combustion on turbulence

The general topic of the impact of turbulence on SI combustion is covered in Fundamental Combustion Modes, and the impact of increased early- and late-cycle mixing processes on CI combustion is addressed in Diesel and Diesel LTC Combustion. Accordingly, we do not discuss combustion-flow interactions in detail here. Nevertheless, given the large influence that the fuel injection event has on the generation of turbulence without combustion, it is relevant to briefly examine if these processes continue to operate with combustion. Figure 20 compares, for a variety of injection pressures and swirl ratios, the evolution in TKE measured with fuel injection only (right-hand side) against the energy measured with fuel injection and combustion (left-hand side). The most striking observation is that the form of the temporal evolution and the trends observed as both injection pressure and swirl ratio are varied are very similar for both cases—although with combustion the increase in energy is larger. A second feature worth noting is that the combustion event, dominated by a rapid premixed burn near 5° after TDC, results in only a small increase in the fluctuating energy which rapidly decays.[4] Apart from this small increase, combustion and heat release, which adds sensible energy but not momentum to the flow, does not appear to fundamentally alter the main turbulence generation processes.

With combustion, however, additional turbulence production mechanisms are present, but these that have not yet been discussed—in particular buoyant production associated with density inhomogeneities embedded in the swirl-induced radial pressure gradient. An order-of-magnitude analysis (Miles *et al*., 2002) indicates that buoyant production may be at least as large as production associated with r–θ plane flow deformation. Apart from turbulence production, buoyancy can also play a role in generating temperature or composition stratification in engines. Lumley (1999) shows that buoyant forces can be expected to transport hot combustion products from the cylinder periphery to the cylinder center in approximately 1/5 of the time required for a piston stroke, and that even without combustion swirl can be effective in promoting fuel stratification because of the higher density of the fuel-containing mixture.

The discussion of turbulence generation in swirling flows relates particularly to light-duty diesel engines operating at light load. However, profiles of swirl velocity measured in heavy-duty engines with swirl-supported combustion systems operating at medium and high loads also show evidence of strong production of turbulence by r–θ plane flow deformation. Dembinski and Ångström (2012) measure a marked decrease in angular velocity within the piston bowl, from peak values approaching 20 times the crankshaft velocity near the bowl centerline to levels of twice the crankshaft velocity at the bowl periphery. Although the mechanism by which these angular velocity gradients are created has not been clarified, it is almost certainly associated with the inward transport of high angular momentum fluid due to displacement by the charge entrained into the fuel jets and with volume expansion due to combustion at the bowl periphery.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

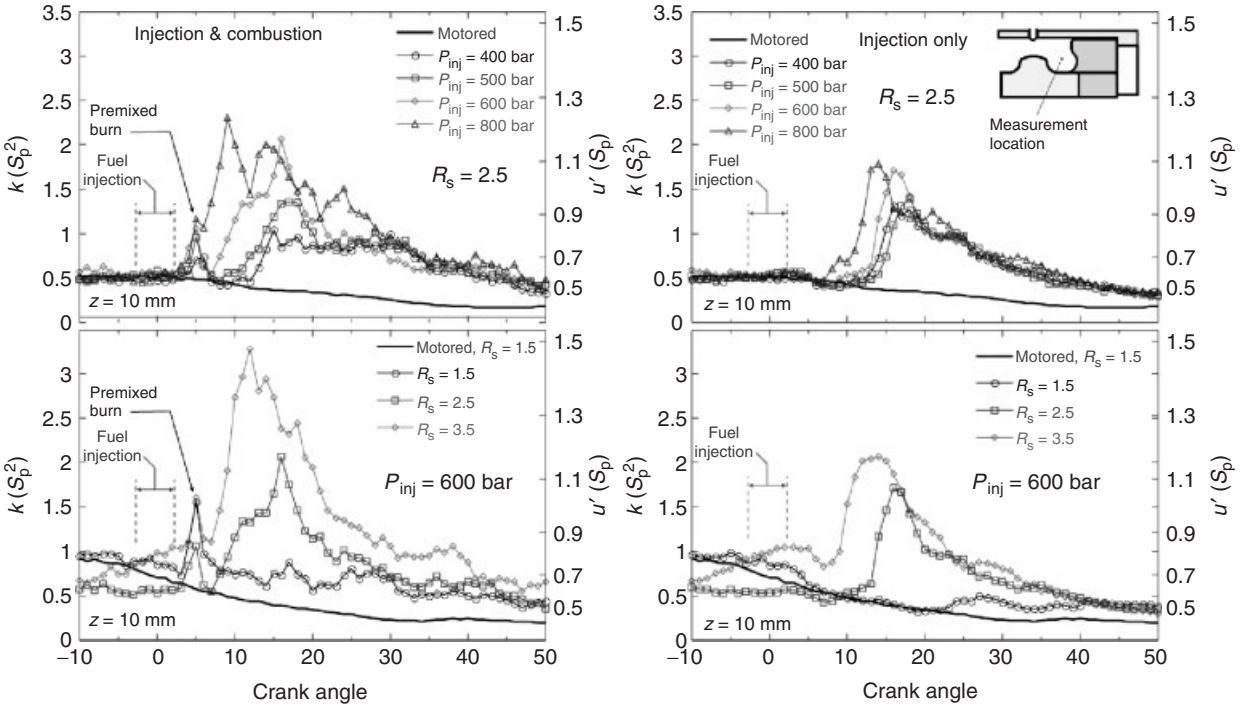


Figure 20. Evolution of turbulent kinetic energy with and without combustion at a simulated idle condition. The energy is estimated from two-component laser Doppler velocimetry measurements.

5 SUMMARY AND CONCLUDING REMARKS

We end this chapter by restating the main properties of in-cylinder flows: they are compressed, rotating, and turbulent. It is our hope that we have provided the background needed to understand some of the basic mechanisms by which these three fundamental properties can be influenced by the engine designer, and that we have provided the engine research community with some guidance as to fruitful areas for future work.

If the flow energy created during the intake process is to be employed to promote mixing and combustion near TDC, it must be stored in large-scale, mean flow structures such as tumble and swirl. Bulk flow compression, which is often considered only in the context of changing the thermodynamic state of the in-cylinder charge, can then act to modify these flow structures. Compression provides a mechanism by which piston motion augments the mean flow kinetic energy stored in tumbling flow structures, deforms and ultimately causes the breakdown of these structures into small-scale turbulence, and then further amplifies the turbulence thus created. Consequently, compression ratio—the selection of which is usually guided primarily by idealized efficiency, knock, ignition delay, and cold-start considerations—also has a significant impact on flow processes, impacting the magnitude and the phasing at which piston motion augments the flow energy. Changes in turbulence energy can go on to impact both burning rates and heat transfer rates, which also affect engine efficiency. Considerable progress has been made in understanding the detailed mechanisms by which these processes occur, but substantial further work needs to be done to understand the topology of desirable flow structures, how to most efficiently form these structures, how to reduce the cycle-to-cycle variability in the formation and breakdown processes, how to phase the breakdown to obtain the most benefit to the ensuing combustion, and how to best take advantage of the interaction of these structures with the fuel sprays in direct-injection engines.

Mean flow swirl has also been shown to impact turbulence generation near TDC, even though bulk compression does not directly impact the energy or stability of the mean flow. In the case of swirl, secondary flow processes such as squish flows or spray-driven flows redistribute angular momentum within the cylinder. It is this redistribution that creates significant departures from solid-body-like velocity profiles, with subsequent high flow deformation and turbulence generation. In some cases, unstable angular momentum distributions that break down rapidly into turbulence are formed. Bulk compression can then act to amplify the turbulence generated. Similar to the tumble

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

flow formation and breakdown, there has been significant progress made in understanding how swirl and squish–swirl or spray–swirl interactions can impact flow turbulence, but much more remains to be done to identify how these processes can best be optimized and adapted as engine speed and load vary.

Finally, it should be clear from the complexity of typical SI and CI induction, compression, and turbulence generation processes that optimization of in-cylinder flow processes—considering flow losses, cyclic variability, turbulence levels, heat transfer, and the ensuing combustion and emissions formation processes—is a daunting task that will be unlikely to be accomplished with "cut-and-try" approaches. Providing the engine designer with a physical understanding of the important flow processes, a task we have tried to accomplish in this chapter, is an important part of achieving this optimization. However, another essential task will be to provide the designer with tools that can be used to accurately test his ideas through truly predictive simulation. Hence, we encourage the continued development of models needed for both RANS-based simulations and LES—such as the wall functions described earlier. We further see need for acquisition of detailed planar or full-field, spatially and temporally resolved experimental data and the development of specific quantitative analysis tools associated with comprehensive physical filters that will support the development of the models and understanding needed to accomplish this goal.

ACKNOWLEDGMENTS

The authors acknowledge the contributions of the research group of Professor Rolf D. Reitz at the University of Wisconsin (Madison), who provided all of the simulation results represented in the figures in this chapter. We are also grateful to Dr. Rémi Manceau of CNRS—University of Poitiers, who kindly reviewed the sections related to production of turbulence in swirling flows and offered helpful suggestions on how it might be clarified. The first author acknowledges his long-term collaboration with Renault SA and the contributions of several PhD students (Y. Cao, I. Cossadia, S. Maurel, and M. Voisine). Support for the second author was provided by the United States Department of Energy (Office of Vehicle Technologies) and by General Motors Corporation (agreement FI083070326). This work was performed at the Combustion Research Facility of Sandia National Laboratories in Livermore, California. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

ENDNOTES

1. Full affiliation of Prof. J. Borée: Institut Pprime, UPR-3346 CNRS, ISAE-ENSMA, Université de Poitiers, Futuroscope Chasseneuil, France.
2. We adopt the usual convention that a repeated index indicates summation over all velocity components.
3. Heywood (1988) provides expressions for the squish velocity in wedge-chamber geometries.
4. Frequency analysis indicates that the combustion event generates small-scale, high frequency fluctuations that likely correspond to true turbulence, not simply cycle-to-cycle fluctuations associated with variations in the phasing or location of the combustion.

REFERENCES

Abraham, J., Williams, F.A., and Bracco, F.V. (1985) A discussion of turbulent flame structure in premixed charges. SAE Technical Paper 850345.

Alharbi, A.Y. and Sick, V. (2010) Investigation of boundary layers in internal combustion engines using a hybrid algorithm of high speed micro-PIV and PTV. *Experiments in Fluids*, **49** (4), 949–995.

Amsden, A.A., O'Rourke, P.J., and Butler, T.D. (1989) KIVA II: a computer program for chemically reactive flows with sprays. Los Alamos National Laboratory Report LA-11560-MS.

Arcoumanis, C., Hadjiapostolou, A., and Whitelaw, J.H. (1991) Flow and combustion in a hydra direct-injection diesel engine. SAE Technical Paper 910177.

Arcoumanis, C., Z. Hu, and J. H. Whitelaw (1990) Tumbling motion: a mechanism for turbulence enhancement in spark-ignition engines. SAE Paper 900060.

Auriemma, M., Corcione, F.E., Macchioni R., and Valentino, G. (1998) Interpretation of air motion in reentrant bowl-in-piston engine by estimating Reynolds stresses. SAE Technical Paper 980482.

Bardina, J., Ferziger, J.H., and Rogallo, R.S. (1985) Effect of rotation on isotropic turbulence: computation and modeling. *Journal of Fluid Mechanics*, **154**, 321–336.

Batchelor, G.K. (1967) *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge.

Bradshaw, P. (1973) Effects of streamline curvature on turbulent flow. AGARDograph 169, AD-768316, Advisory Group for Aerospace Research and Development, North Atlantic Treaty Organization, Paris.

Bonnet, J.P. and Delville, J. (2001) Review of coherent structures in turbulent free shear flows and their possible influence on computational methods. *Flow, Turbulence and Combustion*, **66**, 333–353.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Bonnet, J.P., Delville, J., Glauser, M.N., *et al*. (1998) Collaborative testing of eddy structure identification methods in free turbulent shear flows. *Experiments in Fluids*, **25**, 197–225.

Borée, J., D. Marc, R. Bazile, and B. Lecordier (1999) On the Behaviour of a Large Scale Tumbling Vortex Flow Submitted to a Compression. *European Series in Applied and Industrial Mathematics ESAIM Proceedings*, Vol. 7. http://www.emath.fr/Maths/Proc (accessed 3 September 2013).

Borée, J., S. Maurel, and R. Bazile (2002) Disruption of a compressed vortex. *Physics of Fluids* Vol. 14, 7 2543–2556. (Prof. J. Lumley 70th birthday Symposium Papers).

Brandstätter, W., Johns, R.J.R., and Wigley, G (1985) The effect of inlet port geometry on in-cylinder flow structure. SAE Technical Paper 850499.

Cambon, C., Mao, Y., and Jeandel, D. (1992) On the application of time dependent scaling to the modelling of turbulence undergoing compression. *European Journal of Mechanics, B Fluids*, **11** (6), 683–703.

Chen, M., Zhang, W., Zhang, X., and Ding, N. (2011) In-cylinder CFD simulation of a new 2.0 L turbocharged GDI engine. SAE Technical Paper 2011-01-0826.

Comte-Bellot, G. and Corrsin, S. (1971) Simple-Eulerian time correlation of full- and narrow-band velocity signals in grid-generated, 'isotropic' turbulence. *Journal of Fluid Mechanics*, **48**, 273–337.

Cosadia, I., Borée, J., Charnay, G., and Dumont, P. (2006) Cyclic variations of the swirling flow in a diesel transparent engine. *Experiments in Fluids*, **41** (1), 115–134.

Dannemann, J., Pielhop, K., Klass, M., and Schröder, W. (2010) Cycle resolved multi-planar flow measurements in a four valve combustion engine. *Experiments in Fluids*, **50** (4), 961–976.

Dembinski, H.W.R. and Ångström, H.-E. (2012) Optical study of swirl during combustion in a CI engine with different injection pressures and swirl ratios compared with calculations. SAE Technical Paper 2012-01-0682.

Depardon, S., Lasserre, J.J., Brizzi, L.E., and Borée, J. (2006) Instantaneous skin-friction pattern analysis using automated critical point detection on near-wall PIV data. *Measurement Science and Technology*, **17**, 1659–1669.

El Tahry, S.H. (1982) A numerical study on the effects of fluid motion at intake valve closure on the subsequent fluid motion in a motored engine. SAE Technical Paper 820035.

Enotiadis, A.C., Vafidis, C., and Whitelaw, J.H. (1990) Interpretation of cyclic flow variations in motored internal combustion engines. *Experiments in Fluids*, **10**, 77–86.

Fansler, T.D. and Drake, M.C. (2009) Flow, mixture preparation, and combustion in direct-injection two-stroke gasoline engines in *Flow and Combustion in Reciprocating Engines* (eds C. Arcoumanis and T. Kamimoto), Springer-Verlag, Berlin Heidelberg (Chapter 2).

Fansler, T.D. and French, D.T. (1987) Swirl, squish, and turbulence in stratified charge engines: laser velocimetry measurements and implications for combustion. SAE Technical Paper 870371.

Fansler, T.D. and French, D.T. (1988) Cycle resolved laser-velocimetry measurements in a reentrant-bowl-in-piston engine. SAE Technical Paper 880377.

Fitzgeorge, D. and Allison, G.L. (1962) Air Swirl in a Road Vehicle Diesel Engine. *Proceedings of the Institution of Mechanical Engineers: Automobile Division*, January **16** (1), 151–177.

Fogleman, M. A. (2005) Low dimensional models of internal combustion engine flows using the proper orthogonal decomposition. PhD Thesis. Cornell University.

Fogleman, M., Lumley, J.L., Rempfer, D. and Haworth, D. (2004) Application of the proper orthogonal decomposition to datasets of internal combustion engine flows. *Journal of Turbulence*, **5**, 1–18.

Gatski, T.B., Rumsey, C.L., and Manceau, R. (2007) Current trends in modelling research for turbulent aerodynamic flows. *Philosophical Transactions of the Royal Society A*, **365**, 2389–2418.

Gosman, A.D. (1986) Flow processes in cylinders in *Thermodynamics and Gas Dynamics of Internal Combustion Engines*, vol. 2 (eds J.H. Horlock and D. Winterbone), Oxford University Press, Oxford, pp. 616–772.

Gosman, A.D., Johns, R.J.R., and Watkins, A.P. (1980) Development of prediction methods for in-cylinder processes in reciprocating engines in *Combustion Modeling in Reciprocating Engines* (eds J.D. Mattavi and C.A. Amann), Plenum, New York, London.

Haller, G. (2002) Lagrangian coherent structures from approximate velocity data. *Physics of Fluids*, **14** (6), 1851–1862.

Han, D., Im, H., Han, S.-K., and Kim, H.-J. (2011) The turbocharged theta GDI engine of Hyundai. *Motortechnische Zeitschrift*, **72**, 30–35.

Hasse, C., Sohm, V., and Durst, B. (2010) Numerical investigation of cyclic variations in gasoline engines using a hybrid URANS/LES modeling approach. *Computers & Fluids*, **39**, 25–48.

Haworth, D. (1999) Large-eddy simulation of in-cylinder flows. *Oil & Gas Science and Technology—Revue de l'IFP*, **54** (2), 175–185.

Heywood, J.B. (1988) *Internal Combustion Engines Fundamentals*, McGraw Hill Company, New York.

Hill, P.G. and Zhang, D. (1994) The effects of swirl and tumble on combustion in spark ignition engines. *Progress in Energy and Combustion Science*, **20**, 373–429.

Holloway, A.G.L. and Tavoularis, S. (1992) The effects of curvature on sheared turbulence. *Journal of Fluid Mechanics*, **237**, 569–603.

Jakirlic, S., Hanjalic, K., and Tropea, C. (2002) Modeling rotating and swirling turbulent flows: a perpetual challenge. *AIAA Journal*, **40**, 1984–1996.

Johnston, S., Robinson, C., Rorke, W., and Smith, J. (1979) Application of laser diagnostics to an injected engine. SAE Technical Paper 790092.

Kapitsa, L., Imberdis, O., Bensler, H.P., *et al*. (2010) An experimental analysis of the turbulent structures generated by the intake port of a DISI-engine. *Experiments in Fluids*, **48**, 265–280.

Keromnes, A., Dujol, C., and Guibert, P. (2010) *Measurement Science and Technology*, **21**, 125–404.

Kume, T., Iwamoto, Y., Iida, K., Murakami, M., Akishino, K., and Ando, H. (1996) Combustion control technologies for direct injection SI engine. SAE Technical Paper 960600.

Le Roy, O. and L. Le Penven (1997) *Compression of a turbulent vortex flow*. 11th Symposium on Turbulent Shear flows, Grenoble, France, September 8–11.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Leblanc, S. and Le Penven, L. (1999) Stability of periodically compressed vortices at low Mach number. *Physics of Fluids*, **11** (5), 1–3.

Lumley, J.L. (1967) The structure of inhomogeneous turbulence in *Atmospheric Turbulence and Radio Wave Propagation* (eds A.M. Iaglom and V.I. Tatarski), Nauka, Moscow.

Lumley, J.L. (1990) Some comments on turbulence. *Physics of Fluids A*, **4** (2), 203–211.

Lumley, J.L. (1999) *Engines. An Introduction*, Cambridge University Press, Cambridge.

Lumley, J.L. (2001) Early work on fluid mechanics in the IC engine. *Annual Review of Fluid Mechanics*, **33**, 319–338.

Lundgren, T.S. and Mansour, N.N. (1996) Transition to turbulence in an elliptic vortex. *Journal of Fluid Mechanics*, **307**, 43–62.

Mansour, N.N. and Lundgren, T.S. (1990) Three-dimensional instability of rotating flows with oscillating axial strain. *Physics of Fluids A*, **2** (12), 2089–2091.

Marc, D., J. Borée, R. Bazile, and G. Charnay (1997) Tumbling vortex flow in a model square piston compression machine: PIV and LDV measurements. SAE Paper 972834.

Margary, R., Nino, E., and Vafidis, C. (1990) The effect of intake duct length on the in-cylinder air motion in a motored diesel engine. SAE Technical Paper 900057.

Mathieu, J. and Scott, J. (2000) *An Introduction to Turbulent Flow*, Cambridge University Press, Cambridge.

Maurel, S., Borée, J., and Lumley, J.L. (2001) Extended proper orthogonal decomposition: application to jet/vortex interaction. *Journal of Flow, Turbulence and Combustion*, **67**, 125–136.

Miles, P.C. (2009) In-cylinder turbulent flow structure in direct-injection, swirl-supported diesel engines in *Flow and Combustion in Reciprocating Engines* (eds C. Arcoumanis and T. Kamimoto), Springer-Verlag, Berlin Heidelberg (Chapter 4).

Miles, P.C., Choi, D., Megerle, M., *et al*. (2004) The influence of swirl ratio on turbulent flow structure in a motored HSDI diesel engine—a combined experimental and numerical study. SAE Technical Paper 2004-01-1678.

Miles, P., Megerle, M., Hammer, J., *et al*. (2002) Late-cycle turbulence generation in swirl-supported, direct-injection diesel engines. SAE Technical Paper 2002-01-0891.

Miles, P.C., RempelEwert, B.H., and Reitz, R.D. (2008) Experimental assessment of Reynolds-averaged dissipation modeling in engine flows. *SAE Transactions: Journal of Engines*, **116**, 1540–1549.

Miles, P.C., RempelEwert, B.H., and Reitz, R.D. (2009) Experimental assessment of a non-linear turbulent stress relation in a complex reciprocating engine flow. *Experiments in Fluids*, **47**, 451–461.

Monaghan, M.L. and Pettifer, H.F. (1981) Air motion and its effect on diesel performance and emissions. SAE Technical Paper 810255.

Moreau, J., Borée, J., Bazile, R., and Charnay, G. (2004) Destabilisation of a compressed vortex by a round jet. *Experiments in Fluids*, **37**, 856–871.

Müller, S.H.R., Arndt, S., and Dreizler, A. (2011) Analysis of the in-cylinder flow field/spray injection interaction within a DISI IC engine using high-speed PIV. SAE Technical Paper 2011-01-1288.

Müller, S., Böhm, B., Gleissner, M., *et al*. (2010) Flow field measurements in an optically accessible direct injection spray-guided internal combustion engine using high speed PIV. *Experiments in Fluids*, **48**, 281–290.

Ozdor, N., Dulger, M., and Sher, E. (1994) Cyclic variability in spark ignition engines a literature survey: Journal of Engines, vol. 103. SAE Technical Paper 940987.

Petersen, B.R. and Miles, P.C. (2011) PIV measurements in the swirl plane of a motored light-duty diesel engine. SAE Technical Paper 2011-01-1285.

Pope, S.B. (2000) *Turbulent Flows*, Cambridge University Press, Cambridge.

Reuss, D.L., Kuo, T.-W., Khalighi, B., Haworth, D., and Rosalik, M. (1995) Particle image velocimetry measurements in a high-swirl engine used for evaluation of computational fluid dynamics calculations. SAE Technical Paper 952381.

Ruiz, T., Borée, J., Tran, T., *et al*. (2010) Finite time lagrangian analysis of an unsteady separation using high speed particle image velocimetry. *Physics of Fluids*, **22**, 1–9, **075103**

Sagaut, P. and Cambon, C. (2008) *Homogeneous Turbulence Dynamics*, Cambridge University Press, Cambridge.

Schiestel, R. (1987) Multiple time scale modeling of turbulent flows in one point closures. *Physics of Fluids*, **30** (3), 722–731.

Shadden, S.C., Dabiri, J.O., and Marsden, J.E. (2006) Lagrangian analysis of fluid transport in empirical vortex ring flows. *Physics of Fluids*, **18** (4), 1–11.

Söderberg, F., Johansson, B., and Lindoff, B. (1998) Wavelet analysis of in-cylinder LDV measurements and correlation against heat-release. SAE Technical Paper 980483.

Sullivan, P., Ancimer, R., and Wallace, J. (1999) Turbulence averaging within spark ignition engines. *Experiments in Fluids*, **27**, 92–101.

Tennekes, H. and Lumley, J.L. (1972) *A First Course in Turbulence*, The MIT Press, Cambridge.

Tindal, M.J., Williams, T.J., and Aldoory, M. (1982) The effect of inlet port design on cylinder gas motion in direct-injection engines in *Flows in Internal Combustion Engines*, ASME Fluids Engineering Division, New York.

Tritton, D.J. (1977) *Physical Fluid Dynamics*, Van Nostrand Reinhold, Berkshire.

Tsui, Y.-Y. and Cheng, H.-P. (2007) Tumbling/squish interaction in loop-scavenged two-stroke engines. *Numerical Heat Transfer A*, **32**, 861–876.

Vermorel, O., Richard, S., Colin, O., *et al*. (2009) Towards the understanding of cyclic variability in a spark ignited engine using multi-cycle LES. *Combustion and Flame*, **156**, 1525–1541.

Voisine, M. (2010) Etude expérimentale de l'aérodynamique interne des moteurs. Mise en oeuvre de diagnostics d'analyses spatio-temporels pour un écoulement de rouleau compressé. PhD Thesis, ENSMA.

Voisine, M., Thomas, L., Borée, J., and Rey, P. (2011) Spatio-temporal structure and cycle to cycle variations of an in-cylinder tumbling flow. *Experiments in Fluids*, **50**, 1393–1407.

Wurms, D., Jung, M., Adam, S., Dangler, S., Heiduk, T., and Eiser, A. (2011) Innovative technologies in current and future TFSI engines from Audi. 20th Aachen Colloquium Automobile and Engine Technology, October 10–12, Aachen, Germany.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Xu, H. (2001) Some critical technical issues on the steady flow testing of cylinder heads. SAE Technical Paper 2001-01-1308.

Yamakawa, M., Youso, T., Fujikawa, T., Nishimoto, T., Wada, Y., Sato, K., and Yokohata H. (2011) Combustion technology development for a high compression ratio SI engine. SAE Technical Paper 2011-01-1871.

Zhao, F., Lai, M.-C., and Harrington, D.L. (1999) Automotive spark-ignited direct-injection gasoline engines. *Progress in Energy and Combustion Science*, **25**, 437–562.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto119
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Diesel and Diesel LTC Combustion

Öivind Andersson[1] **and Paul C. Miles**[2]

[1]*Lund University, Lund, Sweden*
[2]*Sandia National Laboratories, Livermore, CA, USA*

1	Introduction	1
2	Modern Trends in Engine Design	2
3	Practical Combustion System Design	3
4	Fundamentals	13
5	Overview of the Combustion Process	21
6	Multiple Injection Strategies	25
7	Low Temperature Diesel Combustion Processes	27
8	Summary	32
Acknowledgments		32
References		32

1 INTRODUCTION

Diesel engine combustion is an enormously complex process. To design and optimize diesel combustion systems, it is necessary to consider thermodynamic processes and their efficiency, fuel spray and atomization processes, multicomponent fuel vaporization processes under potentially supercritical conditions, complex turbulent flow and fuel–air mixing processes, oxidation kinetics of mixtures of large hydrocarbon fuels, pollutant formation and control, as well as heat transfer and radiation processes. The fuel introduction, mixing with ambient charge, ignition, combustion, and pollutant reduction processes must also be completed within a very short time period, while the piston is close to top-dead-center (TDC). In automotive engines, which operate at higher rotational speeds than heavy-duty truck engines, this time can be very short—on the order of 1×10^{-3} s. Beyond these considerations associated with fluid dynamics and the thermal sciences, the combustion system design process cannot be divorced from additional concerns related to materials, manufacturing, durability, and packaging. Finally, if diesel engines are to gain widespread market penetration worldwide, combustion system design must also aim to meet consumer expectations related to cost and refinement (noise, vibration, etc.).

The complexity of the diesel combustion process is compensated for by the higher thermal efficiency that can be achieved with diesel engines. Because diesel engine load control is achieved by regulating the amount of fuel injected, rather than by throttling the charge inducted, diesel engines have considerably lower pumping losses than conventional spark ignition engines. Moreover, with fuel injection near TDC, compression ratios (r_c) are not limited by fuel preignition. High engine r_c that improve thermodynamic efficiency can therefore be employed. The diesel engine combustion process also typically takes place under fuel-lean conditions, which keeps both global average temperatures and concentrations of CO_2 and H_2O in the product gases low—resulting in post-combustion gases characterized by a higher specific heat ratio than is found under stoichiometric combustion conditions. The higher specific heat ratio allows a greater amount of work to be extracted during expansion. Finally, for conventional diesel combustion, the introduction of the fuel into the cylinder near TDC prevents fuel from reaching engine crevices, and the combustion efficiencies are usually very high—typically >98%. These factors translate into a notable efficiency advantage when diesel engines are

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

compared with conventional (port-injected) spark ignition engines. A review of US Environmental Protection Agency (EPA) vehicle fuel efficiency ratings, for the same vehicle equipped with either a diesel or a gasoline engine, indicates that a diesel engine will typically provide a 30–40% improvement in fuel consumption estimated for a combined city/highway test cycle. Similar reductions in CO_2 emissions over a European drive cycle have been reported (Tanaka, 2011).

Because of their efficiency advantage, the potential reward for optimizing diesel engines and increasing their market penetration is large. The US Energy Information Agency (2013) estimates that by 2015, world energy usage by the transportation sector will be 1.1×10^{20} J, corresponding to the energy content of over 50 million barrels of oil per day. Of this, roughly 60% will be consumed by the light-duty sector. Every 1% improvement in the energy efficiency of the light-duty sector, achieved through either optimization of engines or increased diesel engine market penetration, thus corresponds to an energy savings equivalent to approximately 300,000 barrels of oil per day.

Diesel engines used in the light-duty sector through the 1980s were dominated by indirect injection (IDI) or "pre-chamber" designs. In these engines, the fuel is injected into a smaller chamber appended to the main combustion chamber, where it mixes with fresh charge and ignites. After ignition, the burning, fuel-rich mixture is expelled from the pre-chamber into the cylinder, where it rapidly mixes with the rest of the charge. IDI designs are characterized by lower noise and less demanding fuel injection equipment (FIE) requirements. However, throttling and heat transfer losses incurred as the charge is expelled into the main chamber, and higher heat transfer losses in general due to the large combustion chamber surface area, result in a significant loss in efficiency. Hence, virtually all modern engines designed for automotive applications employ direct injection (DI). In this chapter, we restrict our attention to DI combustion systems and refer the reader to Heywood (1988) for a more detailed discussion of IDI diesel combustion.

In the space available here, we cannot hope to summarize the full complexity of diesel combustion processes, and our objective is more modest. First, we will provide an overview of recent trends in diesel engine designs and a summary of the dominant engineering considerations that have factored into the choice of the design parameters. These considerations—which encompass the practical design knowledge that has been accumulated in over a century of building and operating these engines—involve balancing multiple trade-offs as the engineer selects such basic design parameters as bore-to-stroke ratio, r_c, and combustion chamber shape. Second, we provide a brief description of the science behind the diesel engine mixture formation, combustion, and pollutant formation and destruction processes. Lastly, we attempt to link engineering practice with the fundamental science, thereby providing diesel combustion system engineers with a sound foundation on which to base their design and optimization efforts.

2 MODERN TRENDS IN ENGINE DESIGN

The two most notable trends in diesel engines developed over the last decade have been a steady decrease in r_c and an increase in the maximum injection pressure. In a majority of these engines, the decrease in r_c has been driven by a desire to increase the power density, without requiring a large increase in the maximum cylinder pressure rating. Increased power density is enabled by higher injection pressure, which also aids in reducing soot emissions. The steady increase in power density can be clearly seen in the power density of engines released during the last decade, shown in Figure 1. However, there are a number of new engines, designated by the filled markers, which fall outside the general trend. These engines have maintained power densities (and peak torque, or maximum brake mean effective pressure (BMEP)) consistent with levels seen in older engines, although several feature r_c as low as 14.

The lower power density of these engines reflects a basic difference in design philosophy. For the higher power density engines, the engine designers have sought both improved performance and improved brake specific fuel

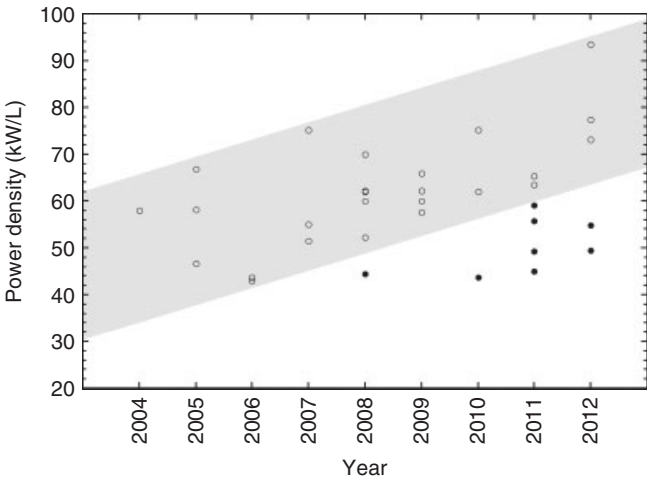


Figure 1. Sampling of power density ratings of light-duty diesel engines released in the last decade.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

consumption (BSFC) by reducing engine displacement. The engine can therefore spend more time operating at more efficient, higher load points, but performance is not compromised due to the high peak power capability. The high peak power density comes with increased cost, however, for improved FIE and boosting technologies. Moreover, if the engine is aggressively downsized, after-treatment costs can also increase. In contrast, lowering the r_c without increasing power density can enable improved BSFC through friction reduction and lower cost due to decreased component and after-treatment system costs. Note that the pursuit of lower engine costs does not necessarily imply poor performance. In developing the Mazda Skyactiv-D engine (Terazawa *et al*., 2011), Mazda's stated objectives in selecting a low r_c were decreased peak firing pressure and lower engine-out emissions through improved premixing. Nevertheless, this engine provides an impressive 59 kW/L and a peak BMEP of over 25 bar, while meeting Euro 6 NO_x emissions regulations without after-treatment. More will be said regarding the trade-offs incurred in engine design in what follows.

Trends in future engine design will also be impacted by technology advances and rising fuel costs. Owing to the introduction of stop–start systems, low load engine performance has become less important, while emissions released shortly after starting are of even greater consequence. The potential future introduction of mild hybridization can be expected to result in a further de-emphasizing of low load operation when considering design choices. Likewise, advances in after-treatment technology may drive different design choices by altering the trade-offs between various pollutants. Engine designs can be expected to change to accommodate these technology advances, which are addressed in greater detail in (Trends—Compression Ignition). However, the fundamental physics behind the selection of many design parameters, and behind the combustion and pollutant formation processes, can be expected to remain unchanged.

3 PRACTICAL COMBUSTION SYSTEM DESIGN

Some practical aspects of diesel combustion systems will be discussed here in order to provide a more concrete basis for subsequent topics. We will point out the typical features of diesel combustion systems and highlight the various factors that must be considered when making design choices regarding variables such as displacement, r_c, bore-to-stroke ratio, and combustion chamber geometry.

3.1 Topology and operating characteristics of light- and heavy-duty engines

Modern DI diesel combustion systems consist of a multi-hole fuel injector mounted in the cylinder head and a cavity in the piston where the main part of combustion takes place. The injector nozzle is normally placed on the cylinder centerline and the fuel jets emanating from it are directed radially into the cavity. A central injector location allows an axisymmetric fuel injection and combustion chamber geometry to be utilized, which facilitates full utilization of the trapped charge.

One reason for placing the cavity in the piston is that it allows a flat cylinder head surface. This increases the mechanical strength, which is favorable for withstanding the high peak cylinder pressures prevalent at high loads. Another reason is that the cavity plays an important role in generating a gas motion that supports the combustion process. The combustion chamber has a number of characteristic features that are described schematically in Figure 2. The area outside the cavity above the piston top is referred to as the *squish region*. Inside the cavity—often called the *bowl* there is a central protuberance or "pip" below the nozzle. If the top of the bowl has a smaller diameter than the maximum bowl diameter (i.e., if it has a "lip"), it is called a *re-entrant combustion chamber*. If there is no lip, it is called an *open combustion chamber*. The re-entrant chamber is the dominating one in light-duty engines, but

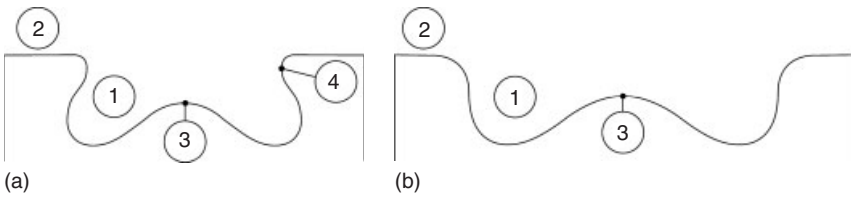


Figure 2. Schematic sections through the pistons in re-entrant (a) and open (b) combustion chambers: (1) bowl, (2) squish volume, (3) pip, and (4) lip. The open combustion chamber is drawn larger to indicate that it is predominantly used in heavy-duty engines, whereas the re-entrant chamber dominates in light-duty engines. (Reproduced from Andersson (2010). © Wiley-VCH Verlag GmbH & Co. KGaA.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

both types may be encountered in both light- and heavy-duty engines.

The re-entrant combustion system was invented in Switzerland by Hippolyt Saurer. His patent from 1934 describes a combustion system that is still surprisingly representative of most modern diesel engines, with an axisymmetric, re-entrant combustion chamber, a centrally mounted DI nozzle, and a gas motion typical of modern engines (Saurer, 1934). Saurer's combustion system is depicted in Figure 3.

Light-duty systems, which are optimized for the low-to-medium loads that characterize urban driving, typically employ relatively short fuel injections, which result in a major part of the heat release often taking place after the end of injection. The turbulence generated during injection is located in the direct vicinity of the spray and decays rapidly after the end of injection. For this reason, the spray cannot deliver all the kinetic energy needed to drive the mixing, and combustion system features promoting enhanced late-cycle mixing and oxidation play a more central role in light-duty systems. This is why such engines tend to be used in combination with a rotating gas motion—swirl—and with re-entrant combustion system geometries.

Toward the end of the compression stroke, the swirling air mass in the cylinder is pushed into the bowl. As the bowl has a smaller diameter than the cylinder, this results in an amplification of the rotational velocity due to the conservation of angular momentum, which aids the mixing process both during and after the end of injection. In the process of compressing the swirling charge into the bowl, large velocity gradients are also created, generating higher levels of turbulence. Both the amplification of the rotational velocity and turbulence generation are greater in re-entrant bowl geometries than in open geometries. Later, as the piston begins to descend, the lip of a re-entrant bowl geometry helps keep the swirling flow within the bowl. Turbulence generation from swirl will be discussed in greater detail in Section 4.5.

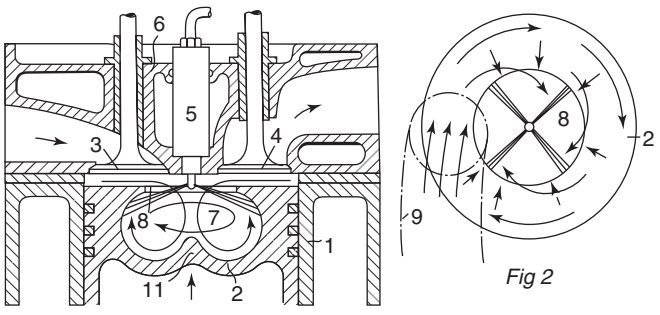


Figure 3. An image from Saurer's patent from 1934, describing the principles of the re-entrant combustion system. (Reproduced from Saurer (1934).)

In contrast to the lower load duty cycles characteristic of light-duty engines, heavy-duty diesel engines tend to be operated at high loads where the long injection durations produce a combustion process that is largely spray-driven. Developers of heavy-duty combustion systems thus tend to focus on spray formation and the quasi-steady phase of combustion, which puts different demands on the combustion systems. Heavy-duty combustion systems frequently employ open combustion chambers. This is partly because the absence of a lip makes the piston more robust to the high thermal loads that are typical for these engines. Another reason is that such systems are often quiescent, that is, there is no organized gas motion (swirl) in the cylinder, thus avoiding the need for a lip. Quiescent systems ideally do not have energy losses associated with setting up the swirl motion. This often gives them a higher volumetric efficiency and hence increased power density, but it also increases the demands on the fuel injection system, which may require higher fuel pressures and smaller nozzle holes (Beer, 1994). Another advantage of quiescent combustion systems is minimized heat transfer losses, as swirl enhances heat transfer to the combustion chamber walls by increasing the turbulence.

During the typical high load conditions in a heavy-duty combustion system, the quasi-steady jet phase is often assumed to play a dominant role in combustion. The conceptual image of the jet depicted in Figure 11 of the next section and the concurrent reasoning about the flame lift-off length are thereby useful models to guide the combustion system design. Although late-cycle oxidation plays a role also in these systems, the long injection process produces intense turbulence during an extended part of the cycle. The guiding principle in these combustion systems is that the sprays should do as much of the mixing work as possible.

In addition to the typical light-duty and heavy-duty combustion systems introduced above, so-called narrow-angle direct injection (NADI) combustion systems have also been introduced (Cursente *et al*., 2008b; Walter and Gatellier, 2002). Although these combustion systems have important advantages for low load, highly premixed operating conditions, they may have difficulty meeting full-load performance targets (Hardy *et al*., 2004). Because these systems have not yet been incorporated in production vehicle designs, we do not describe them in detail here, but refer the reader to the above references.

3.2 Global combustion system parameters

The highest level parameters impacting combustion system design—engine displacement, number of cylinders, and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

power or torque density—will be determined by the power, torque, and packaging requirements needed for the specific vehicle application. These top-level parameters will subsequently influence the choice of additional parameters such as engine geometric compression ratio r_c and bore-to-stroke ratio (B/S).

3.2.1 Displacement

An attractive strategy to improve fuel consumption is to "downsize" the engine, selecting a lower displacement. As mentioned above, this strategy allows the engine to spend more time operating at higher loads, where friction and heat transfer losses are a smaller fraction of the total fuel energy released. To meet peak torque or peak power requirements, however, the peak power and/or torque density must be increased. Although high power or torque density is always desirable, the benefit must be weighed against additional expense related to the block and head structure, boosting requirements, after-treatment systems, and FIE.

Some manufacturers have adopted a very different approach to improve efficiency. Kanda *et al*. (2004) argue that by increasing displacement, lower boost levels, lower injection pressures, and lower swirl ratios can be employed, such that friction and heat transfer losses are reduced and improved fuel economy can be obtained at the same emissions levels. With the larger displacement, the gearing can then be adjusted while maintaining similar vehicle performance levels, leading to a further reduction in fuel consumption. Mazda engineers (Sakono *et al*., 2011; Terazawa *et al*., 2011) follow similar logic but, moreover, argue that due to the nonlinear relationship between torque and NO_x emissions, this engine "upsizing" strategy is also an effective means of NO_x control.

3.2.2 Compression ratio r_c

Increasing the r_c increases the theoretical efficiency of the engine. For an ideal limited-pressure cycle, which encompasses both constant volume and constant pressure cycles, the engine efficiency increases with increasing compression (or expansion) ratio, as indicated in Figure 4. The advantages of a high geometric expansion ratio may not always be achieved, however. Limits on the peak cylinder pressure or control of NO_x emissions may require retardation of the combustion event (Inagaki *et al*., 2011; Yamada *et al*., 2011), resulting in a lower effective expansion ratio. Likewise, a high specific heat ratio γ will increase engine efficiency by increasing the pressure drop that occurs for a given expansion ratio. Because H_2O and CO_2 have higher heat capacities than diatomic molecules such as O_2 and N_2, and hence, lower γ, charge dilution with air is preferred over charge dilution with exhaust gas recirculation (EGR) if maximum efficiency is desired.

There are several additional factors that also prevent achievement of the full advantages of high r_c. Higher r_c results in both higher compression pressures and, for a given amount of heat release, a larger pressure rise due to combustion. In turn, higher cylinder pressures increase ring and journal bearing friction. The surface area-to-volume ratio near TDC (Figure 5) also increases as r_c increases, resulting in higher heat losses to the piston and head surfaces. Finally, air utilization becomes more difficult with higher r_c. It is common to evaluate the potential for full air utilization using the so-called k-factor, defined as the ratio of the volume inside the bowl to the total volume at TDC; see, for example, Fasolo *et al*. (2005). A high k-factor is considered to be beneficial for the air utilization. This is because air outside the bowl, for example in the squish

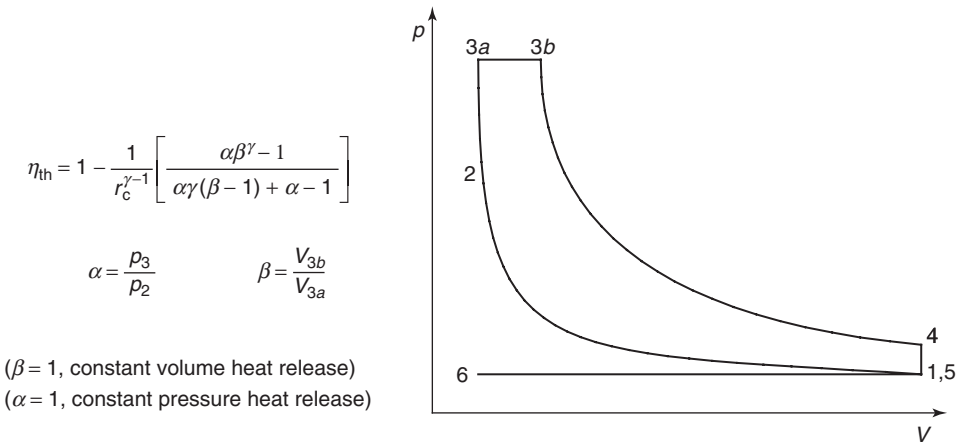


Figure 4. Thermal efficiency of a limited-pressure cycle. (Adapted with permission from J. Heywood, Internal Combustion Engine Fundamentals (Heywood, 1988) © The McGraw-Hill Companies, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

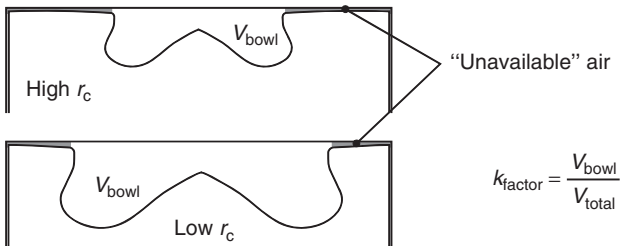


Figure 5. Illustration of the impact of compression ratio on surface area-to-volume ratio and on the *k*-factor.

volume, is assumed to be unavailable for the combustion process and for the oxidation of soot. As shown in Figure 5, for a fixed engine displacement, the volume at TDC is reduced as r_c increases, resulting in a reduced *k*-factor, more difficult air utilization, and consequently lower maximum torque leading to reduced power density and reduced full-load efficiency. Overall, reasonable variations in r_c are expected to have limited impact on efficiency (Heywood, 1988; Inagaki *et al*., 2011; Middlemiss, 1978).

As mentioned in Section 2, lowering r_c allows greater specific power at a given peak firing pressure. This is one of the main reasons for the current trend of decreasing r_c in diesel engines. Another reason to decrease r_c is to lower engine-out emissions by lowering compression temperatures, which lowers peak combustion temperatures (and hence, NO_x formation) and increases ignition delay, thereby enabling more fuel–air premixing which mitigates soot formation. With a typical intake temperature, lowering r_c from 17 : 1 to 14 : 1 will decrease the near-TDC compression temperature, and hence the peak combustion temperature, by approximately 50 K.

The choice of a lower r_c is supported by several additional considerations:

- Owing to the lower expansion, exhaust gas temperatures are generally higher. Higher exhaust gas temperature allows for more energy extraction by the turbocharger, and higher boost levels can be achieved at low speed, improving the maximum low speed torque characteristics (Catania *et al*., 2009; Fasolo *et al*., 2005).
- Diesel oxidation catalysts (DOCs) operate more efficiently with higher exhaust gas temperatures, and despite problems with low load engine-out HC and carbon monoxide (CO) emissions, post-DOC HC and CO levels can be reduced (Cipolla *et al*., 2007). Faster catalyst light-off might also be expected.
- A lower r_c results in slower cooling during expansion, providing more time for oxidation of soot and other products of partial combustion (Inagaki *et al*., 2011; Van den Huevel *et al*., 2006).
- A low r_c, with an attendant large bowl volume, allows the removal of piston top valve pockets with a smaller deterioration in *k*-factor.
- Lastly, at low load, low r_c engines may be able to meet soot emission targets with lower injection pressures, resulting in improved brake specific fuel consumption (Catania *et al*., 2009).

Lowering the r_c can have drawbacks, however:

- The low compression temperatures lead to long ignition delays and low combustion temperatures that can result in high engine-out HC and soot emissions, as mentioned above. These are especially problematic during the first few minutes of operation, before DOC light-off.
- Long ignition delays may also increase combustion noise. Combustion stability suffers as well, and large, robust pilot injection strategies must be adopted to recover acceptable noise and stability levels (Cursente, Pacaud, and Gatellier, 2008a; Terazawa *et al*., 2011). Larger pilots injected early into low density gases can over-penetrate, however, causing problems with oil dilution and cylinder wear.
- Despite the expected greater premixing, low speed, mid-load soot emissions can suffer (Catania *et al*., 2009).
- Low r_c engines exhibit lower BMEP at high speeds, when the maximum torque is limited by exhaust gas temperature.
- Finally, the lower r_c limit is set by demands on startability at low ambient temperatures. In light-duty engines, the limit steadily decreases as new glow plug technology becomes available.

In the end, selection of r_c is driven less by efficiency and fuel consumption concerns and more by a compromise among the desired power density, emission characteristics, and startability.

3.2.3 Bore-to-stroke ratio (B/S)

Bore-to-stroke ratio is another high level combustion system parameter that, like r_c, impacts the combustion system design. Diesel engines are almost always "under-square" or "long stroke". Of the engines represented in Figure 1, the average *B*/*S* is ∼0.91. One reason that *B*/*S* usually falls below unity is that it affects the height and length of the engine, which are usually limited by vehicle constraints. Even differences of a couple of centimeters may be critical for packaging an engine into a vehicle. There are several additional reasons, however, that impact efficiency and combustion system design directly:

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- Friction is the most frequently cited factor that is influenced by the bore-to-stroke ratio. Piston ring friction is thought to scale with S/B^2, which—for a fixed engine displacement—scales as $(B/S)^{-4/3}$ (Heywood, 1988). Likewise, viscous piston friction can be shown to scale with $(B/S)^{-2/3}$. Overall, it is reasonable to anticipate that piston assembly friction scales inversely with B/S, and short-stroke engines are characterized by lower friction.
- Piston speed is reduced with large B/S, thereby allowing greater engine rotational speeds and potential for higher peak power.
- With large B/S, a larger area is available for the inlet valves, allowing higher intake air mass flow rates and thus higher engine power or torque density.
- A larger bore also allows a wider bowl, which could be advantageous at high engine speeds (Fasolo *et al*., 2005) and reduces the risk of liquid fuel wetting the bowl wall. This is mainly a problem during cold conditions when vaporization is impeded (Chartier *et al*., 2009a). Large B/S may be well-suited to low r_c designs due to the larger fuel jet penetration that occurs when the ambient density is low.

There are also disadvantages to having a large B/S:

- With smaller bores, the bowl aspect ratio (depth-to-diameter ratio) more closely approaches 1, as shown in Figure 6. The surface area-to-volume ratio is thus reduced near TDC—even if near bottom-dead-center (BDC), the surface area-to-volume ratio is larger. A smaller B/S is thus expected to result in reduced near-TDC heat losses, although total heat transfer losses may actually increase (Flowers, Martinez-Frias, and Cleeves, 2010). The expected reduction in near-TDC heat loss due to reduced surface area-to-volume ratio will be lessened, however, by somewhat higher convective heat transfer coefficients at smaller B/S.
- Figure 6 also shows that with a smaller B/S, the k-factor is increased, suggesting more efficient air utilization.
- Empirically, turbulent velocity fluctuations are known to scale with mean piston speed, which at fixed displacement scales as $(B/S)^{-2/3}$. Thus, with larger B/S, turbulent fluctuations are reduced, and turbulent mixing times are expected to increase—leading to slower combustion rates. This factor can be quite significant in SI engines (Filipi and Assanis, 2000; Siewert, 1978). Although the impact of B/S on the duration of diesel mixing-controlled combustion has not been explored explicitly, engine speed has been shown to exert a major influence on combustion rates in diesel engines (Timoney, 1987), pointing to the importance of turbulence generated by piston motion.

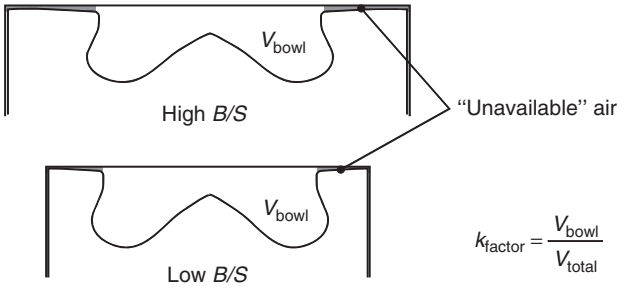


Figure 6. Illustration of the impact of bore-to-stroke ratio (at fixed r_c) on surface area-to-volume ratio and on the k-factor.

A very broad summary assessment of the impact of B/S on combustion system design is that a small B/S is favored for higher engine efficiency, whereas larger B/S favors power density (National Research Council, 2011).

3.3 Combustion chamber geometry

The combustion chamber geometry plays a crucial role in generating the gas motion that supports the combustion process. A successful design provides conditions for a favorable interaction among the sprays, the combustion chamber surfaces, and the swirling gas flow. Geometry also plays an important role in setting up a flow pattern that aids the late-cycle oxidation. Typically, soot emissions are found to be more strongly impacted by combustion chamber geometry than NO_x emissions or fuel consumption, although improvements in all three quantities can be achieved with well-optimized geometries (Andersson *et al*., 2009). In this context, it should be noted that the complex influence of geometry on transport and mixing phenomena in the combustion chamber makes the practical relevance of the k-factor somewhat doubtful. While indicating how large a portion of the air that is available resides within the bowl, two well-optimized bowls with exactly the same k-factor may produce very different soot emissions at exactly the same operating condition (Andersson *et al*., 2009).

Owing to the complexity of the interactions among chamber geometry, sprays, and flows, it is difficult to state general design rules that apply across a range of swirl-based combustion systems. This is reflected in the wide variety of re-entrant designs present on the market. Nevertheless, we attempt to summarize the important considerations in this section.

3.3.1 Axisymmetry

Axisymmetry of the combustion chamber is strongly desirable to promote air utilization, and is a major consideration

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

(along with power density) driving the adoption of four-valve designs (Horrocks, 2010). The presence of valve pockets in the piston top adversely impacts combustion chamber symmetry, although the *k*-factor is improved. Increased symmetry can benefit part-load emissions as well as full-load performance (Fasolo *et al*., 2005), and consequently many new engine designs feature vertical valves and no-valve pockets in the piston top (Abe *et al*., 2004; Bauder *et al*., 2005; Dworschak *et al*., 2009; Langen *et al*., 2010; Steinparzer *et al*., 2007; Van den Huevel *et al*., 2006). As noted above, in four-valve designs, the injector is typically mounted along the cylinder centerline to promote axisymmetry, and the piston bowl is likewise centered in the cylinder. While non-axisymmetric bowls (e.g., square) may have advantages and have been released in production engines (Kihara, Mikami, and Kinbara, 1983), such designs have not been widely adopted.

3.3.2 Piston bowl diameter

The maximum diameter of typical piston bowls is typically greater than ~60% of the bore diameter, and roughly three to four times the maximum bowl depth. Computational studies indicate that bowl diameter is the dominant geometry variable impacting combustion system performance (Genzale, Reitz, and Wickman, 2007). Like many geometry variables, there is no single value that proves optimal at all engine operating conditions. Full-load operation is usually found to benefit from a wide bowl design, whereas narrower bowls are preferred at lower load (Cursente, Pacaud, and Gatellier, 2008a; Ge *et al*., 2010; Lisbona, Olmo, and Rindone, 2000).

A trend that seems to be general is that re-entrant combustion chambers tend to become wider and shallower with each generation, (e.g., Matsui *et al*., 2008; Bauder *et al*., 2005; Crabb *et al*., 2013). This development is dictated by both mechanical demands and combustion considerations. As the specific power and peak combustion pressure increases, the maximum pressure difference between the combustion chamber and the volume below the top piston ring increases, resulting in larger deformation of the piston top during combustion. Considering that there is an oil gallery between the combustion chamber and the piston ring pack, this is one of the weaker regions of the piston. Using a shallower piston bowl makes the top of the piston more robust to deformation, and to maintain r_c, the bowl must be made wider.

There are multiple ways in which a wider bowl diameter impacts the combustion process, most of them positive:

- Wider bowls provide a longer free spray length to reduce the risk of wall wetting (Hadler *et al*., 2007) and over-penetration of fuel vapor. Higher power requires higher fuel flow rates through the nozzle. This translates into larger orifices and longer liquid penetration lengths, especially at cold ambient conditions.
- Increased injection pressures are also employed to support higher power densities, resulting in more rapid penetration of the sprays. Wider bowls are thus thought to more closely match combustion chamber shape to higher injection pressures (Chi *et al*., 2008; Lee *et al*., 2009; Lee *et al*., 2012).
- Wider bowls complement low r_c designs, due to the increased spray penetration associated with lower ambient density (Cipolla *et al*., 2007).
- Wider bowls improve the *k*-factor (improve air utilization) and reduce the heat load on the piston due to a more favorable surface-to-volume ratio (Crabb *et al*., 2013).
- Large diameter bowls are more tolerant to advanced injection timing and help prevent oil dilution (Cipolla *et al*., 2007).
- Wider bowls tend to increase the tolerance of the combustion system to variations in spray targeting (Fasolo *et al*., 2005).
- The soot-NO_x trade-off may suffer at part load with wide bowls (Fasolo *et al*., 2005).
- Low load HC and CO emissions may increase in low r_c engines when wide bowls are employed (Cursente, Pacaud, and Gatellier, 2008a).

3.3.3 Piston bowl re-entrancy

Most light-duty combustion systems feature re-entrant bowls. As noted above, re-entrancy promotes the amplification of the swirl velocity as the charge is compressed into the bowl. It also impacts the strength of the flow from the squish volume into the bowl, as described by Zolver, Griard, and Henriot (1997), and is thus expected to increase turbulence levels and mixing rates within the bowl. Moreover, a re-entrant bowl shape preserves the kinetic energy of the fuel sprays after they impact the bowl wall, redirecting them toward the cylinder center and preventing stagnation of rich mixture near the bottom of the bowl (Terazawa *et al*., 2011). Re-entrancy further retains the swirling flow within the bowl, impeding the spread of burning fluid into the squish volume (Zhang *et al*., 1995).

Experimental investigations confirm that re-entrant bowl geometries increase mixing rates, allowing greater timing retardation before fuel consumption or soot emissions suffer excessively (Ikegami *et al*., 1990a; Kidoguchi, Yang, and Miwa, 1999; Middlemiss, 1978; Saito *et al*., 1986). Although the increased mixing rates may result in higher NO_x, the greater retardability compensates. A greater EGR

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

tolerance is also to be expected from re-entrant combustion systems. Re-entrancy is moreover reported to be advantageous even in large-bore, quiescent combustion systems when highly premixed operation is desired (Cao *et al*., 2009).

Despite these advantages, there appears to be a trend toward lower levels of re-entrancy as existing engines are modified and updated (Hadler *et al*., 2007; Lee *et al*., 2012). Part of this trend is related to combustion performance. Lower values of re-entrancy are reported to promote robustness to spray targeting variations associated with dispersion in spray angle and injector protrusion (Diwakar and Singh, 2009; Fasolo *et al*., 2005). Likewise, re-entrant bowls have higher rim temperatures and can shorten ignition delay (Saito *et al*., 1986)—impeding the premixing of fuel. This disadvantage is expected to be less relevant in contemporary engines where ignition delays are deliberately shortened by means of pilot injections (in order to reduce combustion noise) but may become important in low temperature combustion (LTC) modes, which are based on premixing during an extended ignition delay. Reduced re-entrancy also responds to the higher mechanical and thermal loads in today's downsized diesel engines. Either a decreased degree of re-entrancy or remelting of parts of the piston top during manufacturing (see, e.g., Eidenböck *et al*., 2012) is needed to ensure piston robustness and durability.

3.3.4 Piston bowl lip shape

The various features defining the piston bowl lip shape are shown in Figure 7. It is generally believed that a small radius at the upper lip of the piston is beneficial for soot emissions as well as increasing the combustion system EGR tolerance (e.g., Steinparzer *et al*., 2007)—benefits that may be related to turbulence generation by the reverse squish flow during expansion. The smallest radius that can be practically incorporated, however, is often limited by piston durability issues.

The lower lip shape has been found to impact both smoke and fuel consumption (Middlemiss, 1978). Computational studies indicate that both the angle with which the spray impacts the bowl wall and the bowl radius of curvature near the lip affect the distribution of fuel within the chamber and the strength of the flow structures within the bowl (Diwakar and Singh, 2009). Experiments have likewise shown that the lower lip shape affects the jet spreading rate deep in the bowl (Montajir *et al*., 2001); smooth transitions are preferred to create a large spray volume indicative of increased entrainment. However, the apparent enhancement in spray volume does not translate into improved moderate-load smoke emissions. Lastly, a large outer bowl (torus) radius, resulting in a small lip height, has been reported to improve the soot/NO_x trade-off characteristics at both part and full load (Zhu *et al*., 2004).

3.3.5 "Stepped-lip" bowl geometries

In contrast to the typical bowl lip geometry discussed above, engines recently designed for both light- and heavy-duty applications often have "stepped" or chamfered bowl lips, as shown in Figure 8. These steps are also sometimes referred to as *soot-in-oil rims* (Dreisbach *et al*., 2007).

One objective of stepped-lip bowls is to split the fuel spray, directing a portion of it upward toward the head and the remainder downward into the bowl. By redirecting the radial momentum of the upper portion toward the head, the penetration of the jet into the squish volume is impeded. An obvious result is that less soot is generated near the cylinder walls, where it could find its way into the engine oil. A second benefit is that heat loss to the cylinder liner is reduced, and a third benefit is expected for cold-starting—the glow plug can be located near the head of the spray plumes, where it is less susceptible to variability in the rotational alignment of the fuel sprays (Styron *et al*., 2011). In addition to reduced heat loss to the liner, stepped-lip bowls also favor reduced heat losses to the piston surfaces, due to the improved surface area-to-volume ratio.

Another objective of stepped-lip designs is to enhance air utilization—thereby reducing particulate matter (PM) emissions (Yoo *et al*., 2013). The use of multiple injection strategies is expected to help in maximizing the air utilization. By targeting the upper portion of the bowl with

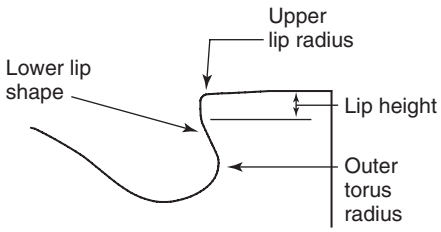


Figure 7. Features defining the bowl lip geometry.

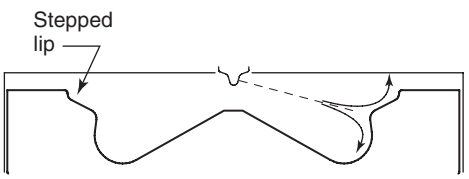


Figure 8. "Typical" stepped-lip bowl geometry. (Reproduced with permission from Ford Motor Company.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

an initial injection and the lower portion with a second injection, mixing of the second injection with O_2-depleted charge is avoided and soot and CO emissions are reduced (Dolak, Shi, and Reitz, 2010). The success of these multiple injection strategies will clearly require the impact of injection pressure, engine speed, and injection timing on spray targeting to be carefully considered. Enhanced air utilization is also reputed to improve the EGR tolerance of the combustion system. Reports of increased robustness of stepped-lip designs to parameter variations, including swirl and spray angle (Styron *et al*., 2011), are likely related to improved air utilization.

3.3.6 Top clearance or "squish" height

The top clearance between the uppermost surface of the piston and the head surface is typically set to roughly 0.6–0.8 mm. From the previous discussion, it is clear that this height should be kept as small as possible in order to promote full air utilization (a high *k*-factor). Indeed, both soot emissions (Ikegami *et al*., 1990b) and low load HC/CO emissions (Aronsson *et al*., 2009) have been found to benefit from a small squish height. However, too small a squish height can place severe demands on manufacturing tolerances, even when multiple head gasket thicknesses are available to compensate. Fuel consumption has also been shown to be minimized with a squish height between 0.6 and 0.8 mm, possibly due to increased heat losses when too small a squish height is employed (Ikegami *et al*., 1990b). It should be noted that, for a given r_c, changes in the squish height must be compensated by changes to the bowl volume. It is thereby difficult to separate the effects of the squish height from those of the combustion chamber geometry.

3.3.7 Bowl pip geometry

Like the lower bowl lip and outer bowl shape, one purpose of the central pip in the piston bowl is to help redirect the fuel jets and avoid stagnation of fuel-rich mixtures in the central region of the bowl. The bowl center is a region where flow velocities and mixing rates are low, making optimal mixture formation difficult (Bauder and Stock, 1990; Middlemiss, 1978)—accordingly, it is thought to be advantageous to fill this area with metal, thereby allowing for increased charge mass along the bowl periphery. This simple consideration is likely at least partly responsible for the observations of increased mixing rates in bowl designs with central pips (Middlemiss, 1978), resulting in less deterioration of fuel consumption and smoke as injection is retarded. However, a central pip is also thought to increase turbulence levels within the bowl (Béard, Mokaddem, and Baritaud, 1998; Kidoguchi, Sanda, and Miwa, 2003), which will also enhance mixing rates.

There are a variety of common bowl pip shapes that have been released in production engines, some of which are illustrated in Figure 9. Computational studies focusing on pip and lower bowl shape find a clear, load-dependent impact of pip shape on emissions and performance, but no profile gives universally better behavior at all loads or for all performance metrics (Juttu *et al*., 2009). The simple W-shaped or conical pip design has been featured in numerous engines and has a high expected reliability (Fasolo *et al*., 2005). So-called step-cone pips displace a greater amount of charge from the bowl center, thereby favoring larger bowl diameters. Bowls with this pip geometry promote better full-load air utilization and tolerate more advanced part-load injection timings without excessive oil dilution (Cipolla *et al*., 2007). However, too high a pip may interfere with air entrainment into the spray (Wickman, Senecal, and Reitz, 2001).

The more complex shapes on the right-hand side of Figure 9 were designed with the stated objective of influencing the bulk flow structures within the bowl late in the combustion process. In the Volvo design (Andersson *et al*., 2009), the designers sought to create a system of two counter-rotating toroidal vortices that serve to transport fuel and oxidant to a common interface (stagnation plane),

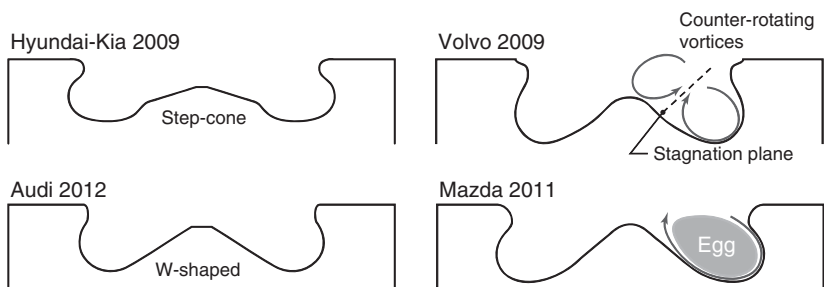


Figure 9. A sample of the variety of central pip shapes found in production engines.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where local turbulence generation rates are high. Enhancing the late-cycle mixing via this dual-vortex system led to reduced soot emissions and reduced combustion duration, thus leading to improved fuel economy. This concept has also been employed in off-highway engines (Crosse, 2010), enabling them to meet off-highway particulate emissions regulations without after-treatment.

In the Mazda design, the focus was on reduction of NO_x (Shimo, Kataoka, and Fujimoto, 2004; Terazawa *et al*., 2011). In this concept, the outer "egg-shaped" vortex transports hot combustion products to the cylinder center where they mix rapidly with cooler surrounding charge, thereby quenching thermal NO_x production. The thermal efficiency is also increased due to a shortening of the combustion duration (Shimo, Kataoka, and Fujimoto, 2004). Toyota has also recognized the importance of the outer vortex on promoting soot oxidation (Hotta *et al*., 2002a).

3.4 Matching piston geometry to flow and fuel injection parameters

The features of the combustion chamber geometry described above are not independent of the flow in the cylinder or of the geometry of the FIE and the maximum available injection pressure. Like the features of the combustion chamber geometry discussed above, it is difficult to state general design rules for matching the geometry to flow and injector characteristics. Consequently, we again endeavor only to summarize the most important considerations.

3.4.1 Swirl level

One purpose of flow swirl, as indicated by the preceding discussion, is to support the development of large-scale flow structures that disperse the burning fuel throughout the bowl. Another is to generate small-scale turbulence that completes molecular level fuel–air mixing on sufficiently short time-scales. As noted above, swirl is particularly important in light-duty engines, where a major part of the heat release often takes place after the turbulence directly generated by the fuel injection event has already decayed. The increased mixing rates associated with flow swirl can result in both lower soot emissions and improved fuel consumption due to shortened combustion duration; see for example, Steinparzer *et al*. (2007).

Employing flow swirl also has disadvantages. First, ports designed to generate swirl typically have greater flow losses and the ensuing reduction in volumetric efficiency results in a loss in power density. Recent engine designs are making greater use of chamfers in the cylinder head, however, which help increase the swirl generated at low valve lifts and reduce interference between the flows from the two intake ports (Bauder, Fröhlich, and Rossi, 2010; Chi *et al*., 2008; Eidenböck *et al*., 2012; Lee *et al*., 2009; Van den Huevel *et al*., 2006). Consequently, flow losses are becoming less severe. Second, flow swirl will increase heat losses, which can adversely affect fuel economy despite the more rapid combustion. The increased heat loss also impacts cold-start behavior. Third, excessive swirl can reduce the penetration of the fuel jets and impede air utilization. Near-nozzle entrainment is not expected to be impacted by swirl due to the low local tangential velocities and high injection velocities (Andersson, 2010). However, excessive swirl has been shown to impede penetration at part load (Sahoo, Petersen, and Miles, 2012), and has been shown to significantly deflect the fuel sprays even at high load (Hotta *et al*., 2002b). Despite these disadvantages, the benefits of swirl are such that, to the authors' knowledge, no modern automotive-scale DI engines have been introduced that do not employ some level of flow swirl.

Swirl can be quantified by a variety of metrics (e.g., Stone and Ladommatos, 1992). The most common is the swirl ratio $R_s = \omega_s/2\pi N$. This is the ratio of the angular velocity ω_s of the solid-body rotating flow with the same average angular momentum to the angular velocity of the crankshaft. N represents the rotational speed of the engine. Light-duty engines are typically designed to operate with either a fixed swirl ratio of between roughly 2 and 2.5 or are fitted with a swirl control valve that restricts one port and allows the flow swirl to vary in the range from approximately 1 to 4. The lower swirl is achieved when both ports are fully open and is appropriate for high engine speeds and loads. At lower speeds and loads, when flow restrictions are less important, the higher swirl levels are used to promote soot oxidation (Chi *et al*., 2008; Cursente, Pacaud, and Gatellier, 2008a). However, high swirl at low loads and high EGR rates can also increase HC and CO emissions (Cursente, Pacaud, and Gatellier, 2008a).

The optimal swirl level is generally found to depend on the number of nozzle holes and is usually lower for a greater number of holes (Akihama *et al*., 2008; Fasolo *et al*., 2005; Kurtz and Styron, 2012; Van den Huevel *et al*., 2006). With a sufficient number of holes, the full-load soot is not overly sensitive to swirl, and it may be possible to use a single swirl level optimized under part-load conditions without a large adverse impact on full-load behavior (Van den Huevel *et al*., 2006). A second parameter that impacts the optimum swirl level is bowl diameter. As the swirling flow is compressed into the piston bowl, its rotational speed is amplified, and combustion systems have been thought to optimize at "bowl" swirl ratios near 5. Larger bowls can be found to require higher swirl levels, because the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

swirl amplification is less (e.g., Hadler *et al*., 2007). This traditional thinking may not be applicable to low r_c designs, in which part-load soot emissions are mitigated primarily through increased fuel–air premixing rather than increased late-cycle oxidation. In this case, low swirl and large bowl diameters can be jointly employed to reduce heat losses and improve both cold-start behavior and fuel consumption (Inagaki *et al*., 2011).

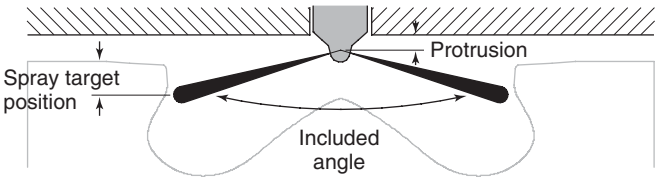


Figure 10. Definitions of variables impacting the interaction of the sprays with the piston.

3.4.2 Fuel injection parameters

Matching the desired power density, the piston bowl geometry, and the swirl level to the characteristics of the fuel injection system is a critical aspect of diesel combustion system design. At the most basic level, the fuel injector flow capability and the maximum injection pressure must be selected to provide sufficient fuel to meet the rated power requirements. The full-load fuel delivery must take place within a 30–35° window, which at rated speed is only about 1 ms in duration. The lowest flow nozzle that meets the full-load requirements is usually selected, as lower flow nozzles generally result in an improved soot/NO_x trade-off (e.g., Fasolo *et al*., 2005). The fuel injector flow rating is determined by the number of nozzle holes, their diameter, and the nozzle hole discharge coefficient. Thus, a small nozzle hole size requires a larger number of holes in order to meet rated power targets. Modern, conical nozzles with rounded nozzle hole inlets exhibit discharge coefficients of ~0.85. The injector needle opening and closing characteristics are also important. Hydraulically actuated injector designs with control chambers close to the nozzle tip and designs that are directly driven by piezoelectric actuators can significantly increase the amount of fuel delivered for the same overall injection duration.

Once the nozzle flow rating is determined, the nozzle hole diameter and number of holes must be fixed, as well as the parameters affecting how the spray interacts with the piston. As noted above, the optimal number of holes is coupled to the swirl ratio. Although a greater number of holes generally optimizes at a lower swirl level, the optimum achieved may be higher for too many or too few holes, or have less robustness to swirl variation (Fasolo *et al*., 2005; Kurtz and Styron, 2012; Van den Huevel *et al*., 2006). The optimal number of holes also depends on the specific engine operating condition and r_c (Cursente, Pacaud, and Gatellier, 2008a; Inagaki *et al*., 2011) as well as the metric used to define the optimum. In general, the higher swirl needed when too few holes are used will adversely affect part-load fuel consumption, whereas with too many nozzle holes, the soot and fuel consumption at higher loads deteriorates. In recent years, there has been a general trend toward an increased number of nozzles holes; 7–8 is now typical for new engine designs.

The spray interaction with the piston is affected by the spray target position, which is set by the injector nozzle tip protrusion and the included spray angle or umbrella angle (Figure 10). Like the optimal nozzle hole size and number, the optimal spray targeting will depend on numerous other combustion system parameters (bowl geometry, swirl ratio, etc.) as well as the engine operating condition. However, a few generalizations can be made.

At higher loads, the choice of spray targeting is mainly concerned with minimizing soot emissions. As described earlier, the shape of the bowl lip in conjunction with targeting location must enforce an appropriate fuel split between the squish volume and the bowl and create sufficiently strong flow structures to avoid stagnation of fuel-rich mixture near the bowl walls or floor. As the load is increased, the optimal targeting that satisfies these conditions moves lower in the bowl (Andersson *et al*., 2009). Targeting lower in the bowl also reduces light-load HC and CO emissions (Andersson *et al*., 2009; Aronsson *et al*., 2009; Cipolla *et al*., 2007), indicating that over-lean mixture formed from fuel injected directly into the squish volume is an important source of these emissions. Spray targeting has also been found to impact secondary liquid atomization by the piston and to have potential for enhancing cold-start behavior (Lippert *et al*., 2000).

Many investigations of spray targeting vary primarily the included angle of the spray, which is cited in modeling studies of heavy-duty engines as the dominant nozzle geometry-related factor impacting NO_x, soot, and fuel consumption (Genzale, Reitz, and Wickman, 2007). In light-duty engines, however, the included angle has been found to be of little importance at mid-to-high loads, provided the same spray targeting position is maintained (Andersson *et al*., 2009). Generally, with typical re-entrant bowls and widely varying spray targeting position, wide included angles are found to give improved part-load soot emissions (Cipolla *et al*., 2007; Siewert, 2007). The impact of nozzle tip protrusion is not typically addressed directly. Here, we note only that increased protrusion will tend to increase the nozzle tip temperature, which is known to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

increase deposit formations within the nozzle (Tang *et al*., 2009).

4 FUNDAMENTALS

This section treats more fundamental aspects of the spray and mixture formation processes, the ignition and heat release processes, and flow processes that drive the design choices discussed above. We will focus on the processes occurring while the intake and exhaust ports are closed. Since the swirling gas motion in the cylinder is of central importance to the combustion process, the intake ports are sometimes considered to be an important part of the combustion system. Port flows and in-cylinder flows are described separately in (Gas Exchange - Breathing and Air Management and In-Cylinder Flow), respectively, and will not be treated here. Peripheral systems for supercharging and cooling the intake air and EGR also have important effects on combustion. These external systems are likewise not treated here, but are covered in (Gas Exchange - Breathing and Air Management, Intake Boosting and Engine Thermal Management).

4.1 The anatomy of a burning diesel jet

It is useful to begin with a description of the history of a fuel element that travels through a burning diesel jet. Figure 11 reproduces Dec's (1997) well-accepted picture of quasi-steady diesel combustion, applicable at moderate-to-high loads in large-bore diesel engines where combustion chamber wall and in-cylinder flow interactions are relatively minor. The picture shows the development of a single diesel fuel jet, which is injected into the cylinder at high pressure. The injection event occurs near TDC, when the ambient, undiluted charge density is $\sim$20 kg/m^3 and the temperature $\sim$1000 K.

On entering the cylinder, the fuel jet atomizes and rapidly entrains the hot ambient charge. Liquid fuel extends downstream to a location where the energy of the entrained charge is sufficient to vaporize the fuel, giving rise to a characteristic "liquid length" (Siebers, 2008). Beyond the liquid length, the vaporized fuel–air mixture continues to entrain hot gases, and eventually undergoes a putative premixed combustion process, generating a large amount of soot precursors and surface growth species—poly-aromatic hydrocarbons (PAHs) and C_2H_2—in the hot products of the fuel-rich combustion. These soot precursors go on to form soot particles. Meanwhile, a diffusion flame has surrounded the hot products and extends upstream to a location denoted the flame "lift-off" length.

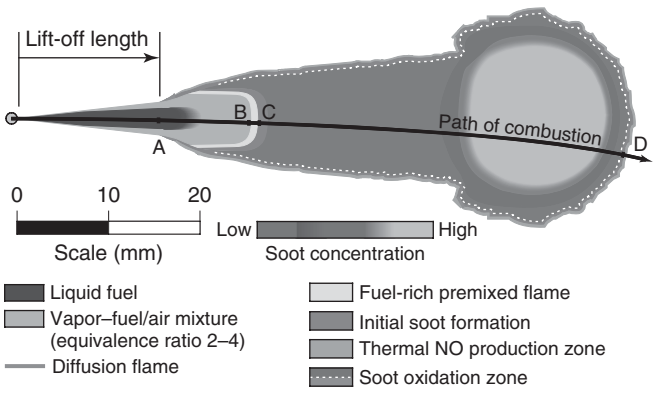


Figure 11. Dec's (1997) quasi-steady model of heavy-duty diesel combustion. The fuel element trajectory shown is correlated to the "path" of combustion in a $\phi-T$ parameter space in Figure 19 below. (From Dec (1997). Copyright © 1997 SAE International. Reprinted with permission.)

The lift-off length is extremely important, as it determines the spatial extent over which the fuel jet is able to entrain fresh charge. The magnitude of the lift-off length is thereby a limiting factor for the equivalence ratio of the mixture that undergoes premixed combustion, and hence, the propensity for soot formation (Siebers, 2008). Beyond the lift-off length, O_2 entrained into the jet is consumed within the diffusion flame, and the hot gases entrained and mixed with the fuel vapor and rich combustion products consist of the products of combustion formed in the diffusion flame. Accordingly, the temperature of the mixture increases, and the mixture fraction decreases, but concentrations of oxygen and key species such as OH are low. Soot formation processes proceed, and although concurrent oxidation of the soot precursors and soot formed also takes place, the oxidation rates are slow as compared to those that occur within the diffusion flame. Pickett *et al*. (2006) provide more information on the soot formation and oxidation processes in the downstream regions of the jet. Owing to the low oxygen concentrations in the downstream regions, NO_x formation likewise takes place predominantly at the diffusion flame, where high temperatures and oxygen concentrations coexist.

4.2 Air entrainment and mixing in diesel fuel jets

Given the central role of the rich premixed reaction zone for the formation of soot, there are in principle two ways to decrease the soot formation in quasi-steady jets. The first, as suggested above, is to increase the lift-off length to provide a longer distance for air entrainment. The second

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is to enhance the air entrainment upstream of the lift-off position.

The following description of the air entrainment follows the analysis of Naber and Siebers (1996), who used an idealized model jet. The jet is assumed to be isothermal and incompressible, to have no velocity slip between the injected fuel and the ambient air, and to have a constant spreading angle, α. It is also assumed to have radially uniform velocity and fuel concentration profiles. Although this is a highly idealized model of a real jet, the analysis focuses on the essential underlying physics of the spray: the conservation of mass and momentum. As a consequence, the fundamental factors limiting the air entrainment process are exposed.

The idealized model jet is depicted in Figure 12. Fuel enters through the orifice on the left and is assumed not to vaporize. As the fuel droplets reach the rightmost boundary of the control volume (dashed line), a certain amount of air has been entrained (white arrows). The mass flow of fuel at the orifice is given by

$$\dot{m}_f = \rho_f A(0) U(0) \tag{1}$$

where ρ_f is the fuel density, $A(0)$ is the effective orifice area, and $U(0)$ is the injection velocity. The air mass flow rate at x, the rightmost boundary of the control volume, is given by

$$\dot{m}_a = \rho_a A(x) U(x) \tag{2}$$

where ρ_a is the ambient air density, $A(x)$ is the jet cross-sectional area, and $U(x)$ is the jet velocity at x. The downstream location, x, is assumed to be large enough that the area occupied by the fuel droplets is negligible compared to $A(x)$.

The fuel mass flow over the control volume boundary at x is the same as the mass flow at the orifice, but the entrained air slows the jet down. The jet velocity at x is given by the conservation of momentum,

$$\rho_f A(0) U(0)^2 = [\rho_f A(0) U(0) + \rho_a A(x) U(x)] U(x) \tag{3}$$

where the left-hand side is the momentum flux at the orifice and the right-hand side is the momentum flux at x. This can be solved for $U(x)$ and inserted into Equation 2. An expression for the fuel-to-air mass flow ratio at x is obtained by dividing Equation 1 by Equation 2. With some algebraic manipulation, this analysis gives a powerful scaling relationship for the mean equivalence ratio in the jet cross section,

$$\overline{\phi} = \frac{2 f_s}{\sqrt{1 + 16\widetilde{x}^2} - 1} \tag{4}$$

where f_s is the stoichiometric air-to-fuel mass ratio for the fuel and ambient air composition. The coordinate $\widetilde{x}$ is a nondimensional axial distance from the orifice,

$$\widetilde{x} = \frac{x}{x^+} \tag{5}$$

where x is the physical distance from the orifice and x^+ is a characteristic length scale,

$$x^+ = \frac{d_f \sqrt{\frac{\rho_f}{\rho_a}}}{\tan(\alpha/2)} \tag{6}$$

In Equation 6, d_f is the effective nozzle diameter (adjusted for cavitating flow) and $\alpha/2$ is half the spreading angle of the model jet, as defined in Figure 12. Although Equation 4 was developed for a non-vaporizing spray, it also applies to vaporizing sprays. Naber and Siebers found that the differences in penetration speed and spreading angle between vaporizing and non-vaporizing sprays were small. As the jet deceleration depends on the air entrainment—through the conservation of momentum—this indicates very similar air entrainment in the two cases.

A number of interesting conclusions can be drawn from Equations 4 through 6. The most obvious one is perhaps that $\overline{\phi}$ decreases with increasing distance from the nozzle, that is, more air is entrained over a longer distance. Equation 6 also implies that:

- Air entrainment increases strongly with decreasing orifice diameter.
- Air entrainment has a weaker dependence on the density ratio between fuel and ambient gas than on orifice diameter.
- The spatial structure of the jet has no explicit dependence on injection pressure. This indicates that air

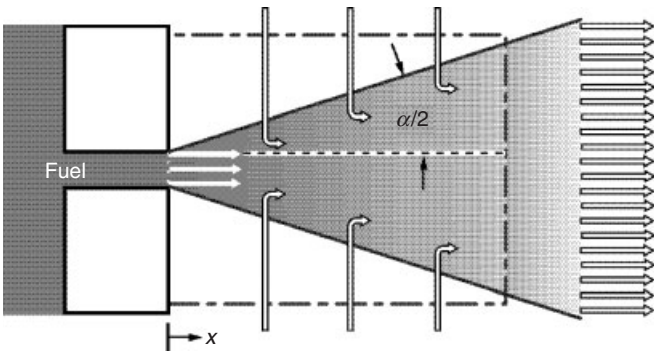


Figure 12. Idealized model jet. (Adopted from Naber and Siebers (1996). Copyright © 1996 SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

entrainment increases in direct proportion to the increase in fueling rate that occurs with increasing injection pressure.

The spreading angle α mainly depends on orifice flow effects, such as cavitation, and on the density ratio between fuel and ambient gas (Siebers, 1999). A larger jet spreading angle reflects more air entrainment, as more air is present inside a wider cone.

In summary, the nozzle hole diameter and the ambient gas density are the main factors affecting the air entrained per unit of fuel, and thus the variation of the average fuel–air equivalence ratio as function of distance from the nozzle.

As previously discussed, the flame lift-off length is an important variable for soot formation as it limits the amount of air entrained into the jet upstream of the premixed reaction zone. A longer lift-off length allows air entrainment over a longer distance. Numerous investigations of the lift-off length have been made in optical engines and spray chambers. A comprehensive experimental database shows that the lift-off length increases as the injector orifice diameter or the injection pressure increases. More specifically, it scales with the square root of the pressure drop over the nozzle, that is, the lift-off length has a linear relationship with the injection velocity U. Furthermore, the lift-off length decreases when:

- The ambient temperature increases.
- The ambient density increases.
- The ambient oxygen concentration increases.

On single, quasi-steady jets under quiescent diesel conditions, the lift-off length, H, scales with these parameters according to

$$H \propto T_a^{-3,74} \rho_a^{-0,85} d^{0,34} U Z_{st}^{-1} \quad (7)$$

where T_a is the temperature of the ambient gas, ρ_a is the ambient density, d is the nozzle orifice diameter, and U is the injection velocity. Z_{st} is the stoichiometric mixture fraction, that is, the ratio of the fuel mass to the total mass of fuel and ambient gas in a stoichiometric mixture. The practical implications of this relationship will be discussed further in Section 5.2 below. These dependencies have been empirically established in a constant volume combustion vessel (e.g., Pickett, Siebers, and Idicheria, 2005), but a fundamental understanding of the underlying mechanisms is still missing. Further research is also needed to clarify the effects of fuel parameters on the lift-off length, such as the cetane number defined below. In an engine, the interaction among adjacent jets on multi-hole injectors (Chartier *et al*., 2013), the influence of the piston bowl, and in-cylinder flow effects such as swirl (Neal and Rothamer, 2012) may also become important.

4.3 Ignition behavior

Ignition delay is an important quantity as it determines the amount of premixing that the fuel undergoes before combustion. This impacts combustion noise, formation of soot, and the oxidation of hydrocarbons and CO formed in lean mixtures. Ignition delay is determined by both physical and chemical processes. After the fuel is injected into the cylinder, it must vaporize and mix with the hot ambient gases before the chemical processes can commence. Several variables impact this initial atomization, vaporization and mixing process, including injector geometry, injection pressure, ambient gas density, and temperature, and the physical properties of the fuel. Once the fuel–air mixture is "prepared," the chemical ignition process commences. Considerable progress has been made in the last two decades in understanding the chemistry of ignition of the various compounds found in typical diesel fuels and in developing detailed chemical kinetic mechanisms to simulate the ignition process. Additional details can be found in (Fundamental Chemical Kinetics). For the purposes of this background, we will focus only on the impact of two major variables on ignition delay: the fuel–air equivalence ratio and the temperature of the prepared mixture.

Although ignition delay is not a property of the fuel alone, fuels possess an inherent ignition quality, which is traditionally described by the fuel cetane number. Cetane numbers are measured in a standardized test employing a variable r_c test engine following ASTM Method D613 (ASTM International, 2008a). Owing to the expense of running the original cetane number test, many fuels are now characterized using a derived cetane number, following ASTM method D6890 (ASTM International, 2008b). High cetane numbers imply high ignition quality, or short ignition delay. The current European Union fuel specification EN 590 requires a minimum fuel cetane number of 51. In North America, the minimum cetane number is 40, although typical cetane numbers average 45–50. Japanese regulations specify cetane numbers >45. In general, fuels rich in normal alkanes have higher cetane numbers, and cetane number increases with the length of a normal alkane. Branched, iso-alkane compounds have lower cetane number, and the cetane number decreases with the number of branches. Naphthenes and aromatic compounds have the lowest cetane numbers.

Figure 13 depicts the variation of ignition delay with temperature computed from simulations of the ignition of a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

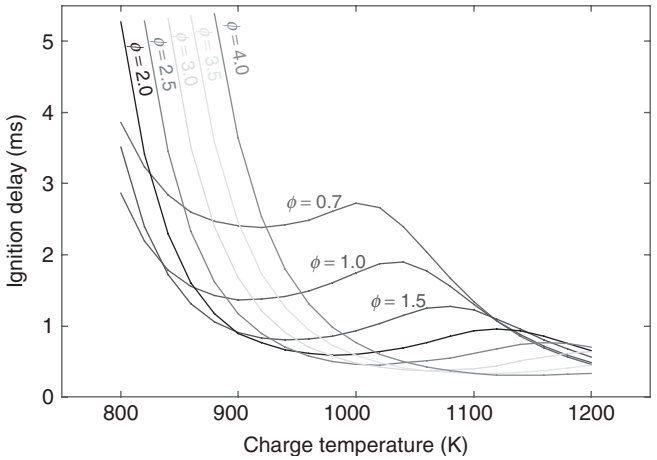


Figure 13. Temperature and equivalence ratio dependency of the ignition delay of n-C_7H_{16}. The ambient composition is 12.7% O_2 and 87.3% N_2.

homogeneous mixture of n-C_7H_{16} at a constant pressure of 50 bar, for a variety of different mixture equivalence ratios. The initial temperature was determined for each ambient temperature and equivalence ratio by assuming an adiabatic fuel vaporization and mixing process. The ignition delay is defined as the time from the start of the simulation until the peak rate of temperature rise associated with the high temperature, OH-driven heat release.

There are several important observations that can be made from Figure 13:

- At lower temperatures, the rapid rise in ignition delay, which is particularly pronounced for mixtures with $\phi > 1$, corresponds to an exponential dependency of ignition delay on T^{-1}. As will be seen below, lowering the charge compression temperature is a powerful method for increasing ignition delay and charge premixing.
- At temperatures typical of conventional diesel combustion in high r_c engines (~1000 K), richer mixtures are expected to ignite first, with very little difference in ignition delay times for ϕ between 2 and 4.
- At temperatures characteristic of low r_c diesel engines (~850 K), near-stoichiometric mixtures are expected to be the first to ignite. These mixtures will reach a high flame temperature, and on subsequent compression by combustion of the remaining mixture, will likely be the hottest regions in the cylinder and the regions where NO_x formation is greatest.
- At intermediate temperatures, the ignition delay is not as sensitive to T, due to the "negative temperature coefficient" or NTC behavior of typical diesel fuels. In this range, decreased temperature can actually lead to reduced ignition delay. The practical consequence of this behavior is that compression temperatures must be reduced below approximately 950 K before decreases in T become particularly effective in increasing ignition delay.

4.4 Heat release in diesel engines

The combustion process in engines is monitored by the rate of heat release as determined from the cylinder pressure trace, following the procedures described in (Pressure and Heat Release Analysis). Traditional heat release curves from diesel engines are divided into three phases. These are delimited by vertical lines in Figure 14, which show the rate of heat release at 50% load in a heavy-duty diesel engine operated with a single injection.

The first phase is the premixed burn and is typically considered to begin when the heat release rate turns positive. Combustible mixture that has been prepared in the cylinder between the start of injection and this point burns in a premixed volumetric reaction zone at a high combustion rate, determined by the rate of the chemical reactions. This produces a steep increase in the cylinder pressure, which is the origin of the characteristic clattering diesel sound. In high r_c engines, the premixed reactions take place in fuel-rich zones, characterized by equivalence ratios ranging from approximately 2 to 4, in the leading portion of the developing diesel jet (Dec, 1997). When a large portion of this mixture has been consumed, the heat release rate temporarily drops, indicating the end of the phase.

The period from the end of the premixed burn until the end of fuel injection is the main mixing-controlled burn

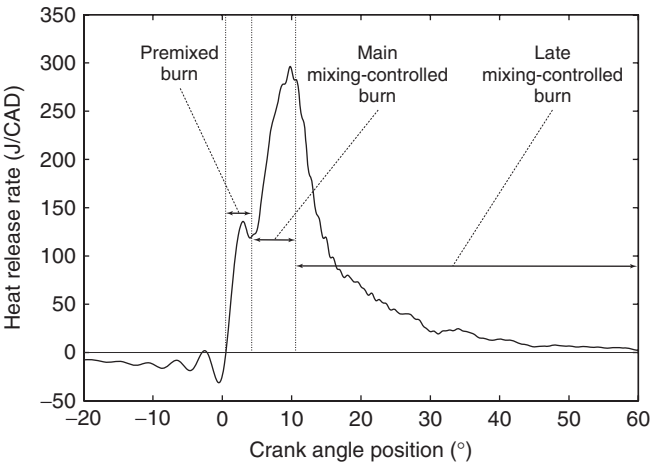


Figure 14. Rate of heat release from a 13-L heavy-duty diesel engine operated at 50% load.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

period. During this phase, combustion occurs in a quasi-steady jet as described in Figure 11. A turbulent diffusion flame has formed around the jet where combustion products from the rich, premixed reactions are consumed. The rate of heat release rate is limited by the rate of mixing between air and these products, which is driven by the spray-induced turbulence. In reality, as described in the context of Figure 11, premixed and mixing-controlled combustion occur simultaneously in different parts of the jet, making the division into these different periods less exact than Figure 14 implies. The curve merely indicates which type of combustion is dominating at different times. The end of injection is not explicitly indicated in Figure 14, but can often be seen as peak or a sudden change in the slope of the heat release rate.

After the end of injection, remaining fuel or combustion intermediates burn in a diffusion flame. This latter period is called the *late mixing-controlled burn*. The main difference from the preceding period is that the spray no longer drives the mixing. As the spray turbulence dissipates, the heat release rate drops rapidly and then gradually tails off to zero. Owing to the expanding motion of the piston, the mixture cools down, which may lead to poor combustion efficiency if the late-cycle mixing is too low.

The period between needle opening and the start of the premixed burn is referred to as the *ignition delay period*, described in the previous section. It is not indicated in Figure 14 as it is not an actual part of the heat release. During this period, the injected fuel is atomized into fine droplets, heated, and vaporized by mixing with hot air until a combustible mixture is formed. As the heat needed for vaporization is absorbed from the in-cylinder gas, it is seen as a negative rate of heat release in the diagram before the start of the premixed burn.

It is common to see a more prominent premixed burn in textbook heat release curves (e.g., Heywood, 1988). Typical injection pressures have increased steadily over the years and have accelerated the mixing and shortened the ignition delay in modern engines. As a result, diesel engines today only display a small portion of premixed heat release. The reason why a 50% load case was chosen for Figure 14 is that at 100% load, the injection pressure has increased enough to render the premixed peak all but invisible. At full load, which is a more representative operating condition for heavy-duty engines, the heat release is dominated by an extended main mixing-controlled phase and the subsequent late mixing-controlled phase.

Typical heat release curves change as engine technology advances. Today, light-duty diesel engines use several injections during the same cycle. This makes it more difficult to divide the heat release into the different modes of combustion outlined above, but it is still instructive to show how advanced injection strategies affect the combustion heat release. Figure 15 shows a heat release rate from a light-duty engine operated at 20% load, which is a typical urban driving condition. The main heat release peak is preceded by a peak associated with an earlier pilot injection and succeeded by another from a later post injection.

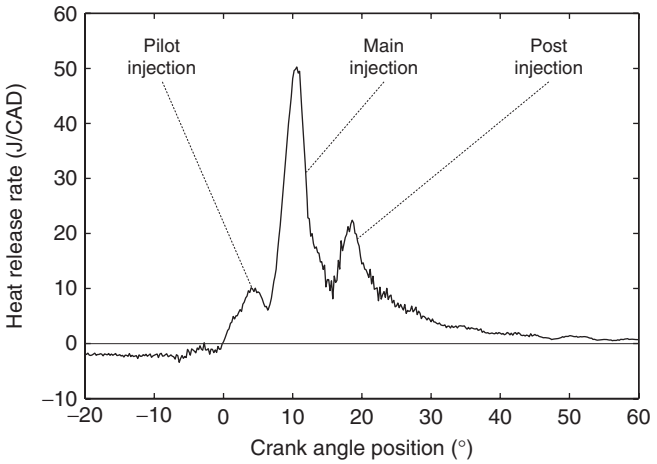


Figure 15. Typical rate of heat release from a light-duty diesel engine operated at 20% load (4 ar IMEP).

The pilot injection is a small quantity of fuel injected before the main injection. Its primary purpose, described in greater detail below, is to decrease the ignition delay of the main injection—thereby decreasing combustion noise and light-load HC and CO emissions. A post injection is a small amount of fuel injected after the main injection, during the tail of the late mixing-controlled burn. The purpose is to improve the conditions for soot oxidation, by increasing either the turbulent mixing rate or the temperature, and the result is decreased PM emissions. Notice that with a post injection, the mixing-controlled tail of the heat release curve is delayed, and efficiency is expected to decrease. Paradoxically, this is not generally the case, as will be discussed further below.

4.5 In-cylinder flow processes

In-cylinder flow processes are covered in detail in (In-Cylinder Flow), and we review only the essentials here. There are conventionally two types of organized flow motion in diesel engines. As previously mentioned, swirl is the organized rotating motion about the cylinder axis, set up by the intake ports. The second type is squish, which will be introduced shortly. Flow swirl can significantly affect both the initial mixture preparation process and the late-cycle oxidation processes, and the optimum level is dependent on both. A third type of flow motion, tumble, is an organized

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

rotating motion with an axis perpendicular to the cylinder axis. While tumble is of great importance in spark ignition engines, it does not appear to improve diesel combustion and will not be considered further here.

Rotating flow structures such as swirl are impacted by compression. For a flow swirling about the cylinder axis, axial compression will have little effect on the angular velocity of the swirl, as the moment of inertia of the rotating flow is not appreciably affected. However, when the flow is compressed radially into the piston bowl, the moment of inertia decreases by approximately the square of the bowl-to-bore diameter ratio, and the rotational velocity must increase accordingly to conserve angular momentum. Measurements indicate that the resulting "bowl" swirl ratio is approximately 60% of the theoretical value computed from the reduction in the moment of inertia (Arcoumanis, Hadjiapostolou, and Whitelaw, 1991). This process results in an increase in the rotational kinetic energy of the flow—work is performed on the flow as it is compressed into the bowl. With a typical B/S ratio of 0.9, a bowl-to-bore diameter ratio of 0.6, and a swirl ratio at intake valve closing of 2.5, the maximum swirl velocity U_θ in the bowl will be roughly seven times the mean piston speed $\overline{V}_p$.

The second organized flow structure is the "squish" flow, formed as fluid is forced out of the squish volume by the rising piston motion (Figure 16). For typical engine geometries, the squish flow (and the reverse squish flow that occurs during expansion) peaks within 10° of TDC and reaches a magnitude closely comparable to the peak swirl velocity—roughly five to six times the mean piston speed.

The squish flow and the flow swirl interact. As a fluid element is displaced inward, conservation of angular momentum dictates that the tangential velocity U_θ increases to keep the angular momentum—the product rU_θ—constant. With sufficient swirl, the increasing centrifugal force acting on the fluid element as it moves inward overcomes the competing influence of the radial pressure gradient $\partial p/\partial r$, and strongly opposes the inward movement. Thus, with low swirl, the squish flow penetrates inward to near the cylinder centerline, whereas with high swirl, the inward penetration of the squish flow is impeded and the squish flow instead descends into the bowl. As a consequence, the rotational structure formed in the vertical r–z plane rotates in the opposite direction as in the low swirl case, as shown on the right-hand side of Figure 16.

The displacement of angular momentum by strong radial flows, which subsequently drives the development of vertical plane flow structures, is an essential feature of swirling flows in diesel engines. The strong, vertical plane flow structures formed by the interaction of the fuel sprays with the lower bowl lip and bowl wall can also displace high angular momentum fluid from near the bowl lip inward, which then attempts to return to the cylinder periphery due to the increased centrifugal forces. This process is illustrated in Figure 17, depicting how the vertical plane vortices shown previously in Figure 9 are formed. Like the squish–swirl interaction, the development of these vortices arises from a balance between the radial momentum imparted by the fuel sprays and the centrifugal forces imparted by the flow swirl.

There are several important consequences of the displacement of high angular momentum fluid inward and the formation of vertical plane vortical structures:

- Like the squish–swirl interaction discussed above, work is done on the high angular momentum fluid and its rotational kinetic energy increases as it is displaced inward by the spray-generated flow structures. This additional rotational energy is provided by the translational kinetic energy of the fuel sprays. Thus, this process allows a fraction of the kinetic energy of the fuel injection event to be stored in the rotating mean flow for later release

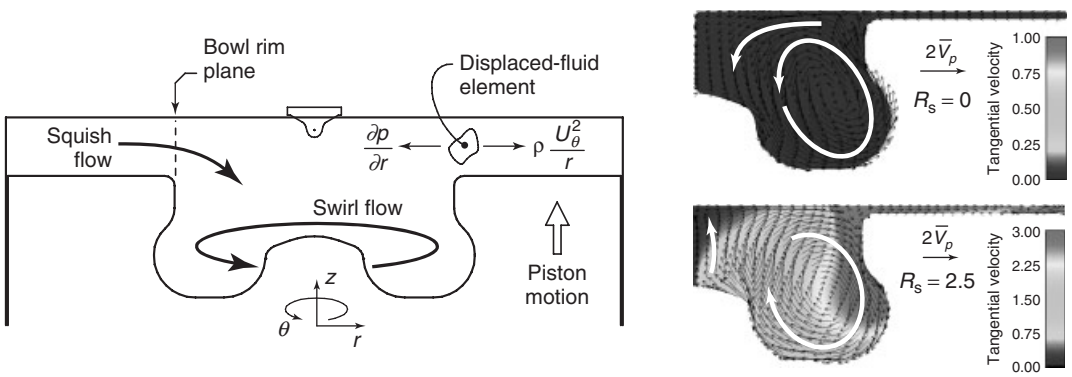


Figure 16. Illustration of the dominant in-cylinder flows and the mechanism by which they interact—conservation of angular and radial momentum.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

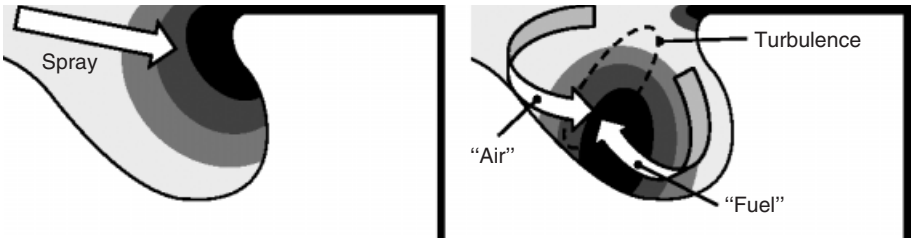


Figure 17. The fluid redistribution in the bowl due to injection. Half the bowl region is depicted from the cylinder centerline to the cylinder wall. The black regions correspond to gas with high tangential velocity. The lighter gray the color, the lower the tangential velocity of the gas. The left and right pictures correspond to the situations before and after injection, respectively. (Reproduced from Andersson (2010). © Wiley-VCH Verlag GmbH & Co. KGaA.)

into turbulence. Re-entrant bowl geometries promote this process. Simple estimates based on conservation of angular momentum (Miles, Rempelewert, and Reitz, 2003) suggest that displacing 25% of the combustion chamber mass from the outer bowl periphery to the bowl center can increase the rotational kinetic energy of the mean flow by over 50%.

- The displacement of high angular momentum fluid inward can result in formation of negative radial gradients of angular momentum. Such flows are inherently unstable (Tritton, 1977) and will result in a rapid breakdown into turbulence. Even when negative radial gradients in angular momentum are not formed, any departure in U_θ from a solid-body-like distribution will result in internal fluid deformation and turbulence production.
- When a dual-vortex structure is formed, a turbulent stagnation plane results (Figure 9), where there is large flow deformation due to radial and axial gradients in U_θ and due to r–z plane gradients in U_r and U_z. This region of high flow deformation is expected to result in considerable turbulence production.

The latter two bullets represent the mechanism by which the kinetic energy stored in the rotational flow is released as turbulence.

The practical importance of these vertical flow structures lies in the fact that they can also provide for bulk transport of unburned fuel or partially oxidized products to the same stagnation plane where turbulence generation and mixing rates are high (Andersson *et al*., 2009; Crosse, 2010; Miles, 2008). The predicted formation of these structures has been shown to correspond to a distinct increase in the measured rate of heat release (Miles *et al*., 2005), and to also correlate closely with combustion system designs providing low soot emissions (Andersson *et al*., 2009; Crosse, 2010).

The balance between radial momentum provided by the fuel injection event and centrifugal forces associated with flow swirl will be impacted by injection pressure, nozzle hole size, targeting on the bowl rim, the shape of the lower bowl lip, outer wall and pip, as well as the swirl ratio. These variables can all be expected to impact the formation of the dual-vortex structure. Excessive radial momentum (higher loads, higher injection pressures) might be expected to result in too large a lower, outer vortex, reducing the size of the upper vortex and its ability to transport O_2 to the stagnation plane (Andersson *et al*., 2009). Conversely, too high a swirl ratio can result in a small outer vortex that traps the partial oxidation products within the bowl (Kook *et al*., 2006).

Several of the practical design considerations cited above are consistent with this momentum balance and suggest a working hypothesis that late-cycle oxidation, perhaps more so than initial mixture preparation, is the process impacted most by flow swirl:

- Since flow velocities scale linearly with engine speed, at high engine speeds, one would expect a lower swirl ratio to be optimal, for the same injected fuel quantity and injection pressure.
- With a large bowl diameter, the increase in swirl as the flow is compressed into the bowl is less, and higher swirl ratios at intake valve closure (IVC) may be required to maintain a proper balance. On the other hand, the jet velocity will decrease at larger radii, which may reduce the need to increase swirl.
- A greater number of nozzle holes will reduce the mass and radial momentum in each fuel spray, and a lower swirl ratio will be required to achieve a balance.

A few additional observations should be made, which are relevant to the below discussion of multiple injection strategies. First, a poorly designed pilot injection can displace the high angular momentum fluid from the region near the bowl lip, such that the fluid entrained and transported inward by the main injection does not possess

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the necessary high tangential velocity needed to create the desired vertical plane vortical structures. Pilot injections may thus increase soot through reducing premixing by reducing the ignition delay *and* by impeding the formation of flow structures that promote late-cycle oxidation. Second, a post injection—intended to increase late-cycle mixing rates—could potentially disrupt beneficial flow structures and ultimately prove harmful.

Lastly, heavy-duty combustion systems are also often swirl supported. We reiterate that in these systems, a guiding principle is to ensure that the sprays do as much of the mixing work as possible. Nevertheless, swirl is often used to reduce soot emissions. Very little is known regarding the impact of swirl and how its interaction with the fuel sprays can impact flow structures in heavy-duty engines. However, recent measurements (Dembinski and Angstrom, 2012) show that similar mechanisms to those observed in light-duty engines are operative. The fuel injection and combustion process displace high angular momentum fluid inward, leading to significantly enhanced angular velocities in the center regions of the cylinder, undoubtedly promoting enhanced turbulence production and mixing.

4.6 Combustion process details related to emissions

If the main advantage of the diesel engine is its efficiency, its main drawback has traditionally been its emissions of oxides of nitrogen, collectively known as NO_x, and PM. Owing to ever-stricter emission legislations, NO_x and PM emissions have decreased steadily over the years. In Europe, the Euro 6 standard for light-duty vehicles brings an almost complete convergence of diesel and gasoline emissions of NO_x and PM by 2014 (Regulation (EC), 2007), whereas in the United States, there is already no distinction between the allowable emissions from gasoline and diesel engines.

Diesel PM consists of a solid part, soluble organic material, and sulfuric acid. The solid part consists of carbonaceous material (soot) originating from the combustion process, ash compounds from lubrication oil additives and particles generated by engine wear, and sulfates. The soot particles dominate (Majewski, 2002) and this is the only part of PM that will be treated in this chapter. Diesel PM is connected with respiratory and cardiovascular diseases (Peng *et al*., 2009). As soot particles are efficient light absorbers, they have been identified as major short-lived climate forcers and may also affect the climate by acting as condensation nuclei for cloud formation (UNEP/WMO, 2011).

Chemically, soot formation is very complex. It involves decomposition and oxidation of fuel, formation and growth of PAHs, coagulation and growth of these into primary particles, followed by aggregation into larger agglomerates, as well as growth on the particle surfaces. The soot formation rate is strongly dependent on temperature and the local mixture equivalence ratio. Generally, ϕ must exceed a value of roughly two for significant soot formation to occur. Soot particles are also oxidized during combustion. This process requires temperatures above 1300 K and takes place at the particle surfaces. Hydroxyl radicals (OH) and O_2 are generally considered to be the most important oxidizing species (Tree and Svensson, 2007).

NO_x emissions can arise through a number of mechanisms, but when using fuels that do not contain nitrogen, oxidation of atmospheric nitrogen by the so-called thermal mechanism is the major source. More specifically, this mechanism is responsible for NO formation, and the remaining part of NO_x (NO_2) is subsequently formed from the NO (Miller and Bowman, 1989); —see (NO_x Formation and Models) for further details. The thermal mechanism is characterized by a strong temperature dependence and relatively slow reaction rates. On timescales typical of engines, significant amounts of NO_x are generally not produced at temperatures below 1800 K (Heywood, 1988). Thermal NO_x is formed in stoichiometric or slightly lean mixtures. In the atmosphere, NO_x contributes to the formation of photochemical smog—a common problem in large cities that is associated with respiratory irritation.

The classic dilemma of diesel combustion is the trade-off between soot and NO_x, meaning that decreasing one of these emissions usually results in the other increasing. This is often explained by these emissions having different temperature dependencies; elevated temperatures promote both NO_x formation and soot oxidation (see, e.g., Heywood, 1988). Figure 18 shows results from a computational fluid dynamics (CFD) simulation of a typical light-duty diesel combustion cycle. It illustrates that soot and NO_x concentrations typically develop in quite different ways as function of time. The NO_x curve is dominated by a steep increase in the early parts of the cycle. This is because mixture that burns early is characterized by high temperatures and has time to form NO_x via the relatively slow thermal mechanism. When expansion cools the in-cylinder gas, the chemistry freezes and the NO_x levels in the exhaust gases thereby tend to be substantially higher than the equilibrium concentration at the ambient conditions (Heywood, 1988). The boot at the beginning of the NO_x curve is due to the pilot injection. Looking at the soot curve, on the other hand, we see a steep increase in the beginning of the cycle and a slower, but large, decrease during the expansion stroke. This is because soot is first formed in fuel-rich zones under high temperature conditions, and later oxidized in near-stoichiometric zones.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

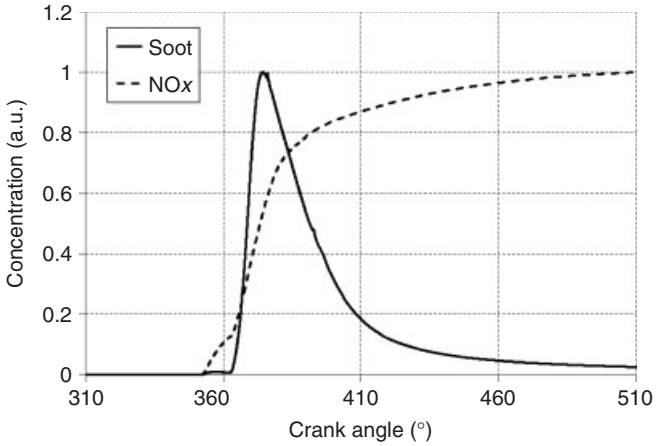


Figure 18. In-cylinder soot and NO_x concentrations as function of crank angle position. Data from a CFD simulation of a typical light-duty diesel combustion cycle. (Reproduced from Andersson (2010). © Wiley-VCH Verlag GmbH & Co. KGaA.)

Soot emissions are thereby dictated by two competing processes of comparable magnitude, whereas trends in NO_x emissions largely can be understood in terms of formation only. As a result, the trade-off between soot and NO_x in a given combustion system can often be improved as these emissions are formed through different mechanisms in different regions of the combustion chamber.

Apart from NO_x and PM, tailpipe emission standards include unburned hydrocarbons (UHCs) and CO, resulting from incomplete oxidation of the fuel. These emissions have traditionally been less challenging for diesel engines than for gasoline engines. Today, tightening NO_x requirements call for decreasing the combustion temperature, for example, by using EGR, and this tends to increase UHC and CO at low loads. UHCs mainly arise from four sources: rich zones, where the oxygen deficit impedes the oxidation process; zones too lean for complete oxidation; cold zones, for example, near walls; and unburned fuel from the sac volume of the injector. CO mainly becomes problematic when the final combustion stages have not been completed before the bulk temperature drops below approximately 1500 K during the expansion stroke. The oxidation step from CO to CO_2 then cannot be completed (Heywood, 1988).

5 OVERVIEW OF THE COMBUSTION PROCESS

In this section, we build on the fundamentals presented above to describe the mixture formation, combustion, and emissions processes that occur during diesel combustion. We subsequently link this picture back to the practical design information presented in Section 3, showing how the design and operating parameter choices impact the overall combustion process.

5.1 Conventional, quasi-steady diesel combustion

The progress of diesel combustion is dictated by the relative rates of two inter-related physical processes: turbulent fuel–air mixing and the kinetics of the fuel oxidation. A useful, albeit qualitative, tool for illustrating the process is a map (Kamimoto and Bae, 1988) illustrating zones of soot and NO_x formation as a function of fuel–air equivalence ratio ϕ and mixture temperature T. An example of such a map, based on the calculations of Kitamura *et al.* (2002), is shown in Figure 19. Also shown on the left-hand side of the map is a dashed gray line that depicts the mixture temperature achieved when the fuel vaporizes and mixes adiabatically with ambient air at 1000 K. Finally, toward the central portion of the map, a solid gray line is shown, which indicates the locus of maximum flame temperatures achieved within 1 ms of ignition. For near-stoichiometric and richer mixtures, the maximum temperature can considerably exceed the adiabatic flame temperature before endothermic reactions lower the final temperature to its equilibrium value.

We will be concerned with the "path" that a fuel element traces through the ϕ–T space as combustion progresses. This mixing process can be naturally characterized by

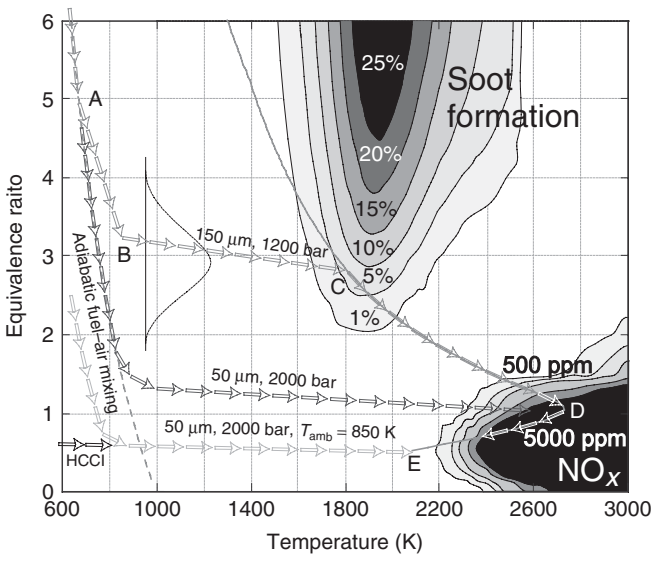


Figure 19. Regions of soot and NO_x formation computed by Kitamura *et al.* (2002) for n-C_7H_{16} at a pressure of 60 bar and a residence time of 2 ms. Potential paths of diesel combustion in an undiluted (21% O_2) ambient charge are also shown.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the mixture fraction Z, which physically represents the fraction of the mixture mass that originated from the fuel. In the absence of combustion, Z can be related to the equivalence ratio by $Z = \phi/(\phi + f_s)$; where f_s is the stoichiometric charge-fuel mass ratio introduced in Equation 4. After the start of combustion, the relationship between ϕ and Z changes, and depends on the progress of the combustion event and on the specific definition employed for ϕ. In this chapter, we will follow common practice in the diesel engine community and refer to the mixture state by the equivalence ratio that would exist in the absence of combustion, such that Z can be readily obtained from the above relationship.

A connection between the progress of combustion in the $\phi-T$ space of Figure 19 and the physical coordinates of space and time is most readily achieved by considering Figure 19 in conjunction with the conceptual model of Dec and coworkers depicted in Figure 11 (Dec, 1997; Flynn *et al*., 1999) and with the mixing rate estimates provided by Siebers and coworkers (Naber and Siebers, 1996; Siebers, 2008). In what follows, we discuss the "path" that a fuel element traces through the $\phi-T$ space (shown by the gray arrows in Figure 19) as the combustion process progresses. Similar descriptions, some of which address changes in the location of the soot and NO_x formation zones as combustion parameters are varied, have been provided by others (Denbratt, 2010; Golovitchev *et al*., 2007; Pickett *et al*., 2006). In general, the path is dictated by the relative magnitudes of the time-scale characterizing the rate of fuel–air mixing experienced by a fuel element and the time-scale characterizing the rate of heat release (temperature increase) due to combustion. Information on the relative magnitude of these time-scales can be obtained from the cross sectionally averaged mixing-rate model by Naber and Siebers (1996) and combustion time-scales predicted using detailed chemical kinetic mechanisms.

Initial fuel–air mixing time-scales are very short. For conventional diesel operating conditions, a "typical" fuel element reaches the lift-off length (point "A" in Figures 11 and 19) in less than 100 μs, at which point the cross-sectional average jet equivalence ratio may be ~4–5 and the temperature ~700 K. The fuel jet subsequently begins to entrain hot combustion products, and the temperature rise per unit of entrained mass can be expected to be roughly double that of a non-combusting jet. With the increased rate of temperature rise, the path of the fuel element departs from the adiabatic mixing line and reaches a threshold temperature of approximately 850 K at an equivalence ratio somewhat greater than 3 (point "B"). For temperatures above 850 K, the chemical heat release rate becomes significant, and the maximum flame temperature is reached in less than 100 μs (point "C"). Fuel–air mixing rates have dropped considerably by this time, and the equivalence ratio is expected to decrease by only about ~0.5 over this period if mixing rates are comparable to a non-combusting jet. Entrainment and mixing are generally impeded by heat release in burning jets, however, and the change in ϕ can be expected to be somewhat smaller. The essential point to recognize is that over the portion of the path from "B" to "C", the time-scales characterizing combustion are considerably smaller than the mixing time-scales, leading to the substantial departure of the path from the adiabatic mixing trajectory.

It is important to recognize that various fuel elements will be subject to stochastic variations in the rates they mix with ambient air or combustion products, and their ignition chemistry will also be impacted by mixing with and diffusion of radicals from mixture that is more advanced in the ignition process. Consequently, at ignition (~850 K), there will be a range of equivalence ratios in the downstream region of the jet, that have been measured to be in the range of 2–4 (Dec, 1997). This distribution is represented by the Gaussian curve sketched in Figure 19. As discussed in Section 4.3, with ambient temperatures of ~1000 K, the ignition delay times for mixtures with equivalence ratios between 2 and 4 are expected to be quite similar.

After the rich premixed combustion, fuel elements continue to mix with additional products of stoichiometric combustion from the diffusion flame, during which time the temperature increases and "ϕ" decreases ("C" to "D"). During this "mixing-limited" period, the process is represented by a path that is similar to the locus of maximum flame temperatures, as is indicated in Figure 19. More rigorous estimates of the temperatures (Pickett *et al*., 2006) experienced by a typical fuel element during this period indicate that temperatures would likely exceed the maximum flame temperature, as the products of combustion from the diffusion flame have not had sufficient time to undergo the relatively slow endothermic reactions that take place when they mix with the products of the rich premixed burn. The process is thus not truly limited by mixing rates, but is also impacted by finite-rate chemistry; higher temperatures are expected when mixing rates are high during this period.

An aspect of the combustion process that is missing from Figure 19 is the cooling of the in-cylinder gases that occurs during expansion, which will predominantly impact the mixing-limited portion of the combustion process. Expansion cooling will lower the maximum flame temperatures reached, and thus reduce both soot and NO_x formation. However, soot oxidation rates will also be impacted, and it is essential that the mixing process be completed before cylinder volume expansion slows chemical reaction rates excessively. The opposing effects of volume expansion, as

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

well as other combustion system parameters, on the relative rates of soot oxidation and NO_x formation give rise to the well-known soot-NO_x trade-off. Because the decay time of the turbulence created by the fuel jets is ~10 μs (a fraction of a crank angle degree), the turbulent mixing rates fall rapidly after the end of injection (EOI)—and completing the path from "C" and "D" with sufficient rapidity to avoid quenching of oxidation reactions by expansion can be challenging. Maintaining high "late-cycle" mixing rates, as discussed in Section 4.5, is essential.

At $\phi = 1$ (point "D"), the fuel oxidation is complete, but thermal NO_x production is near its peak due to the simultaneous presence of high temperatures and high oxygen concentrations. Because NO_x production persists even beyond $\phi < 1$, it is advantageous to maintain high mixing rates after combustion is complete. High rates of post-combustion mixing will cool the hot combustion products and will help reduce NO_x emissions. As discussed above, commercial combustion systems have been designed to accomplish this purpose.

5.2 Impact of design parameters

With this background, the impact of several combustion system design parameters on the combustion process can be clarified. To be able to do so more quantitatively, we will first discuss how the air entrainment scales with important parameters.

5.2.1 *Injection pressure*

As explained in Sections 4.1 and 4.2, the spatial structure of a non-combusting jet is not changed by changes in the injection pressure P_{inj}, as the air entrainment rate changes in proportion to the changes in the fueling rate. However, with increasing P_{inj}, the time-of-flight for a fuel element to reach a given axial location is shorter due to the higher injection velocity. While the physical mechanisms that determine the lift-off length are not yet well-understood, if the lift-off length is dominated by flame propagation processes, then higher jet velocities can be expected to "blow" the flame downstream. Conversely, if the lift-off length is determined by autoignition processes, then the decreased time-of-flight of a fuel element will likewise result in an increased lift-off length. In either case, leaner mixtures undergo premixed combustion and fewer soot precursors are formed. Beyond lift-off and ignition, increased mixing rates will reduce residence times in the soot forming regions, although temperatures may be higher. Overall, increased injection pressure is an effective method of reducing soot formation and subsequent emissions.

5.2.2 *Nozzle hole diameter*

Reducing nozzle hole diameter d_f is also an effective method of reducing the equivalence ratio at the lift-off length and hence soot formation. With smaller diameter holes, the spatial structure of the jet is significantly altered, such that lean mixtures are formed more quickly, at spatial locations closer to the nozzle. Although there is a concurrent shortening of the lift-off length—which, according to Equation 7, scales as $d_f^{0.34}$ (Pickett, Siebers, and Idicheria, 2005)—the more rapid mixing dominates. In addition to formation of leaner mixtures before the lift-off length, reduced nozzle hole size, like increased injection pressure, also impacts soot formation by decreasing the residence time in the soot/precursor formation regions (Pickett *et al*., 2006). With small enough nozzle holes, or a combination of small nozzle holes and increased injection pressure, it is possible to avoid soot formation entirely—even at relatively high ambient temperatures (Chartier *et al*., 2009b; Pickett and Siebers, 2004). The path indicated in Figure 19 by the dashed blue arrows is representative of what might be expected with a nozzle hole size of 50 μm and an injection pressure of 2000 bar. Measurements made under a variety of ambient temperatures, densities, and nozzle sizes confirm the expectation that soot formation will be eliminated under these conditions (Pickett and Siebers, 2004), although NO_x formation is expected to be high.

5.2.3 *Ambient density*

Both the fuel–air mixing rate and the lift-off length are impacted by changes in ambient density, which can be effected by varying intake pressure or r_c. Like smaller nozzle diameters, increased ambient density causes more rapid mixing and alters the spatial structure of the spray—leaner mixtures are formed when the ambient density is higher. Increasing the ambient density also shortens the lift-off length, which scales as $\rho_{amb}^{-0.85}$ (Equation 7). In this case, the impact of lift-off length dominates, and increased ambient density results in richer mixtures at the lift-off length and increased soot formation (Pickett and Siebers, 2004).

5.2.4 *Ambient temperature*

Ambient temperature has a much larger effect on the equivalence ratio at the lift-off length than ambient density, and is a powerful lever that can be used to impact the degree of premixing. For the conditions of Figure 19, a reduction in r_c from 20 : 1 to 16 : 1 would reduce the near-TDC temperature by approximately 75 K, leading to a reduction in ϕ at the lift-off length from ~5 to ~3.5.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Such a change would clearly have a significant impact on the intersection of the path of the combustion process with the soot formation region. Similar benefits can be expected through reduction of intake temperature.

Pickett and Siebers (2004) have shown that, using a combination of small nozzle diameters and reduced intake temperature, lean premixed combustion yielding near-zero soot and low NO_x can be achieved even at 21% O_2—as shown by the green arrows in Figure 19. A similar path can be achieved using very early injection, such that the fuel and air become fully premixed, although not necessarily homogeneous, at low temperature. The "typical" combustion path in this case enters from the lower left of Figure 19, and is labeled "homogeneous charge compression ignition (HCCI)." More information on HCCI combustion systems can be found in (Advanced Compression-Ignition Combustion for Ultra-Low NO_x and Soot).

5.2.5 *Swirl*

Flow swirl can be expected to influence both the initial and the late-cycle mixing processes. An increase in flow swirl has been shown to generally increase the lift-off length, resulting in leaner mixtures and lower soot formation as indicated by reduced soot luminosity (Neal and Rothamer, 2012). The effectiveness of flow swirl in enhancing late-cycle mixing rates has been discussed above in the context of Figures 9 and 17. Like an increase in injection pressure, enhanced late-cycle mixing rates will reduce the residence time in soot forming regions along the path from "C" and "D" in Figure 19, thus likely reducing formation as well as increasing the rate of soot oxidation.

5.2.6 *Fuel properties*

As might be expected, fuel ignition quality, or cetane number, impacts lift-off length—shorter lift-off lengths are to be expected with low ignition delay fuels (Pickett, Siebers, and Idicheria, 2005). Fuel density also impacts mixing rates, although less strongly than ambient density. Lighter fuels form leaner mixtures close to the nozzle. The impact of fuel density on lift-off length has not been studied carefully, although there is no evidence to indicate that it will have a large effect.

5.2.7 *Effect of EGR*

Charge dilution with EGR (recirculated exhaust gas) also has a significant impact on both the fuel–air mixing process and the combustion kinetics. EGR is the foundation of in-cylinder NO_x control, although it is generally associated with an increase in soot emissions. We begin examining the impact of EGR by considering the path in the ϕ–T plane in a quasi-steady mode of combustion such as is shown in Figure 11. For near-full-load conditions, engines operate at a fuel/air equivalence ratio of about 0.7, and an EGR rate of 25% results in an intake $[O_2]$ of about 17%.

Because the diluted charge has a higher heat capacity for a given mass of O_2, during the initial adiabatic mixing process, the temperature is higher at a fixed ϕ. In addition, due to the additional diluent mass, longer mixing times are required to achieve a fixed ϕ. Both of these factors would seem to favor richer mixtures at the lift-off length. However, the lift-off length increases with the addition of EGR, such that ϕ at the lift-off length is very nearly the same as in the undiluted charge. Likewise, the relative rates of mixing processes and oxidation processes are expected to be similar for both ambient O_2 concentrations throughout the premixed burn—both rates are slowed by the presence of EGR. In the end, the path followed through the ϕ–T plane is likely to differ significantly only during the mixing-limited portion, represented by the solid gray line in Figure 19. Since dilution lowers the adiabatic flame temperature, this line will be displaced to the left.

On the basis of the lower temperature path followed with dilution, we might expect dilution to reduce both soot and NO_x formation. This expectation turns out to be only partially true. NO_x formation is certainly reduced, but while increased dilution slows soot formation rates, it increases the residence time in the soot forming regions by slowing the rate at which O_2 needed for soot oxidation is entrained. Moreover, the soot formation regions are shifted toward lower ϕ and T with charge dilution. This shift can be significant. For example, with no dilution, the production of pyrene (an important soot precursor) is 0.1% of the fuel mass with a ϕ of $\sim$1.8 and a T of $\sim$1850 K, whereas these limits move to $\sim$1.3 and $\sim$1700 K, respectively, when $[O_2]$ is reduced to 10% using EGR (Pickett *et al*., 2006). At moderate EGR rates ($[O_2] \gtrsim 15\%$), the net result is increased soot formation (Idicheria and Pickett, 2005). Only for the high EGR rates and/or the lower ambient temperatures employed for low temperature combustion strategies are both soot and NO_x formation reduced simultaneously.

5.3 Light-load operation

While the quasi-steady phase is short or nonexistent at the lower loads typical of light-duty engine operation, the spray physics and chemistry governing the formation of emissions are the same. This means that parameters that control the air entrainment into the fuel jets will affect the formation of soot and NO_x at light loads in roughly the same way as they do under quasi-steady conditions. Higher injection pressure,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

smaller nozzle holes, higher ambient density, and increased swirl all tend to increase fuel–air mixing rates, decreasing the soot formation in light-duty engines. Similarly, lowered ambient temperature slows the chemical rates, allowing still further premixing. While premixing can reduce both soot and NO_x formation, problems can arise with HC and CO emissions stemming from lean mixtures as well as with increasing combustion-generated noise. These problems are primarily addressed using multiple injection strategies, as described below.

A typical light-load heat release rate has an extended tail after the main heat release, well after the end of injection. This is the slow, mixing-controlled burn of combustion products formed earlier in the cycle, for example, soot and partial oxidation products—primarily CO. As previously mentioned, the main challenge at light loads is to maintain sufficient mixing rates after the end of injection to promote oxidation of these products. The advantages of enhanced late-cycle mixing are twofold. First, if the tail of the heat release is shifted toward TDC, the thermodynamic efficiency will increase as it allows for a greater effective expansion ratio. Second, the emissions will decrease as enhanced oxidation permits less partially oxidized products to survive into the exhaust port. Enhancing the late-cycle mixing rate is therefore an important method for reducing both soot emissions and fuel consumption at light loads.

A different approach to suppressing both soot and NO_x is to use so-called LTC (Low Temperature Combustion). This method is enabled by adopting low engine r_c and by the use of large portions of cooled EGR. These measures result in very extensive premixing by substantially extending the ignition delay. Another effect of the EGR dilution and lower compression temperature is that the peak combustion temperature is decreased, slowing both the soot and NO_x formation chemistry. LTC concepts using diesel-like fuel are usually restricted to light-load operation. This is due to the high level of dilution limiting the available air and thereby the attainable load, as well as the in-cylinder temperatures increasing with load and thereby shortening the ignition delays. Although LTC greatly reduces soot and NO_x emissions, CO and HC emissions are typically very high—as is combustion noise. LTC using diesel fuel will be described in greater detail in a separate section below.

6 MULTIPLE INJECTION STRATEGIES

Multiple injection strategies can have a number of variants, as depicted in Figure 20. One or more pilot injections of a small quantity of fuel can be employed before the main injection. When these pilot injections occur some tens of crank degrees before the main injection, they are often called *far pilots*, whereas "close" pilots—sometimes called *preinjections*—typically occur within a few crank degrees of the start of the main injection. A combination of both "close" and "far" pilots may also be used. At idle, there may be no "main" injection, just a sequence of small pilot injections. As a consequence of these numerous variants, the literature describing pilot injections is vast and confusing. Here, we provide only a brief introduction by considering just a single pilot injection.

Post injections similarly are described in a number of ways. A post injection (sometimes called an *after injection*) that is closely coupled to the main injection event is generally employed to reduce smoke emissions, as explained earlier. Post injection events can also be used with the primary purpose of controlling exhaust gas temperature or HC/CO content to support exhaust gas after-treatment devices. These latter injection events occur later in the cycle and do not significantly impact the main combustion event. Likewise, the main injection may be split into multiple components, with the objective of controlling combustion noise as well as soot. Below, we restrict our attention to close-coupled post injections, containing approximately 20% or less of the total fuel mass injected.

6.1 Pilot injections

The pilot injection, introduced in Section 4.4, is used to decrease the ignition delay of the main injection. Although

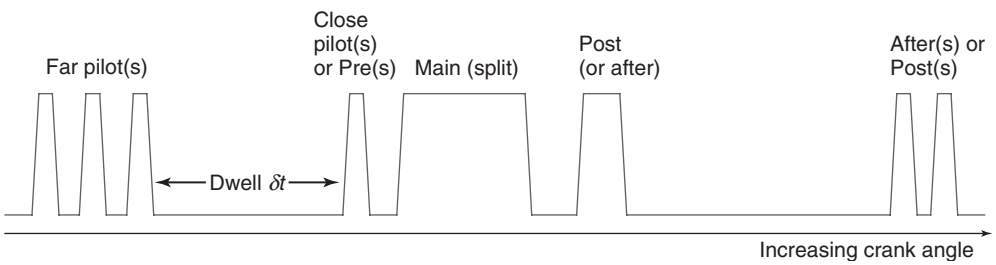


Figure 20. Definition of various injection events.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the details of the impact of the pilot injection on the main charge ignition process are not yet well-understood, modeling studies suggest that the initiation of the combustion is fundamentally changed from an autoignition process to a process resembling flame propagation, and that the nature of this change allows ignition and combustion in regions of the spray with high temperature and concentration gradients that could not otherwise support ignition (Hasse and Peters, 2005). The main impact of the reduced ignition delay is to decrease the amount of fuel–air premixing, and hence, the pressure rise rate and the acoustic noise generated during the main combustion. The pilot injection does not contribute much mechanical work and may even decrease the efficiency of the engine. Its main advantage is found in increased driving comfort, which is an important customer demand.

A second benefit of a reduced ignition delay is a reduction in overmixing of the main injection, which leads to light-load or cold-start HC and CO emissions. However, from the above discussion of the combustion process, it should be clear that increased soot emissions will also be expected when fuel–air premixing is reduced—an outcome that is almost universally observed. Pilot injections can also impact BSFC and NO_x emissions, as discussed below.

In addition to impacting ignition delay, pilot injections are useful for preventing excessive liquid penetration. With early injections, liquid fuel can impact the cylinder liner, resulting in excessive liner wear and oil dilution by the fuel (Dronniou *et al*., 2005; Oinuma *et al*., 2005). By employing multiple, short injections, liquid penetration can be reduced (Pickett, Kook, and Williams, 2009). Even when liquid impingement does not occur, however, limiting the penetration of the vapor phase may be advantageous. Reduced penetration and greater mixture stratification, due to the use of multiple short injections, are likely responsible for reduced HC emissions observed when piezo injectors are applied to low r_c engines (Tomishima *et al*., 2008).

Generally, ignition delay is shortened as the timing of a pilot injection approaches the main injection event or as the heat released from the pilot injection is increased (Badami, Millo, and D'amato, 2001; Carlucci, Ficarella, and Laforgia, 2003; Carlucci, Ficarella, and Laforgia, 2005; Park, Kook, and Bae, 2004). Consequently, light-load HC and CO typically decrease (Hotta *et al*., 2005; Kastner *et al*., 2006). With far pilots, over-penetration of the pilot fuel due to the low in-cylinder density and long mixing times can result in overly lean mixture, resulting in higher HC and CO emissions. Conversely, when the dwell time between the pilot and the main injection is too short, disruption of the pilot ignition process by the main injection event may cause overmixing of the pilot fuel, again increasing HC and CO. Consequently, a minimum HC and CO emissions is observed when the dwell between the pilot and main injection is $\sim 10^\circ$ (Kastner *et al*., 2006). It is unfortunate, then, that the increase in soot emissions is typically greatest with intermediate pilot-main dwell times, between roughly 5 and 20° (Badami, Millo, and D'amato, 2001; Hasse and Peters, 2005; Hotta *et al*., 2005; Kastner *et al*., 2006; Park, Kook, and Bae, 2004; Ricaud and Lavoisier, 2002). Soot emissions can be reduced again, however, with increased rail pressure (Ehleskog, Ochoterena, and Andersson, 2007; Ricaud and Lavoisier, 2002).

The impact of pilot-main dwell on NO_x depends on operating conditions and on the timing of the main injection, and reported results are also influenced by the specific optimization strategy employed or trade-off emphasized. Typically, however, a close-coupled pilot injection is associated with higher NO_x emissions (Hotta *et al*., 2005; Kastner *et al*., 2006), even above levels observed with a single injection (Chen, 2000; Hotta *et al*., 2005). Smaller pilots result in lower NO_x, but can increase BSFC and noise (Badami, Millo, and D'amato, 2001; Hotta *et al*., 2005). As the pilot injection is advanced, NO_x is generally reduced (Badami, Millo, and D'amato, 2001; Hotta *et al*., 2005), albeit at the potential expense of increased UHC, CO, or soot.

The physics responsible for the variation in NO_x emissions is not clearly understood, but the behavior is consistent with the expected effects of combustion phasing. When the timing of the main injection is fixed, the increase in ignition delay as the pilot is advanced retards the main combustion due to the increased ignition delay, which can be expected to reduce NO_x. Moreover, compared to the case with no pilot, combustion is advanced—which produces more NO_x as there is a greater period of time available for NO_x production before cylinder volume expansion reduces the temperature. Advancing combustion will also typically improve BSFC, and a BSFC/NO_x trade-off is observed. Well-optimized pilot injection strategies have been observed to improve the BSFC/NO_x trade-off (e.g., Ishida *et al*., 1994; Kastner *et al*., 2008).

6.2 Post injections

The mechanism by which post injections affect soot emissions is still not clearly understood, and it is not always true that post injections are beneficial. In a recent review article, O'Connor and Musculus (2013) present a compilation of data showing that under many circumstances, post injections can lead to up to an order of magnitude increase in soot emissions. Moreover, the impact of a post injection on soot cannot be considered in isolation—its impact on BSFC and NO_x, CO, and HC emissions must also be considered.

Nevertheless, it is generally accepted that with an appropriate post injection mass and timing, significant reductions

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

in soot can be achieved without a significant adverse impact on fuel economy and other emissions—particularly at low-to-moderate speeds and loads (e.g., Ricaud and Lavoisier, 2002). Modest post injections of a few milligrams that occur within a few °CA of the main injection are usually found to be most effective (e.g., Vanegas *et al*., 2008; Yun *et al*., 2008; Barro *et al*., 2012; Beatrice *et al*., 2003), although they may increase NO_x (Beatrice *et al*., 2003; Badami *et al*., 2003) and adjustments in EGR rate may be required to simultaneously improve NO_x, BSFC, and soot (Hotta *et al*., 2005). Post injections that are too late or too large can increase BSFC, HC, and CO emissions (Chen, 2000; Badami *et al*., 2003); and post injections that are too close to the main have been reported to increase soot emissions (Badami *et al*., 2003; Yun *et al*., 2008).

Several mechanisms have been proposed to explain how a post injection can affect soot emissions, which have been summarized by O'Connor and Musculus (2013):

(1) The post injection enhances mixing rates which can reduce soot formation—by redistributing the fuel from the main injections into smaller and fewer fuel-rich zones—and late-cycle oxidation—by enhancing the rate at which soot laden mixture mixes with fresh O_2. Recall, however, that the possibility exists for a post injection to disrupt flow structures that are beneficial to late-cycle mixing. Such unintended consequences may be partially responsible for reports of increased soot when post injections are used.
(2) Local temperatures are increased as a post injection burns, which increase the kinetic rates of oxidation, may also be capable of reducing soot emissions. The effect of increased temperatures is very difficult to separate from the effects of increased mixing rates, which will also serve to raise the local temperature. Moreover, as discussed in the context of Figure 19, increasing the mixing rate of fuel-rich mixture with embedded soot can raise the temperature due to finite-rate chemistry effects.
(3) Variations in the overall fuel–air equivalence ratio distribution formed when post injections are used may also impact soot formation processes. Owing to a reduction in the mass injected in the main injection, leaner mixtures and reduced soot formation are expected. In this view, fuel from the post injection is thought to form less fuel-rich mixture, and hence form less soot, than if it had been injected along with the main injection.

Lastly, in the discussion of Figure 15, we noted that with a post injection, the mixing-controlled tail of the heat release curve is delayed, and efficiency is expected to decrease. This seems to contradict the studies cited above that report an improvement in both BSFC and soot when post injections are employed. In practice, however, the efficiency is often limited by NO_x constraints, which may call for a later than optimal phasing of combustion. Transferring part of the fuel from the main injection to the post injection decreases the NO_x formation associated with combustion of the main injection. The somewhat counterintuitive result is that the main injection can be advanced, simultaneously increasing the efficiency and decreasing PM for a given NO_x target level. Alternatively expressed, the post injection improves the NO_x-PM trade-off at a given efficiency level.

7 LOW TEMPERATURE DIESEL COMBUSTION PROCESSES

Diesel LTC processes employ extensive dilution of the combustible charge to keep peak combustion temperatures low and minimize NO_x formation. This is accomplished with either extensive premixing to lean equivalence ratios or through use of EGR. LTC systems also seek to minimize soot formation through increased premixing, although not necessarily to fuel-lean conditions.

With very early injection, all of the fuel can be premixed to fuel-lean conditions before the start of heat release. In this case, the phasing of the combustion process is determined by the kinetics of the ignition process—it is decoupled from the timing of the injection event. Early injection can present problems associated with fuel impingement on the cylinder walls, and special fuel injection hardware and atypical diesel combustion chamber geometries are often employed to alleviate these problems. Examples of diesel combustion systems that employ very early-injection strategies have been reviewed by Dec (2003).

Our focus is on low temperature combustion systems that use more conventional diesel engine hardware and that allow some degree of control of the combustion phasing through selection of the fuel injection timing. Generally, this restricts the injection timing to within about 30° of TDC. Within this restriction, we can further subdivide LTC systems into early- and late-injection strategies. Early-injection strategies attempt to achieve a large degree of premixing by injecting into the cool cylinder gases at ~30° before TDC. Extensive charge dilution with EGR is used to delay ignition and to control the maximum pressure rise rate (combustion noise). In contrast, late-injection strategies typically inject fuel near or shortly after TDC. EGR rates are generally lower, and due to the higher near-TDC temperatures and densities, as well as the decreased dilution, the ignition process commences more quickly than with early-injection strategies. However, the slowing of the ignition kinetics by cooling due to volume expansion allows

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

increased premixing before high temperatures characteristic of soot and NO_x formation are achieved, and also moderates the pressure rise rate. For this reason, late-injection systems have been called *modulated kinetics (MK)* combustion systems by their initial developers (Kimura *et al*., 2001; Kimura *et al*., 1999). Modern, low NO_x engine calibrations often employ injection timings and EGR rates that are similar to late-injection LTC strategies. Both strategies are currently useful over only limited speed and load ranges; accordingly, the description provided below is typical of the combustion process under the light engine loads characteristic of the lower half of the loads experienced in an urban drive cycle.

The significant impact of injection timing, and the coupling between the chemistry of oxidation and the engine compression and expansion processes, is illustrated in Figure 21 for a moderately lean mixture with $\phi = 0.7$. Although at first glance, the evolution of the combustion process appears to be considerably different for early- and late-injection strategies, there are substantial similarities.

At low temperatures, the various stages that occur in the ignition and oxidation processes become distinct, and are separated in time by several crank degrees. The first stage of ignition, occurring at temperatures near 850 K, occurs after about 0.5 ms in the early-injection case. With late injection, ignition is advanced due to the higher temperatures and densities at the time of injection. These early reactions produce important combustion intermediates (H_2CO, H_2O_2, and CO among others) that drive the later stages of heat release. After the first stage of ignition, the heat release rate decreases and the temperature rises more slowly. With early injection, the rate of temperature rise gradually increases until at ~1000 K a second, rapid period of heat release occurs. At this time, the remaining UHCs and combustion intermediates are partially oxidized to form CO. The onset of this second, higher temperature period of heat release is considered to be the main ignition event. As temperatures and the OH radical pool rise, the CO is quickly oxidized and NO_x formation increases. Notice that even after completion of combustion, NO_x formation continues due to the high temperature, as was discussed in Section 5.1.

With late injection, the sequence of events is the same, but occurs at very different rates. After the first-stage ignition, the local gas temperature actually decreases slightly due to the expanding cylinder volume. However, beyond about 0.8 ms, the heat release rate begins to increase monotonically. The decreasing rate of temperature rise observed near 1 ms is thus due to the increasing importance of volume expansion, which largely counterbalances the heat release beyond 0.9 ms. Later, beyond 2 ms, the heat release becomes more dominant, the temperature rises more rapidly, and CO formation increases—as seen

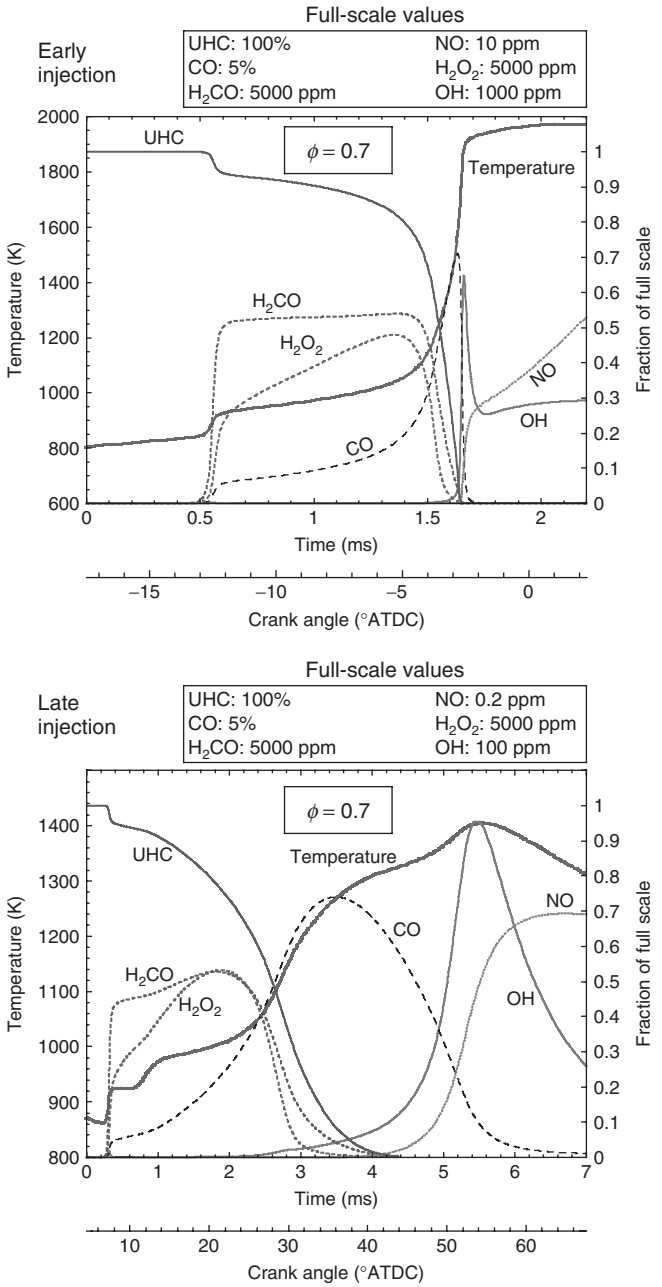


Figure 21. Comparison of the oxidation process for early- and late-injection LTC strategies. The temperatures and species concentrations were simulated using a detailed kinetic mechanism (Lawrence Livermore National Laboratory, 2013) for *n*-heptane mixed with a 12.7% O_2 ambient charge. The pressure was constrained to follow a measured cylinder pressure trace typical of each combustion strategy. (Adapted from Musculus, Miles, and Pickett (2013). © Elsevier.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

also with early injection. Nevertheless, volume expansion significantly delayed the main ignition and slowed the heat release rate and associated pressure rise rate appreciably. The final stages of combustion are also slowed, and the CO oxidation rate, which is highly dependent on the OH concentration, is significantly lower. The low OH concentrations present with late injection will also likely slow the oxidation of soot.

With this understanding of the temporal progress of combustion, the spatial structure of the reacting jet under LTC conditions, which is illustrated in Figure 22, can be more easily understood. Despite the differences in the temporal evolution seen in Figure 21, the spatial structure of the jet does not differ greatly between early- and late-injection strategies, and only a single picture is required. For typical, light-load LTC operation, the duration of the fuel injection process is less than the ignition delay, and a quasi-steady period does not exist. Accordingly, rather than showing a "path" followed by a fuel element through a steady jet, we show the progress of the transient combustion at several times during the combustion event.

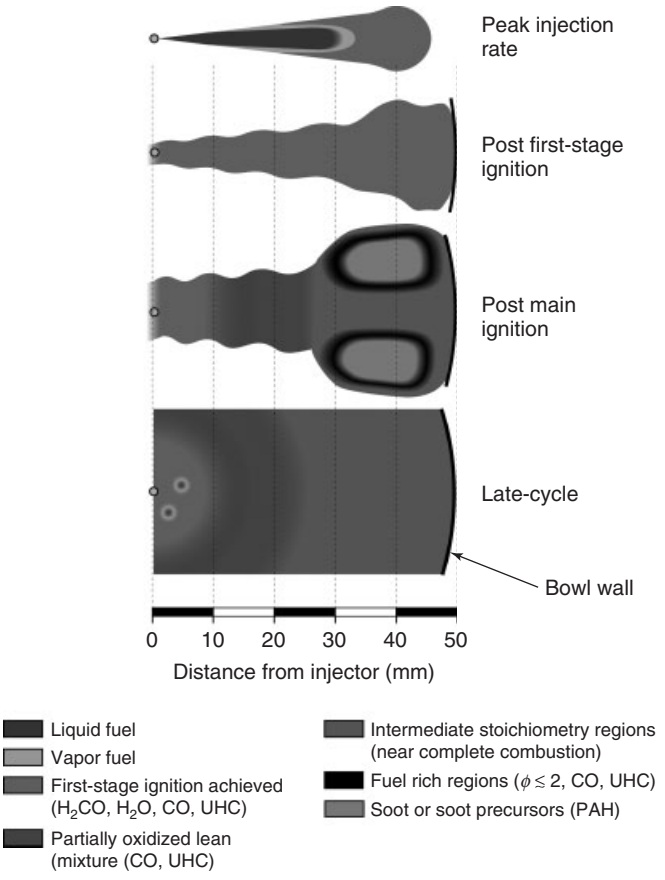


Figure 22. Heavy-duty low temperature diesel combustion at light load. (Adapted from Musculus, Miles, and Pickett (2013). © Elsevier.)

As with conventional diesel combustion, the process starts with the initial penetration of liquid fuel into the gases. For early-injection LTC, the fuel jet penetrates faster and further into the lower temperature, less dense ambient gases, and the maximum liquid penetration observed (~30 mm or more) is somewhat longer than for conventional or late-injection timings. Beyond this "liquid length," the vapor phase fuel continues to penetrate downstream, and if sufficient time has elapsed, undergoes first-stage ignition reactions. These initial reactions are observed to occur nearly simultaneously throughout the fuel vapor, in both rich and lean regions.

As the end of injection approaches, a second important process occurs. The decreasing injection rate creates a mass flux deficit near the injector, and conservation of mass dictates that ambient gas entrainment must therefore increase. The increased entrainment further reinforces the mass flux deficit and drives yet more entrainment (Musculus and Kattke, 2009). This increased entrainment near the end of injection process helps vaporize the remaining liquid and leads to the rapid formation of very lean fuel–air mixtures near the injector, which soon undergo first-stage ignition. Although this process also occurs under conventional diesel combustion, the flame surrounding the jet rapidly propagates toward the injector at the end of injection and prevents the formation of over-lean mixture.

After the first stage of ignition, the temporal evolution of the combustion process depends very strongly on the local equivalence ratio, the degree of dilution, and the injection timing. If the mixture is overly lean ($\phi \lesssim 0.2$–0.7, depending on injection timing and EGR rate), the combustion process is unable to progress beyond the first stage of ignition. Consequently, as shown in Figure 22, UHC, CO, and combustion intermediates persist in the very lean mixture near the injector. For somewhat richer mixtures, combustion progresses further and the UHCs and combustion intermediates are partially oxidized to form CO. However, if the temperature achieved is not sufficiently high, the CO cannot be fully oxidized in the available time before cylinder temperatures drop excessively due to expansion.

Still richer mixtures, characterized by intermediate equivalence ratios near unity, achieve high enough temperatures to enable the second, rapid period of heat release that results in nearly complete combustion. However, if $\phi > 1$, there is insufficient O_2 available to fully oxidize the CO. These intermediate stoichiometry regions are found in the downstream regions of the jet, as shown in Figure 22. Lastly, for sufficiently fuel-rich mixtures, typically found in recirculation zones near the head of the jet or in regions where

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

adjacent jets interact, soot precursors and soot are formed in the period following the high temperature heat release.

As the expansion stroke proceeds, fuel-rich regions mix with leaner regions and the embedded CO, UHC, and soot are oxidized. Lean regions may also mix with hot products of richer combustion, partially oxidizing the remaining products of first-stage ignition. Because the global, in-cylinder equivalence ratios are generally quite lean, however, this is far less probable than the leaning of fuel-rich regions through additional mixing, and extensive regions of UHC, CO, and other partial combustion products remain in the central regions of the cylinder. Fuel droplets that dribble from the injector late in the cycle are also observed near the cylinder center, and likely contribute to engine-out UHC and CO emissions. Additional details are provided by Musculus, Miles, and Pickett (2013) and the references therein.

For light-duty diesel engines, the combustion proceeds in a very similar manner, although the interaction of the fuel jets with the combustion chamber walls adds additional complications, and leads to differences in the way the combustion process unfolds that depend on the injection timing. Consequently, a separate illustration is provided for early- and for late-injection strategies relevant for light-duty engines in Figure 23. The uppermost row of images in Figure 23 illustrates the interaction of the fuel spray with the piston bowl at a time approaching the end of injection, when increased entrainment is acting to rapidly vaporize the remaining liquid fuel. With early injection, the liquid fuel typically impacts the bowl lip and spreads into both the squish volume and the bowl. In contrast, with late injection, the maximum liquid length is ∼20 mm, and significant liquid impingement does not occur.

The delay until the first-stage ignition reactions occur depends on the in-cylinder conditions and fuel type. In the second row of images in Figure 23, we choose to illustrate a case in which the first-stage ignition reactions do not occur until after the end of injection. At this time, most of the liquid has been vaporized, although in the early-injection case, there is evidence of liquid films on the piston top. Large fuel droplets frequently observed near the piston rim may be formed as the flow out of the squish volume shears off liquid from these films.

Shortly after much of the fuel jet achieves second-stage (main) ignition, over-lean mixture containing UHC, CO, and other partial oxidation products remains near the injector, as seen previously in Figure 22. With early injection, the region of over-lean mixture is larger and reaches to the bowl wall. Hence, separate regions of intermediate stoichiometry are seen—one in the squish volume and a second deep in the bowl. Soot is observed forming in rich mixtures near the head of the jet in the bowl, whereas additional over-lean mixture is seen near the tip of and between the jets in the squish volume. Although fuel-rich mixture is measured within the squish volume at the start of second-stage heat release (Petersen, Miles, and Sahoo, 2012), equivalence ratios are sufficiently low that significant soot or PAH is not formed.

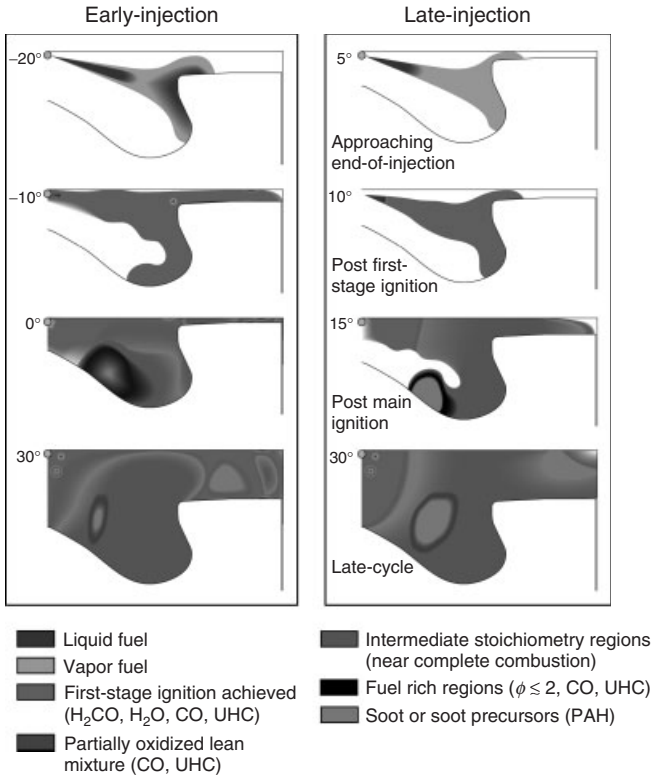


Figure 23. Illustrations of light-duty LTC progress. Separate illustrations are provided for early- and late-injection combustion systems due to differences in the way the fuel jet interacts with the combustion chamber surfaces. (Adapted from Musculus, Miles, and Pickett (2013). © Elsevier.)

With late injection, over-lean mixture is confined closer to the cylinder center, and near complete combustion occurs throughout much of the fuel jet. Owing to the retarded timing, more fuel enters the bowl, and richer mixtures leading to more soot formation than seen in the early-injection case are observed. Although the fuel jets do not initially penetrate far into the squish volume, they are displaced outward as the products of the intermediate stoichiometry combustion from near the bowl lip are drawn in by the reverse squish flow. Mixtures formed within the squish volume are too lean to support soot formation, and over-lean mixture is again observed near the tip of and between the jets (Miles, Petersen, and Sahoo, 2012).

Unlike the late-injection case, the mixture that is drawn into the squish volume with early injection consists largely

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of the first-stage ignition products from the over-lean, upstream regions of the fuel jets. Thus, UHC and CO observed within the squish volume late in expansion stem from both upstream as well as downstream regions of the jet; with late injection, only downstream regions are likely to contribute to emissions within the squish volume. In both cases, like the heavy-duty case, over-lean mixture is also found in the central region of the cylinder, where fuel drops are likewise observed. A discussion of the combustion paths through the ϕ–T map for early- and late-injection combustion systems is provided by (Kook *et al*., 2005).

7.1 Influence of equivalence ratio and combustion temperature

From the temporal progress and physical structure of LTC combustion discussed above, it is apparent that while LTC strategies can significantly reduce NO_x and soot emissions, low temperatures and over-lean mixtures can lead to significant UHC and CO emissions. Figure 24 presents the yield of UHC and CO predicted for the same constant temperature and pressure conditions as the soot and NO_x regions which have been previously presented in Figure 19. Although we expect that emissions of UHC and CO will also depend on the dilution rate and the fuel injection timing, the contours shown in Figure 24 provide a good qualitative indicator of the equivalence ratio and combustion temperature ranges that are likely to lead to UHC and CO emissions.

Figure 24 shows that the yield of UHC is high for all ϕ when the temperature is low. For temperatures below ~750 K, fuel breakdown occurs so slowly that little or no CO is formed. Hydrocarbons trapped in combustion chamber crevices suffer this fate. UHC remains high as the temperature increases from 750 to about 1000 K, although partial oxidation forms substantial amounts of CO. This region of the ϕ–T map represents over-lean mixtures formed at the edges and tail of the fuel jets, or richer mixtures formed as hydrocarbons out-gas from crevices or leak from the injector late in the cycle when the cylinder temperatures have dropped. At temperatures above about 1000 K, and for $\phi \lesssim 2$, the UHC is rapidly consumed and CO rises. Recall that this is the temperature at which the rapid, second-stage ignition processes begin—as was seen previously in Figure 21. For lean to moderately rich mixtures ($\phi \lesssim 1.3$), if the temperature exceeds 1200–1300 K, the partial oxidation of hydrocarbons to CO is sufficiently rapid and UHC yield is very low, even for the rich mixtures. CO oxidation is slower however, and temperatures above 1450–1500 K are required to fully oxidize CO (Sjöberg and Dec, 2005).

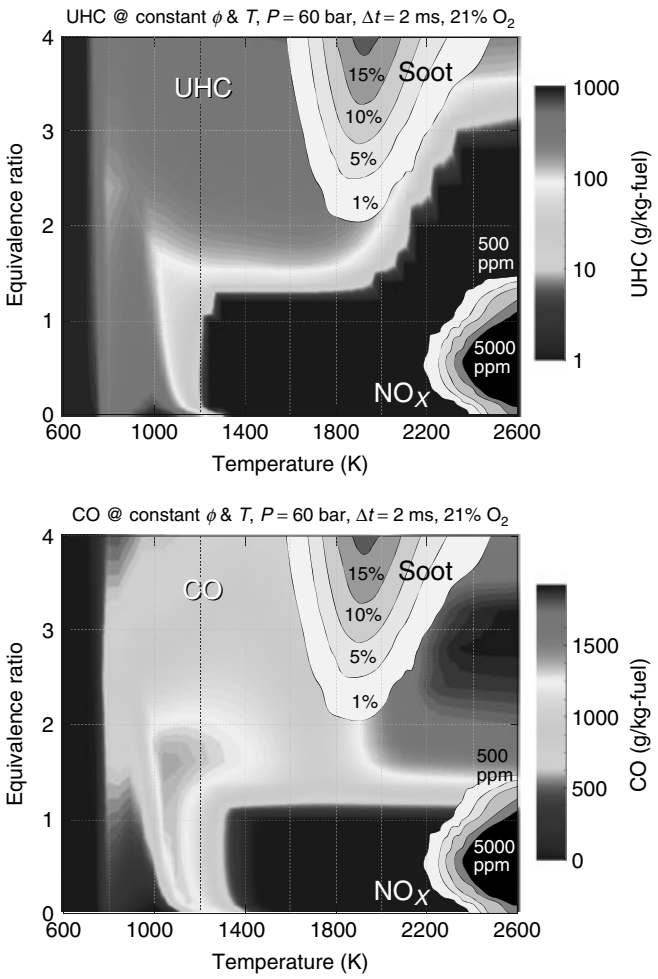


Figure 24. Yield of UHC and CO at constant ϕ, T, and P predicted using a detailed n-heptane kinetic mechanism (Lawrence Livermore national Laboratory, 2013).

The low UHC observed for mixtures with $\phi \lesssim 1.3$ invites comment. At these equivalence ratios, the remaining partial oxidation products are primarily H_2 and CO. Only at equivalence ratios greater than about 1.3 is there significant UHC in the product gases—and then combustion simulations indicate that the UHC is restricted primarily to three species: CH_4, C_2H_4, and C_2H_2. It is not until the equivalence ratio approaches 2 that significant quantities of PAH are formed, in agreement with the soot formation predictions. CO, of course, is present for all $\phi > 1$. Consequently, when adequate bulk-gas mixing is not achieved in the cylinder, CO emissions are likely to be impacted more severely by over-rich mixture than UHC emissions. Additional information regarding CO and UHC emissions is provided in (UHC and CO Formation and Models).

In summarizing this section, we reiterate that low temperature combustion strategies using diesel-like fuels are

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

currently only applicable at light loads. Because soot and NO_x are more problematic at high loads, and HC and CO are of lesser concern due to the higher average cylinder equivalence ratios and temperatures, there is great interest in extending LTC strategies to higher loads. Moreover, there is a natural synergy between LTC strategies and the use of pilot injections, which alleviate excessive combustion noise and UHC and CO emissions through a shortening of the main ignition delay. Although shortening the ignition delay will reduce the premixing characteristic of LTC strategies, high EGR rates and the maintenance of high mixing rates late in the cycle can mitigate the tendency toward increased soot production (Kitamura and Ito, 2010). Considerable work remains to be done in this area.

8 SUMMARY

In this chapter, we have attempted to review the practical design considerations impacting diesel combustion systems and to link them with a fundamental description of the various physical and chemical processes that take place in diesel engines. In the past two decades, significant advances in our fundamental understanding have allowed us to better understand how design and operating parameter choices such as combustion chamber geometry, fuel injection nozzle geometry, fuel injection pressure, and in-cylinder flows impact engine performance and emissions. This understanding has two main benefits. First, it provides combustion system designers with a mental picture to guide their thinking as they attempt to solve problems or optimize certain aspects of the combustion process. Second, it supports the development of more accurate simulation tools that can be used to iteratively optimize engines when human engineering judgment and intuition is insufficient.

The development of accurate simulation tools is particularly important. As we believe has been made clear in this chapter, the multiple interactions between design variables and chemical and physical processes makes the achievement of an optimal design a formidable task. Moreover, the basic design cannot be divorced from the calibration process—that is, the selection of specific settings for the swirl ratio, injection pressure, EGR rate, intake pressure and temperature, injection schedule, and so on, that is employed at each speed, load, and unique set of ambient conditions. An example of this interaction is the design of a bowl geometry to enhance late-cycle mixing. Such a design is clearly dependent on fuel injection parameters that vary from one operating condition to the next, such as injection pressure, injection timing, and the use of pilot and post injections. Ultimately, an optimal design will be a compromise based on a wide range of anticipated operating conditions.

Although significant progress has been made, it is apparent that for many aspects of the design process, both our understanding and the accuracy or speed of current simulation tools is not yet sufficient. Nevertheless, the substantial advantages that diesel engines offer with regard to efficiency, and their certain widespread use for decades to come, provide strong motivation to continue this task.

ACKNOWLEDGMENTS

This work was performed at the Department of Energy Sciences of Lund University in Lund, Sweden, and at the Combustion Research Facility, Sandia National Laboratories in Livermore, California, with support provided by the Swedish Energy Agency, the United States Department of Energy (Office of Vehicle Technologies), and by General Motors Corporation (agreement # FI083070326). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. The authors thank S. Busch, D. Kim, J-P Fayolle, H. Persson, C. Chartier, and R. Winsor for their helpful comments on the draft manuscript.

REFERENCES

Abe, T., Nagahiro, K., Aoki, T. *et al*. (2004) Development of new 2.2-liter turbocharged diesel engine for the EURO-IV standards. SAE technical paper 2004-01-1316.

Akihama, K., Kosaka, H., Hotta, Y., *et al*. (2008) An investigation of high load (compression ignition) operation of the "Naphtha Engine" - a combustion strategy for low well-to-wheel CO_2 emissions. *SAE International Journal of Fuels and Lubricants*, **1**(1), 920–932.

Andersson, Ö. (2010) Diesel combustion in *Handbook on Combustion*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

Andersson, Ö., Somhorst, J., Lindgren, R. *et al*. (2009) Development of the Euro 5 combustion system for volvo cars' 2.4.I diesel engine. SAE technical paper 2009-01-1450.

Arcoumanis, C., Hadjiapostolou, A. and Whitelaw, J.H. (1991) Flow and combustion in a hydra direct-injection Diesel Engine. SAE technical paper 910177.

Aronsson, U., Andersson, Ö., Egnell, R. *et al*. (2009) Influence of spray-target and squish height on sources of CO and UHC in a HSDI diesel engine during PPCI low-temperature combustion. SAE technical paper 2009-01-2810.

ASTM International. (2008a) Standard Test Method for Cetane Number of Diesel Fuel Oil. West Conshohocken, PA ASTM International.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ASTM International. (2008b) Standard Test Method for Determination of Ignition Delay and Derived Cetane Number (DCN) of Diesel Fuel Oils by Combustion in a Constant Volume Chamber. *D6890-08*. West Conshohocken, PA ASTM International.

Badami, M., Mallamo, F., Millo, F., *et al*. (2003) Experimental Investigation on the effect of multiple injection strategies on emissions, noise and brake specific fuel consumption of an automotive direct injection common-rail diesel engine. *International Journal of Engine Research*, **4**(4), 299–314.

Badami, M., Millo, F. and D'amato, D.D. (2001) Experimental investigation on soot and NOx formation in a DI common rail diesel engine with pilot injection. SAE technical paper 2001-01-0657.

Barro, C., Tschanz, F., Obrecht, P. *et al*. (2012) Influence of Post-Injection parameters on Soot Formation and Oxidation in a Common-Rail Diesel Engine Using Multi-Color Pyrometry. *ASME Internal Combustion Engine Division Fall Technical Conference*, Vancouver, BC, Canada: Paper ICEF2012-92075.

Bauder, R., Fröhlich, A., and Rossi, D. (2010) The new generation of the Audi 3.0 L V6 TDI engine. *Motortechnische Zeitschrift Worldwide*, **71**(10), 20–27.

Bauder, R., Gruber, M., Michels, E., *et al*. (2005) The new Audi 4.2.L V8 TDI-engine. Part 2: thermodynamics, application, and exhaust after-treatment. *Motortechnische Zeitschrift*, **66**(11), 898–908.

Bauder, R. and Stock, D. (1990) The new Audi 5-cylinder turbo diesel engine: the first passenger car diesel engine with second generation direct injection. SAE technical paper 900648.

Béard, P., Mokaddem, K. and Baritaud, T. (1998) Measurement and modeling of the flow-field in a DI diesel engine: effects of piston bowl shape and engine speed. SAE technical paper 982587.

Beatrice, C., Belardini, P., Bertoli, C. *et al*. (2003) Downsizing of common rail D.I. engines: influence of different injection strategies on combustion evolution. SAE technical paper 2003-01-1784.

Beer, A. (ed.) (1994) *Diesel Fuel Injection*, Robert Bosch GmbH, Stuttgart.

Cao, L., Bhave, A., Su, H., *et al*. (2009) Influence of injection timing and piston bowl geometry on PCCI combustion and emissions. *SAE International Journal of Engines*, **2**(1), 1019–1033.

Carlucci, P., Ficarella, A. and Laforgia, D. (2003) Effects of pilot injection parameters on combustion for common rail diesel engines. SAE technical paper 2003-01-0700.

Carlucci, P., Ficarella, A., and Laforgia, D. (2005) Effects on combustion and emissions of early and pilot fuel injections in diesel engines. *International Journal of Engine Research*, **6**(1), 43–60.

Catania, A.E., D'ambrosio, S., Finesso, R., *et al*. (2009) Combustion system optimization of a low compression-ratio PCCI diesel engine for light-duty application. *SAE International Journal of Engines*, **2**(1), 1314–1326.

Chartier, C., Aronsson, U., Andersson, Ö., *et al*. (2009a) Effect of injection strategy on cold start performance in an optical light-duty DI diesel engine. *SAE International Journal of Engines*, **2**(2), 431–442.

Chartier, C., Aronsson, U., Andersson, Ö. *et al*. (2009b) Analysis of smokeless spray combustion in a heavy-duty diesel engine by combined simultaneous optical diagnostics. SAE technical paper 2009-01-1353.

Chartier, C., Aronsson, U., Andersson, Ö., *et al*. (2013) Influence of jet–jet interactions on the lift-off length in an optical heavy-duty DI diesel engine. *Fuel*, **112**(0), 311–318.

Chen, S.K. (2000) Simultaneous reduction of NOx and particulate emissions by using multiple injections in a small diesel engine. SAE technical paper 2000-01-3084.

Chi, Y., Park, S., Lee, K., *et al*. (2008) New V6 3.0 L diesel engine for Hyundai/Kia's SUVs. *Motortechnische Zeitschrift Worldwide*, **69**(11), 24–30.

Cipolla, G., Vassallo, A., Catania, A.E. *et al*. (2007) Combined application of CFD modeling and pressure-based combustion diagnostics for the development of a low compression ratio high-performance diesel engine. SAE technical paper 2007-24-0034.

Crabb, D., Fleiss, M., Larsson, J.-E., *et al*. (2013) New modular engine platform from Volvo. *Motortechnische Zeitschrift Worldwide*, **74**(9), 4–11.

Crosse, J. (2010) Going clean off-highway. *Ricardo Quarterly Review*, **Q2**, 16–21.

Cursente, V., Pacaud, P., and Gatellier, B. (2008a) Reduction of the compression ratio on a HSDI diesel engine: combustion design evolution for compliance the future emission standards. *SAE International Journal of Fuels and Lubricants*, **1**(1), 420–439.

Cursente, V., Pacaud, P., Mendez, S., *et al*. (2008b) System approach for compliance with full load targets on a wall guided diesel combustion system. *SAE International Journal of Engines*, **1**(1), 501–513.

Dec, J.E. (1997) A conceptual model of DI diesel combustion based on laser-sheet imaging. SAE technical paper 970873.

Dec, J.E. (2003) Diesel-fueled HCCI engines in *Homogeneous Charge Compression Ignition (HCCI) Engines: Key Research and Development Issues* (eds F. Zhao, T.W. Asmus, D.N. Assanis, J.E. Dec, J.A. eng, P.M. Najt), Warrendate, PA, Society of Automotive Engineers.

Dembinski, H.W.R. and Angstrom, H.-E. (2012) Optical study of swirl during combustion in a CI engine with different injection pressures and swirl ratios compared with calculations. SAE technical paper 2012-01-0682.

Denbratt, I. (2010) Advanced concepts for future light-duty diesel engines in *Advanced Direct Injection Combustion Engine Technologies and Development. Volume 2: Diesel Engines* (ed. H. Zhao), Woodhead Publishing, Cambridge.

Diwakar, R. and Singh, S. (2009) Importance of spray-bowl interaction in a DI diesel engine operating under PCCI combustion mode. SAE technical paper 2009-01-0711.

Dolak, J.G., Shi, Y. and Reitz, R. (2010) A computational investigation of stepped-bowl piston geometry for a light duty engine operating at low load. SAE technical paper 2010-01-1263.

Dreisbach, R., Graf, G., Kreuzig, G. *et al*. (2007) HD base engine development to meet future emission and power density challenges of a DDI™ engine. SAE technical paper 2007-01-4225.

Dronniou, N., Lejeune, M., Balloul, I. *et al*. (2005) Combination of high EGR rates and multiple injection strategies to reduce pollutant emissions. SAE technical paper 2005-01-3726.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Dworschak, J., Neuhauser, W., Rechberger, E., *et al*. (2009) The new BMW six-cylinder diesel engine. *Motortechnische Zeitschrift Worldwide*, **70**(2), 4–10.

Ehleskog, R., Ochoterena, R.L. and Andersson, S. (2007) Effects of multiple injections on engine-out emission levels including particulate mass from an HSDI diesel engine. SAE technical paper 2007-01-0910.

Eidenböck, T., Mayr, K., Neuhauser, W., *et al*. (2012) The new BMW six-cylinder diesel engine with three turbochargers. Part1: drive unit and turbocharger system. *Motortechnische Zeitschrift Wordwide*, **73**(10), 18–24.

Fasolo, B., Doisy, A.-M., Dupont, A. *et al*. (2005) Combustion system optimization of a new 2 liter diesel engine for EURO IV. SAE technical paper 2005-01-0652.

Filipi, Z.S. and Assanis, D.N. (2000) The effect of the stroke-to-bore ratio on combustion, heat transfer and efficiency of a homogeneous charge spark ignition engine of a given displacement. *International Journal of Engine Research*, **1**(2), 191–208.

Flowers, D.L., Martinez-Frias, J. and Cleeves, J.M. (2010) Internal Combustion Engine with Optimal Bore-to-Stroke Ratio. US patent application 20100147269.

Flynn, P.F., Durrett, R.P., Hunter, G.L. *et al*. (1999) Diesel combustion: an integrated view combining laser diagnostics, chemical kinetics, and empirical validation. SAE technical paper 1999-01-0509.

Ge, H.-W., Shi, Y., Reitz, R. *et al*. (2010) Engine development using multi-dimensional CFD and computer optimization. SAE technical paper 2010-01-0360.

Genzale, C.L., Reitz, R.D. and Wickman, D.D. (2007) A computational investigation into the effects of spray targeting, bowl geometry and swirl ratio for low-temperature combustion in a heavy-duty diesel engine. SAE technical paper 2007-01-0119.

Golovitchev, V.I., Montorsi, L., Denbratt, I. *et al*. (2007) Numerical evaluation of direct injection of urea as NOx reduction method for heavy duty diesel engines. SAE technical paper 2007-01-0909.

Hadler, J., Rudolph, F., Engler, H.-J., *et al*. (2007) The new 2.0-L-4V-TDI engine with common rail: modern diesel technology from Volkswagen. *Motortechnische Zeitschrift*, **68**(11), 914–923.

Hardy, J.-P., Lahjaily, H., Besson, M. *et al*. (2004) Diesel combustion optimization at full load by combined CFD and single cylinder tests. SAE technical paper 2004-01-1402.

Hasse, C. and Peters, N. (2005) Modelling of ignition mechanisms and pollutant formation in direct-injection diesel engines with multiple injections. *International Journal of Engine Research*, **6**(3), 231–246.

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*, McGraw-Hill, Inc, New York.

Horrocks, R.W. (2010) Overview of high-speed direct injection engines in *Advanced Direct Injection Combustion Engine Technologies and Development. Volume 2: Diesel Engines* (ed. H. Zhao), Woodhead Publishing.

Hotta, Y., Inayoshi, M., Nakakita, K. *et al.* (2005) Achieving lower exhaust emissions and better performance in an HSDI diesel engine with multiple injection. SAE technical paper 2005-01-0928.

Hotta, Y., Nakakita, K., Fuyuto, T., *et al*. (2002a) Smoke reduction methods using shallow-dish combustion chamber in an HSDI common-rail diesel engine. *R&S Review of Toyota CRDL*, **37**(3), 17–24.

Hotta, Y., Nakakita, K., Fuyuto, T. *et al*. (2002b) Cause of exhaust smoke and its reduction methods in an HSDI diesel engine under high-speed and high-load conditions. SAE technical paper 2002-01-1160.

Idicheria, C.A. and Pickett, L.M. (2005) Soot formation in diesel combustion under high-EGR conditions. SAE technical paper 2005-01-3834.

Ikegami, M., Fukuda, M., Yoshihara, Y. *et al*. (1990a) Combustion chamber shape and pressurized injection in high-speed direct-injection diesel engines. SAE technical paper 900440.

Ikegami, M., Hida, M., Yamane, K., *et al*. (1990b) Influence of top clearance on combustion in direct-injection diesel engines. *JSAE Review*, **11**(3), 10–15.

Inagaki, K., Mizuta, J., Fuyuto, T., *et al*. (2011) Low emissions and high-efficiency diesel combustion using highly dispersed spray with restricted in-cylinder swirl and squish flows. *SAE International Journal of Engines*, **4**(1), 2065–2079.

Ishida, M., Chen, Z.-L., Luo, G.-F. *et al*. (1994) The effect of pilot injection on combustion in a turbocharged D. I. diesel engine. SAE technical paper 941692.

Juttu, S., Thipse, S.S., Marathe, N.V. *et al*. (2009) CFD study of combustion chambers for lower engine exhaust emissions from diesel engines operated in HCCI and conventional diesel mode. SAE technical paper 2009-26-027.

Kamimoto, T. and Bae, M.-H. (1988) High combustion temperature for the reduction of particulate in diesel engines. SAE technical paper 880423.

Kanda, T., Kobayashi, S., Matsui, R., Sono, H. (2004) Study on Euro IV Combustion Technologies for Direct Injection Diesel Engine. SAE technical paper 2004-01-0113.

Kastner, O., Atzler, F., Müller, A. *et al*. (2006) Multiple injection strategies and their effect on pollutant emission in passenger car diesel engines. THIESEL 2006, pp. 61–75, Valencia, Spain.

Kastner, O., Atzler, F., Rotondi, R. *et al*. (2008) Evaluation of Injection Strategies for passenger car diesel engines to meet Euro 6 legislation limits. THIESEL 2008, pp. 27–42, Valencia, Spain.

Kidoguchi, Y., Sanda, M., and Miwa, K. (2003) Experimental and theoretical optimization of combustion chamber and fuel distribution for the low-emission direct-injection diesel engine. *Journal of Engineering for Gas Turbines and Power*, **125**, 351–357.

Kidoguchi, Y., Yang, C. and Miwa, K. (1999) Effect of high squish combustion chamber on simultaneous reduction of NOx and particulate from a direct-injection diesel engine. SAE technical paper 1999-01-1502.

Kihara, R., Mikami, Y. and Kinbara, M. (1983) The advantages of the Isuzu square combustion chamber for D.I. engines. SAE technical paper 830372.

Kimura, S., Aoki, O., Kitahara, Y. *et al*. (2001) Ultra-clean combustion technology combining a low-temperature and premixed combustion concept for meeting future emission standards. SAE technical paper 2001-01-0200.

Kimura, S., Aoki, O., Ogawa, H. *et al*. (1999) New combustion concept for ultra-clean and high-efficiency small DI diesel engines. SAE technical paper 1999-01-3681.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Kitamura, T. and Ito, T. (2010) Mixing-controlled, low temperature diesel combustion with pressure modulated multiple-injection for HSDI diesel engine. *SAE International Journal of Engines*, **3**(1), 461–478.

Kitamura, T., Ito, T., Senda, J., *et al*. (2002) Mechanism of smokeless diesel combustion with oxygenated fuels based on the dependency of the equivalence ratio and temperature on soot particle formation. *International Journal of Engine Research*, **3**(4), 223–247.

Kook, S., Bae, C., Miles, P.C. *et al*. (2006) The effect of swirl ratio and fuel injection parameters on CO emission and fuel conversion efficiency for high-dilution, low-temperature combustion in an automotive diesel engine. SAE technical paper 2006-01-0197.

Kook, S., Bae, C., Miles, P.C. *et al*. (2005) The influence of charge dilution and injection timing on low-temperature diesel combustion and emissions. SAE technical paper 2005-01-3837.

Kurtz, E.M. and Styron, J. (2012) An assessment of two piston bowl concepts in a medium-duty diesel engine. *SAE International Journal of Engines*, **5**(2), 344–352.

Langen, P., Hall, W., Nefischer, P., *et al*. (2010) The new two-stage turbocharged six-cylinder diesel engine of the BMW 740D. *Motortechnische Zeitschrift Worldwide*, **71**(4), 4–11.

Lawrence Livermore National Laboratory n-Heptane, Detailed Mechanism, Version 2 [Online], https://www-pls.llnl.gov/?url=science_and_technology-chemistry-combustion-nc7h16 (accessed 26 January 2013).

Lee, E., Kwak, S., Kim, M., *et al*. (2009) The new 2.0 L and 2.2 L four-cylinder diesel engine family of Hyundai-Kia. *Motortechnische Zeitschrift Worldwide*, **70**(10), 14–19.

Lee, K.W., Jang, K.I., Lee, J.J., *et al*. (2012) The new Hyundia/Kia 1.1-L three-cylinder diesel engine. *Motortechnische Zeitschrift Worldwide*, **73**(9), 16–21.

Lippert, A.M., Stanton, D.W., Reitz, R.D. *et al*. (2000) Investigating the effect of spray targeting and impingement on diesel engine cold start. SAE technical paper 2000-01-0269.

Lisbona, M.G., Olmo, L. and Rindone, G. (2000) Analysis of the Effect of Combustion Bowl Geometry of a DI Diesel Engine on Efficiency and Emissions. THIESEL 2000 Thermo- and Fluid Dynamic Processes in Diesel Engines, pp. 279–293, Valencia, Spain.

Majewski, W.A. (2002) Diesel Particulate Matter: DieselNet Technology Guide [Online], http://www.dieselnet.com/tech/dpm.html (accessed 7 January 2014).

Matsui, R., Shimoyama, K., Nonaka, S. *et al*. (2008) Development of high-performance diesel engine compliant with Euro-V. SAE technical paper 2008-01-1198.

Middlemiss, I.D. (1978) Characteristics of the Perkins 'Squish Lip' direct injection combustion system. SAE technical paper 780113.

Miles, P.C. (2008) Turbulent flow structure in direct-injection, swirl-supported diesel engines in *Flow and Combustion in Reciprocating Engines* (eds C. Arcoumanis and T. Kamimoto), Springer-Verlag, Berlin Heidelberg.

Miles, P.C., Choi, D., Kook, S. *et al*. (2005) The influence of flow structures and mixing on low-temperature diesel combustion. *5th Intl. Symp. on Towards Clean Diesel Engines*, June 2–3, Lund, Sweden.

Miles, P.C., Petersen, B.R. and Sahoo, D. (2012) The impact of injection timing on mixture preparation and chemical kinetics in low-temperature diesel combustion. *Eighth International Conference on Modeling and Diagnostics for Advanced Engine Systems – COMODIA 2012*, July 23–26, Fukuoka, Japan.

Miles, P.C., Rempelewert, B.H. and Reitz, R.D. (2003) Squish-swirl and injection-swirl interaction in direct-injection diesel engines. *ICE 2003: 6th International Conference on Engines for Automobiles*, Sept. 14–19, Capri, Naples, Italy.

Miller, J.A. and Bowman, C.T. (1989) Mechanism and modeling of nitrogen chemistry in combustion. *Progress in Energy and Combustion Science*, **15**(4), 287–338.

Montajir, R.M., Tsunemoto, H., Ishitani, H. *et al*. (2001) A new combustion chamber concept for low emissions in small DI diesel engines. SAE technical paper 2001-01-3263.

Musculus, M.P.B. and Kattke, K. (2009) Entrainment waves in diesel jets. *SAE International Journal of Engines*, **2**(1), 1170–1193.

Musculus, M.P.B., Miles, P.C., and Pickett, L.M. (2013) Conceptual models for partially premixed low-temperature diesel combustion. *Progress in Energy and Combustion Science*, **39**(2–3), 246–283.

Naber, J.D. and Siebers, D.L. (1996) Effects of gas density and vaporization on penetration and dispersion of diesel sprays. SAE technical paper 960034.

National Research Council (2011) *Assessment of Fuel Economy Technologies for Light-Duty Vehicles*, National Academies Press, Washington, D.C..

Neal, N. and Rothamer, D. (2012) An optical study of the impact of swirl ratio on extended lift-off diesel combustion. *Spring Technical Meeting of the Central States Section of the Combustion Institute*, April 22–24.

O'Connor, J. and Musculus, M.P.B. (2013) Post injections for soot reduction in diesel engines: a review of current understanding. SAE technical paper 2013-01-0917.

Oinuma, R., Takuma, S., Koyano, T. *et al*. (2005) Effects of post injection on piston lubrication in a common rail small bore diesel engine. SAE technical paper 2005-01-2166.

Park, C., Kook, S. and Bae, C. (2004) Effects of multiple injections in a HSDI diesel engine equipped with common rail injection system. SAE technical paper 2004-01-0127.

Peng, R.D., Bell, M.L., Geyh, A.S., *et al*. (2009) Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution. *Environmental Health Perspectives*, **117**(6), 957–963.

Petersen, B., Miles, P., and Sahoo, D. (2012) Equivalence ratio distributions in a light-duty diesel engine operating under partially premixed conditions. *SAE International Journal of Engines*, **5**(2), 526–537.

Pickett, L., Kook, S. and Williams, T. (2009) Transient liquid penetration of early-injection diesel sprays. SAE technical paper 2009-01-0839.

Pickett, L.M., Caton, J.A., Musculus, M.P.B., *et al*. (2006) Evaluation of the equivalence ratio-temperature region of diesel soot precursor formation using a two-stage Lagrangian model. *International Journal of Engine Research*, **7**(5), 349–370.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Pickett, L.M. and Siebers, D.L. (2004) Non-sooting, low flame temperature mixing-controlled DI diesel combustion. SAE technical paper 2004-01-1399.

Pickett, L.M., Siebers, D.L. and Idicheria, C.A. (2005) Relationship between ignition processes and the lift-off length of diesel fuel jets. SAE technical paper 2005-01-3843.

The European Parliament and the Council of the European Union, Official Journal of the European Union, Issue 171/1 (2007).

Ricaud, J.C. and Lavoisier, F. (2002) Optimizing the Multiple Injection Settings on an HSDI Diesel Engine. THIESEL 2002 Thermo- and Fluid Dynamic Processes in Diesel Engines, pp. 251–275, Valencia, Spain.

Sahoo, D., Petersen, B.R. and Miles, P.C. (2012) The impact of swirl ratio and injection pressure on fuel-air mixing in a light-duty diesel engine. *ASME 2012 IC Engine Division Spring Technical Conference*, May 6–9, 2012, Torino, Italy.

Saito, T., Daisho, Y., Uchida, N. *et al*. (1986) Effects of combustion chamber geometry on diesel combustion. SAE technical paper 861186.

Sakono, T., Nakai, E., Kataoka, M. *et al*. (2011) MAZDA SKYACTIV-D 2.2L diesel engine. 20th Aachen Colloquium on Automobile and Engine Technology, pp. 943–965, Aachen, Germany.

Saurer, H. (1934) Improvements in and Relating to Internal Combustion Engines of the Liquid Fuel Injection Type. Germany patent application.

Shimo, D., Kataoka, M. and Fujimoto, H. (2004) Effect of cooling of burned gas by vertical vortex on NOx reduction in small DI diesel engines. SAE technical paper 2004-01-0125.

Siebers, D.L. (1999) Scaling Liquid-Phase Fuel Penetration in Diesel Sprays Based on Mixing-Limited Vaporization. SAE technical paper 1999-01-0528.

Siebers, D.L. (2008) Recent developments on diesel fuel jets under quiescent conditions in *Flow and Combustion in Reciprocating Engines* (eds C. Arcoumanis and T. Kamimoto), Springer-Verlag Berlin Heidelberg.

Siewert, R.M. (1978) Engine combustion at large bore-to-stroke ratios. SAE technical paper 780968.

Siewert, R.M. (2007) Spray angle and rail pressure study for low NOx diesel combustion. SAE technical paper 2007-01-0122.

Sjöberg, M. and Dec, J.E. (2005) An investigation into lowest acceptable combustion temperatures for hydrocarbon fuels in HCCI engines. *Proceedings of the Combustion Institute*, **30**(2), 2719–2726.

Steinparzer, F., Mattes, W., Nefischer, P., *et al*. (2007) The new BMW four-cylinder diesel engine, part 2: function and vehicle results. *Motortechnische Zeitschrift*, **68**(12), 932–943.

Stone, C.R. and Ladommatos, N. (1992) The measurement and analysis of swirl in steady flow. SAE technical paper 921642.

Styron, J., Baldwin, B., Fulton, B. *et al*. (2011) Ford 2011 6.7L power stroke® diesel engine combustion system development. SAE technical paper 2011-01-0415.

Tanaka, T. (2011) Toyota's strategy for powertrains including plug-in hybrid vehicles. *2011 SAE/JSAE Powertrain, Fuels and Lubricants Meeting, Keynote Presentation*. Aug. 30-Sept. 2, Kyoto, Japan.

Tang, J., Pischinger, S., Lamping, M., *et al*. (2009) Coking phenomena in nozzle orifices of Dl-diesel engines. *SAE International Journal of Fuels and Lubricants*, **2**(1), 259–272.

Terazawa, Y., Nakai, E., Kataoka, M., *et al*. (2011) The new Mazda four-cylinder diesel engine. *Motortechnische Zeitschrift Worldwide*, **72**(9), 26–32.

Timoney, D.J. (1987) Effects of important variables on measured heat release rates in a D.I. diesel. SAE technical paper 870271.

Tomishima, H., Matsumoto, T., Oki, M. *et al*. (2008) The advanced diesel common rail system for achieving a good balance between ecology and economy. SAE technical paper 2008-28-0017.

Tree, D.R. and Svensson, K.I. (2007) Soot processes in compression ignition engines. *Progress in Energy and Combustion Science*, **33**, 272–309.

Tritton, D.J. (1977) *Physical fluid Dynamics*, Van Nostrand Reinhold Co, Berkshire, England, (UK).

UNEP/WMO (2011) Integrated Assessment of Black Carbon and Tropospheric Ozone. Summary for Decision Makers. DEW/1351/NA.

US Energy Information Administration (2013) International Energy Outlook 2013. Washington, DC, http://www.eia.gov/forecasts/ieo/

Van Den Huevel, B., Willems, W., Krämer, F., *et al*. (2006) Combustion system development for the new diesel engines in light and medium commercial vehicles from ford and PSA. *Motortechnische Zeitschrift*, **67**(9), 606–614.

Vanegas, A., Won, H., Felsch, C. *et al*. (2008) Experimental investigation of the effect of multiple injections on pollutant formation in a common-rail DI diesel engine. SAE technical paper 2008-01-1191.

Walter, B. and Gatellier, B. (2002) Development of the high power NADI™ concept using dual mode diesel combustion to achieve zero NOx and particulate emissions. SAE technical paper 2002-01-1744.

Wickman, D.D., Senecal, P.K. and Reitz, R.D. (2001) Diesel engine combustion chamber geometry optimization using genetic algorithms and multi-dimensional spray and combustion modeling. SAE technical paper 2001-01-0547.

Yamada, T., Haga, H., Matsumoto, I., *et al*. (2011) Study of diesel engine system for hybrid vehicles. *SAE International Journal of Alternative Powertrains*, **1**(2), 560–565.

Yoo, D., Kim, D., Jung, W. *et al*. (2013) Optimization of diesel combustion system for reducing PM to meet tier4-final emission regulation without diesel particulate filter. SAE technical paper 2013-01-2538.

Yun, H., Sellnau, M., Milovanovic, N. *et al*. (2008) Development of premixed low-temperature diesel combustion in a HSDI diesel engine. SAE technical paper 2008-01-0639.

Zhang, L., Ueda, T., Takatsuki, T. *et al*. (1995) A study of the effects of chamber geometries on flame behavior in a DI diesel engine. SAE technical paper 952515.

Zhu, Y., Zhao, H., Melas, D.A. *et al*. (2004) Computational study of the effects of the re-entrant lip shape and toroidal radii of piston bowl on a HSDI diesel engine's performance and emissions. SAE technical paper 2004-01-0118.

Zolver, M., Griard, C. and Henriot, S. (1997) 3D modeling applied to the development of a DI diesel engine: effect of piston bowl shape. SAE technical paper 971599.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto120
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Pressure and Heat Release Analysis

Jaal B. Ghandhi
University of Wisconsin-Madison, Madison, WI, USA

1 Introduction 1
2 Multizone Models 2
3 Single-Zone Models 4
4 Approximate Methods 7
5 Comparison of Calculation Methods 8
6 Summary 12
Appendix A—Combustion Efficiency Calculation 13
References 13

1 INTRODUCTION

The pressure in the cylinder of an internal combustion engine determines the net work transfer to the crankshaft and is, thus, the critical parameter that defines an engine's overall performance. The cylinder pressure, however, is affected by several processes including changes in cylinder volume, heat transfer to the combustion chamber surfaces, mass loss from the combustion chamber due to blowby, mass addition due to fuel injection, and the chemical processes occurring within the gases of combustion chamber. It is this latter quantity that represents the combustion performance of the engine and the thermodynamic analysis used to isolate it is referred to as *heat release analysis*.

The application of thermodynamic principles to the cylinder contents requires a set of simplifying assumptions that will have a strong bearing on the calculated quantities. It is, therefore, important to always keep these assumptions, and their validity, in mind while assessing the results. For example, for diesel engine combustion, where large property gradients are known to exist, the contents of the cylinder are often assumed to be uniform and in thermodynamic equilibrium to allow a tractable analysis. The discrepancies between the model assumptions and reality are important, but very difficult to quantitatively put bounds on. The individual must, therefore, use appropriate caution in their interpretation of the results, especially when comparisons are being made in absolute terms across different engine platforms. One goal of this chapter is to expose and highlight the model assumptions to prepare end users for the proper interpretation of the results.

This chapter is structured based on the fundamental assumption of how the combustion chamber gases are treated in the thermodynamic analysis. Multizone models, where the chamber contents are segregated into (typically two) zones, are discussed first. These models identify the burning rate as the rate at which mass is transferred from the reactant zone to the product zone. Single-zone models that treat the cylinder contents as homogeneous in composition and temperature and are easier to apply than multizone models are discussed second. Finally, approximate methods are discussed. The multi- and single-zone approaches attempt to use a rigorous thermodynamic treatment of the gas properties and, within their assumptions, are quantitative. The approximate methods rely on polytropic fits to the data and are thus not considered to provide absolute quantities.

The topic of heat release calculation has received extensive discussion in the past literature, for example, Krieger and Borman (1966), Gatowski *et al*. (1984), and Foster (1985). This work builds on that knowledge, but does not attempt to comprehensively review it.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 MULTIZONE MODELS

In principle, the combustion chamber gases can be assigned to N zones of uniform composition and temperature, with each zone considered to be in thermodynamic equilibrium. In practice, the utility of this physical model is that it approximates the spark-ignition engine well, where the flame separates the burned gas at high temperature with equilibrium products from the unburned gas at the reactant composition. Thus, the discussion herein is restricted to a two-zone model comprised of burned and unburned zones, identified by subscripts b and u, respectively.

The First Law applied separately to each zone, with a rate of mass transfer between the two zones that represents the rate of conversion of reactants to products, is given by

$$\frac{\mathrm{d}(m_\mathrm{b}u_\mathrm{b})}{\mathrm{d}t} = \dot{Q}_\mathrm{w,b} - p\frac{\mathrm{d}V_\mathrm{b}}{\mathrm{d}t} + h_\mathrm{b}\frac{\mathrm{d}m_\mathrm{b}}{\mathrm{d}t} \tag{1}$$

$$\frac{\mathrm{d}(m_\mathrm{u}u_\mathrm{u})}{\mathrm{d}t} = \dot{Q}_\mathrm{w,u} - p\frac{\mathrm{d}V_\mathrm{u}}{\mathrm{d}t} + h_\mathrm{b}\frac{\mathrm{d}m_\mathrm{u}}{\mathrm{d}t} \tag{2}$$

where m is the mass, u is the internal energy, $\dot{Q}_\mathrm{w}$ is the wall heat transfer to the system, p is the cylinder pressure, V is the volume, and we have used the fact that the flame is occurring at constant pressure so that $h_\mathrm{u} = h_\mathrm{b}$, where h is the specific enthalpy. Equations 1 and 2 are written in terms of temporal derivatives but can be converted to crank angle-based derivatives using the rotational speed of the engine. The solution of Equations 1 and 2 for the mass burning rate is subjected to constraints on the total volume and mass

$$V_\mathrm{u} + V_\mathrm{b} = V \tag{3}$$

$$m_\mathrm{u} + m_\mathrm{b} = m \tag{4}$$

where the unsubscripted quantities represent the total cylinder conditions. It should be noted that in the development of Equations 1 and 2 it has been assumed that there is no heat transfer between the burned and unburned zones and that there is no mass loss to crevice volumes. The former assumption is justified by the treatment of the flame, but the latter assumption is necessitated because of the difficulty of assigning a proportion to each zone.

By measuring the cylinder pressure and knowing the kinematic relationship for total cylinder volume, the solution of Equations 1–4 is possible provided that thermodynamic property data are available for u and h and that a valid correlation for the heat transfer rate is available.

It is important to note that there is no identified "heat release" in this analysis. In fact, the heat release comes about by the chemical conversion of reactants to products. Thus, the internal energy and enthalpy must be represented as their absolute values, that is, the internal energy of species i is $u_i = \Delta u^\circ_{\mathrm{f},i} + u^\mathrm{s}_i$ where $\Delta u^\circ_\mathrm{f}$ is the internal energy of formation and $u^\mathrm{s} = \int c_v \mathrm{d}T$ is the sensible internal energy with constant-volume specific heat c_v and temperature T. The heat release comes about due to the conversion of chemical to sensible energy. The evaluation of the thermodynamic properties also requires an equation of state for the determination of the temperature. The ideal gas equation of state is used, and its applicability will be assessed later.

Each zone is considered to be in thermodynamic equilibrium. For the unburned zone, the composition is frozen at the reactant state, but the internal energy changes with the changes in temperature (consistent with the ideal gas assumption) and $\frac{\mathrm{d}u_\mathrm{u}}{\mathrm{d}T} = \frac{\mathrm{d}u^\mathrm{s}_\mathrm{u}}{\mathrm{d}T} = c_\mathrm{v,u}$. The burned gas composition, however, varies with time; the composition follows a shifting equilibrium. The equilibrium state is affected by both temperature and pressure, and, therefore, $u = u(p, T)$ in spite of the fact that the gas is considered an ideal gas. Further, the changes in composition affect the total amount of chemical energy, so $\frac{\mathrm{d}u_\mathrm{b}}{\mathrm{d}T} \neq \frac{\mathrm{d}u^\mathrm{s}_\mathrm{b}}{\mathrm{d}T}$ and $\frac{\mathrm{d}u^\mathrm{s}_\mathrm{b}}{\mathrm{d}T} \neq c_\mathrm{v,b}$ as both neglect the change in composition and the former neglects the change in chemical energy.

An assessment of the validity of the ideal gas assumption and the necessity of accounting for the shifting equilibrium of the product composition is shown in Figure 1, which shows a pressure–temperature map with contours of the compressibility factor Z (solid) and $\alpha \equiv \frac{\mathrm{d}h_\mathrm{b}}{\mathrm{d}T}/c_\mathrm{p;b}$ (dashed)—the meaning of α is discussed later and $c_\mathrm{p,b}$ is the burned gas constant-pressure specific heat. The data points superimposed on the plot arise from a two-zone model prediction of the unburned (open symbols) and burned (filled symbols) states for a low load (squares) and high load, pressure-charged (circles) condition. The compressibility factor was calculated for a mixture of ideal stoichiometric combustion products (CO_2, H_2O, N_2) using the Peng–Robinson equation of state (Klein and Nellis, 2012). It can be seen in Figure 1 that, as expected, the compressibility factor deviates from unity the most at low temperatures and elevated pressures. As can be seen from the two limiting engine cases, the largest deviation from ideal gas behavior is $Z = 1.03$, and, therefore, the use of the ideal gas equation of state for this global analysis is justified.

The parameter α describes the importance of the shifting equilibrium composition. The numerator was calculated by finding the equilibrium composition and the corresponding absolute enthalpy at a grid of temperature and pressure, then numerically calculating the derivative with respect to temperature at fixed pressure; the denominator was directly evaluated as $\sum y_i c_{\mathrm{p},i}$. For $\alpha > 1$, the shifting composition

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

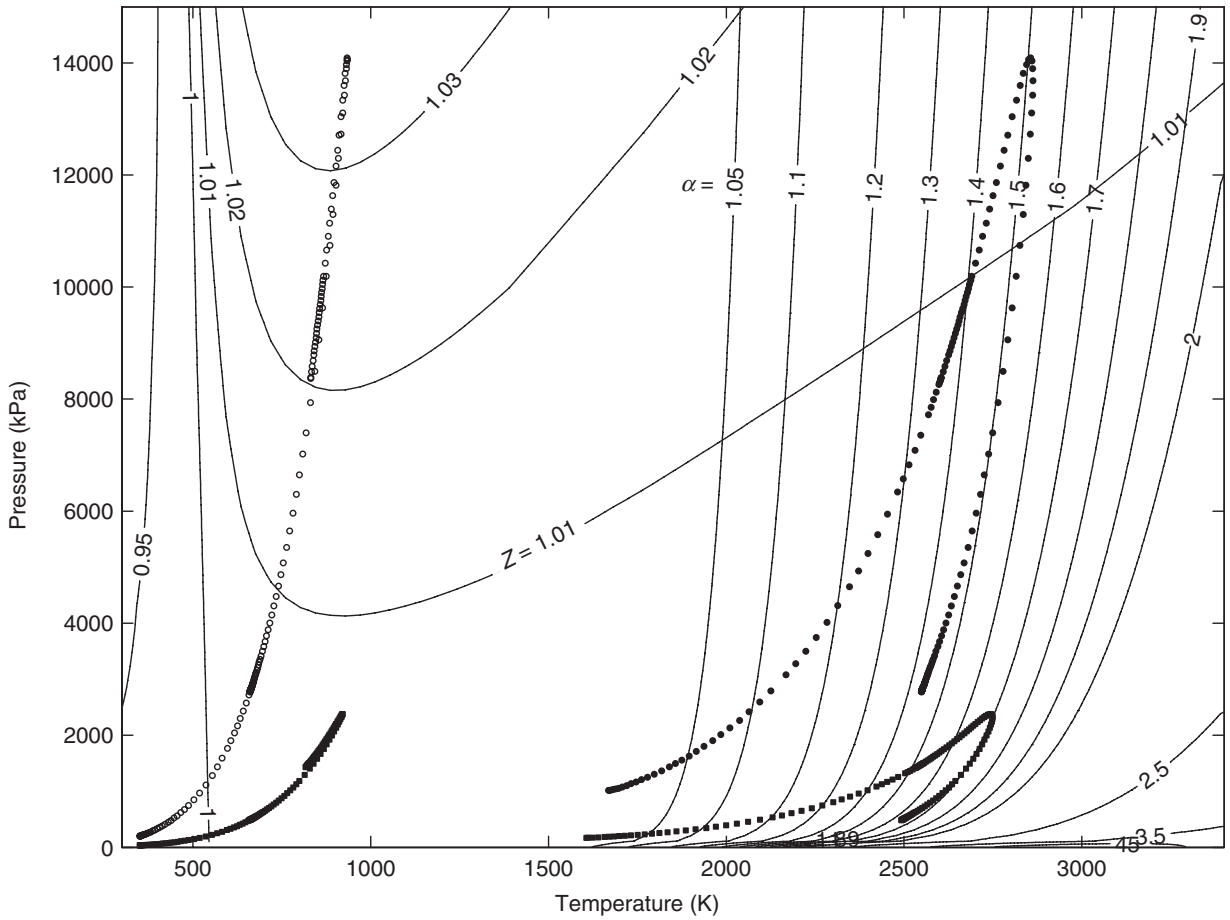


Figure 1. A comparison of the compressibility factor Z (solid) and the quantity α (dashed—see text for definition) as a function of pressure and temperature. Calculations are for a stoichiometric isooctane–air mixture. The data symbols show unburned (open) and burned (filled) states for a low load, stoichiometric condition (squares), and a high load, rich condition (circles).

of the system is significant and materially affects the relationship between enthalpy and temperature, that is, the ideal caloric gas assumption that $h = h(T)$ is insufficient. It can be seen in Figure 1 that for temperatures greater than $\sim$1800 K the magnitude of the iso-α contours are significantly larger than one. Further, the contours are seen to have significant positive curvature, which is an indication of the importance of pressure on the shifting equilibrium and highlights the need to include the pressure dependence in the thermodynamic property calculations. Finally, it can be seen that the burned engine states shown on the figure fall in the regions of $1 \leq \alpha \leq 1.7$, and thus, it is necessary to include the effect of the shifting equilibrium in the calculation of $\frac{\mathrm{d}u_\mathrm{b}}{\mathrm{d}T}$.

On the basis of this information, the necessary equations required to close the problem are the statement of the ideal gas law for the burned and unburned zones and the property relations. The ideal gas equations in differential form for the burned and unburned zones are

$$V_\mathrm{b}\frac{\mathrm{d}p}{\mathrm{d}t} + p\frac{\mathrm{d}V_\mathrm{b}}{\mathrm{d}t} - m_\mathrm{b}R_\mathrm{b}\frac{\mathrm{d}T_\mathrm{b}}{\mathrm{d}t} + \frac{m_\mathrm{b}R_\mathrm{b}T_\mathrm{b}}{W_\mathrm{b}}\frac{\mathrm{d}W_\mathrm{b}}{\mathrm{d}t} - R_\mathrm{b}T_\mathrm{b}\frac{\mathrm{d}m_\mathrm{b}}{\mathrm{d}t} = 0 \qquad (5)$$

$$V_\mathrm{u}\frac{\mathrm{d}p}{\mathrm{d}t} + p\frac{\mathrm{d}V_\mathrm{u}}{\mathrm{d}t} - m_\mathrm{u}R_\mathrm{u}\frac{\mathrm{d}T_\mathrm{u}}{\mathrm{d}t} + \frac{m_\mathrm{u}R_\mathrm{u}T_\mathrm{u}}{W_\mathrm{u}}\frac{\mathrm{d}W_\mathrm{u}}{\mathrm{d}t} - R_\mathrm{u}T_\mathrm{u}\frac{\mathrm{d}m_\mathrm{u}}{\mathrm{d}t} = 0 \qquad (6)$$

where W is the molar mass and R is the specific gas constant ($R \equiv R_\mathrm{u}/W$). It is important to note that the molar mass can change significantly between the reactants and products (discussed more fully later). As previously stated, the effect of pressure shifts the composition and thus $W = W(p,T)$, which gives

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

$$\frac{dW_b}{dt} = \frac{\partial W_b}{\partial T_b}\frac{dT_b}{dt} + \frac{\partial W_b}{\partial p}\frac{dp}{dt} \quad (7)$$

$$\frac{dW_u}{dt} = \frac{\partial W_u}{\partial T_u}\frac{dT_u}{dt} + \frac{\partial W_u}{\partial p}\frac{dp}{dt} \quad (8)$$

where the last equation is trivial for the present example as the composition of the unburned zone is fixed. Collectively, Equations 1–8, along with the measured pressure and volume, provide the relations necessary to calculate the temporal history of mass, temperature, volume, and molar mass of the burned and unburned zones.

The solution of these differential-algebraic equations can be achieved using existing software packages, and property data are available from detailed kinetic mechanisms (see Fundamental Chemical Kinetics). The result of the calculation procedure is the mass burning rate, $\dot{m}_b$, where the over dot represents the time rate of change, which can differ from an energy conversion rate. In the remainder of this chapter, three test cases will be followed. The test cases were computed using a two-zone model with a prescribed mass burning rate given by a Wiebe function (see Zero- and One-Dimensional Methodologies and Tools). Synthetic model data were used because they allow all of the parameters to be known to full precision as compared to using engine data directly for which there are experimental uncertainties in all of the parameters. The three test cases are shown in Table 1 and represent (1) a light-load spark-ignition condition; (2) a high load, heavily boosted spark-ignition condition; and (3) a light-load, lean, low temperature combustion condition. Cases 1 and 2 were shown in Figure 1.

Figure 2 shows a comparison between the mass burning rate and the energy conversion rate as a function of crank angle for all three cases. The mass burning rate is given by $\frac{dy_m}{d\theta} = \frac{\dot{m}_b}{m}$. The energy conversion rate is defined as the rate of change of chemical internal energy, $u^c \equiv \sum y_i \Delta u^\circ_{f,i}$, normalized by the total energy released

$$\frac{dy_c}{d\theta} = -\frac{1}{y_f \eta_c \mathrm{LHV}}\frac{du^c}{d\theta} \quad (9)$$

where y_f is the fuel mass fraction in the reactant mixture, η_c is the combustion efficiency and LHV is the fuel's lower heating value, and the negative sign is needed because a reduction in the chemical internal energy results in an increase in the sensible internal energy, which would be considered as a positive energy conversion rate. Appendix A describes the calculation of η_c and LHV. The mass burning curves shown in Figure 2 were the input to the model calculations.

Table 1. Test cases computed using a two-zone model.

Case	Φ	IMEP (bar)	P_{max} (bar)
1	1.0	2.4	13.2
2	1.2	25.8	141.2
3	0.3	2.4	21.9

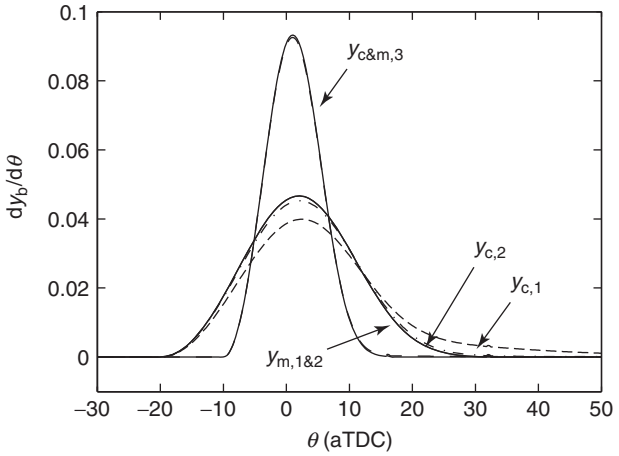


Figure 2. A comparison of the normalized mass (subscript m, and solid) and energy conversion rate (subscript c, and dashed) for the three test cases.

It can be seen in Figure 2 that there is a difference in $\dot{y}_m$ and $\dot{y}_c$, and that the difference is largest for case 1. The discrepancy between the mass burning rate and the energy conversion rate is due to dissociation of the combustion products. Intermediate carbon- and hydrogen-containing species typically have less negative heats of formation than CO_2 and H_2O, causing the chemical energy release rate to be reduced for conditions of high dissociation. As demonstrated in Figure 1, the light-load stoichiometric condition (case 1) has burned gas conditions that have high temperature and relatively low pressure, which promotes dissociation and puts the burned gas states at high levels of α. For the high load condition, the high pressure reduces the effects of dissociation, and the mass burning and energy conversion rates are very close. The lean, low temperature combustion light-load condition has negligible dissociation because of the low peak temperatures, and the mass burning and energy conversion rates are nearly identical.

3 SINGLE-ZONE MODELS

The implementation of a single-zone thermodynamic model is based on six major assumptions, which are enumerated in the text later as (A1–A6). The most significant assumption is that (A1) the cylinder contents are uniform in temperature and composition. This assumption is invalid for virtually all engine combustion regimes during the combustion event,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

that is, at the time of interest. For homogeneous-charge spark-ignition engines, the chamber is more aptly described as having two zones of different composition and temperature; for direct-injection compression-ignition engines, the mixture varies substantially in composition (containing various fractions of air, fuel, and combustion products) and temperature. In both cases, the mass-average composition and temperature, which are available to describe the single-zone's thermodynamic state, are not necessarily realized anywhere in the chamber except before and after combustion. This complicates the assignment of thermodynamic properties to the mixture.

There are two methods of implementing a single-zone model. The first, and most widespread, method assumes that the combustion event can be modeled as a heat addition to the system across the system boundary, which is defined to enclose the cylinder contents. This rate of heat addition is assumed to be the rate of conversion of chemical to sensible internal energy, see Equation 9. Heat transfer from the surroundings to the system can (and should) be accounted for separately. The second implementation method assumes that mass (fuel) is added to the system and that the in-cylinder mixture represents the equilibrium thermodynamic state of the air–fuel mixture. The rate of mass addition then represents the combustion rate. There are significant differences in these treatments. First, as was seen in Figure 2, there is a difference between a mass burning rate and an energy conversion rate. Second, because the former treatment does not use the changing chemical composition of the mixture to assess the mixture's internal energy (it is done through the heat addition term), the *sensible* internal energy should be used, whereas the latter treatment, which does account for the chemical energy in the fuel mass added, uses the *absolute* internal energy.

The First Law for an open system defined by the piston crown and combustion chamber walls can be written as

$$u\frac{\mathrm{d}m}{\mathrm{d}t} + m\frac{\mathrm{d}u}{\mathrm{d}t} = \dot{Q}_{\mathrm{w}} - p\frac{\mathrm{d}V}{\mathrm{d}t} + \dot{m}_{\mathrm{crev}}h_{\mathrm{crev}} + \dot{m}_{\mathrm{f}}h_{\mathrm{f}} \qquad (10)$$

where all thermodynamic properties are their absolute values and provision has been made for mass addition (considered positive) to the system from the crevice (e.g., piston ring crevices), $\dot{m}_{\mathrm{crev}}$, and via the addition of fuel, $\dot{m}_{\mathrm{f}}$, to the system. The value of h_{crev} depends on the direction of the flow, taking its value based on where it flowed from. The heat addition method described earlier is developed by (A2) neglecting the fuel addition term, writing the internal energy explicitly in terms of the chemical and sensible parts, $u = u^{\mathrm{c}} + u^{\mathrm{s}}$, (A3) neglecting the chemical enthalpy of the crevice flow, (A4) neglecting the chemical internal energy of the system mass derivative term, and defining the chemical heat addition, $\dot{Q}_{\mathrm{ch}}$, as

$$\dot{Q}_{\mathrm{ch}} \equiv -m\frac{\mathrm{d}u^{\mathrm{c}}}{\mathrm{d}t} \qquad (11)$$

similar to Equation 9. This gives an explicit relation for chemical heat addition

$$\dot{Q}_{\mathrm{ch}} = m\frac{\mathrm{d}u^{\mathrm{s}}}{\mathrm{d}t} + p\frac{\mathrm{d}V}{\mathrm{d}t} + (u^{\mathrm{s}} - h^{\mathrm{s}}_{\mathrm{crev}})\dot{m}_{\mathrm{crev}} - \dot{Q}_{\mathrm{w}} \qquad (12)$$

The mass addition method requires the assumption that $\dot{m}_{\mathrm{crev}} = 0$ because the composition of the mixture varies with time so it would be difficult to know the crevice composition. Using an overall mass balance, $\frac{\mathrm{d}m}{\mathrm{d}t} = \dot{m}_{\mathrm{f}}$, one finds the following expression for the mass addition rate

$$(u - h_{\mathrm{f}})\ \frac{\mathrm{d}m}{\mathrm{d}t} = \dot{Q}_{\mathrm{w}} - p\frac{\mathrm{d}V}{\mathrm{d}t} - m\ \frac{\mathrm{d}u}{\mathrm{d}t} \qquad (13)$$

The major challenge for the heat addition method is defining the gas properties. The composition of the mixture is known to change with time, but this composition is undefined, so assessing the sensible internal energy and its rate of change by conventional means is not strictly possible. The major challenge of the mass addition method is that the composition of the mixture is assumed to change as a function of the fuel mass addition, which may be a suitable physical description of diesel combustion but does not describe spark-ignition combustion well.

The one added complication that arises in the mass addition method is that the mixture state is changing because of pressure, temperature, and equivalence ratio (as fuel mass is added to the system). As such, all thermodynamic properties are a function of three properties, and the temporal derivatives of a property, such as Equation 7, need to include this equivalence ratio, Φ, effect. The mass addition method is not extensively used and will not be discussed further.

The standard development of the heat addition method involves the additional assumptions: (A5) the sensible internal energy is a function of temperature only so $\frac{\mathrm{d}u^{\mathrm{s}}}{\mathrm{d}t} = c_{\mathrm{v}}\frac{\mathrm{d}T}{\mathrm{d}t}$; and (A6) the mixture has a constant molecular weight, which enables the ideal gas equation of state, for example, Equation 5, to be simplified to find $\frac{\mathrm{d}T}{\mathrm{d}t}$ in terms of known or measured values. Using these assumptions and the specific heat ratio $\gamma \equiv \frac{c_{\mathrm{p}}}{c_{\mathrm{v}}}$, one gets

$$\dot{Q}_{\mathrm{ch}} = \frac{\gamma}{\gamma - 1}p\frac{\mathrm{d}V}{\mathrm{d}t} + \frac{1}{\gamma - 1}V\frac{\mathrm{d}p}{\mathrm{d}t} + (u^{\mathrm{s}} - h^{\mathrm{s}}_{\mathrm{crev}})\dot{m}_{\mathrm{crev}} - \dot{Q}_{\mathrm{w}} \qquad (14)$$

where the mixture composition is still undefined. The assumption (A6) also allows a direct estimate of the gas temperature as $T = pV\,/\,(mR)$.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

As stated earlier, appropriately defining the mixture's properties is difficult. Figure 3a compares the specific heat ratio, γ, calculated for pure air, stoichiometric isooctane–air reactants, and the equilibrium products of stoichiometric isooctane–air combustion. The product γ values were calculated using both the absolute mixture internal energy, $\widetilde{c}_{v,1} = \frac{\partial u}{\partial T}|_p$, and the mixture sensible internal energy $\widetilde{c}_{v,2} = \frac{\partial u^s}{\partial T}|_p$; both values of $\widetilde{c}_v$ differ from $\sum y_i c_{v,i}$ because of the shifting equilibrium composition. From the data of Figure 3a, it can be seen that there is a significant difference between pure air, the reactant mixture, and the product mixture. The product γ is seen to vary substantially depending on the method of calculation. Using the absolute internal energy, one sees a significant decrease in γ above 1500 K, and the value of γ in this temperature range is strongly pressure dependent, consistent with the curvature of the iso-α contours in Figure 1. Using the sensible internal energy, one finds that the product mixture γ is nearly constant for temperatures greater than 1500 K and nearly independent of pressure. It should be noted that the curves shown in Figure 3a for the absolute internal energy method closely resemble those of Gatowski *et al*. (1984), from which they developed the often cited correlation

$$\gamma = \ 1.392 - 8.13 \times 10^{-5} T \tag{15}$$

where the temperature is in Kelvin. The γ in Equation 14 is, however, based on the changes in sensible internal energy, and the behavior of the sensible internal energy is significantly different—instead of continuing to decrease with temperature it plateaus for temperatures in excess of 1500 K—and is nearly independent of pressure, which satisfies assumption (A5).

Chun and Heywood (1987) suggested that a three-part representation for γ should be used. Up to combustion, the reactant mixture value, consistent with Figure 3a, should be used; during combustion, a constant value should be used; and after combustion, the product mixture value should be used. They arrived at this result by back-calculating γ from a single-zone model applied to results acquired from a two-zone model. They argue that during the combustion event the change in internal energy is dominated by the change in composition from reactants to products, that is, the chemical change, and was, therefore, independent of temperature. The end of combustion period was not clearly defined. These results highlight a fundamental problem with the heat addition method. At the time of peak temperature, unless Equation 12 can be satisfied with $\frac{du^s}{dt} = 0$, the mixture specific heat will be infinite. It is interesting to note that the γ results based on the sensible internal energy in Figure 3a show a similar trend of a nearly constant γ value for the combustion products at combustion-like conditions. A simple correlation for γ can be found from data like Figure 3a with a second fit performed to account for the equivalence ratio, Φ, in the range $0.3 < \Phi < 1.2$, as

$$\gamma = \max \begin{cases} \gamma_0 + \gamma_1 T \text{ [K]} \\ \gamma_{HT} \end{cases} \tag{16}$$

where $\gamma_0 = 1.426 - 0.0459\Phi$, $\gamma_1 = 10^{-5} \times (1.02\Phi^2 - 3:30\Phi - 9.44)$, and $\gamma_{HT} = 0.0386\Phi^2 - 0.0838\Phi + 1.33$. The temperature-dependent part of Equation 16 corresponds to the reactant mixture, and the constant part corresponds to the plateau region of the products (with a slight 0.03 offset to better match heat release data). A simple justification for this correlation for γ is that the mixture will likely contain zones of reactants at modest temperature and products at high temperature at the time of combustion.

Figure 3b shows the change in the molar mass between reactants and products of an isooctane–air mixture as a function of the equivalence ratio. It can be seen that there is a significant difference between the two values, with the difference being nearly 10% at stoichiometric, with a small effect of pressure. These differences will minimally affect the results of the heat release calculation—through the γ correlation—but it will affect the mass-average temperature calculated from the ideal gas relation by a proportional amount. The more subtle effect comes through (A6), where the molar mass changes have been neglected in the application of the ideal gas equation.

Wall heat transfer in the engine environment is difficult to model accurately because it varies significantly across the chamber surface and is difficult to measure. The convective heat transfer coefficient, in an analogy to turbulent pipe flows, is usually written as a Nusselt–Reynolds number relation of the form

$$\frac{hL}{k} = a\left(\frac{\rho UL}{\mu}\right)^b \tag{17}$$

where h is the convection coefficient, L is a representative length scale, k is the thermal conductivity, U is a representative velocity, μ is the dynamic viscosity, and a, b are constants. The choice of the characteristic length and velocity scales, in addition to the two constants, differentiate the correlations found in the literature (Borman and Nishiwaki, 1987; Heywood, 1988).

A major issue in the heat transfer model is that the absolute magnitude predicted by correlations such as Equation 17 is not well captured. This has a direct bearing on the estimated heat release rate and can be ameliorated by invoking an overall energy balance. The integrated heat

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

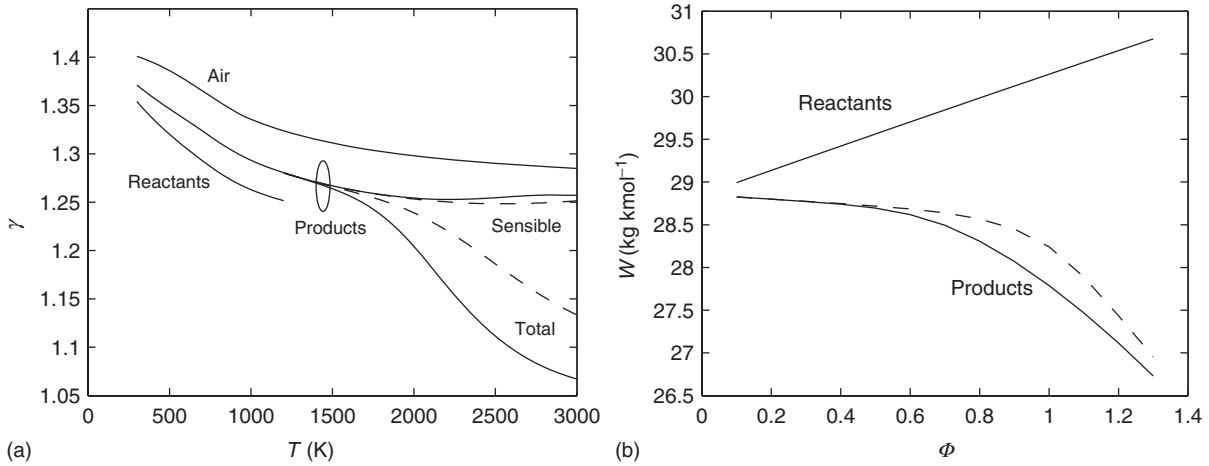


Figure 3. (a) Comparison of specific heat ratio for air, stoichiometric reactants, and the equilibrium products of stoichiometric combustion found from the sensible and total internal energy changes with temperature at 1 atm (solid) and 100 atm (dashed). (b) The molecular weight of reactants and products at the equilibrium flame temperature for 1 atm (solid) and 100 atm (dashed). All calculations performed with isooctane as the fuel.

release is made to equal the total fuel energy, $\int \dot{Q}_{ch} = \eta_c m_f \text{LHV}$, by scaling the wall heat transfer term with a multiplicative constant. The scaling constant that satisfies this condition can be found by iteration or analytically by performing term-wise integration of the parts of Equation 14.

4 APPROXIMATE METHODS

Before the advent of modern computers, several investigators had developed methods of estimating the cumulative burn fraction using the fact that engine pressure data show a polytropic behavior when plotted on logarithmic pressure–volume axes. These methods are described more fully by Amann (1985) and Brunt and Emtage (1997) and are briefly summarized here. All of these methods rely on using a polytropic exponent that is derived from fitting the pressure–volume data, and no direct effort is made to relate the quantities to fundamental thermodynamic properties. Thus, these methods should be considered approximate and used appropriately, for example, for relative comparisons between operating conditions in the same or similar engines.

Assuming a constant-volume combustion process, and neglecting crevice and heat transfer effects, it can be seen from Equation 14 that the heat release rate is proportional to the pressure rise rate. Integrating to an intermediate time, and normalizing to the integral overall time, one finds that the cumulative burn fraction, y_1, where the subscript 1 refers to the first approximate method, is found as

$$y_1 = \frac{p - p_i}{p_f - p_i} \tag{18}$$

where p_i and p_f are the pressures at the start and the end of combustion, respectively. Because the combustion event does not occur at a fixed volume, the pressure terms need to be corrected to their top dead center (TDC) equivalent values using the polytropic assumption $p_{TDC} V_{TDC}^n = pV^n$, where n is the polytropic exponent. Making this adjustment, one finds

$$y_1 = \frac{pV^n - p_i\ V_i^n}{p_f\ V_f^n - p_i\ V_i^n} \tag{19}$$

This approach was originally developed by Marvin (1927). The end of combustion is difficult to define accurately, but Brunt and Emtage (1997) suggest using 10 crank angles past the point that had the maximum value of $pV^{1.15}$.

Rassweiler and Withrow (1938) developed an approximate method of estimating heat release based initially on their high speed flame images, but extended it to an arbitrary case where only pressure data were measured. The burned mass at a given crank angle is compared to that at exhaust valve opening to determine the burned mass fraction. This is combined algebraically with a similar relation based on the unburned gas to find

$$y_2 = \frac{p^{1/n} V - p_i^{1/n} V_i}{p_f^{1/n} V_f - p_i^{1/n} V_i} \tag{20}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5 COMPARISON OF CALCULATION METHODS

The three test cases that were generated from a two-zone model (Table 1) were postprocessed to determine the heat release rate using both the single-zone (denoted First Law) and approximate methods, and the results are shown in Figure 4. The single-zone method was calculated using Equation 16 for the determination of γ, the Woschni correlation (Borman and Nishiwaki, 1987 and Heywood, 1988) for the heat transfer rate, and the heat transfer rate was scaled to enforce an overall energy balance as described earlier. Figure 4 also includes the known mass (m_{act}) and energy (q_{act}) conversion rates. The formulation of the single-zone and approximate methods are more consistent with the energy conversion rate and that will serve as the basis of comparison.

In all three cases, which span a wide range of operating conditions, there is a reasonable agreement between the calculated heat release rate and the known values. There are, however, some differences. For cases 1 and 2, the single-zone method shows the best agreement and is nearly indistinguishable from the actual energy conversion rate curve. For case 3, however, the single-zone method slightly overshoots the real case. The overshoot is corrected at the end of the calculation by the normalization and this effect could be minimized by shortening the calculation window, but this would require *a priori* knowledge of the combustion event.

The approximate methods, although not rigorous treatments, provide results that are relatively close to the single-zone analysis. The largest discrepancy is seen for case 1 where there is a significant amount of dissociation. For this case, the approximate methods are closer to the mass

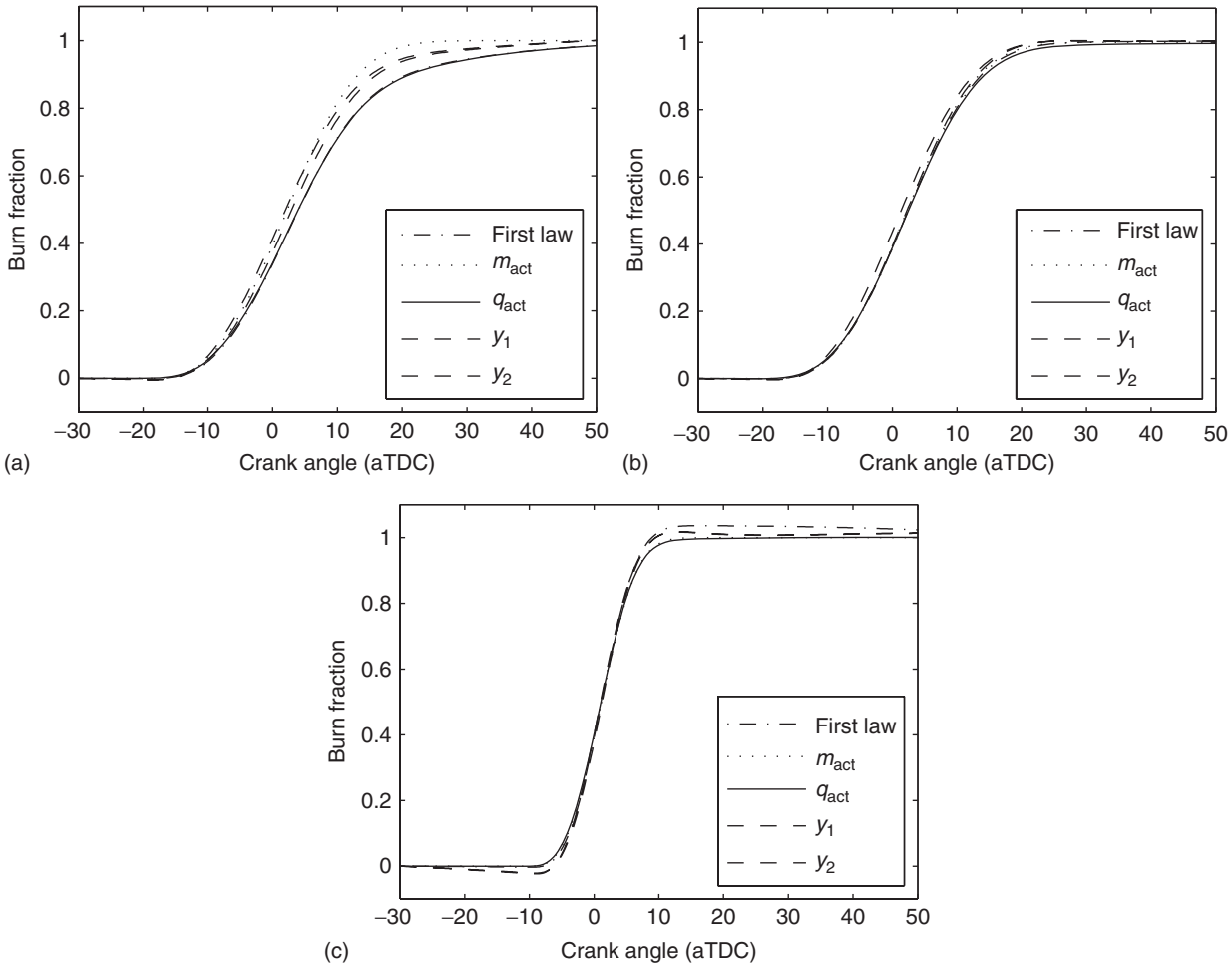


Figure 4. Comparison of actual (mass- and energy-based) and calculated cumulative burn fractions for (a) case 1, (b) case 2, and (c) case 3. The First Law curves were calculated using the heat addition method, and the two approximate cases shown correspond to the numbers given in the text.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Comparison of the accuracy of the single-zone and approximate methods for the calculation of CA_{10}, CA_{50}, and CA_{90}.

	First Law			y_1			y_2		
Case	CA_{10}	CA_{50}	CA_{90}	CA_{10}	CA_{50}	CA_{90}	CA_{10}	CA_{50}	CA_{90}
1	0.23	0.05	− 0.42	− 0.18	− 1.02	− 5.08	− 0.90	− 1.98	− 6.02
2	0.16	− 0.02	− 0.45	0.11	− 0.21	− 1.10	− 0.60	− 1.11	− 1.70
3	0.30	0.04	− 0.61	0.62	0.18	− 0.44	0.46	− 0.06	− 0.60

The value given corresponds to the difference between the calculated value and that derived from the energy release rate from the two-zone model.

conversion rate than the energy conversion rate, but this is considered to be a coincidence rather than a result. The other significant deviation of the approximate methods is for case 3 where both y_1 and y_2 are seen to go negative before the start of combustion. This is a result of a deviation from polytropic behavior because of either heat transfer effects or temperature effects on γ.

Heat release results are often represented by the crank angle at which the cumulative heat release reaches a specified fraction, for example, CA_{50} corresponds to the crank angle at 50% of the cumulative heat release. Table 2 quantitatively compares the accuracy of the CA_{10}, CA_{50}, and CA_{90} values obtained using the single-zone and both approximate methods. The results are presented in terms of the difference between the calculated value and the value directly calculated from the known energy conversion rate, that is, $\Delta\theta = CA_x - CA_{x,\text{known}}$. It can be seen from Table 2 that, for the cases tested, the single-zone model returns values of CA_{10}, CA_{50}, and CA_{90} that are within ± 0.5 crank angles of the actual result. For cases 2 and 3, the approximate methods give reasonable results, within ± 1.5 crank angles, but it can be seen that the errors are largest for CA_{90}. The approximate methods applied to case 1 show the worst performance, especially for CA_{90} where a difference of ~6 crank angles is observed.

5.1 Data collection effects

It is well known that the accuracy of the pressure data has a direct impact on the calculation of heat release and that acquiring accurate cylinder pressure data requires care (Randolph, 1990). Pressure data are typically acquired using piezoelectric transducers that measure the dynamic pressure very accurately, but the dynamic pressure needs to be pegged to an absolute reference pressure once per cycle to obtain thermodynamic pressure, and this pegging pressure value has limited accuracy. The pressure data also need to be referenced to the TDC position of the crankshaft in order to be able to use the kinematic relationship for volume, and there can be an inaccuracy in the location of TDC determined with a rotary encoder. In addition, there can be inaccuracy in knowing the clearance volume (compression ratio) and the trapped mass of fuel and air; the latter are problematic in multicylinder engines where the cylinder-to-cylinder distribution may be imbalanced. For the present calculations, the total trapped mass will be adjusted at a constant equivalence ratio.

The effects of the aforementioned parameters on CA_{10}, CA_{50}, and CA_{90}, obtained using the single-zone method applied to case 1, are shown in Figure 5. The results are presented as the difference relative to the known energy conversion rate equivalent, $\Delta\theta$. Figure 5a shows the effect of an inaccuracy of the pegging pressure on $\Delta\theta$. It can be seen that CA_{10} and CA_{50} are relatively unaffected by an offset in the pegging pressure, but CA_{90} is strongly affected with the calculated CA_{90} retarding as the offset in pegging pressure increases. The effect of the encoder phasing is shown in Figure 5b. There is a noticeable effect on both CA_{10} and CA_{50}, but the two values track in the same direction, that is, they retard with an increase in phasing offset. The CA_{90} value is again more sensitive than both CA_{10} and CA_{50}, and for phasing errors, it moves in an opposite direction from the CA_{90} value, causing even more significant errors in the CA_{10-90}, which is often used as a measure of combustion duration.

The effect of inaccuracies in the TDC volume and trapped mass is presented in Figure 5c and d as a function of the ratio of the assigned value to the actual value. The effect of the TDC volume (or compression ratio) is significant for CA_{10} and CA_{50} and quite strong for CA_{90}. For all three parameters, the direction of the shift is consistent, that is, everything phases in the same direction. The effect of the trapped mass ratio is weak on CA_{10} and CA_{50} but is moderate on CA_{90}, although of lower magnitude than the other parameters investigated.

In order to extend these results to the other operating cases and to the approximate calculation methods, the results have been recast in terms of a required accuracy of the measurement in order to obtain ± 0.5 crank angle precision of CA_{10}, CA_{50}, or CA_{90} independent of the offset shown in Table 2. The accuracy was found by dividing 0.5 crank angles by the slope of the respective CA_x curve at the nominal location (an offset of zero or a ratio of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

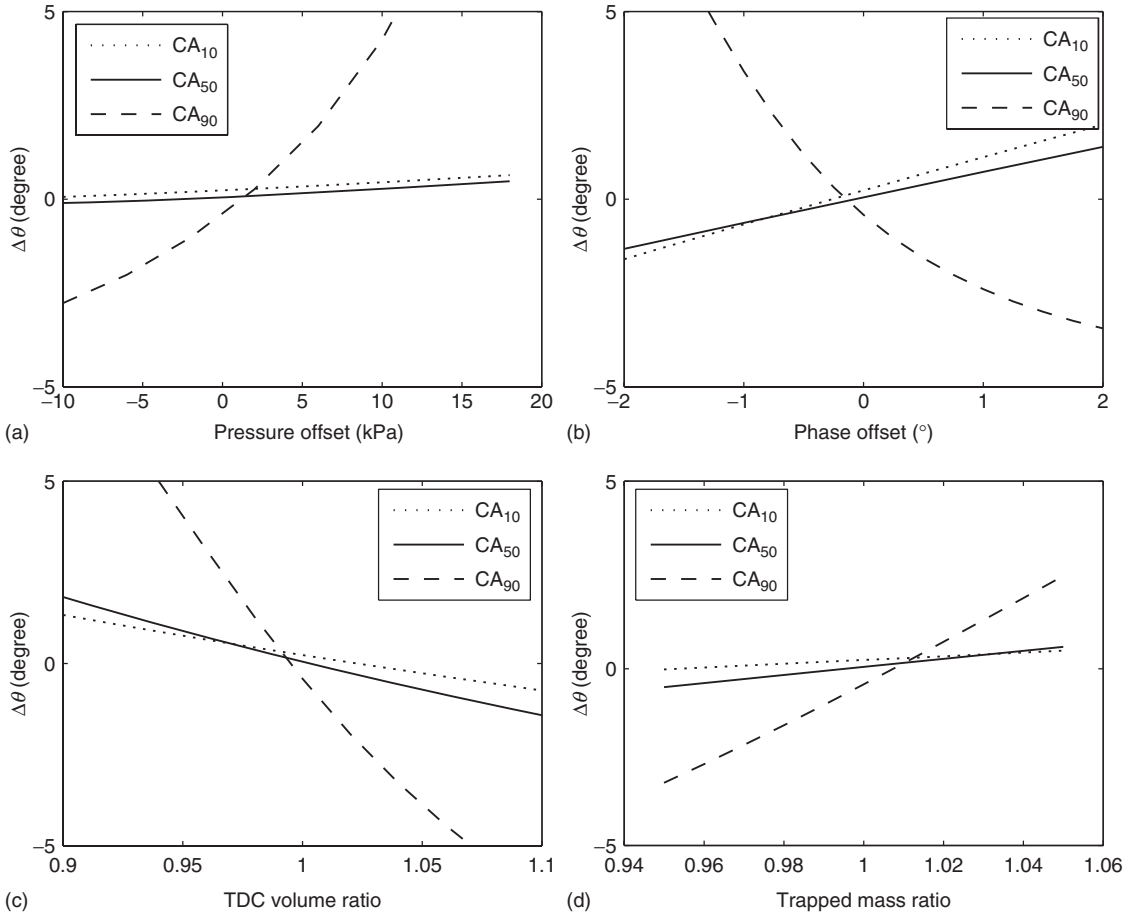


Figure 5. The effect on CA_{10}, CA_{50}, and CA_{90} from incorrect (a) pressure pegging, (b) TDC timing, (c) clearance volume measurement, and (d) trapped mass assignment. All data are for case 1 and were calculated using the single-zone method.

unity). Using this approach, large numbers indicate that low accuracy is required and that the results are very insensitive to a parameter, conversely, small values indicate a high level of sensitivity and the need for high data fidelity. The results for the single-zone and both approximate methods are shown in Table 3.

Table 3 contains a lot of data, and only the general trends will be discussed. For all calculation methods, the throttled,

Table 3. Comparison of the accuracy of the pegging pressure, encoder phasing, trapped mass estimate, and TDC volume estimate required to produce a ± 0.5 crank angle precision in CA_{10}, CA_{50}, and CA_{90}.

		First Law			y_1			y_2		
	Case	CA_{10}	CA_{50}	CA_{90}	CA_{10}	CA_{50}	CA_{90}	CA_{10}	CA_{50}	CA_{90}
Peg (kPa)	1	24.98	25.04	1.56	1.40	0.70	0.10	1.72	0.86	0.13
	2	212.34	262.06	51.90	14.68	7.71	2.33	18.25	9.48	2.62
	3	28.05	46.88	16.75	4.04	2.96	0.95	4.19	3.24	1.00
Phase (CA)	1	0.56	0.74	0.18	0.79	1.61	0.13	0.78	1.21	0.17
	2	0.56	0.67	1.15	0.76	1.26	0.29	0.74	1.03	0.35
	3	0.72	0.67	1.85	1.45	1.07	0.96	1.55	1.03	1.14
Mass (%)	1	10.1	4.7	0.9	—	—	—	—	—	—
	2	11.2	5.2	1.7	—	—	—	—	—	—
	3	26.1	9.6	3.2	—	—	—	—	—	—
V_{TDC} (%)	1	4.9	3.1	0.6	8.6	6.3	2.1	7.7	6.0	2.2
	2	5.2	3.5	1.2	9.2	6.9	2.9	8.1	6.5	3.0
	3	4.3	4.6	1.8	7.1	8.0	3.3	6.8	8.0	3.4

See the text for the calculation procedure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

light-load operating condition (case 1) required the highest accuracy. In addition, the highest accuracy was required to correctly calculate CA_{90}, consistent with the results seen in Figure 5. The two approximate methods gave very similar results. The approximate methods are only sensitive to changes in the polytropic coefficient, which is determined by a power-law fit to the pressure-volume data, and thus are not affected by the trapped mass. The approximate methods are very sensitive to the pegging pressure, requiring on the order of 1 kPa accuracy to correctly estimate CA_{50}, and significantly better to accurately measure CA_{90}. In contrast, the single-zone model is nearly an order of magnitude less sensitive to the pegging pressure. The effect of encoder phasing inaccuracy is comparable for the single-zone and approximate methods, and 0.5 crank angle accuracy provides sufficient results for the range of conditions tested. The trapped mass estimation is relatively unimportant for CA_{10} and CA_{50}, but it does produce a significant effect on CA_{90} for the single-zone method. The effect of the clearance volume is more important for the single-zone model, with $\sim$1% accuracy required to correctly calculate CA_{90}.

In addition to the effects described earlier, another major issue with respect to utilizing either the single- or two-zone methods is the ability to accurately measure the pressure derivative. The accurate determination of the pressure derivative is a function of the frequency content of the pressure signal, the presence of random noise, biases imposed by the data acquisition hardware, and the numerical method used for evaluating the derivative from the measured pressure data. These issues become more demanding when single-cycle analysis of the data is performed and the smoothing effect of ensemble averaging multiple cycles is removed.

Differentiating experimental data amplifies noise, thus, data are often numerically smoothed before differentiation. This can, however, bias the results if rapid pressure transients associated with the heat release event are present. A wide array of data smoothing techniques has been used ranging from simple nearest-neighbor or running averages to more complicated methods. The preferred method is to use a rigorous filter with a well-defined cutoff frequency and roll-off because the cutoff frequency can be tuned to avoid specific fixed frequencies (discussed later) and the effect of the filtering on the gradient can be quantified.

The rapid expansion of a local pocket of gas in a closed chamber can excite pressure oscillations that are associated with the allowable standing wave modes of the enclosure. This occurs when the end gas of a spark-ignition engine autoignites (knock), but it can occur in compression-ignited engines as well. These oscillations occur during the energy release period and affect the pressure derivative calculation. Because this is a resonance phenomenon, the oscillation frequencies are fixed and can be reasonably estimated based on the resonance characteristics of a cylindrical chamber. Therefore, it is possible to design a *time-domain* filter that can alleviate the pressure oscillations from the measured data. Another issue that can arise with high rates of pressure rise (due to either the combustion or chamber resonances) is the excitation of the natural frequency of the pressure transducer, which is typically $\sim$100 kHz. If the data collection rate is significantly higher, this can be filtered out; with sample rates lower than the natural frequency, which are typical for crank-angle-based collection systems, it is possible that the high natural frequency can alias the measured signal. It is recommended to use an analog anti-aliasing filter ahead of the data acquisition system to avoid this possibility. The effects of random noise are normally eliminated by the filtering, and it is typical to have data quality that will enable single-cycle analysis of the heat release.

5.2 Single-zone assumption effects

Directly assessing each of the single-zone method assumptions, that is, (A1)–(A6), is difficult without detailed information from all points within the combustion chamber. Using the two-zone model data for the evaluation of the heat release calculations, assumptions (A2–A4) were automatically satisfied. The assumption that the mixture sensible internal energy is a function of temperature only, assumption (A5), was justified, in part, by the lack of pressure dependence seen for the sensible energy-derived γ in Figure 3. However, this assumes that the single-zone composition is that of the products. In reality, the composition is (in a loose sense) defined by the choice of γ. Figure 6a shows the effect that the correlation used for γ in Equation 14 has on the cumulative energy release for case 2. It can be seen that the use of Equation 16 gives a good match between the calculated and known heat release rates (as seen previously). Further, the choice of $\gamma = 1.28$, which is slightly higher than the high temperature plateau seen in Figure 3a, also gives a good match to the known heat release rate. The fixed value of $\gamma = 1.28$ also gave acceptable results for case 1 but resulted in a significant overshoot for case 3. This occurred because this value of γ is too low for the very lean mixture and low temperatures of case 3. The linear relation of Gatowski *et al*. (1984) (Equation 15) gives results that are significantly different than the actual value; a similar result was found for case 1. This result underscores the sensitivity of the calculation procedure to the choice of γ, but from

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

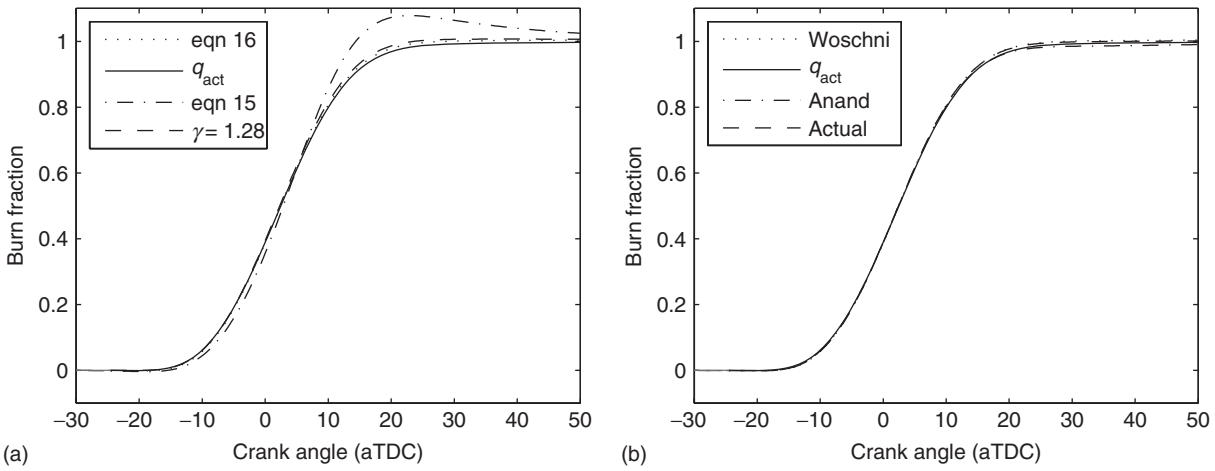


Figure 6. Comparison of the effect of the correlation for γ on the cumulative heat release (a) and (b) the effect of the heat transfer correlation. All data were calculated using the single-zone model for case 2.

the larger perspective, it shows the importance of how one ascribes thermodynamic properties to a single-zone mixture that is in reality not homogeneous in temperature or composition.

Figure 6b shows the effect that the choice of the heat transfer correlation in the single-zone model has on the cumulative energy release; recall that an overall energy balance is used to find a multiplicative constant. Four cumulative energy release curves are shown in Figure 6b: the known value from the two-zone model, the results from the Woschni and Anand correlations, and the result achieved using the known heat loss rate from the two-zone model directly in the heat release calculation. It can be seen that there is very little sensitivity of the cumulative energy release to the heat transfer correlation used. It is interesting to note that the multiplier on the heat transfer varied from 0.5 to 1.7 over these conditions.

6 SUMMARY

The essential elements of the calculation of the heat release rate from experimentally derived cylinder pressure data were discussed. The fundamental equations for a two-zone thermodynamic calculation were developed, and it was shown that the ideal gas equation of state was sufficient, but the shifting equilibrium composition of the gas required that the enthalpy and internal energy be written as functions of both temperature and pressure. A two-zone model (with a prescribed mass conversion rate) was used to generate test data against which a single-zone and two approximate heat release calculation methods were tested. The energy conversion rate for the two-zone model result was shown to be different from the mass conversion rate because of the dissociation of the product species. This effect was most pronounced for high temperature, low pressure conditions as are found in throttled spark-ignition engines; at higher loads or under dilute conditions, there was not a significant effect of dissociation.

The single-zone model for heat release calculations was developed based on modeling the heat release as either a fuel mass addition or an external heat addition; the former method was not explored in detail because of the widespread use of the latter method. The heat addition method does not consider the chemical energy of the gas mixture, thus the sensible internal energy should be used. It was shown that the product-mixture-specific heat ratio calculated based on the sensible internal energy was nearly constant for temperatures in excess of 1500 K and was nearly independent of pressure. A two-part equation for γ that considers the mixture equivalence ratio was presented. The assignment of thermodynamic properties to the (assumed) homogeneous mixture that represents the cylinder contents, which is solely manifested in the specific heat ratio, was seen to have a significant effect on the heat release calculation. The correlation chosen for the wall heat transfer, however, did not strongly affect the heat release rate when an overall energy balance was used to ensure that the integrated heat release matched the known fuel energy converted. The multiplier of the heat transfer term required to close the energy balance differed significantly from unity and highlights the difficulties associated with accurately calculating heat transfer in engines. Using the proposed equation for γ and enforcing an overall energy balance, a good correspondence was seen between the calculated heat release rate and the known energy conversion rate.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The approximate methods were found to provide a reasonable estimate of the cumulate heat release for conditions where there was minimal dissociation. The agreement was worse for the light-load condition, especially at high cumulative burn fractions. The major drawback of the approximate methods is that they do not quantitatively estimate energy, that is, the cumulative burn fraction will always asymptote to unity. Therefore, on a cycle-by-cycle basis, it is not possible to assess the mass of fuel burned, which can be done using the single-zone model when the heat transfer is well modeled.

Both the single-zone and approximate methods are sensitive to the quality of the pressure data. The effects of incorrect pressure pegging, TDC determination, clearance volume, and trapped mass were investigated. The approximate methods were not sensitive to the trapped mass but required high accuracy of the pegging pressure in order to correctly predict CA_{50} and CA_{90}. All methods required ~0.5 crank angle accuracy of the shaft encoder.

APPENDIX A—COMBUSTION EFFICIENCY CALCULATION

The combustion efficiency is calculated as the ratio of the isothermal, constant pressure heat rejection from a given fuel–air mixture to that from a stoichiometric mixture with ideal products, that is, CO_2, H_2O, and N_2. From a simple First Law balance, the heat rejection per mass of fuel is given by

$$\frac{\dot{Q}}{y_f \dot{m}} = \frac{1}{y_f} \sum_{\text{prod}} y_i \Delta h^\circ_{f,i} - \Delta h^\circ_{f,f} \qquad \text{(A1)}$$

and in the case of ideal products, this gives the lower heating value of the fuel, that is, $\text{LHV} = \dot{Q}_{id}/(y_{f,id}\dot{m})$, where the id identifier is for the ideal case. The combustion efficiency is then found as

$$\eta_c = \frac{\frac{1}{y_f}\sum_{\text{prod}} y_i \Delta h^\circ_{f,i} - \Delta h^\circ_{f,f}}{\frac{1}{y_{f,id}}\sum_{\text{prod}} y_{i,id} \Delta h^\circ_{f,i} - \Delta h^\circ_{f,f}} \qquad \text{(A2)}$$

REFERENCES

Amann, C.A. (1985) Cylinder-pressure measurement and its use in engine research. SAE Paper 852067.

Borman, G.L. and Nishiwaki, K. (1987) Internal-combustion engine heat transfer. *Progress in Energy and Combustion Science*, **13**, 1–46.

Brunt, M.F.J. and Emtage, A.L. (1997) Evaluation of burn rate routines and analysis errors. SAE Paper 970037.

Chun, K.M. and Heywood, J.B. (1987) Estimating heat release and mass-of-mixture burned from spark-ignition engine pressure data. *Combustion Science and Technology*, **54**, 133–143.

Foster, D.E. (1985) An overview of zero-dimensional thermodynamic models for IC engine data analysis. SAE Paper 852070.

Gatowski, J.A., Balles, E.N., Chun, K.M., *et al*. (1984) Heat release analysis of engine pressure data. SAE Paper 841359.

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*, McGraw-Hill, New York.

Klein, S. and Nellis, G. (2012) *Thermodynamics*, Cambridge University Press, New York.

Krieger, R.B. and Borman, G.L. (1966) The computation of apparent heat release of internal combustion engines. ASME Paper 66-WA/DGP-4.

Marvin, C.F. (1927) Combustion time in the engine cylinder and its effect on engine performance. NACA Tech. Report 276.

Randolph, A.L. (1990) Methods of processing cylinder-pressure transducer signals to maximize data accuracy. SAE Paper 900170.

Rassweiler, G.M. and Withrow, L. (1938) Motion picture of engine flames correlated with pressure cards. *SAE Journal*, **42**, 185–204.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto122
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Zero- and One-Dimensional Methodologies and Tools

Jerald A. Caton
Texas A&M University, College Station, TX, USA

1 Introduction 1
2 Spark-Ignition Engine Simulations 2
3 Compression-Ignition (Diesel) Engine Simulations 11
4 Summary 12
Acknowledgments 12
References 12

1 INTRODUCTION

Zero- and quasi-one-dimensional engine cycle simulations are well-established tools for modern engine development. These cycle simulations are mainly based on the first law of thermodynamics, and often are called *thermodynamic simulations*. They have evolved over the years from elementary cycle analyses to fairly complex, thorough descriptions of engine processes and fluid properties. Today, a number of commercial codes are available, which are based on these methodologies. This chapter will provide descriptions of these simulations, and some example results to illustrate the type of information that may be obtained from the simulations.

The goal of these simulations is to mathematically simulate the significant engine processes, provide detailed time-resolved information, and predict overall engine performance. The significant engine processes include cylinder heat transfer, combustion, flows, friction, and related items. Benefits of these types of simulations are multiple, which include: (i) they help guide and interpret experimental work, (ii) they are an efficient way to complete extensive parametric studies, (iii) they provide opportunities for optimization, (iv) they may be the only technique to study certain variables, (v) they provide an educational opportunity to better understand engine processes, and (vi) they may contribute to the development of advanced on-board engine controllers.

The evolution of these simulations has been a result of improved computing capabilities, and increased understanding of the individual engine processes and the fluid properties. The first thermodynamic simulations were introduced in the 1960s. A signal work by Patterson and van Wylen (1964) is a good example of these early simulations. Some of the simplifications of this earlier work included idealized flow processes and adiabatic assumptions for some of the processes. Nevertheless, this work was a pioneering effort and an early version of current simulations. Interestingly, they included second-law calculations that were a dozen or more years ahead of any such similar work.

Engine cycle simulation development continued through the 1960s and 1970s, and by the end of the 1970s, these simulations had reached a fairly mature level. Examples of these simulations at the end of the 1970s have been described by Mattavi and Amann (1980) and Blumberg, Lavoie, and Tabaczynski (1979).

An important feature of these simulations is the treatment of the intake and exhaust flows. For cases where the flow dynamics are of less concern, quasi-steady flow and plenum assumptions for the intake and exhaust manifolds are often acceptable. For those cases where the flow dynamics are important, one-dimensional unsteady gas flow equations may be used. This latter feature is necessary where intake and exhaust piping arrangements must be evaluated. More on this topic may be found in Gas–Breathing and Air Management.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

In the mid-1980s, commercial codes began to appear. Examples today include GT Power (Gamma Technologies, Inc.), Wave (Ricardo, Inc.), Boost (AVL, GmbH), and Virtual 4-Stroke (Optimum Power Technologies, Inc.). For the most part, these simulations share the same basic principles that are described in this chapter. Of course, each code has unique features and capabilities that go beyond the basics. They are generally user friendly, but often the algorithms and methodologies are based on assumptions and approximations that are not clearly understood by the user. The information in this chapter is at least partly aimed at helping to improve this understanding.

1.1 Zero- and quasi-one-dimensional cycle simulations

Zero-dimensional cycle simulations are based on a global energy analysis of the cylinder contents (Foster, 1985). From this, time-varying global cylinder pressures and gas temperatures are obtained. Often, multiple zones may be designated to capture some of the spatial differences: but, this is still a zero-dimensional result. These multiple zones, for thermodynamic simulations, have no dimensionality (i.e., these zones are not related to specific locations). This type of simulation requires an input of information on the burning rate. This is often supplied as a mass fraction of fuel burned as a function of the crank angle. This input would need to include the duration of combustion. Since often these items are not known, the zero-dimensional cycle simulation is not generally predictive. In addition, these simulations do not include any dependence on the detail cylinder geometry.

Alternatively, the quasi-one-dimensional cycle simulations are based on attempts to predict the burning rates from more basic inputs such as turbulence, swirl, fuel chemistry, and any fuel jet characteristics or other fuel mixing processes. Some model of the geometry of the flame propagation or fuel jet development is needed as part of the description. Ideally, the start of combustion, the combustion duration, and the instantaneous combustion rate would be predicted. These items would then be used by the cycle simulation along with the other features described for the zero-dimensional simulations. These types of simulations include at least some dependence on the cylinder geometry. Although cycle simulations based on the quasi-dimensional features may be more predictive, often the inputs (turbulence, chemical kinetics, flame speeds) are not well known.

1.2 Multidimensional cycle simulations

For completeness, multidimensional cycle simulations are briefly mentioned next. These types of simulations, often called *computational fluid dynamic* (*CFD*) *models*, are described more fully in Multidimensional Simulation. In principle, these simulations solve the complete governing equations in all dimensions. These simulations require fine grids and small time steps. The inputs are detailed and include all geometry, turbulence, chemistry, and boundary-layer processes. These simulations are computer-intensive and often require hours or days of runtime. In addition, these multidimensional cycle simulations must resolve the complex interactions between the turbulence, combustion, heat transfer, and other key processes. Most of these interactions are still not fully understood. Nevertheless, great progress has been made over the past several decades and these simulations are being used more and more.

2 SPARK-IGNITION ENGINE SIMULATIONS

This section will describe the features of a specific example of a thermodynamic simulation for spark-ignition (SI) engines. Other simulations will possess similar features. We will then present example results, and highlight the strengths and weaknesses of such simulations. Some of the features (e.g., piston motion, heat transfer, flows, and friction) described for SI engines are equally applicable to compression-ignition (CI) engines.

2.1 Description of the thermodynamic simulation

This simulation (Caton, 2001, 2003, 2005, 2006; Shyani and Caton, 2009) is largely based on thermodynamic formulations, and is a complete representation of the four-stroke cycle including the intake, compression, combustion, expansion, and exhaust processes. Many submodels, assumptions, and approximations are integrated into the final simulation. The following describes some of these items for the simulation used here. Other simulations may employ slightly different assumptions and approximations.

The thermodynamic system is the cylinder contents (Figure 1). For portions of the cycle where the cylinder mixture is homogeneous, one zone is used for the thermodynamic system. During combustion, three zones (Caton, 2001, 2003) are used to better capture the high temperatures of combustion. For the purpose of this simulation, the engine is operating at steady state. The thermodynamic properties (including pressure and temperature) vary only with time (crank angle) and are spatially uniform in each zone. The intake and exhaust flow rates are determined from quasi-steady one-dimensional flow equations, and the intake and exhaust manifolds are infinite plenums containing gases

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

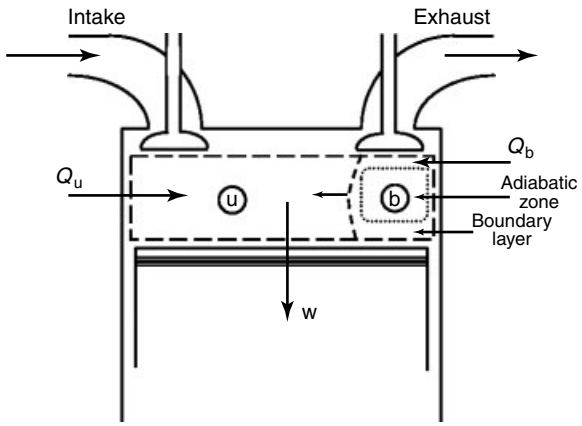


Figure 1. Schematic of the thermodynamic system.

at constant temperature and pressure. These assumptions result in the neglect of any "wave" dynamics in the inlet and exhaust systems.

Although no experimental data are cited in this chapter, previous work (e.g., Mattavi and Amann, 1980; Blumberg, Lavoie, and Tabaczynski, 1979) has demonstrated the success of these types of simulations on a global basis for duplicating experimental results.

2.1.1 Energy equations

As a result of thermodynamic analysis, the governing differential equations are obtained for the gas temperatures, cylinder pressure, volumes, and masses. The instantaneous cylinder conditions (temperatures, pressure, volumes, masses, and thermodynamic properties) as a function of crank angle are obtained by the simultaneous numerical integration of the various differential equations.

For the single-zone cases, the following is the appropriate relation:

$$\frac{\mathrm{d}(mu)}{\mathrm{d}\theta} = \dot{Q} - p\dot{V} + \dot{m}_{\mathrm{in}}h_{\mathrm{in}} - \dot{m}_{\mathrm{out}}h_{\mathrm{out}} \qquad (1)$$

where m is the cylinder mass, u is the specific internal energy, θ is the crank angle, $\dot{Q}$ is the rate of heat transfer, p is the cylinder pressure, $\dot{V}$ is the rate of cylinder volume change, $\dot{m}_{\mathrm{in}}$ is the mass flow rate into the cylinder, h_{in} is the specific enthalpy of the inlet mixture, $\dot{m}_{\mathrm{out}}$ is the mass flow rate out of the cylinder, and h_{out} is the specific enthalpy of the exiting mixture. In conjunction with other relations (such as for the thermodynamic properties, heat transfer, and burn rate (Olikara and Borman, 1975; Woschni, 1968; Wiebe, 1970; Heywood *et al*., 1979; Heywood, 1988; Watts and Heywood, 1980; Sherman and Blumberg, 1977; Sandoval and Heywood, 2003), Equation 1 may be used to find explicit relations for the derivatives of the overall average cylinder gas temperature and cylinder pressure.

The intake and exhaust processes are based on equations for one-dimensional quasi-steady flow that are corrected by an empirical discharge coefficient (see Gas–Breathing and Air Management for related details). The intake and exhaust manifolds are assumed to be at constant pressures. The instantaneous valve lift is approximated with a sinusoidal shape based on valve timings and the maximum valve lift. This approximation has been used for cases where the exact valve lift profile is not known (Sherman and Blumberg, 1977).

For the combustion process, multiple zones are used (Figure 1). First, the formulation is developed for two zones: an unburned zone (u) and a burned zone (b). The burned zone entrains mass from the unburned zone as the flame front proceeds. The next step is to divide the burned zone into the boundary layer and adiabatic core zones. Initially, the boundary layer has zero mass, and the adiabatic core is equal to the burned mass. The boundary layer increases in mass by entraining mass from the adiabatic core. The rate of mass entrainment of the boundary layer is dictated by the mass and energy conservation relations for the adiabatic core and boundary layer zones with a specified temperature definition. In other words, to satisfy the energy and mass conservation relations, a specific boundary layer mass is required. All the burned gas heat transfer is assigned to the boundary layer.

The engine friction includes mechanical friction and pumping friction. Mechanical friction, in turn, includes the rubbing friction (such as from the crank, pistons, and valve train) and the work associated with the auxiliaries (such as oil and water pump, and alternator). Algorithms for each of these items were published by Sandoval and Heywood (2003), and these are used here exactly as presented.

To complete the required input information, the boundary conditions for the inlet (temperature and pressure) and for the exhaust (pressure) are specified. To begin a particular engine cycle calculation, several parameters are not known. The initial amount of exhaust gases left in the cylinder from the previous engine cycle (residual), as well as the initial cylinder gas temperature and pressure, must be assumed. The complete calculation is repeated until the final values agree (within a specified tolerance) with the initial values. Depending on the initial values and the specified tolerance, this procedure usually finds convergence within about three complete cycles.

2.1.2 Exergy parameters

In addition to the first law of thermodynamics, the second law provides further insight about engine operation. The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

second law of thermodynamics is a rich and powerful statement of related physical observations and has a wide range of implications with respect to engineering design and operation of thermal systems. For example, the second law can be used to determine the direction of processes, to establish the conditions of equilibrium, to specify the maximum possible performance of thermal systems, and to identify those aspects of processes that are detrimental to the overall performance. Related to the analysis based on the second law of thermodynamics is the concept of exergy (or available energy). Exergy, a thermodynamic property of a system and its surroundings, is a measure of the maximum useful work that a given system may attain as the system is allowed to reversibly transition to a thermodynamic state that is in equilibrium with its environment. One key aspect of exergy is the fact that a portion of a given amount of energy is "available" to produce useful work while the remaining portion of the original energy is "unavailable" for producing useful work.

The exergy of the fuel and the cylinder gases is described by a number of parameters. For completeness, these parameters are briefly discussed next. Full details are available elsewhere (Caton, 2003, 2005, 2006; Shyani and Caton, 2009). Once the thermodynamic properties are known for a given set of conditions, the determination of exergy is fairly straightforward. In this development, the kinetic and potential energies are neglected (and can be shown to be negligible). At all times, for each zone

$$b = (u - u_o) - [-p_o(v - v_o)] - T_o(s - s_o) \qquad (2)$$

where b is the specific exergy (or exergy for closed systems); u, v, and s are the specific internal energy, the specific volume, and the specific entropy, respectively; u_o, v_o, and s_o are the specific internal energy, specific volume, and specific entropy for the restricted dead state, respectively, and p_o and T_o are the pressure and temperature of the dead state, respectively. The dead state is defined as the conditions of the environment at a temperature of T_o and a pressure of p_o.

For the flow periods (open system), the flow exergy (or exergy for flows) b_f is given by

$$b_f = (h - h_o) - T_o(s - s_o) \qquad (3)$$

where h is the specific enthalpy, h_o and s_o are the specific enthalpy and specific entropy of the restricted dead state, respectively, and s is the specific entropy of the flowing matter. For flows out of the system, the flowing matter is the cylinder contents, and for flows into the system, the flowing matter must be specified.

Exergy is not a conserved property, and hence may be destroyed by irreversibilities such as heat transfer through a finite temperature difference, combustion, friction, or mixing processes. Between any end states, therefore, the change in the exergy may be related to the relevant processes:

$$\Delta B = B_{end} - B_{start}$$
$$\Delta B = B_{in} - B_{out} + B_Q - B_W - B_{dest} \qquad (4)$$

where ΔB is the change of the total system exergy for a process, B_{end} is the total exergy at the end of the period, B_{start} is the total exergy at the start of the period, B_{in} is the total exergy transferred into the system accompanying flow into the system, B_{out} is the exergy transferred out of the system accompanying flow out of the system, B_Q is the exergy transferred accompanying the heat transfer, B_W is the exergy transfer due to work, and B_{dest} is the exergy that is destroyed by irreversible processes. This relation may be used to ascertain the destruction of exergy by solving Equation 4 to find B_{dest}. That is,

$$B_{dest} = B_{start} - B_{end} + B_{in} - B_{out} + B_Q - B_W \qquad (5)$$

$$B_W = W - W_{surr} \qquad (6)$$

where the work done against the surroundings is given by

$$W_{surr} = p_o(V_{end} - V_{start}) \qquad (7)$$

For heat transfer, the exergy that is transferred out of the system is equal to the "available" portion of the heat transfer:

$$B_Q = \int \left(1 - \frac{T_o}{T}\right) \delta Q \qquad (8)$$

where B_Q is the available portion of the heat transfer, and δQ is the differential heat that is transferred at a system (boundary) temperature T. The exergy that transfers into the system (B_{in}) and out of the system (B_{out}) due to flows is given as follows:

$$B_i = \int (\dot{m}_i b_{f,i}) dt \qquad (9)$$

$$B_{fuel} = -(\Delta G)_{T_o, p_o} \qquad (10)$$

where $b_{f,i}$ is the specific flow exergy, and the subscript "i" refers to each individual flow [for this study, intake (in) or exhaust (out)].

For determining the engine efficiency, and for completing the energy and exergy balances, values are needed for the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

energy and exergy of the fuel. For the fuel, the lower heating value (LHV) evaluated for a constant pressure process is used. By definition, the fuel exergy (B_{fuel}) is given by the Gibbs free energy (Equation 10).

2.1.3 Example results

2.1.3.1 Engine specifications and operating conditions. The results provided next are for an automotive, 5.7-L V-8 engine with a bore and stroke of 101.6 and 88.4 mm, respectively. Table 1 lists the engine specifications. The combustion model used is based on the Wiebe function, which has been found to be representative of engine combustion (Wiebe, 1970). The Wiebe combustion parameters used here were recommended by Heywood (1988): $m = 2.0$ and $a = 5.0$.

A base-case operating condition has been selected for most of the following results. This condition is a part-load condition that is near the frequent operating conditions for many light-duty driving cycles, and includes an engine speed of 1400 rpm and an engine load specified by a brake mean effective pressure (BMEP) of 325 kPa. Table 2 lists examples of the values of other parameters (and "how obtained") that were needed in this work. The inlet pressure was determined for obtaining a BMEP of 325 kPa, and the combustion start was selected to maximize the brake torque (MBT). This section provides some examples of the results from the engine thermodynamic simulation, and these results include detailed time-resolved results and the overall performance results.

Table 1. Engine specifications.

Item	Value
Number of cylinders	8
Bore (mm)	101.6
Stroke (mm)	88.4
Crank rad/con. rod ratio	0.305
Inlet valves	
Diameter (mm)	50.8
Max lift (mm)	10.0
Opens (°CA aTDC)	357
Closes (°CA aTDC)	−136
Exhaust valves	
Diameter (mm)	39.6
Max lift (mm)	10.0
Opens (°CA aTDC)	116
Closes (°CA aTDC)	371
Valve overlap (degrees)	14°
Heat transfer multiplier	1.33

Table 2. Some engine and fuel input parameters (base case: 325 kPa, 1400 rpm).

Item	Value Used	How Obtained
Displaced volume (dm^3)	5.733	Computed
Compression ratio	8.0	Input
AF_{stoich}	15.07	Computed
Equivalence ratio	1.0	Input
Inlet (air–fuel) temperature (K)	319.3	Input
Inlet pressure (kPa)	51.2	Input
Exhaust pressure (kPa)	102.7	Input
Start of combustion (°bTDC)	20.0	Determined for MBT
Combustion duration (°CA)	60	Input
Cylinder wall temp. (K)	450	Input
Mech. frictional mep (kPa)	68.5	From algorithm (Sandoval and Heywood, 2003)
Fuel LHV (kJ/kg)	44,400	For *iso*-octane (Heywood, 1988)
Fuel exergy (kJ/kg)	45,670	For *iso*-octane (Heywood, 1988)

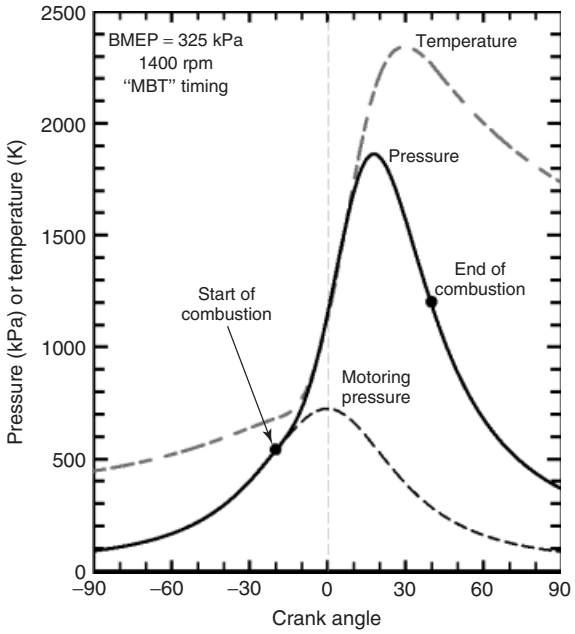


Figure 2. One-zone overall cylinder pressures and temperatures as functions of crank angle for the base case conditions.

2.1.3.2 Basic results. Figure 2 shows cylinder pressures and (one-zone) temperatures as functions of the crank angle for the base case conditions. Both the motoring and firing pressures are included. The start and end of combustion are indicated for reference. The motoring pressure

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is nearly symmetrical about top dead center (TDC) (0° crank angle) with a maximum pressure of about 720 kPa. The firing pressure has a maximum of 1863 kPa at 17.5° after top dead center (aTDC). The one-zone (average of all zones) cylinder gas temperature increases rapidly during the combustion period and reaches a maximum of 2341 K at 29.2° aTDC.

Figure 3 shows the average overall (one-zone) gas temperature (T_{avg}) from Figure 2, and the temperatures associated with the zones defined after combustion begins as functions of crank angle for the base-case conditions. The latter zones include the unburned zone (T_u), the burned zone (T_b), adiabatic core (T_a), and the boundary layer (T_{bl}) zones. The adiabatic core and the boundary layer combine to form the burned zone. At the bottom of the figure is the unburned zone temperature, which remains below about 870 K. The burned zone gas temperature is the energy-averaged temperature of the boundary layer and adiabatic zones. The adiabatic zone has the highest temperatures, and for this case a maximum of about 2600 K is attained. The slightly higher adiabatic zone temperature compared to the burned zone temperature is particularly important for nitric oxide concentration predictions, which are highly temperature dependent (see NO_x Formation and Models). After combustion, the burned-zone temperature and the average temperature are the same. At any crank angle, the average (one-zone) temperature represents the energy-averaged temperature for all zones.

Figure 4 shows the mass flow rates as functions of crank angle (where 0° crank was defined in Figures 2 and 3) for both the intake (solid curve) and exhaust (dashed curve) flows for the base-case conditions. The positive flow rates are the flow rates in the natural directions: the intake flow is positive into the cylinder and the exhaust flow is positive out of the cylinder. The "blow-down" after the exhaust valve opens (EVO) attains a maximum flow of about 42.6 g/s, and then the exhaust flow decreases. The displacement phase of the exhaust flow, which is caused largely by the piston motion, results in a maximum of about 17 g/s at a crank angle of 300°. The flow is back into the cylinder during the valve overlap period—the time between the intake valve open (IVO) and the exhaust valve close (EVC) times. The intake flow begins the flow into the intake manifold during the valve overlap period. The flow reverses shortly after the EVCs. The first matter into the cylinder is the matter that previously flowed into the intake manifold. Once that matter has returned, then the fresh air and fuel vapor mixture flows into the cylinder. The intake flow attains a maximum flow rate of 32 g/s and then decreases. At the end of the IVO period, the flow reverses back into the manifold for a short period.

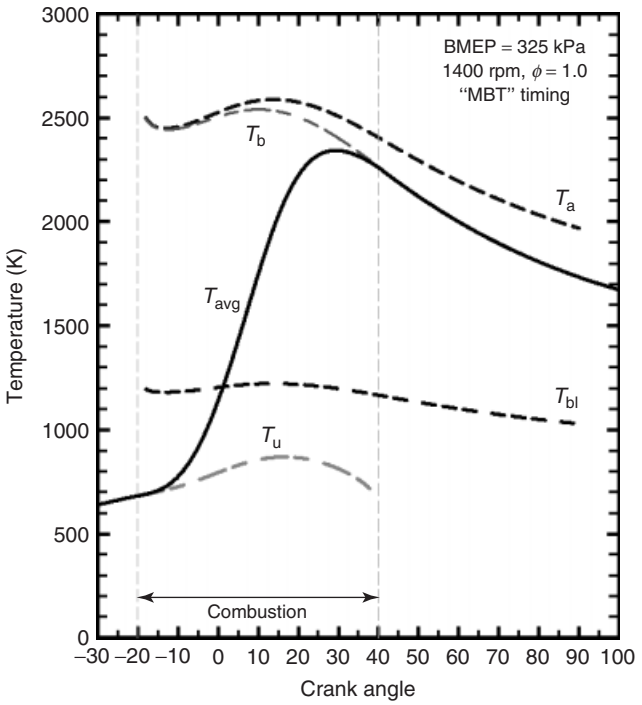


Figure 3. Multiple-zone cylinder gas temperatures as functions of crank angle for the base-case conditions.

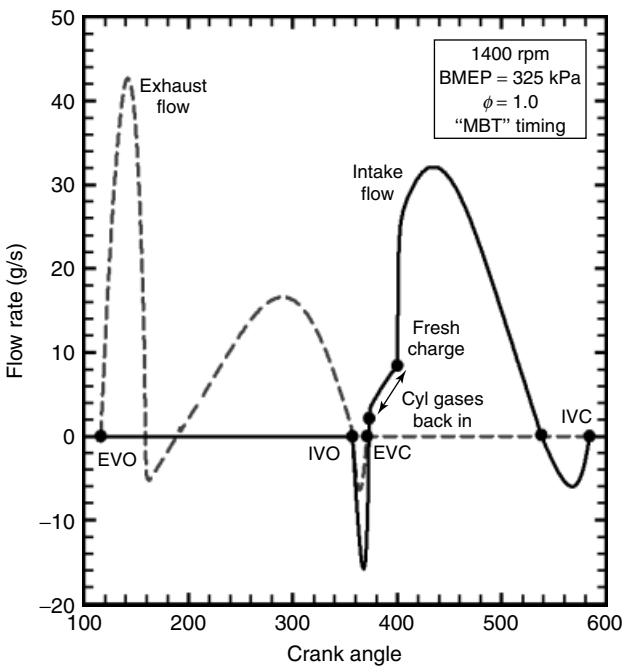


Figure 4. Exhaust (dashed curve) and intake (solid curve) mass flow rates as functions of crank angle for the base-case conditions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

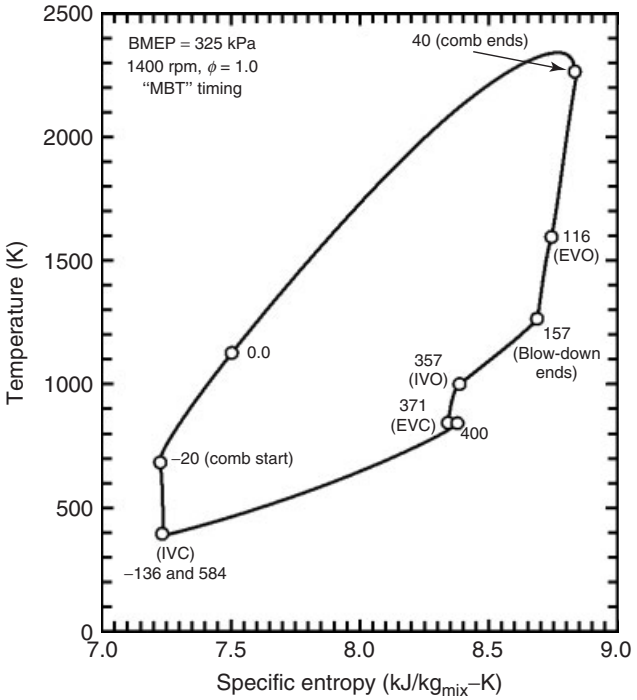


Figure 5. The specific enthalpy for the mixtures as functions of the one-zone overall cylinder gas temperature for the base-case conditions.

Table 3. Results for the base case.

Item	Value[b]	
	Indicated (net)[a]	Brake
mep (kPa)	393.7	325.0
sfc (g/kWh)	264.8	320.7
η (%)	30.6	25.3
Torque (N·m)	179.3	148.1
Power (kW)	26.3	21.9
p_{peak} (kPa)	1863	
CA of p_{peak}	17.5	
Max. T_b (K)	2540	
CA of max. T_b	10.0	
T_{peak} (K)	2341	
T_{exh} (K)	1252	
$\dot{m}_{fuel}$ (g/s)	1.94	
$\dot{m}_{air}$ (g/s)	29.2	
Residual fraction	0.109	
Energy distribution		
Brake work (%)	25.28	
Friction (%)	5.34	
Heat loss (%)	28.51	
Exhaust (%)	40.20	
Unused (%)	0.67	
Total (%)	100.0	

[a]Net indicated is for all four strokes.
[b]Results for $T_{switch} = 1200\,K$ and $\Delta CA = 0.25^\circ$.

Figure 5 shows the average one-zone gas temperature as a function of the specific entropy for the base-case conditions. Specific crank angle "times" are denoted in the figure so that the complete process is shown from intake valve close, through compression, combustion, expansion, exhaust, intake, and back to intake valve close. The temperature increases at nearly constant entropy during the compression process, and then the temperature and entropy increase during the combustion process. The increase of entropy during the combustion process is indicative of the highly irreversible combustion process. Once combustion ends, the temperature decreases rapidly and the entropy decreases slightly during the expansion process. Once the "blow-down" portion of the exhaust period ends, the temperature decreases less rapidly and the entropy decreases a little more rapidly. Once fresh charge starts entering the cylinder, the temperature and entropy continue to decrease until they attain their values at intake valve close.

Although the entropy results are not needed for routine engine performance evaluations, these results are required for determining the exergy values. As described earlier, computing exergy requires a knowledge of entropy. The exergy results are provided later.

2.1.3.3 Engine performance results. Once the cylinder pressure is known as a *function of cylinder volume*, the engine performance may be obtained. The area within the pressure–volume diagram is proportional to the engine work output. From the work output, all the other engine performance and efficiency parameters may be determined using standard relationships (Heywood, 1988).

Table 3 lists results from the simulation for the base condition. The net indicated and brake thermal efficiencies were 30.6% and 25.3%, respectively. The associated net indicated and brake specific fuel consumption values were 264.8 and 320.7 g/kWh, respectively. The energy distribution indicates that about 28.5% and 40.2% of the fuel energy was assigned to heat transfer and exhaust gas energy, respectively. Additional results are listed that serve to illustrate the type of detail that is possible from these simulations. These additional results include peak cylinder pressures and temperatures, average exhaust gas temperatures, average overall fuel and airflow rates, and the residual fraction. Other such results are described below.

Table 4 is a summary of the energy and exergy values for the base case. These values are given both as a percentage of the fuel energy and exergy value, and as the actual value. The fuel energy is divided among work, friction,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Summary of energy and exergy values for base case.

Item	Energy		Exergy	
	Value (kJ)	Percent (%)	Value (kJ)	Percent (%)
Brake work	0.2326	25.28	0.2328	24.66
Friction	0.0491	5.34	0.0492	5.21
Net indicated work	0.2817	30.62	0.2820	29.87
Heat loss	0.2626	28.51	0.2165	22.93
Net flowout	0.3703	40.20	0.2284	24.20
Destruction (comb.)	n/a	n/a	0.1979	20.97
Destruction (intake)	n/a	n/a	0.0128	1.36
Fuel not used	0.0062	0.67	0.0064	0.68
Total	0.9213	100.0	0.9440	100.0

heat transfer, exhaust, and unburned fuel. The fuel exergy is divided among the same items plus destruction during the combustion processes and intake mixing processes. Further comments on these distributions are provided next.

Figure 6 shows a "pie" chart of the energy (a) and exergy (b) distributions for the base-case conditions. The top half of the charts contains the energy and exergy inputs, and the bottom half of the charts contains the outputs. Inputs are from the fuel and the pumping work. The energy and exergy inputs are the fuel energy (or exergy) and the pumping work. The energy outputs are the gross indicated work, heat transfer, net exhaust gas energy, and unburned fuel. The gross indicated work consists of the brake work, mechanical friction, and pumping work. For this operating condition, significant energy is associated with the heat transfer and exhaust gas flow.

The right side of Figure 6 shows the exergy distribution. The exergy inputs are similar to the energy inputs. For the exergy outputs, since the work is exergy, the work values are the same as for the energy chart. The heat transfer and exhaust exergy are smaller portions (compared to the energy portions) since not all of the energy is equivalent to exergy. The new item for the exergy distribution is the destruction terms and especially the destruction during the combustion process, which for this case is about 21% of the fuel exergy.

The next three figures show results for engine performance as functions of operating and design variables. Figure 7 shows percentage of the fuel energy distributed among the various energy output items as functions of the engine load. Engine load is expressed in terms of BMEP. Brake work increases as load increases from idle to peak load. As plotted, the brake work percentage is actually equivalent to the brake thermal efficiency since these items are based on the fuel input energy (LHV). The relative heat transfer decreases as load increases (note that the actual heat transfer is increasing but at a slower rate than the fueling rate increase). Finally, the exhaust gas energy percentage increases as the engine load increases. These results confirm the notion that operating the engine closer to full load improves brake efficiency because of lower relative losses.

As examples of using a thermodynamic cycle simulation to study engine design variables, the following results examine changes in compression ratio (Figure 8) and exhaust gas recirculation (EGR) levels (Figure 9). Figure 8 shows the relative energy (percentage of fuel input energy) as functions of compression ratio for the base-case conditions (Caton, 2007). The brake work increases rapidly as the compression ratio increases to about 10, and then the increases are much more modest as the compression ratio increases to 20. The friction work increases slightly with increases of compression ratio primarily because of the

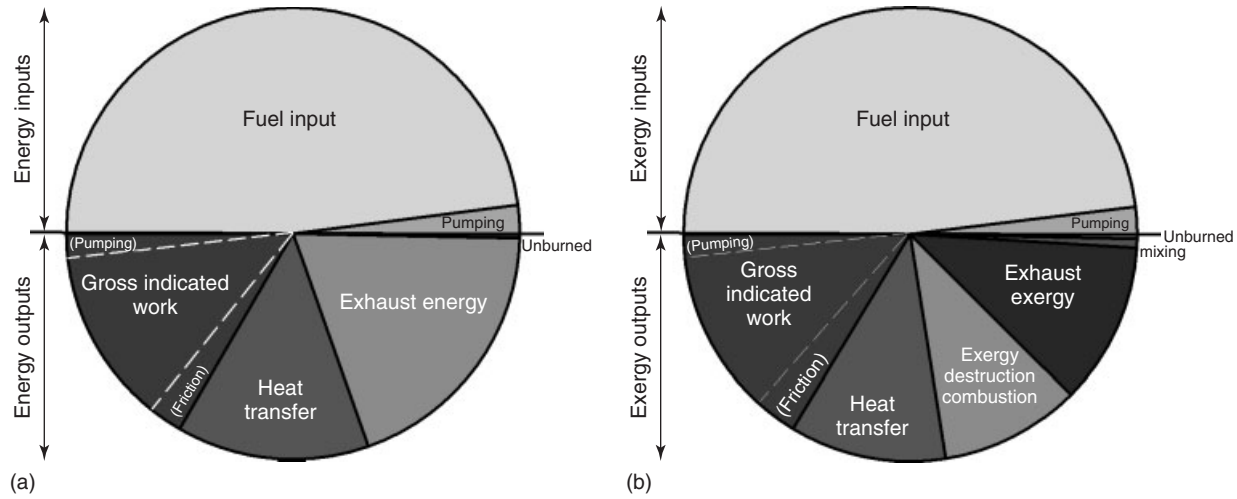


Figure 6. Distribution of the (a) energy and (b) exergy inputs and outputs for the base-case conditions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

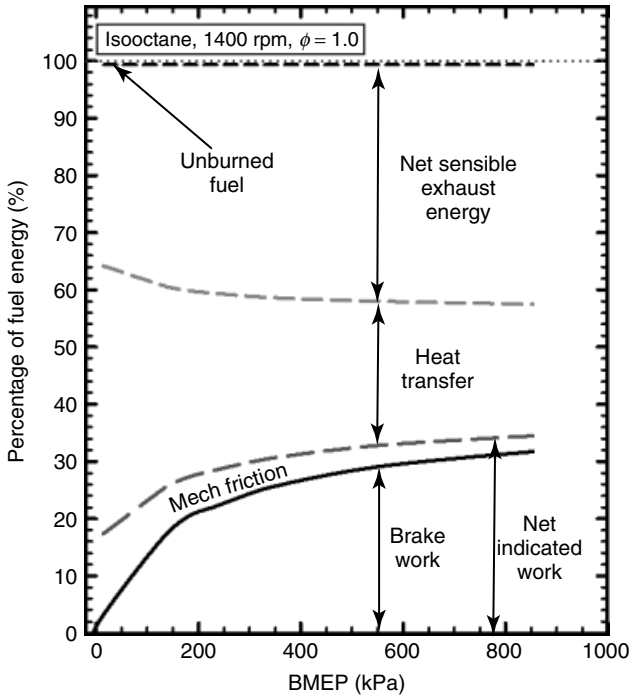


Figure 7. The percentage of the fuel energy for each item for 1400 rpm as a function of load for the base-case conditions.

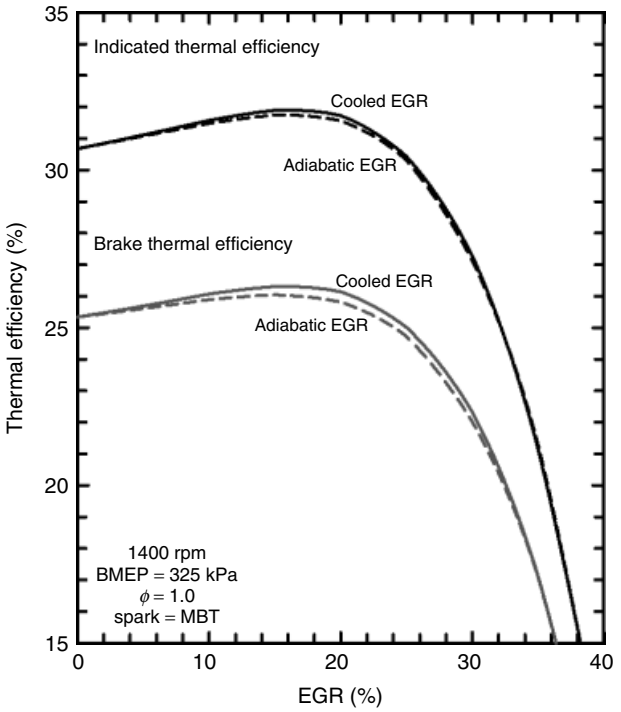


Figure 9. The net indicated and brake thermal efficiencies as functions of EGR for the base-case conditions.

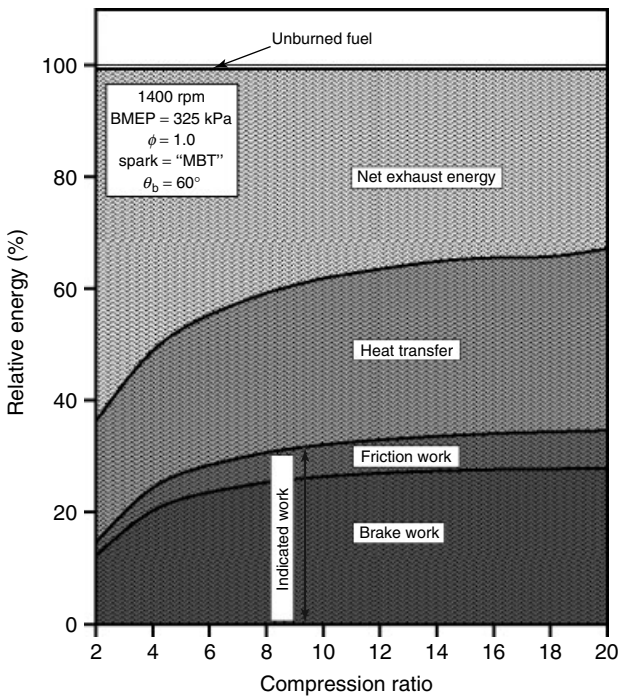


Figure 8. The percentage of the fuel energy for each item for 1400 rpm as a function of compression ratio for the base-case conditions.

higher cylinder pressures. The accuracy of the friction model plays a role in the exact details of this trade-off. The relative heat transfer increases as the compression ratio increases because of higher peak temperatures (even though temperatures during the expansion stroke are lower), higher pressures, and lower fuel consumption (higher efficiencies). Finally, the exhaust gas energy percentage decreases since work and heat transfer increase.

Figure 9 shows the net indicated and brake thermal efficiencies as functions of EGR level for both a cooled and an adiabatic EGR system (Shyani and Caton, 2009). An actual EGR system would be expected to have performance that is in between those of the cooled and adiabatic EGR systems. The efficiencies increase as EGR increases up to an EGR level of about 16%. For higher EGR levels, the efficiencies rapidly decrease because of increasing misfires and eventually incomplete combustion. The increase of the efficiencies for the 0–16% EGR range is largely due to decreasing heat transfer, lower temperatures, and favorable thermodynamics (largely the ratio of specific heats, or "gamma").

2.1.3.4 Low temperature combustion engines. A current example of the utility of these thermodynamic simulations is the assessment of the thermodynamics of low temperature combustion (LTC) engines. Examples of LTC engines are diesel LTC engines (discussed in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Diesel and Diesel LTC Combustion) and homogeneous charge, compression ignition (HCCI) engines (discussed in Advanced Compression-Ignition Combustion for Ultra-Low NO_x and Soot). These engines have been shown to not only possess low emissions but also have significantly increased thermal efficiencies for some operating conditions. These new combustion modes involve various combinations of stratification, lean mixtures, high levels of EGR, high compression ratios, multiple injections, variable valve timings, two fuels, and other such features. A thermodynamic simulation has been used (Caton, 2010) to determine the fundamental thermodynamic reasons for the increases of the thermal efficiencies and to quantify these effects. Results were compared with similar experimental studies (Kokjohn *et al*., 2009), and these comparisons provided confidence that the dominant thermodynamic processes were captured correctly.

By systematically considering a number of the above features in a step-by-step manner, the impact of each feature was quantified. Five specific features were examined: the compression ratio increased from 8 to 20; the burn duration shortened from 60 to 30°CA; the equivalence ratio decreased from 1.0 to 0.7; the EGR level increased from 0% to 50%; and the cylinder wall temperature increased from 450 to 550 K. Each of these changes increased the thermal efficiencies. The greatest increases were obtained from the compression ratio increase, the leaner operation, and the increase in the EGR level. The brake thermal efficiency increased from 32.7% to 41.6% for a 900-kPa BMEP, 2000-rpm operation. Similar improvements were noted for a lower load case.

The thermodynamic simulation provided the following insights concerning the reasons for these improvements. The low temperatures due to the dilute charge were responsible for lower heat losses and higher ratios of the specific heats (gamma). The higher gammas resulted in more effective conversion of the thermal energy into work. The dilute charge operation, however, required higher inlet pressures to produce the 900 BMEP and this resulted in more friction that mitigated some of the thermodynamic gains noted for the net indicated thermal efficiency. Complete details of this study are available elsewhere (Caton, 2010).

2.1.3.5 Overexpanded engine designs. A final example of how these engine cycle simulations may be used is in evaluating engine concepts. An overexpanded engine design is one such concept. (This concept should not be confused with the "Miller cycle" concept that uses valve timing to modify the effective compression ratio relative to the expansion ratio.) The advantages of a longer expansion stroke compared to the compression stroke have been recognized from the very beginning of engine development. The overall cycle for these designs is often called an *Atkinson cycle*. The potential benefits of these overexpanded designs are frequently illustrated with simple air-standard cycle evaluations, but these analyses are too simple and miss some of the important features. The following results show how thermodynamic engine cycle simulations can be used to evaluate such a concept on a quantified basis. Complete details of this study are provided elsewhere (Caton, 2007).

The basic idea was to modify the conventional engine kinematics so as to allow a greater expansion stroke compared to the compression stroke. The engine performance parameters (such as BMEP) are based on the standard engine configuration using the compression stroke to determine the displaced volume. A wide range of compression and expansion ratios have been assumed. These may be outside practical ranges, but the intention was to illustrate the major trends more completely. A multiplier factor for the expansion stroke relative to the compression stroke was defined. For example, if an expansion ratio of 15 is desired and the compression ratio is 10, then the expansion stroke must be increased by a factor of 1.5 relative to the compression stroke.

For the part-load computations, the BMEP was held constant (=325 kPa). So, as the brake efficiency increased, the inlet pressure reduced. For each compression ratio case, as the expansion ratio increased, the thermal efficiency first increased, reached a maximum, and then decreased. The decrease of the efficiency after the maximum value was due to increased heat losses, increased friction, and ineffective exhaust processes (due to the reduced cylinder pressure at the time of exhaust valve opening). The expansion ratio for the highest efficiency appears to be equal to about the compression ratio plus 3. The use of this expansion ratio provides a modest thermal efficiency increase of about 1% absolute (e.g., from 24% to 25%).

For the wide-open throttle (WOT) computations, the inlet pressure remained at 95 kPa and the BMEP (load) was not constant. For WOT, the higher expansion ratios provided significant gains. For example, for a compression ratio of 10, expansion ratios of 10 (conventional engine) and 30 provided brake thermal efficiencies of about 34% and 43%, respectively. Although the net thermodynamic gains are significant, large expansion ratios (such as 30) may not be practical in most applications. Also, for many applications, WOT is not a frequent operating condition and the expense and complexity of an overexpanded engine design may not be justified.

In summary, the use of an overexpanded engine design is most effective for increasing an engine's efficiency at WOT conditions. For these cases, a large expansion ratio is needed to achieve the maximum efficiencies, and these large

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

expansion ratios may not be practical for most applications. Further, the efficiency gains are mitigated for part-load operation because of increased pumping losses associated with the constant load requirement. For all load conditions, the maximum efficiency was found for a specific expansion ratio for each compression ratio. Additional complexity and friction are further reasons why the overexpanded engine design is not more attractive than conventional designs.

3 COMPRESSION-IGNITION (DIESEL) ENGINE SIMULATIONS

This section will briefly describe the use of thermodynamic simulations for CI (diesel) engines. As mentioned above, much of the formulation is similar to that of the SI engine; the main difference is that the conventional diesel engine possesses a highly stratified combustion process that results in highly nonuniform temperatures and equivalence ratios. This may lead to significant problems with respect to considerations of ignition, combustion, heat transfer, and emissions. On the other hand, with good submodels, the overall performance predicted will still be consistent with the thermodynamics.

Historically, CI engines were the first to be modeled using these thermodynamic simulations (e.g., Heywood, 1988). The most important difference from the SI simulations is in the treatment of the heat release. This has been accomplished with empirical models that attempt to capture the various stages of combustion such as premixed and diffusion. More modern approaches have used models of the fuel jet.

3.1 Single-zone models

Single-zone models assume that the cylinder contents are uniform and may be represented by the average property values. Much like with the SI engine, the heat release may be prescribed. Because of the longer duration and different characteristics of the combustion process for the CI engine, alternative models are often used. For example, Watson, Pilley, and Marzouk (1980) described an algebraic function that incorporated two stages: a rapid premixed phase, followed by a slower mixing controlled burning phase. Many such models exist in the literature. The disadvantage is that these types of empirical models may not be correct for different operating conditions or for different engines.

3.2 Fuel-jet-based models

Fuel-jet-based models are attempts to describe the combustion process in a more fundamental way while still retaining the simplicity of the overall one-zone thermodynamics. In this case, the progress of the fuel spray is modeled by allowing "packets" to form as a function of time and for different "locations" in the fuel spray. For this simple simulation, the different "locations" are conceptual and cannot be related to actual spatial coordinates. However, this approach does allow different packets to have different equivalence ratios, different temperatures, and so forth. The total heat release is obtained by summing the heat release from the individual packets. Although this approach appears to be more realistic, items such as the entrainment into each packet is still somewhat arbitrary. Ultimately, these types of models are only as good as the experimental verification, much like the single-zone models described first.

The usefulness of these phenomenological models relies heavily on accurate information and knowledge of the fuel spray, fuel/air mixing, autoignition, and reaction. Development of advanced laser-based diagnostics has provided an abundance of new information and important knowledge concerning the reacting diesel fuel jet development (Espey *et al*., 1994; Dec and Espey, 1992). Dec (1997) proposed a conceptual model by combining these results, and provided a detailed understanding of the temporal and spatial evolution of a reacting diesel fuel jet. This new model has significant differences from old descriptions, and offers completely new insight and a clear picture of how diesel combustion proceeds.

This relatively newer conceptual model has promoted the development of newer and more advanced phenomenological models. Maiboom *et al*. (2009) developed a new phenomenological combustion model based on Dec's conceptual model. They were able to predict relatively accurate heat release rate and provide some important local information. This work did not include chemistry and thus was not able to calculate LTC conditions. In addition, the prediction of pollutant emissions was also not included. Asay (2003) and Ebrahimi, Bazargan, and Jazayeri (2007) created five-zone diesel cycle simulation models with the knowledge of the internal structure of direct-injection diesel fuel jets as well as empirical correlations predicting jet development. This model included all the basic features of diesel combustion. In particular, they extended the chemical equilibrium model to include 21 species, which allowed the program to be valid for equivalence ratios up to at least 8.0 (λ of 0.125). The downside of this model was the lack of some important submodels such as the lift-off length, the gas entrainment after combustion (incomplete combustion products oxidation), and the allocation of heat transfer. The lack of this information resulted in some inaccuracy of local thermodynamic properties as well as the heat release rate.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4 SUMMARY

This chapter has focused on the uses and limitations of zero- and quasi-one-dimensional thermodynamic engine cycle simulations with respect to understanding and development of engines. After a general description of the key items that comprise such a simulation, basic time-resolved cylinder pressures and temperatures and overall engine performance parameters were presented for a part-load operating condition. The temperature as a function of entropy, as well as the distribution of energy and exergy for the base case, was shown. Other example results included the percentage of the fuel energy as functions of load (BMEP) and compression ratio, and the thermal efficiencies as functions of the EGR level. The use of these simulations for LTC engines was briefly described. The use of these simulations to evaluate engine concepts was illustrated by examining an overexpanded engine concept. Finally, the application to CI engines was briefly described.

ACKNOWLEDGMENTS

Over the years, this author and this work have been supported by a number of sponsors. Recent sponsors have included the U.S. Department of Energy and the Oak Ridge National Laboratory. The contents of this chapter, however, do not necessarily reflect the opinions or views of these sponsors or individuals.

REFERENCES

Asay, R.J. (2003) A five-zone model for direct injection diesel combustion. Master Thesis. Department of Mechanical Engineering, Brigham Young University.

Blumberg, P.N., Lavoie, G.A., and Tabaczynski, R.J. (1979) Phenomenological models for reciprocating internal combustion engines. *Progress in Energy and Combustion Science*, **5**, 123–167.

Caton, J.A. (2001), A multiple-zone cycle simulation for spark-ignition engines: thermodynamic details, in Large-Bore Engines, Fuel Effects, Homogeneous Charge Compression Ignition, Engine Performance and Simulation, vol. 2, ICE-Vol. 37–2, *Proceedings of the 2001 Fall Technical Conference*, (ed V.W. Wong), *the ASME Internal Combustion Engine Division*, American Society of Mechanical Engineers, Paper No. 2001-ICE-412, pp. 41–58, Argonne, IL, September 23–26, 2001.

Caton, J.A. (2003) A cycle simulation including the second law of thermodynamics for a spark-ignition engine: implications of the use of multiple-zones for combustion. *Transactions of the Society of Automotive Engineers: Journal of Engines*, **111–113**, Paper No. 2002–01–0007, 281–299.

Caton, J.A. (2005) Use of a cycle simulation incorporating the second law of thermodynamics: results for spark-ignition engines using oxygen enriched combustion air, 2005. *SAE International Congress and Exposition*, Society of Automotive Engineers, Cobo Hall, Detroit, MI. SAE Paper No. 2005-01-1130, 11–14 April 2005.

Caton, J.A. (2006) First and Second Law Analyses of a Spark-Ignition Engine Using Either Isooctane or Hydrogen. *Proceedings of the 2006 Fall Conference of the ASME Internal Combustion Engine Division*, November 5–8, 2006, Sacramento, CA.

Caton, J.A. (2007) The Effects of Compression Ratio and Expansion Ratio on Engine Performance Including the Second Law of Thermodynamics: Results from a Cycle Simulation. *Proceedings of the 2007 Fall Conference of the ASME Internal Combustion Engine Division*, October 14–17, 2007, Charleston, SC.

Caton, J.A. (2010) An Assessment of the Thermodynamics Associated with High-Efficiency Engines. *Proceedings of the 2010 Fall Technical Conference of the ASME Internal Combustion Division*. Paper No. ICEF2010–35037, September 12–14, 2010, San Antonio, TX.

Dec, J.E. (1997) A conceptual model of DI diesel combustion based on laser-sheet imaging. Society of Automotive Engineers, SAE Paper No. 970873.

Dec, J.E. and Espey, C. (1992) Soot and fuel distributions in a D.I. diesel engine via 2-D imaging. Society of Engineers, SAE Paper No. 922307.

Ebrahimi, K., Bazargan, M., and Jazayeri, S.A. (2007) A new phenomenological model for combustion and performance studies of direct injection diesel engine. Society of Automotive Engineers. SAE Paper No. 2007-01-1904.

Espey, C., Dec, J.E., Litzinger, T.A., and Santavicca, D.A. (1994) Quantitative 2-D fuel vapor concentration imaging in a firing D.I. diesel engine using planar laser-induced Rayleigh scattering. Society of Automotive Engineers, SAE Paper No. 940682.

Foster, D. (1985) An overview of zero-dimensional thermodynamic models for IC engine data analysis. Society of Automotive Engineers. SAE Paper No. 852070.

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*, McGraw-Hill book Company, New York, NY.

Heywood, J.B., Higgins, J.M., Watts, P.A., and Tabaczynski, R.J. (1979) Development and use of a cycle simulation to predict SI engine efficiency and NO_x emissions. Society of Automotive Engineers, SAE Paper No. 790291.

Kokjohn, S.L., Hanson, R.M., Splitter, D.A., and Reitz, R.D. (2009) Experiments and modeling of dual-fuel HCCI and PCCI combustion using in-cylinder fuel blending. Society of Automotive Engineers, SAE Paper No. 2009-01-2647.

Maiboom, A., Tauzia, X., Shah, S.R., and Hetet, J. (2009) New phenomenological six-zone combustion model for direct-injection diesel engines. *Energy and Fuels*, **23**, 690–703.

Mattavi, J.N. and Amann, C.A. (eds) (1980) *Combustion Modeling in Reciprocating Engines*, Plenum Press, New York.

Olikara, C. and Borman, G.L. (1975) A computer program for calculating properties of equilibrium combustion products with some applications to I. C. engines. Society of Automotive Engineers, SAE Paper No. 750468.

Patterson, D.J. and van Wylen, G. (1964) A digital computer simulation for spark-ignited engine cycles, in *Digital Calculations of*

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Engine Cycles, SAE Progress in Technology, vol. 7 Society of Automotive Engineers New York, NY.

Sandoval, D. and Heywood, J.B. (2003) An improved friction model for spark-ignition engines. Society of Automotive Engineers, SAE Paper No. 2003–01–0725.

Sherman, R.H. and Blumberg, P.N. (1977) The influence of induction and exhaust processes on emissions and fuel consumption in the spark ignited engine. Society of Automotive Engineers, SAE Paper No. 770880.

Shyani, R.G. and Caton, J.A. (2009) A thermodynamic analysis of the use of EGR in SI engines including the second law of thermodynamics. *Proceedings of the Institution of Mechanical Engineers, Part D, Journal of Automobile Engineering*, **223** (1), 131–149.

Watson, N., Pilley, A.D., and Marzouk, M. (1980) A combustion correlation for diesel engine simulation. SAE Paper No. 800029.

Watts, P.A. and Heywood, J.B. (1980) Simulation studies of the effects of turbocharging and reduced heat transfer on spark-ignition engine operation. Society of Automotive Engineers, SAE Paper No. 800289.

Wiebe, J.J. (1970) *Brennverlauf und Kreisprozess von Verbrennungsmotoren*, VEB-Verlag Technik, Berlin.

Woschni, G. (1968) A universally applicable equation for the instantaneous heat transfer coefficient in the internal combustion engine. SAE Transactions, SAE Paper No. 670931, vol. 76, pp. 3065–3083.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto123
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Multidimensional Simulation

Rolf D. Reitz and Christopher J. Rutland

University of Wisconsin-Madison, Madison, WI, USA

1 Introduction 1
2 Turbulence Models 4
3 Physical and Chemical Submodels 6
4 High Fidelity LES Combustion Models 15
References 16

1 INTRODUCTION

The combustion process in IC engines is characterized by complex multiphase, turbulent reacting flows. In addition, IC engine combustion spans multiple regimes that cover premixed flame propagation, mixing-controlled burning, and chemical-kinetics-controlled processes (Haworth, 2005). Modeling IC engines requires that these physical and chemical processes be described using mathematical models and accurate numerical schemes, as reviewed by Reitz and Sun (2009). The goal is that model predictions can be used to reduce the need for expensive laboratory engine testing and for efficient engine design optimization. The integration of engine modeling with optimization methods has been described by Shi, Ge, and Reitz (2011).

Zero-dimensional (0D) IC engine models based on thermodynamic cycle analysis were introduced in the mid-1900s, and are still useful for interpreting measured engine data (e.g., Zero- and One-Dimensional Methodologies and Tools, Herold *et al*., 2011). These models use simple mathematical formulae such as the Wiebe function (Wiebe, 1956, 1962; Ghojel, 2010) to specify the fuel burning rate. 0D engine heat transfer models are also based on empirical models, such as that of Woschni (1967). However, these empirical models have been shown to require extensive tuning, especially when applied to new engine concepts such as low temperature combustion (LTC) (e.g., Caton, 2011).

Multidimensional computational fluid dynamics (CFD) modeling promises to reduce the amount of model calibration effort required in engine simulation. Instead of describing engine processes empirically, partial differential conservation equations are solved on a finite difference numerical mesh that represents the time-varying engine geometry and the evolution of the combustion process. Submodels are still required to describe processes that occur on time and length scales too small to be resolved in the simulations, but these models have more "universality," as they refer to "microscale" processes that are common to general flows. For example, even though wall boundary layers may be too thin to be resolved in a practical engine CFD mesh, adequate wall heat transfer predictions can be obtained using "law-of-the-wall" submodels that use local information about the mean flow in computational cells at the combustion chamber walls. These submodels are often derived from experiments conducted in more controlled environments than engines, such as in steady pipe flow.

1.1 Multidimensional modeling

The use of multidimensional CFD for engine modeling began to displace simpler 0D models as computers became more and more powerful starting in the 1980s. However, engine CFD modeling was still not generally applied for engine design, as computer capacity was limited and engine submodels were still under development. Instead, in that era, 0D engine modeling was supplemented by

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

phenomenological models to extend the predictive capability of models. These phenomenological models include the quasi-steady spray models of Hiroyasu, Kadota, and Arai (1978), and the soot and NO formation models of Hiroyasu and Kodota (1976), which are still widely used today for rapid trend analysis. Engine phenomenological models for engine applications are reviewed by Lakshminarayanan and Aghav (2010).

A significant step in the progress of multidimensional modeling was the release in the 1980s of engine CFD codes by the Los Alamos National Laboratory, eventually culminating in the widely used open-source KIVA codes (Amsden *et al*., 1985; Amsden, O'Rourke, and Butler, 1989; Amsden, 1993, 1997). This was accompanied by the release of several commercial engine codes, such as Computational Dynamics' Star-CD, AVL's FIRE and Ricardo's VECTIS code. These codes included the capability to represent moving meshes for the piston and valves, and they provided submodels to handle turbulent compressible flows with spray and droplet evaporation, and with fuel combustion chemistry. In a 1995 review, Reitz and Rutland (1995) concluded that multidimensional CFD modeling was able to match experimental engine pressure traces and heat release rates over wide ranges of conditions, and that good quantitative agreements in nitric oxide (NO*x*) and soot emissions were also attainable. Relatively simple combustion chemistry models, such as those based on the Shell Ignition model (Halstead, Kirsh, and Quinn, 1977), were integrated in these CFD codes to describe the combustion process (e.g., Kong, Han, and Reitz, 1995). However, the inclusion of more and more detailed and realistic fuel chemistry models was made possible by the early 2000s with further increases in computer power (e.g., Kong *et al*., 2001). Significant advances have since been made in spray submodel development, as reviewed by Shi, Ge, and Reitz (2011). These improvements include the development of more accurate atomization, secondary drop breakup, collision and coalescence, vaporization and spray-wall impingement models (see also Fuel Introduction: Fuel Introduction). In addition, significant effort has been placed on the development of models that provide reasonably grid-size and time-step independent predictions (e.g., Abani and Reitz, 2010).

Although today's multidimensional engine modeling codes still have quantitative uncertainties, it should be recognized that often the input to codes is itself uncertain, and this can greatly affect predictions (e.g., Yi *et al*., 2000). For example, injector nozzle-hole diameters can have a dramatic effect on pollutant emission predictions. However, typically, only nominal nozzle-hole sizes are specified, and these diameters can even change during engine use because of erosion and cavitation effects. In addition, owing to the hostile engine environment, validation experiments are difficult to perform, and measurement accuracy is affected by compromises that are required in setting up experiments that provide access to the combustion process or real engine phenomena such as cycle-by-cycle variability. In spite of these difficulties, it is now widely recognized that engine CFD model simulations can offer significant advantages to supplement experimental measurements in the engine development process by providing detailed in-cylinder information, which is normally not available or is inaccessible in experiments. Other advantages over engine experiments include lower cost, the ability to explore wider ranges of parameter variation, and the ability to separate and monitor individual physical and chemical processes. In addition, when coupled with optimization tools, previously unimagined engine concepts can be discovered using multidimensional modeling.

Shi, Ge, and Reitz (2011) provide examples that demonstrate that current multidimensional CFD tools are mature enough to guide the development of more efficient and cleaner internal combustion engines. For example, modeling has been applied to study new LTC concepts, such as homogeneous charge compression ignition (HCCI), premixed charge compression ignition (PCCI), and reactivity-controlled compression ignition (RCCI), that offer the promise of dramatically improved engine efficiencies. Optimized dual-fuel RCCI operation (port injection of gasoline together with optimized in-cylinder multiple diesel fuel injections) was discovered using CFD modeling by Kokjohn *et al*. (2011a). The computer simulations predicted high efficiency, low emission operation with excellent combustion phasing control at high and low engine loads without excessive rates of pressure rise. Subsequent engine experiments confirmed the model predictions, and demonstrated that US EPA 2010 NO*x* and soot emissions mandates can be met in-cylinder without after-treatment, while achieving up to 57% gross indicated thermal efficiency. The model and experiment comparisons for RCCI are presented in more detail in Section 3.2.1.

1.1.1 Governing equations

The basic governing equations for the multiphase flows in IC engines are the continuity (mass conservation), momentum (Navier–Stokes equations), energy, turbulence equations (normally the k–ε equations), and equations of state (ideal gas law). The gas mixture consists of multiple species, and the continuity equation for species k is

$$\frac{\partial(\rho Y_k)}{\partial t} + \frac{\partial(\rho Y_k U_j)}{\partial x_j} = \frac{\partial}{\partial x_j}\left(\rho D \frac{\partial Y_k}{\partial x_j}\right) + \dot{\omega}_k + \dot{\rho}_k^{\mathrm{s}} \quad (1)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where Y_k is the mass fraction of species k, ρ the total mass density of the mixture, and U is the fluid velocity, t is time and x_j ($j = 1, 2, 3$) are Cartesian space coordinates. D is the single turbulent diffusion coefficient with the assumption of Fick's law diffusion. $\dot{\omega}_k$ and ${\rho_k}^{\mathrm{s}}$ are source terms due to chemical reaction and spray evaporation/condensation, respectively. Summing Equation 1 over all species yields the continuity equation for the gas phase:

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho U_j)}{\partial x_j} = \dot{\rho}^{\mathrm{s}} \tag{2}$$

where ρ^{s} is the sum of all species spray evaporation/condensation terms.

The momentum equations for the gas-phase are

$$\frac{\partial (\rho U_i)}{\partial t} + \frac{\partial (\rho U_i U_j)}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial \tau_{ij}}{\partial x_j} + F_i^{\mathrm{s}} + F_i^{b} \tag{3}$$

where p is the pressure and τ is the viscous stress tensor:

$$\tau_{ij} = \mu \left(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i} - \frac{2}{3} \frac{\partial U_{\mathrm{k}}}{\partial x_{\mathrm{k}}} \delta_{ij} \right) \tag{4}$$

where δ is the Kronecker delta, and ${F_i}^{\mathrm{s}}$ is a source term due to spray drop drag forces. ${F_i}^{\mathrm{b}}$ is the gravitation body force equal to ρg.

The energy conservation equation can be expressed in terms of sensible energy, e, which is the specific internal energy exclusive of chemical energy, as

$$\frac{\partial (\rho e)}{\partial t} + \frac{\partial (\rho e U_j)}{\partial x_j} = -p \frac{\partial U_j}{\partial x_j} + \frac{\partial J_j}{\partial x_j} + \dot{Q}^{\mathrm{s}} + \dot{Q}^{\mathrm{c}} \tag{5}$$

The heat flux vector $\mathbf{J}$ is the sum of contributions due to heat conduction and enthalpy diffusion:

$$J_j = -K \frac{\partial T}{\partial x_j} - \rho D \sum_{k=1}^{N_{\mathrm{s}}} h_k \frac{\partial Y_k}{\partial x_j} \tag{6}$$

$\dot{Q}^{\mathrm{s}}$ and $\dot{Q}^{\mathrm{c}}$ are source terms due to spray and chemical reaction, respectively, and N_{s} is the total number of species. Assuming an ideal gas, the state equation is used to relate pressure and density:

$$p = \rho R T \sum_{k=1}^{N_{\mathrm{s}}} \frac{Y_k}{W_k} \tag{7}$$

where W_k is the molecular mass of the kth species.

1.1.2 Solution methods

Eulerian–Eulerian or Eulerian–Lagrangian (i.e., Discrete Droplet/Particle) methods have been proposed for solving the governing equations. However, the Eulerian–Lagrangian approach is used in most engine CFD simulations. In this approach, the gas phase is treated as continuous and the Eulerian formulation is used to solve the gas-phase equations. As described in Fuel Introduction, the liquid phase is treated as consisting of discrete particles described using the Lagrangian approach, in which liquid droplets are represented as assemblages of particles using the discrete particle method (Dukowicz, 1980). The Lagrangian description avoids numerical diffusion and allows individual attributes, such as particle size, velocity, position, and so on, to be statistically assigned to each particle. Mass, momentum, and heat transfer between the gas and liquid phases is accounted for using the spray source terms in Equations 1–3, and 5.

Finite volume, finite difference, and finite element methods that use a finite-difference mesh or grid to subdivide the computational region into a number of small cells are used to approximate the continuum partial differential equations of the gas phase. Most engine CFD codes employ the finite volume method because of its inherently conservative nature. The finite volume method applies the integral form of the governing equations, involving surface integrals (fluxes) and volume integrals (sources and sinks) to preserve the local conservation properties of the differential equations. The equations are discretized in both space and time. Usually, implicit differencing schemes are used to permit large numerical timesteps.

Mesh generation can be a tedious process in engine simulations owing to the complex, time-varying geometry of an engine. However, commercial software packages are available to aid in the process. In general, it is believed that body-fitted meshes, whose computational cells faithfully follow the combustion chamber surfaces, yield more accurate results near walls. But this can lead to distorted cells in the interior of the domain, with a corresponding reduction in accuracy, as numerical truncation errors increase on distorted meshes. Several commercial codes use Cartesian meshes whose wall cells require "clipping" or "trimming" in order to match wall boundaries. In this case, truncation error accuracy loss can be reduced by using local mesh refinement, but with corresponding increases in the computer time needed to solve the CFD equations on the resulting fine meshes.

As noted earlier, submodels are still required to describe the spray and combustion source terms in the governing conservation equations as they incorporate processes that occur on time and length scales too small to be resolved

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

on the numerical mesh. Thus, numerous engine-related numerical models have been developed to describe the turbulent flow, injection, spray, fuel/air mixing, ignition, combustion, pollutant formation, heat transfer, and other processes in engines. The fuel injection and spray-related models are described in Fuel Introduction, while this section focuses on turbulence, combustion, pollutant formation, and other models.

2 TURBULENCE MODELS

The governing equations given earlier describe flows in IC engines completely. However, this would require enormous computing resources and the practical use of these equations require averaging. This averaging is viewed as ensemble or cycle averaging for Reynolds-averaged Navier–Stokes (RANS) approaches and local spatial averaging for large eddy simulations (LES). The averaging process is applied to velocities, pressure, temperature, and species concentrations so that they are replaced by decompositions into mean and fluctuating components. For illustration, the velocity vector is replaced by

$$U_i = \overline{u}_i + u_i' \tag{8}$$

where the overbar represents the averaged velocity and the prime indicates the fluctuating component. For engine flows, the averaging is commonly thought of as a density-weighted averaging (Poinsot and Veynante, 2005). When this decomposition of the velocity and other quantities is substituted into Equation 3 and the averaging process is applied to the equation, additional terms occur because of nonlinearities. The most important terms that occur are the Reynolds stresses and they arise from the second term on the left-hand side of Equation 3. The Reynolds stresses are the average (or correlation) between fluctuating velocity components and can be represented by a turbulent stress term:

$$\tau_{ij}^T = -\overline{\rho}\overline{u_i' u_j'} \tag{9}$$

Then, the momentum equation, Equation 3, can be rewritten as

$$\frac{\partial(\overline{\rho u_i})}{\partial t} + \frac{\partial(\overline{\rho u_i u_j})}{\partial x_j} = -\frac{\partial \overline{p}}{\partial x_i} + \frac{\partial \overline{\tau}_{ij}}{\partial x_j} + \frac{\partial \tau_{ij}^T}{\partial x_j} + \overline{F}_i^{\mathrm{s}} + \overline{F}_i^{\mathrm{b}} \tag{10}$$

To solve this equation, a turbulence model for the Reynolds stresses is required. Additional modifications can occur because of nonlinearities in the F terms of Equation 3 as well as nonlinearities in the energy and species equations. All of these nonlinearities result in additional terms in the equations and they all require modeling. The modeling of the Reynolds stresses in Equation 9 forms the basis of most of these other models, so it is examined in detail.

2.1 Reynolds-averaged Navier–Stokes (RANS) models

RANS models are the most prevalent turbulence models in engine simulations. RANS models, specifically the k–ε model and its variants, have been used in engine simulations since CFD was first applied to the engine field. k is the turbulent kinetic energy and ε is its dissipation rate. With these two quantities, the Reynolds stresses are modeled by

$$\tau_{ij}^T = -\mu_{\mathrm{T}}\left(\frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_j}{\partial x_i}\right) \tag{11}$$

where μ_{T} is the turbulent viscosity and is modeled by

$$\mu_{\mathrm{T}} = \overline{\rho} C_\mu \frac{k^2}{\varepsilon} \tag{12}$$

The constant, C_μ, typically has a value of 0.09. An improved k–ε model, which is based on the renormalized group (RNG) theory, was first proposed by Yakhot and Orszag (1986). The k equation in the RNG version is the same as in the standard k–ε model, but the ε equation is based on a more rigorous mathematical derivation, instead of using empirical constants. The RNG k–ε equations are written as

$$\frac{\partial \rho k}{\partial t} + \nabla \cdot (\rho \mathbf{u} k) = -\frac{2}{3}\rho k \nabla \cdot \mathbf{u} + \tau : \nabla \mathbf{u} + \nabla \cdot [\alpha_k \mu \nabla k] - \rho\varepsilon + \dot{W}^{\mathrm{s}} \tag{13a}$$

$$\frac{\partial \rho \varepsilon}{\partial t} + \nabla \cdot (\rho \mathbf{u} \varepsilon) = -\left(\frac{2}{3}C_1 - C_3 + \frac{2}{3}C_\mu C_\eta \frac{k}{\varepsilon}\nabla \cdot \mathbf{u}\right) \times \rho\varepsilon \nabla \cdot \mathbf{u} + \nabla \cdot [\alpha_\varepsilon \mu \nabla \varepsilon] + \frac{\varepsilon}{k}[(C_1 - C_\eta)\tau : \nabla \mathbf{u} - C_2 \rho\varepsilon + C_{\mathrm{s}} \dot{W}^{\mathrm{s}}] \tag{13b}$$

where τ is the stress tensor and μ is the dynamic viscosity, and α's and C's are model constants.

The standard k–ε model includes a source term $\left(-\left(\frac{2}{3}C_1 - C_3\right)\rho\varepsilon\nabla \cdot \mathbf{u}\right)$ in the ε equation to account for length scale changes with velocity dilation, and spray-induced source terms, $\dot{W}^{\mathrm{s}}$. Han and Reitz (1995) modified the constant C_3 in the RNG k–ε model to take

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the compressibility effect into account. In their modified RNG k–ε model,

$$C_3 = \frac{-1 + 2C_1 - 3m(n-1) + (-1)^{\delta}\sqrt{6}C_{\mu}C_{\eta}\eta}{3}$$

where η is a ratio of turbulence and mean flow time scales, and m and n are related gas properties while C_{μ} and C_{η} are constants. Other details of the modified RNG k–ε turbulence model are given by Han and Reitz (1995) who applied the model to engine simulations. It was shown that the model could predict more realistic large-scale flame structures compared with the standard k–ε model. These structures influence in-cylinder temperature predictions and the modified model was able to quantitatively improve NOx emission predictions (Kong, Han, and Reitz, 1995).

2.2 Large eddy simulation (LES) models

The engine fuel preparation and combustion processes are controlled by the details of turbulent fluid motions. Although the simpler k–ε turbulence models are still widely used in industry, they have limited ability to resolve the detailed flow structures that are responsible for mixing and combustion in engines. LES is a numerical technique for simulating turbulent flows, first introduced by Smagorinsky (1963), that addresses this shortcoming. Kolmogorov's theory of self-similarity implies that the large eddies of a turbulent flow are dependent on the device geometry, while smaller eddies are self-similar and have a universal character (Kolmogorov, 1991). Therefore, the basic idea of LES is to solve only for large-scale eddies explicitly (i.e., those that can be resolved on the computational mesh), and to model the more universal small-scale (sub-grid scale, SGS) eddies on large meshes through the use of an SGS model. In contrast, in RANS, all fluctuations about the ensemble mean are modeled.

Therefore, LES has advantages over RANS in terms of predicting the instantaneous large-scale flow characteristics. This means LES is able to simulate important aspects of engine flows that RANS is not able to capture. For example, one of the more important applications of LES is to simulate cyclic variability (Vermorel *et al*., 2009). Another important use of LES is to simulate design sensitivities such as variations in port design (Thobois, Lauvergne, and Poinsot, 2007) and fuel injection characteristics (Giannadakis *et al*., 2009). Because LES captures large-scale flow features, it can also be more accurate than RANS simulations (Richard *et al*., 2007). However, LES is usually more computationally intensive than RANS and often requires the simulation of multiple, consecutive combustion cycles, increasing the computational cost even more (Vermorel *et al*., 2007).

Background information on LES modeling can be found in Pope (2000) for general information and in Rutland (2011) for LES in engine simulations. LES modeling assumes a different decomposition than RANS for Equation 8. In LES, the overbar represents a local spatial average. This can also be viewed as a local filtering of the velocity field that leaves the large eddies and removes the high frequency or sub-grid velocities. Then, $\overline{u_i}$ represents the large eddies and u'_i represents the sub-grid velocities. When this type of averaging is used in the momentum equations, Equation 3, the nonlinear term results in sub-grid stresses:

$$\tau_{ij}^{\mathrm{SGS}} = -\overline{\rho}(\overline{u_i u_j} - \overline{u}_i \overline{u}_j) \tag{14}$$

This replaces the Reynolds stresses τ_{ij}^{T} in Equation 10 and requires modeling. There are many types of LES models available, but the Smagorinsky model is common.

$$\tau_{ij}^{\mathrm{SGS}} = -\mu_{\mathrm{SGS}}\left(\frac{\partial \overline{u_i}}{\partial x_j} + \frac{\partial \overline{u_j}}{\partial x_i}\right) \tag{15}$$

where μ_{SGS} is the sub-grid scale turbulent viscosity and is modeled by

$$\mu_{\mathrm{SGS}} = \overline{\rho}(C_{\mathrm{s}}\Delta)^2|\overline{S}| \tag{16}$$

In this expression C_{s} is the Smagorinsky constant, typically $\sim$0.17. The length scale, Δ, is related to the filter width and is commonly related to the local cell size. The inverse time scale, $|\overline{S}|$, is the magnitude of the strain rate tensor formed using the local filtered velocities. There are several improvements to the Smagorinsky model. For example, there is the dynamic Smagorinsky model, which uses the dynamic procedure of Germano *et al*. (1991) to find the constant, C_{s}, locally as a function of the solution. There is also the one-equation viscosity model (Deardorff, 1980), which includes a transport equation for the sub-grid kinetic energy, k_{SGS}, that is similar to the turbulent kinetic energy Equation 13a and uses this to model the sub-grid viscosity as

$$\mu_{\mathrm{SGS}} = \overline{\rho}C_{\mu k}\Delta k_{\mathrm{SGS}}^{1/2} \tag{17}$$

where the constant, $C_{\mu k}$, is typically given values in the range 0.05–0.1. Additional advancements in LES models make use of technologies that are only available in LES and not in RANS. These include scale similarity concepts (Meneveau and Katz, 2000), Taylor series models (Clark, Ferziger, and Reynolds, 1979), Lagrangian models (Meneveau, Lund, and Cabot, 1996), and nonviscosity models (Rutland, 2011).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3 PHYSICAL AND CHEMICAL SUBMODELS

As mentioned, submodels are required to describe physical and chemical processes that occur on time and length scales too small to be resolved in a practical engine CFD mesh. The following sections review current models that are used to describe ignition, flame propagation and volumetric combustion, and the formation of pollutants. In addition, heat transfer, wall-film, and crevice flow models are briefly discussed. Details of practical spray and droplet submodels are described in Fuel Introduction.

3.1 Ignition and combustion submodels

Combustion models are needed to represent the chemical source terms, $\dot{\omega}_k$, in Equation 1 and $\dot{Q}^c$ in Equation 5. Combustion models have been developed to simulate a wide variety of combustion phenomena in both spark-ignited (SI) and compression-ignited (CI) engines, including spark and autoignition, laminar and turbulent combustion, premixed, non-premixed and stratified-charge combustion, kinetics and mixing-controlled combustion, and flame propagation combustion. For SI engines, the main task of combustion modeling is to describe the spark ignition, flame kernel growth, and flame propagation processes. For CI engines, combustion modeling also involves two steps: low temperature chemistry, which leads to autoignition, and subsequent high temperature reactions that contribute most of the heat release. Advanced LTC strategies combine the premixed charge strategy of the SI engine with the compression ignition of the diesel, as described, for example, by Lu, Han, and Huang (2011).

3.1.1 Ignition models

In SI engines, the flame is initiated by the electrical discharge of the spark plug. As typical length and time scales in engine CFD models are much larger than those of the developing flame kernels during the spark ignition processes, it is not feasible to resolve ignition details in engine CFD calculations. Thus, a phenomenological description is often used in engine simulations. Spark ignition can be simply simulated by adding energy to the ignition cells empirically over the spark duration.

Other models that account for different levels of physics are available, such as the discrete particle ignition kernel (DPIK) model introduced by Fan *et al.* (1999) and later improved by Tan and Reitz (2003). In the original DPIK model, the flame kernel was assumed to be spherical during the spark period and marked by particles. The particles initially move as prescribed by the local laminar flame speed and spark discharge velocities, and then transition to the turbulent burning velocity as the kernel size becomes comparable to turbulent eddy size scales. Precise tracking of the growth of the ignition kernel is thus possible without the need for a very fine mesh near the spark plug. This modeling approach has recently been extended by Dahms *et al.* (2011) to allow the particles to be convected by the local CFD-predicted flow field, thus removing the constraint of a spherical growing spark flame kernel.

Autoignition must be considered to model knock phenomena in SI engines (e.g., conventional gasoline and GDI) and compression ignition in CI engines (e.g., conventional diesel and LTC). The autoignition process is controlled by the chemical properties of the fuel and oxidizer. Relatively simple models that use correlations for ignition delay times are available. More accurate results are found with the use of reduced chemistry models, or with more detailed chemical kinetics mechanisms. Early engine CFD studies used the multistep "Shell" ignition model (Halstead, Kirsh, and Quinn, 1977), which was developed to predict the autoignition of hydrocarbon (HC) fuels. The premise of the Shell model is that degenerate branching plays an important role in determining the cool flame and two-stage ignition phenomena that are observed during the autoignition of realistic fuels. The model assumes that the ignition process can be described with five generic species and eight generic reactions, which represent the initiation, propagation, branching, and termination steps. To model diesel combustion, the Shell model is used only for the low temperature chemistry. After ignition (i.e., when the local gas temperature is greater than $\sim 1000\,\mathrm{K}$), combustion models are used simulate the subsequent second-stage high temperature combustion (e.g., Kong and Reitz, 1993). As described next, in recent studies the ignition process is modeled by integrating it into the combustion model using detailed or reduced chemistry mechanisms that have been validated against shock tube or flow reactor data.

3.1.2 Flame propagation models

As engine combustion occurs in a turbulent flow field, combustion models are often based on RANS and LES turbulence timescales. However, turbulent flames, such as those in SI engines, are thought to comprise interacting laminar flame sheets, and under typical engine pressures and temperatures the laminar flame thickness is less than $50\,\mu\mathrm{m}$, while typical CFD mesh sizes are about 1 mm. Thus, it is not possible to resolve turbulent flame structures and their detailed evolution in practical engine simulations. Instead, combustion models attempt to describe the overall combustion characteristics, such as the turbulent

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

flame propagation velocity and the conversion rate of reactants to products of combustion. The concepts of mixture fraction (Z) or reaction progress (c) are introduced in many combustion models. The mixture fraction quantifies the local mass fractions that originate from the reactant fuel (see Solving Combustion Chemistry in Engine Simulations—for detailed development of mixture fraction). This is useful in the analysis and modeling of non-premixed reacting systems. For premixed combustion systems, a reaction progress variable is often used. The progress variable increases from zero in the unburned reactants to unity in the burned products. The calculation of the progress variable can be associated with the mass fraction of a specific species (e.g., combustion products) or the temperature.

Flamelet and probability density function (PDF) models belong in the category of mixing-controlled combustion models. However, for spatially homogeneous fuel–air mixtures where flame propagation is not dominant, it is reasonable to assume that turbulence plays a lesser role in the combustion process. This is the case under engine operating conditions with low Damköhler numbers (slow chemistry timescales), for example, in LTC of premixed lean mixtures, such as in HCCI engines where chemical kinetics effects dominate. (However, it should be noted that turbulent fuel–air mixing is usually the agency by which the homogeneous fuel–air mixture is established in the engine.) In this case, it is often assumed that combustion occurs volumetrically in each computational cell, as in a "well-stirred reactor," and to account for the complex reaction pathways, detailed or reduced reaction chemistry mechanisms are used.

The simplest combustion model assumes a one-step global reaction (i.e., single timescale) for conversion of reactants to products based, for example, on a temperature-dependent Arrhenius model. This model neglects the effects of complex reaction paths and turbulence on the mean chemical reaction rate, and thus is generally only applicable over a narrow range of operating conditions. In turbulent mixing-controlled combustion models, it is assumed that the burn rate of the mixture is determined by the turbulent mixing rate, instead of by the chemical rate. This applies, for example, in diffusion combustion in conventional diesel engines, where turbulent mixing occurs relatively slowly compared to chemical reactions (high Damköhler number). In this case, it is reasonable to expect that turbulent transport of species and enthalpy to the reaction zone is the controlling factor. Spalding's (1971) eddy breakup (EBU) model was one of the first mixing-controlled models. Here, the burned and unburned gases are assumed to be located in different eddies, and the mean chemical conversion rate is controlled by the eddy's dissipation rate and by fluctuations of the fuel mass fraction or progress variable. In the EBU model, the reaction rate is expressed as

$$\dot{\tilde{\omega}}_{\mathrm{F}} = C_{\mathrm{EBU}} \frac{\overline{\rho}(\widetilde{Y_{\mathrm{F}}''^2})^{1/2}}{\tau_t}$$

where $\widetilde{Y_{\mathrm{F}}''^2}$ is the fluctuation of the fuel mass fraction and C_{EBU} is a model constant. When the k–ε turbulence model or its variants are used, the EBU time is $\tau_t \propto k/\varepsilon$.

For non-premixed turbulent combustion, Magnussen and Hjertager (1976) modified the EBU model such that the combustion rate is limited by the deficient species (fuel, oxidizer, or product depending on whether the mixture is lean or rich or the progress of reaction), and the fuel burn rate is

$$\dot{\tilde{\omega}}_{\mathrm{F}} = \frac{A\overline{\rho}\min\left(\widetilde{Y_{\mathrm{F}}}, \frac{\widetilde{Y_{\mathrm{F}}}}{r}, B\frac{\widetilde{Y_{\mathrm{P}}}}{1+r}\right)}{\tau_t}$$

where A and B are model constants, r is the stoichiometric oxidizer/fuel ratio, and Y_{F} and Y_{P} are the fuel and product mass fractions.

Other hybrid EBU/Arrhenius models have also been proposed. For example, Reitz and Bracco (1983) proposed a laminar and turbulent characteristic timescale combustion (CTC) model, which was applied to engine combustion by Abraham, Bracco, and Reitz (1985). The model uses the k–ε model for turbulent transport and assumes that the characteristic timescale (τ_{c}) at which all the chemical species approach their equilibrium (indicated by * in the equation), is the sum of a turbulence characteristic time (τ_{t}) and a laminar or chemical characteristic time (τ_{l}) of Arrhenius form. The rate change of the concentration of species i is given as (Kong, Han, and Reitz, 1995)

$$\frac{\mathrm{d}Y_i}{\mathrm{d}t} = -\frac{Y_i - Y_i^*}{\tau_{\mathrm{c}}} \tag{18}$$

The CTC model has been used successfully to model both premixed (Abraham, Bracco, and Reitz, 1985) and non-premixed (Kong, Han, and Reitz, 1995) turbulent combustion.

Flamelet models have also been developed for both premixed and non-premixed combustion regimes. For premixed combustion with flame propagation, the models are based on tracking a flame-front interface defined either by a combustion progress variable, c, or by a nonreacting scalar, G, which divides the flow into burned and unburned portions (Peters, 2000). In the Bray–Moss–Libby (BML) model (Bray and Libby, 1994) and the coherent flame model (CFM) (Marble and Broadwell, 1977), the progress variable is viewed either as a normalized temperature or as a normalized product mass fraction. These models assume

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a single-step reaction from unburned reactants to burned products and an infinitely thin flame structure. In the BML model, a PDF of the progress variable is assumed to be a two-delta function distribution and the progress variable is tracked by solving a transport equation.

For premixed combustion with flame propagation, such as stoichiometric combustion in conventional SI engines, the premise of flamelet models is that the principal effect of turbulence on combustion is to increase the effective flame surface area, as described by Haworth (2005). The overall burn rate is assumed to be the product of the fuel burn rate per unit flame area and the active flame area per unit volume, and the local mean fuel burning rate is expressed as

$$\tilde{\dot{\omega}}_{\mathrm{F}} = \rho_{\mathrm{u}} \widetilde{Y_{\mathrm{FR}}} S_{\mathrm{L}} \Sigma$$

where ρ_{u} is the unburned gas density, $\widetilde{Y_{\mathrm{FR}}}$ the fuel mass fraction, S_{L} the local laminar flame speed, and Σ the flame surface density (flame surface area per unit volume). The laminar flame speed can be computed or stored in tables as a function of pressure, temperature, equivalence ratio, and strain rate. The flame surface density can be specified as an algebraic model, such as suggested by Bray and Libby (1994). A more complete approach is to solve a transport equation, as in the CFM. For example, Trouve and Poinsot (1994) proposed a formulation based on DNS data. Flamelet models decouple the chemical kinetics and turbulence, while maintaining local coupling between chemical kinetics and molecular transport (Haworth, 2005). Boudier *et al*. (1992) validated a CFM model by comparing simulation results with in-cylinder flame-front contours in an experimental SI engine.

Recently, flamelet models have been proposed for flame propagation in premixed combustion systems based on a level set method (Sethian, 1999), which tracks interfaces. With application to combustion, Williams (1985) suggested a transport equation of a nonreactive scalar, $G(x, t)$, for laminar flame propagation known as the *G-equation method*. In the G-equation method, flame propagation is driven by the bulk fluid velocity, u, of the unburned mixture ahead of the flame front, and the laminar flame speed S_{L} normal to the flame, n. The rate of change of flame position, x_{f}, that is, the flame propagation velocity is

$$\frac{\mathrm{d}x_{\mathrm{f}}}{\mathrm{d}t} = u + nS_{\mathrm{L}}$$

and the flame front is defined by the $G(x, t) = G_0$ iso-surface with G_0 normally set equal to zero. The domain is divided into an unburned region where $G < G_0$ and a burned region where $G > G_0$. The normal vector $\mathbf{n}$ is defined as

$$\mathbf{n} = -\frac{\nabla G}{|\nabla G|} = -\frac{\nabla G}{[(\nabla G)^2]^{1/2}}$$

and the transport equation for G is derived by differentiating $G(x, t) = G_0$ with respect to t, as

$$\rho\left(\frac{\partial G}{\partial t} + \boldsymbol{u} \cdot \nabla G\right) = (\rho S_{\mathrm{L}})|\nabla G|$$

Peters (2000) extended the G-equation to model turbulent flames as

$$\rho\frac{\partial G}{\partial t} + \rho \boldsymbol{u} \cdot \nabla G = (\rho S_{\mathrm{L}}^0)|\nabla G| - \underbrace{(\rho D_{\mathrm{L}})\kappa|\nabla G|}_{\text{Flame curvature}} - \underbrace{(\rho L)S|\nabla G|}_{\text{strain rate}} \quad (19)$$

In this case, the turbulent flame front is regarded as an ensemble of local laminar flamelets with considerations of flame stretch effects on the flame speed. The turbulent G-equation concept has been successfully applied to combustion simulations for SI engines by Dekena and Peters (1999), Tan and Reitz (2003), (2004), and Liang and Reitz (2006). Tan and Reitz (2004) showed that the G-equation model can also be applied to non-premixed combustion because, in this case, the only change is that $S_{\mathrm{L}} = 0$. This extension also forms the basis of their so-called GAMUT (G equation for all mixtures a universal turbulent) combustion model.

For non-premixed (or partially premixed) combustion, such as turbulent diffusion combustion in conventional CI engines, a representative interactive flamelet (RIF) model has been proposed (see Solving Combustion Chemistry in Engine Simulations for more details). The model is based on the assumption that the smallest turbulent time and length scales are much larger than the chemical ones, and that there exists locally undisturbed flame sheets where the chemical reaction occurs and that the turbulent eddies do not enter the reaction zone (Pitsch, Barths, and Peters, 1996). The composition, mass fraction of each species is assumed to be only dependent on the mixture fraction, Z. Therefore, only the transport equations of the mean and variance of the mixture fraction ($\widetilde{Z}, \widetilde{Z''^2}$) are solved, instead of having to solve the transport equations of each species, and the composition, Y_{i}, as a function of Z is given by laminar flamelet calculations that use detailed fuel chemistry. The flamelet equation for each species is given as

$$\frac{\partial Y_i}{\partial t} = \frac{\chi}{2}\left(\frac{\partial^2 Y_i}{\partial Z^2}\right) + \frac{\dot{\omega}_i W_i}{\rho} \quad (20)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where χ is the scalar dissipation rate. The scalar dissipation rate accounts for strain effects as

$$\chi = 2D\left(\frac{\partial Z}{\partial x}\right)^2$$

In this equation, the diffusivity D is taken as the turbulent diffusivity and thus both turbulent (first term on right hand side) and chemistry (second term) timescales appear in the flamelet Equation 20, similar to the CTC model, Equation 18. However, a single "representative" turbulence timescale is used for the entire computational domain in the RIF model, which limits its accuracy, as discussed by Singh, Reitz, and Musculus (2006). Variations of the RIF model that attempt to remove this restriction by introducing more than one "representative" flamelet in the simulation have been introduced.

To account for the effects of turbulent fluctuation, a presumed PDF can be used for Z. The PDF is usually assumed to be a beta PDF, which has two varying parameters determined from the mean and variance of Z ($\widetilde{Z}, \widetilde{Z''^2}$), and the mean composition is computed by integrating Z from 0 to 1 with the presumed PDF, $P(Z, \widetilde{Z}, \widetilde{Z''^2})$ as

$$\widetilde{Y}_i(\overrightarrow{x}) = \int_0^1 P(Z, \widetilde{Z}, \widetilde{Z''^2}) Y_i(Z) \mathrm{d}Z$$

In the original RIF model, the entire domain was represented by one flamelet (Pitsch, Barths, and Peters, 1996). Hence, the scalar dissipation rate at stoichiometric mixture χ_{st} is defined as the domain-averaged value rather than being spatially resolved. However, in improved flamelet models, the composition Y_{i} is computed as a function of both Z and χ, and a global value for the scalar dissipation rate is not assumed, but is computed using the local properties of each computational cell. Advanced flamelet models also consider fluctuations of the dissipation rate, which are usually modeled using a log-normal PDF. The integration for the mean composition then becomes:

$$\widetilde{Y}_i(\overrightarrow{x}) = \int_0^\infty \int_0^1 P(Z) P(\chi) Y_i(Z, \chi) \mathrm{d}Z \, \mathrm{d}\chi$$

To apply flamelet models to spray combustion, the effects of liquid vaporization on small-scale turbulent mixing and turbulent combustion should also be considered, such as in the model proposed by Demoulin and Borghi (2000).

PDF models can be used for both premixed and nonpremixed turbulent combustion modeling. Other turbulent combustion models determine the PDF by solving a transport equation, instead of specifying a presumed PDF. In these models, joint velocity-composition PDF transport equations for velocity and reactive scalars are solved, usually by using the Lagrangian Monte Carlo particle-based methods (Pope, 1985). The gas-phase flow is represented by a large number of particles, where each particle contains information about its position, velocity, temperature, composition, and so on, which evolve with time. This treatment is similar to the Lagrangian method used for spray modeling (see Fuel Introduction).

3.1.3 Volumetric combustion models

As mentioned, under premixed conditions it is often justified to ignore turbulence–chemistry interactions and to assume volumetric heat release in each computational cell. In this case, a chemical kinetics mechanism is provided and integrated directly into the combustion source terms in the energy and species conservation Equations 1 and 5. This usually involves solving a stiff system of ordinary differential equations (ODEs) that govern the rate of change of the chemical species involved in the reactions. The L reversible reactions in a general chemistry mechanism with N species can be expressed as

$$\sum_{j=1}^{N} \nu'_{ji} X_j \Longleftrightarrow \sum_{j=1}^{N} \nu''_{ji} X_j, \quad \text{for} \quad i = 1, 2, \ldots, L$$

where ν'_{ji} and ν''_{ji} are the stoichiometric coefficients for the reactants and products, respectively, X_j is the chemical symbol, and the system of ODEs, which expresses the net production rate of each species, is

$$\dot{\omega}_j = \sum_{i=1}^{L} (\nu''_{ji} - \nu'_{ji}) \left(k_{\mathrm{f}i} \prod_{j=1}^{N} [X_j]^{\nu'_{ji}} - k_{\mathrm{r}i} \prod_{j=1}^{N} [X_j]^{\nu''_{ji}} \right), \quad \text{for} \quad j = 1, 2, \ldots, N \qquad (21)$$

where $k_{\mathrm{f}i}$ and $k_{\mathrm{r}i}$ are the elementary forward and reverse rate coefficients for the ith reaction, usually expressed in Arrhenius form. The CHEMKIN code (Kee, Rupley, and Miller, 1991) is widely used to solve chemical kinetics problems, and can be integrated into CFD codes (e.g., in KIVA-CHEMKIN by Kong *et al.* (2001).

Various chemical mechanisms have been developed to describe the oxidation of different fuels. Gasoline and diesel fuels are multicomponent fuels and mechanisms are still not available for all of the component fuel species. To simplify the problem, iso-octane and *n*-heptane are often used as surrogates in reduced chemistry models because they have similar chemical characteristics as those of gasoline and diesel, respectively. However, the physical properties of the original multicomponent fuels (e.g., vapor pressure, density, surface tension, etc.) should still be used in modeling

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

nonchemical processes, such as the fuel spray vaporization, as described by Anand *et al*. (2011).

It should be noted that when chemical kinetics is directly integrated into CFD, the effects of turbulence on combustion are still considered at the resolved grid scale. However, each computational cell is treated as a "well-stirred" reactor undergoing volumetric heat release. Therefore, SGS turbulence–chemistry interactions are not considered. As stated earlier, each model has its own applicable range. However, it follows that as numerical grid sizes are reduced, more of the combustion process becomes resolved. Indeed, the direct integration of chemistry approach has been shown to be effective in modeling HCCI combustion (Kong *et al*., 2001) as well as conventional diesel diffusion combustion (Kong, Sun, and Reitz, 2007). Flamelet models such as the G-equation GAMUT model of Tan and Reitz (2004) and Singh, Reitz, and Musculus (2006) are available for cases where SGS turbulence–chemistry interactions might be important.

3.1.4 "Knock" models

Liang *et al*. (2007) modeled knock in spark-ignition engines using the G-equation flame propagation model combined with detailed chemical kinetics. Spark ignition was modeled with the DPIK ignition model, and the turbulent flame propagation was described by the G-equation. A 22-species, 42-reaction iso-octane (iC_8H_{18}) mechanism was adopted to model the autoignition process of the gasoline/air/residual-gas mixture ahead of the flame front, as depicted in Figure 1. In the simulations, the local pressure was monitored at numerous locations within the cylinder to obtain pressure fluctuation information that signaled the onset of knocking combustion or autoignition, as shown in Figure 2. The computed local pressure traces shown in Figure 3 were processed using a digital fourth-order Butterworth band-pass filter, and the maximum peak-to-peak value PP_{max} of the filtered data was used to quantify the knock intensity.

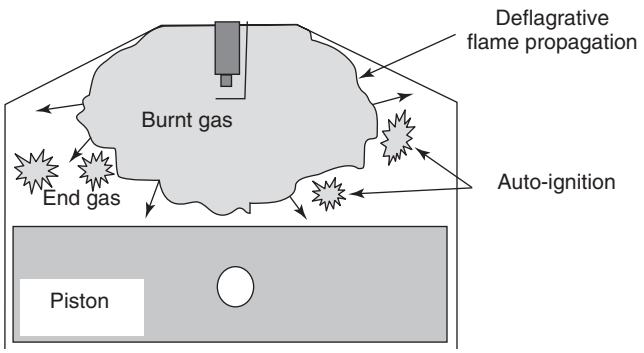


Figure 1. Schematic diagram of combustion modeling in SI engines. The DPIK particle model is used to model spark kernel growth, the G-equation is used to model turbulent flame propagation and a reduced chemistry mechanism is used for autoignition or knock modeling. (Reprinted with permission Copyright © 2007 SAE International (Liang *et al*., 2007).)

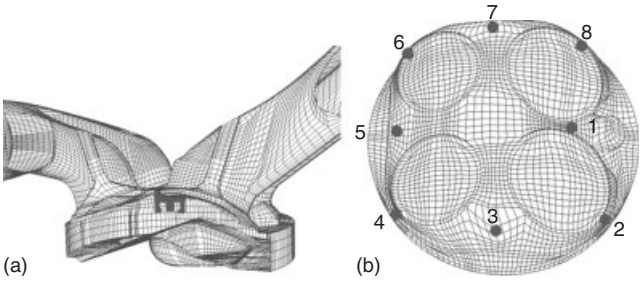


Figure 2. SI engine computational mesh with spark plug modeled using the DPIK particle model (a) and numerical pressure transducer locations (b). (Reprinted with permission Copyright © 2007 SAE International (Liang *et al*., 2007).)

For quantitative comparison of knock intensities under different operating conditions, a knock index KI was defined as the average of the local PP_{max} values at N different locations, that is,

$$KI = \frac{1}{N}\sum_{n=1}^{N} PP_{max,n}$$

The corresponding flame front and temperature evolutions of the -25° after top dead center (ATDC) spark timing case are shown in the left column of Figure 4. In this case, end-gas autoignition does not occur until 25° after spark (or TDC). However, within a further 5°CA almost all the end gas is autoignited simultaneously at around TDC, giving a violent knock. In contrast, in the -10° ATDC case (right column of Figure 4), a considerable amount of end gas is autoignited ahead of the flame front by 20° after spark (much earlier than the 25° ATDC case). However, the autoignited mixture propagates along with the turbulent flame front (indicated by the solid black line), and no explosive autoignition is observed throughout the flame propagation process. Therefore, only a light knock is caused.

When including detailed fuel chemistry mechanisms in CFD codes, the computational time can become prohibitive owing to the large number of chemical species and reactions in Equation 21. Therefore, significant research has been focused on developing methods to save computer time while maintaining acceptable accuracy. For example, much effort has been placed on developing methods to reduce

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

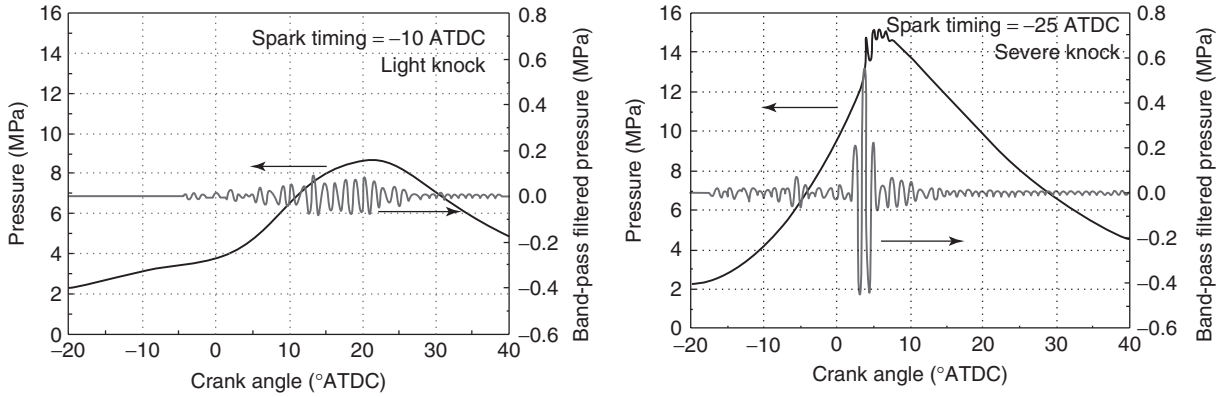


Figure 3. Predicted and band-pass-filtered simulated pressures at location 6 showing onset of light knocking combustion at retarded spark timing (-10°ATDC) and severe knocking combustion at advanced spark timing (-25°ATDC). (Reprinted with permission Copyright © 2007 SAE International (Liang *et al*., 2007).)

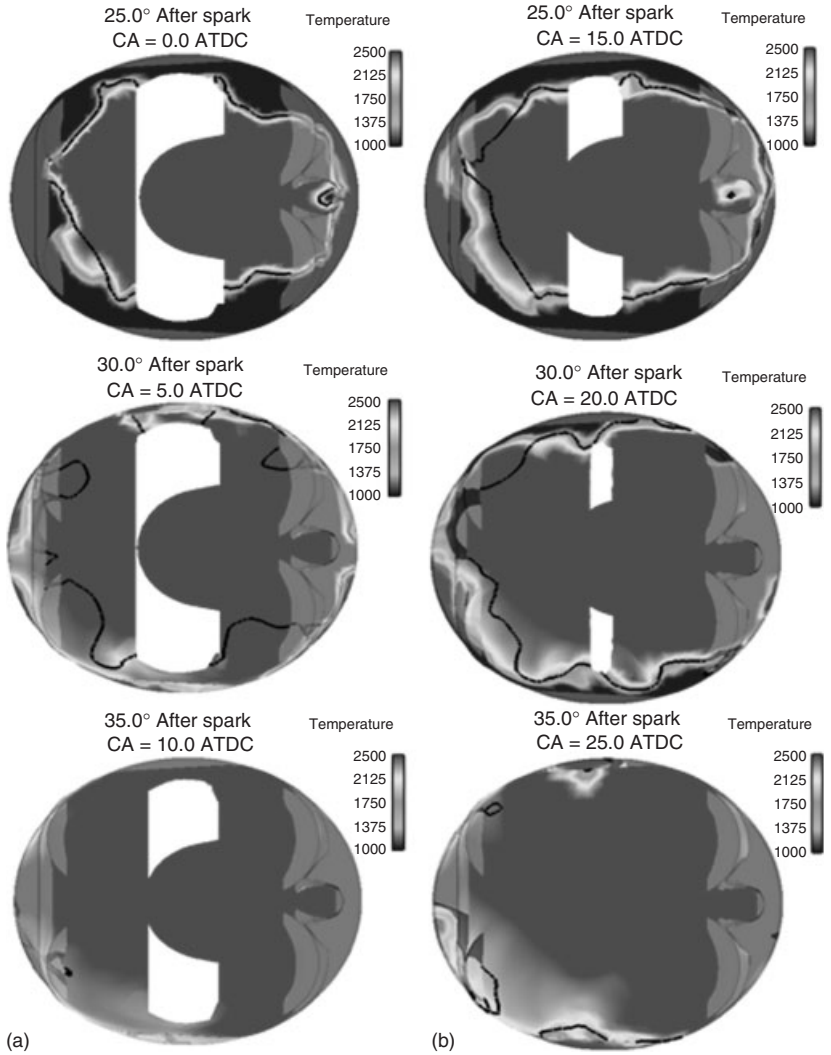


Figure 4. Simulated in-cylinder temperature profiles in a horizontal plane above the piston surface. Location of $G = 0$ surface (flame front) shown by solid black line and high temperature regions ahead of the flame indicate auto-ignition. (a): spark timing -25°ATDC; (b): spark timing -10°ATDC. (Reprinted with permission Copyright © 2007 SAE International (Liang *et al*., 2007).)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

detailed fuel chemistry mechanisms to skeletal mechanisms that eliminate species and reactions that are not crucial and still capture the major combustion characteristics of interest. In addition, reactions can be combined to remove unimportant intermediate species and reactions. In this case, the reaction rate constants of the remaining reactions usually need adjustment to account for the changes (e.g., Patel, Kong, and Reitz, 2004; Ra and Reitz, 2008). A detailed mechanism may contain several hundred species and several thousand reactions, and it can take months to finish a single cycle in engine simulations with realistic engine geometries. In contrast, a skeletal mechanism with tens of species and reactions only requires days or even hours of run time, which makes its application in engine simulations feasible. Other approaches, such as storage/retrieval-based schemes (Pope, 1997) and on-the-fly mechanism reduction during the computation, have also been used to speed up calculations (e.g., Shi, Ge, and Reitz, 2011; Puduppakkam *et al*., 2011).

3.1.5 Turbulence–chemistry interactions

Combustion regimes that range from locally premixed to partially premixed and non-premixed, with and without turbulence–chemistry interactions, can occur simultaneously in an engine. Thus, "universal" combustion models that can describe all types of combustion phenomena simultaneously are of interest. Hybrid models (e.g., Zellat *et al*., 2003; Tan and Reitz, 2004) that combine the models described have been described. Singh, Reitz, and Musculus (2006) assessed the CTC, RIF, and well-stirred or volumetric reaction (CHEMKIN) models for modeling diesel combustion in a heavy-duty (HD) DI diesel engine in different operating modes ranging from conventional high temperature diesel diffusion combustion regimes to so-called LTC regimes. The volumetric heat release model showed superior performance in predicting the LTC behavior and soot emissions as compared to the other two methods.

Kokjohn and Reitz (2011b) compared the performance of the G-equation flame propagation and volumetric heat release models as applied to diesel sprays, and found that turbulence–chemistry interaction effects were only noticeable at the flame lift-off location, the most upstream location of the OH mass fraction contours (surrounding the jets) shown in Figure 5. This work addressed the question of whether the flame lift-off length is governed by autoignition or flame propagation, and demonstrated that both the volumetric heat release and G-equation models described provide accurate lift-off predictions, whether or not flame propagation is considered in the model, as long as the lift-off length is defined as the distance from the tip of the injector to that axial location where significant OH radical formation is seen. As seen in Figure 5, both simulations and the experiments show a thin band of OH surrounding the spray plume, which indicates the presence of a diffusion flame at the periphery of the jet.

3.2 Pollutant emission models

A major goal of engine CFD modeling is to predict pollutant emissions. Of most concern are NO*x* and particulate matter (PM) or soot. The mechanisms of NO*x* production at high temperatures are well known and the extended Zel'dovich mechanism is the most commonly used to describe NO formation (see NO_x Formation and Models and Heywood,

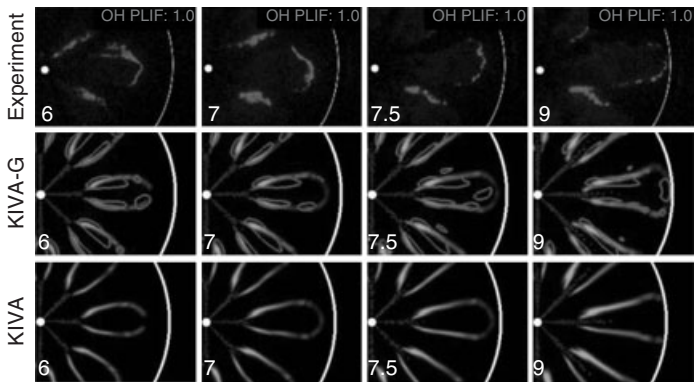


Figure 5. Comparison of measured OH PLIF (top row) and computed OH mass fractions (bottom two rows) for a high temperature, short ignition delay condition from 6 to 9° before TDC. Modeling with (KIVA-G) and without (KIVA) considering flame propagation. Dot indicates injector nozzle location. OH contours surround the jets, predicted $G = 0$ flame surface shown as islands, and spheres in simulation images show location of spray droplets. Flame lift-off is at the most upstream contour location.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

1988):

$$\begin{aligned} O + N_2 &\Longleftrightarrow NO + N \quad (a) \\ N + O_2 &\Longleftrightarrow NO + O \quad (b) \\ N + OH &\Longleftrightarrow NO + H \quad (c) \end{aligned} \tag{22}$$

A factor of 1.533 (the ratio of the molecular weight of NO_2 to NO) is introduced to convert NO to NO*x*, and the steady-state assumption for *N* atoms is usually made. The Zel'dovich mechanism only accounts for the thermal NO. A number of other NO*x* mechanisms are also important under different engine operating conditions, such as the N_2O—intermediate mechanism, which plays an important role in NO production in fuel-lean LTC, and the Fenimore or prompt NO mechanism, which is important in fuel-rich combustion. When a detailed fuel chemistry kinetics mechanism is used for the combustion calculations, a NO mechanism can be integrated into the fuel mechanism to better predict NO*x* emissions. In this case, NO*x* is given as the sum of NO and NO_2. For example, a detailed NO mechanism has been developed and included in the GRI (Gas Research Institute) mechanism (see NO_x Formation and Models and Bowman *et al*., 1997) for methane by the GRI. Sun (2007) reduced the GRI NO mechanism significantly (containing an additional 22 species and 101 reactions, in addition to the fuel chemistry mechanism) and showed that a reduced mechanism, which contains only 4 additional species and 12 reactions, adequately reproduces results obtained with the original GRI NO mechanism in engine simulations.

Soot production involves complicated physical and chemical processes: particle nucleation, surface growth, surface oxidation, and particle coagulation, and so on. (see Particulate Formation and Models). As the mechanism of soot production is still not completely understood, developing an accurate soot model is a subject of much current research. Although soot models with detailed descriptions of the physical and chemical processes leading from the fuel molecules to soot particles have been developed (e.g., Particulate Formation and Models and Kitamura *et al*., 2002, use of such models in engine applications is still not feasible, because solving the soot related gas-phase chemistry itself takes significant computer time. Therefore, empirical soot models are still widely used in multidimensional engine simulations. For example, the Hiroyasu, Kadota, and Arai (1983) two-step soot model considers soot formation and soot oxidation as two competing processes, where:

$$\frac{dM_s}{dt} = \frac{dM_{sf}}{dt} - \frac{dM_{so}}{dt}$$

and M_{sf}, M_{so}, and M_s are the formed, oxidized, and net soot mass in each computational cell at time, *t*. In this model, soot formation is assumed to be proportional to the concentration of fuel vapor and the oxidation step describes the destruction of soot particles by oxygen. Patterson *et al*. (1994) replaced the original Hiroyasu soot oxidation formulation with the Nagle and Strickland-Constable oxidation model (Nagle and Strickland-Constable, 1962) and achieved more realistic predictions. In their model, carbon oxidation occurs by two mechanisms whose rates depend on surface chemistry involving more reactive and less reactive sites. Kong, Sun, and Reitz (2007) pointed out that the concentration of fuel vapor can no longer be associated with soot formation when detailed fuel chemistry is used, as the fuel decomposes to smaller intermediate HC molecules very quickly once the low temperature reactions occur. They suggested use of acetylene (C_2H_2) as the inception species for the soot formation.

Although adequate results can be obtained using the relatively simple two-step soot model, its model constants require some tuning, as described by Wang, Reitz, and Yao (2012). Thus, more detailed models are of interest. Fusco, Knox-Kelecy, and Foster (1994) proposed an eight-step phenomenological soot model in which eight global reactions describe the fuel pyrolysis, particle inception, surface growth, coagulation, and oxidation. Tao *et al*. (2005) improved this model by generating soot precursor species from acetylene, rather than directly from fuel molecules. They also added an additional OH-related oxidation step into the eight-step soot model. This modeling approach has been extended by Vishwanathan and Reitz (2010) to include a reduced kinetics model for PAH species up to pyrene (four-ring aromatic), which serves as the soot precursor species. This model has been found to be able to successfully reproduce soot formation trends in diesel sprays and engines with reduced need for model constant tuning.

3.2.1 Application to optimization of RCCI combustion

To illustrate application of the above-mentioned combustion and emission models, the dual-fuel RCCI results of Kokjohn *et al*. (2011a) are reviewed in Figures 6 and 7. RCCI combustion uses in-cylinder blending of two fuels with differing autoignition characteristics to control premixed charge combustion. In Kokjohn *et al*.'s study, gasoline was delivered using a port-fuel injector and diesel fuel was delivered using a common rail-direct injector using multiple injections in a 2.4 L single-cylinder HD diesel engine operating at 1300 rev/min. The injections' timings and durations were selected using CFD modeling combined with the genetic algorithm optimization tools

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

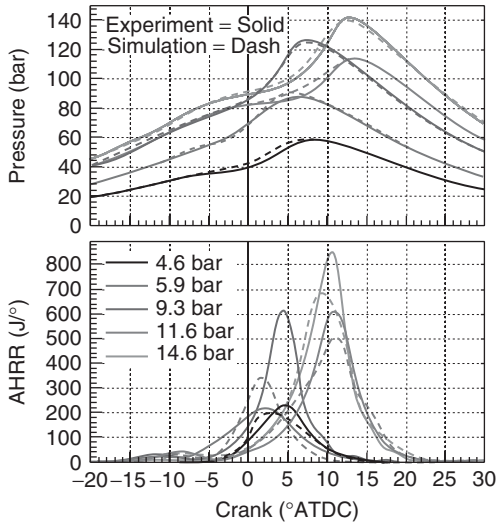


Figure 6. Comparison of measured and predicted dual-fuel RCCI cylinder pressure and apparent heat release rate (AHRR) over the load range from 4.6 to 14.6 bar IMEP. The experimental traces are shown in solid lines and the simulation results are the dashed curves. (Reproduced with permission from Kokjohn *et al*., 2011a. © SAGE Publications Ltd.)

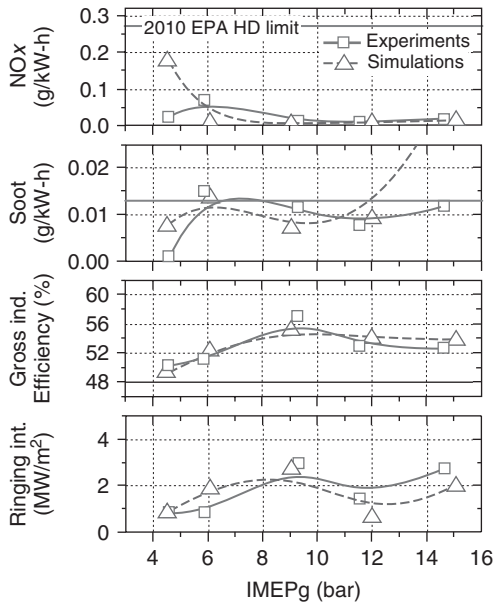


Figure 7. Measured and predicted emissions and performance of dual-fuel RCCI over a range of loads from 4.6 to 14.6 bar IMEP. The solid horizontal lines show the US 2010 environmental protection agency's on-highway truck emissions limits for NO*x* and soot. (Reproduced with permission from Kokjohn *et al*., 2011a. © SAGE Publications Ltd.)

described by Shi, Ge, and Reitz (2011) to reveal the optimum in-cylinder fuel-blending strategy. The first diesel injection, known as the *squish conditioning pulse*, was delivered near 60° BTDC, targeting the squish region of the combustion chamber. The purpose of this injection is to control the fuel reactivity to ensure complete combustion of the premixed gasoline in the outer portion of the combustion chamber. The second injection occurred near 35° BTDC and targets the bowl region of the combustion chamber. This injection event generates a relatively high reactivity region that acts as an ignition source. Note from Figure 6 that second stage ignition indicated by the start of the main heat release occurs near TDC, and thus there is significant time for fuel–air mixing. Thus, RCCI is essentially a low temperature HCCI-like combustion process, but with fuel reactivity gradients providing control of combustion.

Figure 6 shows the measured and predicted cylinder pressures and apparent heat release rate (AHRR) over a load sweep from 4.6 to 14.6 bar indicated mean effective pressure (IMEP). The simulations are seen to be able to accurately capture the combustion characteristics over the entire load range. Figure 7 shows the corresponding measured and predicted emissions and performance. The gross indicated efficiency (defined as gross-indicated work/fuel energy) is seen to be ∼48% at the low load condition, peaks at ∼57% at 9.3 bar IMEP, and levels out near 52% for the higher loads. Also shown in Figure 7 are the US 2010 Environmental Protection Agency's HD on-highway truck limits for NO*x* and soot emissions (i.e., 0.268 g/kW-h and 0.0134 g/kW-h for NO*x* and soot, respectively). Notice that the measured NO*x* is near zero and significantly below the 2010 EPA HD limits. Furthermore, it can be seen that soot is extremely low and below the 2010 EPA HD limits for all but the 6 bar IMEP point, which is very near the limit. The predicted NO*x* and soot emissions also agree well with the measured data, as does the ringing intensity (RI). Because the engine operates under boosted conditions for all but the lightest load, the peak pressure rise rate is not representative of knock; therefore, the RI correlation of Eng (2002) is used. Ringing intensities below 5 MW/m^2 indicate acceptable combustion noise and knock-free operation (Dec and Yang, 2010).

3.3 Heat transfer, wall-film, and crevice flow models

Other models are needed in multidimensional engine simulations, such as wall heat transfer and liquid film models, radiation models, and crevice models, depending on the application. Heat transfer at the cylinder walls

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

significantly affects engine performance, efficiency, and emissions, and is mainly due to gas-phase convection, liquid fuel film conduction, and high temperature gas and soot radiation. In most SI engines, gas-phase convection is the dominant factor for heat transfer. However, in CI engines, radiative heat transfer can become important, as described by Wiedenhoefer and Reitz (2003), who used a discrete ordinates model. Their study showed that nonuniform combustion chamber surface temperature distributions have a significant effect on heat loss through combustion chamber walls and on NOx emissions. A conjugate wall heat transfer model was used by Wiedenhoefer and Reitz (2003) to couple the gas-phase spray and combustion model with the metal component heat conduction.

As wall boundary layers are relatively thin compared to computational mesh sizes, velocity and temperature wall functions are often used to describe the gas-phase near-wall shear stress and heat transfer. Traditional temperature wall functions are derived under the assumptions of a steady and incompressible flow using the Reynolds analogy. However, Han and Reitz (1997) noted that the gas density varies significantly in IC engines and unsteadiness and chemical heat release may invalidate the analogy. Their model accounts for variable-density turbulent flows and was applied to gas/wall convective heat transfer predictions in both a premixed charge SI engine and an HD diesel engine. Satisfactory agreement between predicted and measured wall heat flux data was obtained under both firing and motoring conditions. When liquid films are present, the heat transfer between the gas phase and the liquid phase, the energy used to vaporize the fuel, and the heat conduction between the liquid film and the wall must all be considered. This is considered in wall-film (spray/wall interaction) models, such as that proposed by Amsden (1999).

The crevice region between the piston, cylinder, and piston rings can also have significant effects on engine performance. In engine CFD simulations, the flow through the piston-cylinder-ring crevice can be accounted for by introducing an additional crevice grid, which is attached to the main computational grid in the crevice region (e.g., Lee and Reitz, 2010). The other way is to use a phenomenological crevice flow model, such as the one by Namazian and Heywood (1982) and extended for use in engine CFD by Reitz and Kuo (1989). A major reason for using a crevice grid or a crevice model is to ensure that the compression ratio in the modeled engine is the same as that in the real engine. Another reason is to model unburned HC emissions more accurately. Unburned fuel from the piston-liner ring crevices is one of the major sources for HC emissions in IC engines.

4 HIGH FIDELITY LES COMBUSTION MODELS

When LES is used for turbulence modeling, as described in Section 2.2, the combustion model should be adapted for the higher fidelity simulations. As in RANS, combustion modeling in LES is required to represent the chemical source terms in Equations 1 and 5. In many cases, the LES combustion models are similar to the RANS models, with modifications for LES. Thus, direct chemistry approaches, RIF, characteristic time, PDF, flamelet, and G-equation combustion models can be used in LES as long as the models are updated to be theoretically consistent with LES and adapted to work with LES variables (Rutland, 2011).

A key component of most combustion models is a turbulence timescale. This can be used in combination with "laminar" timescales or inverse reaction rates from direct chemistry solutions (Equation 18). The turbulence timescale can be used in models to account for mixing limited reactions in non-premixed combustion modeling (Equation 20) or flame propagation in premixed combustion modeling (Equation 19). In RANS, turbulent timescales are typically modeled using turbulent kinetic energy, k, divided by the turbulence dissipation rate, ε. In LES, these variables are not available and turbulence timescales are more difficult to obtain. A rudimentary timescale can be found using the inverse large-scale strain rate or the mesh size divided by the turbulent viscosity (Equation 16). LES models that use a transport equation for the sub-grid kinetic energy, k_{SGS}, are better suited to LES combustion modeling because a turbulent timescale is more easily obtained as follows:

$$t_{\mathrm{SGS}} \sim \frac{\Delta}{k_{\mathrm{SGS}}^{1/2}} \tag{23}$$

Another approach that can be used in mixing-controlled combustion modeling is to develop an expression for the sub-grid scalar dissipation rate (see Equation 20 for the basic term). For example, Zhang and Rutland (2012), use scale similarity and a dynamic coefficient to develop a sub-grid scalar dissipation rate model that is used in diesel combustion modeling.

LES modeling results in more flow structures being resolved and appearing in the simulations. An example of this is shown in Figure 8. This shows temperature contours at the start of injection for multiple, consecutive engine cycles. The RANS results show more uniform temperature contours compared to the rich flow detail of the LES results. This figure also demonstrates that LES simulations are able to capture cyclic variability in more detail than RANS simulations. Also, note that the LES results indicate

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

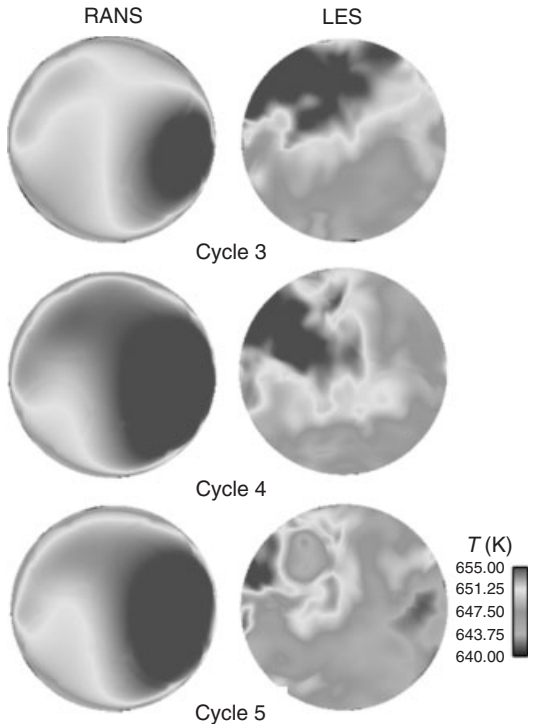


Figure 8. Comparison of RANS and LES simulations of multiple, consecutive engine cycles of a heavy-duty diesel engine showing temperature contours at the start of injection.

a slightly lower spatially averaged temperature than the RANS results owing to higher wall heat transfer resulting from the resolved large scales. These differences indicate that LES is more sensitive to flow details and more likely to correctly respond to changes in combustion system design. In addition, LES gives more accurate results in global parameters such as pressure traces and heat release profiles because more of the physics represented in the equations is solved directly on the mesh rather than being modeled (Rutland, 2011).

REFERENCES

Abani, N. and Reitz, R.D. (2010) Effects of sub-grid scale mixing of vapor in diesel sprays using jet theory. *Atomization and Sprays*, **20**, 71–83.

Abraham, J., Bracco, F.V., and Reitz, R.D. (1985) Comparisons of computed and measured premixed charge engine combustion. *Combustion and Flame*, **60**, 309–322.

Amsden, A.A. (1993) KIVA-3: a KIVA program with block-structured mesh for complex geometries. Los Alamos National Laboratory Report No. LA-12503-MS

Amsden, A.A. (1997) KIVA-3V: A block-structured KIVA program for engines with vertical or canted valves. Los Alamos National Laboratory Report No. LA-13313-MS

Amsden, A.A. (1999) KIVA-3V, Release 2, Improvements to KIVA-3V. Los Alamos National Laboratory Report LA-UR-99-915.

Amsden, A.A., Ramshaw, J.D., O'Rourke, P.J., *et al*. (1985) KIVA: a computer program for two- and three-dimensional fluid flows with chemical reactions and fuel sprays. Los Alamos National Laboratory Report No. LA-10245-MS

Amsden, A.A., O'Rourke, P.J., and Butler, T.D. (1989) KIVA-II: a computer program for chemically reactive flows with sprays. Los Alamos National Laboratory Report No. LA-11560-MS

Anand, K., Ra, Y., Reitz, R.D., and Bunting, B. (2011) Surrogate model development for the FACE fuels. *Energy & Fuels*, **25**, 1474–1484.

Boudier, P., Henriot, S., Poinsot, T., *et al*. (1992) *A model for turbulent flame ignition and propagation in spark ignition engines*. 24th Symposium (International) on Combustion, the Combust. Inst., 503–510.

Bowman, C.T., Hanson, R.K., Davidson, D.F., *et al*. (1997) GRI-Mech http://www.me.berkeley.edu/gri_mech/ (accessed 9 September)

Bray, K.N.C. and Libby, P.A. (1994) Recent developments in the BML model of premixed turbulent combustion, in *Turbulent Reacting Flow* (eds P.A. Libby and F.A. Williams), Academic Press, New York.

Caton, J.A. (2011) Comparisons of Global Heat Transfer Correlations for Conventional and High Efficiency Reciprocating Engines. Paper ICEF2011-60017, *Proceedings of ASME Internal Combustion Engine Division 2011 Fall Technical Conference ICEF201*, Morgantown, WV, October 2011.

Clark, R.A., Ferziger, J.H., and Reynolds, W.C. (1979) Evaluations of subgrid-scale models using an accurately simulated turbulent flow. *Journal of Fluid Mechanics*, **91**, 1–16.

Dahms, R.N., Drake, M.C., Fansler, T.D., *et al*. (2011) Understanding ignition processes in spray-guided gasoline engines using high-speed imaging and the extended spark-ignition model SparkCIMM Part A: spark channel processes and the turbulent flame front propagation. *Combustion and Flame*, **158**, 2229–2244.

Deardorff, J.W. (1980) Stratocumulus-capped mixed layers derived from a three-dimensional model. *Boundary–Layer Meteorology*, **18**, 495–527.

Dec, J.E. and Yang, Y. (2010) Boosted HCCI for high power without engine knock and with ultra-low NOx emissions using conventional gasoline. SAE paper 2010-01-1086.

Dekena, M. and Peters, N. (1999) Combustion modeling with the G-equation. *Oil & Gas Science and Technology-Rev. IFP*, **54** (2), 265–270.

Demoulin, F.X. and Borghi, R. (2000) Presumed PDF modeling of turbulent spray combustion. *Combustion Science and Technology*, **158**, 249–271.

Dukowicz, J.K. (1980) A particle-fluid numerical model for liquid sprays. *Journal of Computational Physics*, **35**, 229–53.

Eng, J. (2002) Characterization of pressure waves in HCCI combustion. SAE paper 2002-01-2859.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fan, L., Li, G., Han, Z., *et al*. (1999) Modeling fuel preparation and stratified combustion in a gasoline direct injection engine. SAE Paper 1999-01-0175.

Fusco, A., Knox-Kelecy, A.L., and Foster, D.E. (1994) Application of a Phenomenological Soot Model to Diesel Engine Combustion. *COMODIA Conference 94*, Yokohama, 571–576.

Germano, S., Piomelli, U., Moin, P., and Cabot, W.H. (1991) A dynamic subgrid-scale eddy viscosity model. *Physics of Fluids A*, **3**, 1760–1765.

Giannadakis, E., Theodorakakos, A., Papoutsakis, A., *et al*. (2009) LES predictions of vortical flow structures in diesel injector nozzles. SAE Paper 2009-01-0833.

Halstead, M., Kirsh, L., and Quinn, C. (1977) The autoignition of hydrocarbon fuels at high temperatures and pressures - fitting of a mathematical model. *Combustion and Flame*, **30**, 45–60.

Han, Z. and Reitz, R.D. (1995) Turbulence modeling of internal combustion engines using RNG k-ε models. *Combustion Science and Technology*, **106**, 267–295.

Han, Z. and Reitz, R.D. (1997) A temperature wall function formulation for variable-density turbulence flows with application to engine convective heat transfer modeling. *International Journal Heat & Mass Transfer*, **40** (3), 613–625.

Haworth, D.C. (2005) A review of turbulent combustion modeling for multidimensional in-cylinder CFD. SAE Paper 2005-01-0993.

Herold, R.E., Wahl, M.H., Regner, G., *et al*. (2011) Thermodynamic benefits of opposed-piston two-stroke engines. SAE paper 2011-01-2216.

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*, McGraw-Hill, New York.

Hiroyasu, H. and Kodota, T. (1976) Models for combustion and formation of nitric oxide and soot in DI diesel engines. SAE Paper 760129.

Hiroyasu, H., Kadota, T., and Arai, M. (1978) *Supplementary comments: fuel spray characterization in diesel engines*. Combustion Modeling in Reciprocating Engines Symposium, General Motors Research Laboratories.

Hiroyasu, H., Kadota, T., and Arai, M. (1983) Development and use of a spray combustion model to predict diesel engine efficiency and pollutant emission. *Bulletin of JSME*, **26** (24 Paper 214–12), 569–575.

Kee, R.J., Rupley, F.M., and Miller, J.A. (1991) CHEMKIN-II: a fortran chemical kinetics package for the analysis of gas-phase chemical kinetics. Sandia National Laboratories Report SAND89-8009.

Kitamura, T., Ito, T., Senda, J., and Fujimoto, H. (2002) Mechanism of smokeless diesel combustion with oxygenated fuels based on the dependence of the equivalence ratio and temperature on soot particle formation. *International Journal of Engine Research*, **3** (4), 223–248.

Kokjohn, S.L., Hanson, R.M., Splitter, D.A., and Reitz, R.D. (2011) Fuel reactivity controlled compression ignition (RCCI): a pathway to controlled high-efficiency clean combustion. *International Journal of Engine Research*, **12** (3), 209–226.

Kokjohn, S.L. and Reitz, R.D. (2011) Investigation of the roles of flame propagation, turbulent mixing, and volumetric heat release in conventional and low temperature diesel combustion. *ASME Journal of Engineering for Gas Turbines and Power*, **133**, 102805-1–102805-10.

Kolmogorov, A.N. (1991) The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **434** (1890), 9–13.

Kong, S.C., Han, Z.Y., and Reitz, R.D. (1995) The development and application of a diesel ignition and combustion model for multidimensional engine simulations. SAE Paper 950278.

Kong, S.C., Marriott, C.D., Reitz, R.D., *et al*. (2001) Modeling and experiments of HCCI engine combustion using detailed chemical kinetics with multidimensional CFD. SAE Paper 2001-01-1026.

Kong, S.C. and Reitz, R.D. (1993) Multidimensional modeling of diesel ignition and combustion using a multistep kinetics model. Paper 93-ICE-22, ASME Transactions, *Journal of Engineering for Gas Turbines and Power*, **115** (4), 781–789.

Kong, S.C., Sun, Y., and Reitz, R.D. (2007) Modeling diesel spray flame lift-off, sooting tendency and NOx emissions using detailed chemistry with phenomenological soot model. *ASME Journal of Engineering for Gas Turbines and Power*, **129**, 245–251.

Lakshminarayanan, P.A. and Aghav, Y.V. (2010) *Modelling Diesel Combustion*, Springer, New York.

Lee, C.-W. and Reitz, R.D. (2010) Predictions of the effects of piston-liner crevices on flow motion and emissions in three-dimensional diesel engine simulations. *International Journal of Engine Research*, **11** (1), 47–59.

Liang, L. and Reitz, R.D. (2006) Spark ignition engine combustion modeling using a level set method with detailed chemistry. SAE Paper 2006-01-0243.

Liang, L., Reitz, R.D., Iyer, C.O., *et al*. (2007) Modeling knock in spark-ignition engines using a G-equation combustion model incorporating detailed chemical kinetics. SAE paper 2007-01-0165.

Lu, X., Han, D., and Huang, Z. (2011) Fuel design and management for the control of advanced compression-ignition combustion modes. *Progress in Energy and Combustion Science*, **37**, 741–783.

Magnussen, B.F. and Hjertager, H. (1976) On Mathematical Modelling of Turbulent Combustion with Special Emphasis on Soot Formation and Combustion. *Proceedings 16th Symposium (International) on Combustion*, The Combustion Institute, **16**, 719–729.

Marble, F.E. and Broadwell, J. (1977) The coherent flame model for turbulent chemical reactions. Project SQUID, Report TRW-9-PU.

Meneveau, C. and Katz, J. (2000) Scale-invariance and turbulence models for large-eddy simulation. *Annual Review of Fluid Mechanics*, **32** (1), 1–32.

Meneveau, C., Lund, T.S., and Cabot, W.H. (1996) A lagrangian dynamic subgrid-scale model of turbulence. *Journal of Fluid Mechanics*, **319**, 353–385.

Nagle, J. and Strickland-Constable, R.F. (1962) Oxidation of Carbon Between 1000–2000 C. *Proceedings of the 5th Carbon Conference*, **1**, 265–325.

Namazian, M. and Heywood, J.B. (1982) Flow in the piston-cylinder-ring crevices of a spark-ignition engine: effect on hydrocarbon emissions, efficiency and power. SAE Paper 820088.

Patel, A., Kong, S.-C., and Reitz, R.D. (2004) Development and validation of a reduced reaction mechanism for HCCI engine simulations. SAE Paper 2004-01-0558.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Patterson, M.A, Kong, S.C., Hampson, G.J., *et al*. (1994) Modeling the effects of fuel injection characteristics on diesel engine soot and NOx emissions. SAE Paper 940523.

Peters, N. (2000) *Turbulent Combustion*, Cambridge University Press, Cambridge, UK.

Pitsch, H., Barths, H., and Peters, N. (1996) Three dimensional modeling of NOx and soot formation in DI-diesel engines using detailed chemistry based on the interactive flamelet approach. SAE Paper 962057.

Poinsot, T. and Veynante, D. (2005) *Theoretical and Numerical Combustion*, Edwards, Philadelphia, PA.

Pope, S.B. (1985) PDF methods for turbulent reactive flows. *Progress in Energy and Combustion Science*, **11**, 119–192.

Pope, S.B. (1997) Computationally efficient implementation of combustion chemistry using in situ adaptive tabulation. *Combustion Theory and Modeling*, **1**, 41–63.

Pope, S.B. (2000) *Turbulent Flows*, Cambridge University Press, Cambridge, UK.

Puduppakkam, K.V., Liang, L., Naik, C.V., *et al*. (2011) Use of detailed kinetics and advanced chemistry-solution techniques in CFD to investigate dual-fuel engine concepts. SAE paper 2011-01-0895.

Ra, Y. and Reitz, R.D. (2008) A reduced chemical kinetic model for IC engine combustion simulations with primary reference fuels. *Combustion and Flame*, **155**, 713–738.

Reitz, R.D. and Bracco, F.V. (1983) Global kinetics and lack of thermodynamic equilibrium. *Combustion and Flame*, **53**, 141–145.

Reitz, R.D. and Kuo, T.W. (1989) Modeling of HC emissions due to crevice flows in premixed-charge engines. SAE Paper 892085.

Reitz, R.D. and Rutland, C.J. (1995) Development and testing of diesel engine CFD models. *Progress in Energy and Combustion Science*, **21**, 173–196.

Reitz, R.D. and Sun, Y. (2009) Advanced computational fluid dynamics modeling of direct injection engines in *Advanced Direct Injection Combustion Engine Technologies and Development* (ed. H. Zhao) **2**: Diesel engines, Chapter 18, Woodhead Publishing Ltd., Cambridge, pp. 676–707.

Richard, S., Colin, O., Vermorel, O., *et al*. (2007) Towards large eddy simulation of combustion in spark ignition engines. *Proceedings of the Combustion Institute*, **31** (2), 3059–3066.

Rutland, C.J. (2011) Large-eddy simulations for internal combustion engines - a review. *International Journal of Engine Research*, **12** (5), 421–451.

Sethian, G.A. (1999) *Level Set Methods and Fast Marching Methods*, Cambridge University Press, Cambridge, UK.

Shi, Y., Ge, H.-W., and Reitz, R.D. (2011) *Computational Optimization of Internal Combustion Engines*, Springer, London, UK. ISBN: 978-0-85729-618-4

Singh, S., Reitz, R.D., and Musculus, M.P.B. (2006) Comparison of the characteristic time (CTC), representative interactive flamelet (RIF), and direct integration with detailed chemistry combustion models against optical diagnostic data for multi-mode combustion in a heavy-duty DI diesel engine. SAE Paper 2006-01-0055.

Smagorinsky, J. (1963) General circulation experiments with the primitive equations: I. the basic experiment. *Monthly Weather Review*, **91** (3), 99–164.

Spalding, D.B. (1971) Mixing and chemical reaction in steady confined turbulent flames. *Proceedings of the Combustion Institute*, **13**, 649–657.

Sun, Y. (2007) Diesel combustion optimization and emissions reduction using adaptive injection strategies (AIS) with improved numerical models. PhD Thesis. University of Wisconsin-Madison.

Tan, Z. and Reitz, R.D. (2003) Modeling ignition and combustion in spark-ignition engines using a level set method. SAE Paper 2003-01-0722.

Tan, Z. and Reitz, R.D. (2004) Development of a universal turbulent combustion model for premixed and direct injection spark/compression ignition engines. SAE Paper 2004-01-0102.

Tao, F., Liu, Y., RempelEwert, B.H., *et al*. (2005) Modeling the effects of EGR and injection pressure on soot formation in a high-speed direct-injection (HSDI) diesel engine using a multi-step phenomenological soot model. SAE Paper 2005-01-0121.

Thobois, L., Lauvergne, R., and Poinsot, T. (2007) Using LES to investigate reacting flow physics in engine design process. SAE Paper 2007-01-0166.

Trouve, A. and Poinsot, T.J. (1994) The evolution equation for the flame surface density in turbulent premixed combustion. *Journal of Fluid Mechanics*, **278**, 1–31.

Vermorel, O., Richard, S., Colin O., *et al*. (2007) Multi-cycle LES simulations of flow and combustion in a PFI SI 4-valve production engine. SAE Paper 2007-01-0151.

Vermorel, O., Richard, S., Colin, O., *et al*. (2009) Towards the understanding of cyclic variability in a spark ignited engine using multi-cycle LES. *Combustion and Flame*, **156** (8), 1525–1541.

Vishwanathan, G. and Reitz, R.D. (2010) Development of a practical soot modeling approach and its application to low temperature diesel combustion. *Combustion Science and Technology*, **182** (8), 1050–1082.

Wang, H., Reitz, R.D., and Yao, M. (2012) Comparison of Diesel Combustion CFD Models and Evaluation of the Effects of Model Constants. SAE Paper 2012-01-1134.

Wiebe, I.I. (1956) Semi-empirical Expression for Combustion Rate in Engines. *Proceedings of Conference on Piston engines*, USSR, 185–191.

Wiebe, I.I. (1962) Progress in engine cycle analysis: combustion rate and cycle processes. Mashgiz, Ural-Siberia Branch, 271.

Wiedenhoefer, J.F. and Reitz, R.D. (2003) Multidimensional modeling of the effects of radiation and soot deposition in heavy-duty diesel engines. SAE Paper 2003-01-0560.

Williams, F.A. (1985) *Turbulent Combustion*, SIAM, Philadelphia.

Woschni, G. (1967) Universally applicable equation for the instantaneous heat transfer coefficient in the internal combustion engine. SAE Paper 670931.

Yi, Y., Hessel, R., Zhu, G., and Reitz, R.D. (2000) The influence of physical input parameter uncertainties on multidimensional model predictions of diesel engine performance and emissions. *SAE Transactions, Journal of Engines*, **109** (3), 1298–1316.

Yakhot, V. and Orszag, S.A. (1986) Renormalization group analysis of turbulence I. Basic theory. *Journal of Scientific Computing*, **1** (3), 3–51.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Zellat, M., Duranti, S., Liang, Y., *et al*. (2003) Towards a universal combustion model in STAR-CD for IC engines: From GDI to HCCI and application to DI diesel combustion optimization. 13th International Multidimensional Engine Modeling Users' Group Meeting, Detroit, MI.

Zhang, Y. and Rutland, C.J. (2012) A mixing controlled direct chemistry (MCDC) model for diesel engine combustion modelling using large eddy simulation. *Combustion Theory and Modelling,* **16** (3), 571–588.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto124
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fuels for Engines and the Impact of Fuel Composition on Engine Performance

Charles J. Mueller[1], William J. Cannella[2] and Gautam T. Kalghatgi[3]

[1] *Sandia National Laboratories, Livermore, CA, USA*
[2] *Chevron Corporation, Richmond, CA, USA*
[3] *Saudi Aramco, Dhahran, Saudi Arabia*

1 Introduction 1
2 Fuel Production 2
3 Fuel Composition 5
4 Key Fuel Properties 8
5 Fuel Specifications 21
6 Summary and Outlook 23
References 24

1 INTRODUCTION

The chemical composition of a fuel determines its properties, and the properties of a fuel determine its engine performance characteristics. Historically, the liquid-phase hydrocarbon (HC) fuels used for transportation and nonroad applications have been refined from crude petroleum recovered from beneath the surface of the Earth. HCs are compounds that contain only the elements hydrogen and carbon. A compound is defined as a pure chemical substance composed of two or more elements bonded in a specific structural arrangement and exhibiting fixed ratios of the different elements. The word "petroleum" is derived from Latin roots; its literal translation is "rock oil."

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

As supplies of readily accessible conventional petroleum are dwindling, the cost of petroleum production is increasing, and environmental and political concerns related to petroleum use are growing. This trend has intensified interest in "renewable" fuels that are currently created from biomass, and in the future might be created directly using solar energy, atmospheric carbon dioxide (CO_2), and water (H_2O). Renewable fuels can be beneficial from a greenhouse gas (GHG) perspective because they essentially recycle CO_2 (a GHG). For example, plants absorb carbon in the form of atmospheric CO_2 as they grow, which offsets the CO_2 released when fuels made from plants are burned. From a GHG perspective, this is preferable to continuing to convert petroleum carbon into new atmospheric CO_2. Although the histories of renewable fuels are significantly different from those of petroleum fuels, renewable fuels can contain many of the same kinds of chemical structures found in petroleum-based fuels, depending on how they are produced. One notable difference is that some common renewable fuels on the market today are oxygenates, that is, compounds that contain oxygen in addition to hydrogen and carbon.

Liquids composed primarily of hydrogen, carbon, and perhaps some oxygen are fuels of choice for transportation applications because of the high amounts of energy they contain per unit mass (specific energy) and per unit volume (energy density). For example, the specific energy of diesel fuel is approximately an order of magnitude larger than that of dynamite. This is primarily because effectively the entire mass of diesel fuel is fuel elements, while the mass of an explosive such as dynamite includes the mass of the oxidizer as well as fuel elements, and the oxidizer does not

carry any extra energy. The specific energies of liquid HC fuels also are many times larger than those of advanced batteries. This high specific energy means that only a small mass of fuel needs to be carried aboard a vehicle or machine to produce a large amount of useful mechanical work. For at least a century, the reciprocating internal combustion engine has been the device of choice for converting the chemical energy stored in liquid fuels into mechanical work for ground transportation and nonroad applications.

Currently, reciprocating internal combustion engines for passenger car and commercial applications fall into either of two broad categories: spark-ignition (SI) engines or compression-ignition (CI) engines. SI engines tend to be used in light-duty applications (e.g., passenger cars) because they are less expensive, quieter, and have lower exhaust emissions [thanks to the highly efficient three-way catalyst (Mooney, 2007)] than their CI counterparts. CI engines tend to be used in heavy-duty applications (e.g., truck, rail, and marine transportation) because of their higher efficiencies (i.e., lower fuel consumption per unit work output and hence lower fuel costs), better durability, and superior low speed torque capabilities.

The specific effects of changing a given fuel property depend upon many detailed parameters of the engine and combustion strategy in which the fuel is used, as well as the conditions over which the engine is operated. The fuel properties required for robust operation of SI and CI engines are different; as a result, different fuels have evolved for use with these different engines. SI engines typically burn gasoline, while CI engines typically burn diesel fuel. The primary differences between gasolines and diesel fuels are that gasolines are resistant to autoignition (i.e., spontaneous reaction of a mixture caused by elevated temperatures and/or pressures) and are more volatile (i.e., require less energy for vaporization and have lower boiling temperatures), while diesel fuels readily autoignite and are less volatile.

Covering all aspects of fuels for reciprocating internal combustion engines and the impact of fuel composition on engine performance in a single chapter of reasonable length is impossible. Hence, the objective of this chapter is to provide a broad overview of the relevant fuel science and a flavor of the key issues, while going into detail in selected areas of particular importance. The chapter starts with a discussion of fuel production techniques and fuel compositional attributes, followed by an overview of fuel properties that affect end-use characteristics and a discussion of the fuel specifications developed to ensure that fuels perform as required in their applications. The chapter concludes by summarizing the main points and offering some thoughts on what the future might hold for fuels and reciprocating internal combustion engines.

2 FUEL PRODUCTION

Fuels and fuel components can be produced from a wide variety of feedstocks including crude oil, oil sands, natural gas, biomass, coal, oil shale, methane hydrates, and even carbon oxides (CO and CO_2) reacted with hydrogen. This section reviews the primary fuel production pathways currently in use.

The huge scale of fuel production is a key factor to keep in mind when evaluating conventional and alternative technologies. In 2011, global oil consumption was ~88 million barrels per day (British Petroleum, 2012), and US consumption was ~19 million barrels per day (U.S. Energy Information Administration, 2012). At 42 US gallons per barrel, global oil consumption is ~3.7 billion gallons per day, or nearly four Olympic-sized swimming pools per minute, representing a market value of ~$9 billion per day at the present crude oil cost of ~$100 per barrel. Most of this oil is used to produce transportation fuels, with the balance used for heating oil and industrial processes.

2.1 Crude oil refining

The vast majority of the liquid HC components of fuels today are produced from petroleum-derived crude oil. Crude oils from different geographic locations and production processes can have a wide range of properties: from low densities and viscosities (“light”) to high densities and viscosities similar to tar (“heavy”), and from low sulfur (“sweet”) to high sulfur (“sour”). All crude oils are primarily composed of HCs of the alkane and aromatic classes (Section 3).

Refining is the process of converting crude oil into higher value products, including fuels for reciprocating internal combustion engines. A typical modern refinery is a complex combination of interdependent processes that can be divided into three basic categories: separation, upgrading, and conversion (Bacha *et al*., 2007; Gibbs *et al*., 2009).

Distillation is the most important and widely used separation process in a refinery. Distillation is a process at the front end of a refinery that divides crude oil into different boiling fractions. Petroleum fractions that undergo little or no processing beyond distillation are called *straight-run* products. In the early days of petroleum refining (when fuel demand was low, engines had low compression ratios, and sulfur specifications were nonexistent), refineries produced mostly straight-run products, but today significant further processing of the individual fractions is typically necessary to provide fuels with the required properties in quantities aligned with market demand.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Upgrading processes remove undesirable impurities such as sulfur, nitrogen, and metal compounds from fuel feedstocks. For example, bitumen (the tar-like HC derived from oil sands and a key feedstock for fuels in North America) requires extensive upgrading, after which it becomes "synthetic crude" that can be processed in a conventional refinery (Yui, 2008). The most common upgrading techniques in use today involve hydrotreating, a broad class of strategies involving reactions of the fuel species with hydrogen in the presence of a catalyst, usually at conditions of elevated temperature and pressure.

Conversion processes are used to enable the refinery to produce fuels meeting specifications in quantities that match the market demand. One class of conversion processes is cracking. Fluid catalytic cracking (FCC) and hydrocracking are used to lower ("crack") the molecular weights of the heavier fractions from the distillation column to those appropriate for the various fuel classes. Other conversion processes such as reforming and alkylation increase the octane number (ON) to levels suitable for gasoline by producing aromatics and branched alkanes, respectively.

Base fuels are produced by blending various intermediate streams in the correct proportions at the refinery. Other components can be blended into a base fuel at a pipeline or rail terminal downstream of the refinery. The blending of additives and/or other components (e.g., ethanol or biodiesel) results in finished fuels, which are subsequently distributed to their various points of sale.

2.2 Fischer–Tropsch processes

Another method of producing fuel components is to synthesize them from a mixture of carbon monoxide (CO) and molecular hydrogen (H_2), called *synthesis gas* or simply syngas. Such catalytic processes typically create HCs (e.g., alkanes, see Section 3.1) plus water (H_2O) as a by-product:

$$(2n+1)H_2 + nCO \rightarrow C_nH_{2n+2} + nH_2O \qquad (1)$$

The process name is derived from two German researchers who discovered effective catalysts and operating conditions to facilitate the reaction in Equation 1 and others like it. Depending on the specific catalysts and reaction conditions used, oxygenates also can be produced by Fischer–Tropsch (or simply F–T) synthesis. Syngas can be produced from coal, natural gas, or biomass feedstocks. Germany commercialized the process using coal to produce fuels during World War II. In South Africa during an oil embargo in the 1970s, Sasol made further technology developments and built F–T plants to produce fuels from coal, which remain operational even today. Several plants have recently been built in which natural gas is the feedstock that is converted to syngas. This can be advantageous because it can convert natural gas that might otherwise be stranded (uneconomical to transport to a market) or flared (burned in an open flame to reduce its global warming impact) into high value liquid fuels.

F–T synthesis can be used to produce diesel or gasoline fuels. Diesel fuel production is maximized by operating at lower process temperatures that favor the creation of waxy, long-chain, normal alkanes, and subsequently cracking the chains into the diesel range. To improve the cold-temperature properties of the diesel fuel, the normal alkanes are "isomerized" to produce branched alkanes (Section 3.1.2). The resulting product has a high cetane number (CN) (70+, Sections 4.1.2 and 4.1.2.2). The production of gasoline is maximized by operating at higher process temperatures and lower pressures.

One potential drawback of F–T fuels is that their near-zero aromatic contents can cause leakage issues for the elastomeric seals present in some fuel system that require aromatics to enable the seals to swell to form a tight fit (Section 4.4.2). Other potential disadvantages of F–T synthesis of fuels include the high capital costs for production, primarily due to the costs of syngas plants (at current fuel prices, F–T fuels are only economically viable if there is long-term availability of a large source of extremely low cost feedstock) and the significant emission of CO_2 during the process if the H_2/CO ratio of the syngas is sufficiently low (e.g., when using coal as a feedstock rather than natural gas). Regarding the last point, Figure 5.3.2-1b in Edwards, Larivé, and Beziat (2011) shows that the well-to-wheels (WTW) GHG emissions (Section 4.5.2) for coal-derived F–T diesel are more than twice as large as those for petroleum diesel or F–T diesel from natural gas.

2.3 Natural gas and liquefied petroleum gas

Natural gas and liquefied petroleum gas (LPG) are lighter petroleum-derived products. Natural gas can be used in compressed natural gas (CNG) or liquefied natural gas (LNG) forms. Compared to gasoline, these fuels have higher ONs and their combustion products have lower GHG footprints. Energy is required, however, for compressing to CNG and cooling to LNG. Furthermore, a principal component of natural gas is methane, and any leaked methane will have a global warming potential 21–25 times that of CO_2 over a 100-year time horizon [see Table 2.14 in Solomon *et al.* (2007)]. Although the lower energy densities of CNG, LNG, and LPG tend to limit the ranges of vehicles in which they are employed, these fuels are used in a number of Asian, South American, and European countries. In the United States, the larger personal vehicles and the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

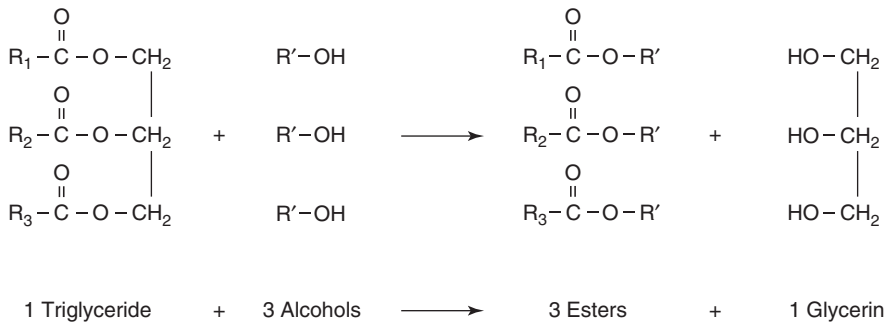


Figure 1. Transesterification reaction used to produce biodiesel esters (the reaction typically employs a catalyst).

limited distribution infrastructure for highly volatile fuels have largely restricted the use of CNG, LNG, and LPG to fleets with centralized refueling facilities. Nevertheless, the growing amounts of recoverable shale gas in some countries, and the corresponding decreased costs of natural gas and light condensed liquids relative to gasoline and diesel fuel, may lead to the wider use of these fuels in the future (NPC, 2012).

2.4 Biomass

2.4.1 Ethanol production

Ethanol is currently added to most of the gasoline sold in the United States and in many countries around the world. It is predominantly produced by fermentation (Schobert, 2013), and the primary feedstocks for ethanol production are the starches and sugars in crops such as corn, sugar beets, and sugar cane. The issues of whether food crops should be used to make fuel components (Section 4.5.4) and the limited life-cycle GHG reduction benefits of corn-based ethanol (Schobert, 2013, Chapter 4) have led to a significant amount of research and development on the conversion to ethanol of cellulosic materials found in sources such as wood and fast-growing nonfood crops such as switchgrass. Commercial production of ethanol from cellulosic material is presently negligible, however, relative to ethanol production from starches and sugars.

2.4.2 Biodiesel production

Biodiesel is an oxygenated bio-derived product composed of fatty acid methyl esters (FAMEs). Esters have the general structure R–(C=O)O–R′, where R and R′ signify HC chains. Typically, the R group is a straight HC chain with 5–21 carbon atoms and containing 0–3 carbon–carbon double bonds (i.e., the chain can be unsaturated, see Section 3.1.3). In the transesterification reactions that create biodiesel, the R-(C=O)O structures come from vegetable- or animal-derived triglycerides (or fatty acids), and R′ comes from an alcohol, as shown in Figure 1. The alcohol used in the reaction is commonly methanol, though sometimes ethanol or higher alcohols are used. Methyl esters are produced when methanol is used in the transesterification reaction, ethyl esters are produced when ethanol is used, and so on. The most common source of triglycerides for biodiesel in the United States is soybean-derived vegetable oil.

2.4.3 Renewable diesel production

An alternative to biodiesel production is to catalytically react vegetable- or animal-derived triglycerides (or fatty acids) with hydrogen instead of alcohols to produce renewable diesel: that is, alkane molecules that are essentially identical to, and completely fungible with, conventional petroleum-derived diesel. The hydroprocessing techniques required to produce these products are very similar to those currently used by refineries to upgrade crude oil, so conceivably these bio-derived feedstocks could be coprocessed with crude oil in refineries.

2.4.4 Other biofuel production approaches

A number of other possible paths are being explored to convert biomass to transportation fuels, and these can be grouped into two broad categories: thermochemical and biochemical. Thermochemical routes typically employ elevated temperatures (and/or pressures) and catalysts. Thermochemical routes such as pyrolysis (reactions occurring at elevated temperature, reduced oxygen content, and perhaps in the presence of catalysts) can convert even solid-phase biomass to liquids, but these liquids can require significant postprocessing (Christensen *et al*., 2011) before being suitable for use in modern engines (Mueller, 2013). Another thermochemical route, gasification, can be used

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to convert organic feedstocks to syngas, which can then be burned directly in engines or converted to liquid HCs through F–T type processes (Section 2.2).

Biochemical routes involve using microorganisms to produce fuels directly, or using microorganisms or their enzymes to break down biopolymers such as cellulose, hemicellulose, and lignin, as intermediate steps in the fuel production process. Examples include the direct production of HCs by *Botryococcus braunii* (Metzger and Largeau, 2005), methane production via anaerobic digestion, and conventional and cellulosic ethanol production processes via fermentation (Section 2.4.1).

3 FUEL COMPOSITION

Because the chemical composition of a fuel determines its properties, it is helpful to have some knowledge of the types of compounds found in modern fuels. Modern gasolines and diesel fuels have a high degree of compositional complexity, with typical fuel samples containing over 100 and 1000 distinct compounds, respectively. The compounds present in greatest abundance in current fuels are HCs and oxygenates, but trace components also are present, and they can have strong effects on engine performance. These three classes of fuel constituents are introduced in the context of liquid-phase fuels and discussed in this section. The section concludes with a brief discussion of gaseous fuel compositions.

3.1 Hydrocarbons

Crude petroleum and the fuels derived therefrom typically contain five classes of HC compounds: unbranched alkanes, branched alkanes, cyclic alkanes, olefins, and aromatics. The relative concentrations of the different HC classes in a given fuel are a function of the feedstocks and processing strategies employed in its production.

The total number of carbon atoms in a compound is known as its *carbon number*, denoted n. A representative molecule from each HC class having $n = 8$ is shown in Figure 2 using several common notational conventions. All elements are shown in the condensed structural formula for *n*-octane and the dash structural formulas for the other HCs in Figure 2. In contrast, only carbon bonds are shown in the bond-line structural formulas. In this latter notation, a carbon atom is assumed to exist at the end of each straight-line segment (unless the symbol of a different element is shown in its place), hydrogen atoms are not shown, two parallel line segments indicate a double bond between adjacent atoms, and a hexagon with a circle inside it indicates a benzene ring.

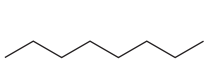
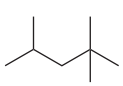
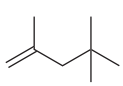
Hydrocarbon class	Unbranched alkane	Branched alkane	Olefin	Cyclic alkane	Aromatic
Representative compound	*n*-Octane	2,2,4-Trimethyl-pentane (*iso*-octane)	2,4,4-Trimethyl-1-pentene	Trans-1,2-dimethyl-cyclohexane	1,2-dimethyl-benzene (*ortho*-xylene)
Chemical formula	C_8H_{18}	C_8H_{18}	C_8H_{16}	C_8H_{16}	C_8H_{10}
Condensed or dash structural formula	$CH_3(CH_2)_6CH_3$	$CH_3-CH(CH_3)-CH_2-C(CH_3)_2-CH_3$	$CH_2=C(CH_3)-CH_2-C(CH_3)_2-CH_3$	CH_2-CH_2, CH_2, CH_2, $CH-CH$, CH_3, CH_3	$CH=CH$, CH, CH, $C-C$, CH_3, CH_3
Bond-line structural formula					

Figure 2. Representative molecules from the five HC classes with carbon number $n = 8$. Molecules within the gasoline boiling range have been selected for simplicity.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Different compounds with the same chemical formula are called *isomers*. As shown in Figure 2, *n*-octane and *iso*-octane are isomers, as are 2,4,4-trimethyl-1-pentene and *trans*-1,2-dimethylcyclohexane. These types of isomers, that is, those that differ in the sequences in which their atoms are arranged, are called *constitutional isomers*. Other types of isomerism exist, such as stereoisomerism, as discussed in the section on cyclic alkanes below. Isomerism is an important concept in fuel science because different isomers can have starkly different physical and chemical properties.

3.1.1 Unbranched alkanes

Also sometimes known as *normal alkanes*, *n-alkanes*, or *n-paraffins*, these "straight-chain" HCs have only single bonds between their carbon atoms, and their carbon atoms are all positioned along a single zigzagged but roughly linear chain (see the bond-line structural representation of *n*-octane in Figure 2). Their chemical formulas are of the form C_nH_{2n+2}, where n is the carbon number of the molecule. Unbranched alkanes are called *saturated compounds* because they contain the maximum possible number of hydrogen atoms for the given number of carbon atoms and general structure.

3.1.2 Branched alkanes

These branched-chain HCs are also commonly known as *iso-alkanes* or *iso-paraffins* (though technically the term *iso-alkane* refers only to an alkane with a single methyl group attached to the second carbon atom of its primary chain). Like unbranched alkanes, the branched alkanes are saturated HCs with chemical formulas of the form C_nH_{2n+2}, but the carbon atoms are no longer all arranged along a single linear chain. Instead, there are one or more alkyl branches extending from the main HC chain. For example, Figure 2 shows that *iso*-octane has three methyl branches, two from the second carbon atom of the main pentane chain and one from the fourth.

3.1.3 Olefins

Also sometimes known as *alkenes*, olefins are HCs that contain at least one carbon–carbon double bond. Olefins can exist as unbranched, branched, or cyclic structures. The olefin shown in Figure 2 has a branched structure with the double bond between the carbon atoms in the first and second positions. An olefin is an unsaturated compound because hydrogen atoms can be added to it without changing the general arrangement of its carbon atoms. For example, the olefin shown in Figure 2 could be converted to *iso*-octane by the addition of two hydrogen atoms. Although olefins are not typically found in crude petroleum in high concentrations, they can be produced during gasoline refining (e.g., via thermal cracking).

3.1.4 Cyclic alkanes

Also called *cyclo-alkanes* or *naphthenes*, cyclic alkanes are saturated compounds with their carbon atoms arranged in one or more ring structures. Cyclo-alkanes may have 3–9 carbon atoms in each ring structure, where the chemical formula of a single-ring cyclo-alkane is C_nH_{2n}. Five- and six-membered rings are the most stable (and hence the most abundant in petroleum-derived fuels) because of their lower carbon-bond strains. The carbon atoms in cyclo-alkane rings are not coplanar. In a five-carbon ring, one of the carbon atoms is raised above the plane of the other four. In a six-carbon ring, a number of conformations are possible, with the two most common being the "chair" and "boat" conformations. These two different conformations of cyclohexane are illustrated in Figure 3.

In these conformations, four of the carbon atoms are coplanar while the remaining two either lie on different sides of the plane (the chair conformation, which is the most stable) or on the same side of the plane (the boat conformation). The potential energy differences among the different cyclohexane conformations are low enough that the molecule can change conformations (or "interconvert") approximately 1 million times per second at standard conditions (Solomons, 1996). These and other conformational differences among molecules can lead to stereoisomerism, that is, molecules with the same chemical formula but different arrangements of their atoms in space that will not interconvert under standard conditions. Cyclo-alkanes with one or more unbranched or branched side chains, called *substituents*, are common and may exhibit stereoisomerism. When both substituents project above the nominal plane of the cyclohexane ring, the compound is a *cis* isomer; when one substituent projects above and one below, the compound is a *trans* isomer. For example, the *trans*-1,2-dimethylcyclohexane shown in Figure 2 has methyl substituents bonded to the carbon atoms in the 1 and 2

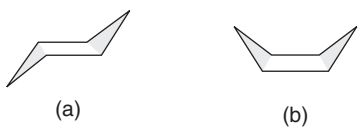


Figure 3. Conformations of cyclohexane. (a) Chair. (b) Boat.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

positions on the ring, with one projecting above the nominal plane of the cyclohexane ring and one below.

3.1.5 Aromatics

Aromatics are unsaturated compounds with their carbon atoms arranged in one or more ring structures. The most common aromatic HCs are those containing one or more benzene rings. A benzene ring has the formula C_6H_6. In contrast to cyclohexane, all six carbon atoms in a benzene ring lie in the same plane. The benzene rings in fuel-derived aromatic compounds can and often do have one or more alkyl substituents.

3.2 Oxygenates

The second most abundant class of compounds found in modern fuels is oxygenates. These are compounds composed of not just of hydrogen and carbon but also oxygen. The oxygenate classes most commonly used in fuels today are alcohols and esters, though other classes also have been and are used. Some common types of oxygenates are discussed below.

3.2.1 Alcohols

Alcohols have the general formula R–O–H, where R signifies an HC radical. Hence, the alcohol formed from the ethyl radical (C_2H_5) is ethyl alcohol, which is also known as *ethanol* (C_2H_5OH). Ethanol is by far the most commonly used oxygenate in gasolines, with nearly all gasoline sold in the United States currently containing some ethanol. Higher alcohols such as butanol isomers also are potential high octane bio-derived blendstocks for gasoline, and methods for scaling up their production are under development.

3.2.2 Esters

Esters have the general structure R–(C=O)O–R′, where R and R′ signify HC radicals. Esters are discussed in Section 2.4.2 on biodiesel production. Although esters most commonly enter fuels as biodiesel FAMEs, they also can be derived directly from various biological and other processes. Raw vegetable- or animal-derived lipids, however, typically are not esters. They are triglycerides and/or fatty acids.

3.2.3 Ethers

Ethers have the general structure R–O–R′, where, again, R and R′ denote HC radicals that may be different. The best-known fuel ether is methyl *tert*-butyl ether (MTBE), which was approved for use at levels up to 15 vol% in gasoline by the US Environmental Protection Agency (EPA) in the 1980s and added to Federal and California reformulated gasolines. This use was banned in California and a number of other states in 2004, after MTBE was found in certain drinking water supplies. Some humans can taste MTBE at concentrations below 0.1 ppm (parts per million) (U.S. Environmental Protection Agency, 2012a). Despite the US experience with MTBE, some countries continue to allow the use of MTBE in gasoline. Other ethers show some promise as additives for diesel fuel, because their oxygen content is beneficial for lowering soot emissions, and they can have CNs >60 (Upatnieks and Mueller, 2005).

3.2.4 Ketones

Ketones have the general structure R–(C=O)–R′. A common ketone is the solvent acetone, or dimethyl ketone: $CH_3(C{=}O)CH_3$. Ketones are not found in current fuels in high concentrations, but they can be produced efficiently by certain biological processes and hence may be important components of future biofuels.

3.3 Trace components

While trace components are, by definition, minor fuel constituents, they can have profound effects on engine performance through a number of mechanisms. These mechanisms include but are not limited to

- corrosion of fuel system components. For example, organic acids can react with iron, copper, and other materials in fuel systems to produce corrosion, which can lead to the failure of fuel injectors, pumps, and other fuel system components.
- deposit formation within fuel injectors, on engine valves, or inside the combustion chamber. For example, trace metals can catalyze liquid-phase polymerization reactions that lead to the formation of lacquers, gums, and products that are insoluble in the fuel that can foul close-tolerance fuel injector components.
- fuel-borne particulates that adversely affect fuel injector operation. For example, if the fuel filter is not functioning properly, fuel-borne particles can plug small orifices in the fuel injector and/or lead to accelerated wear of injector components.
- poisoning of exhaust after-treatment catalysts. For example, sulfur and phosphorus can cause active sites on catalysts in after-treatment devices to become temporarily or permanently inactive (see Section 4.4.3).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- raising the cloud point and/or plugging the fuel filter with crystals. For example, high melting-point impurities can be the first to freeze in colder climates, and their crystals can plug the fuel filter and thereby prevent the engine from running.
- causing fuel blends to separate into two or more distinct liquid phases. For example, even small quantities of water can cause some fuel blends, especially those containing both polar and nonpolar components (Section 4.3.2), to separate.
- decreasing autoignition delay time. For example, naturally formed peroxides can substantially decrease the autoignition-delay times of diesel fuels and gasolines. Potential issues associated with trace components are discussed in more detail in Section 4.

Not all trace components in fuels are detrimental to performance. Cetane improvers, oxidation inhibitors, lubricity improvers, metal deactivators, and detergents are examples of compounds that are routinely added to fuels at ppm levels to improve fuel performance in certain engine applications. The collection of these compounds used in a specific market fuel, called the *additive package*, depends on the base fuel characteristics, the engine application, the fuel provider, and economic factors.

3.4 Gaseous fuels

Gaseous fuels face two primary challenges in transportation and other mobile-engine applications. The first is that they tend to be difficult to be stored at high energy density, often requiring high pressure fuel tanks and/or cryogenic systems. The second is that they tend to lack a distribution infrastructure for transportation applications, that is, there is no nationwide network of gaseous-fuel filling stations as exists for gasoline and diesel fuel. Nevertheless, gaseous fuels are an important piece of the puzzle of meeting global energy demand. The gaseous fuels of primary interest are natural gas and hydrogen.

3.4.1 Natural gas

Natural gas is a mixture of HCs with carbon numbers typically in the range from 1–4 and existing in vapor form at standard conditions. The CN of natural gas is typically very low (<0), but its ON is typically high (>110), and its specific composition can vary widely depending on its geographical origin. The amount of recoverable natural gas in the world has dramatically increased in recent years because of the employment of hydraulic fracturing ("fracking") techniques for its extraction from subterranean deposits.

3.4.2 Hydrogen

There has been a great deal of interest in hydrogen because it does not contain carbon and hence its combustion does not produce the GHG CO_2. Nevertheless, the world currently lacks inexpensive and abundant sources of hydrogen, an established infrastructure for its distribution, and low cost techniques for storing it at high energy density. Until these significant barriers are overcome, hydrogen is likely to remain a fuel of the future.

4 KEY FUEL PROPERTIES

From a high level perspective, the processes occurring in modern internal combustion engines are quite simple: oxygen from the air surrounding the engine is mixed with a fuel in a carefully controlled fashion, and both participate in combustion reactions that produce elevated pressure and temperature in the engine cylinder, plus major and minor product species. The high pressure created within each engine cylinder during combustion causes the engine pistons to move, and this movement transforms some fraction of the chemical energy from the fuel into the desired power at the crankshaft.

Dry air is composed of approximately 78.08 mol% molecular nitrogen (N_2), 20.95 mol% molecular oxygen (O_2), and <1 mol% argon, CO_2, and other species. When a liquid fuel composed of hydrogen, carbon, and perhaps some oxygen, nitrogen, and sulfur is burned in an engine, the vast majority of the atoms react to form the major product species CO_2 and H_2O. The CO_2 and H_2O pass out of the engine through the exhaust valves with N_2 from the intake air, perhaps some O_2 that was not consumed during the in-cylinder reactions (depending on the combustion strategy), and other minor product species including nitrogen oxides (NO_x), particulate matter (PM), carbon monoxide (CO), sulfur oxides (SO_x), and products of incomplete combustion (also known as *unburned HCs*, see Thermodynamic Analysis). The amounts of the minor species that leave the engine can depend strongly on fuel characteristics.

While the processes above are straightforward from a high-level perspective, the devil is in the details, and there are a great many mechanisms by which fuel properties can affect engine operation in subtle and not-so-subtle ways. In this regard, it is sometimes helpful to distinguish between fuel properties that are determined by the bulk fuel composition (e.g., density, aromatic content, distillation curve) versus those that can be significantly affected by trace levels (~1% or less) of certain compounds in the fuel (e.g., autoignition quality, lubricity, sulfur content). For

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

instance, properties in the latter group sometimes can be improved using fuel additives, while those in the former group cannot. It is also helpful to recognize that fuel properties are often highly correlated (e.g., CN, density, and sooting propensity for diesel fuels). As a result, it can be very challenging to determine the effects of changing a certain fuel property when other fuel properties necessarily change with the property of interest.

The objective of this section is to provide an overview of a number of key fuel properties and explain why they are important. The broad categories treated below are combustion properties, physical properties, stability, materials compatibility, and environmental considerations.

4.1 Combustion properties

The primary combustion properties of a reciprocating-engine fuel are its energy content, its autoignition quality, and its emissions-formation characteristics. These topics are treated in turn in the sections below.

4.1.1 Energy content

As mentioned in the introduction, the energy content of a fuel, that is, the amount of chemical energy available per unit mass (specific energy) or per unit volume (energy density), is a key factor in mobile-engine applications, because it plays the primary role in determining the mass and volume of fuel that must be carried aboard the vehicle or machine to satisfy the application requirements. Also, because fuels are typically sold by volume, energy density typically influences operating costs. One way to better understand the effects of fuel energy content is to consider the similarities and differences among the concepts of efficiency, fuel economy (FE), and specific fuel consumption (SFC).

4.1.1.1 Efficiency, fuel economy, and specific fuel consumption. Other chapters of this encyclopedia cover engine and vehicle efficiency, FE, and SFC in greater detail (see Zero- and One-Dimensional Methodologies and Tools, Thermodynamic Analysis, Trends—Spark Ignition, and Trends—Compression Ignition), but certain aspects that are directly related to fuel effects bear mentioning here. For instance, an engine might run more efficiently when fueled with ethanol, but the corresponding FE could be lower and the SFC higher. How could this be? To answer this question, the definitions of efficiency, FE, and SFC must be considered within the context of variable fuel properties.

Efficiency, denoted η, is simply the engine work output per unit of chemical energy input, that is,

$$\eta = \frac{W_{\text{out}}}{m_{\text{f}} q_{\text{LHV}}} \tag{2}$$

where W_{out} is the work output per cycle of the engine, m_{f} is the mass of fuel supplied to the engine per cycle, and q_{LHV} is the lower heating value of the fuel (i.e., the specific energy of the fuel assuming the water in the products remains in the vapor phase). From a thermodynamic standpoint, an optimal engine design is one with maximum efficiency. Nevertheless, liquid fuels are priced on a volumetric basis rather than an energy-content basis, and hence the engine user often cares most about the FE of the vehicle, that is, the distance that can be traveled per unit volume of fuel. Because the distance that can be traveled is proportional to the work output of the engine, Equation 2 can be used to show that

$$\text{FE} \propto \frac{W_{\text{out}}}{V_{\text{f}}} = \frac{\eta m_{\text{f}} q_{\text{LHV}}}{m_{\text{f}}/\rho_{\text{f}}} = \eta \rho_{\text{f}} q_{\text{LHV}} \tag{3}$$

In Equation 3, V_{f} and ρ_{f} are the volume of fuel injected per cycle and the fuel density, respectively. So while the FE is proportional to the efficiency, as expected, it is also proportional to the fuel density and specific energy. Similarly, SFC is often used to quantify efficiency. SFC is the mass of fuel required per unit of work output, and again using Equation 2

$$\text{SFC} = \frac{m_{\text{f}}}{W_{\text{out}}} = \frac{1}{\eta q_{\text{LHV}}} \tag{4}$$

The presence of the specific-energy term in Equation 4 indicates that, like FE, SFC is not a direct measure of efficiency. The subtle differences among η, FE, and SFC can be the source of considerable confusion.

Consider a gasoline-fueled vehicle with an FE of 40 mpg (miles per gallon) and an engine efficiency of 30%. Assuming that the same engine running on neat ethanol achieves an efficiency of 40%, a substantial thermodynamic improvement, the corresponding FE would only be 35.6 mpg. (This analysis assumes that the specific energies of gasoline and ethanol are 42.6 and 27.0 MJ/kg, respectively; the densities of gasoline and ethanol are 749 and 789 kg/m^3, respectively; and all other vehicle parameters are the same.) The lower FE for neat ethanol is because even though the efficiency and density terms in Equation 3 increase when switching from gasoline to neat ethanol, the specific energy of ethanol is low enough to counteract these effects. Hence, even with the increased engine efficiency of neat ethanol and if neat ethanol were available for 10% less

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

cost per unit volume than gasoline, it would not make financial sense to switch from gasoline to neat ethanol based on the fuel cost per mile traveled. In this example, the lower specific energy of ethanol also leads to a higher SFC than that achieved with the gasoline-fueled vehicle. (Note: This is an example that was presented to illustrate some of the potential issues when considering the use of fuels with different energy contents; it is not a true cost/benefit comparison between gasoline and neat ethanol because the efficiency and FE values used were only hypothetical.)

As illustrated in the above example, efficiency (as defined in Equation 2) is the best parameter to use when evaluating fuel effects on the ability of an engine to convert chemical energy into useful work, while FE is likely a better parameter to use for fuel cost comparisons, and SFC might be desirable in situations that are very sensitive to the mass of fuel that must be carried aboard a vehicle.

4.1.2 Autoignition quality

Autoignition has been defined as "a rapid combustion reaction that is not initiated by any external ignition source" (Heywood, 1988). The susceptibility of a fuel to autoignition is called its *autoignition quality*. ONs and CNs are typically used to quantify the autoignition qualities of fuels for SI and CI engines, respectively. Fuels with high ONs are generally resistant to autoignition, and are hence desirable for SI engine applications because autoignition can lead to detrimental knock and preignition, as described in detail in the following section. Fuels with high CNs autoignite easily, and hence are generally desirable for CI engine applications, which rely on autoignition to initiate combustion. On the basis of these characteristics, high ON fuels would be expected to have low CNs, and vice versa, and this is observed in practice. Although ON and CN vary roughly inversely, there is no broadly accepted quantitative relationship between the two parameters.

Whether or not a given fuel mixture autoignites at a specific temperature/pressure condition depends on details of the chemical-kinetic reactions that lead to ignition, and these are strongly dependent on the fuel's molecular structure. An in-depth discussion of these processes is beyond the scope of this chapter but can be found in Fundamental Chemical Kinetics. In general, the presence or absence of certain molecular-structural functional groups can be used to roughly estimate the autoignition quality of a given compound. Straight-chain alkanes tend to autoignite most easily, especially as chain length increases beyond a few carbon atoms. Aromatics and highly branched alkanes/alkenes tend to be the most difficult to autoignite, with cyclo-alkanes being somewhat easier. Oxygenates can have varying effects depending on how the oxygen is bonded into the molecule; peroxides are used as CN improvers, but low molecular-weight alcohols such as ethanol and methanol have high ONs. In practice, the autoignition quality of a fuel is almost always measured because most fuel compounds contain more than one of these functional groups, most fuels are mixtures of many compounds that can interact during ignition, fuel composition is challenging to quantify, and models for estimating the autoignition qualities of mixtures have limited accuracies.

In SI engines, autoignition quality is arguably the single most important fuel property, determining the extent to which knock and preignition phenomena will limit engine performance. Owing to this critical role, fuel effects on SI engine knock and preignition are treated in detail in the following section. Autoignition quality also plays an important role in CI engines, affecting cold-startability, combustion noise, and emissions, and these are discussed after the section on SI engine effects.

4.1.2.1 Fuel autoignition quality effects in SI engines: knock and preignition.

4.1.2.1.1 Knock in SI engines. Knock is an abnormal combustion phenomenon that can damage SI engines (Heywood, 1988) and is determined by the integrated pressure and temperature history of the "end-gas" (i.e., the gas mixture ahead of the advancing flame front) and the antiknock or autoignition quality of the fuel. Knock occurs when the end-gas autoignites, centered on one or more "hot spots," causing a sharp rise in the heat release rate and setting up pressure waves in the cylinder that in turn manifest themselves as a metallic "pinging" sound. Knock intensity is usually defined as the maximum amplitude of this fluctuating pressure signal. When the knock intensity reaches some threshold value (it is common to use a value of 0.2 bar), the engine is said to knock (see Spark Ignition Combustion).

At a given engine operating condition, knock intensity increases as the spark timing is advanced, and the timing at which the engine knocks is known as the *knock-limited spark advance* (KLSA). The larger the KLSA, the more resistant a fuel is to knock at the particular operating condition. When the engine cannot be run at its maximum load and efficiency because of knock, it is said to be *knock-limited*. A modern car includes a knock detector and is knock-limited on normally available fuels over at least some of its operating range. When knock is detected, the engine management system takes corrective action (e.g., it retards the spark timing), and this usually reduces the power. In such engines, fuel antiknock quality also can be quantified by measuring the knock-limited power or acceleration performance at a given operating condition

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

(Kalghatgi, 2005; Kalghatgi, Nakata, and Mogi, 2005). Using a fuel with an antiknock character higher than that required to avoid knock will not generally further improve engine performance, though use of a higher octane fuel may improve FE if the higher octane is achieved by a higher mole fraction of aromatic carbon in the fuel, because this type of fuel will have a greater energy density (Section 4.1.1).

Fuel antiknock quality is described by empirical measures because the fundamental autoignition chemistry cannot be quantified adequately for practical fuels because of their compositional complexity. It is traditionally measured by research octane number (RON) and motor octane number (MON) of the fuel used. The RON test is run in a single-cylinder Co-operative Fuel Research (CFR) engine at an engine speed of 600 rpm and an intake temperature of 52 °C, while the MON test is run at 900 rpm with a higher intake temperature of 149 °C. These tests are run in strict accordance with established procedures (ASTM, 2010b, 2012a, b). The octane scale is based on two alkanes, *n*-heptane (ON ≡ 0) and *iso*-octane (ON ≡ 100). The blends of these components are referred to as *primary reference fuels* (PRFs) and define the intermediate points in the RON and MON scales. The RON or MON is the volume percent of *iso*-octane in the PRF. Thus, a blend of 90 vol% *iso*-octane and 10 vol% *n*-heptane is assigned the ON of 90 in both RON and MON scales. A real-world fuel is assigned the RON (or MON) value of the PRF that matches its knock behavior in the RON (or MON) test. All practical gasolines are more complex than PRF, and are mixtures of aromatics, olefins, naphthenes, oxygenates, and/or alkanes. A practical gasoline will match the PRF of a higher ON in the RON test and hence has a higher RON when compared to the MON test. The difference between the RON and MON of a fuel is known as its *sensitivity*. The sensitivity, S = RON − MON, is a measure of how different the fuel autoignition chemistry is from that of PRF.

4.1.2.1.2 Pressure and temperature development in the unburned gas in an engine. The temperature of the unburned gas increases as pressure increases in the engine cylinder. For instance, at the RON test condition, the temperature at a given pressure is lower than in the MON test condition, as shown in Figure 4 (Kalghatgi, 2005). Modern engines are "beyond RON" because the temperature of the unburned mixture, at a given pressure, is even lower than in the RON test. Alternatively, the pressure for a given temperature is higher in modern engines compared to the RON test. An example of such a condition "beyond RON" is also shown in Figure 4. This is a consequence of the higher efficiency of modern engines. Engine designers have always tried to force more air into the cylinder while minimizing the temperature in order to improve power density and efficiency. Indeed, all the modern trends aimed at improving the efficiency of SI engines, such as increasing the compression ratio, direct injection, and turbo-charging aligned with engine downsizing, will push modern engines further "beyond RON."

4.1.2.1.3 Antiknock quality of gasolines in modern engines. Practical fuels contain components other than alkanes, and they behave like different PRFs at different engine conditions. Thus, the knocking characteristics of a gasoline of 95 RON and 85 MON will match those of an 85 PRF in the MON test, while it will be much less prone to knock, compared to PRF fuels, in the RON test condition and will match 95 PRF. This is because the autoignition chemistry of non-alkane components in gasoline is different from that of PRF. The antiknock quality of practical fuels in an engine is studied by measuring a parameter dependent on the autoignition quality of the fuel such as KLSA. At the same operating condition, many fuels of different chemical compositions and hence different RON and MON, are compared. In such autoignition studies, homogeneous charge compression ignition (HCCI) engines have been valuable complements to knocking SI engines because the fundamental mechanisms of knock and HCCI combustion are the same, as summarized in Kalghatgi (2005).

The true antiknock quality of a gasoline is best described by an octane index, OI (Kalghatgi, 2005; Kalghatgi, Nakata, and Mogi, 2005), which is the ON of the PRF that matches the knock behavior of the gasoline of interest at the conditions of the particular engine test. It is defined as

$$\mathrm{OI} = (1-K) \times \mathrm{RON} + K \times \mathrm{MON} = \mathrm{RON} - K \times S \quad (5)$$

where K is an empirical constant that depends only on the pressure and temperature history of the unburned mixture in the cylinder, and S is the sensitivity. The value of K depends on the design and operating condition of the engine. Thus, $K = 1$ in the MON test, and the antiknock quality of the gasoline is described by its MON, but $K = 0$ in the RON test, which runs at a lower temperature for a given pressure (Figure 4). Modern engines are "beyond RON" and K is negative, that is, for a given RON, a lower MON, higher sensitivity fuel has higher OI and is more resistant to knock. For instance, if $K = -0.5$, a gasoline with 95 RON and 85 MON will have an OI of 100 and will actually match a PRF of 100 (i.e., pure *iso*-octane) for knock. Modern engines usually have negative K values (Coordinating Research Council, 2011; Bell, 2010; Kalghatgi, 2005; Kalghatgi, Nakata, and Mogi, 2005; Mittal and Heywood, 2008).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

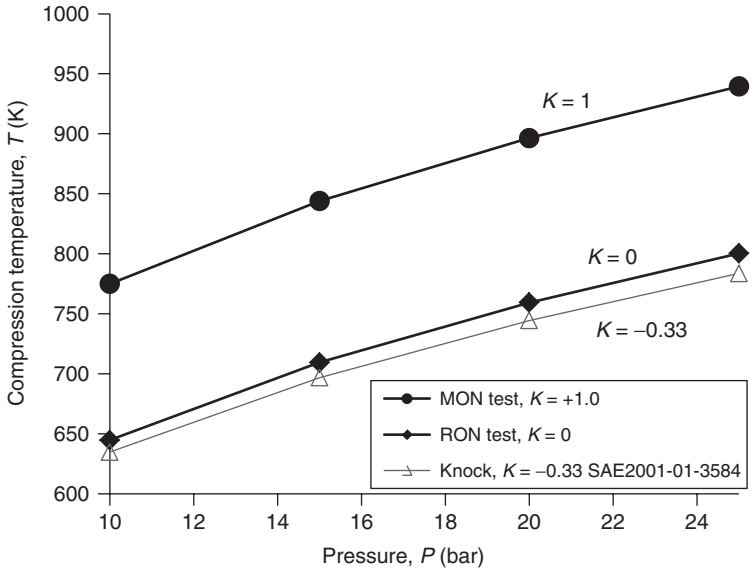


Figure 4. Unburned mixture temperature versus pressure. K decreases from 1 to 0, moving from the MON test condition to the RON test condition and becomes negative "beyond RON." The true antiknock quality is given by $\mathrm{OI} = (1-K) \times \mathrm{RON} + K \times \mathrm{MON}$ (Equation 5).

4.1.2.1.4 The value of K has been decreasing throughout recent history. The RON test was invented in 1930 when the average compression ratio was around 5 (Mittal and Heywood, 2009) and the engines had poor efficiency compared to modern engines. It was soon obvious that the RON test did not predict the antiknock quality of fuels on the road at that time. Hence, the MON test was developed, and it reflected the antiknock behavior of fuels in cars of that time much better. Clearly, the K value of the US car fleet then was around 1. The Coordinating Research Council (CRC) in the United States has conducted frequent surveys to establish the octane requirements of US cars between 1947 and 1996. On the basis of these data and other available sources, Mittal and Heywood (2009) showed that the K value has decreased from around 1 to 0 in 2008 because of reduced end-gas temperatures and higher pressures. A similar decreasing trend for K in the UK car fleet also has been reported in Kalghatgi (2005). Historically, in the United States, fuel antiknock quality has been assumed to be best described by (RON + MON)/2, which translates into $K = 0.5$. The US car fleet, on average, used to have such an octane requirement when this specification was first agreed upon; however, engine technology has moved on and with it, the octane requirement.

4.1.2.1.5 Why is K negative in modern engines? The autoignition chemistry of non-alkane components in practical fuels is significantly different from that of the PRFs used in the RON and MON scales. In general, for a given temperature T, a non-alkane fuel becomes more resistant to autoignition compared to an alkane fuel, as the pressure P increases. A fundamental measure of autoignition quality of a combustible mixture is the ignition delay τ, which is typically measured outside the engine in special equipment such as shock tubes or rapid compression machines (Fikri *et al*., 2008; Gauthier, Davidson, and Hanson, 2004; Heywood, 1988). The smaller the value of τ, the more reactive the mixture. In general, τ can be expressed as a function of temperature and pressure (Fikri *et al*., 2008; Gauthier, Davidson, and Hanson, 2004) in a given pressure/temperature range.

$$\tau = \tau_0 f (T/T_0)(P/P_0)^{-n} \tag{6}$$

Here, τ_0 is the ignition delay measured at some reference pressure P_0 and a reference temperature T_0. It is found experimentally that the value of the pressure exponent n is much smaller in magnitude for non-alkane fuels than for alkane fuels (Fikri *et al*., 2008; Gauthier, Davidson, and Hanson, 2004) at high temperatures ($T > \sim 850$ K). Hence, non-alkane fuels become relatively more resistant to autoignition if the pressure is increased at a fixed temperature. Why this should be so, at the fundamental chemical-kinetic level, is not yet clear. The negative value of K could only be explained in HCCI engines (Bradley, Morley, and Walmsley, 2004) and in knocking SI engines (Kalghatgi, Nakata, and Mogi, 2005) by this difference in the value of the pressure exponent n between alkane and non-alkane fuels.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.1.2.1.6 Specifications for fuel antiknock quality and engine octane requirement. High MON is considered to contribute to the antiknock quality of a gasoline in many areas, for example, in North America and in Europe. In Europe, there is a minimum MON specification of 85, and in North America MON is considered to be as important as RON. However, as discussed above, modern engines have pressure/temperature development regimes that make MON at best irrelevant ($K = 0$) or actually detrimental to fuel antiknock quality ($K < 0$). Thus, these octane specifications are inconsistent with the actual requirements of modern engines. Moreover, as improved engine efficiency is sought, this mismatch between specifications and engine requirements will get wider and will have to be addressed. Steps taken to improve engine efficiency such as downsizing and turbocharging will in general make future engines more prone to knock, necessitating higher antiknock quality in the fuel. Such engines will also be "beyond RON," and the K value will be negative so that for a given RON, a lower MON fuel will be more resistant to knock (Amer *et al*., 2012). The need to bring fuel specifications in line with engine requirements will increase.

4.1.2.1.7 Preignition and "Super-knock". Preignition is an abnormal combustion phenomenon where an expanding flame front is established before the spark plug fires in the engine. Preignition has been of intermittent concern at different stages of the history of SI engines. In the 1950s and early 1960s, compression ratios increased rapidly and preignition became a serious problem. Preignition again seems to have attracted attention in the 1980s primarily because of the problems encountered with using methanol after a relatively quiet period in the 1960s and 1970s. With the trend in downsizing and turbocharging, it has again become a concern in turbocharged DISI (direct injection spark ignition) engines (Kalghatgi *et al*., 2009; Zahdeh *et al*., 2011). The early start of combustion from preignition causes the pressure and the temperature of the unburned gas ahead of the advancing flame front, the end-gas, to rise more rapidly compared to normal spark timing. If autoignition occurs in the end-gas at high pressure and temperature, it can lead to extremely heavy knock, with knock intensities of 100 bar or more. Such events are informally described as *super-knock*, another abnormal combustion phenomenon that could potentially damage the engine (Kalghatgi and Bradley, 2012; Kalghatgi *et al*., 2009).

Preignition is a flame initiation problem and two separate criteria have to be met in order for it to occur. The first is an ignition criterion, which requires that the temperature must reach a minimum level locally so that runaway chemical reactions start. This might occur because of hot spots on the internal surfaces of the engine or because of autoignition and catalytic reactions centered around oil droplets or particulates in modern DISI engines. The additional initiation criterion requires that the incipient flame must reach a critical radius that is proportional to the laminar flame thickness δ before it becomes self-sustaining (Kalghatgi and Bradley, 2012). The smaller the value of δ, the larger is the probability of flame initiation and preignition, all else being equal. In the experiments reported in modern DISI engines (Dahnz *et al*., 2010; Zahdeh *et al*., 2011), the probability of both criteria being satisfied is low as preignition only occurs every 15,000 cycles or so.

δ can be related to pressure, temperature, and fuel properties as follows. δ, in terms of the dynamic viscosity μ, the density ρ, and the laminar burning velocity S_l, is given by (Kalghatgi and Bradley, 2012)

$$\delta = (\mu/\rho S_l) \tag{7}$$

Further, S_l is related to the pressure P and temperature T by

$$S_l = S_{l0}(T/T_0)^n (P/P_0)^{-m} \tag{8}$$

where S_{l0} is the laminar burning velocity at some reference pressure P_0 and temperature T_0, and m and n are constants (Bradley *et al*., 1998; Gulder, 1982). Using Equation 8 and with ρ_0 as the density and μ_0 as the dynamic viscosity at pressure P_0 and temperature T_0, δ can be written as (Kalghatgi and Bradley, 2012)

$$\delta = \left(\frac{\mu_0}{\rho_0 S_{l0}}\right)\left(\frac{P}{P_0}\right)^{m-1}\left(\frac{T}{T_0}\right)^{1.5-n} \tag{9}$$

The kinematic viscosity μ_0/ρ_0 of the fuel/air mixture for different liquid fuels can be assumed to be the same. The values of m have been reported to be between 0.16 and 0.28, whereas n is between 1 and 2.2 for different fuels (Kalghatgi and Bradley, 2012). Hence, for a given fuel (i.e., with S_{l0} fixed), the pressure exponent in Equation 9 is always negative so that, as the pressure increases, the laminar flame thickness decreases and preignition becomes more likely. Experimentally, for a given fuel, preignition is also found to be more likely if the mixture strength is slightly rich of stoichiometric, where S_{l0} is likely to be maximum (Ricardo and Hempson, 1972). We should also expect that, for a given fuel, use of cooled exhaust gas recirculation (EGR) will reduce S_{l0} and hence make preignition less likely, all else being equal. Ricardo and coworkers developed a test to quantify the preignition resistance (PR) of different fuels (Ricardo and Hempson, 1972). For a given operating condition, as we test different fuels, we should expect that, as S_{l0} increases, preignition becomes more likely and PR decreases. This is indeed

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

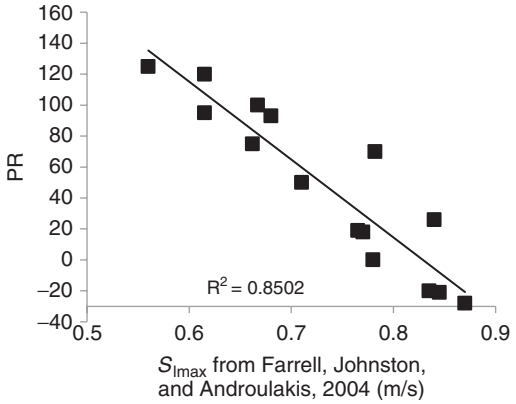


Figure 5. Preignition resistance (PR) from Ricardo and Hempson (1972) versus S_{lmax} from Farrell, Johnston, and Androulakis (2004) for different fuels, see Kalghatgi and Bradley (2012).

found to be the case (Kalghatgi and Bradley, 2012). Figure 5 shows the PR from Ricardo and Hempson (1972) for different fuels plotted against the maximum laminar burning velocity S_{lmax} measured at 30.4 bar (P_0) and 450 K (T_0) in Farrell, Johnston, and Androulakis (2004). PR in Ricardo and Hempson (1972) was measured at a mixture strength slightly rich of stoichiometric where the laminar burning velocity is likely to be maximum. There was little correlation between PR and RON or PR and MON, confirming what has been found in all previous studies (Kalghatgi and Bradley, 2012).

The mixture in the end-gas is never homogeneous and the autoignition that triggers knock is initiated at one or more centers or "hot spots." The pressure wave initiated by the autoignition changes as it traverses through inhomogeneities in the end-gas (Bradley, 2012). Similar phenomena also occur in HCCI engines (Sheppard, Tolegano, and Woolley, 2002). Hot spots are associated with gradients of reactivity, which lead to gradients of autoignition delay time. The pressure wave generated by the hot spot can couple with the autoignition reaction front when it moves into the unburned mixture at approximately the speed of sound. Under some circumstances, the fronts are mutually reinforced to create a large pressure spike propagating at high velocity within the hot spot in a *developing detonation* (Zeldovich, 1980). The probability of developing detonation increases if autoignition occurs at these high pressures and temperatures, and this is the most likely explanation for "super-knock" (Kalghatgi and Bradley, 2012; Kalghatgi *et al*., 2009).

4.1.2.2 Fuel autoignition quality effects in CI engines. In most modern CI engines, most of the fuel is injected directly into the combustion chamber when the piston is near the top of the compression stroke. Within a given time after the start of injection (SOI, i.e., the time at which fuel begins to enter the combustion chamber), fuel-rich regions of the stratified mixture in the cylinder autoignite, creating a rapid pressure increase. The rest of the fuel burns in a quasi-steady lifted jet diffusion flame (Dec, 1997). The time between SOI and autoignition is called the *ignition delay*. In general, a given change in the ignition delay time will have a larger effect in a high speed engine than a low speed engine because the crankshaft rotates through a larger angle per unit time in the high speed engine.

The autoignition quality of fuels for CI engines is typically quantified using the CN. In general, fuels with short ignition delays have higher CNs, and fuels with longer ignition delays have lower CNs. Diesel fuels in developed countries typically have CNs between 40 and 60, with values near the lower and higher ends of this range found in North America and Europe/Asia, respectively. High CN fuels are preferable for use in conventional CI engines. Their tendency to autoignite easily facilitates engine cold-starting. High CN fuels also lead to quieter engine operation, because the shorter time for fuel–air premixing during the ignition delay for high CN fuels leads to a smaller amount of chemical energy release during premixed autoignition (all other factors being equal).

The CN of a fuel is measured using an indirect-injection single-cylinder CFR engine with a pintle-type fuel injector and an adjustable compression ratio operated at a constant speed of 900 rpm, as described in ASTM International Test Method D613 (ASTM, 2008). The conditions used in the cetane engine test do not well represent those in modern CI engines. Other test methods employing constant-volume combustion chambers to measure the derived cetane number (DCN) also are becoming more widely used (ASTM, 2010f, 2012c), but their results are based on correlations with cetane engine data, so they are no more representative of conditions in modern engines. CN and DCN values for a number of pure HC and oxygenated compounds can be found in Murphy, Taylor, and McCormick (2004).

While the ignition delay is a function of the conditions within the engine cylinder (e.g., temperature, pressure, local mixture stoichiometry), it is also a function of the fuel-composition-dependent chemistry that occurs as the fuel begins to oxidize. The autoignition quality of a low CN base fuel can be improved without changing the bulk fuel composition by adding tens to hundreds of ppm of a highly reactive "cetane improver" to the fuel. The most common cetane improvers in use today are 2-ethylhexyl nitrate (EHN) and di-*tert*-butyl peroxide (DTBP). The effects of EHN addition on diesel-fuel CN have been discussed by Ghosh (2008).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

A treatment of the detailed processes leading from possible low temperature (or "cool-flame") reactions through high temperature ignition is beyond the scope of this chapter, but this area has been well addressed in the literature (Hwang, Dec, and Sjoberg, 2008; Mehl *et al*., 2012; Westbrook, 2000; Yang *et al*., 2011). Suffice it to say that kinetic effects on autoignition processes can be complex. Alkanes, aromatics, and oxygenates can and do exhibit markedly different behaviors under certain conditions. As mentioned in the preceding section, the need for new autoignition quality metrics to better characterize the ignition behaviors of modern fuels in modern engines has been acknowledged, but such new metrics have not yet found their way into widespread use.

4.1.3 Emissions formation characteristics

It is often easy to observe that changing the composition of the fuel supplied to an engine can change the emissions characteristics of the engine. Less straightforward, however, is understanding the mechanisms underlying the observed emissions changes. This is because fuel composition changes can directly alter the physical and chemical properties of a fuel, leading to complex and coupled variations in mixture formation, ignition, and combustion processes. Fuel composition changes also can indirectly affect emissions through engine-calibration-specific processes (an example of which is provided later in this section). The engine calibration establishes the exact values of the various input parameters required for optimal operation of the engine at each point in its operating space. Of these, engine-calibration-specific processes can be particularly difficult to understand because calibrations are usually proprietary in nature, costly to create, and only created for an "average" market fuel.

Examples of physical processes that can be affected by fuel composition changes include the following:

- fuel injection (e.g., injection pressure, timing, in-nozzle cavitation);
- fuel vaporization (e.g., liquid- and vapor-phase fuel penetration, entrainment rate, droplet characteristics);
- mixture formation (e.g., degree of fuel and thermal stratification, spray–wall interactions, presence of liquid fuel films).

Examples of chemical processes that can be affected by fuel composition changes include the following:

- fuel autoignition resistance (and dependence on in-cylinder thermal and mixture stratification and pressure);
- combustion kinetics (e.g., heat release rate, emissions formation and removal, combustion stability);
- mixture stoichiometry [i.e., the proximity of the charge gas mixture at each point in space to its stoichiometric condition (Mueller, 2005)];
- after-treatment system reactions and efficiency.

Examples of fuel-dependent engine-calibration-specific processes include the following:

- injection/ignition timing changes;
- turbocharger settings changes;
- EGR rate changes;
- intake manifold pressure and temperature changes;
- differences in any of the above between steady-state and transient conditions for different fuels.

On the basis of the above-mentioned processes, it should be expected that the magnitudes and even the directions of emissions changes will vary with fuel changes depending on which fuel properties are varied (it is almost always impossible to vary only one fuel property at a time); how fuel properties are varied (e.g., if the aromatic content of a fuel is decreased, what kinds of compounds are replacing the aromatics?); the ranges over which fuel properties are varied (the effects of raising CN from 30 to 40 are likely to be larger than the effects of raising it from 70 to 80); the specific engine and combustion strategy (fuel effects observed in a light-duty SI premixed combustion mode are likely to be different from those observed in a heavy-duty CI mixing-controlled combustion mode); the specific operating condition(s) and/or test cycle(s) employed (e.g., high vs low load conditions, steady-state vs transient cycle); and the presence, absence, nature, and/or age of any installed after-treatment devices.

The use of biodiesel in CI engines provides an illustrative example of how changing the fuel can have a number of interrelated, subtle, and complex effects on engine-out emissions. First, biodiesel has a lower volatility than conventional diesel fuel, which can lead to longer penetration of liquid-phase fuel within the combustion chamber and subsequent wall impingement (Fisher, Knothe, and Mueller, 2010). Wall impingement can lead to lower efficiency, reduced durability (due to loss of wear resistance when lubricant is diluted with fuel), and the formation of pool fires that can produce elevated emissions of smoke, NO_x, HC, and CO (Cheng *et al*., 2010; Martin *et al*., 2008). Second, the oxygen bound within biodiesel methyl esters pushes the fuel-rich autoigniting mixtures closer to stoichiometric conditions, leading to higher local temperatures, which are further elevated by less radiative heat loss from the flame due to the lower levels of soot produced

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

(Mueller, Boehman, and Martin, 2009). The higher temperatures and leaner mixtures lead to faster combustion reactions, producing an effective combustion-phasing advance. The higher temperatures and longer residence times at high temperature for biodiesel can lead to increased thermal NO_x formation relative to an HC diesel fuel (Mueller, Boehman, and Martin, 2009). Third, the lower energy content of biodiesel means that a larger quantity must be injected to achieve a given load. An engine calibrated for diesel fuel could interpret the larger quantity of fuel injected as a signal to reduce the EGR level, further increasing the NO_x emissions when fueling with biodiesel (Eckerle *et al*., 2008). In principle, many of these factors can be overcome by recalibrating the engine for biodiesel use, resulting in lower smoke, HC, and CO emissions at equivalent NO_x and efficiency levels relative to HC diesel fuel (Bunce *et al*., 2010). There are costs and challenges involved, however, with recalibrating existing engines, as well as adjusting for different biodiesel blend levels (e.g., B5 vs B100) and differences in biodiesels made from different feedstocks (e.g., soy vs palm).

Trends in mixture stoichiometry and temperature effects on emissions from both CI and SI engines can be estimated using a $\phi - T$ plot, as shown in Figure 6 [from Kim *et al*. (2008)]. Soot production occurs in regions with $\phi > 2$, and the width of the temperature window over which it occurs increases with ϕ. NO_x production occurs in near-stoichiometric to lean mixtures with $T > 1900$ K. HC production occurs in mixtures with $T < 1000$ K and/or higher temperature mixtures with $\phi > 1.2$. CO production occurs mostly in rich, high temperature mixtures, though it can also occur in leaner mixtures where temperatures are insufficient to carry CO oxidation to completion. If the fuel is oxygenated, the oxygen equivalence ratio should be used in place of the conventional equivalence ratio to quantify the mixture stoichiometry (Mueller, 2005).

Emissions trends also can change with fuel composition. Fundamental studies have shown that sooting tendency generally goes in the order poly-aromatics > mono-aromatics > *cyclo*-alkanes > *iso*-alkanes > *n*-alkanes (McEnally and Pfefferle, 2009; Nakakita *et al*., 2005). This seems logical because aromatic rings are the fundamental building blocks of soot. Fuel effects on NO_x emissions tend to be small, provided comparisons are made at the same mixture stoichiometry (Cheng, Upatnieks, and Mueller, 2007), although there is some evidence that NO_x emissions increase with aromatics content (Khalek *et al*., 2002; Mitchell, 2000). Trends for HC and CO emissions tend to be similar. Lower emissions of these pollutants tend to be correlated with higher levels of the more reactive species [e.g., *n*-alkanes (Lilik and Boehman, 2011)]. Other review chapters that may be of interest regarding fuel effects on CI and SI engine emissions include Hochhauser (2008) and Hoekman *et al*. (2011). A good review of fuel effects on 1998 and earlier heavy-duty CI engines can be found in Lee, Pedley, and Hobbs (1998).

4.1.4 *Flash point and flammability limits*

The flash point is a parameter used to quantify the tendency of a fuel to form an explosive mixture, and it is often used to help assess potential safety issues related to fuel storage and handling. The flash point is the temperature at and above which the vapors over a layer of liquid fuel will sustain

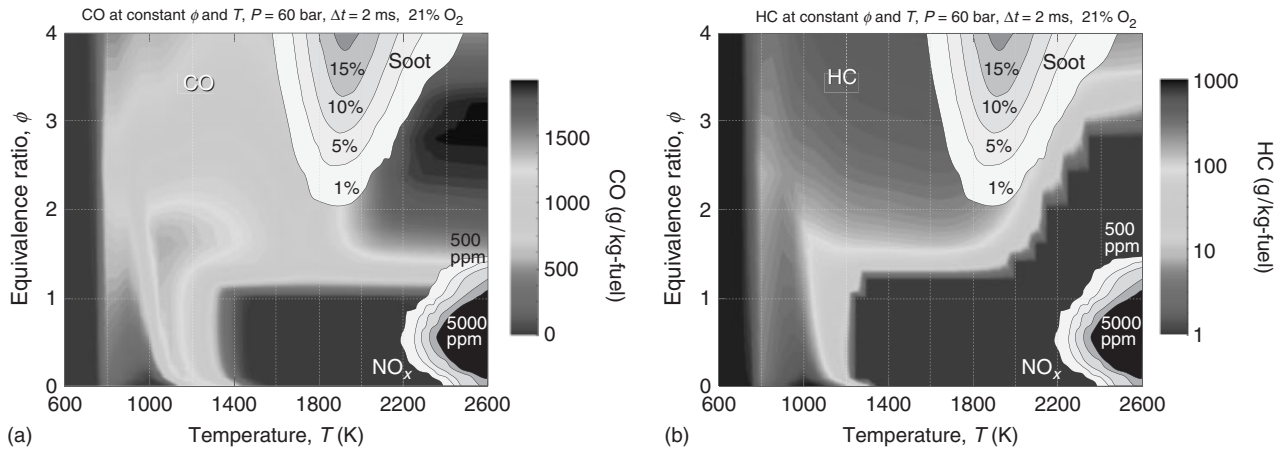


Figure 6. Emissions calculated in a constant temperature and pressure homogeneous reactor at a reaction time of 2.0 ms for *n*-heptane/air mixtures (i.e., no dilution) at the given local equivalence ratio ϕ and temperature T. Contours indicate emissions indices of soot and NO_x (Kitamura *et al*., 2002), whereas shades indicate emissions indices of (a) CO and (b) HC. (Reproduced with permission from the authors. © D. Kim, I. Ekoto, W.F. Colban and P.C. Miles.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a flame if an ignition source is provided. Many methods exist to measure the flash point, and the values yielded by different methods for a given fuel will not necessarily be equal. Nevertheless, the flash point is the most widely used fire safety parameter in commercial practice.

Another way to assess the risk of fire or explosion resulting from an accidental fuel release is to measure the flammability limits of the fuel. Most gasoline and diesel-range HCs will sustain combustion only in a narrow range from ~1 to 7 vol% in air at standard conditions. Fuel storage systems are designed so that the mixture in the fuel tank remains above the rich flammability limit for gasoline-fueled vehicles and below the lean flammability limit for diesel-fueled vehicles over a wide range of ambient conditions. The lean flammability limits for methane and hydrogen are slightly higher than those of gasoline- and diesel-range HCs at 5 and 4 vol%, respectively, but the rich limits are significantly higher at 15 and 74 vol%, respectively. This phenomenon and the high tank pressures often associated with methane- and hydrogen-fueled vehicles require that special safety precautions be taken to protect against fire and explosion in the event of an accidental release when these gaseous fuels are employed.

4.2 Physical properties

4.2.1 Phase-change characteristics

A fuel must satisfy a few primary phase-change requirements to perform well in a given reciprocating engine application. First, the fuel must generally be a liquid when it is pumped to the injectors. Solid and vapor-phase fuels are not compatible with conventional liquid fuel pumps, and will generally lead to filter plugging and "vapor lock" (loss of fuel pressure due to vaporization in the fuel supply system), respectively. Second, any fuel must be vaporized before it can be burned in the combustion chamber. Liquids and solids cannot burn in the conventional sense; some of the condensed phase must be converted to vapor and mixed with a sufficient amount of oxygen before combustion can occur. Third, the fuel should not be so volatile that a significant fraction of it is lost from the tank and/or fuel transfer lines over time (these evaporative emissions can adversely affect fuel consumption and air quality), but it should be volatile enough that the engine starts easily under cold conditions. Most fuel specifications related to phase change characteristics exist to ensure that these three requirements are met.

4.2.1.1 Maintaining liquid-phase fuel in the fuel supply/injection system. Ensuring that the fuel remains in the liquid phase within the fuel system imposes different challenges depending on fuel type, composition, climate at a given geographical location and time of year, altitude, and engine application. For gasolines, the challenges typically occur because of excessive liquid-to-vapor transition. Gasoline volatilities have been limited to prevent vapor lock in hot weather (ASTM, 2011g). The vaporization characteristics of a fuel are typically quantified by measuring its distillation curve. The distillation curve is typically quantified using the ASTM D86 test method (ASTM, 2011a), though alternative methods with an improved grounding in thermodynamics have been proposed (Bruno *et al*., 2010). For diesel fuels, the challenges associated with phase change typically occur because of undesired liquid-to-solid transition. Fuel freezing or gelling can cause fuel filters to plug, thereby stopping the engine. A number of test methods have been designed to quantify the potential cold-weather issues with diesel fuels, including the cloud point (ASTM, 2011e), pour point (ASTM, 2011b), cold filter plugging point (ASTM, 2010e), and the low temperature flow test (ASTM, 2010c).

4.2.1.2 Achieving desired fuel vaporization in the engine. The lower volatilities of diesel fuels also can present problems with in-cylinder vaporization, which can lead to lower efficiencies and higher emissions, especially when advanced or retarded injection timings cause the fuel to be injected into a cooler in-cylinder environment (Cheng *et al*., 2010; Fisher, Knothe, and Mueller, 2010; Martin *et al*., 2008). The main parameter affecting adequate gasoline vaporization is the back-end volatility (i.e., the high temperature end of the distillation curve). Oxygenate addition also can introduce challenges; for example, blending ethanol into gasoline results in increased latent and specific heats and hence more energy required to vaporize the fuel.

4.2.1.3 Preventing excessive evaporative emissions while maintaining cold-start performance. Evaporative emissions are primarily an issue with gasolines, and they are governed by the vapor pressure of the gasoline under the given operating/storage conditions. Vapor pressure can be quantified using a number of different test methods; see ASTM (2011g) for a list. Oxygenate addition can also introduce challenges with vapor pressure. For example, blending ethanol into gasoline results in an increased vapor pressure of the blend. The "boil-off" of cryogenic liquid fuels like LNG and liquid hydrogen also can be a source of evaporative emissions. Excessive vaporization is not typically a problem with diesel fuels because of their lower volatilities; the notable exception is if gasoline has been blended into diesel fuel.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.2.2 Density

Density is a fundamental physical property of a fuel, and it can influence fuel utilization through a number of mechanisms. As discussed in Section 4.1.1.1, the density of a fuel can affect FE and fuel consumption. Fuel density changes also can affect the penetration of liquid- and vapor-phase fuel within the combustion chamber, as well as the engine calibration. Aromatic HCs and oxygenates generally have higher densities than aliphatic HCs of the same carbon number.

4.2.3 Lubricity

The lubricity of a fuel is its ability to prevent the wearing of adjacent wetted parts that move relative to each other. Fuel lubricity is an important parameter, particularly in CI engine applications where the fuel pump and injectors are lubricated by the fuel, include close-tolerance parts, and may operate at pressures exceeding 250 MPa. The lubricity of a fuel is the combination of its hydrodynamic- and boundary-lubrication characteristics. Hydrodynamic lubrication is provided by the fuel viscosity, preventing adjacent parts from touching (higher viscosities are better), while boundary lubrication is provided by the fuel adhering (via intermolecular forces) to the adjacent solid surfaces (Bacha *et al*., 2007). Fuel lubricity can be quantified using a number of test methods, with the high frequency reciprocating rig (HFRR) (ASTM, 2011h) and scuffing load ball on cylinder lubricity evaluator (SLBOCLE) (ASTM, 2010d) being the most common in the United States. Lubricity is a property that can be significantly improved by fuel additives blended at treat rates significantly less than 1 wt%. Lubrication and Friction discusses these issues in more detail.

4.2.4 Viscosity

The viscosity of a fuel affects its hydrodynamic lubrication characteristics, as discussed above. Viscosity also influences fuel pumping and atomization. A fuel with low viscosity will tend to leak through small clearances in a fuel pump, which may lead to reduced delivery pressures and quantities. Lower viscosity (and lower surface tension) has been shown to enhance liquid fuel atomization, but this mechanism may be less important for the injection pressures and ambient conditions in modern CI engines (Dahms *et al*., 2013; Siebers, 1999).

4.2.5 Electrical conductivity

A static charge can accumulate when a fuel is being pumped from one location to another. In the absence of proper grounding, the electrostatic potential difference between a given fuel system component and ground can become large enough for a spark discharge to occur. This spark could act as an ignition source for fuel vapors, and a fire or explosion could ensue. Conductivity-enhancing additives are used when necessary to mitigate this potential hazard.

4.3 Stability

The stability of a fuel refers to its resistance to changes in its chemical composition and/or mixture characteristics over time. A fuel with excellent stability is desirable. There are two main types of fuel instability: chemical instability and mixture instability.

4.3.1 Chemical instability and deposit formation

A fuel that undergoes oxidation, polymerization, or other reactions in the liquid phase that change its composition has chemical instability. The chemical reactions of oxidation or polymerization can produce compounds with generally higher molecular weight that are insoluble in the fuel, such as lacquers, gums, and solids. These species are called *deposits* if they adhere to fuel-system, in-cylinder, or other engine components. Such deposits can plug fuel filters, prevent proper injector operation, foul in-cylinder components like piston rings and valves, and increase engine emissions. For example, the lower oxidative stability of biodiesel compared to conventional HC diesel fuel can lead to fuel system deposits, which is why a number of engine manufacturers only warranty their systems for fuels with up to 5 vol% biodiesel content. In-cylinder deposits in SI engines are formed by various mechanisms and can be correlated with lower volatility and higher aromatic content of the fuel. The deposits can lead to increased knocking that the knock sensor on modern engines will detect, causing the spark timing to be retarded and resulting in degraded acceleration performance as well as higher fuel consumption. Antioxidant additives are formulated to prevent the reactions leading to chemical instability, and detergent additives are formulated to control and, at higher concentrations, remove engine deposits after they have formed.

4.3.2 Mixture instability and the importance of polarity

The second type of fuel instability is mixture instability. A fuel mixture is unstable if it separates into two or more distinct liquid phases over time when exposed to different temperatures or with the addition of a trace contaminant. All

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

common fuels for reciprocating internal combustion engines are mixtures. These mixtures contain compounds composed of covalently bonded atoms. Some of these compounds have large permanent electrostatic dipole moments (i.e., they have a net positive charge in one area of the molecule and a net negative charge in another), some do not, and still others will take on a dipole moment in the presence of an imposed electric field. The magnitude of the permanent or induced electrostatic dipole moment of a compound is called its *polarity*. The polarities of different molecules have a strong impact on the types of mixtures they form and the stabilities of those mixtures.

Compounds with permanent electrostatic dipole moments are called *polar*. A high degree of polarity can be caused by the presence of one or more constituent atoms having different electronegativities. For example, the high electronegativity of oxygen tends to make oxygenated compounds polar. The attractive electrostatic forces between the positive and negative areas of polar molecules tend to hold polar molecules together, making their mixtures true solutions. For example, water and ethanol—both polar compounds—are miscible, that is, they are soluble in all proportions.

Compounds that exhibit small permanent electrostatic dipole moments are called *nonpolar*. HCs are generally nonpolar because the electronegativities of carbon and hydrogen are similar. HCs readily form stable solutions with other HCs. When nonpolar and polar molecules are combined, however, the attractive forces between polar molecules tend to keep them together rather than allowing them to mix with the nonpolar compounds: hence the adage "oil and water don't mix." The two types of compounds can be forced to mix through mechanical means (e.g., by putting them in a blender), producing an emulsion. An emulsion is essentially a suspension of droplets of high polarity in a distributed phase of low polarity, or vice versa. In general, buoyancy-driven stratification due to the different densities of the polar and nonpolar fractions, coupled with a finite rate of coalescence of droplets of similar polarity, will cause the polar (or "hydrophilic") compounds in an emulsion eventually to separate from the nonpolar (or "hydrophobic") compounds, unless other compounds are added to stabilize the mixture.

Finally, some HCs, while nonpolar, are "polarizable." The nonlocalized electron distributions of such compounds can allow them to take on an induced dipole moment because of the presence of polar molecules in the vicinity. Aromatic HCs are a classic example of polarizable compounds. Hence, ethanol is effectively insoluble in *n*-hexadecane, a largely nonpolar and unpolarizable compound, but it is quite soluble in gasolines that contain large concentrations of polarizable aromatic HCs.

Solutions generally will not separate over time, that is, they exhibit mixture stability. Nevertheless, solubility typically depends on factors such as temperature, solute concentration, and the presence of trace species. For example, the addition of a small amount of water to an oxygenate–HC mixture near its solubility limit also can cause the mixture to separate into two or more distinct liquid phases [see Tables 3 and 4 of Natarajan *et al*. (2001)].

4.4 Materials compatibility

Fuels come into contact with many materials as they move from their points of production through the engines they power, including materials in supertankers, pipelines, underground storage tanks, tanker trucks, dispensing pumps and lines, vehicle fuel tanks, fuel level sensors, fuel injection systems, engine components (valves, pistons, cylinder liners), lubricants, and after-treatment devices. The materials in these systems have been carefully selected and optimized over decades of experience. New fuel blendstocks can lead to incompatibilities with existing infrastructure, and these issues can be expensive to rectify.

4.4.1 Corrosivity

One of the most important material-compatibility properties of a fuel is its corrosivity. A fuel is said to be corrosive to a given metal if it chemically reacts with and degrades the surface quality of the metal over time at conditions experienced under normal system operation. It is important to use fuels with acceptable corrosivity characteristics to avoid, for example, the expense of repairing a corroded pipeline or the cost of replacing damaged fuel injectors in tens of thousands of vehicles. Fuels that contain organic or inorganic acids or highly polar compounds such as water or methanol often are corrosive. Fuel corrosivity is commonly quantified using a test method in which a polished strip of copper is submerged in the fuel at an elevated temperature for several hours and then visually compared to reference strips (ASTM, 2010a). Copper is used because it corrodes more quickly than ferrous metals, but in a similar manner. Sometimes, a silver strip is used in place of copper because many fuel level sensors contain silver (ASTM, 2011g).

4.4.2 Elastomer compatibility

Another important factor to consider with new fuels is their compatibility with common fuel system elastomers (i.e., the O-rings, gaskets, and hoses used in fuel injectors, pumps, and lines). Elastomers can swell, shrink, harden, soften,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

split, and even dissolve when exposed to different fuels. For example, nitrile-rubber elastomers used in conventional HC fuel delivery systems swell when wetted by fuels that contain aromatics, and this swell is factored into the system designs. If these systems are used with low aromatics fuels such as F–T diesel or hydrotreated vegetable oil, the seals will not swell adequately, leading to potential fuel system leakage and corresponding safety hazards. Oxygenates (e.g., ethanol and biodiesel) can present elastomer compatibility issues as well. While elastomers that are compatible with a wider range of fuels do exist (e.g., perfluoroelastomers), they tend to be more expensive than conventional rubber elastomers, and it is costly to install them into existing vehicles.

4.4.3 Catalyst deactivation

The third primary material-compatibility consideration with fuels has to do with the deactivation of the catalytic after-treatment systems used to lower the emissions levels of regulated pollutants from most modern powertrains. Catalyst deactivation is a rich area of study, and depends on the type of catalyst and the deactivation mechanism: chemical, fouling, thermal, or mechanical. Fuel properties most directly affect the first two of these mechanisms. Fuels that contain sulfur or phosphorus can "poison" a catalyst (i.e., lower its activity) by bonding or adsorbing to its active sites. Indeed, a key factor in the continual lowering of diesel-fuel sulfur levels in the United States (to 15 ppm by weight currently for on-road engines) is to enable the use of sulfur-sensitive catalytic after-treatment systems [for reference, the current sulfur limit for gasolines sold in the United States is 80 ppm by weight (ASTM, 2011g)]. Another fuel-dependent mechanism for catalyst deactivation is the physical blocking of active sites by fouling. The fouling agent could be soot, which can be burned from the catalyst during a regeneration event, or it could be alkali-metal ash or even silica particles from trace species in the fuel (ASTM, 2011g, Section 3.6.1). Alkali-metal levels in the fuel less than 10 ppm by weight have been found to lead to diesel particulate filter degradation (Williams *et al*., 2011). Gas Aftertreatment Systems and Solid/Condensed Phase Aftertreatment Systems also discuss catalyst poisoning issues.

Other potential fuel material-compatibility issues exist, including lube oil incompatibility, the fuel acting as a solvent for vehicle paint, and incompatibilities with plastic or fiberglass fuel tanks and liners. There are many important material-compatibility details to be considered when developing a new fuel blendstock.

4.5 Environmental considerations

Fuel usage in transportation and nonroad applications provides many benefits in our daily lives: from producing and distributing the food we eat, to building the communities and businesses in which we live and work, constructing roads and bridges, manufacturing and hauling goods for commerce, and enabling personal mobility for work and pleasure. As a result, fuel usage has increased to levels that can be challenging to physically comprehend (Section 2). Some of the environmental issues associated with this tremendous scale of fuel use are discussed in the following sections.

4.5.1 Toxicity of fuel and combustion products

The first major environmental consideration is the toxicity of the fuel and its combustion products. Sometimes, unburned fuel is inadvertently released into the environment; hence, it is important to know the toxicities of various fuel components, their solubility in water, and rates and strategies for biodegradation. Over the years, toxins in petroleum-derived fuels have been reduced. For example, lead (a neurotoxin) was phased out of gasoline between 1970 and 1996, and more recently the concentration of benzene (a known human carcinogen) has been limited in gasoline (Gibbs *et al*., 2009). Toxicity concerns are not restricted to petroleum-derived fuels; some alternative fuels also have toxicity issues. The ingestion of as little as 15 mL of methanol can cause blindness, with lethal dosages in the range 60–240 mL (ChemADVISOR, 2012). The jatropha plant and its seeds, which are considered a potential nonfood feedstock for biofuel production, also are highly toxic. The toxicities of many conventional and alternative fuel compounds have yet to be quantified.

Fuel combustion products are released whenever fuel is burned. The primary combustion products, carbon dioxide and water, are of little or no toxicity concern, but some minor products (e.g., NO_x) are highly toxic. The levels of many toxic combustion products have been on a decreasing trend in the United States since 1990 even though the total number of vehicle miles traveled, population, and energy consumption have been increasing (U.S. EPA (2012b), Figures 3 and 4).

4.5.2 Climate change

A second environmental consideration is global climate change due to human activities including the combustion of fossil fuels. Combustion produces CO_2, a GHG that inhibits radiative heat loss from the Earth by absorbing infrared radiation emitted from the Earth, becoming vibrationally

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

excited, and then reradiating a fraction of this energy back to the Earth when the molecule relaxes to a lower energy level. This "greenhouse effect" could play a role in the observed increase in global average temperatures, loss of polar ice, and rising sea levels in recent decades (Solomon *et al*., 2007). CO_2 is not the only GHG. For example, methane (an alternative fuel) has a GHG potential much greater than that of CO_2 (Section 2.3). The link between atmospheric GHG levels and global climate change is complex and not well understood.

A large number of academic, industrial, and governmental organizations have developed models to estimate GHG emissions. Models that combine all aspects from fuel production through end-use are termed WTW models. Values are generally reported in mass of CO_2 equivalent per unit volume of fuel or per unit distance traveled. Beyond WTW models, life-cycle models additionally include the emissions associated with processes to recycle vehicle components at the end of their useful lives. The specific GHG values for any fuel/engine combination are highly dependent on the modeling assumptions and the GHG emissions estimates for the individual processes that make up a given pathway. As a result, there can be large differences in estimates obtained by different organizations. Generally, compared to petroleum-derived gasoline and diesel fuels, bio-derived fuels have a smaller GHG footprint, while coal- and oil-sands-derived fuels have larger footprints, but there can be significant differences even among the biofuels. For example, in some parts of the world, forests have been cut down to create plantations for biofuel crops such as palm oil. The GHG impacts of such land-use changes can be significant. When land-use changes are included, the GHG emissions for corn-based ethanol rise to levels much closer to those for petroleum-based fuels than ethanol from sugar cane or cellulose.

GHG emissions for electric vehicles are highly dependent on the energy source used for electricity production. Approximately 50% of the electricity generated in the United States comes from coal. Electricity from coal has a larger GHG footprint than electricity from natural gas, and much larger than electricity from hydroelectric, geothermal, or nuclear energy. Of course, all of these energy sources have significant environmental aspects in addition to GHG emissions as well. When the GHG emissions of battery and electronic component production and end-of-life recycling are taken into account in life-cycle analyses, electric vehicles are not as environmentally friendly as they might seem based on the aspect that they produce no GHG emissions at the vehicle.

4.5.3 Sustainability

The key sustainability issue for petroleum-derived fuels is that they are not being created at a rate comparable to the rate at which they are being consumed. Although the discovery of large, new oil fields has diminished, proven oil reserves have continued to grow, even in the United States, because of the development and deployment of new petroleum-recovery technologies. One example is the recent development of technology to extract liquids and natural gas from tight shale formations that previously were inaccessible. For bio-derived feedstocks, sustainability issues include soil erosion and degradation, fresh-water consumption, and energy return on investment (i.e., the energy content of a given amount of fuel divided by the energy required to produce the given amount of fuel). For example, estimates of the energy return on investment for corn-based ethanol are near unity, see Schobert (2013, Chapter 4).

4.5.4 Food versus fuel

The final main environmental consideration is the diversion of resources from food production to fuel production. Examples of this include fuel-ethanol produced from corn, and biodiesel production from, for example, soy, canola, and palm oils. These "food versus fuel" concerns have shifted the focus of biofuel feedstock development from conventional food crops to inedible biomass. This has led to a longer term focus on the use of nonfood crops such as camelina or jatropha as sources of triglycerides for biodiesel production and inedible cellulose for ethanol production.

5 FUEL SPECIFICATIONS

The specific properties that make a fuel well suited to a given engine application have been determined over time through testing and the experience of many end-users. These properties are measured by test methods, and the set of test methods and corresponding fuel property ranges that have been found to be acceptable for a given application are established as fuel specifications. Fuel specifications are developed to increase the likelihood that all fuels marketed for a given application will meet or exceed all relevant requirements. Ideally, a fuel specification would take into account all of the potential fuel-property issues discussed previously in this chapter.

A fuel that meets or exceeds all relevant requirements is deemed "fit for purpose." Fuel specifications are developed on regional, national, and/or international bases by organizations composed of representatives from vehicle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

makers, engine and parts manufacturers, fuel producers, and government agencies. In the United States, fuel properties must conform to ASTM specifications [D4814 for gasoline (ASTM, 2011g) and D975 for diesel fuel (ASTM, 2011c)]. In Europe, most countries require fuel to conform to the Comité Européen de Normalisation (CEN) specifications EN 228 for gasoline (British Standards Institution, 2008b) and EN 590 for diesel fuel (British Standards Institution, 2009). These specifications provide lower and/or upper limits on key fuel properties such as cetane, distillation characteristics, density, aromatic content, viscosity, and sulfur content to ensure that fuels are fit for the engines and the environments in which they will be used. The specification limits can be different for different regions/countries and can vary depending on the season (e.g., winter vs summer). For example, the minimum acceptable value for CN is 40 per ASTM specifications, but 51 per CEN specifications. Differences in specifications typically reflect differences in regional engine characteristics, area climate, and environmental regulations. In the United States, the EPA and local agencies such as California Air Resources Board (CARB) have additional specifications for lower emissions fuels sold in certain areas, and pipeline companies can set supplementary specifications for fuels transported in their pipelines. For example, some pipeline companies have limits for diesel density and pour point that are not covered by the ASTM D975 specification.

As some specifications have a range of acceptable values while others do not have both a lower limit and an upper limit, fuels sold in the marketplace have a range of properties and compositions that reflect differences in refinery conversion units and capabilities, crude-oil slates, and product slates (e.g., whether jet/kerosene and/or home-heating oil are competing for similar fuel components). Fuel composition and properties are also expected to change as larger quantities of alternative-fuel components derived from biofuels, oil sands, and F–T processes (from natural gas, coal, or biomass) are blended with conventional petroleum-derived fuels and enter the market. A considerable amount of testing has been conducted and more still needs to be conducted on these new blend components to relate observed engine performance with measured fuel properties and composition. The results of this testing can be used to improve the relevant fuel specification(s).

The use of detailed fuel specifications and careful engine design has largely mitigated noticeable effects of fuel property changes on engine performance in typical applications. Nevertheless, the growing use of bio-derived and other unconventional fuels with composition and property characteristics substantially different from those of conventional petroleum-derived HC fuels, coupled with the development of advanced high efficiency clean-combustion engines, highlights the need for the continual evolution and improvement of fuel specifications.

5.1 Gasoline

ASTM D4814 covers specifications for 10 gasoline properties, although these do not include ONs (RON and MON) or antiknock index [AKI = (RON + MON)/2], which are set by gasoline refiners and marketers to meet the requirements set by the vehicle original equipment manufacturers (OEMs) and must be clearly labeled on the fuel pumps. The properties covered are the following:

1. volatility on a geographic and seasonal basis (to enable good drivability);
2. solvent washed gum content (to limit the amount of gums formed by oxidation);
3. oxidation stability (to provide additional protection against gum formation);
4. water tolerance (to protect against phase separation due to dissolved water at colder temperatures—usually not a problem for blends of HCs, but can be for blends containing oxygenates);
5. sulfur content (to limit deactivation of exhaust after-treatment catalysts and to protect against engine wear, deterioration of engine oil, and corrosion of exhaust system components);
6. copper strip corrosion (to protect against corrosion of fuel system components due to reactive compounds);
7. silver strip corrosion (to protect against corrosion of silver components present in some fuel gauge systems);
8. lead content (to ensure that unleaded gasolines are free of lead);
9. appearance (to ensure product is visually free of undissolved dirt, sediment, and suspended matter);
10. workmanship (to be free of adulterants or colorants that may make fuel unfit for purpose).

ASTM D4814 references more than 30 other ASTM standards that describe the specific test methods and equipment required to measure the gasoline properties. Specifying a property has little value unless all parties measuring the property use the same procedure and can obtain the same answer within the defined precision of the method.

ASTM (2011f) provides specifications for ethanol that is to be blended into gasoline. Previously, the US EPA designated certain regions as environmental nonattainment areas and required the addition of specific amounts of oxygenates in the gasoline. Recently, the US Congress passed the Renewable Fuels Standard 2 (RFS2), which has delegated the authority to state and local agencies

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to set specific oxygenate blend level requirements, but does require a minimum total volume of ethanol use in gasoline, which increases with time. The EPA continually sets specifications on the maximum fraction of ethanol that can be contained in gasoline. Previously, the maximum amount was 10 vol% (except for the 85 vol% E-85 ethanol blend that is approved for flexible-fuel vehicles), but in 2012 the EPA granted a waiver to allow up to 15 vol% ethanol in gasoline (E-15) for vehicle model years 2001 and newer. OEMs have expressed concerns about the 15 vol% level because some tests conducted by CRC suggest that the fuel systems in some 2001 and newer vehicles not specifically designed for E-15 may be susceptible to failure when fueled with E-15 (Coordinating Research Council, 2012) and also that older vehicles may accidentally be fueled with E-15.

For 2012, the RFS2 total ethanol requirements for the United States are 13.2 billion gallons per year of corn-based ethanol and 500 million gallons of cellulosic ethanol. However, cellulosic ethanol is still in the research and development stages, and the amount produced commercially was essentially zero (Section 2.4.1). In future years, the amount of cellulosic ethanol required will continue to increase to a value of 10.5 billion gallons in 2020, but the amount of corn-based ethanol is capped at 15 billion gallons per year.

5.2 Diesel fuel

The US diesel-fuel specification, ASTM D975 (2011c), covers 11 diesel fuel properties. These include the following:

1. minimum flash point (relates to safety in fuel handling and use);
2. water and sediment (affects fuel filters and injectors);
3. back-end volatility (related to ease of starting and smoke formation);
4. viscosity (affects fuel spray atomization, fuel system lubrication, and fuel system leakage);
5. ash content (can damage fuel injection and after-treatment systems and cause combustion chamber deposits);
6. maximum sulfur content (affects after-treatment catalysts, particulate emissions, and cylinder wear);
7. cetane and maximum aromatic content (relates to autoignition quality and affects cold-start capability, combustion timing and performance, and emissions);
8. cold-temperature properties of cloud point, low temperature filterability, and cold-filter plugging point (affects low temperature operability and fuel handling);
9. carbon residue (measures coking tendency of fuel and may relate to deposit formation);
10. copper strip corrosion rating (indicates potential for corrosive attack on metal parts);
11. lubricity [relates to abrasive wear of components such as fuel pump and injector; has become poorer because of removal of certain compounds during hydrotreating to meet ultra-low sulfur diesel (ULSD) standards].

Two diesel fuel properties that can vary quite a bit in market fuels that meet specifications are CN and aromatic content. In the United States, CNs range from 40 to the mid-50s. The aromatic contents of US #2 ULSD fuels range from <10 vol% to >30 vol%. Although there is an ASTM specification maximum of 35 vol% for aromatics, it can be exceeded if the cetane index value (ASTM, 2011d) is 40 or higher. The aromatic values of 10 vol% and lower are found in some California fuels where the CARB #2 ULSD specification is 10 vol% maximum. However, alternative fuel formulations having aromatic contents higher than 10 vol% are allowed (and actually sold) in California, provided that the emissions of the alternative blends predicted by CARB-specified engine test methods and models are no higher than the official 10 vol% blend (California Environmental Protection Agency, 2004). In addition to wide variation in aromatic content, there also can be a wide variation in the other HC classes [*n*-alkanes, *iso*-alkanes, and mono- and poly-cycloalkanes (Farrell *et al*., 2007)].

The US RFS2 mandates the use of 1 billion gal per year of "biomass-based diesel." Currently, this predominantly is biodiesel composed of FAMEs mostly prepared from soybean vegetable oil. Several state and local governments have set mandates on the minimum fraction of biodiesel that must be blended into each gallon of diesel fuel. Issues related to the limited stabilities, low back-end volatilities, and less-than-ideal cold-temperature properties of biodiesel blends have resulted in engine manufacturers setting upper limits on biodiesel content of 5 vol% for some older engines and up to 20 vol% for newer engines. The use of FAME biodiesel has resulted in the development of ASTM (2010g, 2011i) and CEN (British Standards Institution, 2008a) specifications for the biodiesel component.

6 SUMMARY AND OUTLOOK

The production and distribution of liquid fuels from petroleum has been optimized over more than a century and a half, and the development of reciprocating internal combustion engines to use these fuels has continued for nearly as long. Although fuel production and engine combustion technologies are well established, they continue

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to advance rapidly and play central roles in economy and society. This trend appears likely to continue for at least several more decades (NPC, 2012; Exxon Mobil, 2012). Substantial further improvements in engine efficiency and emissions are possible, especially on an equivalent-cost basis relative to competing technologies.

Significant concerns about environmental quality and energy security are currently driving the coevolution of fuels and engines. Cost-effective, regionally produced, non-food-based, nontoxic, renewable fuels created from biological systems or solar energy and used in high efficiency clean-combustion engines may be able to help address these concerns. These systems are being aggressively researched and developed. The optimal fuel/engine system may be different in different areas of the world, depending on local climate, population density, resources, and myriad other factors. A great deal of research remains to be done to develop optimal fuel/engine systems. The largest hurdle to be overcome in the area of fuels and internal combustion engines is to meet both the tremendous scale of fuel demand and the need for high efficiency, clean combustion engines in a manner that is cost effective and at the same time enhances environmental quality and energy security.

REFERENCES

Amer, A., Babiker, H., Chang, J., *et al*. (2012) Fuel effects on knock in a highly boosted direct injection spark ignition engine. SAE Technical Paper 2012-01-1634.

ASTM Standard D613 (2008) *Standard test method for cetane number of diesel fuel oil*. ASTM International, West Conshohocken, PA.

ASTM Standard D130-10 (2010a) *Standard test method for corrosiveness to copper from petroleum products by copper strip test*. ASTM International, West Conshohocken, PA.

ASTM Standard D2885-10a (2010b) *Standard test method for determination of octane number of spark-ignition engine fuels by on-line direct comparison technique*. ASTM International, West Conshohocken, PA.

ASTM Standard D4539-10 (2010c) *Standard test method for filterability of diesel fuels by low-temperature flow test (LTFT)*. ASTM International, West Conshohocken, PA.

ASTM Standard D6078-04 (2010d) *Standard test method for evaluating lubricity of diesel fuels by the scuffing load ball-on-cylinder lubricity evaluator (SLBOCLE)*. ASTM International, West Conshohocken, PA.

ASTM Standard D6371-05 (2010e) *Standard test method for cold filter plugging point of diesel and heating fuels*. ASTM International, West Conshohocken, PA.

ASTM Standard D6890-10a (2010f) *Standard test method for determination of ignition delay and derived cetane number (DCN) of diesel fuel oils by combustion in a constant volume chamber*. ASTM International, West Conshohocken, PA.

ASTM Standard D7467 (2010g). *Standard specification for diesel fuel oil, biodiesel blend (B6 to B20)*, ASTM International, West Conshohocken, PA.

ASTM Standard D86-11a (2011a). *Standard test method for distillation of petroleum products at atmospheric pressure*. ASTM International, West Conshohocken, PA.

ASTM Standard D97-11 (2011b). *Standard test method for pour point of petroleum products*. ASTM International, West Conshohocken, PA.

ASTM Standard D975-11b (2011c). *Standard specification for diesel fuel oils*. ASTM International, West Conshohocken, PA.

ASTM Standard D976-06 (2011d). *Standard test method for calculated cetane index of distillate fuels*. ASTM International, West Conshohocken, PA.

ASTM Standard D2500-11 (2011e). *Standard test method for cloud point of petroleum products*. ASTM International, West Conshohocken, PA.

ASTM Standard D4806-11a (2011f). *Standard specification for denatured fuel ethanol for blending with gasolines for use as automotive spark-ignition engine fuel*. ASTM International, West Conshohocken, PA.

ASTM Standard D4814-11b (2011g). *Standard specification for automotive spark-ignition engine fuel*. ASTM International, West Conshohocken, PA.

ASTM Standard D6079-11 (2011h). *Standard test method for evaluating lubricity of diesel fuels by the high-frequency reciprocating rig (HFRR)*. ASTM International, West Conshohocken, PA.

ASTM Standard D6751-11b (2011i). *Standard specification for biodiesel fuel blend stock (B100) for middle distillate fuels*. ASTM International, West Conshohocken, PA.

ASTM Standard D2699-12 (2012a). *Standard test method for research octane number of spark-ignition engine fuel*. ASTM International, West Conshohocken, PA.

ASTM Standard D2700-12 (2012b). *Standard test method for motor octane number of spark-ignition engine fuel*. ASTM International, West Conshohocken, PA.

ASTM Standard D7170-12 (2012c). *Standard test method for determination of derived cetane number (DCN) of diesel fuel oils—fixed range injection period, constant volume combustion chamber method*, ASTM International, West Conshohocken, PA.

Bacha, J., Freel, J., Gibbs, A., *et al*. (2007) *Diesel Fuels Technical Review*, Chevron Products Co., San Ramon, CA. http://www.chevronwithtechron.com/products/documents/Diesel_Fuel_Tech_Review.pdf (accessed 4 April 2012)

Bell, A. (2010) Modern SI engine control parameter responses and altitude effects with fuels of varying octane sensitivity. SAE Technical Paper 2010-01-1454.

Bradley, D. (2012) Autoignitions and detonations in engines and ducts *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences*, **370** (1960), 689–714.

Bradley, D., Hicks, R.A., Lawes, M., *et al*. (1998) The measurement of laminar burning velocities and markstein numbers for

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

iso-octane-air and iso-octane-n-heptane-air mixtures at elevated temperatures and pressures in an explosion bomb *Combustion and Flame*, **115** (1–2), 126–144.

Bradley, D., Morley, C., and Walmsley, H.L. (2004) Relevance of research and motor octane numbers to the prediction of engine autoignition. SAE Technical Paper 2004-01-1970.

British Petroleum (2012) *BP Statistical Review of World Energy*. London, http://bp.com/statisticalreview (accessed 14 October 2012).

British Standards Institution Standard EN 14214 (2008a). *Automotive fuels—fatty acid methyl esters (FAME) for diesel engines—requirements and test methods—incorporates amendment a1:* 2009. European Committee for Standardization (CEN), Brussels, Belgium.

British Standards Institution Standard EN 228 (2008b) *Automotive fuels—unleaded petrol—requirements and test methods*. European Committee for Standardization (CEN), Brussels, Belgium.

British Standards Institution Standard EN 590 (2009) *Automotive fuels—diesel—requirements and methods of test*. European Committee for Standardization (CEN), Brussels, Belgium.

Bruno, T.J., Ott, L.S., Lovestead, T.M., Huber, M.L. (2010) Relating complex fluid composition and thermophysical properties with the advanced distillation curve approach *Chemical Engineering & Technology*, **33** (3), 363–376.

Bunce, M., Snyder, D., Adi, G., *et al*. (2010) Stock and optimized performance and emissions with 5 and 20% soy biodiesel blends in a modern common rail turbo-diesel engine *Energy & Fuels*, **24**, 928–939.

California Environmental Protection Agency (2004) Standards for diesel fuel, in *California Code of Regulations*, Title 13, Division 3, Chapter 5, Article 2, Section 2282(g), http://www.arb.ca.gov/fuels/diesel/081404dslregs.pdf (accessed 30 October 2012).

ChemADVISOR, Inc (2012) Material safety data sheet: methyl alcohol. Report No. OHS14280, ChemADVISOR, Inc., Pittsburgh, PA.

Cheng, A.S., Fisher, B.T., Martin, G.C., and Mueller, C.J. (2010) Effects of fuel volatility on early direct-injection, low-temperature combustion in an optical diesel engine *Energy & Fuels*, **24** (3), 1538–1551.

Cheng, A.S., Upatnieks, A., and Mueller, C.J. (2007) Investigation of fuel effects on dilute, mixing-controlled combustion in an optical direct-injection diesel engine *Energy & Fuels*, **21** (6), 1989–2002.

Christensen, E.D., Chupka, G.M., Luecke, J., *et al*. (2011) Analysis of oxygenated compounds in hydrotreated biomass fast pyrolysis oil distillate fractions *Energy & Fuels*, **25** (11), 5462–5471.

Coordinating Research Council (2011) Fuel antiknock quality—engine response to RON versus MON—scoping tests. Coordinating Research Council Report No. 660, Coordinating Research Council, Alpharetta, GA, http://www.crcao.org/reports/recentstudies2011/CRC 660/CRC 660.pdf (accessed 16 September 2012).

Coordinating Research Council (2012) Intermediate-level ethanol blends engine durability study. Coordinating Research Council Report No. CM-136-09-1B, Coordinating Research Council, Alpharetta, GA, http://www.crcao.org/publications/performance/index.html (accessed 23 September 2012).

Dahms, R.N., Manin, J., Pickett, L.M., and Oefelein, J.C. (2013) Understanding high-pressure gas–liquid interface phenomena in diesel engines *Proceedings of the Combustion Institute*, **34** 1667–1675.

Dahnz, C., Han, K.M., Spicher, U., *et al*. (2010) Investigations on pre-ignition in highly supercharged SI engines *SAE International Journal of Engines*, **3** (1), 214–224. DOI: 10.4271/2010-01-0355

Dec, J.E. (1997) A conceptual model of DI diesel combustion based on laser-sheet imaging *SAE Transactions*, **106** (3), 1319–1348. DOI: 10.4271/970873

Eckerle, W.A., Lyford-Pike, E.J., Stanton, D.W., *et al*. (2008) Effects of methyl ester biodiesel blends on NOx emissions *SAE International Journal of Fuels and Lubricants*, **1** (1), 102–118. DOI: 10.4271/2008-01-0078

Edwards, R., Larivé, J.F., and Beziat, J.C. (2011) Well-to-Wheels Analysis of Future Automotive Fuels and Powertrains in the European Context. Publications Office of the European Union, Luxembourg, http://iet.jrc.ec.europa.eu/about-jec/sites/iet.jrc.ec.europa.eu.about-jec/files/documents/wtw3_wtw_report_eurformat.pdf (accessed 21 February 2013).

Exxon Mobil (2012) The Outlook for Energy: A View to 2040. Exxon Mobil, Irving, TX, http://www.exxonmobil.com/Corporate/Files/news_pub_eo2012.pdf (accessed 21 October 2012).

Farrell, J.T., Cernansky, N.P., Dryer, F.L., and Friend, D.G. (2007) Development of an experimental database and kinetic models for surrogate diesel fuels. SAE Technical Paper 2007-01-0201.

Farrell, J.T., Johnston, R.J. and Androulakis, I.P. (2004) Molecular structure effects on laminar burning velocities at elevated temperature and pressure. SAE Technical Paper 2004-01-2936.

Fikri, M., Herzler, J., Starke, R., *et al*. (2008) Autoignition of gasoline surrogates mixtures at intermediate temperatures and high pressures *Combustion and Flame*, **152** (1–2), 276–281.

Fisher, B.T., Knothe, G., and Mueller, C.J. (2010) Liquid-phase penetration under unsteady in-cylinder conditions: soy- and cuphea-derived biodiesel fuels versus conventional diesel *Energy & Fuels*, **24** (9), 5163–5180.

Gauthier, B.M., Davidson, D.F., and Hanson, R.K. (2004) Shock tube determination of ignition delay times in full-blend and surrogate fuel mixtures *Combustion and Flame*, **139** (4), 300–311.

Ghosh, P. (2008) Predicting the effect of cetane improvers on diesel fuels *Energy & Fuels*, **22** (2), 1073–1079.

Gibbs, L., Anderson, R., Barnes, K., and Engeler, G. (2009) Motor Gasolines Technical Review. Chevron Products Co, San Ramon, CA. http://www.chevronwithtechron.com/products/documents/69083_MotorGas_Tech_Review.pdf (accessed 31 May 2012).

Gulder, O.L. (1982) *Laminar burning velocities of methanol, ethanol and isooctane-air mixtures*. Nineteenth Symposium (International) on Combustion, pp. 275–281.

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*, McGraw-Hill, New York.

Hochhauser, A.M. (2008) Review of prior studies of fuel effects on vehicle emissions. Report No. E-84, Coordinating Research Council, Alpharetta, GA.

Hoekman, K., Broch, A., Robbins, C. and Ceniceros, E. (2011) Investigation of biodiesel chemistry, carbon footprint and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

regional fuel quality. Coordinating Research Council Report No. AVFL-17a, Coordinating Research Council, Alpharetta, GA, http://www.crcao.org/publications/advancedVehiclesFuelsLubricants/index.html (accessed 19 July 2013).

Hwang, W., Dec, J., and Sjoberg, M. (2008) Spectroscopic and chemical-kinetic analysis of the phases of HCCI autoignition and combustion for single- and two-stage ignition fuels *Combustion and Flame*, **154** (3), 387–409.

Kalghatgi, G.T. (2005) Auto-ignition quality of practical fuels and implications for fuel requirements of future SI and HCCI engines. SAE Technical Paper 2005-01-0239.

Kalghatgi, G.T. and Bradley, D. (2012) Pre-ignition and 'super-knock' in turbo-charged spark-ignition engines *International Journal of Engine Research*, **13** (4), 399–414.

Kalghatgi, G.T., Bradley, D., Andrae, J. and Harrison, A.J. (2009) The nature of "superknock" and its origins in SI engines. Institution of Mechanical Engineers Conference on Internal Combustion Engines: Performance, Fuel Economy and Emissions. London.

Kalghatgi, G.T., Nakata, K. and Mogi, K. (2005) Octane appetite studies in direct injection spark ignition (DISI) engines. SAE Technical Paper 2005-01-0244.

Khalek, I.A., Ullman, T.L., Vasquez, L. and Guerrero, M. (2002) Hot start transient emissions from a Mercedes OM 366 LA and a Detroit Diesel operated on Chilean, California, and US 2D fuels. SAE Technical Paper 2002-01-2827.

Kim, D., Ekoto, I., Colban, W.F., and Miles, P.C. (2008) In-cylinder CO and UHC imaging in a light-duty diesel engine during PPCI low-temperature combustion *SAE International Journal of Fuels and Lubricants*, **1** (1), 933–956. DOI: 10.4271/2008-01-1602

Kitamura, T., Ito, T., Senda, J., and Fujimoto, H. (2002) Mechanism of smokeless diesel combustion with oxygenated fuels based on the dependency of the equivalence ratio and temperature on soot particle formation *International Journal of Engine Research*, **3** (4), 223–247.

Lee, R., Pedley, J., and Hobbs, C. (1998) Fuel quality impact on heavy duty diesel emissions: a literature review *SAE Transactions*, **107** (4), 1952–1970. DOI: 10.4271/982649

Lilik, G.K. and Boehman, A.L. (2011) Advanced diesel combustion of a high cetane number fuel with low hydrocarbon and carbon monoxide emissions *Energy & Fuels*, **25** (4), 1444–1456.

Martin, G.C., Mueller, C.J., Milam, D.M., *et al*. (2008) Early direct-injection, low-temperature combustion of diesel fuel in an optical engine utilizing a 15-hole, dual-row, narrow-included-angle nozzle *SAE International Journal of Engines*, **1** (1), 1057–1082. DOI: 10.4271/2008-01-2400

McEnally, C.S. and Pfefferle, L.D. (2009) Sooting tendencies of nonvolatile aromatic hydrocarbons *Proceedings of the Combustion Institute*, **32**, 673–679.

Mehl, M., Pitz, W., Sarathy, M. *et al*. (2012) Detailed kinetic modeling of conventional gasoline at highly boosted conditions and the associated intermediate temperature heat release. SAE Technical Paper 2012-01-1109.

Metzger, P. and Largeau, C. (2005) *Botryococcus braunii*: a rich source for hydrocarbons and related ether lipids *Applied Microbiology and Biotechnology*, **66** (5), 486–496.

Mitchell, K. (2000) Effects of fuel properties and source on emissions from five different heavy duty diesel engines. SAE Technical Paper 2000-01-2890.

Mittal, V. and Heywood, J.B. (2008) The relevance of fuel RON and MON to knock onset in modern SI engines. SAE Technical Paper 2008-01-2414.

Mittal, V. and Heywood, J.B. (2009) The shift in relevance of fuel RON and MON to knock onset in modern SI engines over the last 70 years *SAE International Journal of Engines*, **2** (2), 1–10. DOI: 10.4271/2009-01-2622

Mooney, J.J. (2007) The 3-way catalytic converter: (a) Invention and introduction into commerce—impacts and results; (b) Barriers negotiated. Presentation to California Air Resources Board, http://www.arb.ca.gov/research/seminars/mooney/mooney.pdf (accessed 16 September 2012).

Mueller, C.J. (2005) The quantification of mixture stoichiometry when fuel molecules contain oxidizer elements or oxidizer molecules contain fuel elements *SAE Transactions*, **114** (4), 1243–1252. DOI: 10.4271/2005-01-3705

Mueller, C.J. (2013) The feasibility of using raw liquids from fast pyrolysis of woody biomass as fuels for compression-ignition engines: a literature review. *SAE International Journal of Fuels and Lubricants*, **6** (1), 251–262, DOI: 10.4271/2013-01-1691.

Mueller, C.J., Boehman, A.L., and Martin, G.C. (2009) An experimental investigation of the origin of increased NOx emissions when fueling a heavy-duty compression-ignition engine with soy biodiesel *SAE International Journal of Fuels and Lubricants*, **2** (1), 789–816. DOI: 10.4271/2009-01-1792

Murphy, M.J., Taylor, J.D. and McCormick, R.L. (2004) Compendium of experimental cetane number data. National Renewable Energy Laboratory Report No. SR-540-36805, National Renewable Energy Laboratory, Golden, CO.

Nakakita, K., Akihama, K., Weissman, W., and Farrell, J.T. (2005) Effect of the hydrocarbon molecular structure in diesel fuel on the in-cylinder soot formation and exhaust emissions *International Journal of Engine Research*, **6**, 187–205.

Natarajan, M., Frame, E., Naegeli, D.W., *et al*. (2001) Oxygenates for advanced petroleum-based diesel fuels: Part 1. Screening and selection methodology for the oxygenates *SAE Transactions*, **110** (4). DOI: 10.4271/2001-01-3631

NPC (2012) *Advancing Technology for America's Transportation Future*. National Petroleum Council, Washington, D.C., http://www.npc.org/FTF-80112.html (accessed 21 October 2012).

Ricardo, H.R. and Hempson, J.G.G. (1972) *The High-Speed Internal Combustion Engine*, 5th edn, Blackie and Son, Ltd., London.

Schobert, H.H. (2013) in *Chemistry of Fossil Fuels and Biofuels* (ed A. Varma), Cambridge University Press, Cambridge, UK.

Sheppard, C.G.W., Tolegano, S. and Woolley, R. (2002) On the nature of autoignition leading to knock in HCCI engines. SAE Technical Paper 2002-01-2831.

Siebers, D.L. (1999) Scaling liquid-phase fuel penetration in diesel sprays based on mixing-limited vaporization *SAE Transactions*, **108** (3), 703–728. DOI: 10.4271/1999-01-0528

Solomon, S., Qin, D., Manning, M., *et al*. (eds) (2007) *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, UK and New York, NY.

Solomons, T.W.G. (1996) *Organic Chemistry*, 6th edn, John Wiley & Sons, Inc., New York.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Upatnieks, A. and Mueller, C.J. (2005) Clean, controlled DI diesel combustion using dilute, cool charge gas and a short-ignition-delay, oxygenated fuel. SAE Technical Paper 2005-01-0363.

U.S. Energy Information Administration (2012) *Annual Energy Review 2011*. U.S. Energy Information Administration, Washington, D.C., http://www.eia.gov/totalenergy/data/annual (accessed 14 October 2012).

U.S. Environmental Protection Agency (2012a) *Concerns About Methyl Tertiary Butyl Ether (MTBE) in Drinking Water*, http://www.epa.gov/mtbe/water.htm (accessed 16 September 2012).

U.S. Environmental Protection Agency (2012b) Our nation's air: status and trends through 2010. U.S. Environmental Protection Agency Report No. EPA-454/R-12-001, U.S. Environmental Protection Agency, Research Triangle Park, NC, http://www.epa.gov/airtrends/2011/index.html (accessed 14 October 2012).

Westbrook, C.K. (2000) Chemical kinetics of hydrocarbon ignition in practical combustion systems *Proceedings of the Combustion Institute*, **28**, 1563–1577.

Williams, A., McCormick, R., Luecke, J., *et al*. (2011) Impact of biodiesel impurities on the performance and durability of DOC, DPF and SCR technologies *SAE International Journal of Fuels and Lubricants*, **4** (1), 110–124. DOI: 10.4271/2011-01-1136

Yang, Y., Dec, J., Dronniou, N., *et al*. (2011) Partial fuel stratification to control HCCI heat release rates: fuel composition and other factors affecting pre-ignition reactions of two-stage ignition fuels *SAE International Journal of Engines*, **4** (1), 1903–1920. DOI: 10.4271/2011-01-1359

Yui, S. (2008) Producing quality synthetic crude oil from Canadian oil sands bitumen *Journal of the Japan Petroleum Institute*, **51** (1), 1–13.

Zahdeh, A., Rothenberger, P., Nguyen, W., *et al*. (2011) Fundamental approach to investigate pre-ignition in boosted SI engines *SAE International Journal of Engines*, **4** (1), 246–273. DOI: 10.4271/2011-01-0340

Zeldovich, Y.B. (1980) Regime classification of an exothermic reaction with nonuniform initial conditions *Combustion and Flame*, **39** (2), 211–214.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto125
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Intake Boosting

Nick Baines
Concepts NREC, White River Junction, VT, USA

1 Introduction 1
2 Turbocharger Characteristics 4
3 Turbocharger Matching 8
4 Other Boosting Systems 12
References 15

1 INTRODUCTION

1.1 Motivations for intake boosting

The purpose of intake boosting is to increase the density of the air trapped in the cylinder, which enables the fuel flow and energy release from combustion to be increased. The result is an engine of higher power output, or more relevantly to automotive designers, the same power output can be achieved with a smaller engine. This is the direct result of charge boosting, but the indirect effects are often more profound. Boosting allows a smaller engine to be employed that operates at higher loads. The engine is more efficient because higher loads have proportionally less loss because of internal cylinder friction and heat loss due to the engine coolant. A smaller engine will also package better in a vehicle, with improvements to vehicle aerodynamics and a reduction in structural weight. Undoubtedly the greatest effect on engine technology, however, has been the impact on emissions.

All of the developments in engine combustion in recent years that are aimed at reducing emissions involve dilution of the charge. These include exhaust gas recirculation (EGR) that is widely used for this purpose, and other combustion technologies such as stratified charge, homogeneous charge compression ignition (HCCI), premixed charge compression ignition (PCCI), and adjustments to valve timing such as the Miller cycle. Oxides of nitrogen (NO_x) are formed primarily in the high temperature of the combustion flame, and the quantity of NO_x created can be controlled by diluting the charge to limit the flame temperature, reducing the compression ratio, or using excess fuel (although this conflicts with the need for fuel economy). Carbon monoxide, unburnt hydrocarbons, and particulates are the products of partial combustion, which can be controlled using a lean mixture. All of these measures, however, also reduce the specific power output of the engine. In the circumstances, charge boosting has become an essential technology to maintain the engine power output while retaining the advantages of low emissions and fuel economy. The following sections discuss performance characteristics of boosting technologies. Other details can be found in Turbocharging.

1.2 Boosting fundamentals

An internal combustion engine is a positive displacement device, and the mass flow rate of air into the engine is determined by the trapped mass in the cylinder, which is the product of the air density ρ_a, the swept volume of the cylinder V_{sw}, and the volumetric efficiency η_{vol} (Heywood, 1988). Together with the crankshaft speed N, the mass flow rate is

$$m_a = \frac{\eta_{vol}\rho_a V_{sw} N}{2} \quad (1)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

The volumetric efficiency accounts for losses in the induction process. The division by 2 is required for a four-stroke engine because there is only one firing stroke for every two rotations of the crankshaft.

The power that can be delivered is determined by the amount of fuel that is burned m_f, the energy available in the fuel per unit mass Q_f, and a fuel conversion efficiency η_f that allows for the incomplete nature of the combustion process

$$P = \eta_f Q_f m_f \tag{2}$$

The power delivered by a four-stroke engine is therefore

$$P = m_a \eta_f Q_f \left(\frac{m_f}{m_a}\right) = \frac{\eta_{vol} \rho_a V_{sw} N \eta_f Q_f}{2} \left(\frac{1}{AFR}\right) \tag{3}$$

Assuming that the engine combustion system has been well optimized, the only one of these parameters that the designer can influence to a large extent is the inlet air density. This is done by compressing the air before it enters the cylinder.

The simplest method of doing this is by means of a mechanical supercharger, such as a piston, sliding vane, or rotary compressor, driven by the crankshaft of the engine, as shown in Figure 1a. A supercharger enables very considerable increases in output power to be achieved, but at a cost to overall efficiency because the supercharger itself consumes shaft power from the engine. For relatively low boost pressures of about 1.5–2 bar, this may be 10–15% of the engine power, but it increases rapidly, so that at a boost pressure of 4 bar, the fraction is about 40–50%. This cannot be justified economically. The alternative solution is to drive the compressor using the energy of the engine exhaust gas, which is extracted by expanding that gas through a turbine. The turbine is linked to the compressor by means of a shaft that is independent of the engine output shaft. This is the turbocharger concept shown in Figure 1b. In a typical engine, roughly 30–40% of the energy released by combustion of the fuel appears as energy in the exhaust, which is thermodynamically waste energy. If this energy is extracted from the exhaust gases in a turbine, the compressor can be driven without need to draw power from the engine shaft.

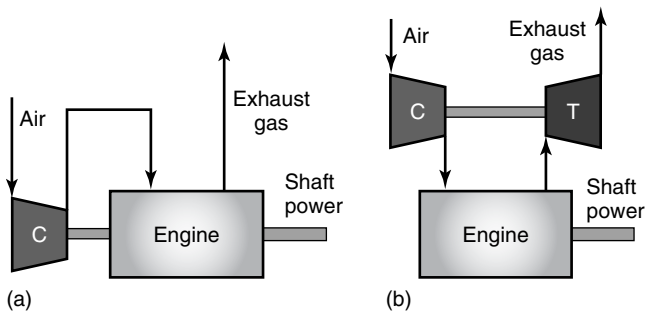


Figure 1. (a) Supercharged engine (b) turbocharged engine.

1.3 Turbocharger

The turbocharger has three principal components: a compressor, a turbine to drive the compressor and linked to it by a shaft, and bearing assemblies to support the shaft. In addition, shaft seals are required to separate the air and exhaust gases from the bearing lubricant, high temperature turbochargers may require cooling passages to limit the casing and bearing temperatures, and bleed valves or other active control devices may be included. Even with the basic turbocharger, considerable differences in detail are encountered.

Figure 2 shows a turbocharger designed for vehicle engines. The compressor is a single-stage centrifugal compressor, with the air entering axially and discharging radially, before being collected in the volute housing and delivered to the engine. The turbine is a single-stage radial turbine. A volute housing is again used to distribute the exhaust gas about the circumference of the turbine. The gas enters the rotor radially and discharges axially. Careful design and precision engineering of all components is necessary to ensure a high performance and long life. Vehicle turbochargers are made in very large numbers and must be designed from the outset for manufacture by mass production methods. The market is competitive, and a commercially successful design is one that achieves the right compromise between performance, service life, and cost.

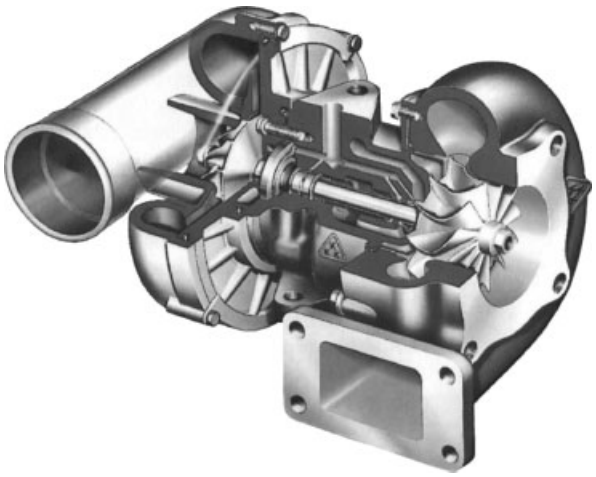


Figure 2. Vehicle engine turbocharger. (Reproduced by permission of BorgWarner Turbo Systems.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The pressure ratios developed by turbocharger compressors range from low values for small gasoline engines where the boost pressure is limited to about 2 bar, up to about 3.5–4 bar for vehicle diesel engines that have to operate over a wide range of speeds and flow rates, or higher for power generation engines where the width of the operating range can be sacrificed for higher boost pressure at design point operating conditions. For high pressures, more than one stage of compression may be employed. Turbocharger speeds can range up to about 300,000 rev/min for very small turbochargers, and as low as 10,000 rev/min for large machines.

To be commercially viable, a turbocharger manufacturer must have products that match a range of engines of varying capacity. The flow range of any given compressor, and to a lesser extent the turbine, is limited, and so manufacturers will normally provide a range of models, or "frame sizes," similar in concept and often developed from the same basic designs, with different flow capacities. Additional variations can be achieved by trimming the compressor and turbine rotors. This involves removing material from the tips of the blade to reduce the flow areas of the blade passages. The rotor is cast to the maximum trim size required and then for smaller trims material is removed by machining. This process invariably introduces some compromise into the blade shape and thickness, and for this reason trimming can only be taken so far. Eventually, a change to a smaller frame size becomes necessary.

1.4 Supercharger

The supercharger is an air compressor, which may be either a positive displacement or a rotordynamic type (pressure wave superchargers are discussed in Section 4.5). Common types of positive displacement compressor are the Roots blower, Lysholm screw compressor, and sliding vane compressor (Figure 3). Positive displacement compressors deliver a fixed volume of air per revolution, and if the supercharger is driven directly from the engine shaft, it is ideally matched to the engine that also has a constant volume flow characteristic. Positive displacement compressors typically have high efficiency at low pressure ratio, but the efficiency decreases rapidly as the pressure ratio increases. They are, therefore, only suited to low boost applications. The drive from the engine shaft may be direct or via a belt or gear drive. In modern engines, a clutch is often included to decouple the supercharger at high engine speed so as to remove the load on the engine and improve the fuel economy. At low engine speed, the supercharger is connected to provide boost and increase the low speed torque of the engine.

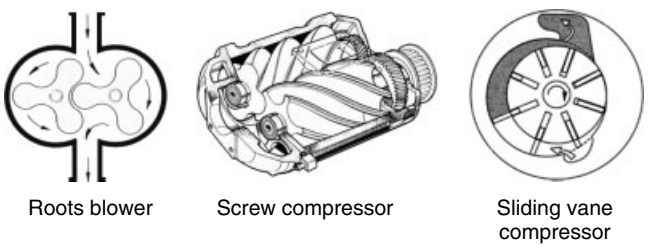


Figure 3. Positive displacement compressors for superchargers.

Rotordynamic superchargers are similar to turbocharger compressors and rotate at speeds much greater than that of the engine, so that a large gear ratio drive is necessary. Alternatively, the compressor may be driven by an electric motor. This requires a source of electricity on the vehicle, which is an engine-driven alternator unless regenerative braking can be utilized, as in a hybrid electric powertrain. Rotordynamic compressors have an airflow characteristic that is less well matched to the engine, making a clutch (for mechanical drive) or speed control (for electric drive) essential.

A two-stroke engine does not have an induction stroke, and a charge booster is essential for operation (Heywood, 1988). Small engines may use the crankcase as a supercharger, but larger engines require a mechanical or electrical supercharger, or a turbocharger, to induce air and scavenge the engine.

1.5 Charge air cooler

An unfortunate effect of the compression process is that the temperature of the air is increased as well as the pressure. This is the same for turbochargers and superchargers. For engine power boosting, the critical parameter is the mass of air trapped in the cylinder, and for a given cylinder volume, this depends on the air density, as is indicated in Equation 1. The increase in temperature reduces the air density. Furthermore, the inefficiency of the compressor increases the air temperature at the exit above what it would be if the compression were ideal.

Reducing the charge air temperature in a separate heat exchanger after compression has a direct effect on the trapped mass in the cylinder and on the mean effective pressure of the engine. Furthermore, it also results in lower temperatures throughout the operating cycle, leading to less thermal stress on the engine and the turbine. The charge air cooler, also known as an *intercooler* or *aftercooler*, is located between the compressor and the inlet manifold. The performance of a heat exchanger is measured in terms of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

its effectiveness

$$\varepsilon = \frac{\text{Actual heat transfer}}{\text{Maximum possible heat transfer}} = \frac{T_{2,\text{in}} - T_{2,\text{out}}}{T_{2,\text{in}} - T_{\text{c}}} \quad (4)$$

where $T_{2,\text{in}}$ is air inlet temperature from the compressor, $T_{2,\text{out}}$ is the air outlet temperature from the cooler, and T_{c} is the coolant temperature. An effectiveness of unity implies that the temperature of the air is reduced to that of the coolant, which is as low as it possibly can be; conversely, an effectiveness of zero means that no cooling occurs. Figure 4 shows the effect on density ratio of cooling effectiveness, assuming that the coolant is at ambient temperature. Even modest levels of cooling have considerable benefit, and this increases with boost pressure ratio.

For an automobile, the sources of coolant that are readily available are the engine cooling water and ambient air. The engine cooling water has the disadvantage that its temperature will be close to boiling for normal engine operation and this limits the scope for cooling. At low boost pressures, the air temperature at exit from the compressor may be lower than the coolant temperature. If operation is required for any length of time under these conditions, it is necessary to turn off the charge air cooler or allow the air to bypass it, otherwise, the charge air density will be reduced rather than increased. Using ambient air as a coolant increases the cooling potential, but air has a much lower convective heat transfer coefficient and thermal capacity than water. This means that to achieve the same amount of heat transfer, a larger surface area and heat exchanger volume is required.

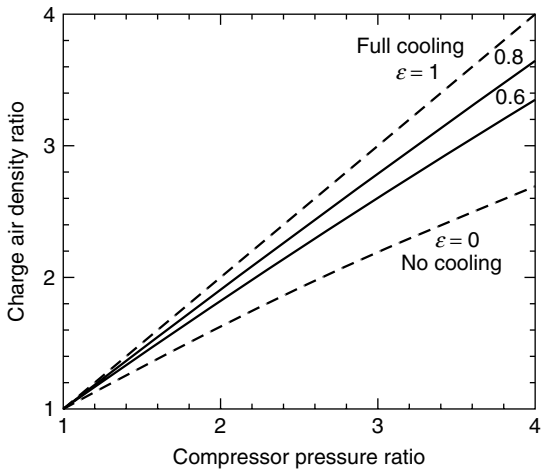


Figure 4. Effect of cooling effectiveness on charge air density ratio. Results are shown for a compressor isentropic efficiency of 0.8 and ambient and coolant temperatures of 288K.

In the circumstances, a separate water cooling circuit for the charge air cooler, with its own circulating pump and cooling radiator, is commonly used because it combines the benefits of the thermal properties of a water coolant with coolant temperatures closer to ambient. The effectiveness of the heat exchanger depends very largely on the area over which heat transfer occurs and thus the size of the heat exchanger. The design of the charge air cooler, therefore, must be a careful balance of cooling effectiveness, installation size, and packaging in the engine compartment.

2 TURBOCHARGER CHARACTERISTICS

2.1 Compressor

2.1.1 Components

A centrifugal compressor stage comprises rotating and stationary components. The air enters the impeller from the axial direction (Figure 5). The flow moves through the impeller and is pressurized by the action of the moving impeller blades. It is also turned toward the radial direction, so that it leaves with a combination of radial and tangential velocities. At this point, it still has substantial kinetic

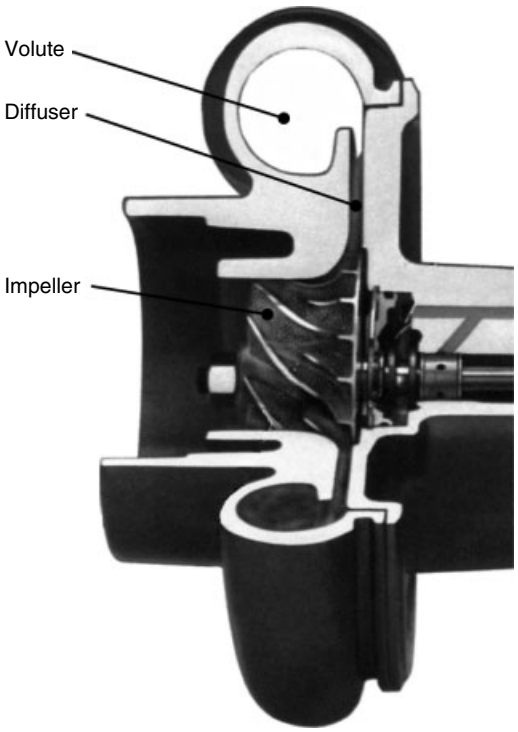


Figure 5. Components of a centrifugal compressor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

energy, and so for performance reasons as much of this as possible is recovered in a diffuser. The diffuser is a stationary component that can take several different forms, including the vaneless diffuser that can be seen in Figure 5, or vanes or channels can be used to guide the flow. Finally, the air is collected in a volute to direct it toward the inlet manifold.

2.1.2 Compressor performance map

To the turbocharger engineer, the compressor performance parameters of greatest interest are the pressure ratio, which determines the boost pressure and brake mean effective pressure (BMEP) of the engine, and the efficiency, which affects the turbocharger efficiency and the overall fuel economy of the system. In addition, engineers are usually concerned with range, which is the range of flow rates over which the compressor will operate stably and with acceptable performance.

A typical compressor operating map is shown in Figure 6. The pressure ratio is plotted as a function of the compressor mass flow rate and the speed of rotation, and contours of efficiency are superimposed on this graph, where efficiency is defined as

$$\eta = \frac{(T_{02}/T_{01}) - 1}{(p_2/p_{01})^{(k-1)/k} - 1} \tag{5}$$

where T is the temperature, p is the pressure, and k is the specific heat ratio. Subscript 0 refers to total conditions, 1 to the compressor inlet and 2 to the compressor exit.

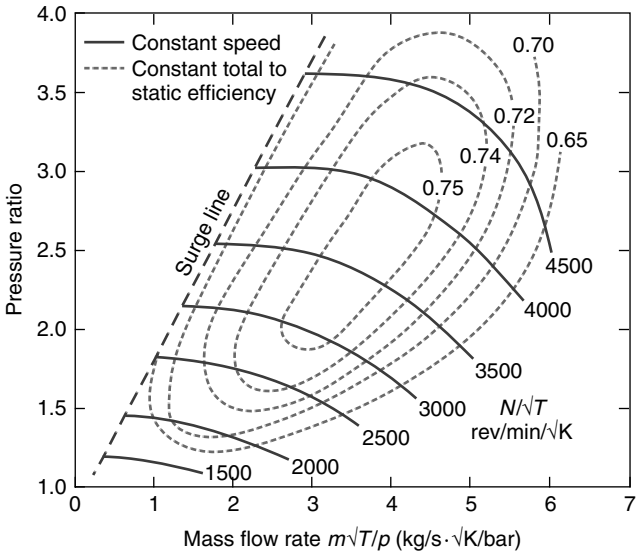


Figure 6. Compressor performance map.

The mass flow and speed parameters used here are corrected for the inlet or ambient pressure p and temperature T. This ensures that the same map can be used for engine matching even when the ambient conditions change, for example, when operating at altitude, and that a map measured on the gas stand at one set of inlet conditions will be applicable when the engine operates at different inlet conditions. It is evident that both pressure ratio and efficiency are strong functions of mass flow rate and speed. Maximum efficiency occurs over a relatively restricted range of operation, and for efficient operation of a turbocharged engine, it is important to ensure that the engine running line is in this region as far as possible.

The range of stable operation of a compressor is limited by choke at high flow rates and surge at low flow rates and is defined as

$$\text{Range} = \frac{m_{\text{choke}} - m_{\text{surge}}}{m_{\text{choke}}} \tag{6}$$

where m_{choke} and m_{surge} are the compressor mass flow rates at choke and surge. Choke occurs when the flow becomes sonic at the throat, or point of minimum area, of the gas path. For a compressor with a vaneless diffuser, choking invariably occurs at the throat of the impeller. For a compressor with a vaned diffuser such as a channel diffuser, the diffuser also has a throat, and depending on the match between the impeller and the diffuser, either component may choke first. Choking sets a maximum limit on the flow capacity of the compressor. As the flow approaches choking, so the aerodynamic losses accumulate, localized patches of sonic and supersonic flows occur, and the efficiency and pressure ratio developed both fall rapidly.

The onset of surge fixes the minimum flow rate of the compressor for any given speed of rotation. Surge is marked by a breakdown of the flow and the initiation of large-scale pressure oscillations that, if unchecked, can cause rapid deterioration and failure of the compressor blades or the turbocharger bearings. Great care must, therefore, be taken to avoid running the compressor into surge. Figure 6 shows a typical lower mass flow rate limit imposed on compressor operation by surge, labeled the surge line (the long-dashed line). The proximity of any stable operating point to surge is the surge margin

$$\text{Surge margin} = \frac{m - m_{\text{surge}}}{m_{\text{choke}} - m_{\text{surge}}} \tag{7}$$

The causes of surge are complex and are not all well understood, but it basically stems from a breakdown in the deceleration of the flow in the compressor. High pressure ratios require more deceleration and compressors are likely to operate closer to this point of breakdown

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

than at low pressure ratios. For modest boost pressures, the range of operation can be made relatively wide and such compressors do not impose severe operating restrictions on turbocharged engines. For high boost pressures, however, the range is narrow and it is this factor, more than any other design or performance limitations, which restricts the boost that can be developed by a single-stage compressor. Very high boost pressures require two stages of compression.

Surge cannot be predicted analytically except in very restricted cases. The problem is complicated by the fact that surge is a system phenomenon and depends not only on the compressor but also on such parameters as the charge air cooler and the inlet manifold volume and configuration. The same compressor may surge at different points on different engines because of the different intake arrangements. The deterioration in performance as choke is approached is also difficult to predict. It is essential to test new compressor designs on the gas stand and on the engine in order to make reliable estimates of the range. Gas stand testing should ideally be done with the actual engine intake hardware to be confident about the measured points of surge (Capon, Leong, and Morris, 2006).

2.1.3 Compressor design

The principal determinant of pressure ratio is the peripheral speed of the impeller, but this also controls the centrifugal stress in the impeller blades and hub. Aluminum alloys are satisfactory materials for pressure ratios up to about 3.5–4, and beyond that, titanium alloy is required. The operating range also decreases as pressure ratio increases, and for automotive applications demanding a wide range, this can be limiting. Two stages may be used to achieve high pressure ratio and range, but for single-stage compressors, other measures such as shroud bleed (sometimes described as ported shrouds) may be used to extend the stable operation to lower flow rates. Backsweep, which is the curvature of the impeller blades at exit away from the radial direction, is also influential in increasing the stable operating range. See Japikse (1996) for a full discussion of centrifugal compressor design.

2.2 Turbine

2.2.1 Components

For automotive turbochargers, the turbine is invariably a single-stage radial turbine or a mixed-flow turbine, which can be considered as a variant of the radial turbine. In the small sizes required for this application, axial turbines suffer unduly from tip leakage losses, and the blades are more prone to high cycle fatigue and foreign object damage. The radial turbine comprises two essential parts: a fixed, swirl generating component or stator in which the working fluid is expanded and turned to give it a circumferential velocity about the axis of the machine, and a rotor, through which the flow passes and, in doing so, does work. In the rotor, the fluid enters in the radial-inward direction, is turned in the meridional plane, and leaves in the axial direction.

The stator may take one of two basic forms. Figure 7a shows a radial turbine with a ring of stator, or nozzle, vanes. Upstream of these vanes is a volute housing that takes the flow from the exhaust manifold of the engine and distributes it around the periphery of the nozzle. The shape of the volute is such that the cross-sectional area normal to the flow direction decreases from a maximum at the inlet as the flow moves around the axis of the turbine. This has the effects of forcing the flow inward into the nozzle and imparting considerable swirl velocity to the gas before it enters the nozzle. The nozzle has only to turn the gas through a fairly small angle. Nozzle blades are often made with no camber at all and rely on the radial inward motion of the gas to accelerate and turn it. Nozzle blades are often arranged to pivot so as to change the spacing between the blades and hence the flow capacity of the turbine. This is a form of variable geometry turbine, which has significant advantages in engine matching.

A nozzleless turbine is shown in Figure 7b, in which the volute alone is responsible for accelerating and swirling the flow. The attraction of this is that by dispensing with the nozzle ring a cheaper assembly can be achieved. The disadvantages are that the volute alone is rarely as aerodynamically efficient as a nozzle and that the larger the expansion ratio, the larger the volute that is required to achieve the necessary acceleration.

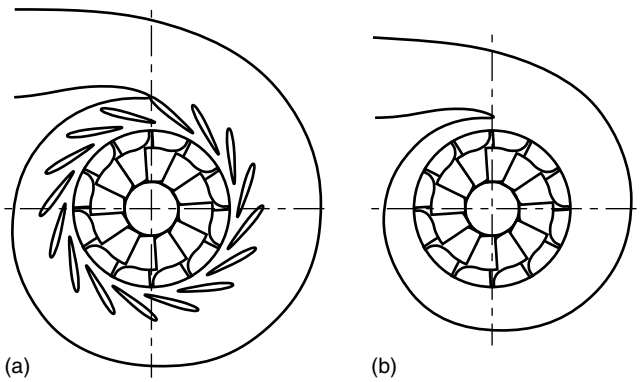


Figure 7. Radial turbines (a) with nozzle (b) nozzleless.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

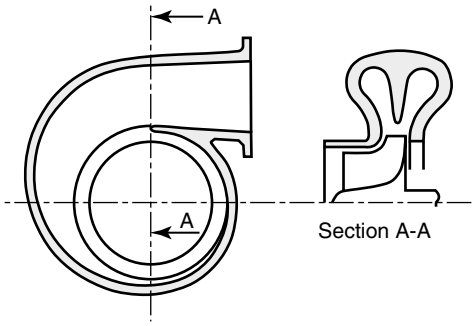


Figure 8. Twin-entry volute for a nozzleless turbine.

Volutes with multiple entries may be required for exhaust pulse energy utilization in multicylinder engines (Baines, 2005; Watson and Janota, 1982). Figure 8 is a section of such a turbine. The two entries are separated by a dividing wall that extends close to the inlet of the rotor. This helps to ensure that the exhaust pulse energy is transferred directly to the turbine rotor while avoiding exhaust interactions between cylinders that can inhibit scavenging.

2.2.2 *Turbine performance map*

The flow capacity of a turbine shows a rising characteristic with pressure ratio that flattens off and becomes constant as the turbine chokes (Figure 9). The flow rate is fairly insensitive to speed of rotation. These characteristics are very important for the matching of the compressor and turbine and the performance of a turbocharged engine. The turbine efficiency is also plotted in Figure 9 against pressure ratio. The efficiency plotted here is often described as the turbine mechanical efficiency and includes the loss of power

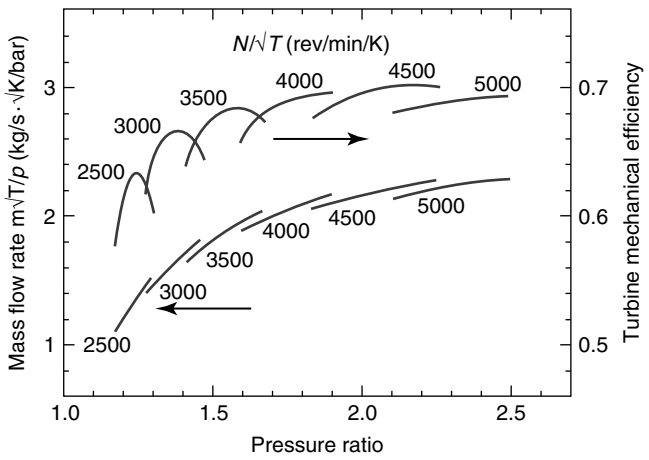


Figure 9. Radial turbine performance maps.

in the turbocharger bearings (as opposed to the aerodynamic efficiency, which excludes bearing loss). When testing a turbocharger, for practical reasons, the shaft power is not measured directly but by means of the temperature increase in the air passing through the compressor. Thus, the turbine mechanical efficiency is based on the power delivered to the compressor rather than the power developed by the turbine. An accurate measurement of the turbine gas temperature change is more difficult to make and is usually avoided. The range of operating conditions covered in testing is much smaller than the potential operating range of a turbine, but is limited by the need to match the compressor and the limits of stable compressor operation. When using the turbine maps in engine simulations, it is usually necessary to extrapolate the map characteristics considerably, and this is a source of possible uncertainty.

2.2.3 *Turbine design*

As with the compressor, the rotor blade speed has an important influence on the power developed by the turbine. However, the polar inertia of the rotating system is also important because it limits the rate of change of turbocharger speed during engine transients, and this has a major effect on "drivability" in automotive applications. Most of the rotating inertia resides in the turbine rotor, which must be made of a temperature-resistant nickel alloy that has a specific gravity approximately three times that of aluminum. The design of a turbine rotor, therefore, is a compromise between performance and inertia, with life and durability also having important roles. See Moustapha *et al*. (2003) for a full discussion of turbine design.

The trade-off between efficiency and inertia is illustrated in Figure 10. Here, the stage loading and flow coefficients are nondimensional expressions of power output and flow capacity, and the contours of efficiency are empirical observations based on a wide range of radial turbines tested. On the basis of this information, radial turbines designed for a flow coefficient of about 0.25 provide the best efficiency, but turbines for small turbochargers are typically designed for much higher values of flow coefficient, as represented by the gray area on the right enclosed in a dashed contour. As flow coefficient is proportional to mass flow rate per unit flow area, this implies smaller turbines for the same engine size. To emphasize this point, superimposed on the figure at the bottom are the approximate scales of turbine rotor size and inertia. Here, the datum case is a turbine designed for a flow coefficient of 0.25 where the achievable efficiency is a maximum. Alternatively, if a larger value of flow coefficient is chosen for the turbine (but still with the same flow rate, and therefore, the same engine application), these scales show approximately how the rotor

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

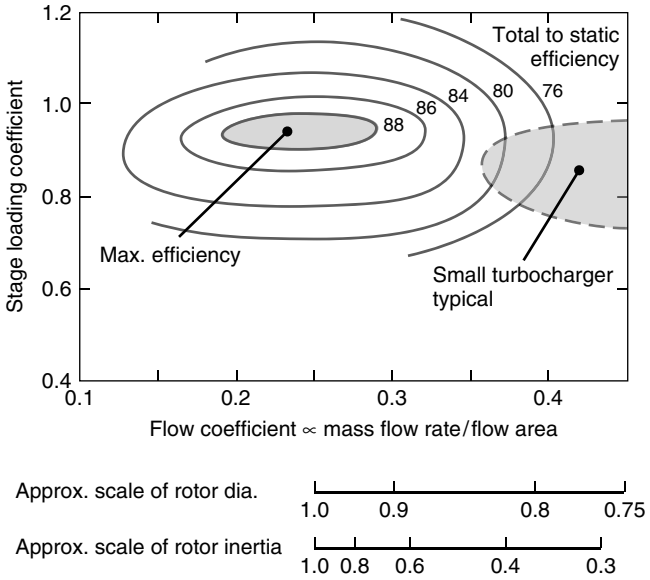


Figure 10. Correlation of stage loading and flow coefficients for conventional radial-inflow turbines, showing estimated scales of turbine size and inertia and the typical range of design for automotive turbocharger turbines. (Adapted from Chen and Baines (1994). © Elsevier.)

diameter and rotating inertia will change. For example, a choice of flow coefficient of 0.4 leads to a diameter that is approximately 80% and an inertia less than 40% of the datum turbine. These would normally be considered important attributes of an automotive turbocharger, but the graph shows that they come with a loss of more than 10 points of attainable efficiency. Large sacrifices in efficiency are often made in order to achieve low inertia. The demands of life and durability in the hot gas environment also impose limitations on the rotor blade design that limits the aerodynamic efficiency.

3 TURBOCHARGER MATCHING

The proper selection of turbocharger components to make up a complete engine–turbocharger system is a complicated balance of many design considerations. For many automotive applications, a single design point cannot be defined because operation is required for a wide range of different operating points. In a competitive market, high system performance will be required and good fuel economy will be necessary. Such criteria must be met with a turbocharger requiring as little space and adding as little weight as possible. To carry out an effective turbocharger matching calculation, it is necessary to have an accurate representation of each component of the turbocharger and the ability to determine how the entire system performs with all components operating together.

In fact, two separate tasks can be identified for the turbocharger engineer. The first of these is to determine the appropriate compressor and turbine that will match a given engine size and operating condition. This task is required whenever a new turbocharged engine project is considered, an existing engine is being rerated, or a change of turbocharger (perhaps for commercial, economic, or supply reasons) is being contemplated. The second task follows the first and is to analyze the performance of a defined engine–turbocharger system. Such analysis enables the engineer to predict the system performance at all of the various operating conditions required. The effect of changes to the system, such as different compressors and turbines, changes to the valve timing, and changes in manifold volume, can also be predicted. With the appropriate analysis tools, it is also possible to predict the transient response of the system.

3.1 Principles of matching

The engine, compressor, and turbine are linked by air and exhaust gas flow, transfers of energy, and a requirement that the air inlet pressure to the system and the exhaust gas exit pressure from the system are both ambient (Watson and Janota, 1982). The purpose of matching is to ensure that these compatibility conditions are satisfied. Mass flow matching can be simply stated as

$$m_e = m_a + m_f \tag{8}$$

or

$$m_e = m_a \left(1 + \frac{1}{\mathrm{AFR}}\right) \tag{9}$$

where m_e, m_a, and m_f are the exhaust gas, intake air, and fuel mass flow rates, respectively, in the engine and AFR is the air/fuel ratio (m_a/m_f). However, the operation of a turbocharged engine is complicated by the different flow characteristics of the various components. A piston engine operates like a positive displacement pump, delivering a certain amount of fluid per cycle, and its volume flow rate increases very nearly linearly with engine speed. For a naturally aspirated engine, the mass flow rate is identically affected. For a turbocharged engine, the mass flow rate increases with engine speed and also with the boost pressure. For the compressor, the mass flow rate of air is controlled by the inlet area and the axial air velocity into the impeller. The rotational speed of the compressor does not directly influence the mass flow rate. However, at high flow rates, the relative velocity within the impeller may

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

become sonic and choke the compressor. The turbocharger speed does have an influence on the maximum flow rate that the compressor can pass. At any compressor speed, there is also a minimum flow rate for stable operation.

The turbine acts like a simple throttle and is nearly independent of rotational speed. The curve of mass flow parameter $m\sqrt{T_0}/p_0$, where T_0 and p_0 are the turbine inlet temperature and pressure, respectively, against expansion ratio increases rapidly at low expansion ratios. Starting from a low engine speed (and low flow rate) condition and accelerating the engine, the turbine expansion ratio initially remains small, so that the exhaust manifold pressure and the energy available to the turbine are limited. At high engine speeds and flow rates, the turbine mass flow parameter reaches a limiting value as the turbine chokes. In this region, because $m\sqrt{T_0}/p_0$ is constant for the turbine, increases in mass flow rate require equivalent increases in the turbine inlet pressure and so the expansion ratio and the power developed by the turbine both increase rapidly.

The energy balance of the turbocharger demands that the power delivered to the compressor is equal to that developed by the turbine, minus power lost in bearing friction and windage. This can be written as

$$\left[\left(\frac{p_2}{p_{01}}\right)^{(k_a-1)/k_a} - 1\right] C_{pa} T_{01} = \left[1 - \left(\frac{p_4}{p_{03}}\right)^{(k_e-1)/k_e}\right] C_{pe} T_{03} \left(1 + \frac{m_f}{m_a}\right) \eta_{tc} \qquad (10)$$

where $\eta_{tc} = \eta_C \eta_T \eta_{mech}$ is the overall turbocharger efficiency, η_C is the compressor efficiency, η_T is the turbine efficiency, and η_{mech} is the mechanical efficiency (allowing for bearing friction, windage, etc.). In this equation, p_2/p_{01} is the compressor pressure ratio, p_4 is the turbine exit pressure, which is equal to ambient pressure plus the pressure loss in the exhaust system downstream of the turbocharger, and T_{01} and T_{03} are the compressor and turbine inlet temperatures, respectively. C_{pa} and C_{pe} are the specific heats at constant pressure, and k_a and k_e are the ratios of specific heat, for the air and exhaust gas, respectively. The turbine inlet pressure p_{03}, which is also the exhaust manifold pressure, is therefore determined by the turbocharger match and cannot be specified independently. This parameter is important because it determines the gas exchange pressure or the pressure difference across the engine, which in turn influences the engine pumping work and the ability to scavenge the engine.

3.2 Matching automotive diesel engines

In a road vehicle, there is normally a direct mechanical connection between the engine shaft and the road wheels, and this imposes considerable constraints on the power–speed characteristics that the engine can follow. For a naturally aspirated engine running at constant air–fuel ratio, the power is almost a linear function of speed and so the torque is nearly constant with speed (curve 1 in Figure 11a). This is maintained from maximum engine speed to some speed near idle. Because the range of road

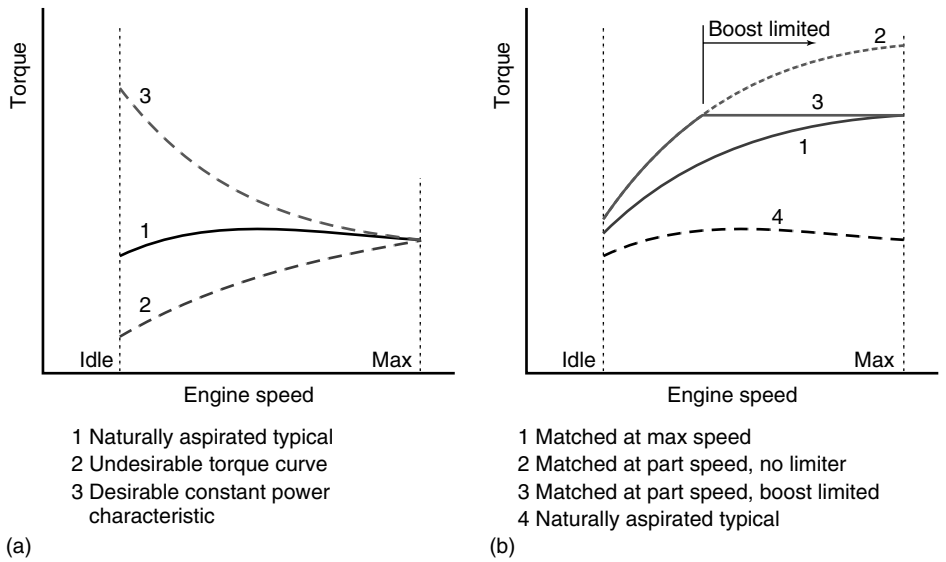


Figure 11. Torque-speed characteristics for (a) naturally aspirated engine (b) turbocharged engine for automotive application.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

speeds required is invariably much greater than the speed range of the engine, a gearbox in the transmission is necessary. A wide torque band is desirable to reduce the amount of gear shifting that is required. This flat torque characteristic is acceptable for normal driving purposes. A falling characteristic of torque with decreasing engine speed (curve 2) is very undesirable because at every grade that the vehicle encounters, the road speed decreases, and the engine speed with it. This would cause the torque to decrease, further reducing the vehicle speed. In every case, a down shift is necessary, and this torque characteristic gives rise to excessive gear shifting. A rising torque characteristic with reducing engine speed (curve 3) would be ideal, providing adequate power at low speeds for acceleration without changing gear, but in a conventional piston engine, this increase in power at low engine speed can only be achieved by a large increase in the fueling rate, leading to poor fuel economy and increased smoke.

For a turbocharged engine, a torque curve that is similar in shape to that of a naturally aspirated engine is desirable, but this requires that the boost pressure at full load is almost constant with speed. This cannot be achieved with a simple turbocharger because of the nonlinear flow characteristic of the turbine. If the turbocharger is matched at the full load, maximum speed condition, then as the engine speed is reduced, the volume flow of the engine decreases approximately in proportion, and the mass flow through the turbine and hence the turbine expansion ratio and the turbine power decrease. This leads to a reduced boost pressure developed by the compressor. At a constant fueling rate, this implies decreasing engine power and torque with speed, which is the undesirable characteristic (curve 1 in Figure 11b). If, alternatively, the turbocharger is matched at a lower engine speed, the turbine flow area is smaller, which means that the pressure ratio across the turbine is higher than it otherwise would be, more turbine power is developed, and the compressor boost pressure is higher at lower engine speed (curve 2). At engine speeds above the matching speed, however, the turbine power output must be limited (curve 3), or the compressor boost pressure will continue to rise. The engine must operate within limits on cylinder pressure, exhaust gas temperature, and turbocharger speed. Any one of these can impose a limit on the boost pressure of the compressor.

Physically, the boost pressure is determined by the power delivered by the turbine. The simplest way to control the turbine power output is by means of a wastegate, which allows some of the exhaust gas to bypass the turbine (Figure 12). By controlling the wastegate opening, it is possible to limit the compressor boost to give a roughly constant engine torque characteristic as shown in curve 3 of Figure 11b. At low engine speed, the wastegate is closed.

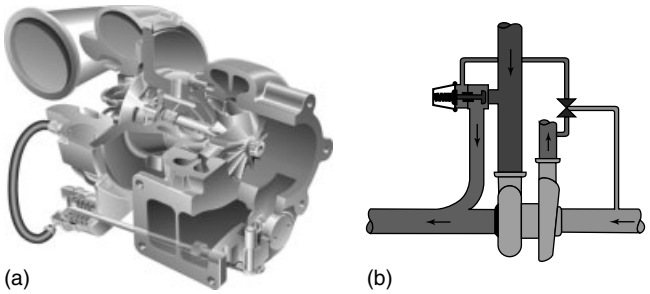


Figure 12. (a) Turbine with wastegate (b) wastegate operation. (Reproduced by permission of BorgWarner Turbo Systems.)

As the engine speed increases from idle, the compressor delivery pressure rises until the boost limit is reached. At this point, the wastegate starts to open and progressively diverts more exhaust from the turbine as the engine speed is increased to its maximum value, holding the boost pressure constant as it does so.

A more sophisticated approach is to use a variable geometry turbine. Variable geometry is a means to control the effective flow area of the turbine, which modifies the turbine mass flow–pressure ratio characteristic as shown schematically in Figure 13. Reducing the effective flow area of the turbine means that the pressure ratio must increase in order to maintain a constant mass flow rate. This increases the engine exhaust manifold pressure, the exhaust energy available, and the power produced by the

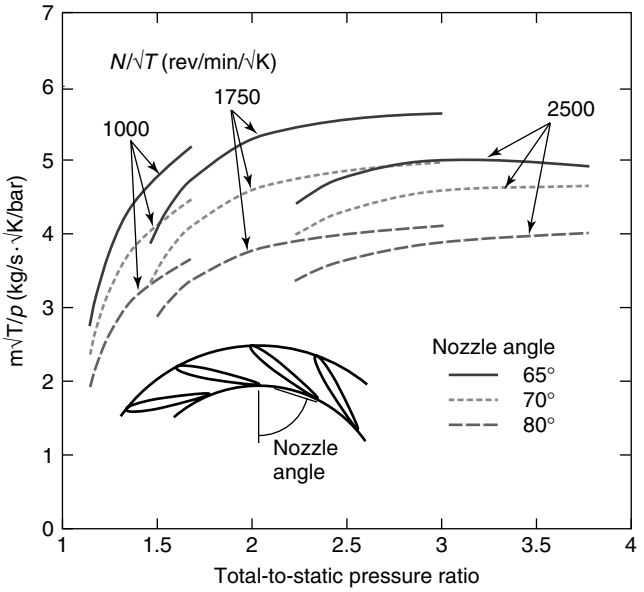


Figure 13. Mass flow–pressure ratio characteristics for a variable geometry turbine.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

turbine. The turbine should be sized to pass the full flow of the engine at the engine maximum speed while the compressor provides the limiting boost pressure. At this point, the variable element of the turbine must be set to give maximum area (a small nozzle angle in Figure 13). At engine idle, the variable element is set to give the minimum area configuration (a large nozzle angle in Figure 13). Thus, as the engine speed is increased, so the turbine area is increased to maintain optimum performance.

The problem of turbine matching can also be addressed by more complex turbocharging systems, including turbocompounding and multiple, switched, turbochargers described in Section 4. All of these are fundamentally methods to modify the basic turbine flow characteristic shown in Figure 9 to make it better matched over a range of engine conditions.

Even if the engine is well matched over most of its range, there is still likely to be an undesirable loss of torque at very low speeds, which can only be mitigated by increasing the fueling rate. This can be done but only within the limits imposed by the need for fuel economy and smoke formation at increasingly rich mixtures. The turbocharger is usually sized to match the engine with the wastegate closed or the variable geometry in the "design" position at about the peak of the engine load versus speed curve, and therefore high boost pressures can be achieved, and the limiting factor here tends to be the cylinder pressure. At higher engine speeds, the exhaust gas temperature and eventually the turbocharger speed (both of which affect the turbine operating life) limit the power developed by the engine.

The implications of this for compressor operation are shown in Figure 14 (Figure 6 shows the complete compressor map, from which details have been omitted here for clarity). For an engine with a fixed geometry turbocharger matched at maximum speed (Figure 14a), the compressor pressure ratio falls with decreasing speed, and the fueling rate is increased as far as possible to limit the decrease. Providing the match point is chosen to give an adequate surge margin, surge at low speeds is unlikely to occur. On the other hand, the match point must not be selected to be so far to the right of the compressor map that the compressor efficiency is poor at full load and full speed. For an engine matched at part speed and boost limited by a wastegate, variable geometry turbine, or some other means (Figure 14b), the match point is much closer to the surge line and the boost pressure at full load and low speed is considerably greater, giving a much better torque characteristic. The fueling rate below the match point speed must not be increased so much as to cause the compressor to surge, even if such an increase lies within the smoke limit. In a turbocharged engine, the available torque band is a function not only of the engine but also of the width of the compressor map. In automotive applications, a wide range of stable operation of the compressor is very important.

At high altitude, the inlet air density is lower and the air mass flow rate falls. As the diesel fuel pump still supplies the same quantity of fuel, the air–fuel ratio falls, the exhaust temperature rises, and the specific available energy of the exhaust gases entering the turbine increases. The turbine exit pressure falls because of the lower ambient

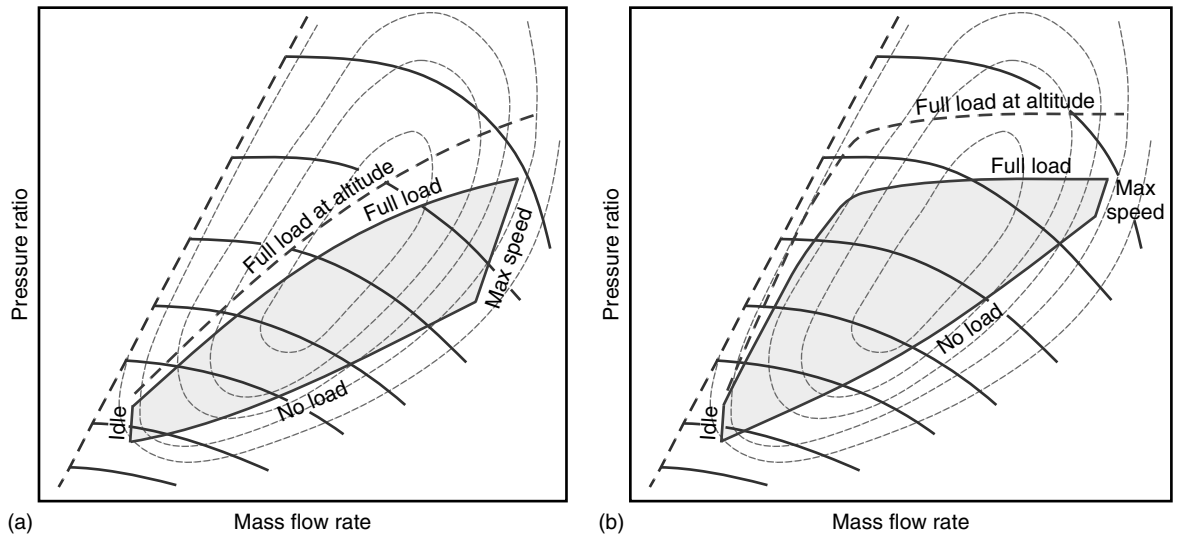


Figure 14. Compressor operation for an automotive diesel engine (a) matched at maximum speed and (b) matched at part speed and boost limited.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

pressure. The combined result is to increase the specific power output of the turbine, and hence, the turbocharger speed and pressure ratio increase. On the compressor maps in Figure 14, the changes in the curves of maximum load with altitude are shown, and the limiting conditions are the turbocharger speed at full load, and in the case of a boost-limited engine, the surge margin near the matching point. If operation at high altitude is envisaged, these limits must be considered when the original turbocharger selection and matching is made.

3.3 Matching automotive gasoline engines

Similar considerations influence the matching of turbochargers to gasoline engines, but this engine has characteristics that give rise to further problems. Gasoline engines generally have a wider speed range than diesel engines and consequently a wider range of airflow rates. The mixture ratio must be carefully controlled within limits to ensure satisfactory combustion.

Because there is less control over the air/fuel ratio, the engine torque is more directly related to the mass of air and hence to the boost pressure. The continued rise of boost with engine speed presents major problems of drivability, and an increase in torque at low engine speed cannot be achieved by substantial increase in the fueling rate as it can in a diesel engine. The solution usually adopted is to use a low pressure compressor design to limit the available boost and to apply some form of boost limiter at high engine speed. A low pressure ratio compressor alone is not sufficient, for if the turbine is matched at the maximum speed and load, at low speed and part load conditions it will be too large, the expansion ratio will be very low and very little boost will be achieved. The use of such a turbocharger will not improve the low speed torque but it will increase the high speed torque, thus making the torque characteristic even more unfavorable.

The solution, therefore, is to match the turbocharger at a relatively low engine speed. This implies a small effective flow area of the turbine. At high engine speed, such a turbine would impose a very high exhaust gas pressure on the engine. To avoid this, a wastegate is used to bypass a fraction of exhaust gas around the turbine, thus limiting the expansion ratio and the power developed by the turbine. In turn, this keeps the exhaust manifold pressure at an acceptable level and prevents the compressor overboosting the engine. Variable geometry turbines would in principle give better control, and while they are widely used in automotive diesel engines, gasoline engines normally run at higher exhaust gas temperature, and the problems of thermal expansion and durability that this gives rise to have limited the application to these engines.

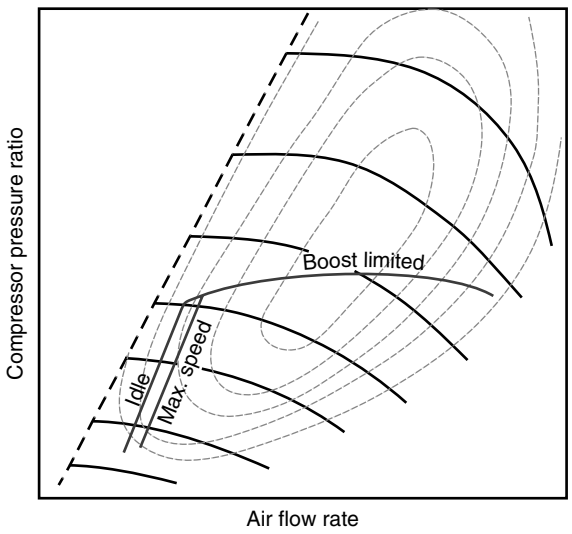


Figure 15. Compressor operation for a turbocharged gasoline engine.

The compressor characteristic for a gasoline engine takes the form shown schematically in Figure 15 (Figure 6 shows the complete compressor map, from which details have been omitted in Figure 15 for clarity). As the engine speed increases from idle, the boost pressure rises to its maximum value when the boost limiter is applied. Thereafter, the boost pressure remains nearly constant. Because of the large speed range of the engine, a wide compressor map width is important. At part speed, the engine is throttled, unlike a diesel engine, and the throttle largely overrides the influence of speed on mass flow rate and so the speed lines collapse into a narrow band.

4 OTHER BOOSTING SYSTEMS

4.1 Series turbocharging

Series or two-stage turbocharging is routinely used in heavy diesel engines for large trucks and off-highway vehicles as a means to achieve boost pressure ratios as high as 5–6 without unacceptable impact on efficiency or range. Such systems are constructed by selecting large and small commercial turbochargers to form the low and high pressure parts of the system. Wastegates or variable geometry may be incorporated in one or both turbines. The relative size of the turbochargers determines how the pressure ratios are split between them, and in general for steady-state operation, there is no advantage in biasing this split strongly in one way or the other.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

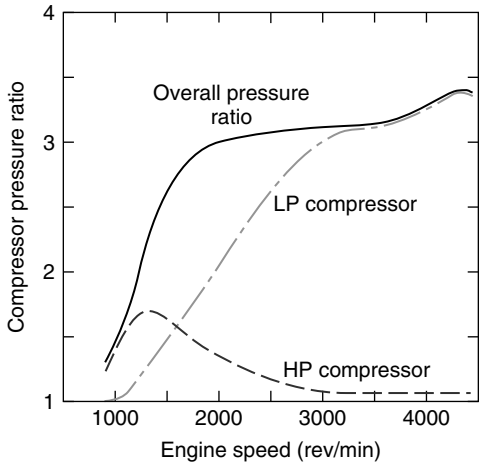


Figure 16. Divisions of pressure ratio in a controlled series turbocharger. (Reproduced by permission of BorgWarner Turbo Systems.)

For smaller diesel engines, the application of series turbocharging is more for low speed torque than for absolute boost level. In such cases, the high pressure (HP) turbocharger is selected to be considerably smaller than the low pressure (LP) turbocharger, so that at low engine speed, most of the exhaust pressure drop occurs across the HP turbine (Pflüger, 1998). The HP compressor can develop the necessary boost pressure and can also respond quickly to engine transients. At high engine speeds, the HP turbocharger must be bypassed completely to avoid overboosting, and the LP turbocharger is entirely responsible for boosting in this region (Figure 16).

4.2 Turbocompound

A turbocompound engine is one in which there is power transfer between the exhaust-driven turbine and the engine. A direct gear drive between the turbocharger shaft and the engine is unsatisfactory for automotive engines, because the limitations that this imposes on the turbocharger speed makes worse the already difficult matching problem. A more feasible solution is a separate power turbine downstream of the turbocharger turbine (Figure 17). At low engine speed, most of the exhaust pressure drop occurs across the turbocharger turbine because it has a smaller flow area than the power turbine, and the power turbine contributes little or no power to the engine. At high engine speed, the power turbine imposes a back pressure on the turbocharger turbine, which limits the turbocharger power and avoids overboosting, while at the same time, the power turbine provides additional power to the engine shaft. This means that the engine fueling can be reduced without a loss of power. Turbocompounding is thus better suited to engines that run for long periods at high load and high speed than for those with a more varied duty cycle (Walsham, 1990).

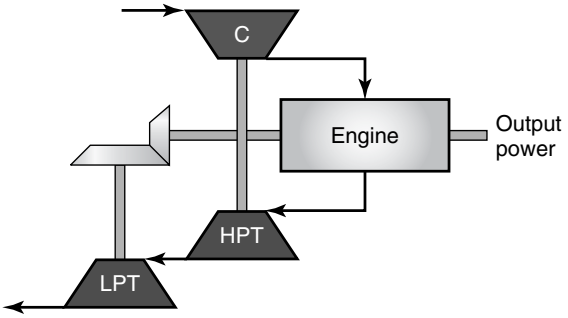


Figure 17. Turbocompound engine.

4.3 Electric assist

Modern developments in electric machines have made feasible small machines with a high power density, which run at shaft speeds comparable to those of turbochargers, and open the way to couple compressors and turbines directly to such machines. With the necessary electric storage capacity and control systems, it is possible to run the electric machine as a motor at low engine speed conditions, giving higher boost pressures and torque, and to run it as a generator at high engine speeds, recharging the batteries and preventing overboosting (Figure 18). The feasibility of such systems is crucially dependent on the

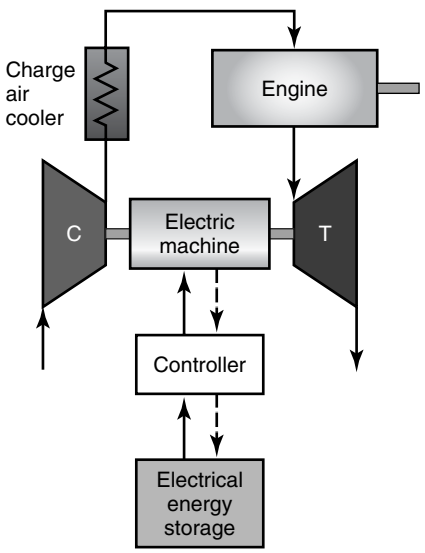


Figure 18. Schematic of an electric assist turbocharger.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

battery capacity, and for this reason may be better suited to hybrid diesel–electric or gasoline–electric powertrains than conventional engines.

Electric machines are also very sensitive to temperature, and cooling the machine in the high temperature environment of the turbocharger is still a considerable challenge. Locating the electric machine between the compressor and turbine gives good packaging but is difficult to manage thermally. Cooling is much easier if the machine is on the compressor end of the shaft, but this makes for a very long shaft with multiple bearings and the potential for rotordynamic instability, and also tends to block the air path to the compressor intake.

The addition of an electric machine also increases the inertia to the rotating system, to the extent that the transient response of the turbocharger following an engine load step may be slower, even though electrical power can be used to assist in accelerating it.

4.4 Hybrid super/turbocharger

A combination of supercharger and turbocharger may be used to boost the engine (Figure 19). The supercharger is required at low engine speed to give adequate boost pressure, but at high engine speed it is declutched and bypassed, and the intake charge is boosted by the turbocharger alone. A bypass or "blow-off" valve is required for the turbocharger compressor to prevent it surging when the supercharger is operating.

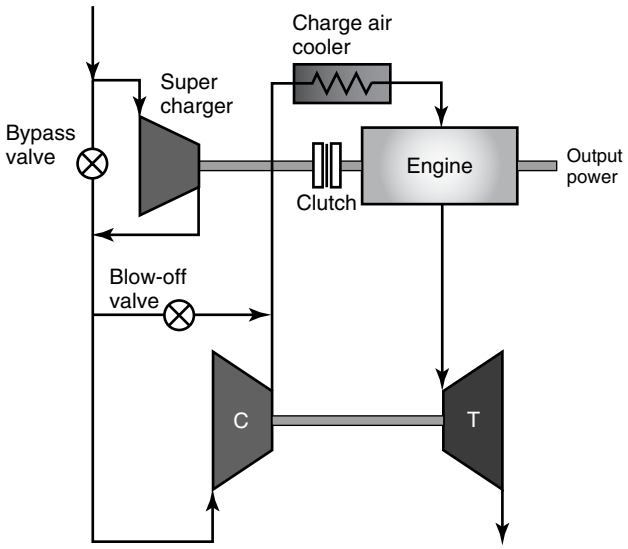


Figure 19. Schematic of a hybrid super/turbocharger.

4.5 Pressure wave supercharger

The pressure wave supercharger uses the pressure waves created by the engine exhaust pulsations to pressurize the intake air. The original and most widely used form of this technology is the Comprex developed by Brown Boveri (now ABB) and shown schematically in Figure 20 (Jenny, Moser, and Hansel, 1986; Jenny *et al*. 1987). A drum containing a set of axially disposed gas passages is rotated by the engine crankshaft. As each passage rotates, it is exposed for a fraction of time to the exhaust, creating a pressure wave that moves along the passage, pressurizing the intake air. Subsequently, the air is released to the intake manifold through a port on the air side of the drum, and the exhaust gas is released through a port on the exhaust side. As it does so, it depressurizes the passage, allowing fresh charge to be drawn in for the next cycle of the drum.

As with other types of supercharger, the device is directly connected to the engine shaft and gives very good transient response, but in this case, only a very small amount of power is required to drive the charger, as the energy to pressurize the intake charge comes from the exhaust. Problems in service include thermomechanical fatigue failures because the drum elements are repeatedly exposed to hot exhaust and cold air, poor off-design performance, and noise created by the pressure waves.

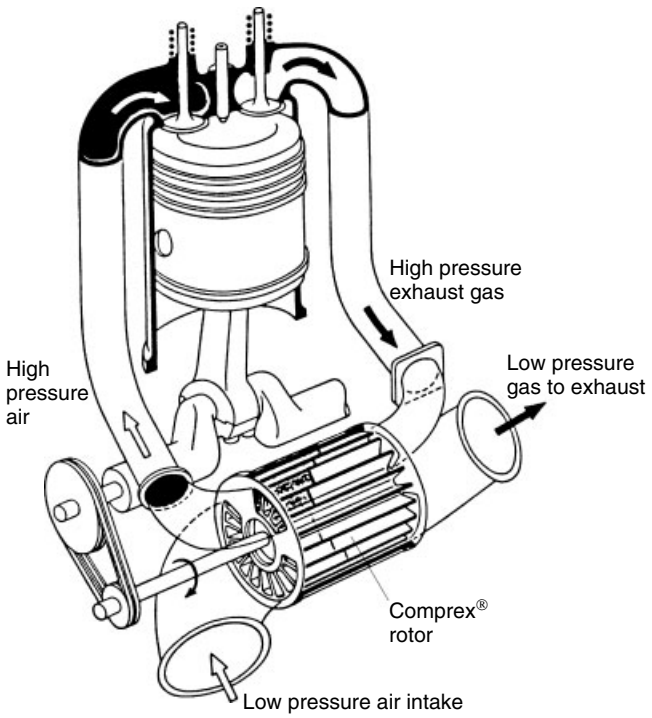


Figure 20. The Comprex pressure wave supercharger.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Cost and installation size have also limited the take-up of superchargers of this type.

REFERENCES

Baines, N.C. (2005) *Fundamentals of Turbocharging*, Concepts ETI, Inc., Vermont, USA.

Capon, G., Leong, A., and Morris, T. (2006) The influence of installation parameters on turbocharged engine performance. *8th Internation Conference on Turbochargers and Turbocharging*, Inst Mech Engrs, pp. 43–54.

Chen, H. and Baines, N.C. (1994) The aerodynamic loading of radial and mixed flow turbines. *International Journal of Mechanical Sciences*, **36**, 63–79.

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*, McGraw-Hill, New York.

Japikse, D. (1996) *Centrifugal Compressor Design and Performance*, Concepts ETI, Inc., Vermont, USA.

Jenny, E., Moser, P., and Hansel, J. (1986) Progress with variable geometry and comprex. *Turbocharging and Turbochargers*, Inst Mech Engrs, pp. 159–70.

Jenny, E., *et al.* (1987) Comprex pressure wave supercharger. *Brown Boveri Review*, **74** (August).

Moustapha, H., Zelesky, M.F., Baines, N.C., and Japikse, D. (2003) *Axial and Radial Turbines*, Concepts ETI, Inc., Vermont, USA.

Pflüger, F. (1998) Regulated two-stage turbocharging—KKK's new charging system for commercial diesel engines. *Turbocharging and Air Management Systems*, Inst Mech Engrs, pp. 127–41.

Walsham, B.E. (1990) Alternative turbocharger systems for the automotive diesel engine. *Turbocharging and Turbochargers*, Inst Mech Engrs, pp. 39–50.

Watson, N. and Janota, M.S. (1982) *Turbocharging the Internal Combustion Engine*, Macmillan Publishers, London.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto126
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Exhaust Gas Energy Recovery

Robert M. Wagner[1], Thomas E. Briggs[2], and Jim P. Szybist[1]

[1] *Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[2] *Southwest Research Institute, San Antonio, TX, USA*

1 Introduction 1
2 Thermodynamics 1
3 Energy Recovery Technologies 2
4 Summary 8
References 9

1 INTRODUCTION

Engine waste heat recovery (WHR) systems have the potential to significantly improve vehicle fuel economy for conventional and hybrid electric power trains spanning passenger to heavy truck applications. The challenge will be to properly match the WHR system to the duty cycle of the power train and constraints of the vehicle. Depending on the circumstances, a WHR system may be relatively small and used only to supplement the alternator for vehicle electrical loads, or it may be a complicated multiloop system that is integrated into a hybrid electric drivetrain or mechanically coupled to the vehicle engine system. The fundamental thermodynamics of the WHR system will be the same regardless of the application or implementation.

Automotive WHR systems are not a new concept. A great deal of work on the design and development of these systems was performed in response to the fuel crises of the 1970s and 1980s in the United States (Patel and Doyle, 1976; Leising *et al*., 1978). Some of this work was performed in conjunction with the adiabatic engine research that was also under way at this time. Ultimately, the sharp decline in fuel prices through the 1980s eliminated the economic incentive for further development or deployment of these systems, and work was largely halted for two decades.

Toward the middle of the first decade of the twenty-first century, interest in WHR systems was rekindled in response to the increasing volatility in petroleum prices, the heavy dependence of national economies on imported oil, and increasingly aggressive fuel economy standards. These factors are providing the necessary economic conditions for a viable WHR business case. In recent years, systems have been developed and demonstrated for both light-duty and heavy-duty engines, including component development for integration of the systems in vehicle infrastructures. These systems are expected to become commercially viable and more available over the next decade.

2 THERMODYNAMICS

The feasibility of implementing an engine WHR system is dictated by the availability of recoverable heat. Figure 1, which shows the distribution of fuel energy usage on a first law basis for a diesel passenger vehicle engine [adapted from Edwards and Wagner (2010)], is an example of potential sources of recoverable heat. In general, about 25–30% of the fuel energy provides useful work through the crankshaft under road load conditions, and the balance is divided between heat losses and thermal energy discarded through the exhaust. These waste heat streams are often very different in terms of temperature and mass flow. Depending on the thermal quality of the waste heat

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto127
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

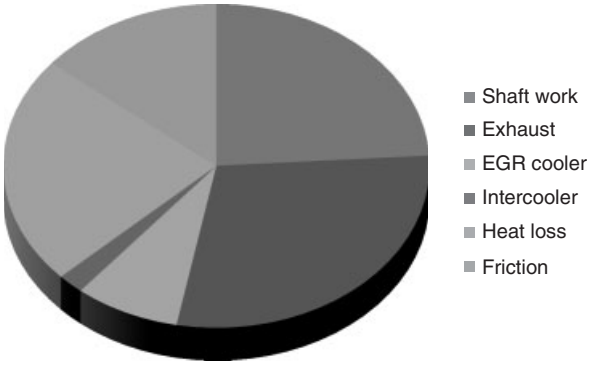


Figure 1. Example energy distribution from a light-duty diesel engine under road load conditions. (Reproduced by permission of ASME.)

streams, the potential for extracting additional work can vary widely.

The necessary approach to characterizing available energy for recovery is through a second-law analysis that focuses on exergy. Exergy is a measure of a system's potential to do useful work because of physical and chemical differences between the system and the ambient environment. An example of a second-law exergy balance for the same engine and operating conditions of Figure 1 is shown in Figure 2. A comparison of the energy and exergy distributions provides a better understanding of the recovery potential of different waste energy streams. For example, Figure 1 [adapted from Edwards and Wagner (2010)] shows a distribution of the fuel energy in a light-duty diesel engine. The energy lost through heating the coolant is the primary component of the "heat loss" segment. Comparing the size of the heat loss segment in Figure 1 with that of Figure 2, which represents the distribution of the fuel exergy, shows that the relatively low temperature of the coolant reduces the potential for extracting work from that heat flow. The exergy portion of the distribution is smaller because of the low temperature of the coolant, which is a much less valuable source of energy for the WHR system compared to the very warm exhaust gas (higher temperature delta with environment). The higher exergy of exhaust gases has made them the primary source of interest for most WHR systems.

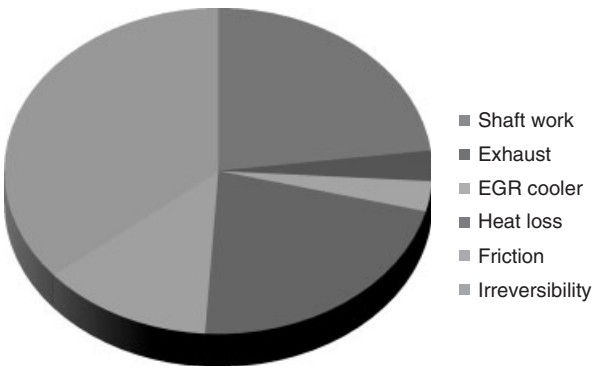


Figure 2. Second-law exergy distribution from a light-duty diesel engine under road load conditions. (Reproduced by permission of ASME.)

3 ENERGY RECOVERY TECHNOLOGIES

There are a variety of viable options for recovering waste heat from an engine system and converting it into usable energy. The three major classifications of energy recovery systems are thermomechanical cycles, thermoelectric (TE) devices, and turbocompounders. A thermomechanical cycle uses waste heat as an input source to a power cycle. These cycles are familiar to most engineers and include Rankine and Brayton cycles. TE devices make use of the Seebeck effect for direct electrical power generation from solid-state devices driven by a thermal gradient. Exhaust turbocompounders make use of the pressure and thermal energy of the exhaust to produce shaft work. These various approaches are discussed here.

3.1 Thermomechanical power cycles

3.1.1 Rankine power cycle

The Rankine vapor power cycle, which typically uses water as the working fluid, provides the basis for many stationary power plants. A diagram of a standard Rankine cycle is shown in Figure 3. The four components of the

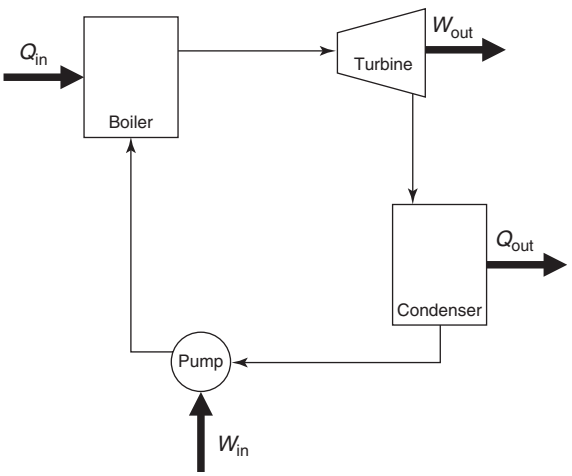


Figure 3. Standard Rankine vapor power cycle configuration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto127
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

standard cycle are a boiler to vaporize the working fluid; an expander to extract work from the hot, pressurized working fluid; a condenser to cool the fluid back to a liquid state; and a pump to bring the fluid back to the boiler pressure. While a more complicated cycle design is used in stationary power generation to obtain the highest practical energy conversion efficiency, a simple cycle is preferable in automotive applications to minimize the weight and volume of the system.

The first critical design choice for a Rankine cycle system is that of the working fluid. Water is the most commonly used fluid in stationary systems because of its combination of thermodynamic properties and low cost. However, it has multiple disadvantages for an automotive WHR system. Most obviously, it freezes at a relatively high temperature, making it unsuited for year-round use in most markets. It also requires relatively large heat exchangers, compromising component packaging. Finally, as the steam expands in a turbine, it can easily expand to the point at which water condenses. This is unacceptable from a performance standpoint as the water droplets can be damaging to turbine blades spinning at high velocity. This two-phase state is readily achieved because of the slope of the vapor dome on the right side of a T-s diagram for water, as seen in Figure 4. A typical expansion process is shown; unless there is a significant degree of superheat, a large expansion ratio will easily cause the water to begin condensing during the expansion rather than remaining a vapor until entering the condenser.

Given the challenges of implementing a water-based Rankine system, many other working fluids have been evaluated. Ethanol is a compound that has received some attention and has been proposed as the working fluid for a heavy-duty vehicle project (Sisken, 2011). Ethanol has a more suitable freezing point than water; however, the thermal properties of ethanol limit the potential efficiency of a WHR system using it as a working fluid. The vapor pressure of ethanol at typical ambient conditions is subatmospheric, which limits the lower temperature limit for the power cycle, thus limiting cycle efficiency. If the system is permitted to operate subatmospherically, there is a significant chance for air to be drawn into the system past the pressure seals. At this point, the flammability of ethanol in air becomes a concern, particularly given the temperature of the exhaust gases and hot surfaces present in a WHR system.

Most experimental WHR system studies have used a refrigerant as the working fluid to obtain a reasonable compromise between thermal performance, working temperature range, vapor pressure, and safety. Early work tended to use R-12, as in a study by Patel and Doyle (1976). More recent studies by Nelson (2009) and Briggs *et al.* (2010) have used R-245fa because of its improved environmental characteristics, including a lower global warming potential (GWP) compared to other fluorinated refrigerants. These studies have noted the restricted operating temperature range of the fluid, but overall system performance within the available range has been acceptable. Newer refrigerants under development promise at least equal performance to R-245fa with even better environmental characteristics, which should permit them to be used into the future without concern for restrictions on their

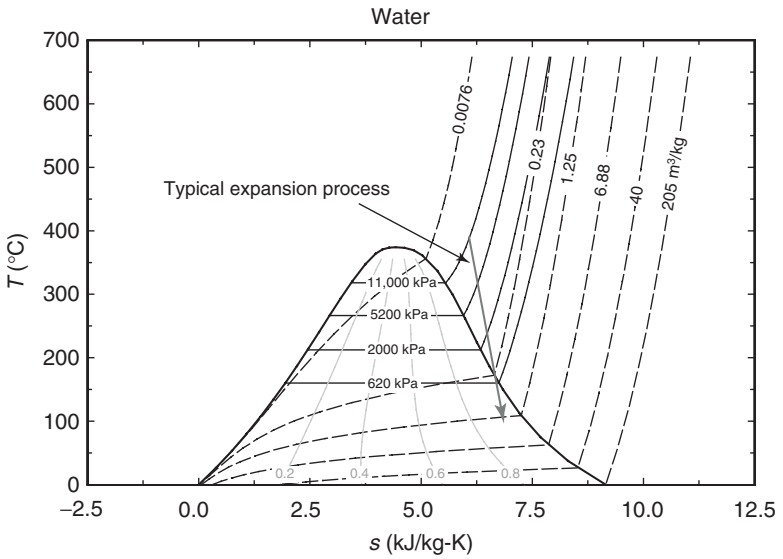


Figure 4. Temperature–entropy diagram for water.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto127
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

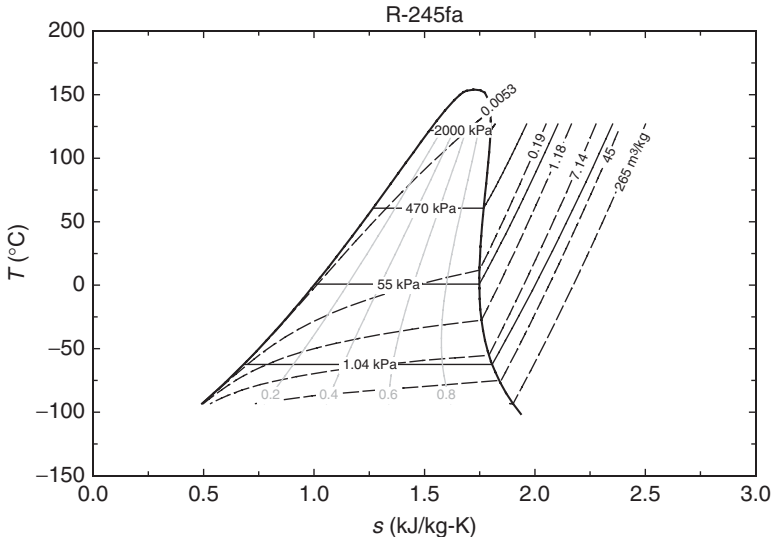


Figure 5. Temperature–entropy diagram for R-245fa.

use. Another advantage is that many refrigerants are "dry" fluids, which means that the slope of the right side of the vapor dome is negative, as seen in Figure 5. For the same type of expansion as described earlier for water, there is no possibility of condensing the fluid during expansion.

The choice of working fluid will impact the choice of expander technology. A wide variety of machinery has been used, including turbines (Nelson, 2009; Briggs *et al*., 2010), pistons (Endo *et al*., 2007), and scroll expanders (Sisken, 2011). Other expander designs will likely continue to be developed. It is well understood from power generation experience that turbines are not compatible with condensing fluids as the droplets lead to pitting and erosion of the blades as a result of the high relative momentum between the drops and the metal blades. Pistons and scroll expanders are more forgiving of liquid drops, making them compatible with condensing fluids. Considering the already-discussed working fluids, water has been shown to be prone to drop formation in an expander, making it more suitable for nonturbine systems. As a dry fluid, R-245fa will be more suitable for use with a turbine, as well as working with other designs.

Expander sizing is a complicated topic and is beyond the scope of this chapter. The approach is analogous to that used for matching a turbocharger to an engine or a pump to an application. The same tools that are used for these common design problems are suitable for the design of a WHR system.

Depending on the working fluid chosen, two or three heat exchangers will be required for a Rankine cycle WHR system. For a system using water or ethanol, the temperature–pressure relationship of the fluid will permit a two-exchanger system as the temperature of the working fluid drops significantly in the expander, leaving little remaining exergy to recuperate within the cycle. For many organic fluids such as R-245fa, the temperature does not drop significantly across the expander, and maximum cycle efficiency demands a recuperator to recycle the remaining thermal energy back to the cold working fluid. The recuperator transfers heat from the fluid leaving the expander to the fluid entering the boiler. This reduces the amount of heat rejected in the condenser and can significantly raise the efficiency of the Rankine cycle at the cost of an additional component.

The performance of the WHR system is highly dependent on the performance of the individual components. Assuming a fixed choice of working fluid, the overall flow rate of the fluid must be matched to the operating condition of the engine and the corresponding available thermal energy sources. An example of this design issue is shown in Figure 6, taken from Edwards, Wagner, and Briggs (2010). The figure shows that there is a significant difference in the optimum working fluid flow rate as the engine conditions are varied. Through the use of a variable rate pump, the working fluid flow can be controlled to better match the engine operation over transient cycles where both the exhaust temperature and the mass flow may vary significantly. The extent of matching engine operation is limited by other design considerations such as the thermal capacity of the system.

The expander efficiency will be the limiting factor for the transient capability of the WHR system. For many

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto127
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

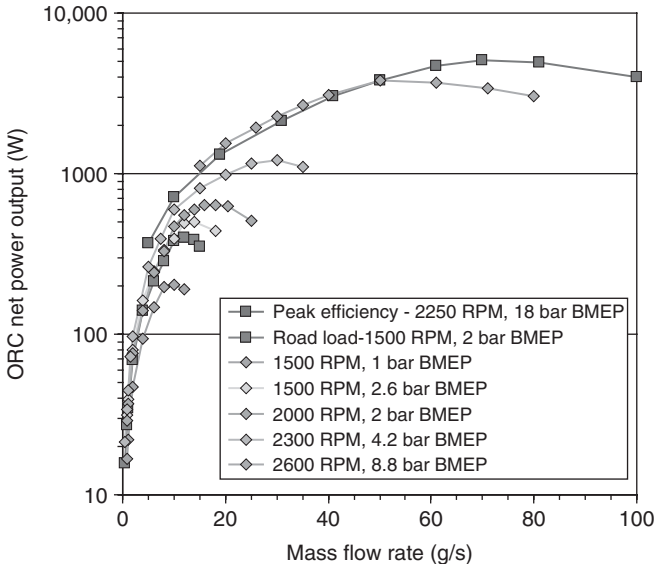


Figure 6. Organic Rankine cycle (ORC) working fluid flow rate for various engine conditions. (Reproduced by permission of ASME.)

expander designs, the efficiency map is quite narrow, with only a small operating range exhibiting high efficiency. The expander must be sized for a maximum efficiency corresponding to the most frequent operating condition of the engine. As the engine operates transiently, the efficiency of the WHR system will vary widely based on the efficiency map of the expander. Depending on the design of the system and operating point, it is even possible that the expander will need to be bypassed at very high engine loads and speeds to flow sufficient working fluid through the system to prevent overheating or over pressurization in the WHR system. Obviously, any bypassed flow produces no power and thus no energy recovery and potentially a parasitic loss.

An example of the performance of a Rankine-based WHR system is found in Briggs *et al*. (2010). In this study, a recuperated Rankine cycle was used to recover thermal energy from the post-turbine exhaust of a GM 1.9-L diesel engine. At the operating point specified in this chapter, the exhaust temperature was 436°C at a flow rate of 317.4 kg/h. The Rankine cycle, using R-245fa working fluid, used a radial inflow turbine expander directly coupled to a high speed electrical generator. The electrical output from the expander was 4.3 kW, and the Rankine pump required 0.3 kW, yielding a net electrical output of 4.0 kW. The authors claimed a Rankine cycle efficiency of 12.7% (thermal to electric) based on heat extraction from the exhaust of 31.4 kW_{th}. This additional output increased the brake thermal efficiency (BTE) of the base engine from 42.4% to 45%. While somewhat low relative to what might be expected in an ideal implementation of a recuperated Rankine cycle, this efficiency is reasonable. An extension of this study was performed by Valentino, Hall, and Briggs (2013) using a model tuned to the experimental data. The system efficiency was found to remain relatively constant over a wide range of operating conditions, with the power output from the WHR system scaling with the enthalpy flow in the exhaust stream.

A study presented by Nelson (2009) on a similar Rankine WHR system installed on a Cummins 15-L on-highway diesel engine indicated an estimated 6–7% increase in the BTE of the engine using waste heat from the exhaust gas recirculation (EGR) and exhaust streams (i.e., from 42% to 48–49%). This study suggested that a shift of 10% in BTE would equate to 1800 gal of fuel savings every year, so the demonstrated increase would yield 1100–1300 gal of fuel savings. The author did not discuss any assumptions about system weight or other details in the calculation of fuel savings.

3.1.2 Brayton power cycle

The Brayton cycle is a gas power cycle version of the Rankine power cycle. The key difference is that the working fluid is a gas and hence does not undergo a phase change in the heating and cooling processes. The standard Brayton cycle is a closed cycle with heat addition and rejection accomplished via heat exchangers. The most common application of the Brayton power cycle is in gas turbine engines such as those in aircraft, large ships, and natural gas fired power plants. These applications use an open cycle with heat addition supplied by combustion. To date, there has not been significant interest in using a Brayton cycle for automotive WHR. Caterpillar conducted a concept study (Kruiswyk, 2008) evaluating the feasibility of using an open Brayton cycle with air as the working fluid, but it was found that the size of the required heat exchangers was unacceptably large for on-truck use. One realization of a closed Brayton cycle is a supercritical CO_2 power cycle, which has been studied for application to nuclear power plants to increase the overall plant efficiency (Wright *et al*., 2010). The high operating pressures present challenges for on-road use as the system would require relatively heavy components to withstand the CO_2 pressure with an acceptable safety margin.

3.2 Thermoelectric devices

TE devices are an attractive option for WHR in vehicles because they are solid-state devices. As a result, their integration into an exhaust system presents fewer additional

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto127
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

moving parts and only enough active management to ensure sufficient cooling. This section focuses on the integration of TE exhaust heat recovery into vehicle systems; it does not provide a comparison or overview of TE materials. TE materials are only discussed to provide an understanding of their functionality in a vehicle system.

3.2.1 Thermoelectric figure of merit

TE materials take advantage of the Seebeck effect so that when a temperature differential is applied a voltage differential and a current flow result. There are a number of different classes of TE materials of interest for automotive WHR, including bismuth telluride, skutterudites, half-Heusler alloys, and nanomaterials. For more information on TE materials for automotive applications, see Nolas, Poon, and Kanatzidis (2006).

TE performance is characterized by a figure of merit, ZT_{AV}, defined by Equation 1, where α is the Seebeck coefficient, ρ is the electrical resistivity, and λ is the thermal conductivity.

$$ZT_{AV} = \frac{\alpha^2}{\rho \times \lambda}\left(\frac{T_{Hot} + T_{Cold}}{2}\right) \quad (1)$$

The TE figure of merit is related to the efficiency of the TE materials by the relationship provided in Equation 2.

$$\eta_{TE} = \frac{(T_{Hot} - T_{Cold})}{T_{Hot}} \frac{[(1 + ZT_{AV})^{1/2} - 1]}{[(1 + ZT_{AV})^{1/2} + T_{Cold}/T_{Hot}]} \quad (2)$$

Thus, TE material efficiency increases as the figure of merit increases and as the temperature gradient between the hot and cold sides of the TE material increase. However, the relationship between TE material efficiency and the efficiency at which a TE system recovers exhaust heat from an engine is complex, and high TE material efficiency does not necessarily mean efficient heat recovery. To illustrate this point, the following section focuses on TE heat recovery systems in vehicles.

3.2.2 Thermoelectric heat recovery systems

A schematic of the heat flow pathways, shown in Figure 7, is helpful in understanding the challenges associated with incorporating TE generators into the automotive exhaust system. First, the exhaust enters a heat exchanger with a thermal energy flux of $Q_{Exhaust}$, a portion of which is transferred to the heat exchanger wall (Q_{HX_wall}). A portion of Q_{HX_wall} is lost to ambient, but the remainder is transferred as heat to the TE devices (Q_{TE_in}). The TE

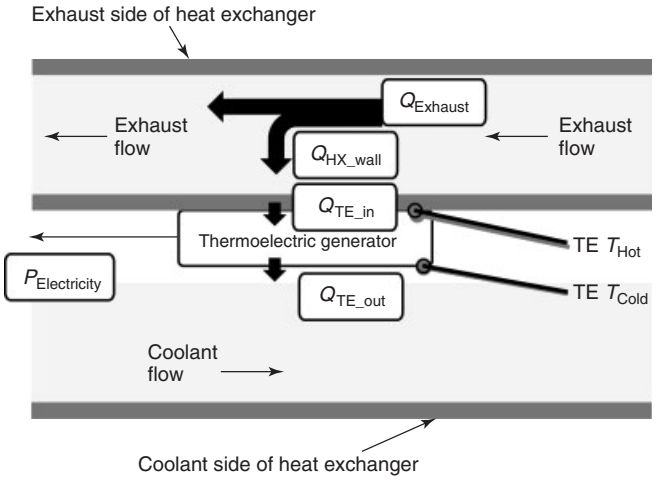


Figure 7. Schematic of heat flows for a thermoelectric heat recovery system in an engine exhaust.

device is capable of converting a portion of the Q_{TE_in} into usable electricity ($P_{Electricity}$), but the remainder exits the device in the form of waste heat (Q_{TE_out}). $P_{Electricity}$ is proportional to the temperature difference across the TE device, from the hot side temperature (TE T_{Hot}) contacting the exhaust duct to the cold side temperature (TE T_{Cold}), which is in contact with an air or liquid heat exchanger for heat rejection. The heat exchanger efficiency (η_{HX}) for this system is $Q_{TE_in}/Q_{Exhaust}$, the TE efficiency (η_{TE}) is $P_{Electricity}/Q_{TE_in}$, and the overall system efficiency (η_{System}) is $P_{Electricity}/Q_{Exhaust}$. Thus, it is desirable to have a high η_{TE}, but if this efficiency comes at the expense of η_{HX}, it will not result in an increase in η_{System}, and it may actually be a detriment to overall system efficiency.

As with other exhaust heat recovery systems, the heat exchanger design and sizing are critical. The exhaust system on a vehicle must be sized to accommodate full load operation, where the exhaust flows and temperatures are highest. An exhaust heat exchanger for TE heat recovery must be sized so that there is not a substantial backpressure at the highest engine loads that detracts from engine efficiency, but it also must operate well at low engine speeds and loads, where light-duty engines typically operate during normal driving cycles. Thus, the heat exchanger system for a TE exhaust energy recovery system must be designed to perform well over a wide range of exhaust flow rates and temperatures.

Extracting thermal energy with a heat exchanger is uniquely challenging for a TE system compared to other heat recovery methods because of the increased thermal resistance in the heat exchanger. In a Rankine cycle, for example, there is typically only one heat exchanger wall between the hot exhaust gas and the working fluid. In a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto127
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

TE system, there are two heat exchanger walls, the TE materials, and typically ceramic coatings to electrically isolate the TE generators, with each layer reducing the system thermal conductivity and the heat flux into the heat exchanger (Q_{TE_in} in Figure 7). The thermal resistance of TE devices is not negligible, and research efforts focused on increasing the figure of merit achieve these improvements by decreasing the thermal conductivity of these materials. This approach results in a more efficient conversion of the heat flux to electricity but simultaneously reduces the heat flux through the TE generator.

Figure 8 further illustrates the relationship between power output of a TE generator module and the figure of merit for a constant temperature differential across the TE device with arbitrary changes in thermal conductivity and electrical resistance. Initially, power output increases with decreases in both thermal conductivity and electrical resistance. However, at $ZT_{AV} = 2.0$, the recovered power reaches a maximum with decreasing thermal conductivity because of the decrease in Q_{TE_in}. Conversely, continued decreases in electrical resistance are effective at continually increasing the TE power output. In reality, there are limits to the extent that thermal conductivity and electrical resistivity can be changed and neither can be changed in isolation from the other. Thus, Figure 8 illustrates that a higher figure of merit does not necessarily translate into an increase in recovered electrical power, but the trade-off point is highly dependent on the particular type of TE material.

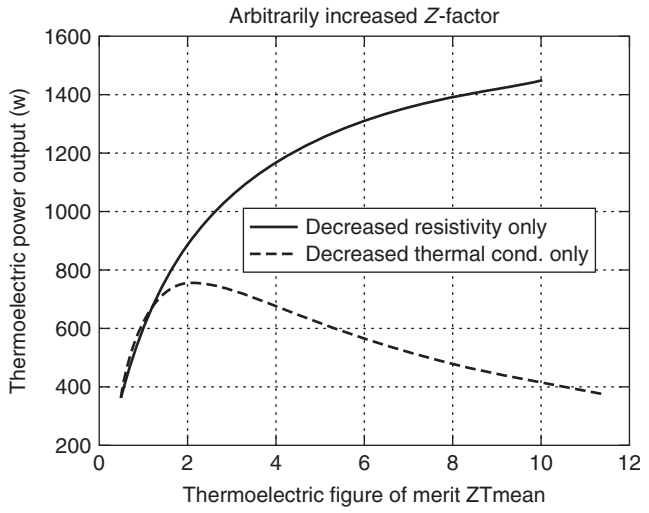


Figure 8. Thermoelectric power as a function of the figure of merit for constant hot- and cold-side temperatures. Changes in the figure of merit are accomplished by arbitrarily increasing the thermal conductivity or decreasing electrical resistivity.

3.2.3 Thermal expansion and thermal management

In other exhaust system heat exchangers, the heat is transferred from one fluid to another fluid (e.g., the exhaust gas to the working fluid of the organic Rankine cycle) through a heat exchanger made from a single material, typically stainless steel. With TE systems, the heat exchanger design is more challenging in many respects because there are different thermal coefficients of expansion for the heat exchanger material; the TE materials; and any other spacers, insulation, or thermal conductors.

In addition, the figure of merit is a function of temperature for each type of TE material, and it is essential not to exceed the maximum temperature of a TE material. For instance, bismuth telluride TE modules typically exhibit an optimal figure of merit between 100 and 200°C, with a maximum temperature around 250°C. As a result, incorporating bismuth telluride TE generators into an exhaust system may require an exhaust bypass to offer thermal protection at high exhaust temperatures. Conversely, skutterudites and half-Heusler alloys can offer good performance at temperatures as high as 800°C but have figures of merit that monotonically increase with temperature and therefore offer very poor performance at low temperatures. Thus, bismuth telluride may be more applicable to compression-ignition (CI) diesel engine TE generators, whereas the higher temperature materials may be more applicable to spark-ignition (SI) gasoline engines.

The last part of the TE system is the cold side of the heat exchanger for heat rejection (Q_{TE_out} in Figure 7). While it is possible to use either an air-cooled heat sink or a liquid heat exchanger, Haidar and Ghojel report that the cold-side temperature would have to be about halved for an air-cooled heat sink to be as effective as a liquid heat exchanger (Haidar and Ghojel, 2002). As a result, liquid-cooled heat exchangers are favored. Further, for the sake of simplicity and minimizing the system weight, it is desirable to use the existing engine liquid coolant system. Engine coolant is typically at a temperature of 90°C or higher, which is higher than an ideal cold-side temperature for a TE device. Figure 9, taken from Ibrahim, Szybist, and Parks (2010), shows that electrical power recovered decreases by more than 25% with a coolant temperature increase from 50 to 90°C. Thus, there is a substantial amount of electrical power to be gained with lower coolant temperatures. However, this again represents a trade-off because a dedicated exhaust coolant system will also require a significant amount of power for a pump and coolant fans. Whether a dedicated cooling system operating at a lower temperature is desirable will be dependent on a number of factors and will have to be assessed on a case-by-case basis.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto127
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

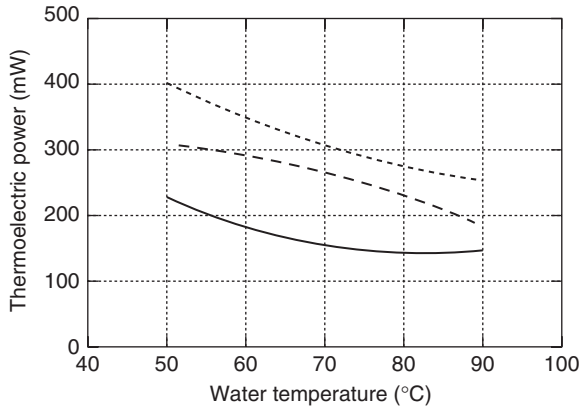


Figure 9. Thermoelectric power as a function of coolant water temperature for three different exhaust flow rates: 40 sL/min (——), 60 sL/min (– – –), and 80 sL/min (. . . .).

3.2.4 Energy recovery potential

The energy recovery potential of TE systems is dependent on the type of TE material used, the placement of the TE generator in the exhaust system, and the size of the TE generator. Ibrahim, Szybist, and Parks (2010) reviewed the TE electrical power recovery in vehicle demonstrations and found a range from 1 to 193 W for light-duty applications, with a large dependence on vehicle operating condition. Considering advancements in materials and system integration, this range of electrical power recovery is in-line with Mori *et al*. (2011) who estimate that a 3% fuel economy enhancement is possible by recovering approximately 450 W in a series hybrid configuration with a TE generator that uses both low and high temperature TE materials. Electrical power output of more than 1 kW from a TE generator has been reported from a heavy-duty truck at a high load condition (Haidar and Ghojel, 2001).

3.3 Turbocompounding

Most modern CI engines and many modern SI engines use turbocharging to recover some of the waste exhaust thermal energy by converting it into additional airflow through the engine. This alone is not generally considered a WHR technique as turbocharging is more useful for improving power density, transient response, and emissions. However, it is possible to use a second turbine after the turbocharger to extract additional energy from the exhaust by more completely expanding it to ambient pressure before exhausting it from the engine. This approach is not well suited to automotive applications though and is only briefly discussed here for completeness.

Turbocompounding has been implemented on a number of on-highway heavy-duty diesel engines such as the 2010–2012 DD15 engine by Detroit Diesel Corporation, n.d. (2012). In a turbocompounded engine, the post-turbocharger exhaust is expanded through a second turbine. This second turbine is not coupled to a compressor wheel but is instead coupled to the engine crankshaft through a fixed-ratio transmission. The expansion work across the turbine is then added to the crankshaft to increase the power delivery for a given fuel rate. Given the constraints of gear ratios, the rotational speed of the turbine is limited both in range and in maximum speed. The narrow operating range of a heavy-duty diesel engine is better suited to match the turbine speed range to aerodynamic designs than a wider-range light-duty engine. For a heavy-duty engine application, a useful increase in efficiency is possible, with a study quoted by Reinhart (2009) suggesting that a typical line-haul truck could see a 2.4–2.8% improvement in fuel economy. Other constraints, principally cost and durability, have limited the market penetration of turbocompounded engines in the heavy-duty market though.

The literature is sparse on the use of turbocompounding systems for light-duty applications. A study by Edwards, Wagner, and Briggs (2010) in which a turbocompound system was simulated on a light-duty vehicle indicated negligible benefit for light-duty applications. This minimal benefit, coupled with the added cost and significant engineering challenges in matching a turbine to the speed range of a light-duty engine, suggests that turbocompounding is not a suitable technology for improving light-duty engine efficiency.

4 SUMMARY

Engine WHR systems will be an important part of future high efficiency engine systems to meet increasingly aggressive CO_2 emissions regulations. The selection of a particular WHR technology will be highly application dependent. More specifically, the appropriate technology will be dictated by the quality of available thermal sources, duty cycle of the engine, acceptable complexity of the recovery system, and cost. The Rankine cycle WHR systems have been demonstrated both in-laboratory and in-vehicle to offer an attractive increase in engine efficiency, but the added cost has prevented manufacturers from putting systems into production. TE systems are available in production for niche uses but still require substantial development of the basic technology to achieve costs and component efficiencies that will justify production. All of the WHR technologies discussed in this chapter have been available for many years, and the technology is advancing

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto127
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

rapidly with specific focus on maximizing the recovery efficiency of low temperature sources and reducing complexity and cost for more widespread implementation. Given the continuing demand for better fuel economy, it is expected that ultimately both Rankine and TE WHR systems will be available for light- and heavy-duty automotive applications.

REFERENCES

Briggs, T.E., Wagner, R.M., Edwards, K.D., *et al*. (2010) A waste heat recovery system for light duty diesel engines. SAE Technical Paper 2010-01-2205.

Detroit Diesel Corporation (n.d.) DD15 TC Engine [online], http://www.demanddetroit.com/engines/dd15tc/default.aspx; accessed November 28, 2012.

Edwards, K.D. and Wagner, R. M. (2010) Investigating Potential Efficiency Improvement for Light-Duty Transportation Applications Through Simulation of an Organic Ranke Cycle for Waste-Heat Recovery. *Presented at the ASME 2010 Internal Combustion Engine Division Fall Technical Conference*, San Antonio, TX.

Edwards, K.D., Wagner, R.M., and Briggs, T.E. (2010) Investigating potential light-duty efficiency improvements through simulation of turbo-compounding and waste-heat recovery systems. SAE Technical Paper 2010-01-2209.

Endo, T., Kawajiri, S., Kojima, Y., *et al*. (2007) Study on maximizing exery in automotive engines. SAE Technical Paper 2007-01-0257.

Haidar, J.G. and Ghojel, J.I. (2001) Waste Heat Recovery of Low-Power Diesel Engine using Thermoelectric Generators. *Proceedings of the 20th IEEE International Conference on Thermoelectrics*, Beijing, China, 413–418.

Haidar, J.G. and Ghojel, J.I. (2002) Optimization of the Thermal Regime of Thermoelectric Generators in Waste Heat Recovery Applications. *International Conference on Thermoelectrics, Proceedings Ict '02*, 427–430.

Ibrahim, E.A., Szybist, J.P., and Parks, J.E. (2010) Enhancement of automotive exhaust heat recovery by thermoelectric devices. *Proceedings of the IMechE Part D: Journal of Automobile Engineering*, **224**, 1097–1111.

Kruiswyk, R.W. (2008) An Engine System Approach to Exhaust Waste Heat Recovery. *Presented at the 2008 Directions in Engine Efficiency and Emissions Conference*, Detroit, MI.

Leising, C.J., Purohit, G.P., DeGrey, S.P., and Finegold, J.G. (1978) Waste heat recovery in truck engines. SAE Technical Paper 780686.

Mori, M., Yamagami, T., Sorazawa, M., *et al.* (2011) Simulation of fuel economy effectiveness of exhaust heat recovery system using thermoelectric generator in a series hybrid. *SAE International Journal of Materials and Manufacturing*, **4** (1), 1268–1276.

Nelson, C. (2009) Exhaust Energy Recovery. *Presented at the 2009 Directions in Engine Efficiency and Emissions Conference*, Detroit, MI.

Nolas, G.S., Poon, J., and Kanatzidis, M. (2006) Recent developments in bulk thermoelectric materials. *MRS Bulletin*, **31**, 199–205.

Patel, P.S. and Doyle, E.F. (1976) Compounding the truck diesel engine with an organic Rankine cycle system. SAE Technical Paper 760343.

Reinhart, T. (2009) *Alternatives for improving heavy truck efficiency*. Presented at the 2009 DERC Symposium, Madison, WI.

Sisken, K. (2011) SuperTruck–50% Improvement in Class 8 Freight Efficiency. *Presented at the 2011 Directions in Engine Efficiency and Emissions Conference*, Detroit, MI.

Valentino, R., Hall, M.J., and Briggs, T.E. (2013) Simulation of organic Rankine cycle electric power generation from spark ignition and gdesel engine exhaust flows. SAE Technical Paper.

Wright, S.A., Radel, R.F., Vernon, M.E., *et al.* (2010) Operation and analysis of a supercritical CO_2 Brayton cycle. Sandia Report SAND2010-0171.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto127
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Lubrication and Friction

Victor W. Wong

Massachusetts Institute of Technology, Cambridge, MA, USA

1 Introduction 1
2 Engine Friction and Mechanical Losses 2
3 Component Lubrication and Friction 5
4 Effects of Lubricants and Additives 15
Acknowledgments 19
References 19

1 INTRODUCTION

The trend toward greater energy conservation and the reduction of green house gases demands that fuel consumption of automotive engines continue to be improved. Although the useful work loss due to engine friction is relatively small for modern engines, the reduction of all parasitic energy losses, including friction, remains as a valuable contribution to overall efficiency improvement. A small gain in fuel consumption, even by 1% over existing levels, is an important achievement. The macroscopic energy and economic savings from improved engine efficiency are huge. Lubrication and friction play essential roles in energy conservation.

There are many moving parts in an engine. Proper lubrication keeps them in good working order, extends component longevity, and minimizes the energy losses due to friction. Many engine durability and reliability issues—such as excessive wear, component seizure, and catastrophic failure—are traced to problems in the inadequate lubrication of essential components. Proper lubrication and low friction are associated with engine integrity and good performance, which are attributes important to the engine user.

Lubrication involves the smoothing of the rubbing process between contacting surfaces. A lubricant film between the surfaces would prevent direct solid-to-solid contact. The degree of solid-to-solid contact and the oil film thickness depend on the applied mechanical load, relative velocity, surface profiles, roughness, textures, as well as lubricant properties. There are different types of lubrication conditions or regimes, the fundamentals of which are illustrated. There are many contacting surfaces in an engine system: in the piston assembly, valvetrain components, and multiple bearing surfaces. The relative magnitudes of friction in these components will be examined.

The lubricant itself is a multi-constituent fluid that strongly influences the lubrication regime of the lubricated parts. Various additives provide different functions in the oil: to maintain the temperature sensitivity of the oil viscosity, to protect against wear through formation of surface films, and to reduce solid-to-solid friction by making the surfaces more slippery. In addition, other additives keep the component surfaces clean and maintain the oil properties to within acceptable levels. In recent years, lubricant-additive derived ash in the exhaust stream has become an important issue in advanced diesel engines equipped with emission aftertreatment control systems. Engine design and lubricant-additive formulation need to be optimized together to simultaneously protect both the engine and the emission-control system from contamination by ash, sulfur, and phosphorus originating in the oil.

This chapter begins with clarifying the common descriptions of mechanical losses and friction in the engine, and the lubrication fundamentals. Then, the various lubrication regimes in the engine component subsystems and individual contacting parts are described. The lubrication regimes span a wide range from predominantly hydrodynamic lubrication

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

in the bearings to mixed-boundary lubrication in the valvetrain cam lobe and follower contacts. The final section discusses the roles of lubricants and additives.

2 ENGINE FRICTION AND MECHANICAL LOSSES

2.1 Types of mechanical losses, rubbing friction and engine work output

When fuel is burned in the combustion chamber of a reciprocating engine, the high pressure that is generated pushes on the piston to generate work. This work on the piston is termed *indicated work*. Owing to losses in the engine system, only part of the indicated work—the brake work—is delivered to the drive shaft to drive the transmission system to propel the vehicle. We define mechanical losses as the difference between the indicated work done on the piston and the brake work delivered to the drive shaft. These losses come from various sources and depend on a range of factors, such as engine operating conditions, mechanical design of the various components (see e.g., Engine Configurations, Piston and Ring Development, Cranktrain Development) and the lubricant.

The total mechanical losses that have been defined are occasionally called *total engine friction* (Heywood, 1988). When used in this context, total friction includes more than the usual connotation of the rubbing friction of the moving parts. Mechanical losses, in general, include the rubbing friction of the moving parts, the pumping work to pump gases into and out of the combustion chamber, and the accessories work to drive the oil and coolant pumps, alternator, and other engine-integrated accessories. Each of these losses contributes to the reduction of net engine work output.

In terms of the mean effective pressure (see Operating Principles), that is, work per engine cycle normalized by the engine displacement volume (Heywood, 1988), the losses and the net work output are related as follows:

$$\text{Total mechanical losses}: \quad \text{tfmep} = \text{rfmep} + \text{pmep} + \text{amep}$$

$$\text{Brake work output}: \quad \text{bmep} = \text{imep (gross)} - \text{tfmep}$$

where

tfmep = total friction mean effective pressure (i.e., mechanical losses)
rfmep = rubbing friction mean effective pressure
pmep = pumping mean effective pressure (of intake and exhaust gases)
amep = accessories mean effective pressure
imep (gross) = indicated mean effective pressure (from fuel combustion)
bmep = brake mean effective pressure (delivered to the engine drive shaft)

This chapter addresses rubbing friction only, that is, the simple connotation of friction, as this is the component of losses that is affected directly by engine lubrication. However, all contributions to total mechanical losses are important in affecting engine efficiency. Rubbing friction will be abbreviated hereafter simply as friction, or mechanical friction.

2.2 Lubrication regimes

Because of the wide range of forces and relative speeds at the lubricated surfaces in an engine, different regimes of lubrication can occur, as distinguished by the thickness of the oil film separating the surfaces compared to the surface roughness. Figure 1 illustrates the levels of separation of two rough surfaces. *Hydrodynamic lubrication*, or *fluid-film lubrication*, refers to the regime where the surfaces are separated by a sufficiently thick oil film so there is no direct solid–solid contact. The mean separation of the surfaces is much greater than the mean asperity height or surface roughness. Friction in this regime is determined by traditional fluid mechanics, where oil viscosity plays a primary role. *Boundary lubrication* occurs when the surfaces come into direct contact with each other as if there were no lubricant in between. Friction is governed predominantly by the surface or surface-film characteristics. This regime occurs when the mean separation between the surfaces is the same in magnitude as the asperity height variations. Between the two regimes is *mixed lubrication*, where the separation of the surfaces is close and, statistically, there

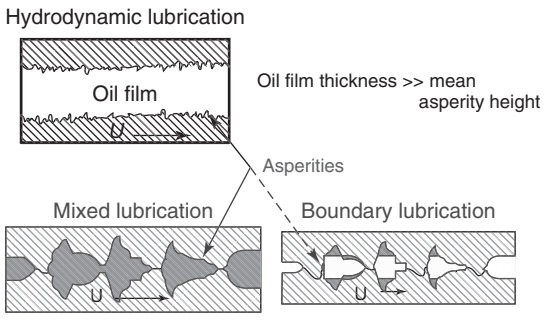


Figure 1. Regimes of lubrication shown on the scale of surface asperities. (From Moughan (2006). Reproduced by permission of Massachusetts Institute of Technology.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

are significant areas of the rough surfaces that come into contact. Both lubricant and surface properties affect friction in this regime.

2.2.1 The Stribeck curve

The lubrication regimes and the friction behavior can be illustrated by a Stribeck curve, shown in Figure 2, where the coefficient of friction (tangential force/normal force) is plotted logarithmically against a dimensionless duty number, $\mu V/P$, where μ is the oil viscosity, V is the relative speed between surfaces, and P is the load per unit length. At a very low duty parameter, which can be a combination of low oil viscosity, low relative speed and/or high loading, film thickness is low, resulting in boundary lubrication with a high friction coefficient. For two sliding wedges, for instance, oil film thickness increases at higher relative speeds and lighter loads or higher viscosities, which correspond to conditions at higher duty parameters, toward the right side of the figure. When the film thickness is sufficiently high, hydrodynamic lubrication occurs. Notice that in the mixed lubrication region where an oil film is formed even at close separation of the surfaces, the coefficient of friction significantly drops from that of boundary lubrication. The Stribeck curve illustrates one fundamental effect that lubricant viscosity has on friction: reducing oil viscosity (thinner oil) lowers viscous drag in hydrodynamic lubrication, but will eventually increase friction when the film thickness becomes so thin that solid-to-solid contact begins to occur in the mixed lubrication regime and reaches a maximum in the boundary lubrication regime.

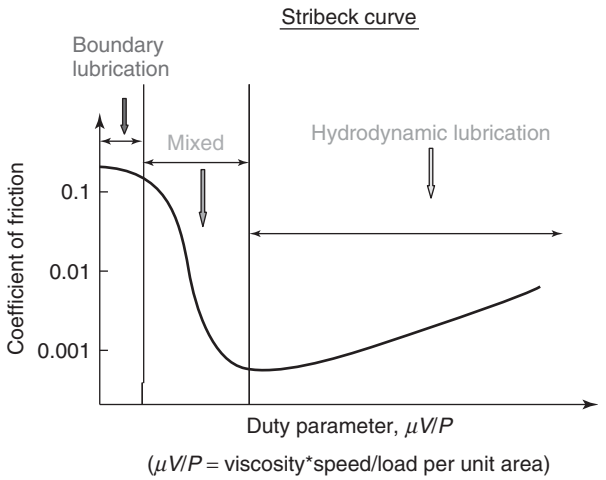


Figure 2. Stribeck curve.

2.2.2 Boundary lubrication

Under boundary lubrication, friction is determined by the characteristics of the surfaces, such as the asperities, material properties, and roughness or surface textures. Solid or nonsolid films, formed by various surface reactions, also affect friction and wear. These surface films include chemically formed tribofilms originating from antiwear additives, adsorbed layers of friction-modifier additives or other surface-active agents. The friction coefficient is relatively constant in this regime and does not vary significantly with the duty parameter, especially the oil viscosity. Lubricant additives can affect the coefficient of friction via formation of the surface films.

2.2.3 Hydrodynamic lubrication

Hydrodynamic lubrication occurs when the oil film thickness is much higher than the surface roughness or asperity heights, and the surfaces do not touch each other. Accordingly, the effects of material properties or surface characteristics are negligible. Friction comes primarily from viscous shear stress. The viscous shear is proportional to the oil viscosity to some exponent, considering film thickness variations between the particular surfaces, and to the relative velocity of the surfaces. Thus, the coefficient of friction shows an approximate linear slope in the Stribeck curve.

2.2.4 Mixed lubrication

This is an intermediate regime where the separation of the surfaces is close, and an effective oil film separates the surfaces from major direct contact. However, statistically, owing to asperity height variations, parts of the surfaces do make direct contact. The extent to which direct solid-to-solid contact occurs determines the friction coefficient. Friction reduces significantly as the oil film thickness or the duty parameter increases from pure boundary lubrication.

2.3 Relative magnitude of friction and distribution by components

This section first discusses magnitudes of friction relative to fuel energy consumed and relative to engine work output, and then summarizes estimates of the distribution of friction losses by engine components.

2.3.1 Friction compared to engine work output, fuel energy use and to other losses

While friction is a strong function of engine speed (rpm), it varies less directly with engine load. Increasing the power

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

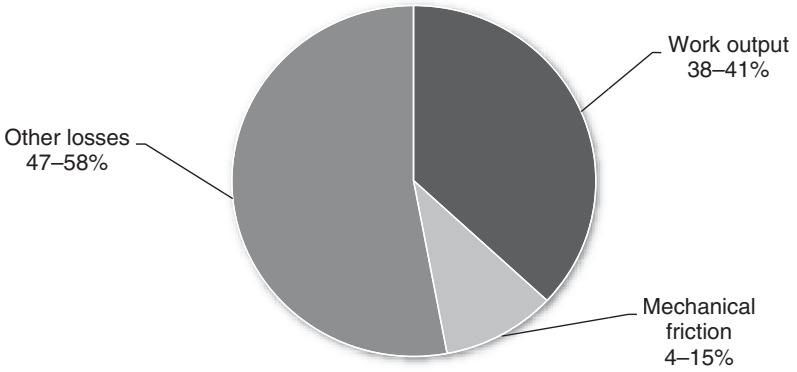


Figure 3. Distribution of total energy in a fired engine. (Reproduced with permission from Richardson (2000). © ASME.)

output for a given sized engine at a given speed (viz. increasing the bmep, see Operating Principles) is a typical strategy for reducing friction as a percentage of engine work output. There are typical estimates of the relative magnitude of friction for common engine size and power output classes; however, these mostly empirically based estimates (Sandoval and Heywood, 2003; Taraza *et al*., 2007; Patton, Nitschke, and Heywood, 1989) span a wide range and do not point to a simple distribution quantitatively.

A typical estimate of friction for a fired engine (diesel or SI) as a fraction of total fuel energy used is shown in Figure 3 (Richardson, 2000), in which *mechanical friction is* shown to take up roughly 4–15% of the *total fuel energy*. This general estimate reflects typical in-use engine conditions on the aggregate over various operating conditions. It does not apply to unique extreme conditions such as at idling and at very light loads where most of the fuel energy is consumed to overcome friction, with no net power output. Peak thermal efficiencies (work output/fuel used) of modern engines vary between 33% and 50%, with passenger-car engines at the lower end of the range and larger bore engines doing better and targeting 50% as a common development goal (Koeberlein, 2012). Accordingly, *mechanical friction* is typically 10–30% of engine *power output*, although it could be 100%, at idling, at the extreme.

The above-mentioned estimate of mechanical friction is consistent with other estimates of total mechanical losses in an engine. Those estimates include pumping and accessory losses in addition to mechanical friction itself, at up to 40% of the gross (indicated) power output from the engine (Quillen *et al*., 2006, 2007; Nakada, 1995; Taylor, 1993). Most of the mechanical losses, about 75%, are rubbing friction, although the relative pumping losses become more significant at lighter loads (Heywood, 1988).

The mechanical efficiency, a measure of the mechanical losses, is defined as

$$\text{Mechanical efficiency} = \text{bmep or (brake power)/gross imep (or indicated power)} = 1 - \text{mechanical power loss/gross indicated power} \times 100\%$$

As engine power output from a given engine increases, friction becomes less as a percentage of power output. Therefore, mechanical efficiency typically increases with engine load, as shown in Figure 4. Friction could be a small fraction of engine power output, at 10% or less at high loads. Its relative importance increases at lighter loads, at 30% or more at part loads.

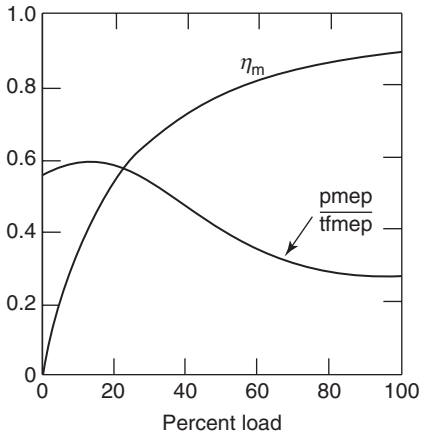


Figure 4. Mechanical efficiency η_m and ratio of pumping mep (pmep) to total friction mep (tfmep) as a function of load for a typical spark ignition engine at fixed speed (Reproduced with permission from J. Heywood, Internal Combustion Engine Fundamentals (Heywood, 1988) © The McGraw-Hill Companies, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.3.2 Breakdown of friction by engine components

This section compares the relative contributions to the total mechanical (rubbing) friction in the engine by the various component groups. The mechanisms of friction and lubrication of the various component systems are discussed in a later section.

Pumping losses result from the flow of intake and exhaust gases. Accessories include coolant and lubricant pump, fans, and other pneumatic systems that may be powered directly by the engine. The losses in these systems depend on parameters other than the traditional concept of lubrication or a lubricant. They comprise 20–30% of total mechanical losses for accessories for heavy-duty diesels and 30–50% for pumping loss for gasoline engines, depending on the operating speed and load. While important, these losses are not included in the current focused discussions on mechanical or rubbing friction.

With these exclusions, the three major subsystems of the engine contributing to mechanical friction are thus: (i) piston-ring-liner system, (ii) crankshaft and bearings system, and (iii) valvetrain system. The exact distribution of the friction among these three groups depends on the particular engine, the component design details, and operating conditions. However, prevalent reported results show that the crankshaft system (main bearing and seals) contributes roughly 50–100% higher friction than the valvetrain system, and the power cylinder friction approximately equals that from the valvetrain and crankshaft bearing systems combined. Figure 5 shows a typical partitioning of the mechanical friction in the engine, among the three major component groups (James, 2012; Richardson, 2000; Taylor, 1993). Friction and lubrication in these components groups are discussed next.

3 COMPONENT LUBRICATION AND FRICTION

3.1 The piston-assembly system

The piston assembly (see Piston and Ring Development) consists of the piston, piston rings, piston pin, connecting rod and bearings, as shown schematically in Figure 6. There are three main friction and lubrication groups: (i) the piston-skirt surfaces sliding up and down the liner, (ii) the ring-face surfaces of the ring pack likewise in reciprocating motion along the liner, and (iii) the bearing surfaces in rotating motion in the wrist pin and connecting rods. The friction and lubrication in the bearings are similar to that in the crankshaft main bearings and thus are discussed in the next section. Most of the piston-assembly friction comes from either (i) piston–skirt/liner interaction, or (ii) ring–pack/liner interaction. Strictly speaking, there is also lubrication and friction as the rings slide radially against the inside surfaces of the ring grooves in which the rings reside. However, the ring–groove interactions are only intermittent and do not contribute significantly to energy losses, but rather to ring–groove wear issues.

3.1.1 The piston-skirt-liner subsystem

Because of the kinematics of the connecting rod transmitting the piston reciprocating motion to rotating crank motion, side forces act on the piston laterally, causing what is termed the *secondary motion of the piston* inside the cylinder. Piston secondary motion results primarily in (i) a variable slight tilt of the piston as it rotates about the piston-pin, and (ii) an impact force, commonly called *piston slap*, of the piston as it switches from sliding up on one side of

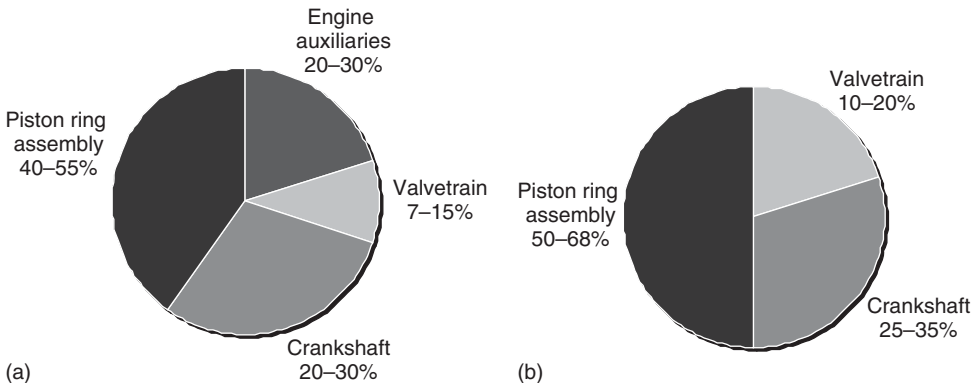


Figure 5. (a) Distribution of total mechanical losses (From James (2012). Reproduced by permission of Massachusetts Institute of Technology.) and (b) friction in a diesel engine (Reproduced with permission from Richardson (2000) © ASME and Taylor (1993). © Elsevier.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

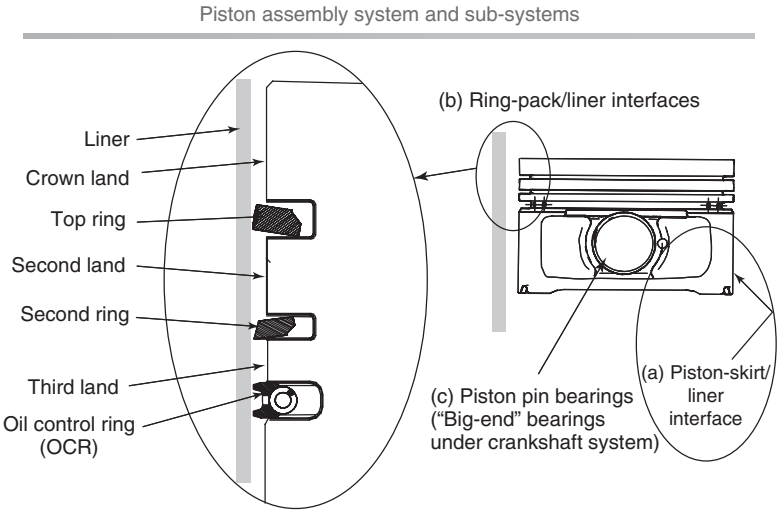


Figure 6. Piston assembly system showing (a) piston-skirt/liner subsystem, (b) ring-pack/liner subsystem, and (c) piston-pin/piston bearing surfaces. Con-rod "big end" bearings under crankshaft section.

the liner (minor-thrust or anti-thrust side) to sliding down on the other side (major-thrust or, simply, thrust side) of the liner. The piston tilt is affected to a large extent by the skirt profile, while the operating clearance between the piston and liner and the thickness of the oil film thickness between them significantly affect the side impact force. Although the piston rings provide vital sealing functions, the side forces on the piston are supported mainly by the piston skirts instead. The rings move relatively freely in their grooves and do not exert much side force on the piston other than through the friction on the ring groove surfaces.

The lubrication regimes and friction losses in the piston-skirt-liner subsystem are significantly influenced by the piston secondary motion. As one would expect, skirt-liner friction is higher when there is solid–solid contact in the boundary lubrication and mixed lubrication regimes. The axially barrel-shaped skirt profile, acting like a hydrofoil, generates the hydrodynamic pressure to sufficiently separate the skirt from the liner. Hydrodynamic lubrication is thus maintained during most parts of the piston travel. However, when the piston speed approaches zero at the ends of the piston travel up or down strokes, the hydrodynamic lift due to the sliding motion fades. Instead, the squeeze-film damping, occurring when the skirt moves sideways toward the liner, becomes the essential mechanism for maintaining a reasonable oil film, although often not thick enough to avoid solid–solid contact.

The important parameters governing piston-skirt-liner friction include the surface characteristics, such as textures or waviness patterns on the skirt and surface roughness; skirt design details such as ovality and axial profile, and lubricant thickness and rheology.

The piston skirt is considered compliant and flexible in response to mechanical loads such as the oil film pressure itself. The mechanical deformations add challenge in predicting skirt-liner lubrication. It was reported on the basis of modeling results (Mansouri and Wong, 2004) that a more compliant skirt provides a greater separation between the skirt and liner surface, thus lower friction, as shown in Figure 7, where computationally the flexibility (deformation response to applied load) of the skirt was reduced to zero (rigid skirt) or made several times more compliant in the model.

Figure 8a (Moughon, 2006) illustrates conceptually typical effects of increasing viscosity in piston-skirt friction, where the hydrodynamic friction increases and boundary friction decreases with increasing oil viscosity for a skirt design with a fair amount of boundary lubrication. In this case, a thicker oil maintains a larger skirt-liner separation and consistently reduces friction. In Figure 8b, however, in a different skirt design with less boundary lubrication, increasing oil viscosity would increase friction beyond an optimal point, as hydrodynamic lubrication becomes dominant and a lower viscosity would decrease friction.

The key in reducing piston skirt-liner friction lies in maintaining hydrodynamic lubrication of the skirt. With an adequate oil supply to the skirt, most other issues of skirt profile design and surface characteristics affecting boundary lubrication would disappear or diminish. This is illustrated in Figure 9, where it is shown (Mansouri and Wong, 2004) that piston-skirt friction can be reduced by reducing primarily the boundary contacts between the liner and skirt surfaces; this can be achieved by providing ample oil supply to the skirt.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

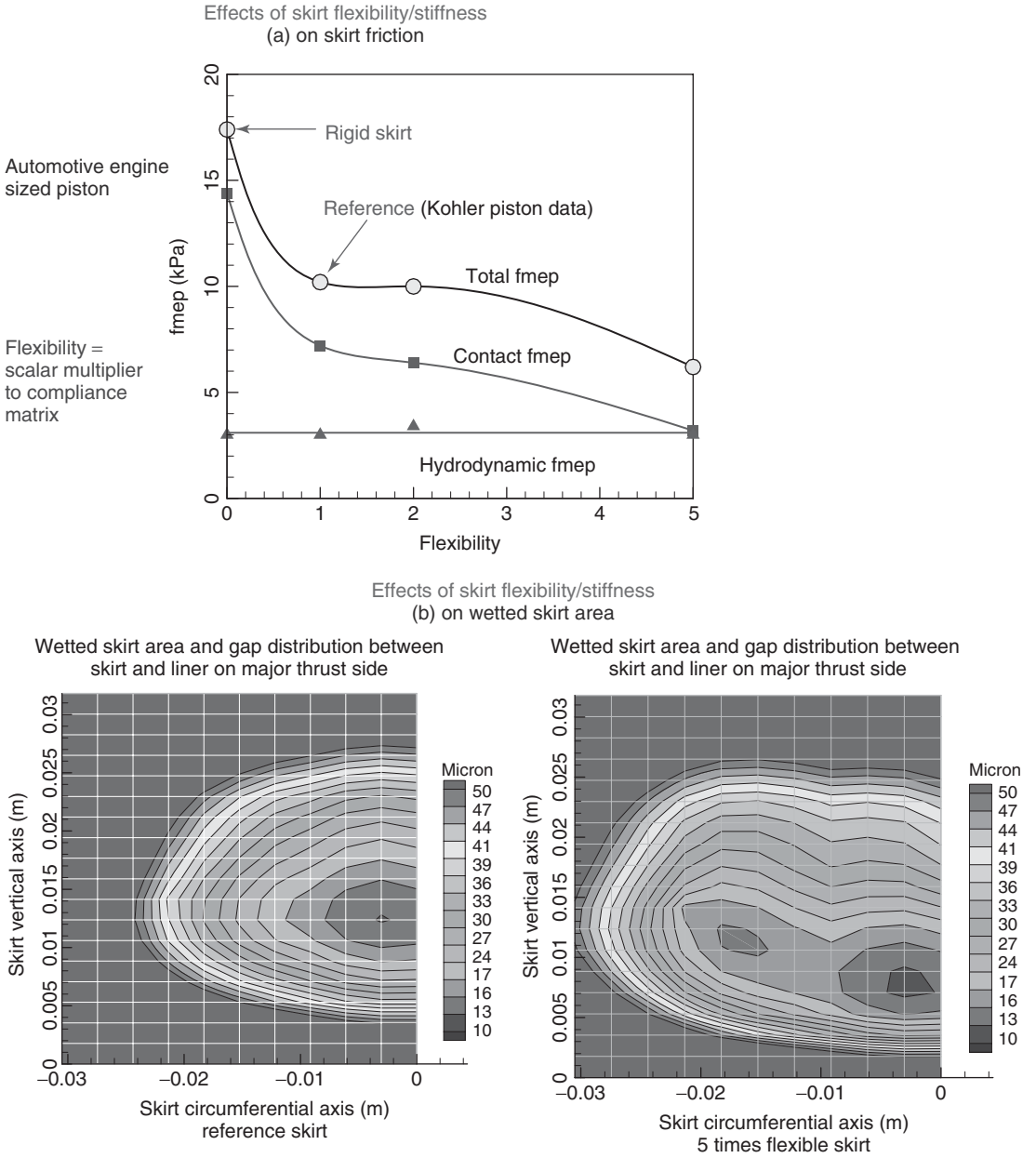


Figure 7. Effects of piston-skirt flexibility/stiffness (a) on skirt-liner friction, and (b) on wetted area of piston-skirt and skirt-liner separation. (From Mansouri and Wong (2004). Copyright © 2004 SAE International. Reprinted with permission.)

3.1.2 The piston ring-pack subsystem

An automotive-engine piston ring pack usually consists of three rings as shown in Figure 6 (see Piston and Ring Development). Uninstalled, the top two rings have diameters larger than the cylinder bore. When compressed and installed in the grooves in the piston and fit into the cylinder liner, they expand against the liner, and this force is called the *ring tension*. The third ring from the top is the oil-control ring, which is either a two-piece design in many diesel engines or a three-piece design in gasoline engines. Ring tension in the oil control ring is provided by an expander piece. The top ring, or compression ring, primarily seals the combustion chamber gas from leakage past the ring. This action is accomplished by the ring tension and combustion gas pressure at the back of the ring. The function of the second ring, or scraper ring, is more intricate: first, the second ring performs an additional sealing function and its face profile is shaped to scrap oil

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

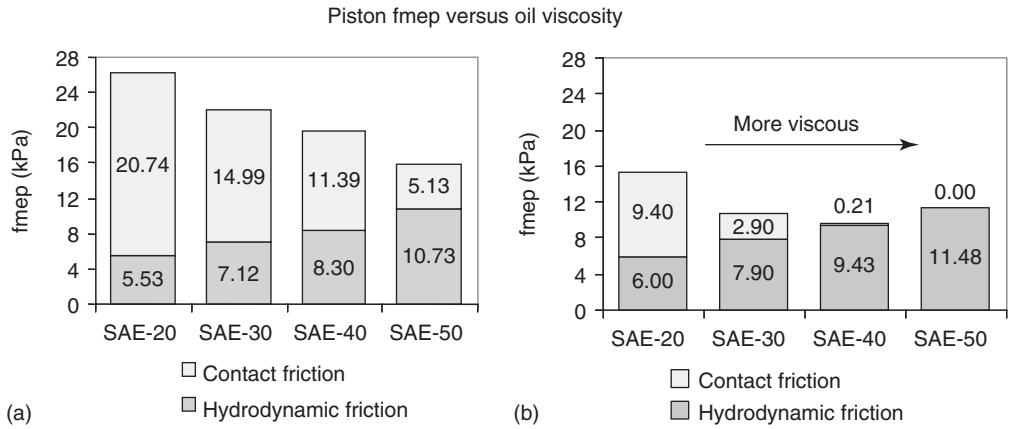


Figure 8. Computer calculations showing effects of oil viscosity on piston-skirt/liner friction, illustrating dependence on degree of mixed/boundary lubrication (a) significant mixed-boundary lubrication in sharp-curvature skirt profile, (b) moderate mixed-boundary lubrication in shallow-curvature skirt profile. (Reproduced from Moughan and Wong (2005) © ASME.)

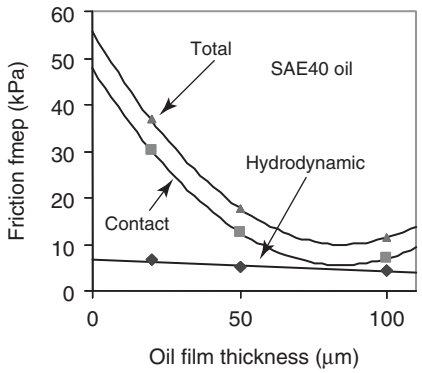


Figure 9. Computations showing effects of adequate upstream oil film thickness (oil supply) on piston-skirt friction for an 18-L natural gas engine at 1800 rev/min full load. Boundary contact friction diminishes rapidly as skirt becomes adequately lubricated. Curves between points are fitted. (Reproduced with permission from from Mansouri and Wong (2004). © ASME.)

on the liner down, away from the combustion chamber. Furthermore, the second ring controls the inter-ring gas pressures, and thereby the flow of blow-by gases toward the crankcase or their reverse flow back into the combustion chamber. This subtle action is accomplished by the careful balance of a combination of design factors of the ring, such as ring twist (preferential bending and resulting contact with the ring groove), ring gap, and mass and geometry of the ring. As the name implies, the oil-control ring controls the amount of oil available to the upper rings for adequate lubrication but minimum oil consumption.

The relative sliding speed between the rings and the liner varies substantially over the engine cycle, and so does the lubrication regime for each of the rings. Boundary friction is dominant near the end strokes where the relative rubbing velocity is zero and oil film thickness minimal. Near the mid-strokes of piston travel, the reverse is true. While the exact proportion of boundary versus hydrodynamic friction varies with specific mechanical design and operating parameters, the oil-control ring is expected to operate preferentially with more boundary lubrication overall. Contributing to this oil-control-ring performance is the high ring tension and the relatively small rails against the liner. In general, both the top-ring and the oil-control ring friction are significant, while the second-ring friction is generally considered the smallest in the ring pack, owing to the relatively lower ring tension and lower gas pressure behind the ring.

Although shown with parallel top and bottom groove and ring surfaces in Figure 6, various shapes of rings are used in practice. The Keystone ring and groove, characterized by a tapered ring and groove for diesel engines, facilitates the removal of combustion residues due to the radial movement of the rings relative to the grooves. Moreover, the running surfaces of rings are often coated with wear-resistant materials. Significant engineering has gone into piston-ring designs (see Piston and Ring Development); the following subsections can only cover the general lubrication and friction characteristics of the rings rather than their detailed design and engineering.

3.1.2.1 Lubrication and friction in the ring pack.

3.1.2.1.1 Fundamental ring-liner slider analysis. The basic understanding of ring-liner lubrication is shown conceptually by a slider arrangement in Figure 10, where a slider (ring) executes a reciprocating motion relative to the liner. The radial load on the ring consists of the pressure force at the back of the ring in the groove acting

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

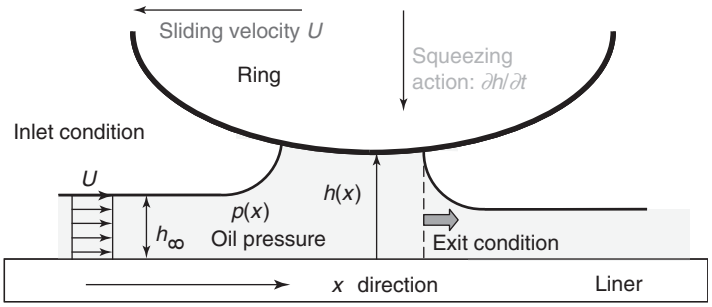


Figure 10. Fundamental ring-liner lubrication and friction model.

perpendicularly toward the liner surface as shown, plus the ring tension that tends to expand the ring against the liner. A hydrodynamic pressure is generated in the oil film that, from hydrodynamic theory, strongly depends on the sliding speed of the barrel-shaped wedge. A simple form of the Reynolds equation, with surface roughness and other features omitted, for the ring slider is shown in Equation 1, in reference to Figure 10, where x is the distance in the lubricant-flow direction, h the film thickness, U the sliding speed, and p the hydrodynamic pressure in the oil film, and μ the oil viscosity:

$$\frac{\partial}{\partial x}\left(\frac{h^3}{\mu}\frac{\partial p}{\partial x}\right) = 6U\frac{\partial h}{\partial x} + 12\frac{\partial h}{\partial t} \tag{1}$$

To account for surface characteristics such as surface roughness, textures, or waviness, flow factors (Patir and Cheng, 1979) that modify the first two terms in Equation 1 can be added. The force balance of the radial load against the oil pressure, together with pressure and mass continuity boundary conditions at the wetted edges of the ring, determine the minimum oil film thickness. When the film thickness becomes small enough where boundary or mixed lubrication may occur, an asperity contact model (Greenwood and Tripp, 1971) is commonly used to determine the boundary contact pressure, which will also take part in the radial force balance. Similar analysis is carried out for the second ring and for each of the rails of the oil-control ring. The film thickness on the liner left by the passage of one ring provides an inlet film thickness condition for the following ring or rail. The effects of any piston tilt or groove angle when the rings rests on the ring groove will be to change the relative orientation of the ring-face profile relative to the liner.

3.1.2.1.2 Friction behavior of individual rings. Obviously, the exact magnitudes of the film thickness of the rings and friction depend on the ring design parameters, surface characteristics, lubricant properties, and operating conditions. Figure 11 shows an example of predictions of ring-pack friction in an advanced reciprocating engine (Smedley, 2004) that illustrates some basic features of ring-pack friction:

(a) Two types of friction power loss peaks occur in this illustration—(i) friction peaks periodically at mid-stroke that correspond to periods of high sliding velocity, as shown by the oil-control ring friction, and (ii) peak friction power loss around the top ring reversal position near top center, shown by the top ring. Upon closer examination, the peaks correspond to a high level of solid–solid contact in boundary lubrication.
(b) The other general feature is that the magnitude of the second ring friction is significantly less than either that of the top ring or oil-control ring, primarily due to the subdued inter-ring pressure adding to the outward radial load on the ring against the liner.

3.1.2.2 Ring dynamics and gas flows in the ring pack. In addition to the radial forces of ring tension and gas pressure holding the rings against the liner, providing ring-liner seals, axial forces (gas pressure, inertia, and friction) also act on the rings, pressing the rings against the surfaces of the grooves, sealing the combustion gases from leaking around the rings in the grooves. The rings are carefully designed with a positive or negative twist angle (relative to the ring groove edges), as shown in Figure 12, to control the point of sealing and the pressure distribution around the ring. The axial forces and moments determine the ring's axial motion and its tilt in the ring groove. These axial forces include primarily the gas pressure forces acting on the flanks of the ring—intricately controlled by the designed twist (static twist)—balanced against the inertial force on the ring due to the reciprocating piston motion. The rings typically sit flat on the bottom groove flank about three-quarters of the time and on the top groove flank about a quarter of the time because of the higher

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

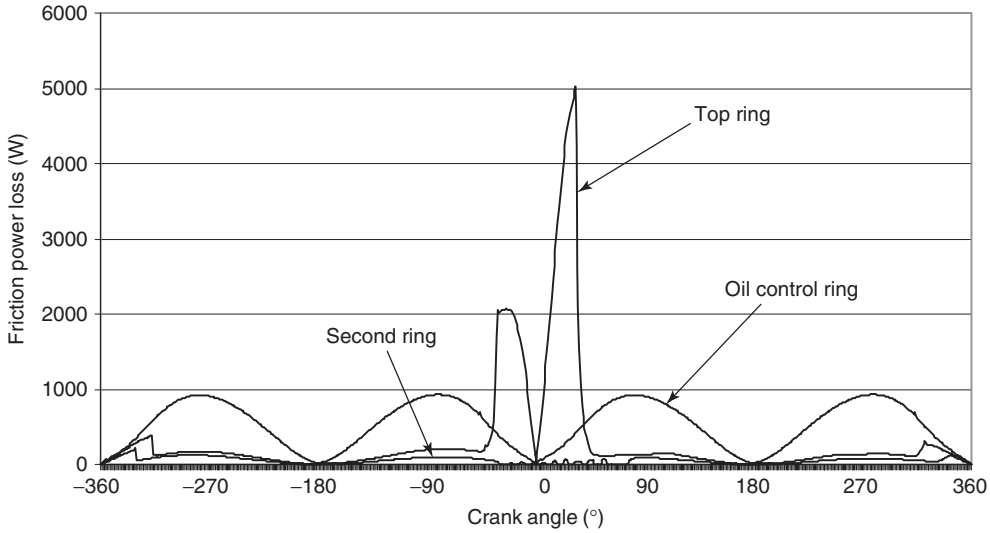


Figure 11. Friction power loss contributions in the piston ring pack for an 18-L natural gas engine at 1800 rev/min full load. (Reproduced from Smedley (2004). © Massachusetts Institute of Technology.)

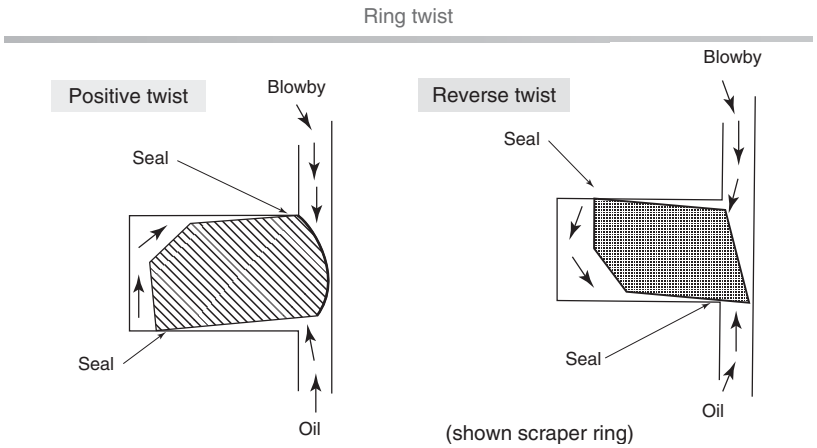


Figure 12. Schematic illustrating positive and negative static ring twists.

gas pressure on the combustion chamber side. There are two narrow time intervals, of a few crank angle degrees each, where a ring makes a transition from primarily one side of the groove toward the other. During ring transition, enhanced leakage of gases occurs. If the flow is toward the bottom, toward the crankcase, there is increased blow-by. Reverse flow can also occur because of the inter-ring pressure variations. This happens when the cylinder pressure decreases faster than the reduction of inter-ring pressures. Oil consumption could increase when reverse flow occurs, either due to flow around the grooves or through the ring gaps.

The stability and dynamics of the rings in the grooves affect the leakage of gases and oil around the rings in the grooves. Figure 13 (Smedley, 2004) shows a second ring with a negative twist (drooping downwards at the outer ring edge) and the rapid up-and-down motion of the ring in the groove (flutter) between −90 and 90 crank angle degrees that can occur. The second-ring instability or flutter occurs when the inertial force on the ring is high near TDC compared to the pressure force holding it down in the groove. When the ring is seated at the bottom of the groove with a negative twist, as shown in Figure 13, it is held down with a small pressure differential as the pressure above the ring penetrates around the ring to the bottom, as there is no flow. When the ring begins to lift, the gas flow and pressure differential forces the ring to seat at the bottom again. The ring flutter thus generates the additional

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

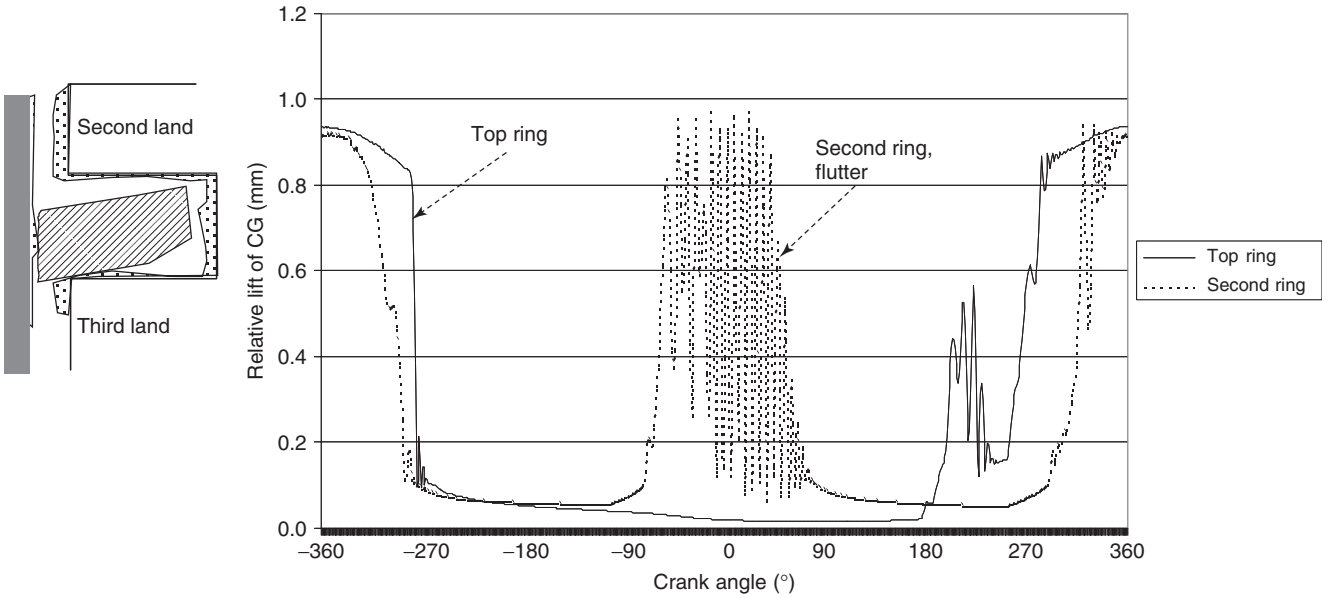


Figure 13. Second ring with negative static twist and ring flutter. (Reproduced from Smedley (2004). © Massachusetts Institute of Technology.)

flow path for gases and oil. The second ring instability incidentally relieves the second land gas pressure, which potentially could have caused reverse flows through the top ring gap (or top-ring groove when it flutters). The top-ring reverse flow contributes to oil consumption and emissions.

3.2 The crankshaft and connecting-rod bearing systems

The lubrication modes at the main bearings of the crankshaft, at the connecting-rod/crankshaft interface (big-end bearings), and at the interfaces between the piston pin and the piston pin bosses, and between the connecting rod and the piston pin are all journal-bearing lubrication. Hence, categorically they are discussed under this section.

3.2.1 The crankshaft main-bearing subsystem

3.2.1.1 Journal-bearing friction. Apart from its interfaces with the connecting rods, the crankshaft's friction comes primarily from the main bearings that support the crankshaft in its rotational motion. The bearing seals also generate some friction attributable to the crankshaft. The crankshaft rests on a layer of oil between the shaft and the outer bearing shell. The axis of the crankshaft is off center from that of the bearing center. This offset, called *bearing eccentricity*, generates the hydrodynamic pressure during shaft rotation. Oil is amply supplied to the bearing surfaces through oil feeds along the crankshaft. With adequate oil supply and under normal loads, the lubrication at the main bearings is primarily in the hydrodynamic regime. Journal-bearing calculations usually apply the Reynolds equation, in cylindrical coordinates, to the lubricant in the journal bearing in determining the oil pressure distribution, the locus of the journal relative to the bearing surface, and thus the minimum oil film thickness. The minimum oil film thickness is an important design parameter and is usually kept larger than the surface asperity heights to avoid mixed or boundary lubrication. The dynamic loading originates from the rapidly varying cylinder pressure pushing against the piston and the connecting rod, and then to the crankshaft.

The friction in the journal bearing is proportional to the surface shear stress integrated over the entire bearing surface area. A dimensional analysis indicates the following functional dependence holds (Heywood, 1988):

Average shear stress $\approx \mu(\pi D_b N/h_m)$

where

μ is the oil viscosity
D_b is the bearing diameter
N is the rotational speed
h_m is the mean radial clearance
bearing surface area $\approx \pi D_b L_b$
L_b is the bearing length

Accordingly, the bearing friction, $F_b \approx \mu(\pi^2 D^2{}_b L_b N/h_m)$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The mean oil film thickness, h_m, in the journal bearing is a function of the applied loading and other geometric factors of the journal bearing and lubricant viscosity. Thus, the friction scaling law for journal bearings often used is simply (Sandoval and Heywood, 2003; Patton, Nitschke, and Heywood, 1989):

$$F_b \alpha \mu D^2_b L_b N \quad (2)$$

The proportionality constants are often empirically determined and are specific for a certain bearing design. Thus, the connecting-rod/piston-pin bearing takes on a different proportionality constant from the con-rod big-end bearing, which is also different from that for the crankshaft.

3.2.1.2 Main-bearing seal and other friction. The bearing seal lips and the crankshaft surfaces are generally considered to be in constant solid-to-solid contact, with a constant friction coefficient, as in boundary lubrication, and a constant normal force, thus constant friction force. Obviously, the friction power loss from the seals is thus proportional to the rotational speed and the bearing diameter and the total contacting surface area. The proportionality constants depend on how tightly the seals are maintained and the surface characteristics of the surfaces. These constants, which vary from seal to seal, are determined empirically. Some researchers (Patton, Nitschke, and Heywood, 1989) consider another loss mechanism due to the power loss from pumping oil through the crankshaft oil feeds. However, strictly, this is not "rubbing" friction as discussed earlier in this chapter, but could actually be considered part of the power losses of the accessories.

3.2.2 The connecting-rod subsystem

3.2.2.1 The connecting-rod/piston-pin friction. The cylinder pressure force on the piston is transmitted to the crankshaft via the connecting rod, the top end of which connects to the piston via a piston pin and pin bosses that form part of the piston (see Piston and Ring Development). Figure 14 shows the piston, pin, and connecting rod system. There are actually two sets of interfaces: (i) the bearing between the piston-pin and the small end of the connecting rod, and (ii) the bearing between the piston pin and the pin bosses. However, as in the crankshaft bearings, lubrication in either case is characterized as a dynamically loaded journal-bearing system. For high cylinder pressures—a trend in producing more power for given engine displacement—the pin/boss, pin/connecting-rod bearing interfaces represent some of the most highly loaded areas in the engine. Lubrication is predominantly in the hydrodynamic regime, although analyses assuming the more general condition of mixed lubrication have been reported (Nishikawa, 2012; Wang, Du, and Zhang, 2011; Ligier and Ragot, 2005). Some measurements of pin friction (Takiguchi, Suhara, and Tsuneo, 1998) also suggested mixed lubrication. However, in those experiments, the pin showed significant bending, perhaps partially explaining the observed mixed or boundary lubrication. Assuming predominantly hydrodynamic lubrication, then the friction coefficient is roughly proportional to the term $\mu V/P$, where μ is the oil viscosity, V is the relative speed between surfaces which is proportional to engine RPM, and P is the load per unit area.

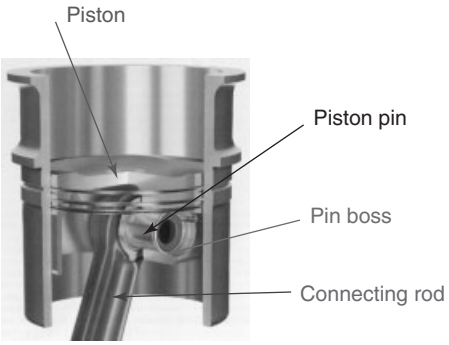


Figure 14. Schematic of piston, piston-pin, connecting rod, showing bearing interfaces.

3.2.2.2 The connecting-rod/big-end friction. The connecting-rod big end refers to the connection between the connecting rod and the crank. Lubrication here is also primarily in the hydrodynamic regime. Adequate oil is supplied to the bearing surfaces through feeds along the crankshaft. As in the case with crankshaft main bearing and piston-pin bearings, the bearing friction is proportional to the bearing surface area and mean linear velocity, which in turn is proportional to the bearing diameter at a given engine RPM.

For both the piston-pin and con-rod big-end bearing lubrication, the friction is proportional to the square of the bearing diameter, D_b, and the bearing length, L_b, (i.e., $D^2_b L_b$). Earlier estimates indicate that connecting-rod bearing friction is comparable to but somewhat less than piston-skirt friction (Richardson, 2000). In view of the increasing trend of higher cylinder pressure (bmep) engine operation for improved efficiency, the contribution to total friction from the connecting-rod bearings could become more significant, especially when and if asperity contacts in the mixed lubrication mode begin to be felt. However, the design of an adequate lubricant flow to the bearings will keep solid–solid contact to a minimum.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

An early analysis by Bishop (1965) derived an expression for the crankshaft and connecting-rod journal-bearing friction combined, where the additive terms from the various subsystems are apparent. The friction, normalized by the displacement volume (α bore$^2 \times$ stroke), in the form of friction mean effective pressure (fmep), in kilopascal, is given by (Lee *et al*., 1999; Bishop, 1965):

$$\text{Combined bearings fmep} = 41.4 \left(\frac{N}{1000}\right) \frac{(D^2_{\text{mb}}\, L_{\text{mb}} + D^2_{\text{rb}}\, L_{\text{rb}}/m + D^2_{\text{as}}\, L_{\text{as}})}{(B^2\, L)}$$

where N is the crankshaft rotational speed in rev/min, $B =$ bore, $L =$ stroke, $D_{\text{mb}} =$ the main bearing diameter, $L_{\text{mb}} =$ the total main bearing length divided by the number of cylinders, $D_{\text{rb}} =$ the connecting-rod bearing diameter, $L_{\text{rb}} =$ the total connecting bearing length, $m =$ the number of pistons per rod bearing, $D_{\text{as}} =$ the accessory shaft bearing diameter, $L_{\text{as}} = k$ the total length of all accessory shaft bearings divided by the number of cylinders.

3.3 Valvetrain system

The valvetrain system consists of a series of mechanical parts that serve primarily to open and close the intake and exhaust valves. The valvetrain converts the rotary motion of the camshaft, at one end, to oscillatory motion of the valves at the other end. The cam lobes on the camshaft determine the valve timings. There are several prevalent configurations of the component layouts (primarily of the rocker arm) between the camshaft and the valves themselves, as shown in Figure 15:

1. Direct-acting, overhead cam (OHC): cam lobe on tappet directly, no rocker arm
2. End-pivot rocker arm, OHC: cam lobe drives follower between pivot and valves
3. Center-pivot rocker arm:
   (a) OHC: cam lobe acts on end of rocker arm directly
   (b) OHC: cam lobe acts on end of rocker arm via lifter
   (c) Cam in block, overhead valve (OHV), cam lobe acts on rocker arm via extended pushrod

These configurations differ in simplicity, the number, size, and mass of the parts involved, and thus the stiffness of the system. The stiffness determines how fast the response of the oscillatory valve motion follows the actuating cam motion. They also differ in size/packaging and inertia, and thus their suitability varies depending on specific engine applications.

3.3.1 Tribological contacts and sources of friction

There are four main categories of contacts and sources of friction in the various configurations of valvetrains described. The lubrication modes range from predominantly hydrodynamic to boundary lubrication and mixed lubrication. The major contact and friction categories are as follows:

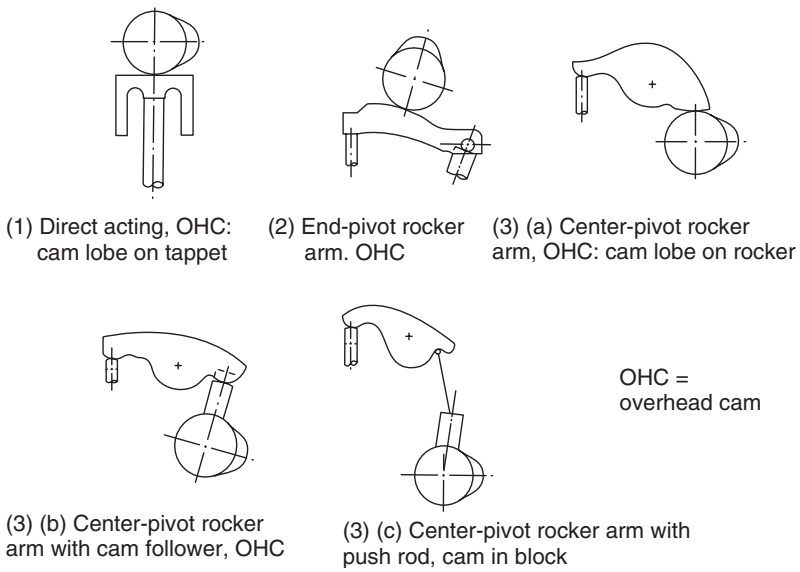


Figure 15. Major types of valve train configurations. (From Armstrong and Buuck (1981). © SAE International. Copyright © 1981 SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

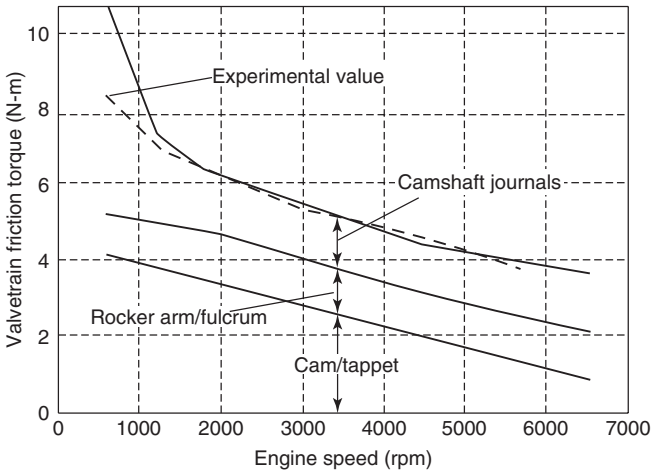


Figure 16. Effect of engine speed on valvetrain friction components for nonfriction-modified SAE 30 oil at 100°C. (From Wang (2007). © SAE International. Copyright © 2007 SAE International. Reprinted with permission.)

1. The camshaft bearing friction: the camshaft is supported by camshaft bearings (journals) similar to the crankshaft main bearings. The applied bearing load on the camshaft is significantly less than the load from the cylinder pressure through the connecting rod to the crankshaft. The journal-bearing lubrication at the camshaft bearings is mostly hydrodynamic. Estimates of the percentage contribution to total valvetrain friction from camshaft bearings varies from 12% or higher in earlier estimates (Staron and Willermet, 1983; Wang, 2007), as shown in Figure 16. Recent estimates (Comfort, 2003), however, suggest a smaller contribution from camshaft bearings. In reality, the relative contribution is a function of engine speed and it depends on the magnitude of the other components, specifically that of the cam followers.
2. The cam/follower interface friction: the cam–follower interface can be the cam lobe against a flat follower or a roller follower. In the flat follower configuration, the local load, film thickness, and friction vary with the relative position of the cam lobe to the follower. At the tip of the lobe, the local load is high and concentrated in a small area, and boundary lubrication is dominant. For the rest of the cam–follower contact, mixed lubrication prevails. The cam/follower interface is often modeled as a narrow elliptical or line contact from which the contact pressures are calculated. In the mixed lubrication regime, the viscosity of the lubricant depends on the pressure and elastohydrodynamic lubrication is assumed. Friction in the cam/flat-follower interface, consisting mostly of boundary-contact friction and some viscous drag, contributes to most of the valvetrain friction (Comfort, 2003). Roller followers, however, significantly reduce the cam/follower friction recently, by an order of 50% or better (Beloiu, 2010).
3. The rocker arm pivot/shaft friction: similar to the crankshaft seals, lubrication at the rocker arm pivot/shaft interfaces is mostly boundary lubrication. This is also due to the fact that there is very little lubricant supply to the surfaces. The boundary friction force at these interfaces is proportional to a constant friction coefficient and the applied contact load. Overall, friction at the rocker arm pivots can be as low as 10%, as Figure 16 shows at low speeds. Obviously, this percentage depends on the friction in the other components, which have come down as well in recent years, making the rocker arm pivot/shaft friction not negligible.
4. Friction in linearly oscillatory components: the components in the valvetrain in this category include the valve stem and seals, valves and guides, valve lifter and lifter bore—components that experience relative reciprocating or oscillatory motion. When the velocity in the oscillatory motion is small, we assume boundary lubrication. In general, the oscillatory motion, similar to the piston rings against the liner, also shows hydrodynamic lubrication at higher speeds during parts of the oscillatory motion. Both experiments and computations show that the percentage of valvetrain friction contributed from oscillatory motion is of the order of a few percent (Comfort, 2003).

Valvetrain friction has been studied in great detail computationally and by experiments (Gangopadhyay, Soltis, and Johnson, 2004). Typically, the reported contribution of valvetrain friction to overall engine mechanical losses is on the order of 15–20%, although estimates of valvetrain friction as high as 40% have been reported (de Paula Pignatti, Mizaiara, and da Cunha, 2011). Valvetrain friction is relatively more significant at lower speeds, indicating that most of the valvetrain friction comes from boundary and mixed lubrication.

3.3.2 Lubrication, lubricant composition, and supply to the contacts

Similar to the piston-motion reversal positions along the liner, the valvetrain cam/follower interface also exhibits significant boundary lubrication. The resulting metal–metal contact increases wear, through various wear mechanisms. Antiwear additives are formulated for lubricants to reduce valvetrain wear substantially. Friction modifiers are also

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

effective in reducing valvetrain friction. An adequate supply of fully formulated lubricant to the valvetrain is required. Engine lubricant is supplied to the camshaft through passages along the camshaft where, through oil relief holes, oil is airborne and made available to the cam/follower interfaces as well as other oscillating valvetrain components. The oil film thickness in these valvetrain interfaces depends on the oil entrainment rate into the contacting zones. Excess lubrication does not increase oil film thickness beyond the fully flooded level. However, inadequate oil supply, however minimal required, would reduce oil film thickness. The thinner oil films and reduced supply of antiwear agents in the lubricant could be detrimental to the contacts in both friction and wear. In view of the significant boundary and mixed lubrication in the valvetrain system, the proper design and selection of the chemistry and supply of oil to the valvetrain system is particularly important.

4 EFFECTS OF LUBRICANTS AND ADDITIVES

An engine lubricant is expected to perform several functions in the engine and to maintain satisfactory levels of performance for a considerable time period between oil change intervals. These functions are categorically: (i) to reduce friction and wear between rubbing surfaces; (ii) maintain cleanliness: minimize deposits, corrosion; contain contaminations in particulate (such as soot), liquid (such as fuel), or gaseous (such as acidic gases from products of combustion) forms; and (iii) maintain the oil properties, such as flow, vaporization, and other characteristics. The lubricant also acts as a sealant to minimize leakage of gases, and a coolant to remove heat from some components. To perform all these functions, the base lubricant, base oil, or "base stock," is blended with typically about 5–25% of additives, each of which performs one or more characteristic functions. The major constituents of the lubricant are discussed here, together with how the engine combustion process may contaminate the oil and the effects of the oil byproducts on the emission-control systems.

4.1 Base oils

A traditional mineral base oil is typically derived from the heavier hydrocarbons during the refining processes. A synthetic base oil is one which is synthesized from highly processed chemicals beyond those directly from the crude-oil refining stream. Some base oils being studied use exploratory fluids such as ionic—for example, water based—liquids (Fox and Priest, 2008), and some environmentally friendly lubricants use biodegradable base stocks (Schramm, 2004; Perez, Cheenkachorn, and Lloyd, 2002).

The most significant performance parameter of base oils is the viscosity. The oil viscosity characteristics include the sensitivity of changes in the viscosity to temperature, such as the viscosity index, VI. Lubricants tend to become thinner (lower viscosity) as temperature increases and become more viscous at low temperatures. American Society for Testing and Materials (ASTM) D2270 provides formulas for quantifying the VI given kinematic viscosities at 40 and 100°C. A high VI means a lubricant does not thin out much as it heats up nor becomes too thick at cold temperatures. These characteristics are important to ascertain that the oil film at the critical regions in the engine, such as at the hottest point of the top piston-ring (see Engine Thermal Management) and cylinder-liner contact region, does not become too thin at peak engine loads. The same oil also cannot be too thick to hamper its circulating freely around the engine during low temperature start-up operation. Another characteristic is the dependence of the oil viscosity on shear rate measured by the relative velocities and the film thickness between the parts. Specifications for the limits on the viscosities of the oil, including the high temperature high shear viscosities, at low temperatures and high temperatures, are given in the Society of Automotive Engineers (SAE) oil grade classification system (Haycock, Caines, and Hillier, 2004). Oils showing viscosity variations at high and low temperatures are multigrade oils. By carefully controlling the engine oil-film temperature via strategic thermal management techniques (such as by increased or decreased cooling of the liner), piston-liner friction can be affected (James, 2012); computations show friction improvements of 20–30% by increasing the temperature of the oil in the midsection of the liner (James, 2012).

For mineral oils, the major classes of heavy distillates deriving from the crude oil for the lubricant are paraffinic or napthenic hydrocarbons. Paraffinic oils show high VI, and napthenic oils show low VI—large viscosity variation with temperature. Depending on the relative composition of the base oil, the VI can vary. The American Petroleum Institute (API) designates different groups of base oils based on the level of saturates and sulfur in the oil and the VI. Groups I to III represents increasing level of saturates (either below or over 90%), decreasing sulfur (either greater than 0.03% or less than 0.03%), and increasing VI (between 80 and 120 or over 120). Group IV is for synthetic oils such as polyalphaolefins (PAO), and Group V is for all others, such as napthenics, polyalphaglycols, esters, and so on (Haycock, Caines, and Hillier, 2004).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

VI improvers can be added to the base oil to modify the VI of the formulated oil. Designing or using a proper oil for a given application or optimizing engine operation for given base oil characteristics, or both, provide opportunities for reducing engine friction and improve engine efficiency.

4.2 Additives

Additives are materials added to the base oil to improve the performance or properties of the oil. There could be 10–15 additives in the engine oil (Rudnick, 2009). These additives perform different functions, such as to affect friction and wear, or to maintain engine cleanliness, or to maintain the fluid properties, such as pour point or antifoam properties. Most of these additives are organic compounds and several involve metallic compounds that could adversely affect the emission aftertreatment system operation. The major additive types, especially those that affect friction and wear, are discussed in turn in two categories. The list is not exclusive as it does not include minor or special additive types.

4.2.1 Additives that directly affect friction and wear

This category of additives includes primarily the viscosity modifiers, friction modifiers, and antiwear additives.

4.2.1.1 Viscosity modifiers. Viscosity modifiers are additives, typically high molecular-weight polymers, added in small quantities to the base oil to reduce the temperature sensitivity of oil viscosity. As discussed in the section on base oils, high VI index is needed to ensure that the oil does not become too thin at high operating temperatures nor too thick for start-up and low temperature operations. Hence, viscosity modifiers have been called *VI* improvers in the past. Without viscosity modifiers, the viscosity of most mineral oil base stocks increase sharply with temperature. Generally, viscosity modifiers are added to increase the viscosity of the base oil, but more so at elevated temperatures, thus lowering the variation of viscosity with temperature. Typical viscosity modifiers are olefin copolymers (OCP), polymethacrylates (PMA) and styrene-butadiene copolymer. Some of these compounds also affect oil properties to some extent in other ways, such as the ability of the oil to flow freely at low temperatures and the ability to suspend particles in the oil as well (Rudnick, 2009).

4.2.1.2 Friction modifiers. Friction modifier additives in crankcase oils are designed to reduce friction in mixed or boundary lubrication, in contrast to transmission applications where increased friction is desired for additional traction. These friction-reducing additives are designed to form a slippery layer on the surfaces, and the layers have very low shear strength and thus low friction coefficient. Typical materials are long-chain molecules with polar heads that anchor to the metal surface. Common friction-reducing organic modifiers include oleylamide and glycerol monooleate (GMO).

4.2.1.3 Antiwear additives. The most common antiwear additive is zinc dialkyldithiophosphate (ZDDP) in its various forms, which also has antioxidant properties. The additive first decomposes thermally into reactive byproducts that form tribofilms or deposits on the surfaces. These surface films serve as protective layers during the wear process, which are continuously formed from the tribochemical reactions between the additive and the metal surface. During high pressure operating conditions, such as in gears, special sulfur and phosphorus compounds are used in place of ZDDP; these are called *extreme-pressure* (*EP*) antiwear additives.

4.2.2 Additives that maintain engine cleanliness and oil properties

This category of additives includes detergents, dispersants, antioxidants, pour point depressants, antifoam agents. Brief descriptions follow.

Detergents are designed to control engine deposits and to neutralize acids generated from the products of combustion that have migrated into the oil. Detergents are usually made up of metal salts of organic acids—the sulfonates, phenates, or carboxylates—of which sulfonates of calcium or magnesium are very common. They are characterized by two essential parameters: the ash level and the basicity. ASTM procedures have been defined to quantify the sulfated (i.e., treated with sulfuric acid and burned) ash level, as well as to determine the total base number (TBN). The TBN is a measure of the detergent's ability to neutralize acids in the lubricant.

Dispersants are designed to keep particles from separating from the oil and forming sludge and deposits that may block lubricant passageways. These particles in the oil could come from combustion products such as soot from diesels, wear debris, condensed water, or other particulate matter. Dispersants have polar heads that attach to the particles and keep them from agglomerating and thus dispersed as fine particles in suspension in the lubricant. Many totally organic (ash free) formulations of dispersants are available in the market, such as succinimides and OCP, which also show viscosity-modifying properties.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In maintaining or regulating the lubricant properties, antioxidants are designed to protect the oil itself from oxidation and thereby degradation. ZDDP, discussed earlier, also has excellent antioxidation properties. Various forms of ZDDPs can also be combined to enhance the overall thermal stability of the lubricant. Besides ZDDP, various amine and phenol compounds are available as supplemental antioxidant additives. Other additives are also available to extend the low temperature at which the oil remains fluid—pour point depressants—or to minimize the formation of foam, especially in the engine crankcase with significant churning and foaming of the lubricant.

Additives are essential to lubricant performance in engine lubrication. The oil will not function the way we expect it to function in the absence of additives. The various additives are carefully formulated to ensure that antagonistic interactions among the additives are minimized.

4.3 Effects of engine operation on lubricant properties

Besides the warm-up period during which the oil viscosity gradually increases with increasing temperature, the oil composition as well as oil properties vary during engine operation in time and location in the engine because of various other processes. These processes include vaporization of volatiles, local fuel dilution, oxidation, and contamination. The friction and wear at critical components depend on the local lubricant conditions at the points of contact under actual operating conditions. The processes that affect the local oil conditions include the following:

(a) Increased vaporization of the light hydrocarbon fractions of oil near the top piston ring zone, for instance, resulting in oil having different rheology at the top-ring zone compared to that referenced at the sump. Additive concentrations are also expected to vary. Figure 17 (Watson, 2010) shows "enrichment factors" of metallic compounds of oil sampled from the piston-ring groove. The enrichment factor is defined as the ratio of the concentration of a given elemental mass to its concentration in the sump. The metallic concentrations in the piston ring zone are much higher than those in the sump.

(b) During combustion, acidic gases such as oxides of nitrogen and sulfur are generated and absorbed in the lubricant, gradually decreasing the basicity of the oil, as evidenced by typical decreases in the TBN over time.

(c) Fuel and particulate matter contaminate the oil, especially from direct injection diesel (or gasoline) engines where some level of fuel impingement on the cylinder liner walls may occur. Some fuel may get into the oil via entrainment or diffusion, resulting in fuel dilution of the oil. This phenomenon also occurs in diesel engines equipped with diesel particulate filters (DPFs) that employ late fuel injection as a means to raise the exhaust temperature. Repeated fuel impingement on the cylinder walls will significantly increase the fuel fraction in the crankcase oil. It has been generally understood that soot from diesel combustion continuously increases the soot content in the oil over time.

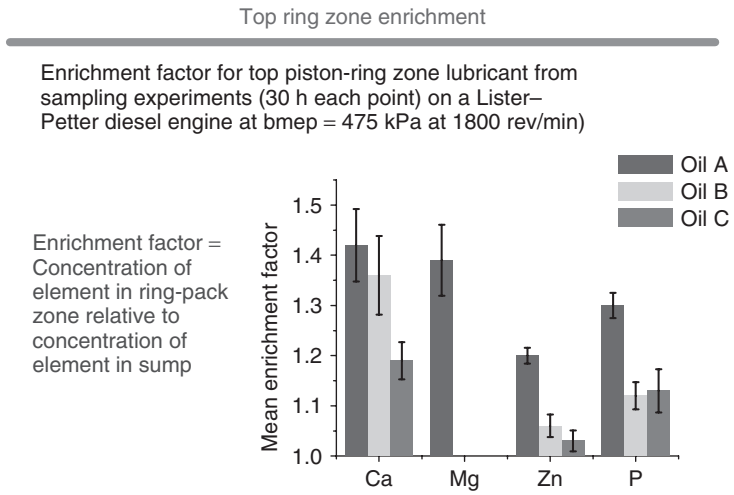


Figure 17. The average enrichment factors of metallic elements in top piston-ring zone oil samples. (From Watson, Huang, and Wong (2007). Copyright © 2007 SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

(d) Blowby gases traveling from the combustion chamber to the crankcase carry unburned fuel, and products of combustion in gaseous or particulate form. These heavily polluted gases contaminate oil in the oil sump. The contaminants are, in turn, carried to the valvetrain system in the lubricant circulation circuit.

(e) There is significant oxidation and degradation of oil on the upper portions of the cylinder liner and in the top-ring areas. High shear operation at the piston-ring/liner interface, as well as in the bearings, also contributes to lubricant degradation. As a result, lubricant properties, friction and wear in the lubricated parts, and the overall engine fuel economy change over time between oil changes.

(f) Foam and aeration: oil bubbles form in the oil churning action in the sump or cavitation elsewhere in the lubricant system. Actual lubrication performance is affected by this phenomenon. Antifoam additives alleviate the problem somewhat, but not entirely.

4.4 Major effects of additives on component friction and wear

The effects of additives on engine components depend on the lubrication regimes at the prevailing conditions at the local contacts. The lubrication regimes at the various components in most warmed-up conditions are: hydrodynamic (for bearings); mostly boundary (for valvetrain cam-follower), and mixed for the piston/ring-liner interface except around the mid-stroke of the piston travel, where significant hydrodynamic lubrication is expected in most cases. There are variations among the different rings and the piston-skirt surfaces, however. Accordingly, the effectiveness of the different additives—viscosity modifiers versus friction modifiers—varies at the different components and operating conditions.

Lubricant formulation affects friction primarily via (i) viscosity control—base oil selection and viscosity (VI) modifiers, which can change the shear and temperature dependency of the viscosity, and (ii) (boundary) friction modifier additives, which affect the boundary friction by forming surface layers with low shear strength. Some organic polymers both affect the viscosity and form surface layers, for example, Polyalkylmethacrylates, PAMAs (Muller, Fan, and Spikes, 2007; Hedrich *et al*., 2003). The performance of these additives varies greatly with temperature and in the presence of other additives.

4.5 Lubricant/additives effects on engine emission-control system

While lubricants and additives perform vital functions in an engine, the lubricant-derived emissions have serious impact on the exhaust aftertreatment system (Sappok, Munnis, and Wong, 2012). For gasoline engines, the three-way catalyst (TWC) has been around for close to 40 years, where it has been shown that significant levels of phosphorus from engine oils could deteriorate the TWC prematurely. Since 2007, world-wide diesel regulations have been in place at mandated particulate emission levels that essentially need to be met with DPFs. Since 2011, use of NO*x* emission aftertreatment devices—mostly SCRs (selective catalytic reductions) for heavy-duty diesels and LNTs (lean NO*x* traps) for light duty diesel engines—has also become widespread in the United States, Europe, and Japan.

Lubricant-derived compounds in the exhaust that affect the emission-control aftertreatment system include the following:

- Incombustible ash from lubricant additives
- Sulfur and phosphorus compounds

Ash is problematic because it can build up inside the channels of DPFs. Unlike soot, ash cannot be oxidized into gaseous species. In as little as 56,000 kilometers, there is more ash accumulated in a DPF between regeneration intervals (for active regenerations) than soot (Sappok, Kamp, and Wong, 2012). The ratio of ash to soot in the DPF is even higher for continuously regenerated DPFs. Figure 18 (Aravelli *et al*., 2006) shows a DPF with channels clogged with incombustible ash matter.

Over the past several years, lubricant specifications have been in place to limit the sulfur, phosphorus, and ash levels in lubricants, as well as volatility limits in the CJ-4 oil category (McGeehan *et al*., 2006). The API, the European Automobile Manufacturers Association (ACEA), and the

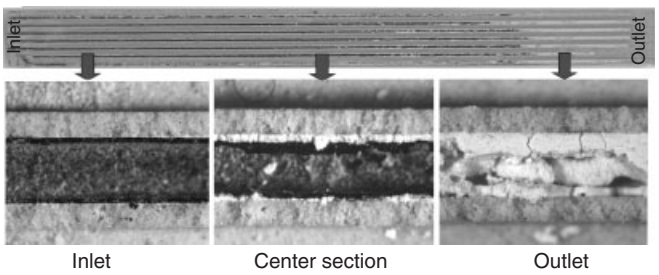


Figure 18. Ash distribution in a channel of a diesel particulate filter. (From Aravelli *et al*. (2006). Copyright © 2006 SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

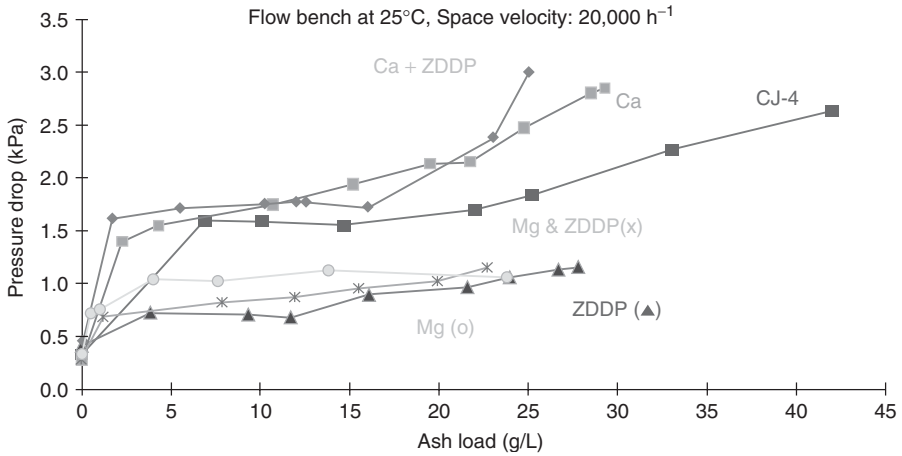


Figure 19. Sensitivity of pressure drop across diesel particulate filter for several lubricant formulations. (Reproduced with permission from Sappok, Munnis, and Wong (2012). © ASME.)

Table 1. International "low ash" engine oil specifications.

Specification	Sulfated ash (%)	Sulfur (%)	Phosphorus (%)
API CJ-4	1.0	0.4	0.12
ACEA E6	1.0	0.3	0.08
JASO DH-2	1.0	0.5	0.12

Japanese Automotive Standards Organization (JASO) have all introduced new "low ash" heavy-duty diesel engine oil specifications (Table 1).

Significant studies have been conducted to characterize the ash compounds in the DPF. It has been shown that the engine back pressure doubles in about 290,000 kilometers of normal operation (Bodek and Wong, 2007), and that the type of lubricant additive seems to have a difference in how DPFs are affected. Figure 19 shows very clearly that for the same mass quantity of ash in the DPF, verified gravimetrically, ash from calcium compound shows the highest flow restriction, while zinc compounds produce the least effect (Sappok, Munnis, and Wong, 2012).

Lubricant-derived sulfur compounds affect LNT performance, as SO_2 does compete with NO*x* for storage sites in the LNT system. However, desulfation cycles that will drive the occupation of catalyst sites by SO_2 can be designed. However, the repeated high temperature desulfation cycles could compromise the DPF substrate integrity in the long term. It is not clear that phosphorus chemically interferes with conversion efficiencies of NO*x* reduction systems. However, phosphorus affects the catalytic operation of diesel oxidation catalysts (DOCs), which are important in the conversion of NO to NO_2—a step that is important in both NO*x* reduction and soot oxidation.

Tremendous efforts are continuing to understand the characteristics of the lubricant-derived compounds in the emission aftertreatment systems, so that optimum formulations of lubricants and additives that meet the simultaneous requirements of emission control and adequate engine protection can be further developed. Moreover, these requirements are expected to continue to evolve as engine and aftertreatment systems advance to meet increasing fuel economy and tighter emissions standards in the coming decades.

ACKNOWLEDGMENTS

The author acknowledges and thanks the support of the many sponsors of a multitude of engine lubrication and lubricant-related programs at MIT, in which the author served as PI or co-PI in the past 25 years. The author is also grateful to the wonderful work of many graduate students and colleagues at MIT over the years that contributed to the understanding of the subject matter in this chapter.

REFERENCES

Aravelli, K., Jamison, J., Robbins, K., *et al*. (2006) Improved lifetime pressure drop management for DuraTrap RC filters with asymmetric cell technology (ACT). DEER 2006 Conference, August, http://www1.eere.energy.gov/vehiclesandfuels/pdfs/deer_2006/session5/2006_deer_aravelli.pdf (accessed 4 September).

Armstrong, W.B. and Buuck, B.A. (1981) Valve gear energy consumption: effect of design and operation parameters, SAE Paper 810787, Society of Automotive Engineers, Warrendale, PA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Beloiu, D.M. (2010) Modeling and analysis of valve train, part I—conventional systems. SAE Paper # 2010-01-1198. Society of Automotive Engineers, Warrendale, PA: *SAE International Journal of Engines*, **3**(1), 850–877.

Bishop, I.N. (1965) Effect of design variables on friction and economy. *SAE Transactions*, **73**, 334–358.

Bodek, K.M. and Wong, V.W. (2007) The effects of sulfated ash phosphorus and sulfur on diesel aftertreatment systems—a review. SAE Paper # 2007-01-1922. JSAE # 20077200, Society of Automotive Engineers, Warrendale, PA.

Comfort, A. (2003) An introduction to heavy-duty diesel engine frictional losses and lubricant properties affecting fuel economy-part I. SAE Paper # 2003-01-3225, Society of Automotive Engineers, Warrendale, PA.

Fox, M. and Priest, M. (2008) Tribological properties of ionic liquids as lubricants and additives: part 1: synergistic tribofilm formation between ionic liquids and tricresyl phosphates. *Journal of Engineering Tribology*, **222**(13), pp. 291–303.

Gangopadhyay, A., Soltis, E., and Johnson, M.D. (2004) Valvetrain friction and wear: influence of surface engineering and lubricants. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, **218**, 147–156.

Greenwood, J.A. and Tripp, J. (1971) The contact of two nominally flat surfaces. *Proceedings of the Institution of Mechanical Engineers*, **185**, 625–633.

Haycock, R.F., Caines, A.J., and Hiller, J. (2004) *Automotive Lubricants Reference Book*, 2nd edn, Society of Automotive Engineers, Warrendale, PA.

Hedrich, K., Scherer, M., Herzog, S.N., *et al*. (2003) The influence of dispersant PAMA on soot handling, wear and fuel economy in heavy-duty diesel oils. SAE Paper No. 2003-01-1959, Society of Automotive Engineers, Warrendale, PA.

Heywood, J. (1988) *Internal Combustion Engine Fundamentals*, McGraw-Hill, New York, NY.

James, C.J. (2012) Analysis of parasitic losses in heavy duty diesel engines. M.S. Thesis. Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, June.

Koeberlein, D. (2012) Supertruck technologies for 55% thermal efficiency and 68% freight efficiency. DEER 2012 Conference, October 2006, http://www1.eere.energy.gov/vehiclesandfuels/pdfs/deer_2012/tuesday/presentations/deer12_koeberlein.pdf (accessed 4 September).

Lee, S., Shannon, B. A., Mikulec, A., *et al*. (1999) Applications of friction algorithms for rapid engine concept assessments. SAE Paper # 1999-01-0558, Society of Automotive Engineers, Warrendale, PA.

Ligier, J. and Ragot, P. (2005) Piston pin: wear and rotating motion. SAE Paper # 2005-01-1651, Society of Automotive Engineers, Warrendale, PA.

Mansouri, S.H. and Wong, V.W. (2004) Effects of piston design parameters on piston secondary motion and skirt-liner friction. SAE Paper # 2004-01-2911, Society of Automotive Engineers, Warrendale, PA.

McGeehan, J.A., Moritz, J., Shank, G., *et al*. (2006) API CJ-4: diesel oil category for both legacy engines and low emission engines using diesel particulate filter. SAE Paper # 2006-01-3439, Society of Automotive Engineers, Warrendale, PA.

Moughon, L. (2006) Effects of piston design and lubricant selection on reciprocating engine friction. M.S. Thesis. *Department of Mechanical Engineering, Massachusetts Institute of Technology*, Cambridge, MA, June.

Moughan, L. and Wong, V. (2005) Effects of Lubricant and Piston Design on Reciprocating Engine Friction. ASME Paper 2005–1343, *Proceedings of ICEF2005*, ASME Internal Combustion Engine Division Fall Technical Conference, September 11–14, Ottawa, Canada.

Muller, M., Fan, J., and Spikes, H. (2007) Influence of polymethacrylate viscosity index improvers on friction and wear of lubricant formulations. SAE Paper No. 2007-01-198, Society of Automotive Engineers, Warrendale, PA.

Nakada, M. (1995) Piston and Piston Ring Tribology and Fuel Economy. *Proceedings of the International Tribology Conference*, Yokohama.

Nishikawa, C. (2012) Optimization of semi-floating piston pin boss formed by using oil-film simulations. SAE Paper # 2012-01-0908, Society of Automotive Engineers, Warrendale, PA.

Patir, N. and Cheng, H.S. (1979) Application of average flow model to lubrication between rough sliding surfaces. *ASME Journal of Lubrication Technology*, **101**, 220–230.

Patton, K.J., Nitschke, R.G., and Heywood, J.B. (1989) Development and evaluation of a friction model for spark-ignition engines. SAE Paper # 890836, Society of Automotive Engineers, Warrendale, PA.

de Paula Pignatti, T., Miziara, W., and Esteves da Cunha, R. (2011) Reduced friction for a four cylinder two valve Otto engine valve train. SAE Paper # 2011-36-0208E, Society of Automotive Engineers, Warrendale, PA.

Perez, J.M., Cheenkachorn, K., and Lloyd, W.A. (2002) A comparison of some biodegradable hydraulic fluids and engine oils. SAE Paper # 2002-01-1498, Society of Automotive Engineers, Warrendale, PA.

Quillen, K., Stanglmaier, R.H., Moughon, L., *et al*. (2006) Friction Reduction by Piston Ring Pack Modifications of a Lean-Burn 4-Stroke Natural Gas Engine: Experimental Results. ASME Paper ICES2006-1327, *Proceedings of ASME ICED Spring Technical Conference*, May 8–10, Aachen, Germany.

Quillen, K., Stanglmaier, R.H., Wong, V., *et al*. (2007) Friction reduction due to lubrication oil changes in a lean-burn 4-stroke natural gas engine: experimental results. ASME Paper JRCICE2007-40128, Joint Rail Conference & Internal Combustion Engine Division Spring Technical Conference, March 13–14, Pueblo, Colorado.

Richardson, D. (2000) Review of power cylinder friction for diesel engines. *Journal of Engineering for Gas Turbine and Power, Transactions of the ASME*, **122**(4), 506–519.

Rudnick, L.R. (2009) *Lubricant Additives—Chemistry and Applications*, 2nd edn, CRC Press, Boca Raton, FL.

Sandoval, D. and Heywood, J. (2003) An improved friction model for spark-ignition engines. SAE Paper # 2003-01-0725, Society of Automotive Engineers, Warrendale, PA.

Sappok, A., Kamp, C., and Wong, V. (2012) Sensitivity analysis of ash packing and distribution in diesel particulate filters to transient changes in exhaust conditions. SAE Paper #2012-01-1093, Society of Automotive Engineers, Warrendale, PA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Sappok, A., Munnis, S., and Wong, V.W. (2012) Individual and synergistic effects of lubricant additive components on diesel particulate filter ash accumulation and performance. ASME Paper# ICES2012-81237. *ASME 2012 Internal Combustion Engine Division Spring Technical Conference*, Turin, Italy, May 6–9.

Schramm, J. (2004) Application of a biodegradable lubricant in two flexible fuel vehicles. SAE Paper # 2004-01-2988, Society of Automotive Engineers, Warrendale, PA.

Smedley, G. (2004) Piston ring design for reduced friction in modern internal combustion engines. M.S. Thesis. *Department of Mechanical Engineering, Massachusetts Institute of Technology*, Cambridge, MA, June.

Staron, J.T. and Willermet, P.A. (1983) An analysis of valve train friction in terms of lubrication principles. SAE Paper # 830165, Society of Automotive Engineers, Warrendale, PA.

Takiguchi, M., Suhara, T., and Tsuneo, S. (1998) Reduction of friction for piston pin boss bearing of automotive gasoline engine by utilizing oil around the boss. ASME Paper # 98-ICE-101, American Society of Mechanical Engineers, New York, NY.

Taraza, D., Henein, N.A., Ceausu, R., *et al*. (2007) Engine friction model for transient operation of turbocharged, common rail diesel engines. SAE Paper # 2007-01-1460, Society of Automotive Engineers, Warrendale, PA.

Taylor, C.M. (1993) *Engine Tribology*, Elsevier Science Publishers, Amsterdam, The Netherlands.

Wang, Y. (2007) *Introduction to Engine Valvetrains*, Society of Automotive Engineers, Warrendale, PA.

Wang, X., Du, J., and Zhang, J. (2011) Mixed lubrication analysis of piston pin bearing in diesel engine with high power density. ASME Paper # IJTC2011-61185. *ASME/STLE 2011 International Joint Tribology Conference*, Los Angeles, CA.

Watson, S.A.G. (2010) Lubricant-derived ash—in-engine sources and opportunities for reduction. Doctoral Thesis. *Department of Mechanical Engineering, Massachusetts Institute of Technology*, June.

Watson, S.A.G., Huang, W., and Wong, V.W. (2007) Correlations among ash-related oil species in the power cylinder, crankcase, and the exhaust stream of a heavy-duty diesel engine. SAE Paper 2007-01-1965, Society of Automotive Engineers, Warrendale, PA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto129
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Gas Aftertreatment Systems

James E. Parks II
Oak Ridge National Laboratory, Knoxville, TN, USA

1 Introduction 1
2 Pollutant Concerns 1
3 Catalyst-Based Aftertreatment Solutions 2
4 Emission Compliance Testing (Driving Cycles) 6
5 Degradation Issues and Deterioration Factors 6
6 Common Catalysts for Vehicle Applications 7
7 Ongoing Challenges 15
8 Conclusion 16
Endnotes 16
References 17

1 INTRODUCTION

The internal combustion engine (ICE) serves a vast array of transportation needs. Modern ICEs are finely controlled to achieve optimal fuel combustion inside the cylinder; however, gaseous pollutants are emitted and need to be controlled to protect the environment and public health. This chapter describes the control of gaseous emissions from engines in the exhaust system with catalyst technologies.

2 POLLUTANT CONCERNS

During the combustion of fuel inside the engine cylinder, vaporized fuel is mixed with oxygen in air and oxidized to produce power. Although the process is highly controlled, combustion is not perfect and does not complete entirely; thus, some unburned or partially burned fuel products are emitted into the exhaust system as hydrocarbons (HCs) and carbon monoxide (CO) (see UHC and CO Formation and Models for more detail). Furthermore, during the high pressure and high temperature conditions of combustion, nitrogen (N_2) in the intake air reacts with oxygen in the intake air to form nitrogen oxide (NO) and nitrogen dioxide (NO_2) emissions that together are referred to as NO_x (*oxides of nitrogen*) (see NO_x Formation and Models for more detail). Together, HCs, CO, and NO_x are the primary "criteria" gaseous pollutants and the focus of emission control exhaust systems. Particulate matter pollutants are emitted by the engine, too, and in some cases, must be controlled in the exhaust (see Solid/Condensed Phase Aftertreatment Systems for more on the control of particulate matter emissions).

Emission regulations exist to insure that minimal amounts of HCs, CO, and NO_x pollutants are emitted into the atmosphere. Release of HCs, CO, and NO_x pollutants into the environment may cause various health and environmental effects that depend on the pollutant type, pollutant level, and often the atmospheric conditions where the pollution occurs. In the United States, the Clean Air Act, enacted in 1963 and 1970 with major amendments in 1977 and 1990, establishes the authority for federal regulations to protect the environment and, thereby, human health (http://epa.gov/oar/caa/caa_history.html). Current U.S. EPA (United States Environmental Protection Agency) emission regulations are known as *Tier 2 Exhaust Emission Standards* (http://www.epa.gov/otaq/standards/light-duty/tier2stds.htm.), and a new set of emission standards with lower required emissions from vehicles was proposed in 2013 and is known as "Tier 3".

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2.1 Hydrocarbon pollution concerns

HC emissions consist of a variety of HC-based chemicals that are derived either directly or indirectly from fuel or oil in the engine. The health and environmental concerns are specific to the HC chemistry. Direct exposure to HCs can affect human health with health effects ranging from irritants for respiratory systems to carcinogens, which can cause cancer (http://www.atsdr.cdc.gov/phs/phs.asp?id=422&tid=75). Indirectly, HCs can affect human health by interacting with NO_x in the atmosphere to form ozone (Seinfeld and Pandis, 1997a), which affects respiratory systems and causes smog. In addition, smog limits the visibility in the environment and can have negative impacts on the economy especially for tourism-based companies.

U.S. EPA Tier 2 emission regulations refer to HCs as non-methane organic gases (NMOG) and include a wide range of HCs except for methane under this category. Specific species of HCs are more carefully monitored because of their potential harmful effects. The U.S. EPA defines 21 pollutant species as mobile source air toxics (MSATs) because of their potentially harmful nature (http://www.epa.gov/oms/toxics.htm). Formaldehyde is one of the MSATs and receives specific attention in emission regulations. U.S. EPA Tier 2 regulations require minimum levels of formaldehyde emissions to be met. Although formaldehyde is not generally present in fuel, it can be formed in the engine as an intermediate combustion product and thus, formaldehyde emissions can result in the exhaust.

2.2 CO pollution concerns

CO is a poisonous gas that directly impacts human health. CO exposure to humans can cause headaches, nausea, and other effects at low exposure levels. At higher levels, CO exposure can lead to death by limiting oxygen uptake (http://www.osha.gov/SLTC/healthguidelines/carbonmonoxide/recognition.html). While CO pollution dilutes readily in the atmosphere during normal driving conditions, its potential danger to human health requires careful control. Eventually, CO will react with hydroxyl radicals in the atmosphere to form CO_2, but the reaction leads to the formation of the hydroperoxyl radical HO_2, which reacts with NO to form ozone precursors (Seinfeld and Pandis, 1997b).

2.3 NO_x pollution concerns

The primary concern related to emissions of NO_x is the formation of ozone in the atmosphere. NO_x together with energy in the form of sunlight are directly involved in the chemical reactions to form ozone (Seinfeld and Pandis, 1997c), and ground-level ozone can have harmful effects on human respiratory systems. NO_x can also affect particulate formation in the atmosphere by forming nitric acid and reacting with ammonia to form ammonium nitrate particles (Seinfeld and Pandis, 1997d) (Figures 1, 2).

3 CATALYST-BASED AFTERTREATMENT SOLUTIONS

Careful and precise control of engine operation can minimize the amount of pollutants emitted by the engine, but

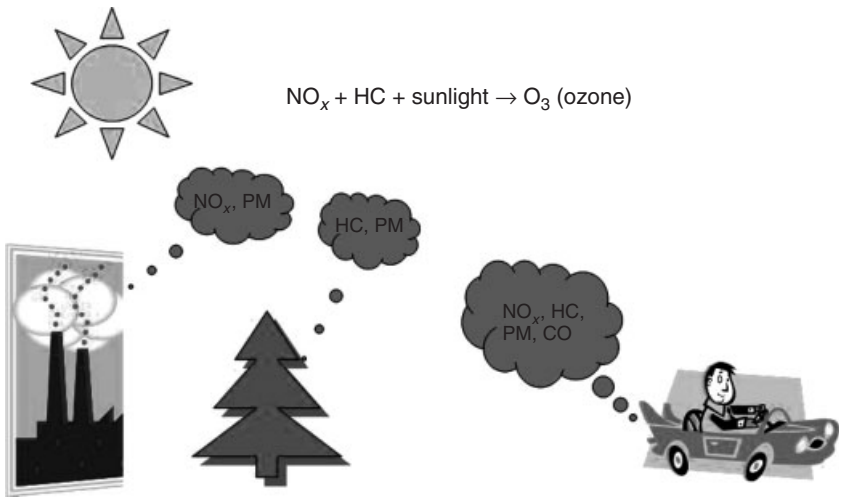


Figure 1. Pollutant emissions interact with energy in the form of sunlight in the atmosphere to form ground-level ozone.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 2. Image of New York City clouded by smog. (Reproduced from CBS New York, 2011. © Mario Tama/Getty Images.)

catalyst-based aftertreatment is required to obtain near-zero levels of pollution emissions at the tailpipe of the vehicle. The term *aftertreatment* is used to note that the emission control is occurring postcombustion or downstream of the engine. Primarily, pollutant reduction with aftertreatment occurs by passing the hot engine exhaust over a catalyst or series of catalysts. The catalyst increases the rate at which the chemical reactions occur that consume pollutant species.

The catalyst, also commonly called a *catalytic convertor*, is made of a variety of materials that can both *function* and *survive* under the exhaust conditions over many years and hundreds of thousands of miles of operation. In general, the catalyst is composed of a substrate that is coated with a mixture of metal oxides and metals. The purpose of the substrate is to hold the active coating materials in the exhaust flow and maximize the interactions between the gases and the catalyst materials in the coating. The coated substrate is mounted in a metal can with a vibration damping thermally insulating fibrous material between the hard substrate and can to adsorb shock and provide room for the differential expansion of the metal can and ceramic part (Figure 3).

Substrates are typically made of the ceramic material cordierite, which has a low expansion coefficient that

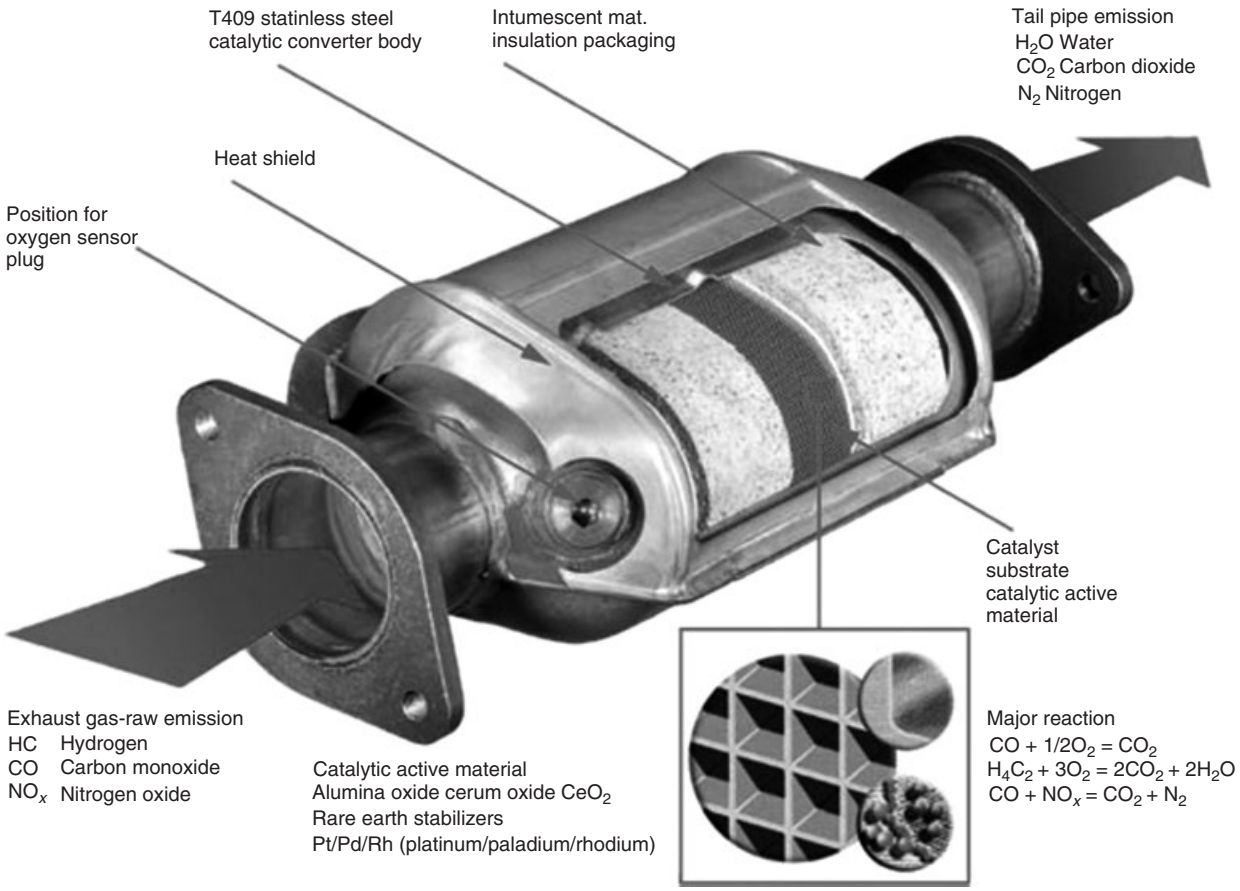


Figure 3. Diagram of a typical catalytic convertor. (Reproduced from UK Catalysts, n.d. © UKCatalysts.com.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

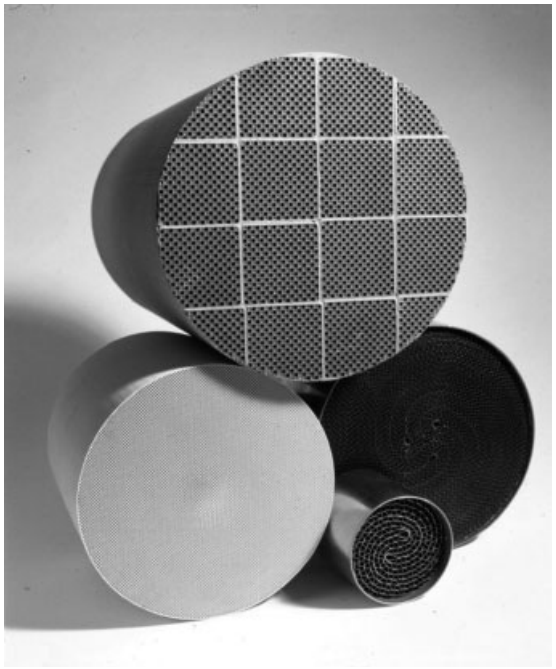


Figure 4. Various substrates used to support catalysts. The bottom three substrates have flow-through channels, and the top substrate is a particulate filter geometry that requires gases to flow through the porous walls of the device. (Reproduced from Johnson Matthey, n.d(a). Photograph reproduced with permission from Johnson Matthey.)

enables durability against thermal expansion and shock. In some cases, substrates are made of metal foil, which is oxidized to form a surface that the catalyst coating can adhere to. In both cases, the substrate, alternatively called a *monolith*, is composed of thousands of small channels that the exhaust gas flows through. The walls of the channels are coated with the catalyst material, so that the exhaust gases interact with the catalyst as they flow along the length or "flow axis" of the catalyst. Ceramic substrates are manufactured with an extrusion process that forms an array of channels. Metal substrates are generally manufactured by rolling corrugated metal foil into a circular or elliptical structure that contains thousands of channels for exhaust flow. While more channels improve the surface area available for the catalyst coating, care must be taken to avoid too many channels and thick coatings, which restrict exhaust flow and increase backpressure on the engine leading to some loss in engine efficiency (Figure 4).

While the substrate forms the structural foundation, the coating contains the actual catalytic materials that accelerate chemical reactions; often, the coating is called the "*washcoat*".[1] In general, coatings are composed of a mixture of metal oxides and active components. The primary role of the metal oxides is to form a high surface area support to disperse the active components on the catalyst surface for maximum reactivity with the exhaust gases. Surface areas for the support metal oxides commonly exceed $100\,m^2/g$, and a coated substrate with these metal oxides can provide over an acre of surface area ($>4000\,m^2$) on a catalyst that can fit in one's hand. This large surface effectively enables the chemical reactions to occur between the exhaust gases and the catalysts when the exhaust is flowing at high rates over the catalyst in the engine exhaust system. In the manufacturing process, the coating materials are ball milled in an aqueous suspension that resembles paint in viscosity and color. The applied coating is dried and calcined (heated to a high temperature) to form a strong bond to the substrate. The milling process controls the size of the particles that make up the porous coating. High porosity improves gas diffusion into the coating layer to maximize the interaction between the exhaust and the catalyst surface.

The active components of the catalyst are coated onto the high surface area metal oxide support. Different materials form the active components for different catalysts. Common active components include precious noble metals such as the platinum group metals (PGMs) Pt, Pd, and Rh. These metals are unique in their ability to lower the activation energy of many chemical processes and thereby accelerate the kinetic rate of chemical reactions. However, while excellent catalytic materials, the PGMs, are also very expensive. In order to cost effectively utilize them in commercial products, PGMs such as Pt are finely dispersed over the high surface area metal oxide support to the point where Pt exists as tiny nanometer-sized balls on the coating surface. The high dispersion of Pt results in a high surface area of Pt per total mass of Pt applied and thereby maximizes the performance as a function of cost. Thus, the importance of the high surface area metal oxide support not only provides overall surface area to enable exhaust gas reactions to the surface materials but also provides a high surface area for effective dispersion of the active components across the surface (Figure 5).

During the development of catalysts, researchers utilize specialized tools to measure the surface area of both the support metal oxide and active components such as PGMs. A technique named "BET" after the work of three scientists Brunauer, Emmett, and Teller (1938) is utilized to measure the total surface area of catalyst coatings by determining the amount of N_2 gas that adsorbs on the surface. A similar technique measures the chemisorption of H_2 onto PGMs such as Pt to specifically measure the active component surface area. These measurements are routinely made by catalyst researchers and form the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

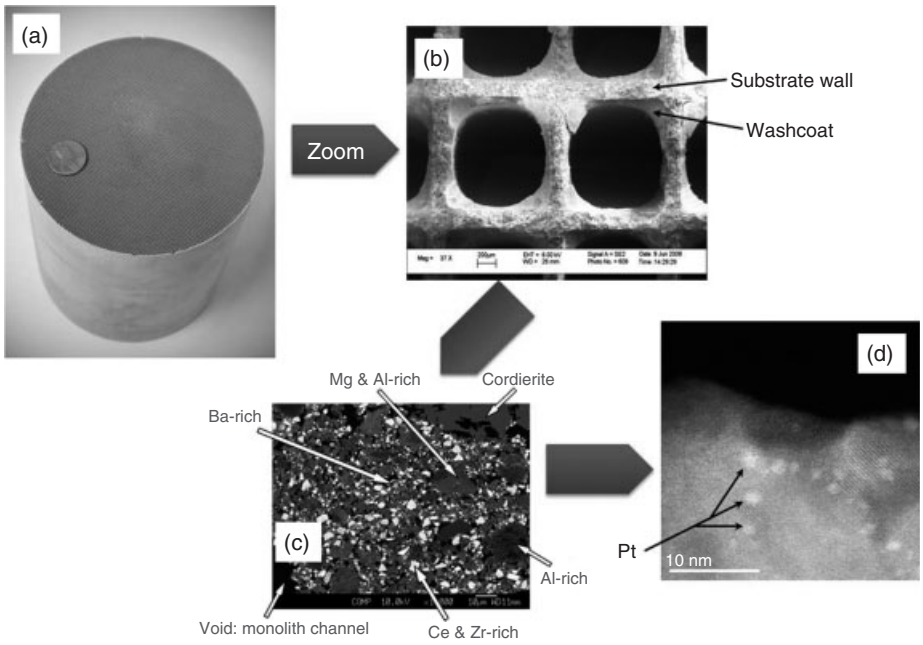


Figure 5. Various scales of catalysts. (a) A full-size catalyst contains hundreds of channels, called *cells*, per square inch of the catalyst face for gas to flow through. (b) Each cell contains an inner coating of catalyst materials collectively known as the *washcoat*. (c) The composition of the washcoat can be quite complex with 1–10 micron size particles of various metal oxides that serve different functions for the catalyst. (d) Ultimately, gases flowing through the catalyst interact with finely dispersed nanometer scale particles of platinum group metals and other active components to complete the chemical reactions that process the pollutant gases.

basis for practical development of catalyst technologies that depend on nanoscale materials.

In many catalysts, exhaust gases briefly chemically adsorb onto the active component where they react with other chemisorbed gases and then release into the gas stream. However, in some catalysts, the active component may also form a stronger chemical bond and effectively store an exhaust gas for further control of the chemical processes. For instance, in three-way catalysts, metal oxides such as ceria are used to store oxygen, which better enables the three-way catalyst to reduce NO_x pollutants. Thus, catalyst coatings tend to be a complex mix of various materials that often provide multiple aspects of functionality for treating pollutant species in the exhaust gas.

The critical physical and chemical reactions occur on the nanometer scale on the catalyst surface. Efficient interchange of gas species to the active catalyst surface sites is essential to enable the catalysis to occur. Limitations to the gas–surface interactions are due to kinetic and mass transfer issues. If the catalysis reaction rate is kinetically limited (or slow), the catalyst site will remain ineffective for other reactions, as the unreacted chemisorbed gas species block access to the surface site for other gas species. As temperatures increase, kinetic rates of reaction generally increase too. Mass transfer limitations occur when the gas species of interest cannot get to the catalyst surface site at a rate rapid enough to fully utilize the surface site. Diffusion into coating pores is one mass transfer limiting process. Once the gas does interact with catalyst surface site at the suitable temperature, the activation energy of the desired chemical reaction is lowered and the reaction can occur. The rate at which the catalyst sites can enable the reactions of the chemisorbed species and receive more gas species for more reactions is called the *turnover frequency* and is one parameter of interest in studying catalysis.

While in action on the vehicle, the catalyst is exposed to a wide range of conditions that challenge meeting the requirements for high emission control efficiency and durability. Cold vehicle starts in winter climates cause rapid extreme catalyst temperature changes. Transient driving conditions cause highly variable exhaust gas chemical conditions. In addition, vehicles driven over rough roads in all types of climates create challenging mechanical vibrations and stress conditions. All of these conditions need to be addressed through the design and control of the catalyst system while maintaining durability and longevity of the catalyst, so that emissions can be controlled over the useful life of the vehicle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4 EMISSION COMPLIANCE TESTING (DRIVING CYCLES)

While a great deal of effort goes into the development of a catalyst, the complete vehicle is subjected to emission regulations and thus must pass emission compliance tests known as *driving cycles*. Driving cycle tests are conducted on a chassis dynamometer, which is similar to a treadmill for vehicles. During the drive cycle, the speed and load of the engine are varied by fixing the vehicle speed to a standard transient speed profile that mimics real world driving. An example U.S. EPA drive cycle is the federal test procedure (FTP). The FTP lasts 1874 s with vehicle speeds ranging from 0 to 57 mph. The test represents 11.04 miles of driving and includes a cold start (http://www.epa.gov/nvfel/testing/dynamometer.htm). Emissions over the course of the test are diluted with air and collected in bags, which are analyzed after the test for pollutants. The resulting emissions measured are expressed in mass (grams) per mile and compared to the emission standard (Figure 6).

A key challenge for virtually all catalyst technologies is the cold start portion of the drive cycle test. Catalysts are essentially inactive below 100°C. During cold-start conditions, the catalyst warms up but first must desorb condensed H_2O from the surface before becoming effective. Above 100°C, the catalyst surface is free from adsorbed H_2O but the catalyst will still not be very effective until the "light-off" temperature is reached. The *light-off* temperature is a term used to describe the point at which the kinetics of the catalytic reaction reach a level at which significant catalytic reactivity is present. It varies with catalyst technology and pollutant species. For instance, for a particular catalyst, CO may begin to oxidize at 150°C, but HCs may not start oxidizing to 220°C. As very little pollutant reduction occurs before light-off, emissions during the cold start portion of the drive cycle test are major contributions to the overall integrated emissions measured over the entire drive cycle. Thus, engine control strategies are designed to rapidly warm the catalyst, and catalysts are designed for improved low temperature performance (Figure 7).

5 DEGRADATION ISSUES AND DETERIORATION FACTORS

In addition to meeting the emission regulation level by demonstrating low emissions over the prescribed drive cycle test, catalysts must prove durability over the "full useful life" of the vehicle. The definition of useful life

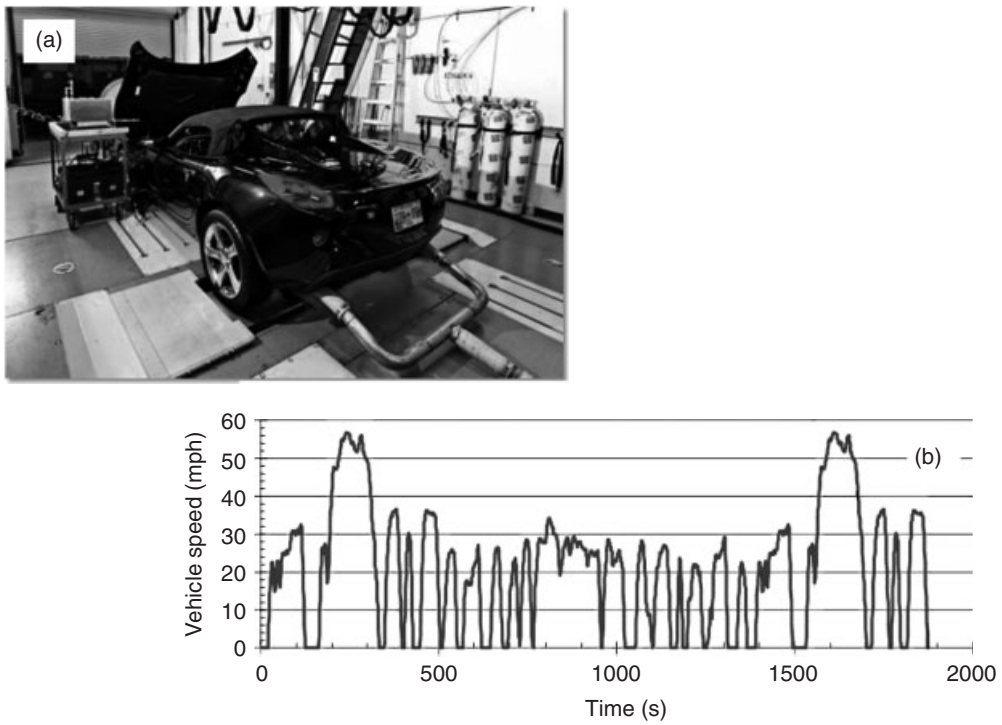


Figure 6. Emission compliance for passenger cars involves driving the vehicle on a chassis dynamometer (a) over a set driving cycle with transient speeds (b) and measuring total vehicle emissions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

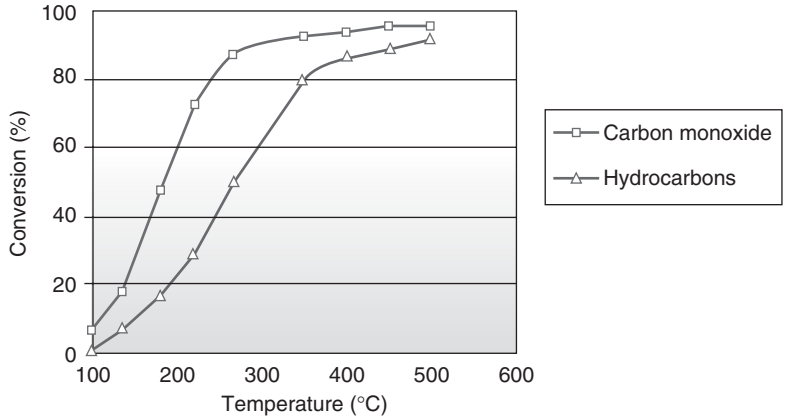


Figure 7. Example data showing the "light-off" in catalyst activity of CO and hydrocarbon conversion as temperature increases. At temperatures greater than the light-off temperature, excellent conversion is obtained. (From Nett Technologies Inc, n.d. Reproduced by permission of Nett Technologies Inc.)

is part of the regulation; U.S. Tier 2 emission regulations define full useful life as 120 k miles and also include a standard for 50 k miles (http://www.epa.gov/otaq/standards/light-duty/tier2stds.htm). In order to demonstrate catalyst durability to meet the useful life requirement, automotive companies together with catalyst suppliers perform extensive aging tests of catalysts on vehicles and on test rigs. The testing enables the measurement of the catalyst "deterioration factor," which expresses the loss in emission reduction efficiency per vehicle miles traveled (VMT). In effect, new catalysts must reduce emissions to a level significantly below the emissions standard, and the deterioration factor must be low enough that the rate of increase in emissions per VMT does not result in emissions higher than the standard at the end of the useful life period.

Many degradation processes can contribute to catalyst performance loss. The catalyst operates in a harsh environment with rapidly changing temperature and flow conditions. Furthermore, mechanical vibrations and shock experienced over the course of driving can potentially cause loss of the active catalyst coating. Studies of catalyst performance durability have identified several degradation mechanisms that include poisoning, fouling, thermal degradation, vapor compound formation, vapor–solid reactions, and attrition/crushing (Bartholomew, 2001).

Poisoning occurs when specific chemicals in the exhaust permanently degrade catalytic activity by covering the catalyst surface. An example poison is phosphorous, which is an additive in oil. Over time, some oil is consumed and combusted by the engine, which results in the release of phosphorous and other additive chemicals to the exhaust. The poison chemically bonds to the catalyst surface and prevents other gas species from interacting with the catalyst, thereby reducing catalyst effectiveness. Fouling is a similar process but is characterized by physical blockage of catalyst pores, which leads to similar damaging effects.

Thermal degradation affects almost all catalysts because of the harsh temperature exposure in automotive applications. When extreme high temperatures occur, the catalyst metal oxide structure can collapse and the coating can be severely compromised. However, careful thermal management generally prevents such catastrophic failure, and instead, the more common thermal degradation effect that occurs is sintering or the loss of dispersion of the active catalyst components (Figure 8). For example, the ideal catalyst has a well dispersed array of precious metal spheres distributed over the metal oxide surface to maximize the interactions with the exhaust gas; however, at high temperatures, these precious metal spheres can migrate and coalesce to form larger spheres, which have less total surface area and, therefore, lower overall catalytic performance. This effect is known as *Ostwald ripening* after the German scientist Wilhelm Ostwald (http://en.wikipedia.org/wiki/Ostwald_ripening#cite_note-1). Overall, careful control of the catalyst exposure conditions and design of durable catalysts must be combined to meet the durability required by regulations.

6 COMMON CATALYSTS FOR VEHICLE APPLICATIONS

The most common individual catalyst technologies for vehicle emission control are presented here with follow on sections for degradation issues and ongoing catalyst research and development challenges.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

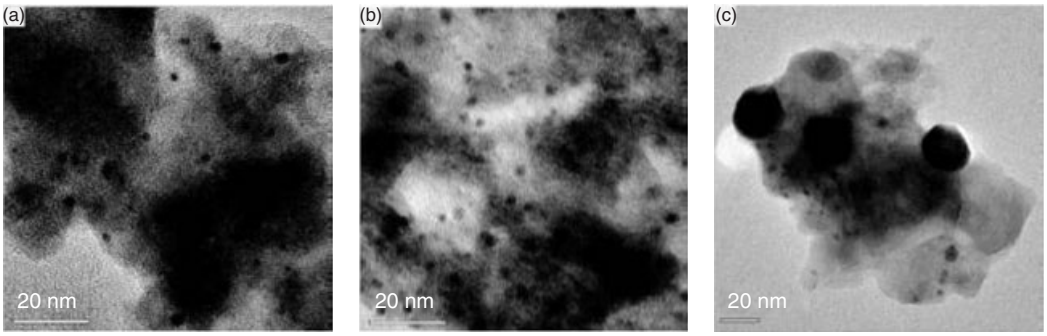


Figure 8. Electron micrographs showing the occurrence of sintering. With operation at high temperatures, the finely dispersed 1–2 nm Pt particles shown by the small dark specks in (a) agglomerate together to form larger particles (b) and eventually, particles >10 nm (c), which have significantly less surface area and, thereby, less catalytic activity. (With kind permission from Springer Science+Business Media: *Catalysis Letters*, Coarsening of Pt particles in a model NO_x trap, 93, 2004, pp. 129–134, G. W. Graham, H.-W. Jen, W. Chun, H. P. Sun, X. Q. Pan and R. W. McCabe.)

6.1 Three-way catalysts

Three-way catalysts are the most common catalysts on automotive vehicles. The name *three-way catalyst* comes from the fact that a single catalyst takes care of three pollutants (NO_x, CO, and HCs) simultaneously. The active components in three-way catalysts are PGMs. The PGM sites oxidize CO and HCs and reduce NO_x to N_2. Equation groups 1 through 4 show the main chemical reactions that occur.

[CO oxidation]

$$2CO + O_2 \rightarrow 2CO_2 \quad (1)$$

[HC oxidation, for example, HC propene]

$$2C_3H_6 + 9O_2 \rightarrow 6CO_2 + 6H_2O \quad (2)$$

[NO reduction]

$$2NO + 2CO \rightarrow N_2 + 2CO_2$$
$$10NO + 4HC \rightarrow 5N_2 + 2H_2O + 4CO_2$$
$$2NO + 2H_2 \rightarrow N_2 + 2H_2O \quad (3)$$

[NO_2 reduction]

$$2NO_2 + 4CO \rightarrow N_2 + 4CO_2$$
$$10NO_2 + 8HC \rightarrow 5N_2 + 4H_2O + 8CO_2$$
$$2NO_2 + 4H_2 \rightarrow N_2 + 4H_2O \quad (4)$$

Accurate control of engine air-to-fuel ratio (AFR) is critical to enable the TWC (three-way catalytic *converter*) to function. If the engine is operated with a fuel lean AFR, excess oxygen will be in the exhaust causing CO and HCs to be oxidized, but preventing NO_x from being reduced to N_2. Conversely, if the engine is operated with a fuel-rich AFR, NO_x is reduced to N_2, but CO and HCs are not oxidized. Only control of AFR to near stoichiometric conditions will allow simultaneous CO and HC oxidations with NO_x reduction (Figure 9).

In modern gasoline-fueled vehicles, the AFR is carefully controlled with feedback from exhaust gas oxygen (EGO) sensors in the exhaust system. Fuel injection into the engine can be adjusted based on the EGO sensor reading to create stoichiometric combustion. Typically, two EGO sensors are located in the exhaust system—one upstream and one downstream of the TWC. To further improve the efficiency of the TWC, the AFR of the engine is modulated in a narrow band around the stoichiometric combustion point. The modulating condition serves two purposes. Firstly, oxygen storage components of the TWC will adsorb and desorb O_2 from/to the exhaust, which effectively adds more accurate control to the stoichiometric condition on the TWC catalyst surface; this leads to lower NO_x, CO, and HC emissions. Secondly, by comparing the upstream and downstream EGO sensor signals, changes in the modulation of AFR due to the TWC can be monitored, and if damage occurs to the TWC, EGO signal analysis can determine whether the TWC has been damaged. This second function related to EGO sensing of air-to-fuel modulation is an important element of on-board diagnostics (OBD) for insuring continued emission control as the vehicle ages.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

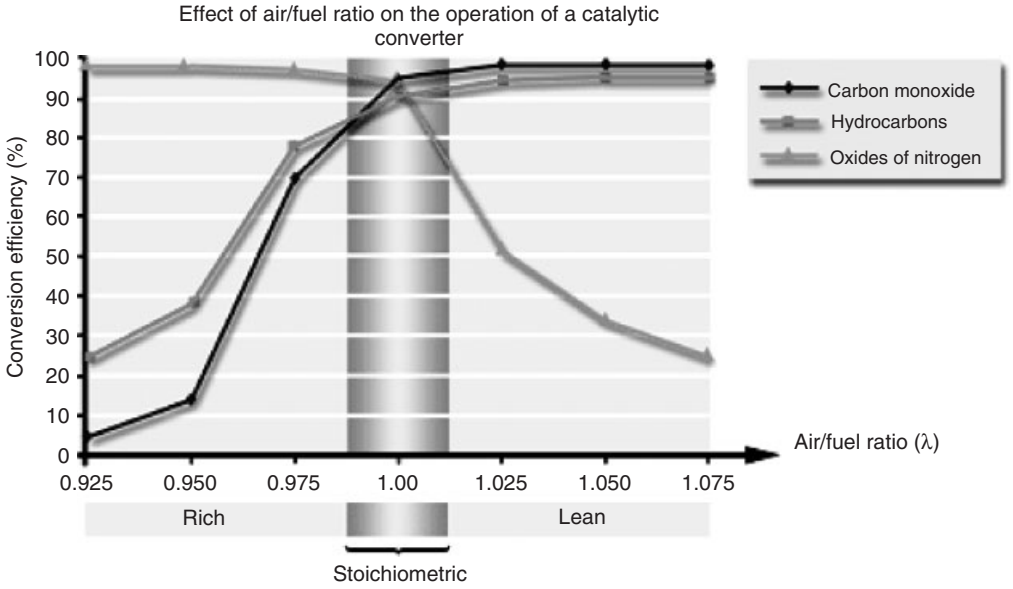


Figure 9. The three-way catalyst is named for its effective treatment of the three pollutants CO, NO_x, and hydrocarbons; however, proper air-to-fuel ratio control to stoichiometric combustion conditions must occur for >90% conversion of all the three species. (From Blackthorn, UK. Reproduced by permission of Blackthorn.)

For spark-ignited gasoline engine vehicles, the TWC is typically the only catalyst required to meet emission regulations. Often, two TWCs will be used in the exhaust system: one very close to the engine in the "close-coupled" position and the other further downstream in the "underfloor" position of the vehicle. The temperature of the close-coupled catalyst will increase more rapidly during cold start of the vehicle and enable the vehicle to have lower emissions during engine warm-up. More space is available on the vehicle for the underfloor catalyst, and lower overall temperatures in the underfloor position help enable better durability of the underfloor catalyst.

6.2 Diesel oxidation catalysts

Diesel oxidation catalysts (DOCs), as their name implies, control pollutant emissions from diesel engine vehicles. The primary pollutants that DOCs control are CO and HCs; the DOC oxidizes CO and HCs in the oxygen-rich exhaust that is produced by the fuel lean operation of diesel engines. In addition, some control of particulate matter is achieved by DOCs. As the diesel engine's main mode of operation is fuel lean, NO_x emissions reduction is not achieved by the DOC. Instead, additional catalysts are required to reduce NO_x from the engine exhaust. However, the DOC does play a critical role in supporting the NO_x reduction in downstream catalysts by oxidizing NO to NO_2; control of the NO to NO_2 ratio is important in optimizing NO_x reduction performance over downstream catalysts.

In addition to controlling CO and HCs, the DOC also serves other functions on diesel engine systems. Often, the DOC is utilized to heat the exhaust for thermal management of diesel particulate filters (DPFs), which require periodic heating to oxidize stored particulate matter. Such exhaust heating is achieved by adding extra fuel to the exhaust upstream of the DOC and converting the fuel to heat via exothermic oxidation reactions of the fuel over the DOC.

DOCs primarily consist of PGMs that are dispersed over high surface area metal oxide coatings. In some cases, the metal oxide support will be composed of oxygen storage materials. The primary goals of DOC design are to enable low emissions with minimal cost and to enable low temperature functionality. There is an extra challenge, as diesel engines have inherently lower temperature exhaust as compared with stoichiometric gasoline engines. Other than thermal management, the DOC requires little control by the engine system to function well for CO and HC emission control.

6.3 Selective catalytic reduction

A challenging emission control objective for automotive catalysts is reducing NO_x emissions from fuel lean engines. Engines that operate with lean fuel and air mixtures

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

inherently intake more oxygen than can be consumed in the combustion process; thus, the exhaust from lean engines contains significant (percentage) levels of oxygen. For NO_x emission control, the goal is to reduce the NO_x to N_2, but achieving NO_x *reduction* in the *oxidative* oxygen-rich exhaust atmosphere is chemically unlikely. Thus, special catalyst approaches are required to reduce NO_x in oxygen-rich exhaust. One such approach is selective catalytic reduction (SCR). In SCR, the catalyst utilizes a reductant gas to reduce the NO_x. Two SCR technologies based on different reductant chemistry are presented here.

6.3.1 Ammonia SCR

SCR can effectively be performed with ammonia (NH_3) as the reductant; this is the most common form of SCR-based NO_x control. During NH_3 SCR, NH_3 is stored on the SCR catalyst and interacts with NO_x on the catalyst surface to reduce NO_x to N_2 via the chemical reactions shown in Equations 5 through 7. NH_3 is an excellent reductant for NO_x and enables NO_x reduction in oxygen-rich exhaust over a broad range of exhaust temperatures. Typically, the NH_3 required is equal to the NO_x rate on a mole basis to achieve maximum NO_x reduction efficiency; the ratio of NH_3 to NO_x is known as *ANR* (*ammonia to NO_x ratio*) or "α" with $\alpha = 1$ being optimal. Equation 7 shows the "fast" SCR reaction that occurs with equal NO and NO_2 and has higher kinetic rates than the "standard" reaction (Equation 5) that only involves NO as a reactant.

["standard" SCR reaction]

$$4NO + 4NH_3 + O_2 \rightarrow 4N_2 + 6H_2O \quad (5)$$

["NO_2" SCR reaction]

$$8NH_3 + 6NO_2 \rightarrow 7N_2 + 12H_2O \quad (6)$$

["fast" SCR reaction]

$$NO + NO_2 + 2NH_3 \rightarrow 2N_2 + 3H_2O \quad (7)$$

NH_3 can be stored on a vehicle in various ways for use in the emission control system. Gaseous tanks are one option for NH_3 storage but not commonly utilized because of potential dangers of ammonia leakage and challenges with storing large amounts of NH_3 in pressurized tanks. A newer technology utilizes solid-state materials for onboard NH_3 storage (Johannessen *et al.*, 2008). The material releases NH_3 when heated to control dosing of the NH_3 into the exhaust system. $SrCl_2$ is one of the solid-state materials utilized for the NH_3 storage. The solid-state NH_3 approach has safety advantages because of the containment of NH_3 even under rupture of the solid-state container.

The most common form of NH_3 SCR utilizes urea as the NH_3 storage medium. Urea is represented by the chemical formula $CO(NH_2)_2$. It is a solid material but readily dissolves in water. Thus, in practice for SCR applications, the term *urea* generally refers to an aqueous solution of urea. The industry standard concentration of urea in the solution is 32.5% by mass,[2,3] and the solution is commonly called *diesel exhaust fluid* or *DEF* (http://www.discoverdef.com/).

The urea solution is stored onboard the vehicle in a liquid form with some heating capability in the tank and lines to prevent freezing, which occurs at $-11^{\circ}C$. The urea solution is injected directly into the exhaust upstream of the SCR catalyst. Once in the exhaust, the urea solution evaporates and decomposes to NH_3 via the hydrolysis reactions shown in Equations 8 and 9. Of the two reactions, Equation 8 occurs more rapidly after the injected urea evaporates, and Equation 9 occurs slower and commonly occurs in the SCR catalyst (van Helden *et al.*, 2004). Careful engineering of the injection system and exhaust flow upstream of the SCR catalyst is required for the appropriate mixing of urea into the exhaust, so that the hydrolysis reaction can occur without urea deposits forming along the exhaust pipe.

$$CO(NH_2)_2 \rightarrow NH_3 + HNCO \quad (8)$$

$$HNCO + H_2O \rightarrow NH_3 + CO_2 \quad (9)$$

Once hydrolysis is complete, NH_3 is in the exhaust system and can be utilized by the SCR catalyst. The SCR catalyst is an excellent storage media for NH_3, and the storage functionality greatly improves performance under the transient conditions experienced during vehicle driving. By maintaining a steady supply of NH_3 on the catalyst via storage, the SCR catalyst can reduce NO_x emissions that vary rapidly with time without having to match injection of urea exactly to the rapid NO_x emission variations. The urea–SCR system containing urea storage, urea injection capability, and the SCR catalyst has been successfully commercialized on diesel engine passenger cars as well as heavy-duty diesel trucks. Often a DOC and DPF are used in conjunction with the SCR for control of CO, HC, and particulate emissions. An upstream DOC can also improve SCR performance by oxidizing some NO to NO_2 to achieve overall higher NO_x reduction efficiencies via the "fast" SCR reaction shown in Equation 7 (Figure 10).

SCR catalyst formulations have evolved over time as material advancements have occurred. Early SCR catalysts based on stationary power NO_x control were composed of vanadia (V_2O_5) on a titania (TiO_2) support; these catalysts were often modified with tungsten oxide (WO_3) or molybdenum oxide (MoO_3) and other materials for improved

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

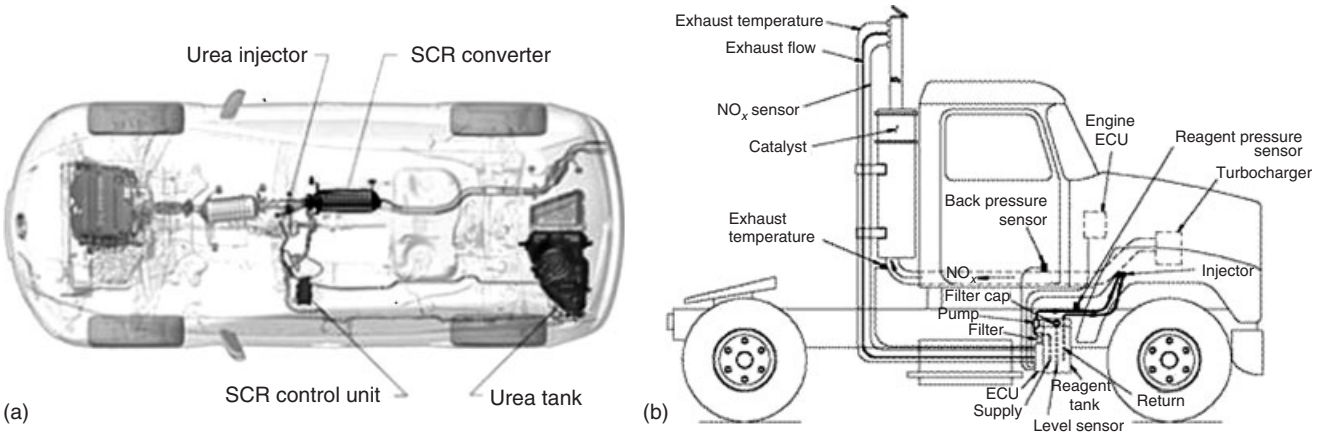


Figure 10. Diagrams of a passenger car (Reproduced from Japanese Sport Cars, 2009. © Japanesesportscars.com) (a) and a heavy-duty truck (b) that utilize SCR to control NO_x (Reproduced with permission from Krishnan, R and Tarabulski, 2005. Source: Dieselnet.com © Ecopoint Inc., 2005). In both cases, an on-board tank supplies urea, which is injected into the exhaust and converted to NH_3 upstream of the SCR catalyst where NH_3 reduces NO_x to N_2 in the lean oxygen-rich exhaust.

performance (Bartholomew and Farrauto, 2006a). The V-based site acts as the catalytic reduction of NO_x with NH_3. Modern SCR catalysts are based on zeolite materials. Zeolites are a class of aluminoslicate materials that have unique pores in their structure on the scale of the size of gas molecules (Bartholomew and Farrauto, 2006b). They are also known as *molecular sieves* because of their ability to filter and/or trap various gas molecules of different sizes based on the zeolite pore size, which can be as small as 0.3 nm. Owing to these unique characteristics, they can adsorb large quantities of NH_3 for the SCR reaction. Metals are ion-exchanged with zeolite materials to add catalytic function. The most common metals added are Fe and Cu; thus, SCR catalysts are relatively lower in cost compared with catalysts that rely on PGMs. Cu-zeolite SCR catalysts generally have better low temperature NO_x reduction performance when compared with Fe-zeolite SCR catalysts (Figure 11). Further improvement has also occurred with the zeolite materials themselves. Original zeolites have been replaced with newer chabazite zeolite materials that have different structural features that significantly improve their thermal durability and limit the adsorption of HCs on the SCR catalyst. The adsorption of HCs can be detrimental to SCR performance by fouling the catalyst and preventing the SCR NO_x reduction reaction from occurring.

Although it is important to maintain a supply of NH_3 on the SCR catalyst during operation, the stored level of NH_3 is constrained by the control strategy to prevent excess NH_3 from being emitted downstream of the SCR catalyst. While not regulated as a mobile source, NH_3 emissions are not desired and can lead to NO_x formation in the atmosphere; moreover, NH_3 can react with sulfates in the atmosphere

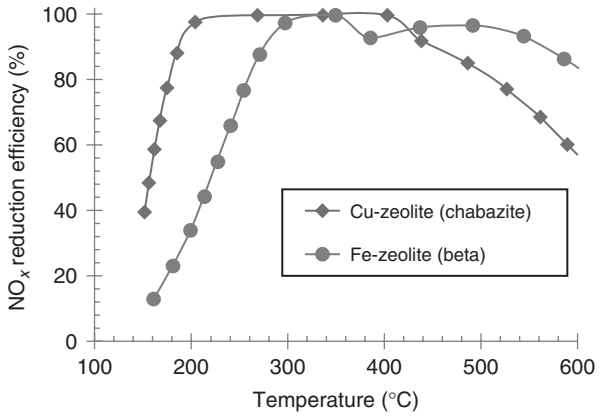


Figure 11. The NO_x reduction efficiency of two SCR catalysts as a function of catalyst temperature. The terms *chabazite* and *beta* refer to the type of zeolite that forms the basis for these catalysts.

to form ammonia sulfate particulate matter. To further minimize the potential for NH_3 emissions to occur, some SCR catalyst systems utilize a downstream NH_3 oxidation catalyst, which oxidizes the NH_3 to N_2 before the exhaust exits the tailpipe. NH_3 oxidation catalysts are typically based on Ni or Fe metals to provide a low cost oxidation function.

6.3.2 Hydrocarbon SCR

Although urea-based SCR effectively reduces NO_x in oxygen-rich exhaust, the requirement of a second tank on vehicles to store, handle, and refill adds cost and complexity to the system. Another form of SCR catalysis utilizes HC

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

reductants for the NO_x reduction process and is called *HC-SCR*. The intent of HC-SCR is to utilize the vehicle fuel itself as the reductant for the NO_x reduction process. As fuel contains a wide range of HC chemistries, the HC-SCR approach can be quite complex, and the HC-SCR efforts have primarily been limited to research activities. Research has shown that the effectiveness of HC-SCR approach is extremely dependent on the HC type. The most effective HCs are alcohols such as ethanol, and Ag-based catalysts supported on Al_2O_3 have been shown to give the highest NO_x reduction efficiencies. Other HCs studied include diesel fuel, propane, and propene; other catalyst materials studied include Pt on Al_2O_3 and Cu on Al_2O_3. A key challenge for all HC-SCR fuel and catalyst combinations is achieving high NO_x reduction efficiencies over a broad temperature range. Typically, peak NO_x reduction efficiencies do not exceed 80% and occur only over a narrow temperature window associated with the light-off temperature of the HC-catalyst combination.

Improving the NO_x reduction performance of HC-SCR has been a key objective for catalyst researchers. New approaches such as the addition of H_2 to the exhaust gas stream upstream of the HC-SCR catalyst and modification of catalyst formulations can greatly improve the low temperature NO_x reduction performance of the HC-SCR process (Burch and Coleman, 2002). Further research in this area is ongoing in hopes of achieving sufficient NO_x reduction efficiency to implement fuel-based SCR systems in the marketplace.

6.4 Lean NO_x trap catalyst

Another catalyst technology specifically designed for controlling pollutants from lean engines is the lean NO_x trap (LNT) catalyst. The LNT operates by storing or "trapping" NO_x on the catalyst surface when the engine operates lean; then, during brief rich modes of engine operation, the LNT releases and reduces the stored NO_x to N_2. The LNT is also known by the name *NO_x storage and reduction catalyst*, which also characterizes its operation. LNTs have been commercialized for diesel and lean gasoline passenger car applications with the lean gasoline commercialization occurring primarily in European countries.

An LNT catalyst is composed of materials similar to a TWC but with the addition of a storage component for NO_x. The catalyst is commonly coated to a ceramic monolith, as high coating levels are preferred to maximize NO_x storage capability. The coating composition is similar to the TWC with high surface area metal oxides supporting PGMs. The NO_x storage component is typically an alkali or alkaline earth material and is often based on K (potassium) or Ba (barium). The storage chemical form of the storage component changes during the operation of the LNT. For a K-based LNT system, K may be in the form of K_2CO_3 (potassium carbonate) originally then convert to KNO_3 (potassium nitrate) or KNO_2 (potassium nitrite), as the K stores NO_x under lean conditions. Then, under rich conditions, the NO_x is released from the K and reduced to N_2 with the K left as KOH (potassium hydroxide) on the surface, which quickly converts back to K_2CO_3 with the adsorption of CO_2 in the exhaust. The cases shown in Equations 10 through 12 are example reactions for a K-based LNT. In actuality, other surface chemistry forms of K may exist and differ based on nanoscale properties of the dispersed K and PGMs on the catalyst surface. Similar reactions occur for Ba-based LNTs with BaO also being a potential moiety of interest. Ongoing studies seek to precisely characterize these critical surface reactions that are central to the operation of the LNT technology (Broqvist *et al*., 2004; Toops, Smith, and Partridge, 2006).

[lean NO_x adsorption]

$$K_2CO_3 + 2NO_2 \rightarrow KNO_3 + KNO_2 + CO_2 \quad (10)$$

[rich NO_x release and reduction catalyzed by Pt]

$$KNO_2 + KNO_3 + 4H_2 \rightarrow 2KOH + N_2 + 3H_2O \quad (11)$$

[revert back to original carbonate form]

$$2KOH + CO_2 \rightarrow K_2CO_3 + H_2O \quad (12)$$

The LNT operates in a cyclic manner that parallels the surface chemistry changes of the catalyst. During lean operation of the engine or "storage phase," NO_x in the oxygen-containing exhaust travels into the catalyst where it is trapped by the alkali or alkaline earth component. The NO_x is bound to the alkali/alkaline earth material as a nitrate or nitrite species. Trapping of NO and NO_2 can differ, but the PGM in the LNT can also convert NO to NO_2 for greater trapping efficiency. In general, the NO_x trap reaction causes CO_2 to be released by the catalyst, as the carbonate form of the alkali or alkaline earth component changes. Ultimately, after a period of time, the LNT begins to fill up with NO_x. The upstream section of the LNT will become saturated with NO_x first, and as the downstream sections of the LNT fill with NO_x, NO_x begins to slip through the catalyst untreated. At this point, a process known as *regeneration* is initiated to reduce the stored NO_x to N_2 and ready the LNT for more NO_x storage (Figure 12).

The "regeneration" phase of the LNT cycle is achieved by creating net-rich exhaust conditions that contain reductant species and depleted oxygen levels. Typically, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

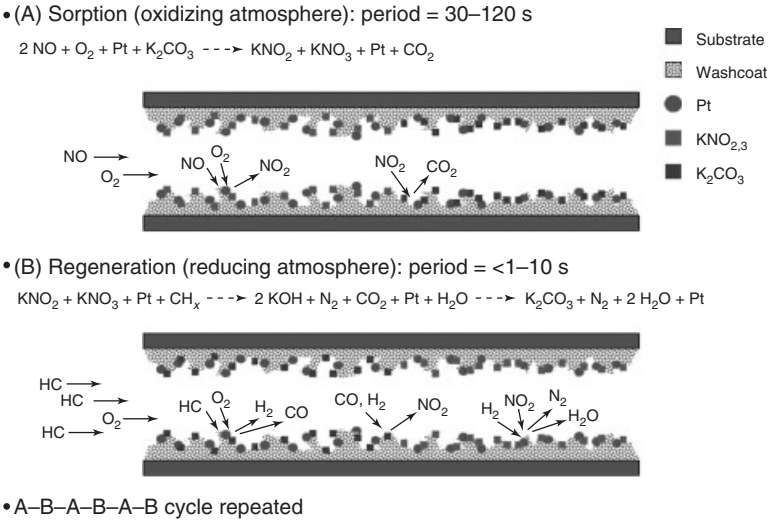


Figure 12. Diagram of the processes that occur for an LNT catalyst. The catalyst operates in a cyclic manner alternating between sorption and regeneration phases to trap and reduce NO_x, respectively.

net-rich exhaust is created by operating the engine with rich AFRs; this brings oxygen down to very low levels and creates CO, H_2, and HC reductants in the exhaust. Any stray oxygen can be further eliminated in the LNT by oxidation over the PGM component with the reductants. Once the rich exhaust conditions are achieved, the alkali/alkaline earth storage components will release the NO_x and the NO_x will be reduced to N_2 over the PGM component of the catalyst. The alkali/alkaline earth component rapidly reforms as a carbonate and becomes ready for more storage once the lean portion of the cycle begins.

The period of both the lean and the rich portions of the LNT cycle are carefully controlled to minimize the downstream pollutant levels. The lean period ends when NO_x slip downstream of the catalyst begins to occur; map databases or downstream NO_x sensors can be used by the vehicle to control the period end. The storage capacity depends on temperature and other parameters; LNTs are sized by application with typical lean periods ranging from 30 to 120 s. In contrast, the rich period is comparably short and generally lasts 1–5 s. As the overall objective of lean engine operation is to reduce fuel consumption, minimizing rich operation is important. Any extra fuel required to operate the engine rich for LNT regeneration is referred to as a *fuel penalty*. Shorter rich periods reduce the fuel penalty and also minimize unwanted pollutants downstream. During the rich conditions of regeneration, the NO_x is released rapidly from the alkali/alkaline earth components and can be rapidly reduced with the reductants in the exhaust stream; however, once the bulk of NO_x is released, lower levels of NO_x present in the reductant-rich exhaust stream tend to get further reduced to NH_3. Therefore, shortening the rich period to end just after the core NO_x storage sites are regenerated helps to minimize NH_3 emissions; smaller levels of N_2O emissions are also minimized in this manner. Furthermore, once the core NO_x storage sites are regenerated, less reductants are consumed; therefore, CO, H_2, or HCs can be released from the LNT. Although NH_3 and reductants can be oxidized by additional "clean up" catalysts downstream, the extra cost and complexity associated with this solution is not as preferred as the more cost-effective careful control solution (Figure 13).

Many aspects of LNT performance are affected by temperature. As mentioned, the storage capacity of the LNT varies with temperature. Different levels and mixtures of alkali/alkaline earth components are designed into LNTs to optimize storage performance for a given application. While higher levels of alkali/alkaline earth materials increase storage capacity, negative impacts can occur as well when the PGM component is physically covered by the alkali/alkaline earth materials. Moreover, at high temperatures, the alkali/alkaline earth materials, especially in nitrate form, can essentially melt and migrate into the cordierite substrate where they weaken the cordierite material. Ideally, the level of alkali/alkaline earth material is controlled to the point where evenly dispersed coverage of the metal oxide surfaces is achieved; this maximizes NO_x storage while minimizing other negative effects.

At higher temperatures, the nitrate species are not stable and NO_x storage diminishes. The actual temperature range where this occurs depends on the alkali/alkaline earth species type. Ba tends to have a higher temperature of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

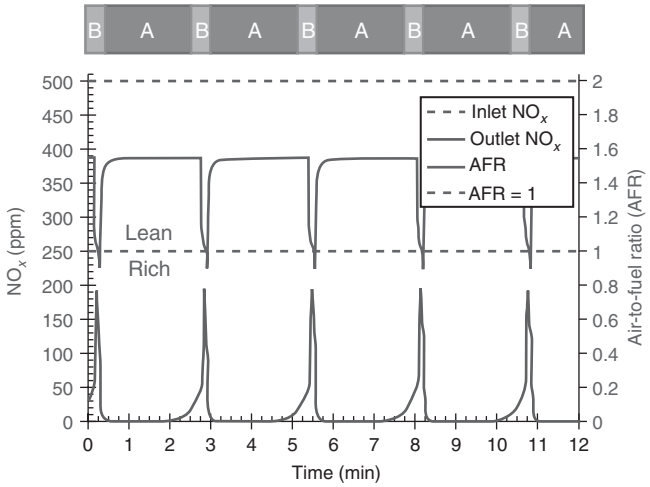


Figure 13. Plot of the NO_x and air-to-fuel ratio (AFR) for an LNT over the course of several cycles of operation. During the sorption phase ("A"), NO_x is trapped on the LNT under the lean conditions. Toward the end of the sorption phase, NO_x begins to slip past the catalyst as apparent by the rise in NO_x levels in the outlet NO_x data, and the regeneration phase ("B") must be conducted to clear the LNT for more NO_x trapping to occur. During regeneration, rich conditions enable the reduction of the stored NO_x except for a short spike of NO_x emitted as seen in the outlet NO_x data. On average over time, the LNT converted 97% of the inlet NO_x to N_2.

instability and is thus preferred in many automotive applications where high exhaust temperatures can occur.

At low temperatures, the LNT effectively stores NO_x, but the regeneration process can limit the overall performance. The release of NO_x is limited at low temperatures meaning that the LNT cycle can fail to completely regenerate the catalyst. The degree of regeneration that does occur on the LNT is also a function of the reductant chemistry in the exhaust. H_2 is the most effective reductant for regenerating the LNT at low temperatures, and CO is the second best reductant at low temperatures. HCs are often not effective for LNT regeneration below the 200–250°C range where the PGMs on the LNT have difficulty partially oxidizing the HCs to produce more reactive reductant species. The level and types of PGMs on the LNT have a major effect on low temperature LNT performance, as the PGMs affect the reductant utilization during regeneration and the NO to NO_2 oxidation efficiency, which effects low temperature NO_x storage (Figure 14). PGMs affect reductant utilization by enabling reforming-type reactions to occur such as the water–gas shift reaction (Equation 13) and steam reforming (Equation 14); these reactions alter the reductant chemistry mixture for LNT regeneration.

[water–gas shift reaction]

$$CO + H_2O \rightarrow H_2 + CO_2 \qquad (13)$$

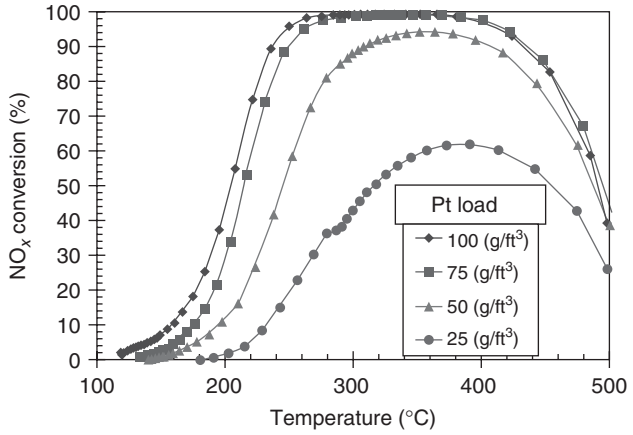


Figure 14. NO_x conversion as a function of temperature for LNTs with varying Pt load. As the Pt load decreases, conversion at the lower temperatures is lost first.

[steam reforming, for example, HC methane]

$$CH_4 + H_2O \rightarrow CO + 3H_2 \qquad (14)$$

A notable drawback to LNTs is their sensitivity to sulfur poisoning. Sulfur is a known catalyst poison and negatively impacts many catalytic processes, but for LNTs, sulfur has major impacts. Sulfur present in the exhaust, typically as SO_2, can be trapped by the alkali/alkaline earth storage components, and the resulting sulfate species that are formed on the catalyst surface are more thermodynamically stable than the nitrate or carbonate moieties. The more stable sulfates do not release SO_2 at typical regeneration temperatures, which prevents the sulfated alkali/alkaline earth site from trapping NO_x. Thus, upon significant sulfur exposure, the LNT cycle is effectively broken and overall NO_x reduction efficiency of the LNT is diminished. The negative impact of sulfur on LNTs is managed on several fronts. First, low sulfur fuels are required. For instance, the U.S. EPA reduced the maximum level of sulfur in diesel fuel from 500 to 15 ppm to enable emission control technologies such as LNTs. Even with ultra-low sulfur fuels, some sulfur will be present in the exhaust, and over time, sulfur will accumulate on the LNT (oil consumption by the engine also contributes to sulfur in the exhaust, as oil has high levels of sulfur). Once this occurs, the sulfur can be removed during vehicle operation in a process called *desulfation*. Desulfation is essentially the same process utilized for NO_x regeneration of the LNT except that the SO_x removal requires higher catalyst temperatures (typically >500°C) and longer time periods. Owing to the extra complexity of desulfation, it is only done as required at intervals of tens of hours or thousands of miles of operation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

As with all catalysts, the most effective application of the LNT catalyst technology is in an optimized system. The engine controls have a major effect on LNT performance, as the rich conditions for LNT regeneration are created primarily by engine operation. Fuel-rich combustion is monitored in the exhaust by oxygen sensors, and downstream oxygen and NO_x sensors also aid control of the LNT process. Engine operation is also controlled for careful thermal management of the LNT during both trapping and desulfation processes. In general, an upstream catalyst is paired with the LNT for optimal operation. In diesel engine applications, a DOC upstream of the LNT can assist in reductant control and thermal management of the LNT. In gasoline engine applications, a TWC is commonly placed upstream to control NO_x during stoichiometric operation and also aid in reductant optimization for LNT regeneration. Moreover, as DPFs require occasional high temperature exposure for regeneration, often the LNT desulfation process is coupled to the DPF process to reduce the overall fuel penalty for the emission control system.

6.5 Hydrocarbon trap catalysts (cold start)

HC trap catalysts are a specialty catalyst that can greatly assist cold start emission control. HC traps operate by trapping HCs during cold temperature conditions; then, at higher temperatures, the HCs are released and oxidized by another catalyst on the system. They are typically composed of zeolite materials or other high surface area metal oxides that can provide a large storage capacity for HCs. HC traps are designed for cold start emission reduction and function by trapping HCs during the warm-up phase of the driving cycle; subsequently, after warm-up occurs, the HCs are thermally desorbed from the HC trap and released to a downstream oxidation catalyst or TWC where they are oxidized to CO_2 and H_2O. Commercialization of HC traps has been limited to date but may increase, as lower exhaust temperatures from more efficient engines and stricter HC emission regulations demand more effective HC emission control.

7 ONGOING CHALLENGES

Although great progress in catalysis for automotive applications has been achieved, a multitude of challenges still exist in the application of catalysts for automotive emission controls. Current and future vehicles are addressing the dual challenges of better fuel economy and lower emissions in an atmosphere of overall concern for the environment. These challenges are spurring the introduction of new drivetrain technology as well as alternative fuels.

Hybrid electric vehicles (HEVs) are gaining ground in the marketplace and greatly improving fuel economy especially for city driving conditions where brake energy recovery enables greater efficiency. The engine for HEVs is predominantly a stoichiometric gasoline engine, and TWCs are utilized to control emissions. Future movement toward more electrification in HEVs including the plug-in hybrid electric vehicle (PHEV) may challenge the TWC technology during repeated near-cold start conditions, which are caused by greater time operating in pure electric modes. Thus, advanced catalyst technologies to address repeated starting in HEVs is of interest.

While growth of HEVs in the marketplace is occurring, vehicles with only ICE power sources are expected to persist in the marketplace. Several approaches will drive improvements in the fuel efficiency of ICE vehicles. Two major trends expected for emissions control of future ICE vehicles are lower exhaust temperatures and lean operation. Lower exhaust temperatures will result from vehicle fuel efficiency gains via both smaller boosted engines and more efficient combustion processes. In both cases, lower exhaust temperatures and increased time for warm-up result. The challenges for catalysts will be lower temperature operation and advanced strategies for cold start emission control.

Lean engine operation will be driven by the need for greater fuel economy, and as described, lean engine exhaust presents unique challenges for NO_x control. While urea-based SCR and LNT catalyst technologies have been commercialized for lean diesel and gasoline engine vehicles, more development and advancements can be made on these technologies that have relatively less experience on automotive applications in comparison to the mature TWC technology. Key development thrusts for lean engine emission controls will be expansion of the temperature window for operation, lower cost, and improved reductant supply and control. Better integration with other catalyst components is also an area of interest. In some cases, hybrid catalyst systems that combine different catalyst technologies may prove to give better emission control performance. One example system is the approach of combining an LNT catalyst with an SCR catalyst; NH_3 formed by the LNT during the regeneration process can be used for further NO_x reduction by the downstream SCR catalyst (Theis, Dearth, and McCabe, 2011).

In addition to drivetrain changes in the future, a growing variability in fuel supply is also occurring. Driven in large part by efforts to produce sustainable fuels from biomass feedstocks, new fuels are likely to have greater biofuel content, which results in changes in fuel chemistry and exhaust chemistry. The major exhaust gas chemistry changes will occur to HC emissions. Fuel HC chemistry differences will carry through to exhaust unburned HCs

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

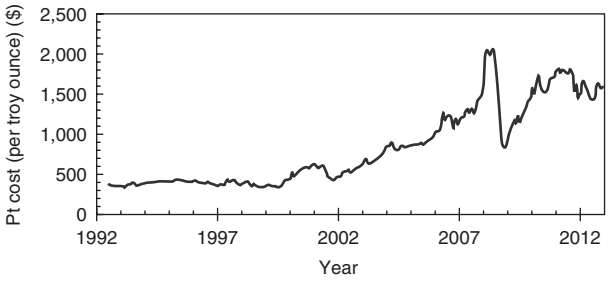


Figure 15. The monthly average market price of Pt has been volatile over the past decades. (Reproduced from Johnson Matthey, n.d.(b). Figure reproduced with permission from Johnson Matthey.)

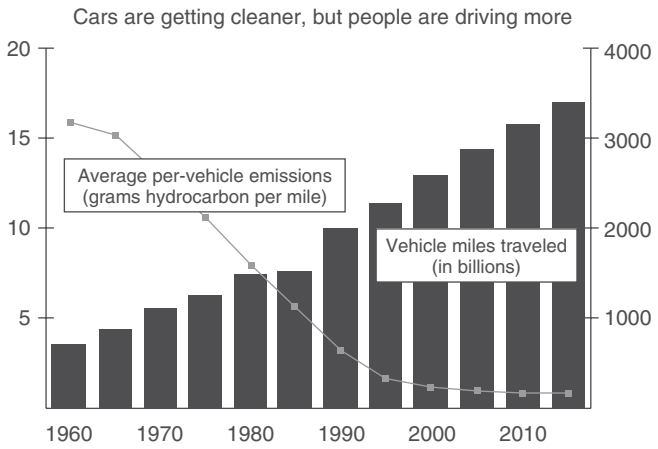


Figure 16. Increase of total vehicle miles traveled in the United States over time. (Reproduced from United States Environmental Protection Agency, n.d.)

and challenge oxidation catalysts during cold start, and some durability issues from impurities introduced in biofuel processing will need to be addressed.

A continued focal point for catalyst development will be lower cost. The competition for marketplace acceptance of new vehicle technologies will be affected greatly by price; therefore, efforts to minimize catalyst system costs will continue by the automotive companies and suppliers. Complicating the cost issue are highly volatile PGM prices that are affected by global supply and demand. Continued interest in the practice of optimizing PGM content and composition in TWCs and other catalysts (also known as *thrifting*) is expected to continue to allow products to manage swings in PGM pricing (Figure 15).

8 CONCLUSION

The collective total number of vehicle miles traveled (VMT has continuously increased over decades of years with a VMT in the United States exceeding three trillion miles in 2010. Catalyst technology has played a critical role in minimizing the emissions from vehicles that meet our transportation needs. Highly effective automotive catalysts have reduced pollutants from vehicles and protected our air quality, which has led to improved human health and an improved quality of living with clearer views of our natural surroundings. As global needs for transportation continue to increase, the automotive catalyst will continue to play a central role in enabling personal transportation in an environmentally sustainable way (Figures 16, 17).

ENDNOTES

1. Unable to determine the origin of the term "washcoat," but this term is commonly used in the field.

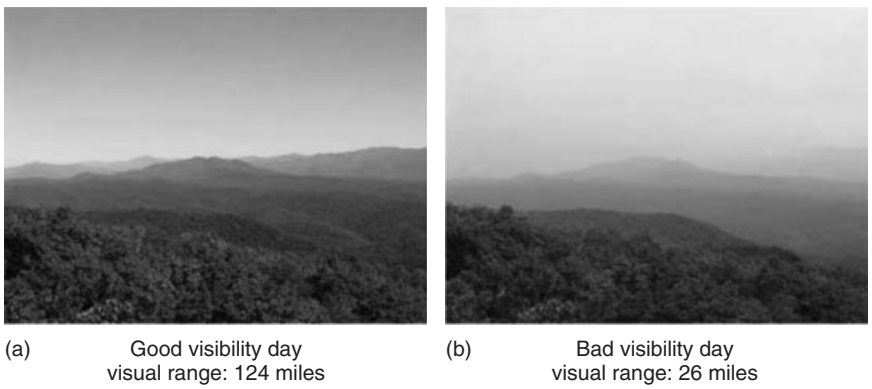


Figure 17. Photos of Smoky Mountains from Look Rock, TN on (a) good and (b) bad visibility days. (Reproduced from National Park Service, n.d.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2. DIN 70070 by Deutsches Institut für Normung e. V. is the European standard for urea solutions as diesel exhaust fluids for urea-based SCR catalysis; see http://www.naautomobil.din.de/cmd?artid=82816886&bcrumblevel=1&contextid=naautomobil&subcommitteeid=69429897&level=tpl-art-detailansicht&committeeid=54738955&languageid=en.
3. Currently, the American Society for Testing and Materials known as *ASTM International* is investigating a new international standard for urea solutions as diesel exhaust fluids for urea-based SCR catalysis; see http://www.astmnewsroom.org/default.aspx?pageid=2316&year=2010&category=General+Topics.

REFERENCES

Bartholomew, C.H. (2001) Mechanisms of catalyst deactivation. *Applied Catalysis A: General*, **212**, 17–60.

Bartholomew, C.H. and Farrauto, R.J. (2006a) *Fundamentals of Industrial Catalytic Processes*, 2nd edn, pp. 756–761. ISBN: 978-0-471-45713-8

Bartholomew, C.H. and Farrauto, R.J. (2006b) *Fundamentals of Industrial Catalytic Processes*, 2nd edn, pp. 68–77. ISBN: 978-0-471-45713-8

Blackthorn (n.d.) http://www.blackthorn.eu.com/html/catalytic-converters-gas-engines.aspx

Broqvist, P., Grönbeck, H., Fridell, E., and Panas, I. (2004) NO_x storage on BaO: theory and experiment. *Catalysis Today*, **96**, 71–78.

Brunauer, S., Emmett, P.H., and Teller, E. (1938) Adsorption of gases in multimolecular layers. *Journal of the American Chemical Society*, **60**, 309–319.

Burch, R. and Coleman, M.D. (2002) An investigation of promoter effects in the reduction of NO by H_2 under lean-burn conditions. *Journal of Catalysis*, **208**, 435–447.

CBS New York (2011) New Push To Clear The Air After Report On Tri-State Area Smog in CBS New York. http://newyork.cbslocal.com/2011/09/22/new-push-to-clear-the-air-after-report-on-tri-state-area-smog/ (accessed 18 November 2013)

Graham, G.W., Jen, H.-W., Chun, W., *et al*. (2004) Coarsening of Pt particles in a model NO_x trap. *Catalysis Letters*, **93**, 129–134.

Japanese Sports Cars (2009) Mazda, the First Japanese Automaker to Develop a Urea SCR System for Cars in Japanese Sports Cars. http://www.japanesesportcars.com/mazda-urea-scr-system-for-cars_5629.html

Johannessen, T., Schmidt, H., Svagin, J., *et al*. (2008) Ammonia storage and delivery systems for automotive NO_x aftertreatment. SAE Technical Paper 2008-01-1027, doi:10.4271/2008-01-1027.

Johnson Matthey, n.d.(a) Multimedia Library. http://www.matthey.com/media/imagelibrary/images.htm

Johnson Matthey, n.d.(b) Platinum Today. http://www.platinum.matthey.com/

Krishnan, R. and Tarabulski, T.J. (2005) Economics of Emission Reduction for Heavy Duty Trucks, http://www.dieselnet.com/papers/0501krishnan/ (accessed 12 February 2012).

National Park Service (n.d.) Great Smoky Mountains National Park. http://www.nature.nps.gov/air/webcams/parks/grsmcam/grsmcam.cfm#

Nett Technologies Inc (n.d.) What Is a Diesel Oxidation Catalyst? http://www.nett.ca/faq/diesel-4.html

Seinfeld, J.H. and Pandis, S.N. (1997a) *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, pp. 300–302. ISBN: 0-471-17815-2

Seinfeld, J.H. and Pandis, S.N. (1997b) *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, pp. 240–241. ISBN: 0-471-17815-2

Seinfeld, J.H. and Pandis, S.N. (1997c) *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, pp. 292–298. ISBN: 0-471-17815-2

Seinfeld, J.H. and Pandis, S.N. (1997d) *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, pp. 531–533. ISBN: 0-471-17815-2

Theis, J., Dearth, M. and McCabe, R. (2011) LNT+SCR catalyst systems optimized for NO_x conversion on diesel applications. SAE Technical Paper 2011-01-0305, doi:10.4271/2011-01-0305.

Toops, T.J., Smith, D.B., and Partridge, W.P. (2006) NO_x adsorption on $Pt/K/Al_2O_3$. *Catalysis Today*, **114**, 112–124.

UK Catalysts (n.d.) http://www.ukcatalysts.com

United States Environmental Protection Agency (n.d.). http://www.epa.gov/otaq/

van Helden, R., Verbeek, R., Willems, F. and van der Welle, R. (2004) Optimization of urea SCR deNO_x systems for HD diesel engines. SAE Technical Paper 2004-01-0154, doi:10.4271/2004-01-0154.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto130
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Solid/Condensed Phase Aftertreatment Systems

George G. Muntean, Mark L. Stewart, and Maruthi N. Devarakonda

Pacific Northwest National Laboratory, Richland, WA, USA

1 Introduction 1
2 Particulate Matter from Internal Combustion Engines 1
3 Diesel Particulate Filtration 6
4 Integration of Particulate Filtration with Vehicle Systems 11
References 16

1 INTRODUCTION

This chapter examines the issues surrounding the nature of particulate matter (PM), its regulation and the technology used for its control. The topic is introduced by first giving the legal and physical definitions of PM, followed by a discussion of the impacts of particulates and a brief overview of the regulatory landscape. The discussion then turns to the primary control technology for particulates, that is, filtration. The mechanisms of filtration are described along with the physical description of real-world approaches. Finally, this chapter concludes with a discussion of how particulate filtration devices are integrated into the overall vehicle system, the performance implications, and practical concerns regarding their reliability.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 PARTICULATE MATTER FROM INTERNAL COMBUSTION ENGINES

2.1 What is diesel particulate matter?

If one would ask the typical motorist to describe diesel particulates, they would likely point to the exhaust stack of an older diesel-powered truck and highlight the cloud of black smoke that billows out as the driver does a hard acceleration. What the motorist is primarily witnessing are large carbonaceous particles formed from the incomplete combustion of fuel. This smoke is not unlike the soot that comes off the tip of a flickering candle flame. In reality, diesel particulate matter (DPM) is not so easily described. While it is true that a large weight percentage of the particulates are in the form of this visible smoke, there are many other components that are potentially more harmful and are not visible. Ironically, it has been pointed out that the large visible particles are potentially the least harmful as they settle out of the atmosphere quickly and are not easily transported into the alveolar sacks in the human lung during respiration. This observation is at the root of the current debate on regulation of PM, which is discussed in more detail later. In addition to the carbonaceous compounds (both visible and invisible), particulate emissions also include metals, ash, sulfates, and adhered organic compounds as illustrated in Figure 1. These components may potentially have even more severe health impacts than the black carbon. As such, the control of particulate emissions would ideally address all these constituents. It is not sufficient to simply eliminate the visible nuisance.

A further complicating fact of DPM is its dynamic nature. From the first moment of nucleation to the moment of deposition in the lung, the PM evolves and changes its characteristics. The life cycle of a DPM particle begins

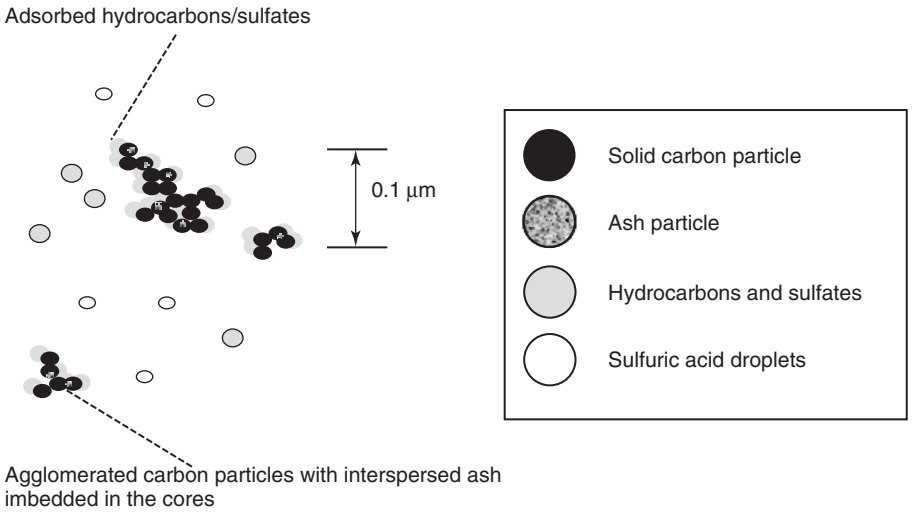


Figure 1. Image depicting the composition of diesel particulate matter.

in a high temperature, high pressure oxygen lean environment inside the combustion chamber. This particle is quickly quenched as it exits into the exhaust manifold and is transported through the exhaust system of the vehicle. At this stage, the particle may undergo agglomeration with other particles, it may adsorb vapor-phase compounds, or entirely new particles may begin nucleating. The actual specific sequence of events is controlled by the design of the vehicle, the fuel quality, operating conditions, and the presence of catalytic emissions control devices. As the particles exit the stack (or tailpipe), they undergo yet another sudden quench and a dramatic dilution within the atmosphere. At this point, the particles are transported through the air while undergoing further reactions with other compounds in the atmosphere under the action of sunlight, rain, and other naturally occurring PM. In addition to the PM emitted from the exhaust, new particulates may also form in the atmosphere from gas-phase emissions of sulfur and nitrogen compounds. These compounds interact with materials in the air to form what are known as secondary DPM. This evolution continues until the final moment of deposition. It is only at this moment in time when the specific nature of the particulate has a potential health impact. It quickly becomes apparent that the precise understanding of this evolution is a daunting challenge.

In the face of these uncertainties, researches have (and continue to) grapple with the most impactful way to monitor, control, and regulate diesel particulate emissions. At the heart of the debate are the precise conditions under which the particulate sample is derived and the metric by which the particulate standards are promulgated. Ideally, the sampling process would exactly mimic the entire transport chain of the DPM to the moment of inhalation and the metric would encompass the specific qualities of the DPM that have the most detrimental impact on human health. In practice, the research community has settled on standardized testing that can be done repeatably and cost effectively while capturing many of the salient phenomena described earlier, especially atmospheric dilution. Perhaps, the more contentious issue relates to the metric by which the standards are set. Traditionally, the metric has been gravimetric, that is, by the weight of the PM released, as this is a convenient and economic approach toward measurement. Clearly, this is driven more by the practicality of implementation than by specific health effect consideration. Because of this, there have been recent moves to implement a metric based on the particulate number density of the smaller diameter particles that are more likely to penetrate into the deepest recesses of the lung. The United States Environmental Protection Agency (US EPA) has not yet promulgated these standards. If they ever do become law, it has the potential to dramatically impact the technologies currently deployed for emissions control.

2.2 Why do particulate emissions need to be controlled?

Diesel particulates have both known and suspected impacts to our environment. There is little dispute that excessive soot causes issues with visibility and deposition on structures such as buildings and monuments. There is some recent speculation that DPM in suspension in the atmosphere may have an impact on global climate change but

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

there is still much work to be done before definitive cause and effect can be determined. The most recent results have indicated that the PM may have either a net cooling or a net heating effect depending on the specific characteristics of the particles (Solomon *et al*., 2007). While these are important environmental issues, they have not been the primary driver for regulation of particulate emissions.

The suspected impact that has been the dominant issue over the past several decades has been the impact on human health. There is a large body of evidence implicating DPM in human morbidity and mortality. However, the very complexity of particulates makes the precise understanding of human health effects challenging. The many studies of DPM health effects (mostly toxicological and epidemiological) have often been inconclusive and, in some instances, contradictory. However, when taken as a whole, the body of literature led the Environmental Protection Agency (EPA) to conclude that sufficient evidence existed to take regulatory action. What has become clear is that particular attention needs to be placed on the smallest particles, especially those which can penetrate into the deepest portions of the lung, as these are the most likely to contribute to potential health effects. Once in the alveolar sacks, these particles may attack the lung cells through either physical or chemical action. As alluded to already, the issues regarding size effects are of paramount concern and are driving ongoing discussions of PM2.5 number regulations in addition to the current PM10 gravimetric regulations (the numbers 2.5 and 10 refer to the effective diameters of the PM in micrometers). In 1997, EPA promulgated National Ambient Air Quality Standards (NAAQS) specifically for PM2.5 while retaining the existing PM10 standards. In 2006, EPA made further revisions to the NAAQS by tightening the PM2.5 standards to 35 mg/m^3 for a 24-h period while maintaining a 15 mg/m^3 annual standard. At the same time, EPA eliminated the annual PM10 standard in recognition of the fact that there was a lack of evidence linking these coarse particle exposures to health effects. It did, however, retain the existing 24-h PM10 standard of 150 mg/m^3. These modifications to the NAAQS are the first step toward the processes of implementation of potentially more stringent automotive emissions regulations.

In addition to size effects, there is concern with carbonaceous particulate material because of its propensity for adsorbing other chemical components, which may be most harmful. For example, polycyclic aromatic hydrocarbons (PAHs) and nitro-PAH are known carcinogens and have been detected on DPM. Organic compounds are also suspected of inducing allergic reactions and acting as irritants. The metallic particles found in DPM are known to participate in oxidation–reduction reactions while the smallest of particles ("ultra-fines") are believed to be able to translocate to other organs in the human body. The earlier work on health effects focused primarily on the carcinogenic aspects of PM, whereas more recent work has expanded to encompass cardiovascular effects and a focus on asthma and other degradations in lung function.

There remains much work to be done in understanding the impacts of PM on the environment and, especially, on human health. Pathophysiologic studies have typically involved model particles at high concentrations casting some uncertainties on the results. Similarly, epidemiologic studies have difficulties in precisely quantifying exposures to DPM as unique markers for diesel particulates are difficult to devise leading to uncertainties in exposure correlation. Regardless of these uncertainties, the preponderance of evidence does support some form of control over the release of excessive PM into the environment.

2.3 Legal definition

A fundamental understanding of the primary PM is critical for the design of efficient controls technologies, for example, its composition, morphology, and kinetics. These properties are required for the efficient capture and regeneration of filtration devices as described later in this chapter. For a comprehensive discussion of particulate formation and modeling (see Thermodynamic Analysis). However, the legal definition of DPM is just as important to the engineer as it sets the standard by which the effectiveness of the controls technology can be assessed. The EPA defines DPM implicitly through a prescribed sampling method and through the formulation of the legal standard, for example, 0.10 g/bHp-h. In this approach, particulates are anything that adhere to a filter patch (typically made of polytetrafluoroethylene, i.e., Teflon™, with 2 mm pore size) subjected to diluted exhaust flows and that add weight to the gravimetric measurement. The standards dictate that the exhaust is cooled and diluted to a temperature at or below 52°C in a system known as a *constant volume sampler* (*CVS*). The "grams" in the numerator do not discriminate as to size, density, or composition of the actual particulate material. It is simply the difference in weight between a dirty patch and a clean one. As the emissions standards have been reduced, this procedure has required modifications. Current engines, especially those equipped with particulate filters, now emit such low levels of PM that extreme care and precision must be used to obtain reliable and consistent gravimetric measurements. In fact, there have been concerns whether this testing approach is still reliable. To address these concerns, in 2007, the EPA modified the definition of PM by making numerous changes to the sampling procedure, most notably a requirement to maintain a narrow

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

range of temperatures on the filter face from 42 to 52°C. Despite these changes, meeting satisfactory coefficients of variation (COV) in repeatability tests remains a challenge. Clearly, an intimate understanding of this sampling process and the testing cycles are critical to the design engineers. At these very low emissions levels, even the subtlest changes to a design can have a profound impact on the measured results.

Over the years, researchers have spent countless hours developing an agreed on procedure for these particulate measurement tests. As stated earlier, the ultimate objective is for this single weight measurement to have a correlation with real-world hazardous particulate emissions. Regardless of the remaining uncertainties, a legal consensus has been reached and is prescribed in the Code of Federal Regulation (CRF). The clarity and stability of this agreement is critical to engine and vehicle manufacturers as it provides the basis for the design and optimization of the emissions control systems. For specific details of the CVS sampling procedure, the reader is directed to Title 40 of the CFR at the National Archives and Records Administration (Note: access to the CFR can be found at http://www.gpoaccess.gov/fr/index.html).

2.4 Regulation standards

The tremendous complexity of DPM regulation standards throughout the world prohibits a detailed survey in this work. The reader is directed toward the relevant regulations for the class of vehicle and the specific locale of interest. In the United States, the federal regulations for automotive particulate emission standards, as with the test procedures, are also found in Title 40 of the CFR and are promulgated by the EPA. The EPA maintains web sites that are valuable sources of information, and the interested reader is encouraged to explore those resources. A brief overview of the topic is given here to give a general sense of the approach, especially as it pertains to the design of particulate controls systems.

Typically, vehicle emissions regulations are set at the national level, as vehicles are mobile sources that often travel across regional boundaries. The United States has a notable exception in the authority that has been given to the State of California's Air Resources Board (CARB). This has a historic origin rooted in the fact that CARB promulgated regulations before the formation of the EPA and was, thus, "grandfathered" into the federal program. Both the CARB and EPA programs share many similar attributes although the CARB standards tend to be somewhat more stringent. Both programs partition the transportation sector along functional lines. Standards for on-road vehicles are separated from standards for off-road. Further distinctions are made within each category according to use. The on-road standards are largely divided into light-duty (LD) and heavy-duty vehicles. The off-road vehicles are divided into functional areas such as rail, marine, aircraft, and specialty vehicles such as mining, agricultural, and construction. This final category is typically further subdivided by horsepower ratings. For the on-road categories, the primary subdivisions are based on gross vehicle weight with numerous variations in each subdivision. The need for this complexity in regulation is unfortunate but understandable when viewed in light of the realization that each segment has very unique attributes and market drivers. Similarly, despite long-standing efforts at global standardization, global regulations are varied and reflect diverse geographical and historical influences.

Although specific regulations are unique, they all share general features that are ubiquitous. All regulations attempt to test the engine or vehicle in ways which best represent the actual in-use application. Countless studies have been performed to determine the drive cycles that would be most applicable. For particulate regulation, the transient cycles are most problematic as sudden, hard accelerations are the source of the majority of emitted DPM. In addition, all regulations have useful life provisions that set durability targets for the emissions controls system. Of particular concern to diesel particulate filtration are limitations on maintenance intervals as this has implications on ash build up in the devices. Many regulations have mandates for on-board diagnostics (OBD) to ensure the viability of the control system while in operation. These mandates implicitly drive high reliability standards while proper methods of sensing filtration device effectiveness are an active area of research. Finally, most regulators employ multiple test procedures that force stringent product variability requirements.

For illustrative purposes, the on-road heavy-duty segment will be briefly described, as it has been the segment of the transportation market at the forefront of particulate regulations in the United States.

Heavy-duty vehicles in the United States are defined as vehicles exceeding a gross vehicle weight rating (GVWR) of 8500 lbs (3855 kg) [note: CARB definition is 14,000 lbs (6350 kg)]. This category consists primarily of vehicles involved in the enterprise of moving large volumes and/or mass of materials. The tremendous variety of forms has led to the need to certify the engines as opposed to the vehicles themselves. In this segment, engines are certified by the engine manufacturer on motoring dynamometers in specialized test cells equipped with constant volume sampling equipment as described earlier (see Engine Performance and Exhaust Emissions for discussion on test cells and emission measurements, respectively). Furthermore, in this

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

segment, the emissions standards are normalized by the horsepower output of the engine. This is done in recognition of the primary function of the vehicles, that is, their function is to do work. In contrast, LD vehicles are dominated by passenger cars and light trucks where the entire vehicle is certified and the emissions standards are weighted by miles, for example, grams per mile. The heavy-duty segment is further divided into three categories by GVWR. Of these, the heavy heavy-duty diesel engines (HHDDEs) are the most numerous and are defined as vehicles above 33,000 lbs GVWR, typified by the line-haul tractor-trailers (aka "18 wheelers"). Historically, engines in this category would undergo the "transient FTP (Federal Test Procedure)" while the exhaust was diluted, then measured via filtration in a CVS tunnel. This transient test requires a motoring dyno that can not only adsorb energy but also drive power into the engine to simulate vehicle decelerations. This test cycle contains four unique portions that were derived from three individual cycle simulations (the fourth portion is a repeat of the first portion) as illustrated in Figure 2. They attempt to simulate driving in urban settings of New York City and Los Angles as well as a freeway portion in Los Angles. Combined, these portions total 1200 s of prescribed speed and torque as a percentage of rated values.

Beginning in 1998, two additional tests were incorporated and are run in addition to the transient FTP. The first is known as the *supplemental emission test* (*SET*). The emission limits are the same as for the FTP test but the SET is a steady-state test that attempts to better simulate line-haul operations on the freeway. The second test is the not-to-exceed (NTE) test. This test attempts to cap maximum emissions in a prescribed operating regime of the engine and is run in both steady and transient modes with varying ambient conditions. As of 2007, the particulate standards for heavy-duty trucks in the United States are 0.01 g/bhp-h. This level must be achieved on both the transient and SET tests, whereas the NTE limit is a factor of 1.25 higher. For additional details on the US regulatory standards, the reader is directed to Title 40 of the CFR at the National Archives and Records Administration (note: access to the CFR

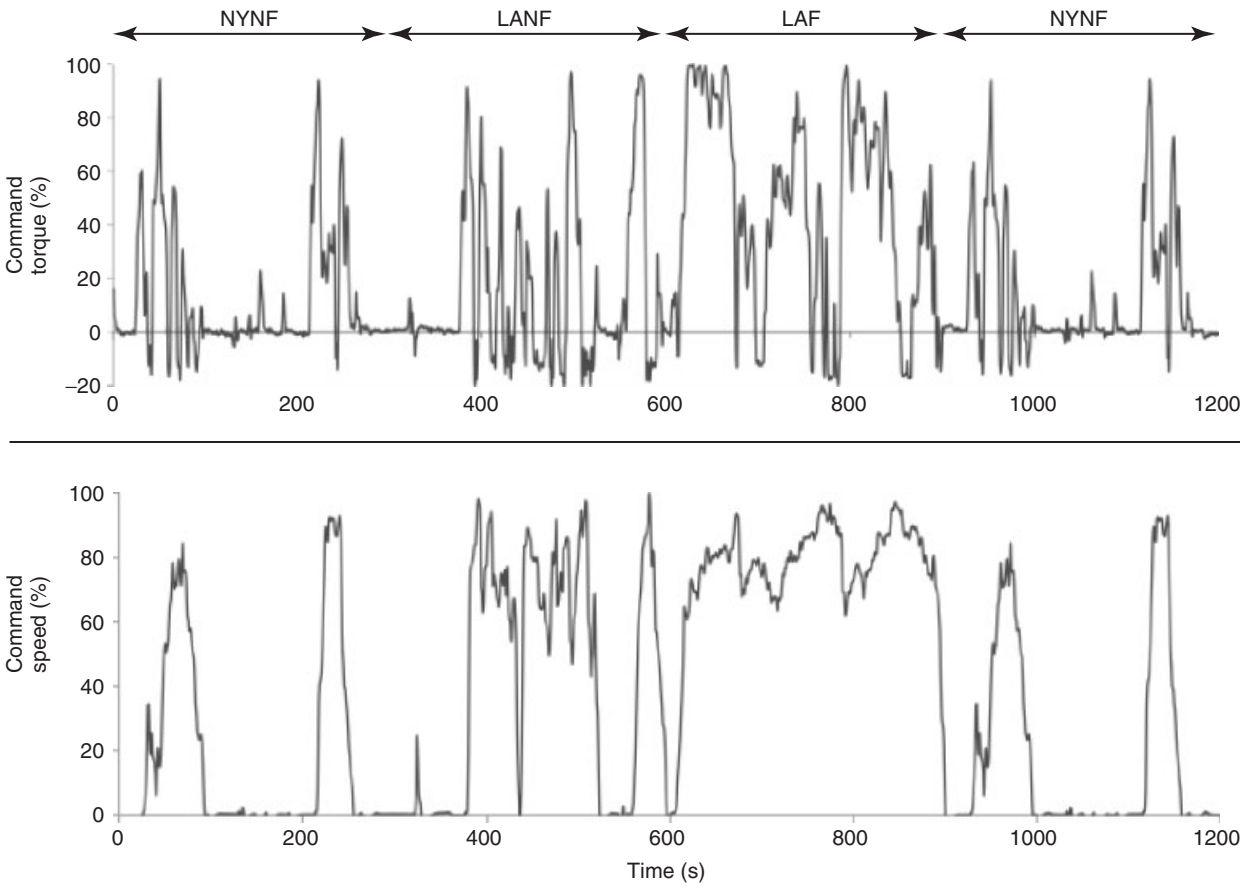


Figure 2. The US EPA heavy-duty transient Federal Test Procedure. NYNF is the "New York Non Freeway" section. LAF is the "Los Angles Freeway" section. LANF is the "Los Angles Non Freeway" section.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

can be found at http://www.gpoaccess.gov/fr/index.html or through the EPA web site).

Taken as a whole, the three test cycles attempt to ensure that particulate control is maintained throughout the entire operating range of the engine. Anomalous release at off-cycle operating points is restricted and transient operations are thoroughly covered. Little leeway is given for designs that attempt to simply optimize to the original transient certification cycle.

2.5 Control of diesel particulate emissions

Over the years, the primary approach to reducing particulate emissions has been through the control of its initial formation in the combustion chamber. Tremendous advances were made early on through improvements in air handling and fuel injection systems. As described in earlier chapters, higher injection pressures, optimized swirl, four valve cylinder heads, central vertical injection, and combustion chamber designs helped improve the mixing of the fuel with the air in the cylinder. Similarly, air handling equipment, such as variable geometry turbocharging, helped optimize air/fuel ratios to control particulate formation. In the future, continued improvements in combustion systems will likely further reduce particulate formation. Advanced fuel systems employing rate shaping and/or multiple injections will further optimize mixing, whereas strategies enabling low temperature combustion [e.g., PCCI (premixed charge compression ignition)] will help eliminate stratification of air/fuel mixtures. Ultimately, however, particulate emissions standards were reduced to levels that required additional control. Most significant were the changes to ultralow sulfur fuel (which lowered sulfate particles and allowed catalyzed emission control systems) and the use of physical filtration systems in the exhaust stream, that is, diesel particulate filters (DPFs). Wide-scale introduction of these devices in the United States began with the implementation of the 2007 heavy-duty on-highway rulemaking by the EPA. The fundamental processes associated with filtration in such devices are discussed in the next section followed by a brief discussion of practical issues encountered in these devices when deployed on vehicles.

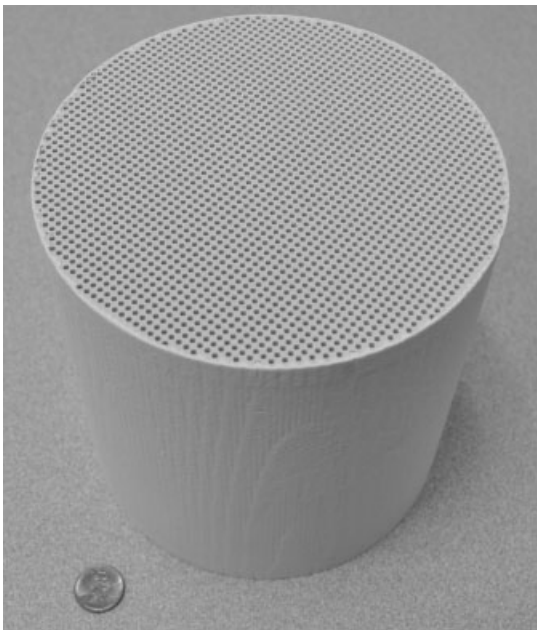


Figure 3. Ceramic diesel particulate filter.

3 DIESEL PARTICULATE FILTRATION

3.1 DPF overview

3.1.1 Description of a typical DPF

A number of different materials and configurations have been proposed as DPFs, but porous ceramic honeycombs (Figure 3) are currently the dominant technology for new cars and trucks. These devices evolved from the flow-through catalyst supports used in three-way catalytic converters on gasoline vehicles. Alternate channels are plugged at either end of the monolith to force the exhaust flow through the porous filter walls, as shown in Figure 4.

Typical properties of commonly used filter substrates are as follows

- wall porosities of around 50%
- mean pore sizes of around 10–20 μm
- filter wall thicknesses in most applications fall within the range 0.30–0.48 mm (12–19 thousandths of an inch)
- filters are available in sizes ranging from 143.8 mm (5.66 inch) in diameter by 152.4 mm (6 inch) long, for LD automotive applications, to 508.0 mm (20 inch) in

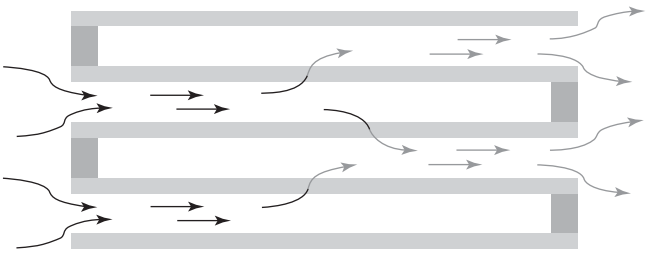


Figure 4. Exhaust flow through a ceramic DPF.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

diameter by 381.0 (15 inch) long, for large displacement heavy-duty applications
- 31 cells per square centimeter (200 cells per square inch) is a common channel density, but other cell densities are also available (Corning, 2010a).

3.1.2 Basic operation

As soot collects in a DPF, the backpressure increases. Accumulated soot must be burned away before the backpressure becomes excessive. This process is referred to as *filter regeneration*. Some oxidation of the soot in the filter happens during normal operation (passive regeneration), but rates are dependent on exhaust temperature and composition. In many systems, it is necessary to periodically raise the exhaust temperature to insure oxidation in order to reduce the inventory of soot in the filter (active regeneration).

3.1.3 Design goals

A DPF must have high enough filtration efficiency to allow the vehicle to meet particulate emissions limits. Backpressure must be minimized over the entire operating cycle in order to limit the negative impact on fuel economy. Similarly, maximizing overall system efficiency requires keeping the amount of energy used to raise the exhaust temperature during active regenerations to a bare minimum. Balancing these competing goals is a significant challenge for designers of DPF systems.

The total backpressure caused by a DPF unit is the sum of several resistances to flow, including entry and exit effects, resistance to flow down the inlet and outlet channels, and resistance to flow across the soot layer and filter wall (Konstandopoulos and Johnson, 1989; Masoudi, 2002). The relative importance of each of these resistances depends on the system design and operating parameters. Higher cell densities not only lead to higher channel resistance but also provide more filter area per monolith volume. Flow resistance through porous media, such as soot and filter substrates, is often expressed in terms of permeability (Masoudi, 2002), as described in Equation 1.

$$\frac{\Delta p}{t} = \frac{\mu v}{k} \tag{1}$$

where Δp is the pressure drop (N/m^2), t is the thickness of the layer through which the flow is passing (m), μ is the dynamic viscosity ($N{\cdot}s/m^2$), v is the velocity of the flow (m/s), and k is the permeability (m^2).

An approximate permeability value for a cordierite filter wall is 3E-13 m^2 (Konstandopoulos *et al*., 2000). Permeability becomes greater for larger pore sizes and higher porosities (Geankoplis, 1993). This means that the resistance of a clean filter can be lowered using a substrate with higher porosity and larger pores. Unfortunately, these features would also lower the clean filtration efficiency.

Durability is also a very important design consideration for DPFs. Ceramic monoliths are subjected to stresses during canning for mounting in exhaust systems. They can also experience high temperatures and steep temperature gradients during active regeneration events. Improper management of regenerations can lead to temperatures above the substrate melting points. Thermal gradients can cause abrupt cracking, whereas thermal cycling may lead to long-term degradation of mechanical properties. High melting points, high strength, high thermal capacitance, and low coefficients of thermal expansion (CTE) are therefore desirable from the standpoint of mechanical durability.

3.2 Soot filtration

3.2.1 Capture mechanisms

In filtration, suspended particles in the fluid are removed when they adhere to the solid structure of the filter medium. If particles had infinitesimal size and followed the fluid streamlines perfectly, no contact would be made with the walls of the filter medium pores, and none of the particles would be captured. Several physical mechanisms bring real particles into contact with the solid structure of the filter medium. Which mechanisms come into play in a given system depends on particle size, the geometry of the porous filter material, and a variety of other factors. Usually, one or two mechanisms dominate for particles of a given size. Important capture mechanisms include

- Diffusion: Diffusion moves particles between streamlines via the random Brownian motions arising from collisions with individual gas molecules. Diffusion capture often dominates for the smallest aerosol particles.
- Interception: Interception refers to the capture of particles that follow fluid streamlines almost perfectly, but are brought into contact with solid surfaces because of the finite size of the particles. In other words, capture by interception occurs when a particle's center of mass follows a streamline that passes within one particle radius of a solid surface. Interception tends to be important for particles of intermediate size.
- Momentum: A particle's inertia may cause it to move between fluid streamlines as they change direction

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

moving through a tortuous network of pores. This mechanism becomes important for heavy particles moving at high speeds.
- Gravity: Gravity forces cause heavy particles to drift downward within the flow field. This mechanism is often called *settling* in aerosol systems or *sedimentation* in liquid systems and becomes important for heavy particles in regions of slower flow.
- Other mechanisms, such as thermophoresis and electrostatic attraction, can also dominate in aerosol systems where the associated forces are present.

For flow rates and fluid and particle properties typically seen in DPF systems, the diffusion and interception mechanisms are usually considered to be the most important (Konstandopoulos and Johnson, 1989).

A unit collector model uses a simple shape to represent elementary obstructions making up a porous filter medium (Elimelech, 1995). Cylindrical unit collectors are often used to represent fiber filters, whereas spherical unit collectors are often used to represent granular beds. The flow field around the unit collector is usually assumed to be adequately approximated by some known analytical solution. This allows estimates to be made for capture rates of particles moving with the fluid flowing around the collector. Capture rates are estimated for fundamental capture mechanisms individually using simplifying assumptions, and then the capture rates are summed to estimate the total efficiency. Figure 5 illustrates several mechanisms for particle capture on a spherical unit collector. Unit collector models have been used extensively to predict DPF behavior and explain experimental data (Konstandopoulos and Johnson, 1989; Miyairi *et al*., 2006; Thalagavara *et al*., 2005; Tandon *et al*., 2010).

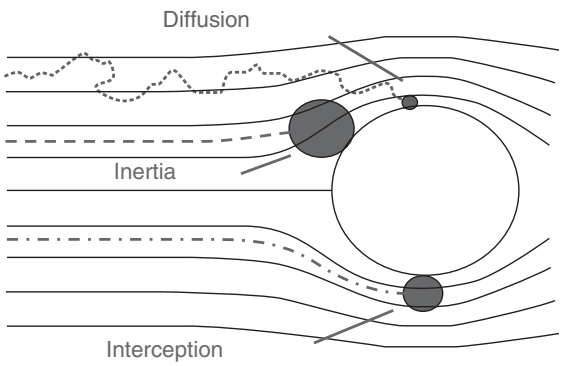


Figure 5. Several mechanisms leading to particle capture on a spherical unit collector.

3.2.2 DPF system mechanics

Soot deposits have very high porosities, between 90% and 99%, depending on operating conditions (Konstandopoulos, Skaperdas, and Masoudi, 2002). However, as the feature sizes of the soot particles are as small as tens of nanometers, soot permeability is much lower than that of a clean DPF wall, perhaps between 2.0×10^{-15} and 5.0×10^{-14} m^2 (Konstandopoulos, Skaperdas, and Masoudi, 2002; Suresh, Khan, and Johnson, 2000).

When a clean DPF is first exposed to particle-laden exhaust, there is some penetration of the soot particles into the porous filter wall. This is referred to as *depth filtration*. Initially, soot can form a thin coating on the inside pore surfaces, resulting in modest increases in backpressure per mass of soot captured. This is not always seen in practice, however, because the soot quickly begins to block pore throats within the medium. This blockage results in more dramatic increases in pressure drop per mass captured, as the soot deposits form in locations where flow restrictions result in high fluid velocities. At some point, all of the pore throats on the upstream filter wall surface become blocked, and soot begins to accumulate as an even cake on the filter wall, which is referred to as *cake filtration*. Backpressure then increases less quickly per mass of soot collected, as the incremental soot volume is now spread over a relatively even layer, where fluid velocities are lower than in the pore throats (Merkel *et al*., 2003). These stages of filtration are illustrated in Figure 6, which shows a computational simulation of the pressure drop through a silicon carbide (SiC) DPF wall as a function of filter loading.

From the description earlier, it can be seen that not all of the soot trapped in a filter contributes equally to

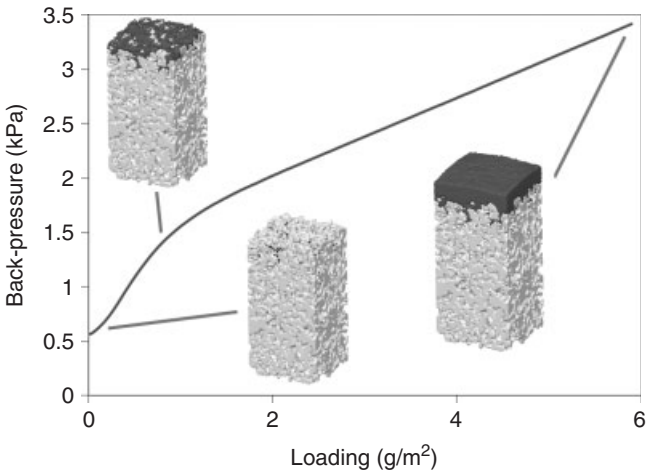


Figure 6. Pressure drop through a silicon carbide DPF as a function of filter loading.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

flow resistance. One practical result is that it is difficult to estimate the soot inventory in the DPF solely from the backpressure. Hysteretic effects can occur when operating conditions cause changes in the distribution of soot with respect to the filter walls. For example, passive regeneration can burn away soot within pore throats, whereas the soot cake remains intact because of continuing collection of new soot. In this case, the observed backpressure could decrease significantly, whereas the soot inventory in the filter remains constant (Premchand *et al*., 2007).

3.3 DPF regeneration

3.3.1 Soot oxidation

Removal of soot by oxidation is a requirement for continued operation of a DPF. Lean combustion in diesel engines means that there is ample oxygen present in the exhaust stream for oxidation of soot. Unfortunately, reaction rates at exhaust temperatures are generally too low to keep up with rates of soot accumulation. Exhaust temperatures must therefore be raised significantly to initiate soot combustion by oxygen in the filter, either by modifying engine operation or by injecting fuel into the exhaust stream before an upstream oxidation catalyst. Soot can also be oxidized by NO_2 in the exhaust. Reaction rates are much faster than for oxidation by oxygen, so NO_2 oxidation is very important for passive regeneration at normal exhaust temperatures.

Regeneration strategies that depend heavily on the participation of nitrogen oxide (NO_x) can require a minimum ratio of NO_x to soot in the engine-out exhaust. As NO_x is also a regulated pollutant, downstream NO_x abatement processes may be necessary to comply with emissions standards. As fuel efficiency, soot production, and NO_x production cannot be independently set by engine design and operation, close coordination is necessary between engine and aftertreatment engineers to achieve a system that balances competing factors.

The rate of soot oxidation (mass per time) at a given temperature and exhaust composition is proportional to the total mass of soot in the filter. A "balance point" is a set of conditions where the rate of new soot accumulation is equal to the rate of soot oxidation. Soot oxidation rates also depend on the extent of soot oxidation because of changes in surface area and composition. Initial soot oxidation rates are typically higher, possibly because of the presence of reactive functional groups and semivolatile organics. Oxidation rates drop off after the initial stages of oxidation and then increase again as shrinking particles result in higher surface areas (Messerer, Niessner, and Poschl, 2006). Oxidation rates and behavior of soot during oxidation can also vary significantly depending on fuel and engine operating conditions.

3.4 Catalysts

Catalysts to promote oxidation of soot have played an important role in DPF systems, and their continuing development is an active research area. Most approaches to the application of catalysts in DPFs fall within one of two broad categories:

- Direct, or contact catalysis, which accelerates oxidation where physical contact exists between the soot and catalytic sites
- Indirect catalysis, where oxidation is mediated by an active gaseous species, namely NO_2.

An early approach to direct catalysis of soot oxidation was to introduce the catalyst in the fuel (Konstandopoulos *et al*., 2007), resulting in small catalytic particles embedded within the soot particles. Although effective in accelerating oxidation rates, fuel-born oxidation catalysts have not found common application.

Contact catalysts such as ceria have an oxygen storage function, making molecular oxygen available for reactions with soot very close to the catalyzed surface. Direct oxidation catalysts have shown great promise in a number of laboratory studies (Southward and Basso, 2008), but maintaining a consistent degree of soot/catalyst contact is a major challenge in field applications. Studies have shown that soot loaded onto a catalyzed monolith reacts much more slowly than soot mixed with a powdered sample of the same direct oxidation catalyst.

Indirect catalysis has been used extensively to promote passive oxidation of soot by NO_2. Catalytic function is similar to that provided by diesel oxidation catalysts (DOCs). Less reactive NO is converted to the more reactive NO_2. Back diffusion of NO_2 from a catalytic site within the filter wall upstream to a soot deposit means that a single molecule of NO_x can be "recycled" multiple times to participate in soot oxidation (Vlachos *et al*., 2008). It should be noted that oxidation catalysts in the DPF can also play an important role in oxidizing CO and hydrocarbons, which are themselves regulated pollutants. In this way, a catalyzed DPF can augment functions of an upstream DOC, as well as oxidizing CO, which might be produced by partial oxidation of soot in the filter itself.

3.4.1 Ash

While the majority of the PM is consumed by oxidation and removed with the gas phase during DPF regeneration,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a small fraction [0.5–1.0% (Sappok and Wong, 2010)] stays behind as ash. Ash accumulation is a major design consideration for DPF systems. At 150,000 miles, ash may account for as much as 80% of the particulate mass trapped in a DPF (Sappok and Wong, 2010). DPF ash originates from a variety of sources including engine wear and corrosion, trace impurities in the fuel, and lubricant additives. Ash derived from the lubricant is thought to comprise the majority of the mass accumulated and is composed primarily of oxides, sulfates, and phosphates of metals including Ca, Zn, and Mg (Sappok *et al*., 2009).

Observations suggest that ash initially forms a layer on filter walls, but then is subsequently transported to the back of inlet channels, where it forms a porous plug. Ash deposits act to increase filter backpressure by obstructing surface pores and taking up volume in the inlet channels. However, ash deposits are very porous (Sappok *et al*., 2009) (roughly 85% or higher void fraction) and may have much higher permeability than soot or the underlying filter substrate (Young *et al*., 2004). A thin coating of ash on the filter wall can actually result in lower backpressures during subsequent soot loading cycles and limit the hysteric effects discussed earlier by reducing penetration of soot into pore throats (Zarvalis, Lorentzou, and Konstandopoulos, 2009; Sappok *et al*., 2009). Higher ash loadings, however, can dramatically increase the sensitivity of backpressure to soot loading level as the ash comes to occupy a significant fraction of the inlet channel volume (Sappok and Wong, 2010). The presence of ash can also significantly improve filtration efficiency over an ash-free filter (Zarvalis, Lorentzou, and Konstandopoulos, 2009).

3.5 DPF substrates

Table 1 lists properties for several example DPF products. Note that the manufacturers can vary properties such as porosity and mean pore size over a considerable range by tuning manufacturing parameters, and development of new products and processes continues.

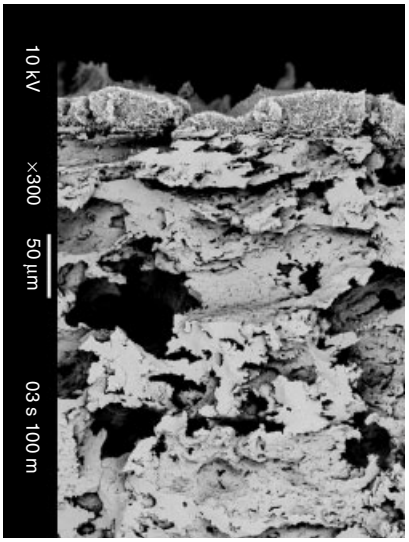


Figure 7. Cordierite filter, lightly loaded with diesel soot.

Cordierite (Figure 7) is possibly the most important ceramic substrate for current DPFs. It is favored by OEMs, especially in the United States, because of its relatively low cost. Cordierite is less susceptible to thermal shock than some competing materials, but it also has a lower melting point. Pore size distributions tend to be somewhat broader than in sintered granular substrates, such as SiC, and the connectivity of the pores is not as uniform.

SiC DPFs (Figure 8) have been deployed mostly in Europe and Japan. SiC filters have a higher melting point than cordierite, but are much more sensitive to thermal shock. This limitation is overcome by manufacturing the DPF bricks in multiple longitudinal segments, which are

Table 1. Examples of ceramic DPF products with properties.

Example Product	Composition	Chemical Formula	Porosity (%)	CTE (1/°C)	Melting Temperature (°C)	Mean Pore Size (µm)
Corning DuraTrap® AC	Cordierite	$2MgO-2Al_2O_3-5SiO_2$	50[a]	5E−7[a]	1450[a]	19[a]
Ibidin	Silicon carbide	SiC	42[b]	4E−6[b]	2200[b] (sublimation)	8.7[b]
Corning DuraTrap® AT	Aluminum titanate	Al_2TiO_5	50[c]	1E−6[c]	1600[c]	15[c]
Dow Aerify™	Acicular mullite	$3Al_2O_3-2SiO_2$		5E−6 (600°C)[d]	>1600[d]	

[a] Corning (2010a).
[b] Ohno *et al*. (2000).
[c] Corning (2010b).
[d] Dow (2011).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

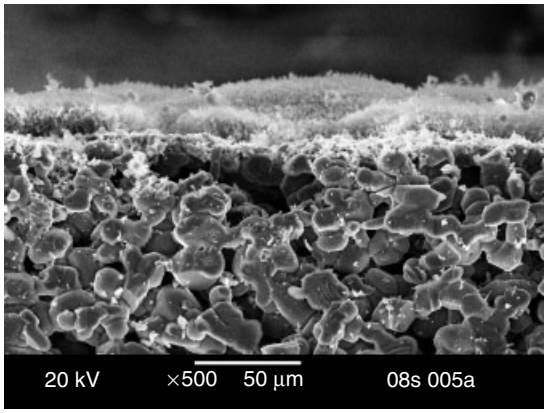


Figure 8. SiC filter, lightly loaded with particulate.

then cemented together with materials that can absorb thermal strain.

A number of porous ceramic substrates have been explored for use in DPFs, some of which are at various stages of commercialization. Aluminum titanate was introduced as an alternative to SiC filters. This material has a significantly higher melting point than cordierite. It is also more robust with respect to thermal shock than SiC, making segmented construction unnecessary. Acicular mullite (Li *et al*., 2004) has a unique microstructure characterized by narrow, interpenetrating crystals (Figure 9). Crystals that extend from the filter walls into the inlet and outlet channels provide additional surface area for coating and filtration, as well as high average porosity. Thermal shock resistance is somewhat less than in cordierite filters, and large filter bricks may be segmented, although the segments may be larger than those typically used in SiC filters.

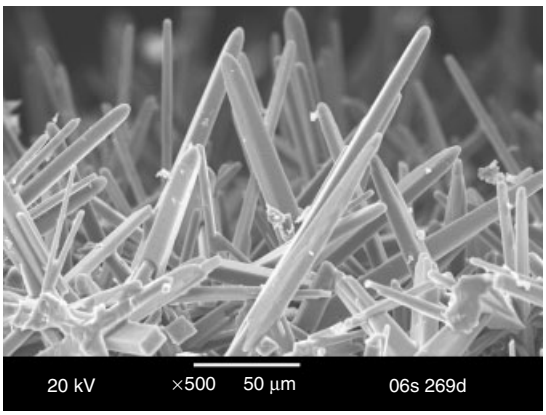


Figure 9. Acicular mullite filter wall surface. This image was taken at Pacific Northwest National Laboratory.

Although ceramic wall-flow monoliths have received most of the attention thus far for meeting stringent particulate emissions limits on new vehicles, many other concepts have been proposed for vehicle exhaust filtration. Partial flow structured metal filters (Emitec, 2009) represent a compromise between no filtration and ceramic wall-flow monoliths. Although they do not have the extremely high mass filtration efficiencies of wall-flow DPFs, they can nevertheless significantly reduce particulate number by capturing the majority of very small particles. As active regeneration is not necessary, such technologies are especially attractive for retrofit applications.

4 INTEGRATION OF PARTICULATE FILTRATION WITH VEHICLE SYSTEMS

4.1 Soot mass estimation, sensing, and control

As previously discussed, accumulation of soot in DPFs leads to an increase in exhaust flow resistance. To alleviate this resistance, it is necessary to regenerate the filter (or oxidize the soot in the filter) periodically when the vehicle is in operation. This can be achieved by either passive regeneration or active regeneration. Passive regeneration is when the soot is oxidized by the NO_2 present in the exhaust stream and occurs at temperatures less than 400°C. Active regeneration, on the other hand, is when the soot is oxidized by excess O_2 and occurs at higher temperatures ($T > 500$°C). In the following section, we discuss techniques to estimate the soot mass in the filter, state-of-the-art soot sensors, and control strategies that are widely used in the emission control industry. We also discuss the consequences of incorrect estimation of the soot that can lead to filter deactivation and/or performance degradation. Various aging mechanisms, owing to both thermal and chemical effects, the effect of ash plugging, and diagnostic methods to adapt the filter to deal such scenarios, are also discussed. Finally, we conclude with a discussion of the impacts of alternate fuels such as biodiesel on DPF performance and with a discussion on the impact of other aftertreatment system components such as DOC and SCR on DPF performance.

4.1.1 Soot mass estimation

Accurate estimation of soot in a DPF is critical for its smooth operation and is one of the key inputs for the filter regeneration control strategy. This is important because underestimating the soot mass could lead to larger accumulation of soot, and on regeneration, the faster soot burn-out

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

rate could lead to filter melting and cracking. Overestimating soot mass results in frequent DPF regenerations, leading to lower fuel economy and a higher CO_2 penalty. Frequently, regenerating the DPF also exposes the filter to excessive thermal cycling that could result in cracking and catalyst deactivation because of cumulative high temperatures exposure.

Various open-loop and closed-loop soot mass estimation strategies are proposed in the literature (Rose and Boger, 2009; Dabhoiwala *et al*., 2008; Benaicha *et al*., 2011; Perrin *et al*., 2004; Su, Gordon, and Hong, 2007). Almost all of these strategies are model based. Open-loop models estimate the engine-out soot as a function of engine speed and load and determine the DPF soot mass from the engine-out calculation. A limitation of such an approach is the ability to map soot emissions during transient engine operating conditions, which is extremely complex and involves extensive parameterization efforts (Rose and Boger, 2009; Wang *et al*., 2011). The closed-loop models, on the other hand, rely on the pressure drop measurement from the sensors upstream and downstream of the DPF, airflow measurements from the flow meter, and temperature measurements upstream of the DPF. The challenge in using a pressure drop measurement for soot estimation is that the pressure drop is a cumulative effect of various phenomena such as friction in the inlet and outlet channels and flow resistance through wall and soot layers, whose individual contribution to the overall pressure drop vary based on the filter operating conditions. Several efforts are ongoing in this area to quantify and develop high fidelity modeling tools, based on engine dynamometer tests, to understand the relationship between the pressure drop and soot mass and its precise location (Dabhoiwala *et al*., 2008).

Advanced state estimation/observer techniques for DPF systems are also reported in the literature (Benaicha *et al*., 2011; Perrin *et al*., 2004; Su, Gordon, and Hong, 2007; Hsieh and Wang, 2011; Surenahalli *et al*., 2012). A majority of these techniques are based on a reduced order DPF model and use linear/nonlinear systems theory to estimate the soot mass and help in OBD. An estimator to estimate the soot mass based on the pressure drop measurement across the filter was developed by Benaicha *et al*. (2011). The estimator uses an extended Kalman filter (EKF) technique and was validated against data collected on both engine dynamometer and vehicle. Motivated by future OBD legislations, studies on detecting filter degradation through model-based algorithms were also reported (Perrin *et al*., 2004).

A reduced order model of the physical system forms the basis of an estimator design. It is important to realize a reduced order model that is mathematically consistent and maintains its accuracy, while being implementable in real time. One advanced realization technique was developed by Su, Gordon, and Hong (2007). The authors used a singularly perturbed sliding manifold approach to reduce a DPF model and then designed a sliding mode observer to estimate the soot mass. Model-based estimators were also developed for DOC to understand the impact of DOC states such as temperature, NO, NO_2, and hydrocarbons (during active regeneration) on DPF performance (Hsieh and Wang, 2011; Surenahalli *et al*., 2012).

4.1.2 Soot (PM) sensors

As the accuracy of the soot mass estimate is affected by the types and combinations of sensors, feedback from PM sensors is vital for soot estimation and regeneration control strategy development, especially in light of the future 2014–2018 fuel consumption regulations (US EPA, 2009). Various types of PM sensors are currently being explored for diesel aftertreatment applications (Moos, 2005; Ochs *et al*., 2010; Roesch *et al*., 2010; Kondo *et al*., 2011; Katsuyama *et al*., 2011; Sappok *et al*., 2010).

PM sensors that are reported in the literature are broadly classified into two categories—(i) sensors that accumulate the soot over a period of time and provide a response and (ii) sensors that provide the soot measurement in real time. Sensors that belong to the first category mostly use the conductometric principles where a good relationship between the sensor resistance and accumulated soot mass measurements (from measuring the pressure difference across the filter) was reported (Moos, 2005). Bosch developed an EGS-PM (exhaust gas sensor for particulate matter) sensor based on similar principles for OBD purposes (Ochs *et al*., 2010; Roesch *et al*., 2010), and a diagnostics concept for DPF state of health monitoring was proposed based on the sensor signal. Soot deposits between the electrodes of a sensing element form an electrical contact, decreasing the resistance. After a predefined current threshold is reached, the heater element within the sensor heats the sensor above typical soot regeneration temperatures ($>600^{\circ}C$) and regenerates the adsorbed soot. The sensor signal, which correlates the exhaust soot concentration, is the time between regeneration and the time when the current reaches a predefined threshold.

A PM sensor capable of detecting lower soot concentrations, typically found during New European Drive Cycle (NEDC) and FTP modes, was developed by NGK insulators (Kondo *et al*., 2011; Katsuyama *et al*., 2011). The sensor operates in three sequential stages including forced collection, detection, and burn-off. In the first stage, soot is forcibly collected using an electric field and a thin

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

soot layer is formed on the surface of the detecting electrode. The second stage involves soot detection where the accumulated soot is determined by measuring electrostatic capacitance changes followed by the third stage when the accumulated soot is burnt off allowing the sensor to return to its initial state.

An example of a sensor that measures the accumulated soot mass in real time is an RF (radio-frequency) sensor developed by Filter Sensing Technologies (Sappok *et al*., 2010) where antennas are mounted upstream and downstream of the filter and an RF signal from the sensor control unit is used to detect the soot in the filter. The characteristics of the signal are influenced by the dielectric properties of the material through which it propagates. While accurate sensor feedback forms a key component in triggering filter regeneration, the other important component is the control strategy that ensures that the filter is regenerated only when required, considering the key filter dynamics discussed in the next section. The RF measurement system was tested on a 1.9-L GM turbocharged diesel engine under a variety of engine operating conditions and advanced combustion modes and has generated encouraging results demonstrating the feasibility for measuring the soot levels directly in the DPF using RF sensing.

4.1.3 Regeneration control strategies

Active regeneration control strategies reported in the literature are mainly based on the temperature and soot mass dynamics in the filter and the temperature and pressure measurements across the filter. Singh *et al*. (2005) proposed a model-based control strategy that filters the high frequency components in DPF inlet temperature and pressure drop to avoid the negative effects associated with responding to high frequency measurements. On the basis of their strategy, the fuel penalty associated with DPF active regeneration was reduced with higher soot mass and higher DPF inlet temperatures. An advanced control strategy for DPF regeneration was proposed by Van Nieuwstadt and Tennison (2006) that includes both feed-forward and feedback components. The *feed-forward* term computes a nominal hydrocarbon (diesel) injection quantity to raise the predetermined DPF inlet temperature based on the oxidation catalyst temperature, whereas the feedback quantity is the output of a PI controller, which takes the error in temperature as the input signal.

Such active regeneration control strategies in DPFs are crucial to realize the future fuel economy targets (US EPA, 2009) for heavy-duty vehicles. In addition, they also help avoid unnecessary regenerations that could affect the health and durability of the filter, discussed in the next section.

4.2 DPF failure modes and diagnostics

A DPF needs to survive its vehicle lifetime and hence ensuring its durability is a critical component in DPF product development. Several filter durability studies and performance evaluations are reported in the literature (Zhan, 2008; Wilcox *et al*., 2004; Watkins *et al*., 2009). This section summarizes the common failure modes in DPFs such as thermal deactivation, chemical poisoning, and cracking. Further, techniques to diagnose the failure modes and adaptive measures are discussed.

4.2.1 Thermal deactivation

The DPF is typically exposed to large exotherms (650–750°C) during active regeneration. Therefore, it is important to understand the impact of such high temperatures on the filter substrate. For example, Wilcox *et al*. (2004) investigated thermal stresses on the DPF and explored the likelihood of cracking, given the part geometry, material properties, temperature distribution, and failure criterion. A 3D DPF model was developed using Ansys, and a continuum approach was followed accounting for anisotropy, core, plugs, and skin. The authors report that the thermal stress analysis was successfully used to develop base materials and coatings, to impose the max soot mass criterion for failure, and for durability studies. Similar analysis tools were developed by Watkins *et al*. (2009) using nondestructive evaluation (NDE) techniques (applied to DPF ceramic substrates), to compare the properties of field-aged DPFs with virgin filters.

4.2.2 Impact of aging

DPFs age because of both thermal and chemical effects (DaCosta, Shannon, and Silver, 2006). One of the serious issues with DPF performance degradation is filter aging because of ash. Ash particles, which originate from the lubricant oil accumulate in the filter, may cause false triggers for filter regeneration. A rapid ash aging technique that evaluates the ash-aged filter samples as a function of soot loading behavior, regeneration, and filter efficiency was developed (Zarvalis, Lorentzou, and Konstandopoulos, 2009). While ash layer caused a significant increase in filtration efficiency, it has been concluded that the impact of ash depends on the catalyst layer morphology.

4.2.3 Diagnostic methods

To meet the future OBD legislations, state of health of DPF systems needs to be constantly monitored. A survey of various diagnostic methods for DPFs was presented

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

by Mohammadpour, Franchek, and Grigoriadis (2011). The diagnostic method based on pressure drop measurement across the filter was found to be of limited capability because of sensor noise not measured by the engine control unit. Suggested by the California Air Resources Board, researchers explored various signal processing and statistical analysis techniques to correlate the pressure drop measurement to model-based DPF diagnostics (Cunningham, Meckl, and Shah, 2007; Van Nieuwstadt and Brahma, 2008). A DPF fault detection algorithm based on energy spectral analysis obtained through temperature and pressure measurements pre- and post-DPF was developed by Surve (2008). The approach was successfully implemented to detect failure in a lightly failed DPF (small crack), and the dynamic pressure signal analysis was extended to transient engine operations as well. A similar model-based failure detection technique was discussed (Gupta *et al*., 2011), which uses orthogonal least squares method to estimate the model coefficients based on the dynamic pressure drop signal. By comparing the coefficients as a function of time, the filter failure can be detected. This approach does not require additional sensors, is robust to sensor noise and process variability, and was validated using FTP data in a test cell for both healthy and failed filters.

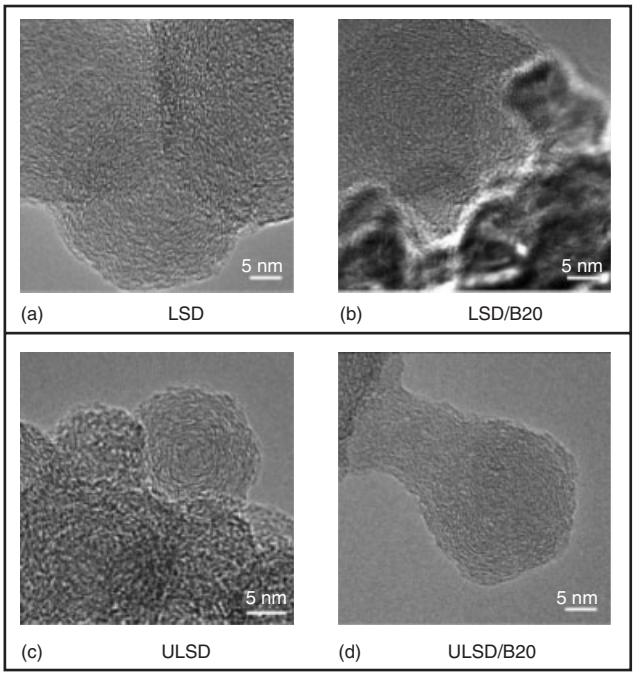


Figure 10. (a–d) HR-TEM Images of soot nanostructure showing the effect of biodiesel blending for four test fuels. As mentioned by (Boehman *et al*., 2005), the soot particles from B20 blending in ULSD have a less ordered structure. (Reprinted with permission from Boehman *et al*. 2005. Copyright (2005) American Chemical Society.)

4.3 Effect of alternate fuels on DPF performance

Biodiesel is attractive for its renewable properties and reduces dependence on foreign oil. The US EPA conducted a comprehensive analysis of biodiesel on vehicle exhaust emissions in 2002 based on statistical regression analysis, correlating the biodiesel content in diesel fuel to regulated and unregulated pollutants (US EPA, 2002). Biodiesel consisting of mono alkyl esters is derived from various vegetable oils; it reduces tail-pipe PM, CO, and HC emissions whereas increases NO_x emissions. Recent EPA analysis indicates that PM emissions are reduced by as much as 12% using the B20 blend and by 47% using B100 (US EPA, 2002). This section introduces the impact of alternate fuels such as biodiesel on DPF performance.

A comprehensive review on the impact of biodiesel fuels on engine emissions with specific focus on PM and NO_x was reported in Lapuerta, Armas, and Fernandez (2008) and Szybist *et al*. (2007). This body of work gives a big picture comparison between the conventional diesel fuel and the biodiesel blends. In particular, there has been a lot of work in characterizing the soot from biodiesel exhaust. Boehman, Song, and Alam (2005) observed that the soot generated from biodiesel blend (B20) varies in nanostructure and oxidative reactivity as shown in Figure 10. They concluded that the amorphous structure and higher oxygen content in biodiesel soot (B20) are partly responsible for soot reactivity.

The break-even temperature (BET) or balance point temperature (BPT) is defined as the temperature at which soot deposition is equal to the oxidation rate. This equilibrium point is typically lower for biodiesel when compared to conventional diesel fuel, indicating that it may improve the passive regeneration process in the DPF (Buono, Senatore, and Prati, 2011). In this report, the DPF regeneration procedure was reported as different for the biodiesel fuel and needs to be recalibrated and optimized for better fuel economy. These findings were supported by the findings from Williams *et al*. (2006) who concluded that lower BET temperatures and higher regeneration temperatures for biodiesel were due to higher soot reactivity in biodiesel fuels. Vehicle tests conducted on a 2006 MY 2.8 L, 120 kW VM Motori DI diesel engine with a DOC and CPF or cDPF (catalyzed particulate filter) aftertreatment system using B20 fuel indicated lower regeneration temperatures ($\sim 33^{\circ}$C) and higher loading times because of less engine-out PM when compared to ULSD fuel (Peterson *et al*., 2009). This could be advantageous from a fuel

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

economy standpoint, decreasing the number of active regenerations. However, constant monitoring of the biodiesel dilution in engine oil is required to prevent engine deterioration. Even though NO_2 concentrations upstream of the DPF increased by 6% (due to higher engine-out NO_x from biodiesel fuels), lower regeneration temperatures were attributed to higher soot reactivity for biodiesel blends.

Owing to the potential for residual alkali and alkaline earth metals in biodiesel to form ash, it is critical to understand the impact of biodiesel on ash emissions (Sappok and Wong, 2007). Researchers concluded that the presence of metals might impact ash loading and catalyst performance and noticed increased debris and metals in biodiesel ash and attributed to solvent properties of the fuel.

4.4 Systems integration and interaction

A typical 2010 aftertreatment system for heavy-duty diesel engines consists of an oxidation catalyst, an NO_x reduction catalyst, and a DPF as shown in Figure 11.

In addition to the transients imposed by engine operating conditions in terms of flow rate, temperature, and concentrations, the reductants such as diesel and urea add interesting dynamics to the overall aftertreatment system. Of particular interest are the DOC and SCR dynamics that would impact the performance of DPF. Selective catalytic reduction (SCR) is a widely used technology to reduce NO*x* emissions from engine exhaust through injection of a reductant such as urea solution, ammonia, or hydrocarbons upstream of the SCR catalyst.

4.4.1 Impact of DOC and SCR on DPF

The DOC that oxidizes hydrocarbons, NO and CO on the active platinum sites, is important for passive regeneration of soot in the DPF, as it converts NO to the more active oxidizer NO_2. As the catalyst ages and the oxidation conversion efficiency of DOC decreases, it would have a direct impact on the DPF performance, especially during a heavy-duty scenario (as shown in Figure 11) where DPF is located right after the DOC. If the DPF is catalyzed (also called as *CPF* or *cDPF*), then the impact of aged DOC could be compensated by the catalyst on the filter to a certain extent. It is also reported that DOC depletes the organic compounds in the soot (Lizarraga *et al*., 2011).

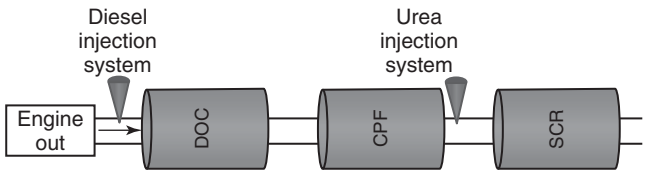


Figure 11. A schematic of a typical 2010 aftertreatment system for heavy-duty diesel engines with an oxidation catalyst, diesel particulate filter, and a urea-SCR catalyst.

Under an LD scenario where active regeneration in DPF is much more frequent, DPF is placed downstream of the SCR catalyst. Here, the impacts of both the DOC and SCR catalysts on the filter should be considered. Passive regeneration in DPF is not as important in this case as NO and NO_2 are consumed in the SCR catalyst. As a result, the DPF has to heavily rely on active regeneration to burn the accumulated soot. However, when the SCR catalyst ages, the NO oxidation (especially for base metal zeolites) and the SCR capability reduce and there could be some benefit to the DPF for achieving passive regeneration. On the downside, oxidation of NH_3 from the SCR process decreases because of aging and the DPF might experience clogging problems because of NH_3 and other urea derivatives from the SCR such as HNCO (isocyanic acid), occupying the filter pores.

4.4.2 Integrated DPF/SCR (2-way) devices

Urea-SCR catalysts and DPFs are proved technologies in reducing the NO_x and particulate emissions from diesel engines, respectively. These are the technologies at the forefront to meet the HD emission standards of NO_x (0.2 g/bhp-h) and PM (PM—0.01 g/bhp-h). From an exhaust aftertreatment standpoint, a minimum of three catalytic converters, that is, an oxidation catalyst, a particulate filter, and an NO_x reduction catalyst, are necessary to meet these standards. To decrease the overall volume of the catalysts and the costs of the precious metals in the catalysts to mitigate these harmful pollutants (CO, HC, NO_x, and PM) in the exhaust, researchers have been investigating an integrated NO_x/PM aftertreatment system on a single substrate. One such technology involves the integration of urea-SCR and DPF in a single substrate. The following paragraph summarizes the current efforts in developing and deploying such an integrated device.

The overarching goal of these efforts has been to develop an integrated device that reduces the overall aftertreatment volume while maintaining optimal NO_x conversion efficiency with minimal backpressure. Tan, Solbrig, and Schmieg (2011) have evaluated a combined SCR/DPF system to meet 2010 and beyond EPA/CARB regulations that reduces SCR catalyst volume, improves vehicle packaging, and reduces cost. Experiments were carried out on a Euro IV compliant V6 diesel engine operating on an engine dynamometer and on a fixed bed reactor in the laboratory. The system was able to meet the HD dynamometer and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

chassis certification requirements with significant reduction in overall aftertreatment catalyst volume. The impact of EGR on the SCR/DPF system was also evaluated and compared to a conventional DOC-CSF-SCR system (Naseri *et al*., 2011). Cu/Z catalyst was coated on a high porosity filter similar to many other reports found in the literature (Guo *et al*., 2010). With EGR system turned off, SCR/DPF systems performed better than the conventional because of higher SCR catalyst volume that allows flexibility for HD engines to operate at higher out NO_x. SCR/DPF systems showed better light-off and demonstrated better passive regeneration capabilities (due to higher engine-out NO_x and hence high NO_2) resulting in lesser active regeneration events and hence better fuel economy. On the LD side, Boorse *et al*. (2010) tested SCR/DPF system on a 2.7 L V6 Land Rover to achieve T2B5 standards for LD trucks. Two systems, DOC-SCR-SCRF (SCR on filter) and DOC-SCRF-SCR, were tested (washcoat loading on SCRF was 60% of the washcoat loading on a flow-through SCR). A combo cycle involving FTP-75, HWFET, and US06 was tested to evaluate the two systems. DOC-SCR-SCRF showed higher NO_x conversion efficiency because of high washcoat loading allowing higher NH_3 storage. Oven aging tests indicate that the DOC-SCRF-SCR was able to meet the T2B5 tailpipe NO_x and TMHC standards through 120 k miles. Alternate DPF substrates such as SiC have also been tested for the integrated device and reports indicate better SCR catalyst utilization because of its microstructure (Boorse *et al*., 2010).

Even though a few reports indicate a minor effect of soot on NO_x conversion (Bush, Iretskaya, and Tadrous, 2009), there exist significant design challenges in catalyst coating process optimization and washcoat loading to develop an optimal, integrated SCR/DPF system for simultaneous NO_x and PM conversion. In addition, bench reactor evaluations and modeling tools (Park, 2011) are invaluable in understanding the impact of species diffusion and competitive adsorption in the integrated system and could lead to a better SCR/DPF system design.

REFERENCES

Benaicha, F., Bencherif, K., Sorine, M., and Vivalda, J.C. (2011) Model based mass soot observer of diesel particulate filter. *Proceedings of the 18th IFAC World Congress*, pp. 10647–10652.

Boehman, A., Song, J., and Alam, M. (2005) Impact of biodiesel blending on diesel soot and the regeneration of particulate filters. *Energy and Fuels*, **19**, 1857–1864.

Boorse, R.S., Dieterle, M., Voss, K., *et al.* (2010) Two in One—SCR on Filter. *Directions in Engine Efficiency and Emissions Research Conference*.

Buono, D., Senatore, A., and Prati, M.V. (2011) Particulate filter behavior of a diesel engine fueled with biodiesel. *Applied Thermal Engineering*, DOI: 10.1016/j.applthermaleng.2011.08.019 (Article in Press).

Bush, P., Iretskaya, S., and Tadrous, T. (2009) Investigation on Continuous Soot Oxidation and NOx Reduction by SCR Coated DPF. *Directions in Engine Efficiency and Emissions Research Conference*.

Corning (2010a) Clean Diesel Made Possible, Corning DuraTrap AC Filters (product brochure).

Corning (2010b) Clean Diesel Made Possible, Corning DuraTrap AT Filters (product brochure).

Cunningham, P., Meckl, P., and Shah, C. (2007) Correlating dynamic pressure signal features to diesel particulate filter load. SAE Technical Paper 2007-01-0333.

Dabhoiwala, R., Johnson, J.H., Naber, J., and Bagley, S. (2008) A methodology to estimate the mass of particulate matter retained in a catalyzed particulate filter as applied to active regeneration and on-board diagnostics to detect filter failures. SAE Technical Paper 2008-01-0764.

DaCosta, H., Shannon, C.M., and Silver, R. (2006) Durability of Diesel Particulate Filters—Bench Studies on Cordierite Filters. *Diesel Exhaust Emissions Research Conference*.

Dow (2011) Aerify Diesel Particulate Filters (product brochure).

Elimelech, M. (1995) Particle Deposition and Aggregation: Measurement, Modelling, and Simulation, Butterworth-Heinemann, Oxford; Boston, MA.

Emitec (2009) Partial flow diesel particulate filters (P-DPF) are easy to retrofit. Press Release, 63rd IAA.

Geankoplis, C.J. (1993) Transport Processes and Unit Operations, PTR Prentice Hall, Engelwood Cliffs, NJ.

Guo, G., Warner, J., Cavataio, G., *et al.* (2010) The development of advanced urea-SCR systems for tier 2 bin 5 and beyond diesel vehicles. SAE Technical Paper 2010-01-1183.

Gupta, A., Franchek, M., Grigoriadis, K., and Smith, D.J. (2011) Model Based Failure Detection of Diesel Particulate Filter. *American Control Conference*.

Hsieh, M. and Wang, J. (2011) NO and NO_2 concentration modeling and observer based estimation across a diesel engine aftertreatment system. *ASME Journal of Dynamic Systems, Measurement and Control*, **133** (4) paper 041005.

Katsuyama, K., Kitoh, K., Sakuma, T., *et al.* (2011) Particulate Matter Sensor, United States Patent 7,977,955, July 12, 2011.

Kondo, A., Yokoi, S., Sakurai, T., *et al.* (2011) New particulate matter sensor for on board diagnosis. SAE Technical Paper 2011-01-0302.

Konstandopoulos, A. and Johnson, J.H. (1989) Wall-flow diesel particulate filters - their pressure drop and collection efficiency. SAE 890405.

Konstandopoulos, A.G., Kostoglou, M., Lorentzou, S., *et al.* (2007) Soot oxidation kinetics in diesel particulate filters. SAE 2007 World Congress, 2007-01-1129.

Konstandopoulos, A.G., Kostoglou, M., Skaperdas, E., *et al.* (2000) Fundamental studies of diesel particulate filters: transient loading, regeneration and aging. SAE Technical Paper Series 2000-01-1016.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Konstandopoulos, A.G., Skaperdas, E., and Masoudi, M. (2002) Microstructural properties of soot deposits in diesel particulate traps. SAE 2002-01-1015.

Lapuerta, M., Armas, O., and Fernandez, J. (2008) Effect of biodiesel fuels on diesel engine emissions. *Progress in Energy and Combustion Science*, **34** (2), 198–223.

Li, C.G., Mao, F., Swartzmiller, S.B., *et al.* (2004) Properties and performance of diesel particulate filters of an advanced ceramic material. SAE World Congress, 2004-01-0955.

Lizarraga, L., Souentie, S., Boreave, A., *et al.* (2011) Effect of diesel oxidation catalysts on the diesel particulate filter regeneration process. *Environmental Science and Technology*, DOI: dx.doi.org/10.1021/es2026054 (Article in Press).

Masoudi, M. (2002) Hydrodynamics of diesel particulate filters. SAE 2002 World Congress, 2002-01-1016.

Merkel, G.A., Cutler, W.A., Tao, T., *et al.* (2003) New Cordierite Diesel Particulate Filters for Catalyzed and Non-Catalyzed Applications. *9th Diesel Engine Emissions Reduction Conference*, Newport, Rhode Island.

Messerer, A., Niessner, R., and Poschl, U. (2006) Comprehensive kinetic characterization of the oxidation and gasification of model and real diesel soot by nitrogen oxides and oxygen under engine exhaust conditions: measurement, Langmuir-Hinshelwood, and Arrhenius parameters. *Carbon*, **44**, 307–324.

Miyairi, Y., Noguchi, T., Hiramatsu, T., *et al.* (2006) Diesel particulate filter (DPF) trapping efficiency improvement under no-soot-layer conditions obtained through pore size distribution optimization. SAE 2006-01-1528.

Mohammadpour, J., Franchek, M., and Grigoriadis, K. (2011) A Survey on Diagnostic Methods for Automotive Engines. *American Control Conference*.

Moos, R. (2005) A brief overview on automotive exhaust gas sensors based on electroceramics. *International Journal of Applied Ceramic Technology*, **2** (5), 401–413.

Naseri, M., *et al.* (2011) Development of SCR on diesel particulate filter systems for heavy duty applications. SAE Technical Paper 2011-01-1312.

Ochs, T., Schittenhelm, H., Genssle, A., and Kamp, B. (2010) Particulate matter sensor for on board diagnostics (OBD) of diesel particulate filters. SAE Technical Paper 2010-01-0307.

Ohno, K., Shimato, K., Taoka, N., *et al.* (2000) Characterization of Sic-Dpf for passenger car. SAE 2000 World Congress, 2000-01-0185.

Park, S.Y. (2011) Development and Validation of a Model for 2-way DPF/SCR. *Cross-Lean Exhaust Emissions Reduction Simulation (CLEERS) Workshop*.

Perrin, O., Basseville, M., Sorine, M., and Zhang, Q. (2004) On-Board Diesel Particulate Filter Fault Detection using an Adaptive Observer. *IFAC Proceedings Series*, pp. 367–372.

Peterson, A., Lee, P., Lai, M., *et al.* (2009) Impact of biodiesel emission products from a multi cylinder direction injection diesel engine on particulate filter performance. SAE Technical Paper 2009-01-1184.

Premchand, K.C., Johnson, J., Yang, S., *et al.* (2007) A study of the filtration and oxidation characteristics of a diesel oxidation catalyst and a catalyzed particulate filter. SAE 2007-01-1123, 31-57.

Roesch, S., Ochs, T., Kamp, B., and Schittenhelm, H. (2010) Sensor Element for Particle Sensors and Method for Operating Same, United States Patent 7,770,432, Aug 10, 2010.

Rose, D. and Boger, T. (2009) Different approaches to soot estimation as key requirement for DPF applications. SAE Technical Paper 2009-01-1262.

Sappok, A. and Wong, V. (2007) Impact of Biodiesel on Ash Emissions and Lubricant Properties Affecting Fuel Economy and Engine Wear. *Diesel Exhaust Emissions Research Conference*.

Sappok, A. and Wong, V.W. (2010) Ash effects on diesel particulate filter pressure drop sensitivity to soot and implications for regeneration frequency and DPF control. SAE World Congress, 2010-01-0811.

Sappok, A., Parks, J. II,, and Prikhodko, V. (2010) Loading and regeneration analysis of a diesel particulate filter with a radio frequency based sensor. SAE Technical Paper 2010-01-2126.

Sappok, A., Santiago, M., Vianna, T., and Wong, V.W. (2009) Characteristics and effects of ash accumulation on diesel particulate filter performance: rapidly aged and field aged results. SAE World Congress 2009-01-1086.

Singh, N., Johnson, J.H., Parker, G., and Yang, S.L. (2005) Vehicle engine aftertreatment system simulation model: application to a controls design strategy for active regeneration of a catalyzed particulate filter. SAE Technical Paper 2005-01-0970.

Solomon, S., Qin, D., Manning, M., *et al.* (eds) (2007) *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press.

Southward, B.W.L. and Basso, S. (2008) An investigation into the NO_2-mediated decoupling of catalyst to soot contact and its implications for catalysed DPF performance. SAE World Congress, 2008-01-0481.

Su, F., Gordon, B., and Hong, H. (2007) A Sliding Mode Observer for a Typical Diesel Engine Particulate Aftertreatment System. *Proceedings of the IEEE International Conference on Mechatronics and Automation*, pp. 1184–1189.

Surenahalli, H., Parker, G., Johnson, J.H., and Devarakonda, M. (2012) A Kalman Filter Estimator for a Diesel Oxidation Catalyst During Active Regeneration of a CPF. *American Control Conference*.

Suresh, A., Khan, A., and Johnson, J.H. (2000) An experimental and modeling study of cordierite traps - pressure drop and permeability of clean and particulate loaded traps. SAE Technical Paper Series 2000-01-0476.

Surve, P. (2008) Diesel particulate filter diagnostics using correlation and spectral analysis. Master thesis. Purdue University.

Szybist, J.P., Song, J., Alam, M., and Boehman, A. (2007) Biodiesel combustion, emissions and emission control. *Fuel Processing Technology*, **88** (7), 679–691.

Tan, J., Solbrig, C., and Schmieg, S. (2011) The development of advanced 2-way SCR/DPF systems to meet future heavy duty diesel emissions. SAE Technical Paper 2011-01-1140.

Tandon, P., Heibel, A., Whitmore, J., *et al.* (2010) Measurement and prediction of filtration efficiency evolution of soot loaded diesel particulate filters. *Chemical Engineering Science*, **65**, 4751–4760.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Thalagavara, A.M., Shende, A.S., Johnson, J., *et al.* (2005) The effects of two catalyzed particulate filters on exhaust emissions from a heavy duty diesel engine: filtration and particulate matter oxidation characteristics studied experimentally and using a 1- D 2- layer model. SAE 2005 World Congress, 2005-01-0950.

US EPA (2002) A Comprehensive Analysis of Biodiesel Impact on Exhaust Emissions, EPA420-P-02-001.

US EPA (2009) Control of Air Pollution from New Motor Vehicle Engines.

Van Nieuwstadt, M. and Brahma, A. (2008) Uncertainty analysis of model based diesel particulate filter diagnostics. SAE Technical Paper 2008-01-2648.

Van Nieuwstadt, M. and Tennison, P. (2006) Control System and Method for Diesel Particulate Filter Regeneration. United States Patent 7,047,729, May 23, 2006.

Vlachos, N., Patrianakos, G.N., Kostoglou, M., and Konstandopoulos, A. (2008) Micro-simulation of NO-NO_2 transport and reaction in the wall of a catalyzed diesel particulate filter. 2008-01-0442.

Wang, D., Liu, Z., Han, Y., *et al.* (2011) Experimental Studies on Pressure Drop Performance and Regeneration Safety of Diesel Particulate Filter. *Proceedings of the IEEE Conference on Electric Information and Control Engineering*, pp. 2175–2178.

Watkins, T., Shyam, A., Lin, H.T., *et al.* (2009) Durability of Diesel Particulate Filters. DOE-VT Annual Merit Review and Peer Evaluation Meeting.

Wilcox, D., Aniolek, K., Parsamian, G., and Blauvelt, J. (2004) Predicting Thermal Stress in Diesel Particulate Filters. *Diesel Exhaust Emissions Research Conference*.

Williams, A., McCormick, R., Hayes, R., *et al.* (2006) Effects of biodiesel blends on diesel particulate filter performance. SAE Technical Paper 2006-01-3280.

Young, D.M., Hickman, D.L., Bhatia, G., and Gunasekaran, N. (2004) Ash storage concept for diesel particulate filters. SAE World Congress, 2004-01-0948.

Zarvalis, D., Lorentzou, S., and Konstandopoulos, A.G. (2009) A methodology for the fast evaluation of the effect of ash aging on the diesel particulate filter performance. SAE World Congress, 2009-01-0630.

Zhan, R. (2008) *DPF Durability*. SwRI Symposium.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto131
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fundamentals

Mark P.B. Musculus[1], Lyle M. Pickett[1], and Sebastian A. Kaiser[2]

[1] *Sandia National Laboratories, Livermore, CA, USA*
[2] *Universität Duisburg-Essen, Duisburg, Germany*

1 Optical Engines and Other Facilities 1
2 Optical Diagnostic Techniques 7
References 19
Further Reading 21

1 OPTICAL ENGINES AND OTHER FACILITIES

1.1 Optical engines

An "optical engine" is designed with transparent windows to allow processes occurring inside the cylinder of the running engine to be viewed. A normal production metal engine may be modified for optical access, or a stand-alone optical engine may be specifically designed for the task. With optical access, not only can the cylinder contents be passively observed, but laser-based measurement techniques may also be applied to extract specific information about the physical and chemical in-cylinder processes occurring rapidly in the harsh, high pressure, high temperature environment inside the engine.

Optical access to combustion chambers can be created in different ways and to different extents. Small-scale (a few millimeters) windows can accommodate laser beam input and endoscopic viewing through the cylinder head or cylinder wall, whereas large-scale (tens of millimeters or more) windows can provide access to the whole cylinder liner (using a toroid window) or the piston-top (using a disk window) for full-field imaging. In keeping with the generally accepted meaning of optical engine within the engine research community, this section describes engines with large-scale optical access that requires significant engine modifications. In some optical engines, the windows alter the engine geometry somewhat (such as using flat windows to avoid optical distortion), and in other engines, the windows are contoured to conform to the production engine geometry, although image postprocessing may be required to correct for optical distortions, as described later.

Optical engines are typically single-cylinder versions of multicylinder production engines. A typical example of a modern design with optical access through the cylinder liner, piston crown, and cylinder head is shown schematically in Figure 1. The key features are as follows:

- Optical access through a spacer ring with flush mounting, curved spacer-ring windows. Typical window materials are fused silica (excellent UV light transmission) or sapphire (ultrahard and strong).
- Correspondingly lowered piston ring pack keeping the rings from sliding over the spacer-ring windows (and creating a large ring-land crevice volume and lower compression ratio).
- Piston rings from self-lubricating polymer mixtures to allow operating the upper engine nearly oil free. Oil would greatly contribute to window fouling.
- Optical access through a window in the piston crown, which forms the top part of an extended piston with a tall vertical slot for viewing access ["Bowditch" piston (Bowditch, 1961)]. A range of compression

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

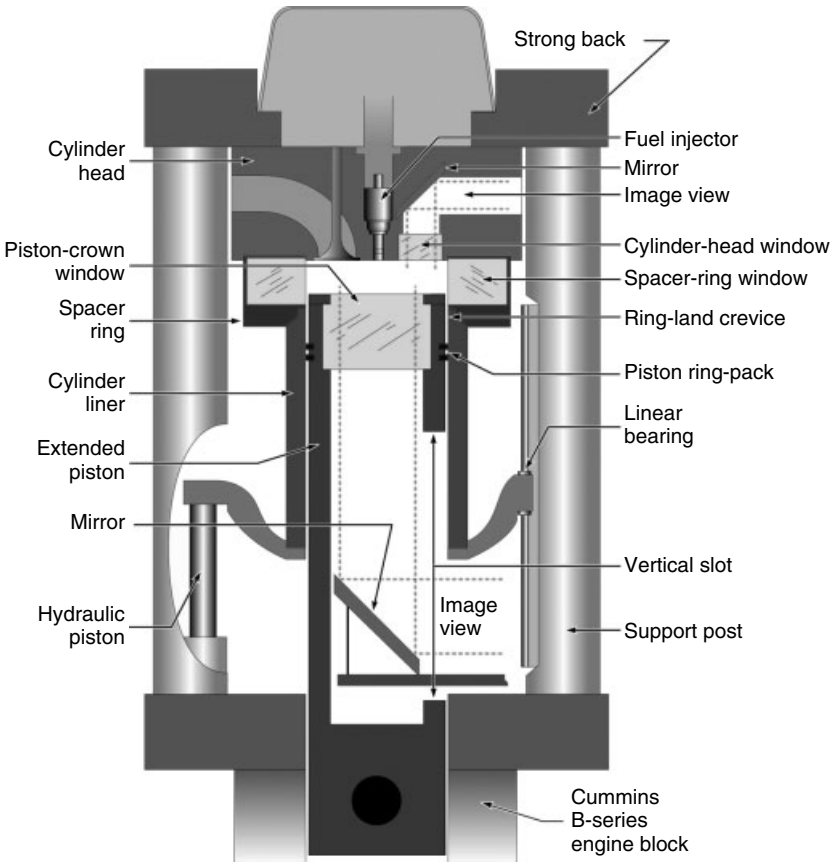


Figure 1. Schematic example of an optically accessible engine with piston extension (and cylinder-head window illustrated). All key components of the upper engine are shown. The base engine block (here a Cummins B-series) is below, with its optical piston extending upward into the optically accessible part of the engine. (Adapted with permission from Dec, Hwang and Sjöberg, 2006. © J.E. Dec, W. Hwang, and M. Sjöberg.)

ratios and combustion chamber configurations can be realized by installing pistons of different dimensions and geometries.

- A stationary mirror at 45°, inserted from the side into the piston slot, allowing upward viewing through the piston window at all times in the cycle.
- Additional small-scale windows and mirror in the engine head, for example, inserted in place of one of the exhaust valves. Figure 11 originates from such an arrangement.
- A means of quickly separating the engine head from the cylinder liner to allow for cleaning the inside surfaces of windows. Here, the cylinder liner can be lowered and raised hydraulically. Other solutions exist.
- A stiff connection of the base engine's crankcase to the cylinder head, which is elevated because of the extended piston. In Figure 1, four massive support posts are bolted to a "strong back" that holds the cylinder head down against the cylinder pressure and upward force from the hydraulic pistons that hold the liner tight against the cylinder head.
- A base engine providing the piston kinematics and forces. For the engine in Figure 1, the first cylinder of a suitably modified six-cylinder block is used. More typical bases are robust single-cylinder research engines.

Many variations of the concept outlined earlier are possible. Perhaps, the most important ones pertain to the optical ports. Compared to Figure 1, other designs have the following:

- A short transparent ring instead of the metal spacer with small windows. This large-scale window option gives 360° optical access to the top 10–20% of the combustion chamber but limits the safe peak pressure to about 50–100 bar.
- A full-height transparent liner with the piston rings sliding over the liner. This configuration provides 360°

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

optical access to the entire piston stroke, but it is much more fragile than other designs and is almost exclusively used in motored operation only.
- A shaped piston-crown window. For example, in a diesel engine, a bowl with a complex geometry may be machined into a suitably mounted glass crown. Such a feature may be necessary to investigate salient aspects of in-cylinder phenomena, but the accompanying optical distortions make imaging more difficult.
- No piston-crown window. Particularly, for laser-based imaging, this not only severely limits the possible configurations but also obviates the need for a piston extension.

Owing to the optical access modifications, the many aspects of optical engine performance differ from those of a similar production engine. Replacing liquid-cooled metal components with largely uncooled windows having different material properties alters the heat transfer between walls and the gas phase. In addition, oil-free polymer piston rings can have much more blow-by than their production metal counterparts. Gapless rings can help to reduce blow-by, but their friction is very temperature-dependent because they expand significantly when heated. Polymer rings, limited cooling, and lower thermal gradient tolerance of optical materials limit the thermal load the engine can tolerate, so optical engines are often "skip fired." In this mode of operation, one or several fired cycles for data acquisition are followed by (∼10) motored cycles to allow the engine and optical components to cool down somewhat. With skip firing, engine systems that use the exhaust gas stream, such as turbocharging or exhaust-gas recirculation (EGR), cannot function as in a normal engine. To simulate turbocharging or EGR, optical engines typically are supplied with an externally pressurized and heated mixture of air with nitrogen, carbon dioxide, and/or water vapor and other gases properly metered to mimic EGR.

Further limitations in the operating range come from the mechanical properties of the optical materials (fused silica or sapphire), which are reasonably strong, but very brittle, with a low coefficients of thermal expansion that are mismatched with most metals. When an extended piston is used, the additional weight of the piston requires not only redesign of the engine's mass balancing but also a reduction in maximum speed. Typical top speed of a light-duty optical engine may be 2000–2500 rpm, compared to 5000–8000 for similar metal engine designs. Together with blow-by, the reduced axial stiffness of the slotted, extended piston also means the real compression ratio can be significantly different from the geometric compression ratio (Kashdan and Thirouard, 2009). Particularly, the calculation of apparent heat-release rate is affected by this (Aronsson *et al*., 2008, 2011). Adjusting intake temperature and pressure allows matching thermodynamic conditions of the intended all-metal counterpart over a limited range of crank angles, which may be sufficient to study, for example, injection and ignition in diesel combustion under realistic conditions (Colban *et al*., 2008).

Because of these thermodynamic differences and often much greater crevice volumes, exhaust gas composition will also differ from that of a nominally identical all-metal engine. With careful operation, however, trends over a parameter variation can be reproduced well with an optical engine (Colban *et al*., 2008).

Naturally, then, an important line of advancement in optical engines is to extent their limits of operation. For example, the engine shown in Figure 1 is designed to withstand 155 bar peak pressure. Other designs have been developed for operation at peak pressures as high as 20 MPa (Fuyuto *et al*., 2012). In the past, less emphasis has been put on extending the speed limit, but it is possible to design optical engines to operate at 6000 rpm (Fuyuto *et al*., 2012). Progress has also been made in improving the realism of the optical engine, that is, the correspondence to its all-metal counterpart in the aspects that are important for a particular investigation. For example, if the details of the flow in and around the bowl of a diesel engine's piston bowl are to be examined, the bowl of the optical engine needs to retain the production geometry. This is mechanically possible, but the downside can be optical distortion that is difficult to correct, as shown in Figure 2.

1.2 Rapid compression machines

A rapid compression machine (RCM) is similar to an engine, in that it features a closed cylinder with a sliding piston that compresses gases within the cylinder. Unlike an engine, however, the RCM is designed so that the piston remains stopped at the end of compression, so that the compressed gases do not expand. The compressed gases thus remain at high temperature and pressure for an extended period of time (10–100 ms), during which in-cylinder processes may be studied.

Optical RCMs generally share the same five operational elements: a "driver section" that applies force to accelerate a sliding piston, a closed piston-cylinder "driven section" in which gases are compressed, a mechanism to decelerate and hold the piston, a compression (combustion) chamber at the end of the piston's stroke, and transducers/windows for optical and other diagnostics. A schematic representation of the operation of one RCM, based on a design from the University of Michigan (Donovan *et al*., 2004), is shown in Figure 3. Many other designs for RCMs exist, a number of which are listed in Table 1.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

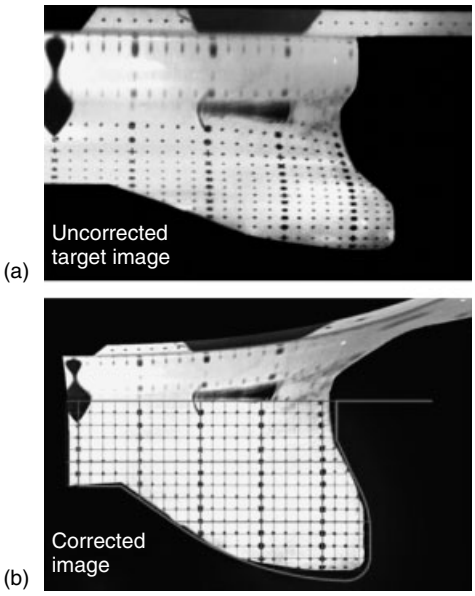


Figure 2. Raw (a) and distortion-corrected bowl region (b) image of a calibration grid in the bowl of a quartz piston crown with the geometry of a production diesel-engine bowl. (Reprinted with kind permission from Springer Science+Business Media: Experiments in Fluids, "The influence of fuel injection and heat release on bulk flow structures in a direct-injection, swirl-supported diesel engine," 43(2–3), 2007, 273–283, Miles, P. C., Hildingsson, L., and Hultqvist, A., Figure 4, Copyright Springer-Verlag, 2007.)

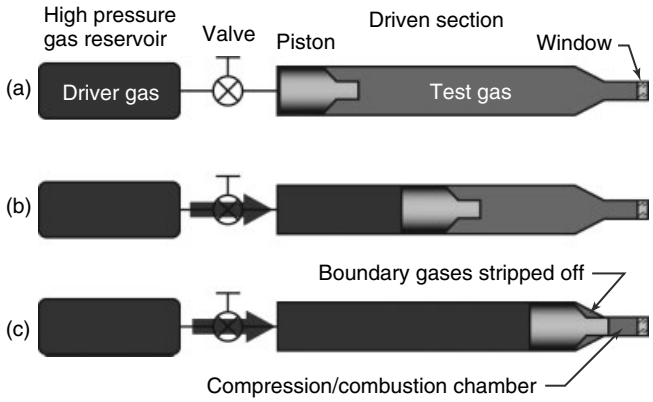


Figure 3. Example of RCM operation for design similar to that of the University of Michigan (Donovan *et al*., 2004). (a) The driven section is filled with the test gas (light gray) to be compressed. (b) A control valve is opened to release high pressure gas (dark gray) that accelerates piston, which compresses the test gas mixture. (c) At the end of its travel, the piston decelerates and locks in place as it wedges into the end of the driven section. Cooler and more turbulent boundary gases are stripped off into the shoulder region outside the compression/combustion chamber.

The RCM driving system is typically pneumatic or hydraulic, and it can act on the piston directly, as in Figure 3, or it can drive a second piston-cylinder arrangement that is mechanically connected to the compression piston. Other less-common driving systems include mechanical cams/ramps or even gravity-driven free-fall/impact. After compression, the piston must be decelerated at the end of its stroke and held in place to maintain compression. Deceleration may be accomplished by crushing/wedging elements as in Figure 3, by controls on the drive side, or by hydraulic or pneumatic braking. Piston rebound after deceleration also must be avoided to maintain compression, by wedging elements that hold the piston tight (Figure 3), feedback control of the driver system, or mechanical locks.

RCMs are free from residual gas effects and can generate more quiescent charges with less cycle-to-cycle variation than reciprocating engines, which can make the RCM a better platform for fundamental measurements, such as flame speed for spark-ignition (SI) applications. RCMs are often utilized to study chemical kinetics, for which it is desirable to create a uniform gas temperature and composition as quickly as possible. Precompression nonuniformities, heat losses creating cooler gases near the cylinder wall, and flows generated by compression all can contribute to nonuniformities in the compressed gases. Some RCM designs, such as the one illustrated in Figure 3, use a tapered/stepped piston to strip off the cooler and more turbulent boundary gases and prevent them from entering the compression chamber, thereby improving charge uniformity.

Most RCM compression times vary from 20 to 200 ms, whereas engine compression times range from 5 to 50 ms for crankshaft speeds from 600 to 6000 rpm. The fastest RCM listed in Table 1, from Nihon University in Japan (Watanabe *et al*., 2008), uses a piston-collision design to generate a rapid 5 ms compression time. RCMs typically measure the piston position and velocity during the compression stroke for diagnostics and/or feedback control of the driver system. Optical access to the compression/combustion chamber is typically achieved either with a large-scale disk window at the top of the compression chamber or by full large-scale optical ring or small-scale windows around the side of the compression chamber.

1.3 Static high temperature, high pressure spray chambers

While experimentation in optical engines and RCMs simulates the time history of compression of practical engines, uncertainties in the charge gas temperature and EGR

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Brief summary of a number of RCMs.

Institution	Acceleration/ deceleration	Bore (mm)	Compression time (ms)	Compression ratio	References
University of Michigan	Pneumatic/wedge	50.8	70	37	Donovan *et al*. (2004)
Massachusetts Institute of Technology	Pneumatic/hydraulic	50.8	10–30	19	CNF 132:219-39, 2003
Hosei University	Pneumatic	80	—	—	SAE 2001-28-0010
Université Pierre et Marie Curie	Hydraulic	40	29	15	SAE 2007-01-1869
University of Leeds	Pneumatic/hydraulic.	45	22	14.6	CNF 39:255-268, 1980
University of Science and Technology of Lille	Pneumatic/cam	50	60	10	PCI 23:1753-58, 1990
Nihon University	Collision/crush	56	5	18.7	Watanabe *et al*. (2008)
Keio University	Pneumatic/hydraulic	145	185	14.6	SAE 2009-32-0085
University of Tokushima	Pneumatic	60	—	—	SAE 2003-01-1792
Norwegian University of Science and Technology	Hydraulic	80.5	—	11.2	SAE 2005-24-009
Seoul National University	Hydraulic	60	20	20	SAE 2007-01-0218
Kyushu University	Pneumatic/cam	50	53	22.4	SAE 2002-01-2867
Tokyo Institute of Technology	—	200	—	14.7	SAE 811004
Technical University of Lodz	Pneumatic	60	20	16	SAE 2001-01-1338
Nagoya Institute of Technology	—	65	20	∼5	COMODIA:189,1994
University of Tokyo	Pneumatic/lock	50	10	12.2	CNF 54:33-47, 1983
Honda	Pneumatic/cam	80	50	16.1	SAE 2007-01-2038
Gunma University	Pneumatic/lock	80	55	10	SAE 2005-26-358
University of Karlsruhe	Pneumatic	88 sq	—	6	SAE 2005-01-0238
Wayne State University	Pneumatic/hydraulic	38	—	∼19	SAE 2001-01-2005
Argonne National Laboratory	Pneumatichydraulic	63.5	17	12	SAE 2005-01-2189
Case Western Reserve University	Pneumatic/ hydraulic	50.8	30	21	CST 179:497-530, 2007

Abbreviations: CNF = Combustion and Flame; SAE = Society of Automotive Engineers, PCI = Proceedings of the Combustion Institute, CST = Combustion Science and Technology, COMODIA = International Conference on Modeling and Diagnostics for Advanced Engine Systems.

distribution at the time of fuel injection can be significant. Consequently, the details of combustion that unfold can also have significant uncertainty. Furthermore, the maximum pressures and temperatures achievable in reciprocating optical facilities are limited, as discussed in Section 1.1. As an alternative, devices that generate sustained charge-gas conditions typical of full compression have been developed, intending to isolate certain variables and to provide a more fundamental understanding of combustion than can be achieved in engine experiments and at more extreme operating conditions. Fundamental understanding over a wide range of conditions is especially needed for development of improved CFD models where specification of exact boundary conditions is critical. Many constant-volume chambers equipped with ignition systems have been developed to understand the details of flame propagation for various fuel–ambient mixtures (e.g., flame speed as a function of pressure, temperature, and equivalence ratio). These chambers can also be utilized to prepare well-controlled high temperature ambient conditions that will subsequently mix with the spray. Different research facilities that produce high temperature, high pressure environments for subsequent spray injection are reviewed in this section.

A schematic for one type of pressurized high temperature spray chamber is shown in Figure 4. The chamber shown

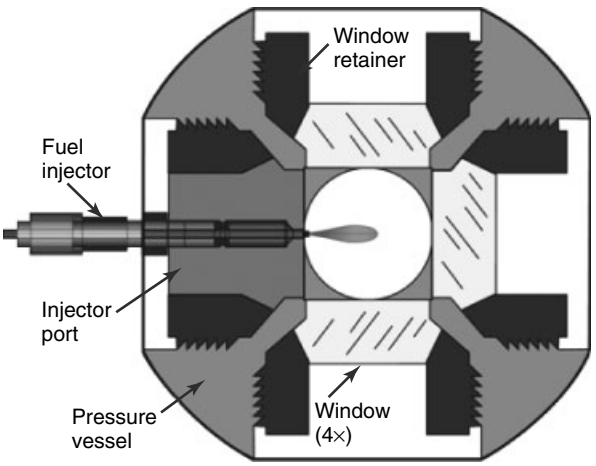


Figure 4. Schematic of high temperature, high pressure constant-volume preburn chamber (Siebers and Higgins, 2001).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is cube shaped with ports available on the six faces of the cube. In Figure 4, the chamber is configured with four optical windows to allow full view of the spray, by line-of-sight, head-on, or side access. A custom fuel injector with a single axial hole is mounted in one of the other ports, producing a spray that penetrates into the middle of the chamber. Intake and exhaust of gases occurs via small valves located at the chamber corners. The valves are closed during injection such that flow within the vessel is mainly generated by the spray itself. While highly simplified, this type of static chamber offers advantages in terms of optical access, maximum pressure and temperature capability, and isolation of variables compared to a running engine configuration.

The operational challenge for these types of chambers is to simultaneously generate high pressure and high temperature conditions representative of engines at full compression. Several different approaches have been developed, each with different design tradeoffs and capabilities. A summary of current device capabilities in terms of pressure and temperature is given in Table 2.

The first type of device listed is constant-volume heating (CVH), wherein the gases are directly heated by the walls of the chamber or by heating elements placed inside a closed chamber. An alternative is to highly preheat gases and quickly ingest them into the chamber before valve closure. Either way, the vessel walls must be heated significantly. Because of the high optical and metal material temperature and associated lower mechanical strength, the maximum pressure and temperature rating for these devices tends to be relatively low. Table 2 indicates that CVH design ratings fall below 80 bar and 900 K, which does not encompass the full range applicable to engines. Another disadvantage is that significant time (~10 min) is typically required to prepare a single charge, which limits the rate of experimentation.

The second approach listed is constant-pressure flow (CPF). For these devices, the system is open and controlled to a specific pressure. A flow of hot, pressurized gases is continuously delivered to the injector region. Fuel is injected into these gases periodically and then naturally scavenged away from the injector by the flow of air. Like CVH chambers, a significant challenge is attaining high temperature and pressure conditions because of material limitations. A solution is to separate and insulate components that direct the flow of high temperature gases from

Table 2. Summary of reported constant-volume heated (CVH), constant-pressure flow (CPF), and constant-volume preburn (CVP) chambers.

Institution	Type	Maximum temperature (K)	Maximum pressure (bar)[a]	References
Colorado School Mines	CVH	873	50	Rev.Sci.Inst.76(3):035108 (2005)
Doshisha University	CVH	700	30	SAE 2004-01-0529
Hiroshima University	CVH	873	41	SAE 1999-01-3600
Paul Scherrer Institute	CVH	850	80	Appl.Phys.B 80:1039 (2005)
RWTH Aachen University	CPF	800	50	SAE 2007-01-0020
Caterpillar	CPF	1000	150	ILASS 2011-177
Chalmers University	CPF	900	100	SAE 2004-01-1917
CMT Polytechnic University of Valencia	CPF	1000	140	ILASS 2011-163
University of Erlangen	CPF	1000	100	SAE 2011-01-1928
General Motors	CPF	900	100	ILASS 2011-170
Brigham Young University	CVP	1400	100	SAE 2005-01-0381
Doshisha University	CVP	1400	100	SAE 2003-01-0073
Eindhoven University	CVP	1400	350	SAE 2009-01-0649
Ghent University	CVP	1400	350	SAE 2012-01-0461
IFP Energies Nouvelles	CVP	1400	150	SAE 981069
University of Illinois Urbana–Champaign	CVP	1400	180	SAE 2004-01-1411
Kyoto University	CVP	1400	150	SAE 2005-01-0364
Michigan Technological University	CVP	1400	350	ASME ICEF2011-60034 (2011)
North Carolina State	CVP	1400	70	SAE 2011-01-1380
Sandia	CVP	1400	350	SAE 980809
Seoul National University	CVP	1400	60	SAE 2011-01-0684
Tokyo Institute of Technology/Meiji University	CVP	1400	80	IJER 11:79 (2011)
University of Wisconsin	CVP	1400	100	SAE 2001-01-3495

Abbreviations: SAE = Society of Automotive Engineers, ILASS = Institute for Liquid Atomization and Sprays, ASME ICEF = American Society of Mechanical Engineers Internal Combustion Engine Fall Conference, IJER = International Journal of Engine Research.
[a] Maximum pressure: stated design or maximum demonstrated.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

that of a cooled, outer pressure vessel, thereby creating a double-hull structure of hot and cold windows and housing. Table 2 shows that temperatures as high as 1000 K and pressures as high as 150 bar are possible with this type of CPF design. A substantial benefit of a CPF chamber is that injections may be repeated quickly (∼one per second), permitting fast experimentation. A disadvantage for a CPF chamber is that a pressure rise event is not available to correlate with heat release.

The third device is a constant-volume preburn (CVP) chamber, which is the type depicted in Figure 4. Table 2 shows that CVP chambers have been designed to reach temperatures of 1400 K and pressures as high as 350 bar, which is significantly higher than either CVH or CPF chamber. CVP chambers use SI and combustion of a premixed, fuel–lean mixture to raise the temperature within the vessel (Oren, Wahiduzzaman, and Ferguson, 1984). The gases and walls may be preheated, as in CVH or CPF design, but it is the "preburn" combustion that primarily raises the temperature and pressure. Chamber wall and gas heating to only 450–500 K is typical, whereas lean preburn flame temperatures approach 2000 K. Because of this temperature differential, the preburn combustion products cool over time and the vessel pressure and temperature slowly decrease. Fuel injection may occur at anytime during this cool down when the desired pressure and temperature is reached, thereby producing an extensive range of possible gas states at the time of injection. However, in practice, a time delay after completion of the premixed burn is needed to finish reaction and produce a more homogenous gas core to mix with the spray; hence, the upper estimate of 1400 K for peak temperature listed in Table 2 is lower than the lean flame temperature of about 2000 K. With fuel injection times of only a few milliseconds, the temperature and pressure at the time of fuel injection is approximately constant (Siebers and Higgins, 2001) because the mass of the ambient gases is much larger than that of the fuel injection mass.

For CVP chambers, the composition of the initial reactant mixture is chosen so that, following its complete combustion, the product gas composition simulates air (21% O_2) or even EGR-diluted engine intake conditions (e.g., 10–19% O_2). As such, the spray will autoignite as it does in diesel combustion. Alternatively, a preburn mixture that yields 0% O_2 can be produced to simulate high temperature and pressure conditions, but as the mixture is inert, no spray combustion occurs. In this way, the spray mixing and vaporization processes may be separated from that with ignition and combustion. In addition, when spray combustion does occur in an oxygenated environment, the measured pressure rise within the constant-volume chamber (CVP or CVH) may be related to the spray heat release rate, including ignition delay and combustion duration. While CVP chambers can traverse a wide range of temperature, pressure, and ambient compositions compared to other devices and provide quantitative heat release rate, a disadvantage is that the preparation steps to produce the preburn take several minutes, thereby limiting the rate of experimentation, particularly compared to continuous-flow CPF chambers.

Beyond being a tool for understanding spray mixing and combustion, constant-volume preburn CVP chambers have contributed significant understanding about SI and flame propagation processes at the conditions most appropriate to engines. Heat release from the flame propagation continuously increases the pressure and temperature of the unburned reactants, or end gas, which modifies the local conditions of the flame as in an actual engine. Idealized experimental configurations, such as the development of a laminar, spherical flame originating from an ignition site, are utilized to understand the temperature of the unburned gases with respect to time, thereby providing a measurement of flame speed as a function of reactant temperature and pressure (Metghalchi and Keck, 1982). Such CVP experiments provide the most pertinent data (e.g., laminar flame speed, autoignition or knock, and soot formation dependency on equivalence ratio) for fuels, charge-gas compositions, and charge-gas temperatures and pressures appropriate to engines.

2 OPTICAL DIAGNOSTIC TECHNIQUES

A variety of in-cylinder phenomena may be probed by optical diagnostics, including flow structures and turbulence associated with gas exchange, compression/expansion, and transient fuel jets; liquid-fuel atomization and vaporization; mixing of vapor fuel with ambient gases; ignition and combustion; and SI kernel growth, premixed flame propagation and quenching, and pollutant formation and destruction. The light measured by the optical diagnostics may arise naturally from combustion processes, or it may be artificially introduced and/or induced using external light sources, including lasers, flashlamps, arcs, and light-emitting diodes. Many optical diagnostics are available, and it is beyond the scope of this chapter to present all of them. The following is a brief overview of several of the more commonly used diagnostics.

2.1 Naturally occurring luminescence

Combustion processes emit light from chemical reactions (chemiluminescence) and combustion-heated particles may

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

also emit light (incandescence). Both the intensity and the spectrum (color) of the naturally occurring luminescence can provide information about the local temperature and concentration of the emitters.

2.1.1 Chemiluminescence

Chemical reactions during combustion can emit light either directly, during the reaction, or indirectly, by producing electronically excited species that later emit light when they spontaneously return to an electronic ground state. For direct chemiluminescence, excess reaction energy is emitted as light (photons). One important direct chemiluminescence reaction is the CO continuum

$$CO + O \rightarrow CO_2 + hv \quad (1)$$

In Equation 1, *hv* represents the emission of a photon from the reaction. The wavelength (color) of the emitted photon depends on the excess energy in the reaction collision, which is not quantized. Hence, the photon wavelength can vary continuously. In the case of the CO continuum, the emission is in the blue to ultraviolet (UV), over the approximate range from 350 to 450 nm.

For other chemiluminescence reactions, excess chemical energy is not released immediately, but rather it is temporarily stored in the electron cloud of one of the product species. A few important reactions producing electronically excited species are

$$CH + O_2 \rightarrow CO + OH^* \quad (2)$$

$$C_2 + OH \rightarrow CO + CH^* \quad (3)$$

$$CH_2 + C \rightarrow H_2 + C_2^* \quad (4)$$

$$CH_3O + OH \rightarrow HO_2 + CH_2O^* \quad (5)$$

In Equations 2–5, the "electronically excited" species are denoted by an asterisk superscript. In the excited species, one electron occupies a quantum manifold (shell) that is one or more electronic levels above the lowest, or ground manifold. The electron does not remain in the excited manifold indefinitely, but rather spontaneously "relaxes" to a lower (usually ground) manifold, and in doing so, emits a photon. Both the excited and ground manifolds can have a multitude of slightly different, yet discrete (quantized), energy states, depending on the vibrational and rotational quantum numbers of the occupied states. The spontaneous transition of each excited molecule must occur between discrete quantum states. As a result, the wavelength (color) of the emitted photon must correspond exactly to the energy difference between the two states. Among the population of all emitting molecules in a probed region, many different transitions will occur between many different vibrational and rotational states, so that the collection of emitted photons will have many different, yet discrete, colors. The chemiluminescence emission thus bears the spectral "fingerprint" of the excited species.

The intensity of chemiluminescence emission depends on the reaction rate and/or the concentration and temperature of the emitters. High temperature combustion reactions that produce CH^*, OH^*, and C_2^* are often detectable with photodiodes or unintensified cameras at bandwidths or frame rates on the order of 10 kHz. For weaker low temperature (cool flame) combustion reactions, emission from CH_2O^*, or for conditions with weaker high temperature reaction emissions (e.g., with dilution by exhaust gases or late in combustion), photomultiplier tubes or intensified cameras may be required.

A few notable examples of chemiluminescence measurements are using chemiluminescence of OH^* as a marker for high temperature reactions in diesel sprays to measure the lift-off length (Siebers and Higgins, 2001); using relative CH^*, C_2^*, and OH^* chemiluminescence intensities to infer the local equivalence ratio (Ikeda, Kaneko, and Nakajima, 2001); tracking ignition and combustion in compression-ignition engines (Table 3) (Augusta *et al*., 2006; Kokjohn, Musculus, and Reitz, 2012); and kernel development and flame propagation in SI engines (Smith and Sick, 2005; Aleiferis *et al*., 2004).

Limitations of chemiluminescence include the line-of-sight nature of the signal and interference by other emitters. Emitters all along the line of sight contribute to the emission that reaches the detector, so that the signal represents an accumulation along the line of sight. Hence, two-dimensional images are effectively projections of a three-dimensional source. Spectral measurements also represent an integration along the line of sight. Spectral filters or dispersion techniques (e.g., spectrographs) can help to isolate specific emitting species, but the spectral emission of multiple species often overlap, which limits the specificity of the chemiluminescence signal. Furthermore, other emission sources, especially including soot incandescence, can be orders of magnitude stronger and overlap spectrally with the chemiluminescence emission.

2.1.2 Soot incandescence

Under fuel-rich conditions, combustion reactions can synthesize carbonaceous soot particles, some of which

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Brief summary of a number of optical engines.

Institution	Engine/Head	Displacement (cm^3)	Bore (mm)	Compression Ratio	References
Compression-Ignition Engines					
University of Illinois	FEV engine, Ford DIATA head	0	70	19.5	SAE 2002-01-2666
Bosch	—	0	68	18	SAE 2005-01-0181
University of Erlangen-Nürnberg	Audi V6 TDI	0	78.3	15.5, 19.5	SAE 2001-01-3499
Renault	—	0	80	18.1	SAE 2003-01-3083
University of Erlangen-Nürnberg	VW 1.9 L	474	79.5	20.5	SAE 922204
Imperial College London	VW 1.9 L	0	79.5	19.5	SAE 950850
Sandia National Laboratories	General Motors 1.9 L	477	82	13.95, 16.6	SAE 2008-01-1066
Lund University	Volvo D5	0	81	13.75	EF 43:273-283, 2007
Korea Advanced Insitute (KAIST)	—	0	83	18.9	SAE 2004-01-0127
University College London	Lister Petter Diesel	498	89	—	SAE 2009-01-1921
Brunel University of Hiroshima	Ricardo Hydra, Ford 2.0l HSDI head	499	86	15.88	SAE 2005-01-0915
PSA Peugeot Citroen	PSA DW 10	0	85	17.6	SAE 2002-01-1162
Institut Français du Pétrole (IFP)	PSA DW 10 TED4	499	85	14	SAE 2003-01-3174
University of Hiroshima	—	500	86	16.7	SAE 2007-01-4050
Wayne State University	AVL Model 5402	0	85	15	SAE 2009-01-2712
Istituto Motori	General Motors Euro5 2.0 L	0	85	16.5	SAE 2011-01-1381
Royal Institute of Technology (KTH), Stockholm	AVL FM 528	0	85	—	SAE 2000-01-2862
Daimler	—	0	88	18.5	SAE 1999-01-3646
Doshisha University	—	0	91.1	17.2	SAE 1999-01-3652
Imperial College London	Ricardo Hydra, Ford York 2.5 L head	0	93.67	16.75	SAE 2000-01-1183
Istituto Motori	—	750	100	22.3	SAE 2001-01-1258
Sandia National Laboratories	Cummins B	0	102	14, 18	SAE 2006-01-1518
Wayne State University	AVL 520	0	120	19.27	SAE 952366
Sandia National Laboratories	Caterpillar HEUI	1720	125	11.8	SAE 2009-01-1792
Sandia National Laboratories	Scania DSC14 V8	1773	127	18	SAE 1999-01-3650
Korea Advanced Insitute (KAIST)	—	1818	128	17	SAE 2009-01-1354
Eindhoven University of Technology	DAF MX	1939	130	15, 16	PCI 31:765-773, 2007
Helsinki University of Technology	commercial six-cylinder off road	1939	111	—	SAE 2009-01-0710
Lund University	Scania D12	1950	127	15.1	SAE 2009-01-1353
Daimler	Daimler-Chrysler BR500	0	130	17.25	SAE 1999-01-3647
Mitsubishi	—	2000	135	12.0, 18.5	SAE 1999-01-0185
Sandia National Laboratories	Cummins N-14	2340	139.7	10.75	SAE 2006-01-0079
University of Politecnica Valencia (CMT)	—	3000	150	—	SAE 2003-01-1110
Helsinki University of Technology	Wärtsilä W20	8796	200	—	SAE 2008-01-2477

survive combustion and are ultimately exhausted from the engine, along with adsorbed hydrocarbons and sulfates, as particulate matter (smoke) (see Particulate Formation and Models for more information about particulate matter). The soot is of practical interest because it is a primary component of particulate matter, and emissions of particulate matter from engines are regulated in most markets. Hence, knowledge of the in-cylinder processes that control the formation and destruction of soot is important for designing combustion systems that comply with air quality standards.

While inside the engine, the combustion-generated soot particles are heated to high temperatures such that they glow by incandescence. In an aerosol (cloud) of soot particles, some of the light emitted by each particle is scattered and absorbed by other particles, and the amount of absorption and scattering within the aerosol depends on the amount of soot and the light wavelength. The spectral intensity

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. (*Continued*).

Institution	Engine/Head	Displacement (cm^3)	Bore (mm)	Compression Ratio	References
Spark-Ignition Engines					
City University of London	Yamaha	250	73	9.5	SAE 2001-01-3556
University of Tokai, Japan	Honda SOHC 4	0	80	—	SAE 960266
University of Brighton	Ricardo Hydra	0	74	9.2	SAE 2009-32-0072
Imperial College London	Honda BKR-4	360	75	7.9	SAE 2000-01-1207
Instituto Motori	—	0	79	10	SAE 2009-01-0697
University of Hiroshima	—	399	78	10.0(8, 3)	SAE 932641
Renault	—	0	79.5	9.6	SAE 971643
University of Rouen (CERTAM/Renault)	—	400	79.46	6.7	SAE 2000-01-1794
University of Erlangen-Nürnberg	BMW BDE	400	77	—	SAE 2009-01-0656
Jaguar	Jaguar 2.5 L V6	0	81.64	12.9	SAE 2005-01-2129
Institut Français du Pétrole (IFP)	Renault F7P	441	82	8.4	SAE 961928
Volkswagen	VW	0	81	6.5	SAE 961122
Brunel University	Ricardo Hydra	0	80	11	SAE 2004-01-1354
University of Oxford	Mitsubishi GDi	0	80	—	SAE 2002-01-0839
Bosch	—	0	82	9.6	EF 48:281-290, 2010
Institut Francais du Petrole (IFP)	—	0	85	7.8	SAE 2001-01-1926
University of College London	MAHLE Powertrain	0	82.5	9.8	SAE 2008-01-0073
Chalmers U. Tech., Göteborg	AVL block, Volvo head	485	83	10.5	CST 183:1266-1281, 2011
University of Rouen	—	0	87	4.7	SAE 941988
BMW	BMW GDI	0	84	7	SAE 2000-01-1793
University of Loughborough	Jaguar V6 3.0 L	495	89	11.3	SAE 2003-01-1859
Nissan Motor Co., Ltd.	—	498	93	8	SAE 2006-01-1202
Renault	—	0	88	9.2	SAE 952457
University of Duisburg-Essen	AVL 528, BMW head	0	84	10	PCI 34:2911-2919, 2013
University of Orléans (PRISME/LME)	—	0	85	10.2	Lisbon 2008
Institut Français du Pétrole (IFP)	—	0	82.7	10.6	SAE 2010-01-0346
Toyota	—	500	86	10	SAE 2011-01-2050
University of Michigan	—	0	86	11	SAE 2003-01-0068
University of Loughborough	Lotus	500	88	10.5	SAE 2007-01-1414
Lund University	AVL 528	0	85	10	SAE 2004-01-0609
University of Birmingham	Jaguar V8	0	89	11.15	SAE 2007-01-4033
University of Oxford	Jaguar V8	0	89	11.15	SAE 2007-01-4033
University of College London	Jaguar V8	562	89	11.15	SAE 2007-01-4033
Sandia National Laboratories	Labeco CLR	0	92	11.8	SAE-2000-01-2900
General Motors	—	0	92	8	SAE 2000-01-0246
University of Brighton	Ricardo engine, Volvo B234 head	575	92	10.1	SAE 2010-01-1458
University of College London	Short block Lister ST-1	0	90.2	—	SAE 961149
Stanford University	GM	0	92	11.5	SAE 2008-01-1067
Lund University	Volvo TD 102	0	120.65	10(15)	SAE 1999-01-3649

of light that is ultimately emitted by a soot aerosol can be approximated as a fraction ε (emissivity) of that of a blackbody, according to the Planck equation

$$I = \varepsilon \frac{375\ [\mathrm{W \cdot nm^2}]}{\lambda^5 \left(\exp\left({}^{14.4\cdot 10^6[\mathrm{nm\cdot K}]}/_{\lambda T}\right) - 1\right)} \tag{6}$$

In Equation 6, I is the total spectral intensity in watts per cubic nanometer (radiative power per source area per wavelength) emitted into a hemisphere, λ is the light wavelength in nanometers, T is the soot temperature in Kelvin, and ε is the emissivity (which would have a value of one at all wavelengths for a blackbody).

Assuming scattering is negligible, the emissivity ε of the soot aerosol can be expressed as

$$\varepsilon = 1 - \exp(-k_a L) \tag{7}$$

where k_a is the soot absorption coefficient and L is the path length of the light traveling through the cloud. The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

product k_aL is often termed the *KL factor*. The absorption coefficient depends on both the soot concentration in the aerosol and the soot optical properties, which vary with the light wavelength. Various approaches, either fundamental or empirical, can provide a functional relationship between k_a and the soot aerosol concentration. The optical properties of soot and the importance of scattering are highly uncertain, however, which translates to high uncertainty in the relationship between k_a and soot concentration. Nevertheless, using the two above-mentioned equations together, the broadband spectrum and intensity of the soot aerosol emission thus becomes characteristic of the soot temperature and concentration. Hence, simultaneous measurements at multiple light wavelengths can provide sufficient information to determine both the soot temperature and its concentration. Measurements at a minimum of two wavelengths are required, which yields two equations for the two unknowns (temperature and concentration).

Figure 5 shows an example of soot KL and temperature maps for *n*-heptane diesel spray combustion (Svensson *et al*., 2005). The KL and temperature data were calculated from two-color soot incandescence images acquired with a single color camera with relatively broad blue (400–550 nm) and red (570–700 nm) channel sensitivities. The two-color data were compared with laser extinction measurements of soot (Section 2.3.1), which showed that the two-color KL values were consistently lower than the extinction measurements, although the trends were consistent.

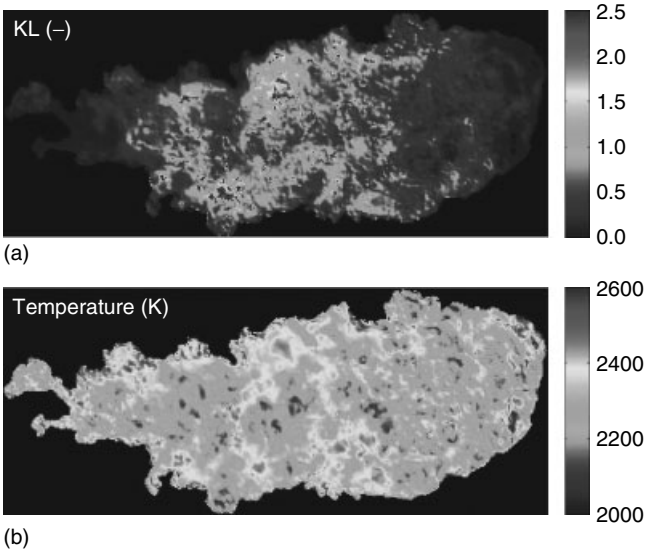


Figure 5. (a, b) Soot KL and temperature in an *n*-heptane diesel jet measured by the two-color method. (From Svensson *et al*., (2005). Copyright © 2005 SAE International. Reprinted with permission.)

In addition to the study in Figure 5, a few other notable examples of soot incandescence measurements are total soot incandescence imaging of diesel soot (Mueller and Martin, 2002); total soot incandescence, two-color, and laser-induced incandescence (LII) imaging in diesel spray flames (Pastor *et al*., 2005); and two-color uncertainty assessment with computer modeling and measurements under conventional and low temperature diesel combustion conditions (Musculus, Singh, and Reitz, 2008).

While broadband measurements (imaging) of the soot emission can provide some information about in-cylinder soot formation and destruction, the data are confounded because variations in soot intensity may be due to variations in either soot concentration or temperature, or both. Two-color techniques, in theory, can separate the temperature and concentration effects on the emission intensity, so that soot formation and destruction can be quantified. In practice, however, the measurement uncertainty can be considerable. In addition to uncertainty in the soot optical properties and absorption and scattering processes, variations in the soot temperature along the line of sight, additional extinction from window deposits, and the contribution of reflections off in-cylinder surfaces and other luminous interferences can further increase the uncertainty.

One other laser-based option for measuring temperature in engines is coherent anti-Stokes Raman spectroscopy (CARS). The technique is based on interaction of several laser beams in the probe volume and yields highly resolved anti-Stokes Raman spectra of (usually) nitrogen, which are fit to calculated data to obtain temperature. Its main advantages are high accuracy, broad applicability, and resistance against interfering signals. To its disadvantage, it is mature only as a point measurement and data evaluation requires a relatively high degree of spectroscopic expertise. A recent application example is Birkigt *et al*. (2011).

2.2 Beam refraction: shadowgraphy and schlieren

Variations in local density and composition in combustion systems and evaporating sprays create gradients in the refractive index, so that light passing through the region is bent. Optical techniques sensitive to refractive index gradients can probe such regions. Schlieren and shadowgraph imaging systems exhibit different sensitivities to the variations in refractive index and are conveniently applied wherever there is line-of-sight optical access. The lighting and imaging arrangement also lends itself easily to high speed imaging, wherein the entire cycle or injection and combustion event can be resolved with hundreds or thousands of images.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

A Z-type schlieren or focused shadowgraph setup may utilize inexpensive lighting, such as lamp or LED point sources, coupled to camera imaging systems (Settles, 2001; Pickett, Kook, and Williams, 2009). The diverging point source light is collected and collimated by a concave mirror, passed through the combustion chamber, refocused using a second concave mirror, and then directed into a camera equipped with lenses appropriate for the desired image magnification.

Schlieren stops are placed at the collection focal point to adjust system sensitivity to refractive index gradients. A knife-edge is the classic schlieren stop. When rays are steered by refractive index gradients within the combustion chamber, they may be steered into or away from the knife-edge, thereby identifying these regions in the collected image. The degree of knife-edge cutoff affects the sensitivity. With 0% cutoff (no knife-edge), the sensitivity is weakened and the system is technically no longer a schlieren setup but rather a "focused shadowgraph" (Settles, 2001). With high cutoff, the sensitivity to refractive index gradients increases. A round aperture can be used as a schlieren stop producing a "bright field", as it tends to leave the background bright, and darken regions where there is a schlieren effect. Alternatively, a full light cutoff at the focal spot with a pin shape produces a "dark field" because schlieren-effected regions appear bright above a dark background.

For vaporizing fuel sprays, schlieren imaging shows the boundary between vaporized fuel and background ambient gases because (i) refractive index differences exist between the fuel and ambient gases and (ii) density gradients are created in the ambient gases as the vaporized fuel spray cools the ambient. An example is shown in Figure 6. The technique also works well to identify the boundary of nonvaporizing sprays because of light extinction by droplets (Naber and Siebers, 1996) (Section 2.3.2). Light extinction by droplets is clearly shown in the upstream region of the spray in Figure 6. Strong density (and refractive index) gradients are also easily detected at combustion interfaces as high temperature combustion forms distinct low density regions. In addition, first-stage ignition in a fuel jet, where the parent fuel breaks down and the temperature rises only modestly, produces a vanishing schlieren effect as these ignition steps change the refractive index of the jet to match the ambient more closely (Pickett, Kook, and Williams, 2009).

Flame fronts in premixed combustion, such as in an SI engine, create steep thermal gradients that can be detected with schlieren methods. In particular, the development of the early flame kernel has been examined this way, both in combustion vessels (Metghalchi and Keck, 1982) and in engines (Salazar and Kaiser, 2011).

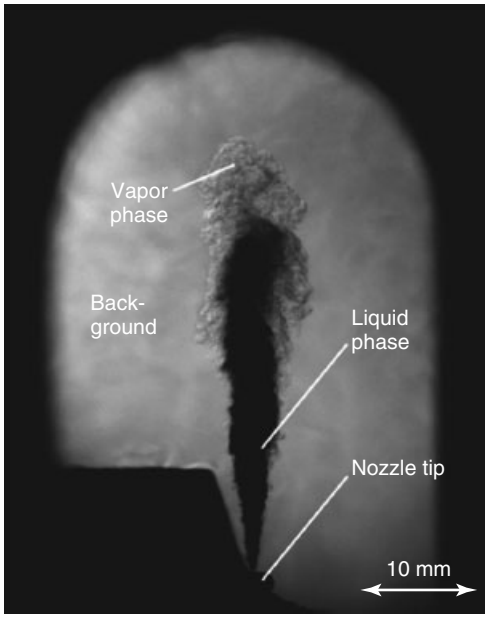


Figure 6. Schlieren imaging of a diesel spray penetrating into ambient gases at 800 K and 50 bar. (From Pawlowski *et al*., 2008. Copyright © 2008 SAE International. Reprinted with permission.)

2.3 Beam extinction

Light passing through an engine combustion chamber may be absorbed and/or scattered by the in-cylinder contents. The sum of absorption and scattering, termed *extinction*, can be indicative of the properties and quantity of matter along the beam path. The transmitted intensity of a light beam directed through the combustion chamber can be measured to quantify the amount and/or temperature of in-cylinder soot particles, liquid fuel sprays, and in-cylinder gases.

2.3.1 Soot

The transmitted intensity I of a light beam passing through a uniform soot cloud (aerosol) can be described by the Beer–Lambert law

$$I = I_0 \exp(-k_e L) \tag{8}$$

In Equation 8, I_0 is the incident intensity of the light entering the soot cloud, L is the path length of the light passing through the cloud, and k_e is the extinction coefficient, which can include both scattering and absorption.

With measurements of the incident and transmitted intensities, the value of $k_e L$, often called the *KL factor*, can be determined using Equation 8. The KL factor is a relative measure of the amount of soot along the path of the beam. An absolute measure of the soot may be estimated using

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

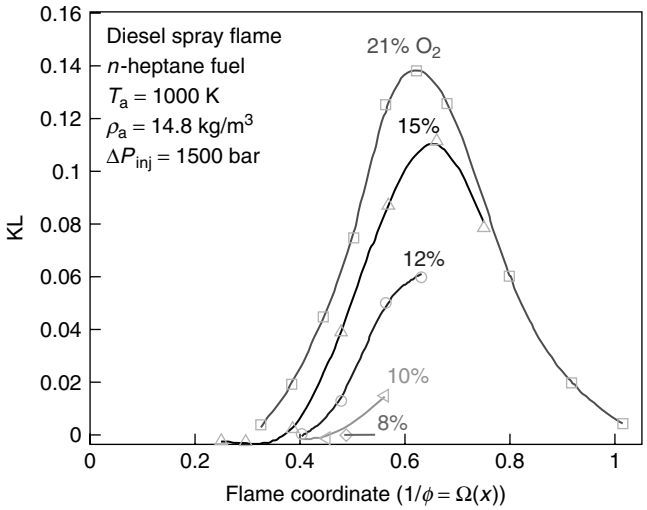


Figure 7. Laser extinction measurements of soot KL in a diesel spray flame using *n*-heptane fuel at different ambient oxygen concentrations, plotted as a function of flame coordinate. (From Idicheria and Pickett, 2005. Copyright © 2005 SAE International. Reprinted with permission.)

correlations for k_e that are either empirically based or from scattering theory using the optical properties of the soot particles. For small soot particles, extinction is typically dominated by absorption, such that the extinction coefficient in Equation 8 can be approximated as the absorption coefficient k_a used in Equation 6 of the two-color incandescence method.

The light source for extinction measurements is usually a laser source, although other light sources could be used as well. Figure 7 shows soot KL measurements in a diesel spray flame using *n*-heptane fuel at several different ambient oxygen concentrations (Idicheria and Pickett, 2005). The data are plotted versus the "flame coordinate, $\Omega(x)$," which is the reciprocal of the nominal equivalence ratio, ϕ, along the jet axis. The data show how soot is formed and consumed along the jet axis and how dilution by EGR reduces soot formation.

In addition to the study highlighted in Figure 7, a few notable examples of soot extinction measurement are combined soot extinction and LII in a gasoline direct injection engine (de Francqueville, Bruneaux, and Thirouard, 2010), fundamental soot extinction measurements in several laminar diffusion flames (Williams *et al*., 2007), and a modified forward-illumination extinction technique that can be particularly useful in engines with limited optical access (Xu and Lee, 2005).

Properly applied, the soot extinction technique can provide quantitative measurements with relatively low uncertainty, and hence, it has been used frequently as a calibration for other soot measurement techniques. Nevertheless, many of the same uncertainties in the soot optical properties and window transmission involved in the two-color soot incandescence measurements described in Section 2.1.2 must also be considered in the functional relationship between the KL factor and the total soot mass or volume fraction along the beam path. Furthermore, the technique usually needs to be designed to minimize interference from combustion luminosity, and the sensing optics must accommodate any bending of the extinction beam that may occur as it passes through regions of varying density within the combustion chamber.

2.3.2 *Sprays*

Like soot, fuel spray droplets cause light extinction of transmitted laser beams or other light sources. The attenuation of light is a convenient diagnostic to characterize the spray position and concentration of droplets or to determine the position where all liquid droplets are vaporized as they mix with hot, in-cylinder gases. Liquid droplet extinction is labeled in the vaporizing spray of Figure 6. Another example of light extinction in a low temperature hollow-cone gasoline spray is given in Figure 8. Rather than a collimated light source, as with a laser beam or schlieren imaging (Section 2.2), in this case, the spray is illuminated from the back with a diffuse light source (Ghandhi and Heim, 2009; Gasoline Fuel Injection Standards Committee, 2007). Dark regions indicating light extinction by droplets are readily apparent.

Quantifying the total liquid fuel volume along the light path is difficult because of complex extinction relationships for fuel droplets. The measured intensity and the

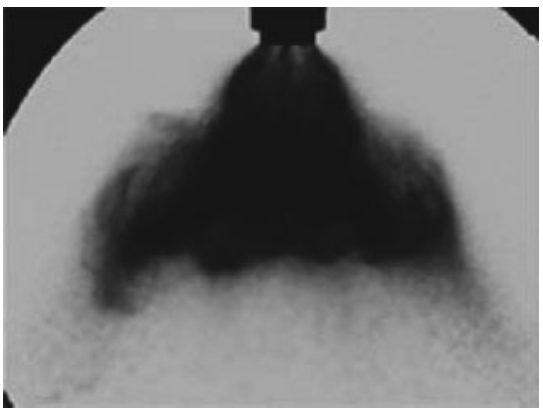


Figure 8. Diffuser back-illumination image of an air-assisted gasoline spray. (Reprinted with permission from Ghandhi and Heim, 2009. Copyright 2009, American Institute of Physics.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Beer–Lambert law (Equation 8) can be applied to assess the extinction coefficient k_e for a cloud of droplets, with $k_e = \sigma_{ext}N$ where σ_{ext} is the extinction cross section of a droplet and N is the number density of droplets. However, fuel droplets have much different σ_{ext} characteristics compared to soot particles. Fuel droplet extinction is dominated by light scattering, rather than absorption. In addition, fuel droplet size is much larger (tens of micrometers) than that of soot particles (tens of nanometers). As shown in Section 2.4.1, Mie scatter theory predicts significant changes in σ_{ext} depending on the size of the particle/droplet, thereby making quantification difficult, especially as droplet size is nonuniform in most applications. Another challenge is that many engine sprays are dense, with either high σ_{ext} or N, causing nearly complete extinction, as well as multiple photon scattering with illumination originating from many directions other than the original source. Techniques to suppress multiple scattering are therefore being developed and applied, with both advanced laser systems (Berrocal *et al*., 2008; Linne *et al*., 2009) and X-ray sources (Kastengren *et al*., 2009). While difficult to quantify the fuel volume within the spray, light extinction is particularly useful for rapid characterization of the spray envelope as depicted in Figure 8.

2.4 Elastic scattering

Direct elastic scatter from laser or other light sources is applied to characterize in-cylinder fuel–air mixing, liquid evaporation, and combustion. Unlike the diagnostics covered thus far in previous sections, light collection at angles to a precisely directed light source allows spatially resolved, rather than line-of-sight, measurement. The application of elastic scattering depends strongly on the size of the particle. Complex Mie scatter theory is required for particles and droplets, whereas a simpler Rayleigh scatter assumption can be applied to small particles and molecules.

2.4.1 Mie scatter

Gustav Mie developed an analytical solution to describe the scattering of light by spherical particles or droplets, given the light wavelength, the droplet diameter, and refractive index. Figure 9 shows the angular scattering intensity calculated from Mie's solution for an *n*-dodecane droplet with a diameter of 5 μm, a size that is typical for engine fuel sprays. Figure 9a indicates that the strongest scatter (by a factor of 10 or more) occurs in the forward direction (0°), but some light is also scattered sideward and backward. Scattering peaks and valleys (or lobes) also exist at discrete angles, for example, near 30°, and these angles vary with droplet diameter.

The total scatter is also very sensitive to droplet diameter. Figure 9b shows the extinction cross section σ_{ext} versus droplet diameter d, where σ_{ext} is proportional to the summation of light lost through the scattering process at all angles. Larger droplets more typical of a spray exhibit a d^2 dependency on σ_{ext}, whereas smaller droplets and particles with diameters approximately less than the light wavelength fall into the Rayleigh-scatter regime and exhibit a d^6 dependency, as well as a simpler angular scattering pattern.

Because of the difficulties associated with understanding the exact droplet diameter and collection angle, as well as the actual incident intensity after attenuation and scatter by

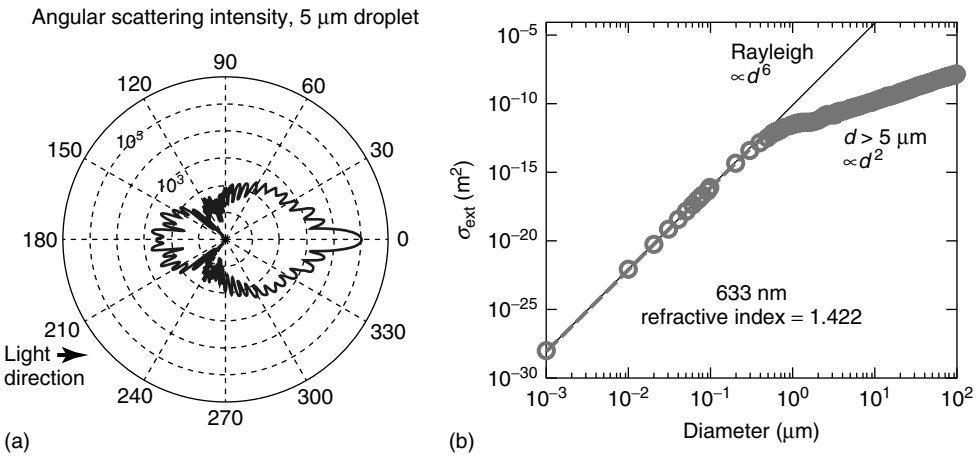


Figure 9. (a) Angular scattering intensity distribution (logarithmic scale) for a 5-μm *n*-dodecane droplet in air. (b) Extinction cross section as a function of droplet diameter (logarithmic scale). Wavelength is 633 nm and unpolarized; droplet refractive index is 1.422.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

other droplets, Mie scatter is typically not used to quantify total droplet concentration in engine applications. However, it can be used to identify if droplets do or do not exist at a particular location and is therefore widely used for imaging in engines. In addition, Mie scatter optical techniques, such as phase-Doppler interferometry, capitalize on the existence of angular scattering lobes indicated in Figure 9 as means for measuring droplet diameter size distributions within sprays (Sankar *et al*., 1991).

2.4.2 *Rayleigh scatter*

Elastic scatter from small, nanometer-sized particles exhibits Rayleigh scattering behavior when using typical light sources. Soot particles or even gaseous fuel molecules fall into this regime. For example, the molecular size of vaporized *n*-dodecane is less than 2 nm and the extinction cross section $\sigma_{ext} = 8.5 \times 10^{-29}$ m^2 (Pickett *et al*., 2011), consistent with the relationships shown in Figure 9. By comparison, σ_{ext} for a typical droplet is more than 15 orders of magnitude higher. This major difference effectively means that Rayleigh scattering measurements must be performed in an environment without droplets or larger particles, and other scattering sources such as window reflections must be minimized. If this can be achieved, Rayleigh scattering offers the potential for quantitative measurement because, unlike droplet Mie scattering, the scattering cross sections for gaseous fuel and oxidizer molecules are known (Idicheria and Pickett, 2007).

Figure 10 shows a composite of a Mie scattering (left) and Rayleigh scattering images (top right) acquired in a vaporizing *n*-heptane fuel spray. The Mie scattering and Rayleigh scattering images are shown as a composite, but they were actually obtained at different instances with different experimental setups. The Mie scattering image shows that as the spray entrains and mixes with the hot ambient, all of the liquid fuel is eventually vaporized downstream of about 10 mm. A laser sheet positioned downstream of this liquid droplet region induces Rayleigh scattering from fuel vapor and ambient gases without interference from liquid droplets.

The Rayleigh scattering signature from the ambient gases outside of the spray creates reference data that permits corrections for laser sheet intensity variations caused by energy variation in the delivery beam as well as beam steering in the high pressure combustion vessel. The laser energy variation is apparent in the raw Rayleigh image. Not only is the incoming laser intensity nonuniform, but as the laser beam propagates through the vaporized fuel spray from bottom to top, there is additional beam refraction caused by temperature variation within the spray (see Section 2.2 for more discussion on beam refraction). The success of these corrections is indicated by image at the bottom of the figure, where intensity in the right image is proportional to the local mixture fraction (or fuel mass fraction).

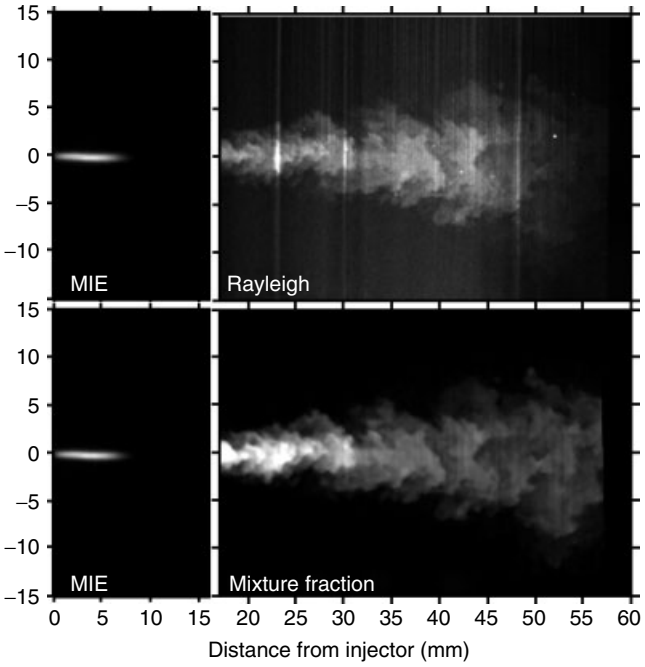


Figure 10. Composite image of Mie scattering (left) and Rayleigh scattering (right). Ambient conditions: 1000 K, 14.8 kg/m^3, 4.2 MPa, and 0% O_2. Injector conditions: 0.100-mm orifice, 154 MPa, and *n*-heptane. (From Pickett *et al*., 2011. Copyright © 2011 SAE International. Reprinted with permission.)

2.5 Fluorescence

Like chemiluminescence (Section 2.1.1), fluorescence refers to light emitted when an electronically excited molecule spontaneously transitions to a lower energy state. For fluorescence, the molecule is excited not by chemical reaction but by absorption of light photon(s), typically from a laser source (laser-induced fluorescence, LIF). The photon absorption is typically very strong, so combustion intermediate species can be detected down to ppm-level concentrations. As described in Section 2.1.1, each molecular species only absorb photons at specific wavelengths (colors), so specific species may be targeted if the available laser output wavelengths can be matched to the absorption wavelength of the target species. In combustion, many of the practically useful absorption lines are in the UV wavelength region.

After excitation into a specific energy state, various energy transfer processes (e.g., collisions) populate other

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

rotational, vibrational, and even electronic states, typically with lower energy than the directly excited state. When these lower energy states transition back to the ground state, the emitted photon has less energy than the original excitation photon. Spectrally, this means the fluorescence emission occurs at a longer wavelength than excitation. This redshift is of great practical importance, because it allows *blocking* the incident laser wavelength by suitable color-selective filters while still *detecting* the fluorescence signal. Thus, the detector sees only the redshifted fluorescence signal, not unwanted reflections and scattering of the laser light at the unshifted color off windows, particles, and so on.

Not all excited molecules emit a fluorescence photon. Collisions with other molecules can deexcite the excited molecule without emission of a photon, a process called *quenching*. The likelihood of quenching depends on pressure, temperature, and the collisional partner, that is, on the gas composition. There are also other noncollision deexcitation mechanisms, such as intersystem crossing and internal conversion, which depend on temperature. The overall likelihood of absorbing light and then returning to the ground state via fluorescence, or equivalently, the fraction of excitation photons that yield fluorescence photons, is called *quantum efficiency*.

As a result of the multitude of energy pathways in the fluorescence process, determination of the quantum efficiency requires knowledge of both the photophysical properties and the thermochemical environment of the molecule to be detected. The former is often not available from the literature for high temperature and pressure; the latter may not be known in combustion. Hence, quantification of fluorescence signals to determine concentration can be challenging or even impossible. On the other hand, these dependencies can sometimes be exploited to gain additional knowledge about the local temperature or gas composition. For these reasons, advancement of fluorescence measurements in engines crucially depends on improving our understanding of photophysics at engine-relevant conditions via fundamental experiments in pressurized, heated cells and shock tubes.

2.5.1 Combustion intermediates

Among the educts and products of complete combustion, nitrogen, oxygen, water, or carbon dioxide cannot be detected by fluorescence in engines in a practically useful manner. However, some important combustion intermediates and fuel components have accessible electronic transitions making LIF a powerful tool for their in-cylinder measurements. Table 4 summarizes key features of fluorescence measurements in engines for five commonly detected combustion intermediates. All have fluorescence signals strong enough for spatially resolved, two-dimensional planar laser-induced fluorescence (PLIF), although barely so in the case of CO. Note that, except for C_2, all excitation wavelengths are in the UV, which is typical for electronic absorption bands of small molecules.

The hydroxyl radical (OH) is one of the most common targets for fluorescence measurements of combustion intermediates. Strong absorption features around 283.5 nm can be excited by use of tunable UV lasers such as a frequency-doubled, Nd:YAG pumped dye laser or an optical parametric oscillator (OPO). In nonpremixed combustion (classic diesel), OH is formed in hot, stoichiometric or slightly fuel–lean mixtures. In premixed auto-ignition (HCCI), OH indicates the active second-stage ignition and the main phase of heat release. In premixed SI, the initial appearance of OH delineates the flame front, and it persists later in the burnt regions.

In addition to those listed in Table 4, many other species can be detected, although most of them not in the single-shot measurements that are necessary to "freeze" turbulent fluid motion. The main limitations are the lack of powerful, tunable, UV lasers that could utilize the weak transitions of the more "exotic" species and also interference of more common species.

Quantitative concentration measurements of combustion intermediates in engines are difficult and correspondingly

Table 4. Qualitative overview of the most commonly used schemes for fluorescence measurements of combustion intermediates in engines.

Species	λ_{exc} (nm)	λ_{det} (nm)	Laser	Indicative for	Signal
OH	(248) 283	310	(KrF excimer) dye, OPO	High T reactions	Strong
H_2CO	353, 355	400–475	dye, Nd:YAG	Low T reactions	Strong
NO	226 (248)	225–300	OPO (KrF excimer)	High T, pollutant	Weak
CO	230	484	OPO	Incomplete or rich combustion	Very weak
C_2	355, 532	435–640	Nd:YAG	Soot and its precursors	Medium

The excitation and detection wavelength given in columns 2 and 3, respectively, are not the only possible choices. Like all other entries, they merely represent the dominant literature values. Entries in parentheses represent less common choices.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

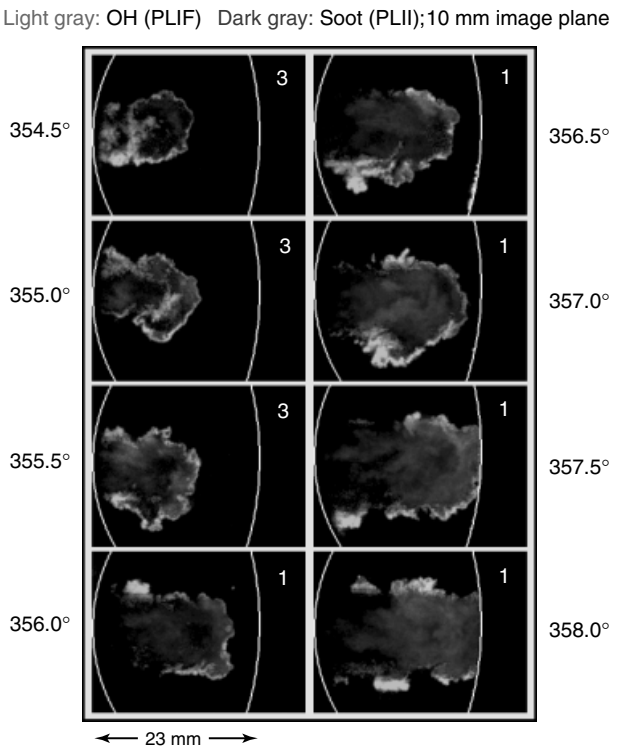


Figure 11. Series of simultaneous OH PLIF (light gray) and soot PLII (dark gray) images. The numbers beside the images indicate the crank angle of acquisition, with 360° at top-dead center. The (white) curve at the left marks the edge of a window in the cylinder head and the curve at the right marks the piston bowl wall. (Reproduced with permission from Dec and Tree, 2001. © J.E. Dec and D.R. Tree.)

uncommon. Major challenges are the lack of literature data on the species' high pressure, high temperature photophysics, difficulties in calibrating against a known concentration (particularly for unstable intermediates), complicated dependencies of fluorescence on usually unknown local temperature and gas composition.

Fortunately, much can be learned from qualitative, spatially resolved measurements. Even more information can be extracted from combined measurements, for which LIF is very suitable, because often the signal is limited to a narrow spectral region and can be separated from other, simultaneously detected channels. For instance, consider the images in Figure 11. We learn *where* OH is located in the spray flame, *when* it appears during the injection, and *where* it is in *relation* to the soot. From this, we can infer that the edge of the spray flame is hot and stoichiometric to slightly fuel−lean and that this remains the case throughout the entire injection. Furthermore, OH and soot have very little spatial overlap, which indicates that the OH-containing edge is likely actively oxidizing soot.

2.5.2 Fuel and fuel tracers

Most commercial fuels, with the notable exception of biodiesel, emit strong fluorescence when illuminated by UV lasers. Therefore, LIF can be used to detect both liquid- and gas-phase fuels with good sensitivity. Two-dimensional, PLIF is possible with several laser types, including excimer or Nd:YAG lasers. In some circumstances, conversion of the LIF signal to fuel concentration is possible; however, in general, quantification is difficult because commercial fuels are mixtures that contain multiple fluorescing components, many of which have unknown photophysical properties.

In quantitative LIF measurements, a nonfluorescing *model fuel* is typically doped with a *tracer*, which exhibits fluorescence with known characteristics. Ideally, the tracer behaves like the fuel in all physical and chemical aspects and the mixture of nonfluorescing model fuel and the tracer behaves like the commercial fuel it is supposed to model. Conveniently, none of the aliphatic components in the traditional primary reference fuels for gasoline and diesel (*n*-heptane + iso-octane and *n*-cetane + iso-cetane, respectively) fluoresces significantly at near UV wavelengths. The most common types of tracers are one- or two-ring aromatics, ketones, aldehydes, and amines. The aromatic compounds also are also found in significant quantities in commercial fuels. Unlike the small combustion intermediate molecules with narrow absorption lines in Table 4, the larger tracer molecules have broad absorption spectra, covering tens of nanometers. Therefore, complicated and weak tunable lasers are usually not necessary. Tracer LIF is most often used to image the mixing of evaporated fuel and intake air, as for example shown in Fuel Introduction.

In engine applications, the photophysical processes of LIF are dependent on excitation wavelength, gas composition, pressure, and temperature. Each of these dependencies can be an obstacle in quantification of the measurements but, once known, can also give the opportunity to measure the quantity of influence. For example, the dependence of toluene fluorescence on temperature has been characterized in shock tube and flow cell experiments (Koban *et al*., 2004). It can be exploited to determine that quantity based on either the redshift of the emission spectrum with increasing temperature (Luong *et al*., 2008) or the overall decrease of signal with increasing temperature (Dec and Hwang, 2009).

Two important limitations in the applicability of tracer LIF are the presence of liquid-phase fuel (spray) and the onset of combustion reactions. Fluorescence from a simultaneously present liquid phase is hard to quantify, although techniques exist to at least reliably separate it from gas-phase fluorescence ("Exciplex" fluorescence). The chemical

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

reactions of combustion convert the tracer via usually unquantifiable intermediates into nonfluorescing species.

2.6 Laser-induced incandescence of soot

As described in Section 2.1.2, soot heated to high temperatures by combustion emits light by incandescence, and the intensity of the incandescence increases nonlinearly with temperature. For instance, according to Equation 6, a 10% increase in temperature from 2400 to 2500 K results in a 60% increase in light emission at 500 nm (green color). Using high energy pulsed lasers, the soot particles can be heated by absorbing laser light to much higher temperatures. The laser heating limit is the vaporization temperature of the soot, which is approximately 4000–4500 K. At 4000 K, the emission at 500 nm increases by a factor of 120 relative to 2400 K. As a result, LII emission captured over the duration of the short (typically ~10 ns) high power laser pulse can be dominated by the laser-heated soot.

In theory, the LII technique can be quantitative. The soot temperature should remain at a steady level during vaporization (akin to boiling water), so that the signal becomes independent of laser intensity. Hence, above the vaporization threshold intensity, the LII emission should be independent of the local laser intensity. At a constant temperature, the LII signal is approximately proportional to the soot volume fraction.

One advantage of the LII technique over other line-of-sight techniques such as two-color soot incandescence technique described in Section 2.1.2 is that the laser-induced emission arises only from sooty regions that intersect with the path of the laser beam. Hence, LII of soot can provide spatial resolution along the line of sight of the detector.

To achieve the laser light intensities required to heat soot particles to their vaporization temperature, even high power beams typically must be focused on a small region. For imaging techniques, the laser beam is focused into a thin sheet (1 mm thick or less) for PLII of soot. Figure 11 shows a series of composite images of PLII of soot (dark gray) with PLIF of OH (light gray) for one of the eight diesel jets in an optical diesel engine (Dec and Tree, 2001).

In addition to the images in Figure 11, a few other notable examples of soot LII measurements are soot LII in a DI diesel engine (Pinson *et al*., 1993), combined soot extinction and LII in a gasoline direct injection engine (de Francqueville, Bruneaux, and Thirouard, 2010), and high speed soot LII in a laboratory burner and an optical engine (Sjöholm *et al*., 2011).

Although soot LII can be quantitative in principle, several factors introduce significant uncertainty. The vaporization temperature signal plateau predicted by theory is not observed in practice because of nonuniform laser beam intensity distributions and obliteration of soot particles by excessive vaporization at high laser beam intensities. Interferences from other effects, including fluorescence of photolysis products such as C_2 and C_3, as well as fluorescence of polycyclic aromatic hydrocarbon soot precursor species, can also spectrally and temporally overlap with the LII emission. In addition, while the LII emission is much stronger than the combustion-heated soot emission within the measurement volume, integration of the combustion-heated soot along the whole line of sight from the detector can be a significant interference, depending on the soot cloud geometry.

2.7 Velocimetry

In-cylinder convection and turbulence are two flow properties of paramount importance to engine operation because they determine transport and mixing, discussed in detail in In-Cylinder Flow. Two families of diagnostics have been applied in engines to measure velocity. The two-dimensional techniques are particle image velocimetry (PIV) and particle tracking velocimetry (PTV), and a point-measurement technique is laser-Doppler velocimetry (LDV).

The velocity of spray droplets can be measured directly. On the other hand, in engines, gas-phase velocity almost always needs to be measured indirectly using small particles seeded into the intake gas. Seeding is far from trivial. Particles need to be

- present at all measurement locations and timings in the right density, which is dependent on the technique;
- small enough to follow the flow with the required fidelity, but large enough to enough scatter light for reliable detection;
- not interfere with the measurement by fouling the windows or disturbing the flow itself;
- survive the high temperatures of compression or even combustion as required by the measurement timing.

In fact, ongoing development aside, the diagnostics described later are fairly mature. The challenge in engines is proper seeding. Typical seed particles are droplets from liquids with high boiling point, such as silicone oil or di-ethylhexyl-sebacate (DEHS), or (less commonly) solid particles from silicon dioxide, graphite, or hollow glass spheres. Liquid seeding does not survive combustion, which

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

can also be an advantage, because the absence of seeding then approximately marks the burned gas.

For engines, particles with diameters on the order of 20 μm follow the mean bulk flow with sufficient fidelity, while investigating small-scale turbulence or highly rotational flows may require particle sizes of less than 1 μm. If quantitative results are the goal, a case-based comparison of the expected accelerations (i.e., particle inertia) and aerodynamic drag is required.

The development of velocimetry has greatly profited from the broad range of applications. Uniquely among the diagnostics discussed here, complete systems including lasers, detectors, and computer units for timing, data acquisition, and data evaluation are commercially available from several companies. This should not detract from the fact that reliable measurements in engines are still not easy.

2.7.1 PIV and PTV

Both PIV and PTV are based on the same principle. Particles in an extended probe volume, typically a plane of about 1 mm thickness, are illuminated by two (or more) laser pulses separated by known time interval(s). An image of the light scattering from the particles is acquired with each laser pulse. The measured particle displacement divided by the interpulse time yields velocity vectors throughout the field of view. In the basic configuration, the two in-plane components of the velocity vector are determined, but extensions exist for capturing all three, and potentially not only in a plane but also within a volume. Extracting more information requires more cameras and better optical access, which limits the applicability in engines.

The difference between PIV and PTV is in the data evaluation, which corresponds to a difference in seeding density. In PTV, the seeding density is low, thus bright spots (i.e., particles) in the images are sparse. This allows for tracking the displacement of individual particles between the two images with little ambiguity. Each particle yields one velocity vector. In PIV, seeding is dense in the sense that the displacement of individual particles cannot be determined unambiguously. Instead, the image is partitioned into many "interrogation areas," each large enough to contain at least several particles. Via cross correlation of two corresponding interrogation areas in the two images, the average displacement within the area is calculated. Each velocity vector relies on data from all particles in the interrogation area, increasing precision. In addition, PIV can achieve higher resolution when seeding is sufficiently dense. When dense seeding cannot be achieved, PTV is advantageous. Hybrid techniques for spatially very inhomogeneous seeding exist. Figure 6 in In-Cylinder Flowshows examples of flow fields acquired by PIV.

In engines, viewing the probe volume through curved windows introduces particular challenges for PIV and PTV. As for measurements of scalar quantities, distortion needs to be corrected. However, the cross-correlation algorithm relies on information on the pixel level. Residual aberrations, inevitable in a case like that shown in Figure 2, degrade the accuracy of the cross correlation and thereby the velocity measurement. Distortion also can result in significant bias of the two measured in-plane velocity components by the unknown out-of-plane component (Miles, Hildingsson, and Hultqvist, 2007).

2.7.2 LDV and PDA

While PIV and PTV provide instantaneous measurements across an extended field of view, LDV delivers a nearly continuous stream of velocity data from particles passing through a small, point-like probe volume. In LDV, also called *laser Doppler anemometry* (LDA), two coherent laser beams of the same (or very nearly the same) wavelength are focused into the probe volume at a small angle such that an interference pattern is formed. Particles moving through this spatial intensity modulation cause temporally varying scattering, which is detected by receiving optics. The transverse component of the particle velocity can be determined from the modulation frequency. Each passing particle yields one data point. The data rate depends on flow velocity and seeding density, but rates of tens of kilohertz are achievable. Extensions with multiple lasers and detectors can measure more than one vector component.

Another important extension of LDV is phase Doppler anemometry (PDA), applicable to transparent particles such as droplets from a spray (Sankar *et al*., 1991). In this technique, a suitably located second receiver allows for determination of droplet size in addition to velocity. PDA is a powerful tool for the quantitative study of injection sprays and as such is indispensable for the development of computational spray models.

REFERENCES

Aleiferis, P.G., Serras-Pereira, J., van Romunde, Z., *et al*. (2004) The nature of early flame development in a lean-burn stratified-charge spark-ignition engine. *Combustion and Flame*, **157** (4), 735–756.

Aronsson, U., Chartier, C., Horn, U., *et al*. (2008) Heat release comparison between optical and all-metal HSDI diesel engines. SAE Paper 2008-01-1062.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Aronsson, U., Solaka, H., Chartier, C., *et al*. (2011) Impact of mechanical deformation due to pressure, mass, and thermal forces on the in-cylinder volume trace in optical engines of Bowditch design. SAE Paper 2011-26-0082.

Augusta, R., Foster, D.E., Ghandhi, J.B., *et al*. (2006) Chemiluminescence measurements of homogeneous charge compression ignition (HCCI) combustion. SAE Paper 2006-01-1520.

Berrocal, E., Kristensson, E., Richter, M., *et al*. (2008) Application of structured illumination for multiple scattering suppression in planar laser imaging of dense sprays. *Optics Express*, **16**, 17870–17881.

Birkigt, A., Michels, K., Theobald, J., *et al*. (2011) Investigation of compression temperature in highly charged spark-ignition engines. *International Journal of Engine Research*, **12** (3), 282–292.

Bowditch, F.W. (1961) A new tool for combustion research: a quartz piston engine. *SAE Transactions*, **69**, 17–23. SAE Paper 610002.

Colban, W., Kim, D., Miles, P.C., *et al*. (2008) A detailed comparison of emissions and combustion performance between optical and metal single-cylinder diesel engines at low temperature combustion conditions. *SAE Transactions*, **117** (4), 505–519. SAE Paper 2008-01-1066.

Dec, J.E. and Hwang, W. (2009) Characterizing the development of thermal stratification in an HCCI engine using planar-imaging thermometry. *SAE International Journal of Engines*, **2**, 421–438. SAE Paper 2009-01-0650.

Dec, J.E. and Tree, D.R. (2001) Diffusion-flame/wall interactions in a heavy-duty DI diesel engine. *SAE Transactions*, **110** (3), 1618–1634. SAE Paper 2001-01-1295.

Dec, J.E., Hwang, W., and Sjöberg, M. (2006) An investigation of thermal stratification in HCCI engines using chemiluminescence imaging. *SAE Transactions*, **115** (3), 759–776. SAE Paper 2006-01-1518.

Donovan, M.T., He, X., Zigler, B.T., *et al*. (2004) Demonstration of a free-piston rapid compression facility for the study of high temperature combustion phenomena. *Combustion and Flame*, **137**, 351–365.

de Francqueville, L., Bruneaux, G., Thirouard, B. (2010) Soot volume fraction measurements in a gasoline direct injection engine by combined laser induced incandescence and laser extinction method. *SAE International Journal of Engines*, **3**, 163–182. SAE Paper 2010-01-0346.

Fuyuto, T., Matsumoto, T., Yoshiaki, H., *et al*. (2012) A new generation of optically accessible single-cylinder engines for high-speed and high-load combustion analysis. *SAE International Journal of Fuels and Lubricants*, **5** (1), 307–315. SAE Paper 2011-01-2050.

Gasoline Fuel Injection Standards Committee SAE Standard J2715 (2007) Gasoline fuel injector spray measurement and characterization.

Ghandhi, J.B. and Heim, D.M. (2009) An optimized optical system for backlit imaging. *Review of Scientific Instruments*, **80** (5), 056105.

Idicheria, C.A. and Pickett, L.M. (2005) Soot formation in diesel combustion under high-EGR conditions. *SAE Transactions*, **114** (4), 1559–1574. SAE Paper 2005-01-3834.

Idicheria, C.A. and Pickett, L.M. (2007) Quantitative mixing measurements in a vaporizing diesel spray by Rayleigh imaging. *SAE Transactions*, **116** (3), 490–504. SAE Paper 2007-01-0647.

Ikeda, Y., Kaneko, M. and Nakajima, T. (2001) Local A/F measurement by chemiluminescence of OH^*, CH^* and $C2^*$ in SI engine. SAE Paper 2001-01-0919.

Kashdan, J.T. and Thirouard, B. (2009) A comparison of combustion and emissions behaviour in optical and metal single-cylinder diesel engines. *SAE International Journal of Engines*, **2**, 1857–1872. SAE Paper 2009-01-1963.

Kastengren, A.L., Powell, C.F., Wang, Y.J., *et al*. (2009) X-Ray radiography measurements of diesel spray structure at engine-like ambient density. *Atomization and Sprays*, **19**, 1031–1044.

Koban, W., Koch, J.D., Hanson, R.K., and Schulz, C. (2004) Absorption and fluorescence of toluene vapor at elevated temperatures. *Physical Chemistry Chemical Physics*, **6**, 2940–2945.

Kokjohn, S., Musculus, M.P.B. and Reitz, R.D. (2012) Investigation of fuel reactivity stratification for controlling PCI heat-release rates using high-speed chemiluminescence imaging and fuel tracer fluorescence. SAE Paper 2012-01-0375.

Linne, M.A., Paciaroni, M., Berrocal, E., and Sedarski, D. (2009) Ballistic imaging of liquid breakup processes in dense sprays. *Proceedings of the Combustion Institute*, **32**, 2147–2161.

Luong, M., Zhang, R., Schulz, C., and Sick, V. (2008) Toluene laser-induced fluorescence for in-cylinder temperature imaging in internal combustion engines. *Applied Physics B*, **91** (3-4), 669–675.

Metghalchi, M. and Keck, J. (1982) Burning velocities of mixtures of air with methanol, isooctane, and indolene at high pressure and temperature. *Combustion and Flame*, **48**, 191–210.

Miles, P.C., Hildingsson, L., and Hultqvist, A. (2007) The influence of fuel injection and heat release on bulk flow structures in a direct-injection, swirl-supported diesel engine. *Experiments in Fluids*, **43** (2-3), 273–283.

Mueller, C.J. and Martin, G.C. (2002) Effects of oxygenated compounds on combustion and soot evolution in a DI diesel engine: broadband natural luminosity imaging. *SAE Transactions*, **111** (4), 518–527. SAE Paper 2002-01-1631.

Musculus, M.P.B., Singh, S., and Reitz, R.D. (2008) Gradient effect on two-color soot optical pyrometry in a heavy-duty DI diesel engine. *Combustion and Flame*, **153** (1-2), 216–227.

Naber, J.D. and Siebers, D.L. (1996) Effects of gas density and vaporization on penetration and dispersion of diesel sprays. *SAE Transactions*, **105** (3), 82–111. SAE Paper 960034.

Oren, D.C., Wahiduzzaman, S. and Ferguson, C.R. (1984) A diesel combustion bomb: proof of concept. SAE Paper 841358.

Pastor, J.V., Garcia, J.M., Pastor, J.M., and Buitrago, J.E. (2005) Analysis methodology of diesel combustion by using flame luminosity, two-color method, and laser-induced incandescence. SAE Paper 2005-24-012.

Pawlowski, A., Kneer, R., Lippert, A., and Parrish, S.E. (2008) Investigation of the interaction of sprays from clustered orifices under ambient conditions relevant for diesel engines. *SAE Transactions*, **514** (4), 514–527. SAE Paper 2008-01-0928.

Pickett, L.M., Kook, S., and Williams, T.C. (2009) Visualization of diesel spray penetration, cool-flame, ignition, high-temperature combustion, and soot formation using high-speed imaging. *SAE*

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

International Journal of Engines, **2** (1), 439–459. SAE Paper 2009-01-0658.

Pickett, L.M., Manin, J., Genzale, C.L., *et al*. (2011) Relationship between diesel fuel-jet vapor penetration/dispersion and local fuel mixture-fraction. *SAE International Journal of Engines*, **4**, 764–799. SAE Paper 2011-01-0686.

Pinson, J.A., Mitchell, D.L., Santoro, R.J. and Litzinger, T.A. (1993) Quantitative, planar soot measurements in a D.I. diesel engine using laser-induced incandescence and light scattering. SAE Paper 932650.

Salazar, V. and Kaiser, S. (2011) Influence of the flow field on flame propagation in a hydrogen-fueled internal combustion engine. *SAE International Journal of Engines*, **4** (2), 2376–2394. SAE Paper 2011-24-0098.

Sankar, S.V., Weber, B.J., Kamemoto, D.Y., and Bachalo, W.D. (1991) Sizing fine particles with the phase Doppler interferometric technique. *Applied Optics*, **30**, 4914–4920.

Settles, G.S. (2001) *Schlieren and Shadowgraph Techniques*, Springer-Verlag.

Siebers, D.L. and Higgins, B.S. (2001) Flame lift-off on direct-injection diesel sprays under quiescent conditions. *SAE Transactions*, **110** (3), 400–421. SAE Paper 2001-01-0530.

Sjöholm, J., Wellander, R., Bladh, H., *et al*. (2011) Challenges of in-cylinder high-speed two-dimensional laser-induced incandescence measurements of soot. *SAE International Journal of Engines*, **4** (1), 1607–1622. SAE Paper 2011-01-1280.

Smith, J. and Sick, V. (2005) Crank-angle resolved imaging of fuel distribution, ignition and combustion in a direct-injection spark-ignition engine. *SAE Transactions*, 114.

Svensson, K.I., Mackrory, A.J., Richards, M.J. and Tree, D.R. (2005) Calibration of an RGB, CCD camera and interpretation of its two-color images from KL and temperature. SAE Paper 2005-01-0648.

Watanabe, Y., Morikawa, K., Kuwahara, T., and Tanabe, M. (2008) Evaluation of homogeneous charge compression ignition at high engine speeds using a super rapid compression machine. SAE Paper 2008-01-2403.

Williams, T.C., Shaddix, C.R., Jensen, K.A., and Suo-Anttila, J.M. (2007) Measurement of the dimensionless extinction coefficient of soot within laminar diffusion flames. *International Journal of Heat and Mass Transfer*, **50** (7-8), 1616–1630.

Xu, Y. and Lee, C. (2005) Investigation of fuel effects on soot formation using forward illumination light extinction (FILE) technique. SAE Paper 2005-01-0365.

Eckbreth, A.C. (1996) *Laser Diagnostics for Combustion Temperature and Species*, Gordon and Breach, Amsterdam.

Espey, C., Dec, J.E., Litzinger, T.A., and Santavicca, D.A. (1997) Planar laser Rayleigh scattering for quantitative vapor-fuel imaging in a diesel jet. *Combustion and Flame*, **109**, 65–86.

Engine Combustion Network. http://www.sandia.gov/ECN.

Gaydon, A.G. (1974) *The Spectroscopy of Flames*, 2nd edn, Chapman and Hall.

Guibert, P., Keromnes, A. and Legros, G. (2007) Development of a turbulence controlled rapid compression machine for HCCI combustion. SAE Paper 2007-01-1869.

Kohse-Höinghaus, K. (1994) *Progress in Energy and Combustion Science*, **20** (3), 203–279. DOI: 10.1016/0360-1285(94)90015-9.

Musculus, M.P.B. and Pickett, L.M. (2005) Diagnostic considerations for optical laser-extinction measurements of soot in high-pressure combustion environments. *Combustion and Flame*, **141** (4), 371–391.

Nori, V. and Seitzman, J. (2008) Evaluation of chemiluminescence as a combustion diagnostic under varying operating conditions. AIAA Paper 2008-953.

Raffel, M., Willert, C., Wereley, S., and Kompenhans, J. (2007) *Particle Image Velocimetry: A Practical Guide*, Springer, New York.

Santoro, R.J. and Shaddix, C.R. (2002) *Laser-induced incandescence* in *Applied Combustion Diagnostics*, *Chapter 9* (eds K. Kohse-Höinghaus and J.B. Jeffries), Taylor and Francis.

Schulz, C., Kock, B.F., Hofmann, M., *et al*. (2006) Laser-induced incandescence: recent trends and current questions. *Applied Physics B*, **83**, 333–354.

Schulz, C. and Sick, V. (2005) Tracer-LIF diagnostics: quantitative measurement of fuel concentration, temperature, and fuel/air ratio in practical combustion systems. *Progress in Energy and Combustion Science*, **31** (1), 75–121. DOI: 10.1016/j.pecs.2004.08.002.

di Stasio, S. and Massoli, P. (1994) Influence of the soot property uncertainties in temperature and volume-fraction measurements by two-color pyrometry. *Measurement Science and Technology*, **5**, 1453–1465.

Zhao, H. and Ladommatos, N. (1998) Optical diagnostics for soot temperature measurement in diesel engines. *Progress in Energy and Combustion Science*, **24**, 221–255.

Zhao, H. and Ladommatos, N. (2001) *Engine Combustion Instrumentation and Diagnostics*, Chapter 11, Society of Automotive Engineers.

FURTHER READING

Baert, R., Frijters, P., Somers, B. and Luijten, C. (2009) Design and operation of a high pressure, high temperature cell for HD diesel spray diagnostics: guidelines and results. SAE Paper 2009-01-0649.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto132
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Material and Process Selection—Cylinder Blocks and Heads

Marc C. Megel and George E. Bailey
Southwest Research Institute, San Antonio, TX, USA

1 Introduction 1
2 Material Selection 2
3 Casting Process Development 8
Reference 10
Further Reading 10

1 INTRODUCTION

Cylinder heads and blocks are subjected to cyclic pressure and thermal loading. There are also assembly loads that must be withstood. In blocks, the cylinder head is clamped tightly to the block with head bolts, and there is also pressure from the head gasket that must seal the combustion pressure. The main bearing caps are installed at a high torque to carry the loads from the crankshaft. The cylinder head is also subjected to the loads from the head bolts and head gasket. Valve seats and valve guides are often made of a different material than the head, and are installed with interference fits.

These assembly loads that may at first seem constant are also cyclically loaded as the members react loads from both combustion and changes in dimensions because of thermal expansion. Thermal gradients can be great, for example, when a cold engine is started. When an engine is started, the combustion gases rapidly heat the walls of the combustion chamber. The walls will try to expand because of the nonzero thermal expansion coefficient of the material, but the bulk material of the block and head that is still cold will constrain this expansion putting the bulk material in tension. The material must have sufficient strength to withstand this tension, on top of the forces from combustion and assembly loads. If the material has high thermal conductivity, the heat will travel through the block and head faster and these stresses will be reduced. This brief discussion begins to illustrate how the properties of the block and head interact. If the material has a higher thermal expansion coefficient, it will need higher strength. However, if thermal conductivity is high, then the strength of the material will not need to be quite so high.

The design of the head and block influences the material choice, as thinner sections will have lower thermal gradients but will require stronger material. As the loading is cyclical, the material will need to possess adequate fatigue life at all operating temperatures. There are three fatigue cases that need to be considered. The engine is subjected to high cycle fatigue (HCF) loading from combustion forces. The structure must be designed for infinite HCF life. The engine is subjected to low cycle fatigue (LCF) loading during start-up. The LCF durability target needs to be chosen according to a projected number of cycles determined for a particular application. Thermal shock loading also needs to be considered. This occurs when the engine is started from its coldest condition and needs to make maximum power rapidly.

Except in rare cases, the cost of the engine must be kept as low as practically possible for the manufacturer to capture and maintain a share in the market. The block and cylinder head have complicated shapes as they must house the moving components, have passages for coolant and lubricant, and provide mounting points for accessories,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto135
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

pumps, manifolds, fuel injectors, ignition systems, and so on. In most cases, it is impractical to fabricate blocks and heads from plates or machine them from solid, so castings are employed. In the casting process, molds are prepared, which include cores. The inside of the mold forms the outer shape of the component to be cast. Cores form the interior passages, such as water jackets and gas ports. When a component is cast, liquid metal alloy is either poured or injected into the mold, where it flows into the mold around the cores until the cavity is filled, and all of this is happening as the molten alloy is cooling. The resulting casting must have consistent properties and not possess significant voids or porous regions. The cooling rate of the metal alloy will affect its material properties, so that rate must be understood and controlled. The collective behavior of materials during the casting process is referred to as *castability*. Some alloys behave well when cast in thin sections, whereas others are required for thick sections. Of course, these different alloys have different mechanical properties, which must be considered during the design of the component.

The choice of materials for blocks and cylinder heads is thus a very involved and complicated process. The proper choice is the difference between an engine that only runs well when it is new, and a durable engine that has consistent performance over its intended lifetime.

2 MATERIAL SELECTION

The primary choices for engine blocks and heads are cast irons and aluminum alloys. Magnesium alloys are sometimes used, but are not as common. A recent example would be the hybrid magnesium/aluminum cylinder block casting produced by BMW, which claims to reduce the block weight by 24% versus 100% aluminum alloy using a magnesium skeleton. Cast irons do not conduct heat as well as aluminum or magnesium alloys; they take longer to reach operating temperature, and they are heavier but have significantly higher tensile and fatigue strengths depending on the grade. Typical densities of cast iron and aluminum are 7.2 g/cm^3 and 2.7 g/cm^3, respectively, so the same volume of cast iron is almost three times as heavy. While the strength of cast iron and aluminum casting alloys varies significantly, the strength of both cast irons and heat-treated aluminum alloys typically used in lower cost cylinder heads is 250 MPa. The raw material cost for common gray cast iron grades can be roughly half the cost per ton of common aluminum alloys. The specific heat of cast iron (0.46 kJ/kg-K) is roughly half that of aluminum (0.96 kJ/kg-K), but the melting point is almost double (1200°C for cast iron and 600°C for aluminum), so the energy cost per ton is similar. As aluminum castings are lighter, the energy cost per casting will be less, but aluminum castings often require thicker walls than cast iron (high peak cylinder pressure (PCP) applications), which complicates energy and material cost comparisons. Differences in costs of molds, machining, and yield must also be considered. There are cases where cast iron blocks and heads can be cheaper to manufacture than aluminum castings. These are primarily cases where (i) volume does not support permanent die molding, (ii) regional technology for product localization does not support high quality aluminum casting, or (iii) PCPs in the engine start become high enough to require significantly more material in aluminum to have the same structural integrity, noting that some applications may require cast wall thicknesses beyond what can be achieved with aluminum alloys. For light-duty, lower PCP automotive or industrial castings, aluminum can possess significant piece price advantages, primarily through high volume and the use of die molds. Cast iron also possesses higher internal damping than the nonferrous alloys, which influences engine NVH.

2.1 Specific weight

Specific weight is a key consideration when selecting a cylinder block or head material during a new engine design program. The units of specific weight are mass/power (e.g., kW/kg). This parameter helps a designer consider the impact performance requirements such as high PCP have on the overall design and the resultant weight of an engine and not just the volume and density of the material. Therefore, an optimal design should have the highest specific weight, but may not be cast from the lightest material (this is especially common for highly boosted diesel and gasoline engines).

2.2 Geometric considerations

When considering cylinder block materials for highly boosted engines, one is likely to find trade-offs between packaging dimensions and cylinder spacing. A good illustration of this comes from a comparison of the Mercedes 4.0L V8 with an aluminum block to the Audi 4.2L V8 with a CGI block (Dawson and Indra, 2007). The larger Audi engine exhibits a specific weight of 0.94 versus 0.89 for the Mercedes with a 7 mm smaller bore spacing and 120 mm shorter overall length. Each factor is extremely important when designing a light-duty, passenger car engine.

As the bore diameter of an engine decreases, the structure becomes inherently stiffer and deflects less under cylinder pressure. This deflection is what generates HCF in the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto135
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

cylinder head and block. It is this phenomena, coupled with a light-duty cycle, that allows high speed direct injection (HSDI) engines to still operate at cylinder pressures in the 180–200 bar range reliably, even with the reduced tensile and fatigue properties of aluminum.

A final geometric consideration is desired/required wall thickness. Typically, cylinder heads utilizing cast iron have a minimum production wall thickness limit of 5–6 mm with some allowances for locally thinner sections. It is possible to design cylinder heads in cast iron around a 4 mm general wall thickness, but these designs require significant development time in the foundry to ensure quality and may have cost trade-offs in scrap to achieve the lighter weight component. Aluminum can typically be sand cast in a cylinder head with a minimum wall thickness of 3–3.5 mm.

2.3 Power density

Power density, power/engine displacement (e.g., kW/L), is another critical parameter to consider when selecting materials, especially for cylinder heads. Heavy-duty diesel (HDD) engines, which normally run cylinder pressures over 200 bar, usually target specific powers of 30–35 kW/L to meet the extended durability (typically 1 million miles between overhaul) and reliability requirements of the demanding duty cycle for that application. At these levels, cast iron has sufficient thermal conductivity to transfer the heat generated by combustion. At around 40 kW/L, cast iron becomes limited in its ability to transfer sufficient heat out of the cylinder head to maintain fire deck temperatures at the material limit ($<400^{\circ}$C) even with the consideration of mixed-mode cooling (forced convection plus nucleate boiling). At these levels, a limited-duty cycle and reduced durability requirement for higher cylinder pressure operation verses heavy duty must also be assumed. At or above this level, aluminum should be considered for light-duty engine designs. These applications are also typically very weight sensitive, which also makes aluminum attractive. State-of-the-art gasoline and high speed, light-duty diesels (HSDI) operate at specific powers in the range 75–100 kW/L. At these levels, aluminum must be utilized as the cylinder head material to transfer sufficient heat from the fire deck. Even with the high thermal conductivity of aluminum verses cast iron, mixed-mode cooling will likely be prevalent at these levels.

2.4 Overview of aluminum alloys

Alloying elements are added to aluminum to change or enhance its properties. Alloying elements have different degrees of solubility in aluminum and diffuse at different rates. The combination of these properties allows various alloying elements to have different influences on aluminum alloy properties. Major aluminum alloying elements are zinc, copper, manganese, magnesium, silicon, and iron. Zinc and copper are used in high strength alloys, and manganese, magnesium, silicon, and iron are used to tailor strength, ductility, and toughness. Silicon also improves the fluidity of molten aluminum. Iron hurts ductility but improves mold-sticking behavior. Adding manganese to half the levels of iron corrects iron's influence on ductility. Copper adds high temperature strength but reduces ductility, corrosion resistance, and thermal conductivity.

Aluminum casting and wrought alloys are generally considered separately because of the requirements for castability and hot formability. Therefore, there is a different designation system for each. The European, US, and ISO designations for wrought alloys are similar but there are different systems for cast alloys. The US (Aluminum Association or AA) system will be used in this chapter. AA cast alloy designations have a decimal, which designates whether the alloy properties being reported are for castings (.0), ingot (.1), or ingot with narrow composition ranges (.2). All of the properties in this report apply to castings so the .0 suffix is implied and not written. The AA temper designation system is also used in this chapter.

Aluminum alloys classified by the AA system are given a three-number designation, where the first number represents the combination of major alloying elements. The 100 series are pure aluminums with less than 1% alloying elements. The 200 series are aluminum with copper. The 300 series are aluminum and silicon with magnesium, copper, or both as major alloying elements. The 400 series is aluminum and silicon, the 500 series is aluminum and magnesium, the 700 series is aluminum and zinc, and the 800 series is aluminum and tin. The 100 series have low strength and are primarily used where electrical conduction is important. The 800 series casting alloys are bearing grade aluminums. The 700 series alloys are age hardening alloys and do not have very good castability. The 500 series are used for automotive frames, cookware, and applications where a good surface finish is required. These alloys generally have poor pressure tightness, which makes them unattractive for use in engine blocks and cylinder heads. Some of the 400 series alloys are used for pistons. The majority of engine block and cylinder head castings are from the 300 series. As cylinder pressures and operating temperatures increase, however, there is greater interest in the 200 series, which maintain their high strength at elevated temperatures, although these alloys have limitations as is discussed below.

Some common aluminum alloys for light-duty engines are 319, 356, and 357. As cylinder pressures and temperatures increase, alloy 319 is being replaced by 356 as

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto135
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

that alloy allows tighter control of microstructure. Alloy 357 is similar to 356 but some beryllium is added to improve strength and high temperature performance. For high performance engines, these alloys are modified to enhance certain properties. Copper is added to increase strength, particularly at high temperatures, but it decreases thermal conductivity. The amounts of silicon and magnesium are also adjusted to achieve the desired balance of properties for a given application. To illustrate the differences between alloys, some properties at room temperature are compared to 319 alloy in Figure 1.

There are many variables to consider when casting aluminum that directly affect the alloy choice. Cooling the alloy faster generally produces a finer microstructure and higher strength. However, if the outside of the alloy cools too fast while the bulk is shrinking, it can tear or crack the surface. Hypoeutectic alloys (with silicon concentrations less than 11.7%) such as 390 will cool as they solidify, whereas hypereutectic alloys will tend to solidify isothermally. Therefore, it will be more difficult to cast large or complex shapes with hypoeutectic alloys. Hypereutectic alloys such as 390 will have a hard primary silicon phase, which adds to wear resistance. Compared to a hypoeutectic alloy such as 356, they have lower thermal expansion and lower thermal conductivity. Considering the case of a cylinder head, a decrease in thermal conductivity will result in a hotter combustion chamber surface, and higher thermal gradients. As strength and fatigue life decrease with temperature, these locally higher temperatures can result in lower durability.

The 200 series of casting alloys retain more strength at higher temperatures, but there are several challenges for their use in high volume production engines. Alloy 201 has exceptional strength after heat treatment and maintains this strength at high temperatures. Silver is added to improve

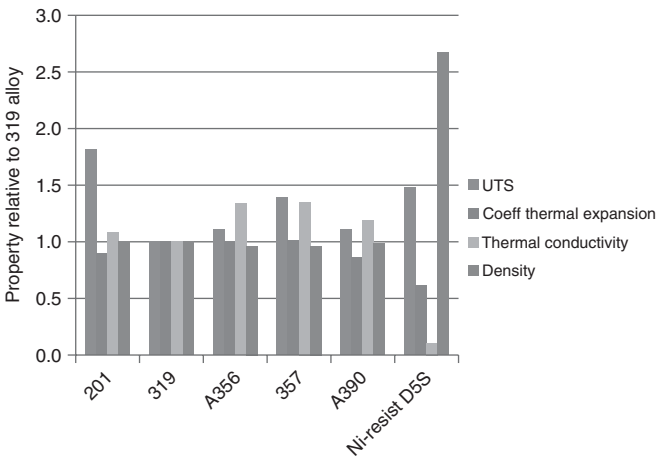


Figure 1. Room temperature properties relative to 319 alloy.

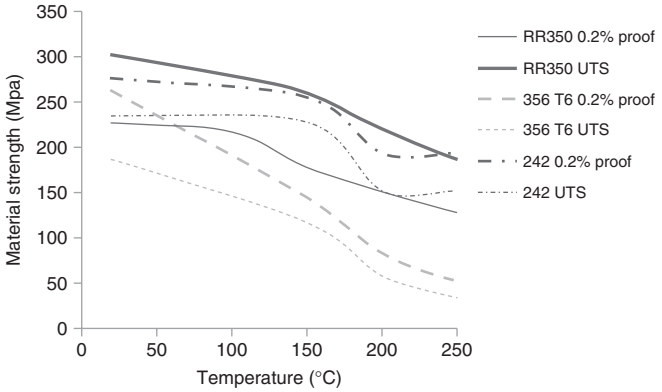


Figure 2. Strength versus temperature for three aluminum alloys.

resistance to stress corrosion cracking and to increase the response to heat treatment. This alloy is used for pistons and cylinder heads, and many aerospace applications. Alloys 242 and RR350 (in the same alloying family and also referred to as *Hiduminium* or *203*) are used for pistons and air-cooled cylinder heads. They do not contain silver such as alloy 201, so the price premium over 356 on a per-ton basis is not as great. They are low silicon alloys, so they maintain their mechanical properties at higher temperatures. Grainger and Worrall (G&W) Foundry in the United Kingdom estimated that the per ton raw material cost of RR350 would be 1.7 times that of 356. By contrast, 201 is five times that of 356 and a 200 series alloy specifically optimized for high performance engines is six times that of 356. The strength of 242 is similar to 356 but it retains much more of its strength up to 250°C as shown in Figure 2. Creep properties are better than 356 but the thermal conductivity is lower. It is more difficult to cast than 356, which could lead to increased production costs, but the biggest drawback is its lack of corrosion resistance. RR350 is stronger yet than 242 but also suffers from a lack of corrosion resistance.

2.5 Overview of cast iron alloys

Cast iron is a general term used to describe the family of iron-based casting alloys. One method of classifying the various types of cast iron is by the structure of the carbon in the microstructure. Gray iron contains lamellar graphite flakes. Ductile iron (DI) contains spheroidal graphite. Compacted or vermicular iron contains graphite that is interconnected within eutectic cells such as gray iron, but the graphite is coarser and more rounded, that is, its structure is between gray and DI. These three types of cast iron are the most common utilized in cylinder block and head castings. Within these three basic types of cast iron are

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto135
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

various strength grades, which are formed based on the quantity of alloying elements added to the melt.

2.5.1 Gray iron (GI)

Basic gray iron is probably one of the cheapest and most commonly used casting materials. Grade 250 (250 MPa tensile strength) iron is used for low cylinder pressure, low cost cylinder blocks and heads. This iron is typically good up to cylinder pressures of 160 bar and normally results in a heavier engine because of the addition of material to make up for the low tensile and fatigue properties. An exception, however, would be heavy-duty, in-line cylinder blocks (even at higher cylinder pressures) where the simplified geometric structure allows the use of a lower strength and lower cost material, and weight is not a chief design concern.

2.5.2 High strength gray iron (HSGI)

Alloyed high strength gray iron (HSGI) is the standard cylinder head material for HDD engines. This material has a flake graphite structure and derives its elevated tensile and fatigue strength levels from a minimized carbon equivalence (CE) and the use of alloying elements such as chromium (Cr) and molybdenum (Mo). While these elements are effective at increasing the strength of gray iron, there is a limit to which they can be applied and thus a limit to the maximum tensile strength, which can be achieved in gray iron for complex engine castings. Increasing the level of Mo significantly increases the risk of shrinkage porosity within the casting and thus increases production scrap rates. An increased level of Cr along with a reduced CE level increases the hardness of the material and introduces the risk of carbide formation. Therefore, production with these irons requires a foundry with good process and quality control. Typical tensile strength of these grades used in block and head castings ranges from 275 MPa for mild amounts of alloying to a maximum of 320 MPa. Fatigue strength of gray iron is highly influenced by graphite flake size, cooling rate, and the amount of alloying elements. Development of a material should target minimum fatigue strength of 42% of the target tensile strength (e.g., 135 MPa minimum fatigue strength for 320 HSGI). The casting should exhibit Type A flake graphite with a size of 3 or finer to help maximize the fatigue strength/tensile strength ratio. Figure 3 shows an example of evaluating graphite flake size during casting development.

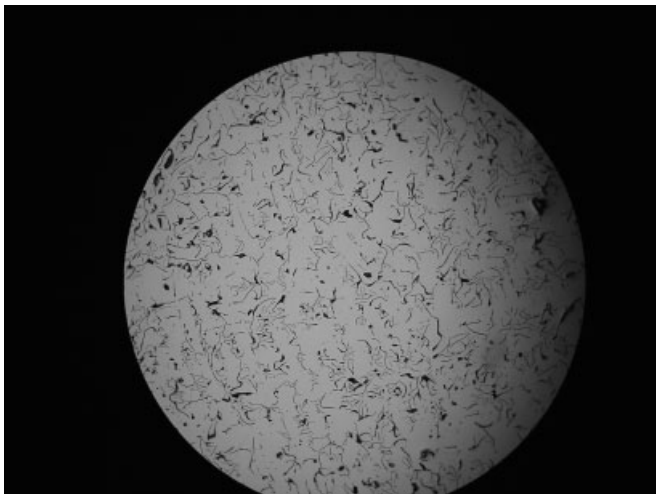


Figure 3. Graphite flake size evaluation during casting development (type A flake, size 2–4).

2.5.3 Ductile iron (DI)

DI is also known as *spheroidal iron* because of the shape of the graphite contained in the microstructure. DI has significantly improved tensile and fatigue strength over standard gray iron. DI derives its name from the fact that the spheroidal graphite structure makes the material significantly more ductile than gray iron where the interface between the flake graphite and the pearlitic structure essentially forms interstitial cracks within the material. While examples exist where DI has been used for unit cylinder heads (one head per engine cylinder), no slab heads have been cast in DI because of the issues associated with casting DI in a large volume, complex geometry. DI is most commonly used for exhaust manifolds and turbine housings in HDD engines as it also exhibits good high temperature fatigue properties when alloyed with silicon (Si).

2.5.4 Compacted graphite iron (CGI)

Compacted graphite iron is essentially a cross between flake gray and DI. The graphite formation yields a "worm"-shaped graphite structure. As a result, its physical and strength properties lie between gray and DI. CGI is created by inoculating the molten iron with magnesium (Mg) just before pouring into the casting mold pack. The resultant microstructure and thus strength properties are highly influenced by cooling rates within the mold pack. As cylinder heads have extremely nonuniform geometry and internal wall thicknesses, this results in a casting with varying material properties. In addition, if the iron cools too quickly, a lay of flake graphite can be formed on the surface where maximum fatigue properties are desired. This layer of flake graphite can reduce the effective fatigue strength by up to 20%. Even with this reduction, however, the fatigue strength is still higher than traditional HSGI. Controlling

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto135
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the variation of these properties is the biggest challenge for using CGI in block and head castings successfully.

Table 1 shows some example material properties obtained from separately cast test bars for versions of HSGI and CGI typically used for cylinder block and head castings. This data would be typical for castings that achieve a target UTS of 320 and 400 MPa respectively in critical sections.

One of the key factors associated with CGI or DI is the decrease in thermal conductivity as compared to high alloy gray iron. This reduction in thermal conductivity will cause an increase in peak fire deck temperatures when compared to high alloy gray iron. As a result, thermal stress in the cylinder head can increase for CGI or DI, which creates higher fatigue stress levels. Therefore, it may be difficult to realize the full strength advantage of these materials in cylinder head castings. In order to minimize the reduction in thermal conductivity of CGI versus HSGI, nodularity must be carefully controlled and minimized.

For cylinder blocks on the other hand, CGI presents many opportunities for reduced weight and package size in high cylinder pressure applications. Cylinder blocks typically exhibit more uniform material cross sections with larger sand cores for more uniform cooling versus cylinder heads, yielding more consistent material properties throughout the casting when using CGI. While thermal conductivity is still a concern for parent cylinder bore designs, the increased stiffness of CGI verses gray iron or aluminum allows for a significantly thinner cylinder wall thus counteracting the heat transfer impacts of reduced thermal conductivity while still retaining proper bore distortion characteristics. For wet-liner block designs such as heavy-duty and large-bore diesel/natural gas engines, thermal conductivity is not a concern as the cylinder heat transfer occurs through the wet-liner material. Another inherent advantage of using CGI versus gray iron or aluminum for cylinder block castings is the ability to have fractured main bearing caps because of the higher fatigue strength of the base material. Common cylinder block gray iron alloys do not exhibit the required fatigue strength required for the main bearing cap design.



Figure 4. Heavy-duty diesel main bearing cap comparison: machined-ductile iron cap with ring dowel locators (left) and fractured-CGI cap (right).

Therefore, gray iron blocks (especially for heavy-duty and large-bore diesel engines) normally use DI main bearing caps. This leads to additional cost, assembly, and machining steps in the manufacturing of the block. In addition, a common issue with high cylinder pressure blocks is main cap fretting. The use of a fractured cap eliminates all of these issues. Figure 4 shows a common, fully machined, DI main bearing cap from a heavy-duty diesel gray iron block verses a fractured CGI cap from a similar design.

2.5.5 Heat treatment

Typically, cast iron alloys will undergo some form of heat treatment process, minimally to stress relieve the casting. During solidification, the casting can develop significant residual stress, especially in the internal structure of cylinder head castings around injector and bolt boss sections. The most common form of stress relieving is in-mold stress relieving where the casting is left in the

Table 1. Example material properties for HSGI and CGI obtained from separately cast bars.

Material Property	Units	Material	
		G320	CG1400
UTS	MPa	360	483
Modulus of elasticity	GPa	115	132
Density	g/cm^2	7.2	7.1
Specific heat	J/g—°C	0.46	0.47
Thermal conductivity	W/m-K		
	At 100°C	46	39
	At 400°C	43	38

Data Courtesy of Grainger and Worrall, Ltd.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto135
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

sand mold for an extended period of time to allow the casting to cool slowly and minimize residual stress. For prototype castings, a furnace heating process is typically used as the prototype mold packs are often smaller and cool faster than the production molds thus not providing sufficient stress relief. Once broken out of the mold, the castings will be reheated and held at temperature until sufficient stress relief is achieved. This is a critical part of iron casting development as high residual stresses can lead to failures in parts at cylinder pressure and thermal loads much lower than what might be predicted by FEA.

2.6 Material testing

Material testing must be performed to qualify the casting process. Test bars are frequently cast to qualify the material chemistry, strength, microstructure, chosen cooling rate, and heat treatment. While test bars provide useful data, there are many influences on the final material property distribution in the castings that need to be understood to produce durable engines. It should be noted that textbook values for properties of cast materials are typically obtained from these separately cast specimens. The microstructure of the casting material is a function of the cooling rate, which will never be constant throughout a casting. While impurities can be present, voids, shrinkage porosity, and cracking will be present, and the distribution of the various phases of the alloy will never be completely evenly distributed. As these separately cast bars cool at a significantly faster rate than the actual casting, they typically do not exhibit the same impurity characteristics and utilize different processing; they will normally exhibit vastly different material properties than the actual part. This is especially true for cast irons where it is not uncommon to have a 20–30% difference in tensile strength between the separately cast test bars and the actual part. Therefore, it is critical to understand the material properties (strength and microstructure) at the critical high stress areas within the actual cast part. Once a new casting is fully developed and a correlation between the cast part and separately cast bars has been developed for microstructure and strength, then the results from the separately cast test bars can be utilized for continuing process quality audit purposes.

2.6.1 Cylinder block specimen locations

Specimens for material testing are typically cut from the bulkhead area (Figure 5) as this is one of the highest stress locations in the block. While high stresses do occur around the top deck and cylinder wall in parent bore blocks, it is normally difficult to find a section thickness in the bore or top deck sufficient for a properly sized specimen, especially in cast irons.

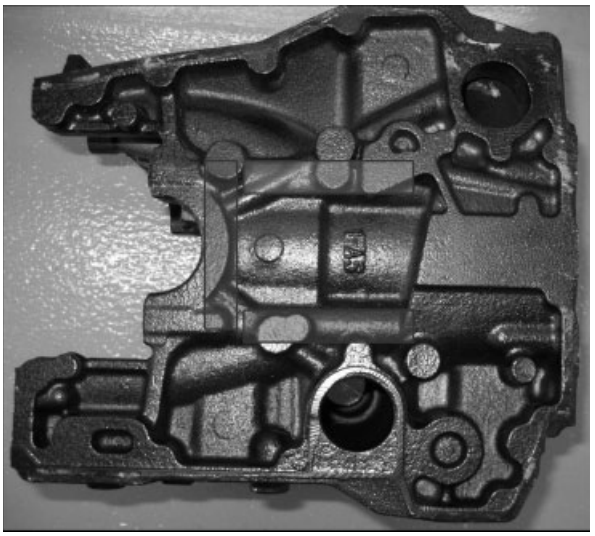


Figure 5. Cylinder block material test specimen location.

2.6.2 Cylinder head specimen locations

Specimens for material testing are typically cut from the fire deck in the valve bridge areas (Figure 6) as this the most critical location for material properties within the head. Samples can also be cut from the top deck, which play a critical role in cylinder head structure (especially for high PCP diesel designs), and head bolt bosses to understand material strength variation throughout the casting.

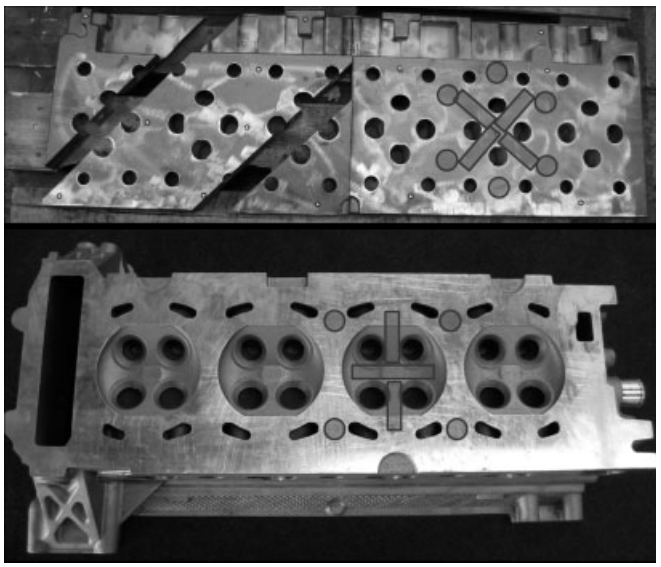


Figure 6. Cylinder head material test specimen locations (cast iron HD diesel-top, aluminum LD gasoline-bottom).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto135
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

During the development process, complete characterization of the material is necessary to understand the tensile strength, yield, and fatigue life, and any significant influences of temperature. With cast irons, the material properties can often be assumed constant over the operating temperature. This is not the case with the nonferrous alloys. The modulus of elasticity, tensile strength, yield strength, ductility, thermal conductivity, and fatigue strength are all a function of temperature and change significantly over the operating temperature range of the engine for most alloys.

At a minimum, the development material testing program should include tensile testing, compression testing, and fatigue testing with specimens taken from critical areas in production-intent cast parts. For nonferrous alloys, this testing should be repeated at ambient temperature and an elevated temperature equal to the maximum predicted operating temperature. The microstructure should be inspected from the same locations. Typically, specimens are cut using the following ASTM test procedures: ASTM E-8(04) for tensile testing and ASTM E-468-90 for fatigue testing. Rotating beam fixtures are commonly used for fatigue testing of cast irons, whereas axial testing is preferred for fatigue of nonferrous alloys. Figure 7 shows an example of tensile and fatigue test specimens for gray iron cylinder head, whereas Figure 8 shows a sample fatigue specimen from an aluminum cylinder head.

Figure 9 shows an example of the scatter in gray iron fatigue strength, which may be expected throughout a cylinder head casting and between castings of different batches for a controlled/developed process.



Figure 7. Example failed gray iron fatigue (left) and tensile (right) test specimens.

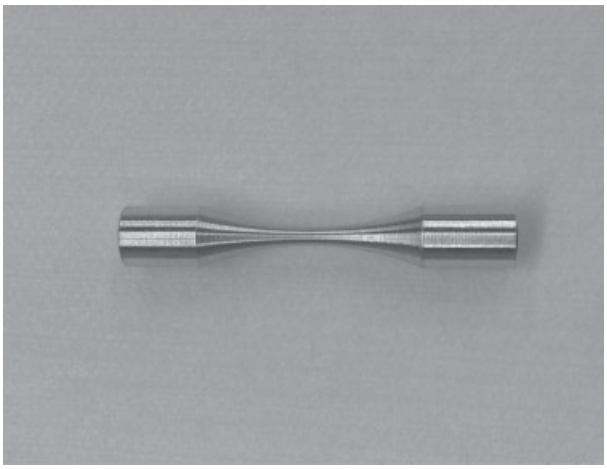


Figure 8. Example aluminum fatigue (left) and tensile (right) test specimens.

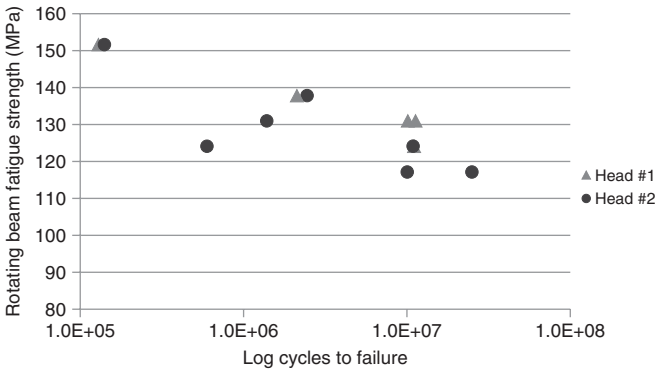


Figure 9. Example gray iron fatigue test data.

3 CASTING PROCESS DEVELOPMENT

There are many casting processes including sand casting using molds made of clay, water, or chemically bonded sand, permanent mold casting, and pressure die casting. Each of these processes has its variants, and the casting process needs to be selected at the same time the casting material is selected. Not all materials can be cast using any of these methods, and not all of these methods are suitable for all geometries and wall thicknesses. Manufacturing volume is a major factor in choosing a process. Small volumes are often produced by sand casting, but as volumes increase permanent mold or die casting becomes more attractive. The costs for molds required for permanent mold or die casting can be absorbed in large volume production. The sand casting process can be automated, using machines to produce and fill sand molds. Quality is another consideration when choosing a process as solidification influences the material properties.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto135
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

It is absolutely vital to get the production foundry, and if applicable, the prototype foundry involved in the design process at the very beginning to prevent rework or delays during the casting development process.

The molten alloy can be introduced into the mold by pouring or injecting it at elevated pressure. The pressure casting methods typically use permanent molds or dies that are nonexpendable, but do have a finite life because of wear, fatigue, or cracking. The molten metal is usually introduced into the mold using a gating system. The metal is poured into the top of the mold and travels down the side in a chute called a *sprue*. It then runs horizontally along the bottom of the mold, then up into the cavity through runners called *gates*. As the metal enters the cavity from the bottom, any gasses will rise out of the mold. This method also creates less turbulence of the molten metal. Gating design is absolutely critical in achieving castings with premium material properties and minimal defects. Additional areas called *risers* (Figure 10) are added to control shrinkage porosity during solidification by providing pockets of molten metal at strategic locations. Bodies of cool material, called *chills*, can be placed on the side of the mold to remove heat quickly as the molten metal first comes in contact to rapidly solidify specific areas of the casting to locally improve material properties. This is often done in the combustion chamber, for example, to provide improved strength in the material between the valves called the *bridges*. Casting orientation within the mold also plays a critical role in the final material properties. For example, cylinder heads are typically cast fire deck down (fire deck at, horizontal and parallel to gating) to achieve the best properties in the critical sections. Orientation can also have positive or negative affects on factors such as core float and gas venting. Therefore, these factors must be thoroughly evaluated and quantified during casting development.

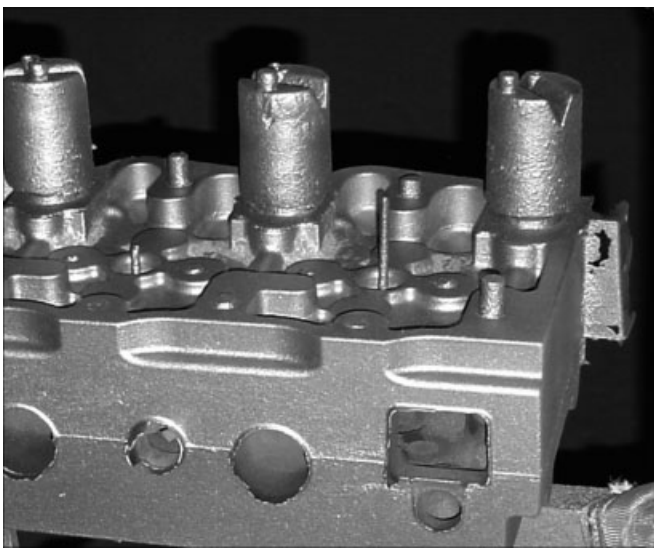


Figure 10. Example showing risers placed on the top deck of a cast iron diesel cylinder head casting to control shrinkage porosity.

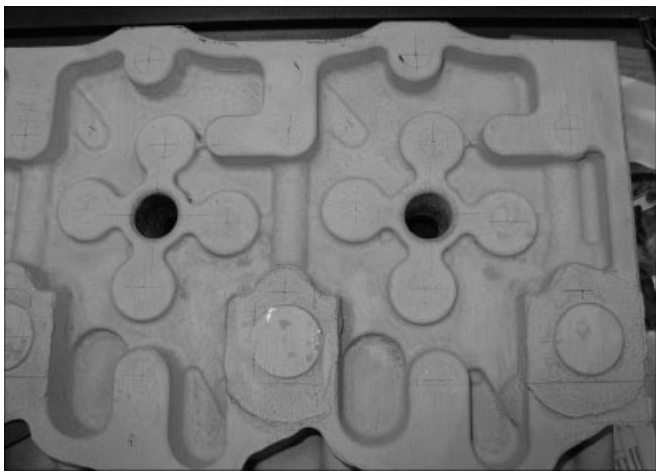


Figure 11. Conventional casting geometric qualification using markup technique.

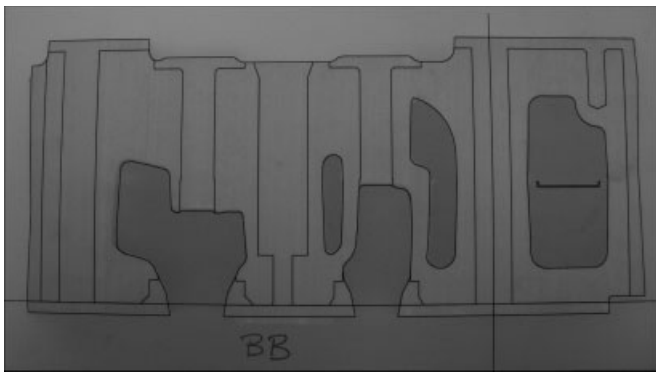


Figure 12. Casting geometric qualification of internal features using mylar sections from CAD geometry (injector bore cross section).

As the integrity of block and head castings is critical to the quality of an engine, and as the casting process has so many variables, casting process simulation is employed to ensure quality and help develop the process, locate risers, and identify areas to enhance cooling. Dedicated simulation packages have been developed over many years, including Magma, PAM/CAST, and EKK CAP. These sophisticated packages simulate the flow of the molten metal into the mold and the solidification process.

Typical steps in the development of the cylinder head or block casting process are to:

1. Select a candidate alloy.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto135
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

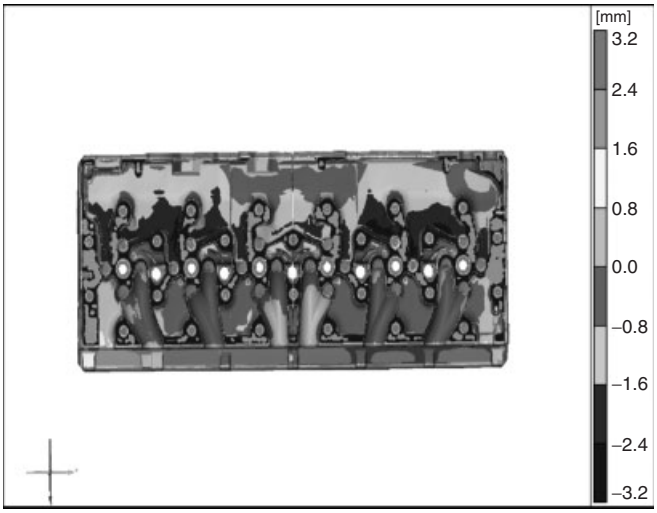


Figure 13. Laser scan of head casting section to map core shift (vertical direction shift).

2. Design the head or block.
3. Choose an appropriate casting process considering manufacturing volume, cost, cycle time, and quality.
4. Perform a casting analysis studying filling and solidification rate and predicting locations of high residual stress and potential shrinkage porosity.
5. Iterate through steps 1–4 as required, changing the gating system or using chills (aluminum for strength) or risers (cast iron for porosity) as required.
6. Make a test melt of material and cast test bars and cylinder heads or blocks using production-intent molds.
7. Perform tensile, compression, and fatigue tests and check hardness and microstructure of both the test bars and the samples cut from critical areas in the casting as identified by analysis.
8. From the results, develop procedures for periodic checks in production. These checks may include chemical analysis of the molten metal before casting, casting and testing of test bars, and mechanical testing of samples taken from castings. Wherever possible, checks should be performed on the inputs to the process, such as checks on raw materials, measurement equipment, and molds. Special care should be taken with components subject to wear, such as molds. If a process is fully understood and the inputs are controlled, the output will remain in specification.
9. Make any required changes to the process, material, or design.
10. Make a limited-volume casting run using production-intent tooling and processes for production qualification. Use the actual production line if possible. The quantity of parts in this run must be representative of actual production volumes. Perform all of the defined process checks.
11. Thoroughly inspect and test the qualification castings. If issues are found, improve the process controls or make process or design changes as needed and requalify the modified process. In addition to material testing, X-ray (cast iron) or CT scan (aluminum) must be performed to look for shrinkage porosity and voids along with geometric qualification. Geometric qualification is performed to ensure that the part meets the designed tolerance (typically ISO CT 7–8), does not exhibit excess shrinkage, and validate that all of the cores are in the correct place and did not shift. This qualification can be performed using several methods. The classic method involves painting the casting with layout paint and then scribing centers for all machined features on the exterior of the casting (Figure 11). These centers are then compared to their respective bosses for shift during the casting process. Other methods to validate internal wall placement include sectioning the casting and then comparing to mylar sections generated from the CAD geometry (Figure 12) or using laser scanning techniques to generate a 3-D map of any core shift (Figure 13).
12. Have all involved personnel and departments sign off on the process.
13. Commence production, performing the process checks and material quality audits at regular intervals.

REFERENCE

Dawson, S. and Indra, F. (2007) *Compacted Graphite Iron – A New Material for Highly Stressed Cylinder Blocks and Cylinder Heads*, Internationales Wiener Motorensymposium.

FURTHER READING

Goodrich, G.M. (ed.) (2003) *Iron Castings Engineering Handbook*, USA, American Foundry Society. ISBN: 978-0-87433-260-5

Davis, J.R. (ed.) (1996) *ASM Specialty Handbook: Cast Irons*, USA, ASM International. ISBN: 978-0-87170-564-8

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto135
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Block and Head Analysis

Marc C. Megel
Southwest Research Institute, San Antonio, TX, USA

1 Introduction 1
2 Effective Applicaton of Advanced Analysis Techniques 1
3 Computational Fluid Dynamics Analysis 5
4 Finite Element Analysis 7
5 Geometry Considerations 14
Acknowledgments 16
Related Articles 16
References 16

1 INTRODUCTION

Computer-aided engineering (CAE) analysis is a key component of the design/development process for engine components. While not a complete replacement for mechanical development testing, effective application of advanced analysis can significantly reduce design cycle times and minimize prototype redesign iterations during development testing. Two of the most complex and time-consuming components to design and analyze within an engine system are the cylinder block and head. With respect to these two components, this type of analysis can normally be broken into two main categories: computational fluid dynamics (CFD) and finite element analysis (FEA). Figure 1 shows a high level overview of the normal process for a new engine design and development. Figure 2 outlines the detailed process commonly utilized for a new cylinder head design and identifies how various analyses and tests feed into the design and development process. A similar approach is utilized for the cylinder block. From this flowchart, it can be seen how instrumental CAE analysis has become in modern cylinder head or block design and development.

A few brief comments regarding CAE software packages are probably warranted given the multitude of products in the market today. Packages such as ANSYS, ABAQUS, STAR CCM+, and NASTRAN are well established and commonly utilized in the engine design field for advanced analysis. These programs are widely used by numerous engine and vehicle original equipment manufacturer (OEM)'s (both light duty and heavy duty) and as such, face continual review and scrutiny from a multitude of users. Caution should be exercised when considering an entry-level package for the analysis of complex components such as blocks or heads as all software packages are not created equally. Users should always bear in mind that finite element or CFD analysis of blocks and heads is a mathematical approximation of various physical phenomena within a geometrically complex structure. Key considerations for selecting a software package should include accuracy (especially regarding entry-level systems), transparency, and control of operations and functions such as meshing and contact, and the ability run the program from controlled user macro's and scripts. Ease of use should never be a primary consideration when choosing a software package.

2 EFFECTIVE APPLICATON OF ADVANCED ANALYSIS TECHNIQUES

The most effective analysis that can be applied to engine design and development is that which produces the necessary outcome in the required timeframe at the lowest cost.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

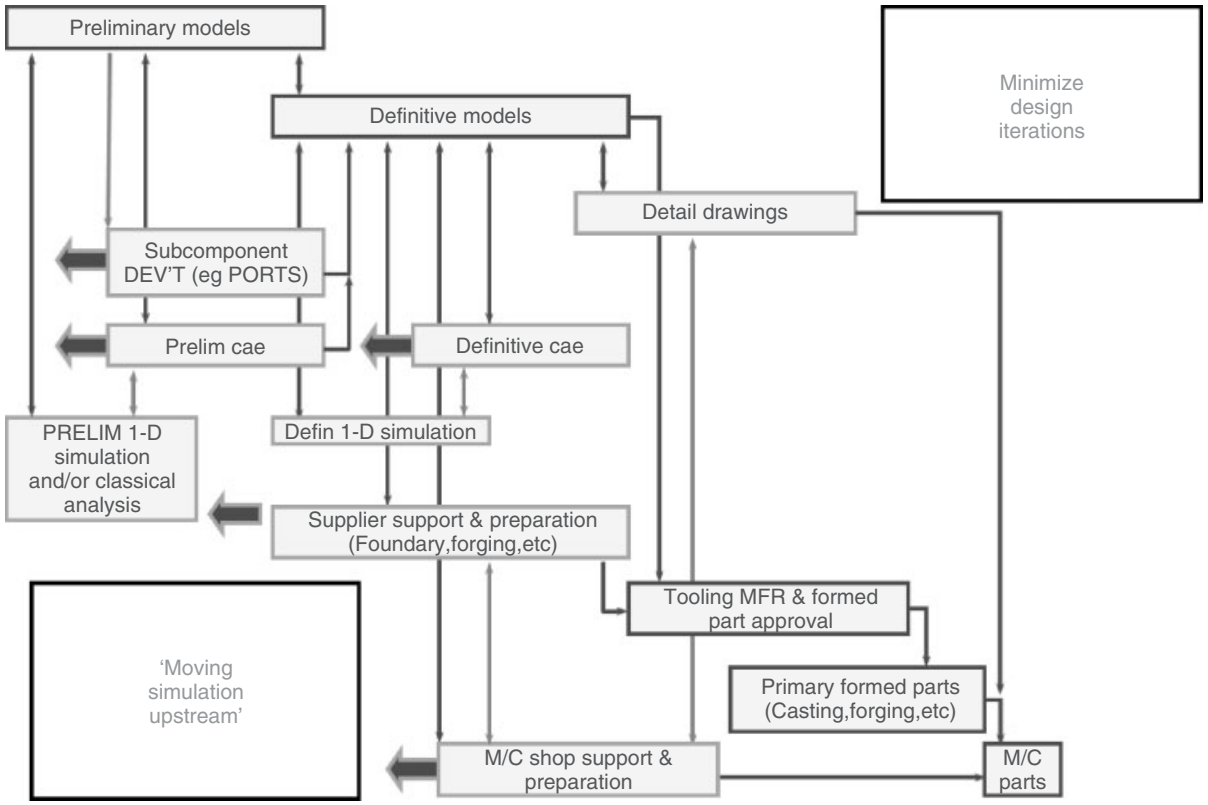


Figure 1. New engine design and analysis process.

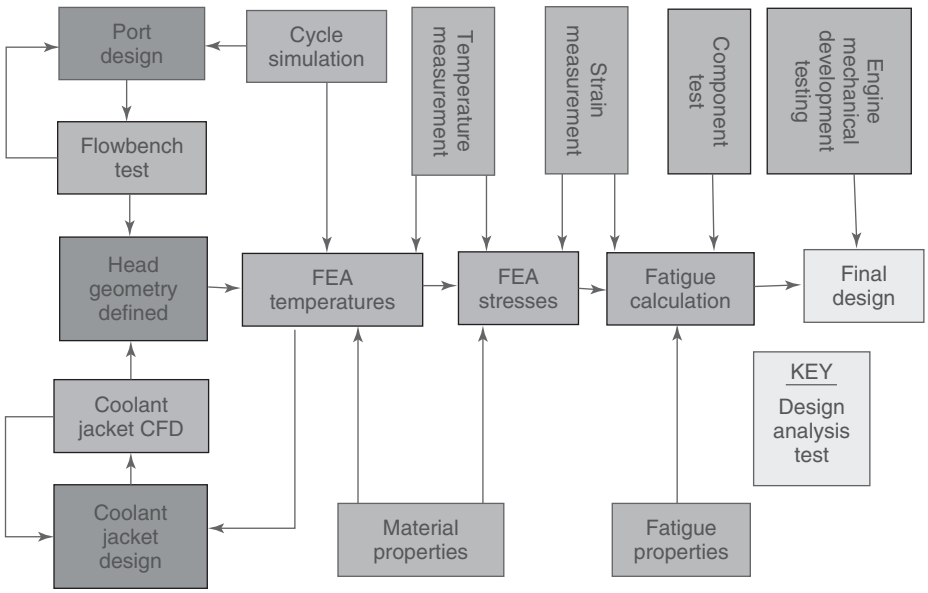


Figure 2. Detailed cylinder head design and development process.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Simply put, this means the analysis with the fewest features and minimum necessary level of complexity that gives an adequate result is the best. The time spent on simplifying the analysis processes to the minimum complexity required and developing robust procedures will pay off during the design process or when issues arise during production.

It has been proposed that analysis (CFD or FEA) can be broken into four basic categories: relative–comparative, relative–predictive, absolute–comparative, and absolute–predictive.

2.1 Relative analysis

This technique has proven to be a very effective approach for the analysis of engine components. One of the major obstacles to absolute analysis of engine components is lack of input data. This primarily stems from the fact that engines are relatively low cost, general consumer commodity items. This type of product does not have a cost structure to support extensive research and testing to exactly determine all of the necessary material behavior and boundary condition inputs required for accurate absolute analysis.

In a relative analysis approach, standardized process are developed and utilized consistently for all product development. They are intentionally seldom changed to maintain a connection for all products to a historical database. The objective is to create a process that correctly captures the general physics of the component operation. This process allows for the use of a nominal set of boundary conditions and material properties. Normal variation in boundary conditions and properties experienced by the component in service are accounted for in the process using a historical failure database to determine a component fatigue limit rather than a material fatigue limit. This process has proven especially valuable for blocks and heads, which are manufactured using either an iron or an aluminum casting process. Casting processes contain significantly more variability in end material properties than those used to manufacture billet material. As a result, design and foundry engineers work to control properties within an allowable range, provided that the range and processes do not shift for a given product database, a nominal value can be assumed and variability is addressed in the component limits.

Relative–comparative analysis refers to a process whereby a modification to a product (either for a new product requirement or for addressing an issue) is being compared back to the current design by the way of a standardized analysis process to determine reductions in stress at critical locations.

Relative–predictive analysis refers to a process whereby a standardized process is used to analyze a clean sheet design and the results are compared to a historical limits database to determine the acceptability of a design.

2.2 Absolute analysis

The most recognizable application of absolute analysis is probably the aerospace industry, although others exist for similar reasons. Factors driving product development in the aerospace industry are in direct contrast to those discussed previously for automotive engine components. Reliability and durability of an aircraft turbofan engine are paramount above cost and development schedule given the ramifications of failures. Exact fatigue behavior of materials must be thoroughly understood to ensure robust designs and develop maintenance/repair schedules.

Occasionally, situations will arise in engine design and development where an absolute type analysis approach may be required. Bracket analysis is a common instance where common designs seldom exist. Even at that, standardized processes with common approaches to meshing and load application will help reduce variability. The other most common instance is during failure resolutions activities where a unique problem has arisen that maybe a previous analysis process did not identify. In this instance, mechanical development mapping such as strain gauge measurements may be required to calibrate a model to identify the proper failure mode. Even in this scenario, absolute–comparative (establish a baseline for the current issue and use the calibrated model to compare proposed solutions to the current design) analysis is probably the best approach.

The following discussions throughout this chapter assume a relative (comparative or predictive) approach to block and head analysis unless otherwise specified.

2.3 Standardized analysis processes

Standardized analysis processes are critical to implement a relative analysis approach and achieve consistent results between multiple analysts as well as expediting the analysis process. Development of simplified analysis approaches is a key to effectively move the analysis upstream in the design process to have maximum impact on the end product design and minimize redesign iterations. While inclusion of every detail and nonlinearity in a given system "might" improve the accuracy of the model, it has little benefit if the time to perform the calculation causes delays in the overall design cycle or is not completed in sufficient time to influence predetermined engine build milestones. In addition, continual changes such as these invalidate a relative analysis approach. A simplified process that can rapidly

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

produce reasonable results and evaluate multiple design options provides significantly more value to the overall engine development program than a complex analysis of a single design. Another key to develop an effective analysis processes is to have properly trained analysts who understand pitfalls of the work and software and have significant experience in component development, so the process focuses on evaluating the component. Application of analysis to cylinder block and head design/development should be process and component centric, not software centric. The software should be viewed simply as a tool to acquire the necessary information to make design decisions and progress component development.

Development of a standard analysis process that is used consistently will also allow creation of a historical product database over time, which allows an analyst to compare results from a new design to previous benchmarks. It is this database which is utilized to develop analysis limits rooted in real historical engine failures and component fatigue as opposed to material specimen fatigue. The value of this approach is twofold. First and foremost, this approach allows the analyst to account for variations in manufacturing (casting and machining), material properties, duty cycle, on a given product and process. Uses of analysis processes with limits such as this have been shown to correlate extremely well to crack locations throughout the water jacket and provide valuable insight into head gasket joint integrity.

To the extent possible, a good standardized process should control all key aspects of FEA or CFD modeling.

Meshing
Boundary condition application
Load steps
Solution and solution controls
Post processing

Variation in any of these parameters will result in variation between analysts and analysis runs and invalidate the use of a component-based fatigue evaluation approach. Specifically when considering meshing, the maximum amount of manual mesh control possible should be utilized. Automated meshers with default settings should never be used as they exhibit unacceptable random behavior. While the development of automated meshers within CAE codes greatly improved the ability to mesh complex geometries, they also introduced one of the most significant sources of variation within CAE modeling. The results calculated in a CAE model are entirely dependent on the mesh. Therefore, variation in mesh equals variation in results and prohibits the use of a relative analysis process. A good approach is to manually define the line element size for all lines and the target area element size. From there, the automated mesher can grow the mesh to fill the areas. Once areas have been meshed, then a volume mesh can be generated, again by setting a target element size. This method has proven to generate consistent meshes acceptable for a relative analysis process. This approach does require the analysts to maintain the connection between the solid geometry and the mesh.

It should be noted that application of this type of analysis process to aluminum components can be more difficult given the extreme nonlinear behavior and material property sensitivity to temperature compared to iron (see Material and Process Selection—Cylinder Blocks and Heads). However, many of the aspects can still be applied.

2.4 Component versus system analysis

With development of good analysis procedures, component analysis of blocks and heads has proven to be extremely effective (both cost and time) for initial analysis and rework of designs, especially those cast from iron materials. This use of component analysis is critical to move the analysis upstream in the design process. One of the major advantages of component-only analyses is the fact that they do not require highly nonlinear elements, such as contact, springs, bolt pretention, or gasket elements. These types of elements, especially contact, significantly increase the computational requirements of the model from both a solution time and hardware perspective. As a result, an analyst can progress through significantly more design iterations if required in a given amount of time. In addition, as these models only require input loads and not details of components such as head gaskets or bolts, they can be employed very early in the design process. Only a basic CAD model with completed port, valve, injector, and water jacket geometry is required to start the component CFD or FEA. In this approach, the designer will first focus on modeling the basic geometric features including internal water jacket rounds to provide a CAD model to start CFD and FEA, then work to incorporate details such as bolting and flanges, which has minimal to no impact on the key thermal and structural areas of the block and head. Having a well-developed component analysis process capable of rapid design iterations is also very valuable for rapidly addressing test or field failure issues, which may arise during development or over the life cycle of the product.

Through the analysis process development, methods to model interactions among the block, head, gasket, liner, and bolts using forces, displacements, and reactions can be established. This process development may require

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

mechanical development testing such as pressure-indicating film, strain gauging, ultrasonic length measurements, or hydraulic rig testing.

Should the situation arise where a detailed analysis of the head gasket joint or cylinder bore distortion is required, then a system model must be utilized. Owing to the level of complexity, resulting assembly and solution time, and lack of sufficient design details for supplier components (head gasket, liner, and bolts), this type of model has little value early in the process to affect the design. If performed correctly, this type of modeling can be very valuable in setting up simplified component analysis procedures or when more detail is required to investigate a particular problem.

2.5 Developing limits and guidelines

2.5.1 Empirical component/process-based versus material fatigue limits

The development of empirical, component-based fatigue limits is critical to the relative analysis process. By utilizing a historical component performance database, all of the normal variations for a given material, process, duty cycle, and so on are accounted for in a single limit value. This then allows for the ability to compare analysis fatigue results against a single limit for a given type of engine and application thus significantly simplifying the analysis by reducing the sources of variation that must be accounted for/analyzed. Once this limit is in place, the only variation that must be accounted for are those which produce significantly different boundary conditions to the baseline such as duty cycle or application.

This process also allows for the assessment of different materials by the scaling of the single limit, provided the two have similar production variations. This is most commonly used for evaluating various grades of cast iron in a design assuming similar casting processes.

2.5.2 Duty cycle considerations

Understanding of duty cycle and load variations (peak cylinder pressure for example) for a given product is absolutely critical to developing a robust, effective relative analysis process. Different empirical component-based fatigue limits will need to be developed for distinctly different duty cycles and applications. Figures 3 and 4 are included below to help illustrate the differences in duty cycle that a given engine can be exposed to. Although not exact duty-cycle plots from a specific application, emissions certification test cycles provide a good general approximation of the average duty cycle to be expected in a given application for the purposes of this discussion. When developing component fatigue limits for relative analysis approaches, engineers should always rely on actual average duty cycle data obtained for their specific application and not general approximations. Figure 3 shows the differences between the light-duty chassis certification and heavy-duty engine certification test cycles for a medium-duty engine, which could fall into either or both of these applications. Figure 4 shows a similar comparison, but this time for a heavy-duty diesel engine used in either Europe or North America. Note the difference in average engine speed and load that may be expected between the two cycles. This leads to the development of significantly different component fatigue limits and even material choices (see Material and Process Selection—Cylinder Blocks and Heads) depending simply on the geographic use of the engine.

3 COMPUTATIONAL FLUID DYNAMICS ANALYSIS

CFD analysis should be utilized to optimize flow distribution throughout the block and head and maximize coolant velocities in critical areas such as valve bridges and around injectors/spark plugs. The CFD model is also used to calculate heat transfer coefficients (HTC's), which will be inputted into the FEA models as coolant side boundary conditions. CFD should be performed as a system model to include the following water jacket cavities: block, head, thermostat housing, oil cooler housing, and water header if equipped.

Most modern engines use orifices in the head gasket for control between the block and head water jackets to optimize flow distribution front-to-rear and side-to-side in the block. These orifices are typically modeled as the connection between the block and head water jackets with a length equal to the thickness of the head gasket. The head gasket orifices along with cast-in geometric features are used to control flow distribution and velocities in the head. Typically, both the inlet and outlet meshes are extended for some distance from their actual location to reduce numerical instabilities at the boundaries which otherwise has no impact on the calculated results. Figure 5 shows an example of the mesh for a heavy-duty, in-line six-cylinder water jacket configuration.

Typical boundary conditions for the water jacket system model would be coolant inlet flow rate and system outlet pressure. Critical coolant physical properties such as density, viscosity, specific heat, and thermal conductivity must be input as well. If the CFD analysis is being performed stand-alone from the FEA (non-conjugate heat transfer analysis), then wall

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

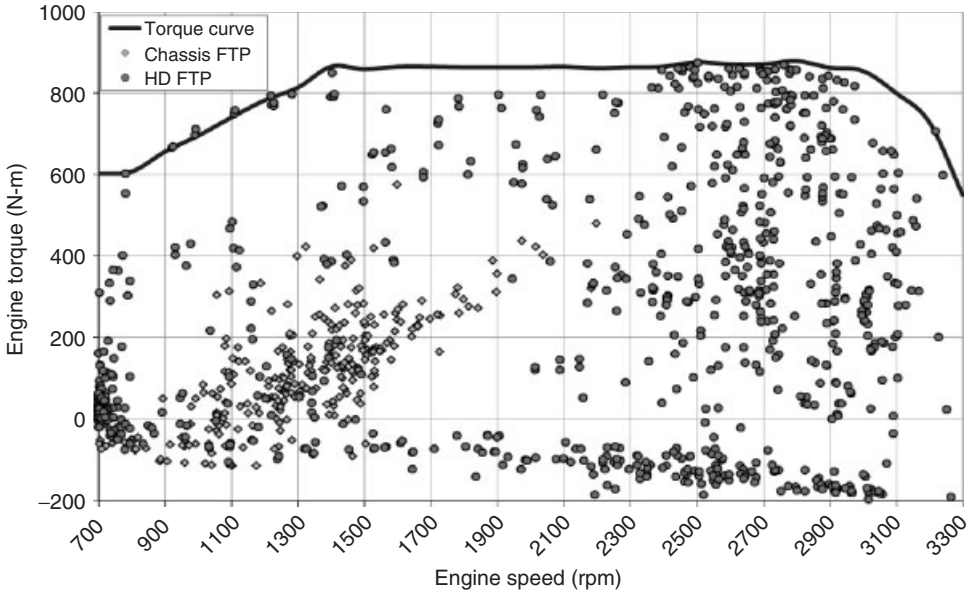


Figure 3. Comparison of EPA certification test cycles for medium-duty engine installed in a light-duty verses heavy-duty application.

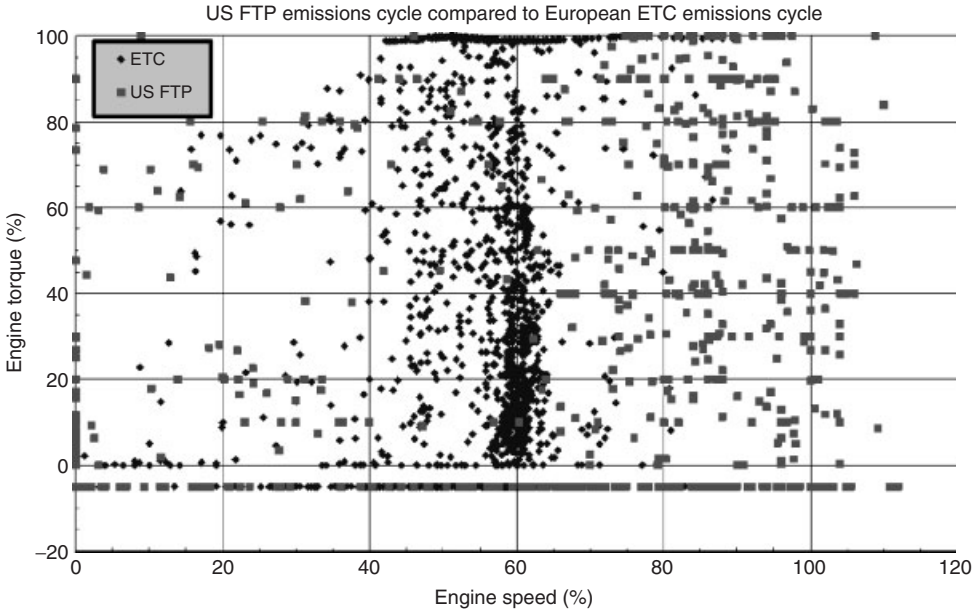


Figure 4. Comparison of US versus European certification test cycle for a heavy-duty diesel engine.

temperatures must be inputted for the calculation of HTC's. From these boundary conditions, the model solution will develop the pressure drop profile across the engine and provide input to the required water pump flow versus pressure characteristics. In addition, coolant velocities and resultant HTCs will be calculated for input into the FEA thermal models. Figure 6 shows a typical plot of coolant velocities and associated HTCs for a cylinder head lower water jacket on a heavy-duty engine.

On the basis of the component thermal profiles calculated from FEA, iterations on the head gasket orificing should be performed as necessary to achieve both the desired component temperature distribution and the uniform temperature distribution, front-to-rear and bank-to-bank (vee engine configurations).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

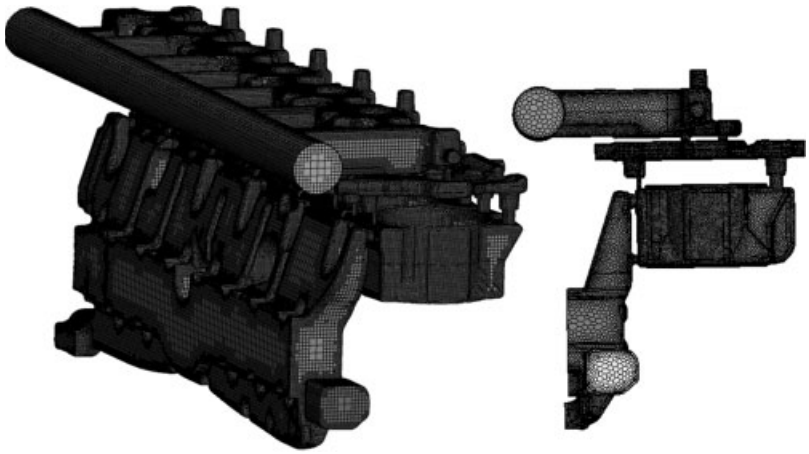


Figure 5. In-line six-cylinder system water jacket CFD mesh.

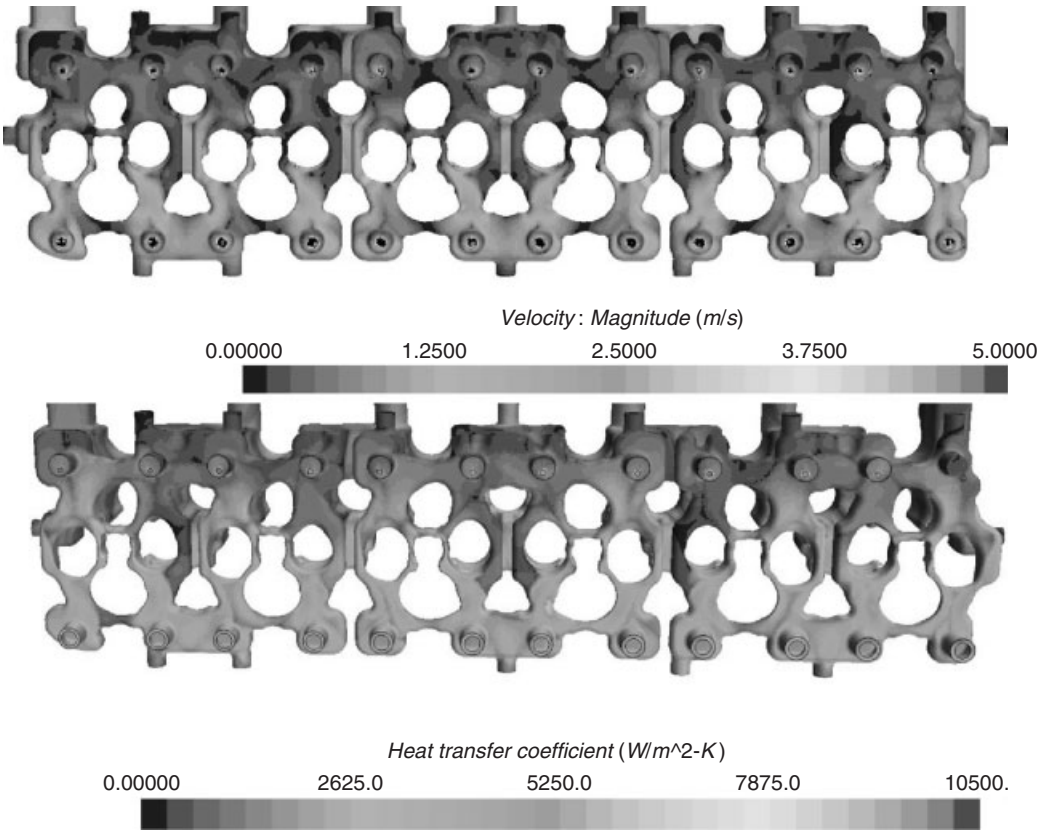


Figure 6. Example cylinder head lower water jacket coolant velocity and HTC distribution.

4 FINITE ELEMENT ANALYSIS

Discussion in this section focuses primarily on cylinder block or head component simulation. When establishing a process for a given configuration of cylinder head or block, the benefit of symmetry should be utilized to the maximum extent possible. For instance, an in-line six-cylinder head configuration utilizing a repeating pattern of structure for each cylinder (as opposed to a pattern mirrored about the cylinder head centerline) can be cut through

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the center head bolts on the cylinder 3–4 split plane. This instantly reduces the model size and computational requirements by half over analyzing the entire cylinder head, which produces no added value. The new three-cylinder section utilizing symmetry allows analysis of the structure utilized on the four interior cylinders and the end cylinder configuration. Stresses and displacements are evaluated on the two cylinders farthest from the symmetry plane as they are not affected by the local application of boundary conditions at the symmetry plane (St. Venants Theory) (Shigley and Mischke, 1989) and the symmetry boundary conditions replicate what those cylinders actually see in the full configuration. A similar approach can be used to analyze the cylinder block bulkhead and main cap.

In addition to the use of symmetry, further simplification of the model and reduction of solution times can be achieved through targeted meshing of critical locations. Take for instance, the example stated earlier with symmetry already applied. Utilizing meshing controls, the analysts can create a consistent, refined mesh that encompasses the central and/or end cylinder to allow proper calculation of stresses and mesh the remainder of the geometry with a coarser grid to reduce computational time and requirements but still allow sufficient resolution for calculations of temperatures and displacements. An example as seen in Figure 7 would be to fine mesh the central cylinder with 3–5 mm elements while meshing the remainder of the head at 7–8 mm. Figure 5 shows how a CFD mesh can be coarsened as you move away from the wall to the center of the flow stream.

4.1 Cylinder head component analysis

The primary objective and focus of the cylinder head component FEA should be to optimize water jacket design to satisfy cooling and structural requirements. It is not uncommon for a new design to require multiple (5–10) analysis/design iterations to achieve a structure that satisfies the component fatigue limits established in a given procedure. With a well-developed, simplified, relative analysis approach, iterations can normally be completed in 1–5 days.

FEA of the cylinder head component can be broken into two primary tasks: thermal and structural analyses. Thermal analysis is utilized for prediction of component metal temperatures and input to the structural analysis for calculation of thermal stress.

4.1.1 Thermal analysis

Thermal FEA utilizes the cooling side HTC values obtained from CFD coupled with the combustion side boundary conditions developed from empirical correlations, 1D cycle simulation such as GTPower, 3D in-cylinder combustion, CFD simulation such as Kiva and Converge, or a combination of all the three. Well-calibrated empirical correlations have proven very effective for developing combustion boundary conditions for a given type of combustion. This method is a critical piece of moving the FEA simulation upstream in the design process as in-cylinder CFD models tend to have lengthy solution times as you have to run the entire two- or four-stroke cycle to develop cycle-averaged (not instantaneous) HTC values as that is what the FEA is simulating. Even with the advanced 3D combustion CFD capability, experience has shown that these results may still need adjustment through the empirical relationships to achieve temperature predictions from FEA, which match measured values. The difficulty in accurately predicting in-cylinder, cycle-averaged HTC values likely stems from the fact that in-cylinder heat transfer is a complex interaction

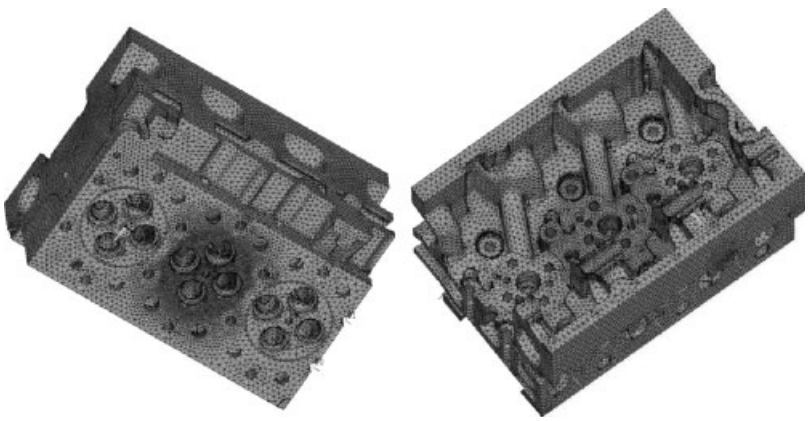


Figure 7. Example of in-line six-cylinder head utilizing symmetry and focused mesh refinement.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of convection and radiation/adsorption from combustion chamber surfaces and combustion particles (most notably soot in diesel and gasoline direct injection). In addition to HTC values for combustion surfaces, the thermal FEA model must also include all other heat input or extraction sources such as ports, oil splash, and ambient conditions. Within the cylinder head component model, the option exists to include or exclude valves and seats. As with the other boundary conditions, the option exists to develop empirical boundary conditions, which address the heat transfer across the valve/seat/head interfaces thus allowing for further simplification of the model.

Primary results of the thermal analysis will be predictions of peak fire-deck temperatures (a critical design constraint) and temperature distributions along the length of the cylinder head. As mentioned in Section 3, an iterative loop among design, CFD, and FEA will likely be required to try and maximize the uniformity of temperature distribution along the length of the head. Designs that produce a significant gradient along the length of the head should be avoided. Nevertheless, some variation along the length of the block or head is often present in many designs. Any comparison of peak fire-deck temperature or stresses against limits must consider this variation and be based on the hottest cylinder. Figure 8 shows an example of predicted fire-deck temperatures from a thermal FEA simulation.

4.1.2 Structural analysis

Cylinder head component structural analysis incorporates nodal temperature results with the necessary loads such as cylinder pressure, head bolts, injector clamping, press-fits, and valve train loads (if required) along with reactions from the head gasket and symmetry conditions to calculate deflections and stresses. The four critical load steps for a cylinder head structural analysis are as follows.

Assembly loading (A)
Assembly + cylinder pressure loading (AP)
Assembly + thermal loading (AT)
Assembly + cylinder pressure + thermal loading (ATP)

Using principle (cast iron materials) or equivalent stress (ductile materials) calculated from these four primary load steps, all critical fatigue states can be analyzed. Two of the most important and most commonly analyzed fatigue states are ATP-AT for high cycle fatigue and AT-A for low cycle fatigue.

Both stress-life and strain-life fatigue analyses are utilized for the evaluation of cylinder block and head stresses, especially with iron components. Stress life is highly effective at evaluating high cycle fatigue because of cylinder pressure loading. This method is commonly utilized for the fatigue analysis of iron components but can be used for aluminum components (provided the strain state is still relatively elastic and likely employing a Neuber type correction (Bannantine, Comer, and Handrock, 1990). In addition, stress life fatigue analysis of low cycle fire-deck stresses for iron components can provide a useful initial assessment assuming that the thermal stress state is lower than the compressive yield point of the material. Stress life analysis of FEA results can be performed very efficiently using user subroutines or macros. Owing to the simple calculations involved, macros can be written to process the entire set of results in a minimal amount of time. One very effective method is to calculate an equivalent fully reversed or EFR stress for each critical location based on modified Goodman theory with basic material data for true fracture or tensile, yield, and fatigue strength. This calculation converts



Figure 8. Example of cylinder head peak fire-deck temperature predictions from thermal FEA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the various mean and alternating stress states found at different locations throughout the cylinder head into a zero-mean equivalent stress (Equation 1). Using this method, the fatigue stress at each critical location can then be compared against a single set of limits for the component.

EFR stress under tensile loading can be calculated using Equation 1.

$$\sigma_{efr} = \frac{(\sigma_{tfs} \times \sigma_{alt})}{(\sigma_{tfs} - \sigma_{m})} \quad (1)$$

where: σ_{efr} = EFR stress
σ_{tfs} = true fracture strength
σ_{alt} = calculated alternating stress
σ_{m} = calculated mean stress

Note: for brittle materials such as cast iron, TFS can be approximated as ultimate tensile strength (UTS)

Strain life calculations are much more involved and require significantly more material input data to perform. Strain life analysis of a full set FEA results for a cylinder head typically requires the use of additional software (such as fe-safe) to perform calculations in a timely manner. Strain life fatigue analysis should be reserved for locations or conditions in which stress life analysis is not valid. Situations that warrant strain life fatigue analysis are the use of highly nonlinear material such as aluminum or conditions occasionally found on fire-decks of iron cylinder heads where the thermal stress is high enough to induce compressive yielding either from high thermal loads, an insufficient material cross section through the valve bridge or a combination of the two. The outcome of the strain life analysis is a prediction of cycles to failure.

For either the stress or the strain life fatigue analysis approaches, the resultant parameter (EFR stress or cycles to failure) is compared with the component fatigue limit determined during the development of the analysis process and not a specific set of endurance cycle data.

For naturally aspirated, spark-ignited (SI) applications, cylinder head loading is primarily thermal driven (both low and high cycle thermal fatigues) as component temperatures can be very high and cylinder pressures are very low. Thermal loading, creep and the resultant low cycle fatigue are also very important considerations when dealing with aluminum cylinder heads. With high cylinder pressure applications such as diesel or highly boosted SI engines, cylinder pressure loading, and high cycle fatigue become the predominant fatigue mechanism and play a much more significant role in component durability.

Critical areas for low cycle thermal fatigue stress are primarily located on the fire-deck of the cylinder head between valves (also known as the valve bridge area) and adjacent to the injector or spark plug. Depending on the cylinder head design, certain locations between the port walls and fire-deck on the water jacket side of the head will also exhibit sensitivity to thermal low cycle fatigue loading. Figure 9 shows an example of an EFR stress plot on the cylinder head fire-deck for AT-A low cycle fatigue case.

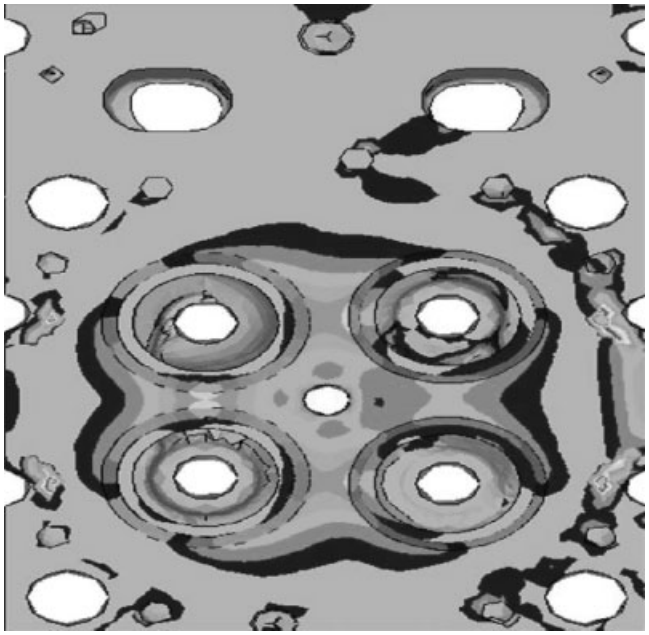


Figure 9. AT-A low cycle EFR stress plot example (cylinder head fire deck).

Critical areas for high cycle fatigue stress are typically found throughout fillets in the water jacket and the top deck of the cylinder head. In addition, rare situations exist in certain cylinder head designs where the fire-deck actually experiences issues with high cycle fatigue. This condition occurs when the thermal mean stress component exceeds the yield strength of the material (high thermal loading or insufficient material to react load). Figure 10 shows an example of an EFR stress plot in a cylinder head lower water jacket for ATP-AT high cycle fatigue case.

Before finalizing any fatigue results, it is good practice to verify that the mesh quality is sufficient at all high stress locations identified. One approach is to check the error bounds of stress calculation on the highest stressed element. It should be noted that this assessment must be performed on the calculated principle or equivalent stress for a given load case and not EFR stress as this is a post-processed parameter. This is a feature of most software programs and provides a quick method for ensuring mesh quality and consistency for a standardized procedure. If the error bounds are always kept, for example, below 10%, then any variation between analyses can be assumed to be accounted for in the process and does not invalidate the relative analysis approach.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

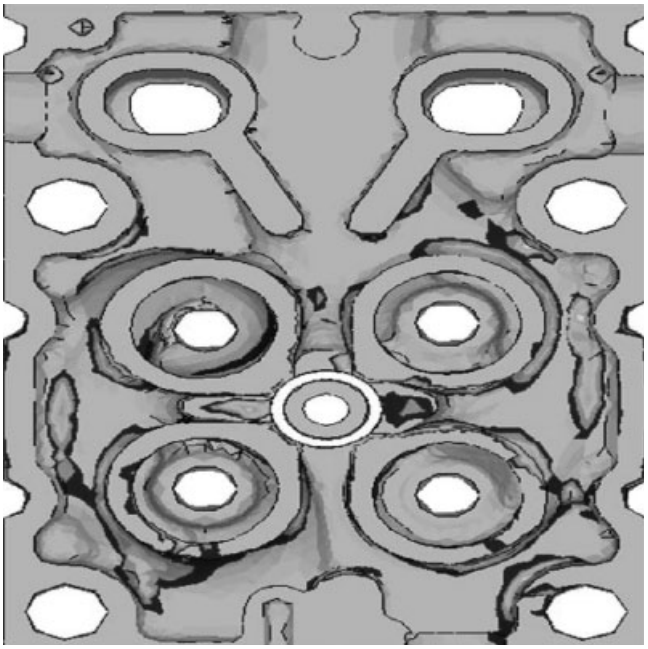


Figure 10. ATP-AT high cycle EFR stress plot example (cylinder head lower water jacket).

In addition to evaluation of fatigue stress, correlations can be developed, which utilize parameters such as displacement of the cylinder head fire-deck (or block if a parent bore design) above the combustion seal under assembly loading to evaluate the head gasket bolted joint. This displacement is a direct indication of the stiffness (or lack thereof depending on the design) of the clamp members in the bolted joint and thus a direct indication of potential leak spots for the combustion seal. The more uniform the block and head stiffness around the circumference of the combustion seal is, the more uniform will be the head gasket combustion seal loads and unloads minimizing locations and conditions prone to leakage. Therefore, limits for variability in displacement at this location can be developed and applied early in the design to ensure head gasket durability, even before the head gasket has been designed.

4.2 Cylinder block component analysis

The most critical cylinder block locations requiring FEA can be broken down into the following component analyses.

4.2.1 Liner/cylinder bore thermal analysis

This analysis is used to predict the peak cylinder surface temperatures and assess temperature distributions to identify hot spots, which could create issues with bore distortion or wear. As with the cylinder head thermal analysis, HTC values from CFD are transferred onto the water jacket surface of the liner, and boundary conditions for combustion are applied to the cylinder surface. The most critical temperature on the cylinder liner surface occurs at the top ring reversal location. This temperature must be kept below design limits to ensure proper wear behavior of the ring/liner interface and prevent oil coking on the liner and top ring.

4.2.2 Bulkhead analysis

This analysis focuses on the bulkhead area of the cylinder block (Figure 11). The bulkhead area of the block forms the load reaction path for cylinder pressure from the head bolts to the main cap. The bulkhead contains numerous locations required for critical features such as main bores and oil drillings, which result in high stresses under application of the cylinder pressure load. Stresses in these locations must be kept below applicable limits to ensure a robust design. Failure at this location typically results in a complete separation of the main bores from the cylinder section of the block. Like the cylinder head, analysis of the bulkhead can typically utilize symmetry (especially for in-line engine configurations) to simplify and speed up the analysis process. As the areas of concern are typically far away from the load reaction points, this model can utilize

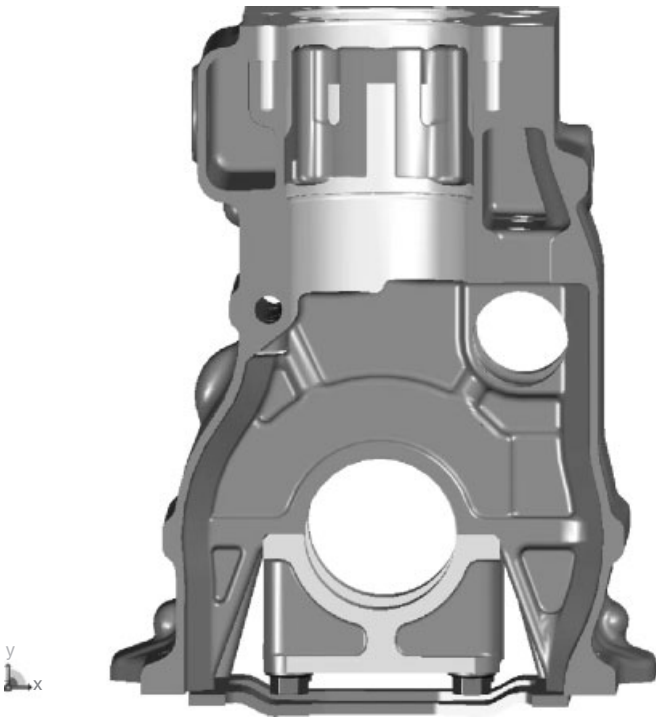


Figure 11. Cylinder block bulkhead section (in-line engine configuration).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a rigidly attached main cap to react the cylinder pressure load. Analysis of stresses around the main cap detail must be evaluated with a main cap analysis model. Because potential high stress locations within the bulkhead can be distributed throughout its entirety, a mesh suitable for stress analysis must be utilized for the entire component. Given that thermal loading does not affect stresses in the bulkhead and main cap areas of the block, the same high cycle fatigue analysis utilizing the EFR stress approach outlined in the cylinder discussion can be employed for iron blocks.

4.2.3 Main cap analysis

This analysis is used to evaluate the main cap design and ensure stresses are within an acceptable range. With proper meshing, cylinder block stresses in the region of the main cap can also be evaluated. Proper modeling of contact is required for the main cap to ensure correct modeling of load transfer in the bolted joint. All processing operations assorted with the main cap and main bore must also be accounted for.

4.2.4 Cylinder bore-top deck stress (parent bore block)

Analysis to evaluate stresses between the top deck and the cylinder wall should be considered for parent bore block designs depending on operating conditions. Under high thermal loading or increased combustion seal load conditions, the stress in this location can increase significantly and lead to fatigue cracks. This is not an issue for wet-liner engines as the top of the liner is not connected to the top deck of the block. This analysis will require application of the loads and boundary conditions from the head gasket. These may be developed empirically, from testing, or from system modeling of the complete head gasket joint.

4.3 System analysis

Certain situations exist where a system analysis may be required. The two primary conditions aside from boundary condition development discussed earlier are assessment of head gasket loading or bore distortion. As discussed earlier, this type of modeling is very time consuming and should not be utilized to drive the early stages of design. This type of modeling has proven very useful for analyzing issues that may occur in the field or during development testing regarding head gaskets loading or bore distortion. One of the most beneficial outputs for this type of analysis is to provide a visual understanding of complex physical loading and reactions happening within the engine. Indecently, this also turns out to be a very beneficial use of in-cylinder CFD analysis. Figure 12 shows the assembled geometry and mesh for a typical head gasket system-level analysis.

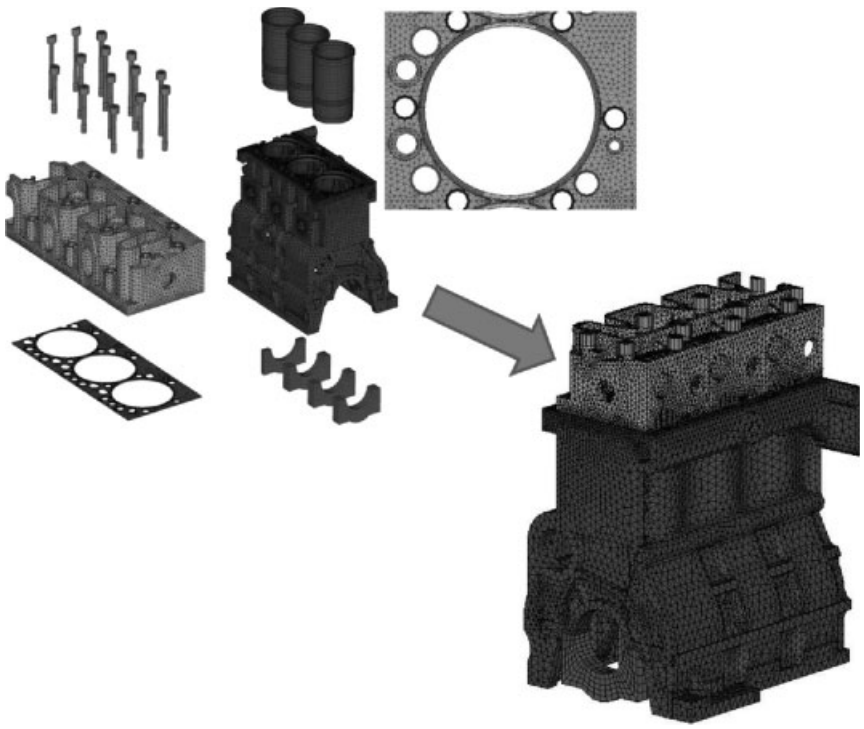


Figure 12. Example of block/head/gasket system FEA model assembly.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.3.1 Mechanical development testing feedback

Mechanical development testing provides valuable feedback to the design and analysis process in the form of calibration data and design validation. Use of each type of testing on existing mature products is essential for developing effective analysis procedures. Mechanical development testing can be broken down into three main categories: mapping tests, rig tests, and engine overload tests.

4.3.1.1 Mapping tests. Two of the most common mapping tests performed during cylinder head or block development are thermal and strain mappings. During thermal mapping, the cylinder head, valves, valve seats, and cylinder liner are instrumented using thermocouples, temperature check materials, or a combination of both. Installation of thermocouples in the cylinder head fire-deck and cylinder liner allows real time temperature measurement across the full operating range of the engine. The thermocouples do not however measure the peak component temperature at either the fire-deck or the cylinder surface. The measured thermocouple values must be used to calibrate the thermal FEA models at the specific measurement points and subsequently revise peak component temperature predictions.

Temperature check materials are materials that exhibit a specific characteristic of hardness verses exposure to a specific temperature. They come in the form of small plugs (1.6–3 mm diameter) which can be installed in components and actual components such as valves and seats machined from the temperature check materials. Most of these materials require stable, constant operation at a given engine operating condition for at least 2 h to develop the proper hardness for the exposed temperature. The advantage of temperature check materials is that they effectively measure the surface temperature of the component. The down side is that they can only measure temperatures for a single engine operating point. New plugs or components must be installed to evaluate different engine operating points.

4.3.1.2 Strain gauge. Strain gauges can be installed in various locations on the block and cylinder head to measure strain during either rig or engine operation. The measured strain values can then be used to help develop analysis procedures. Strain gauges are still the most effective method for evaluating residual stress caused by the casting process. Strain gauges are installed on a complete component in the area of concern and zeroed. The strain gauged location is then sectioned out of the full component and the resultant strain reading indicates the residual stress at that location.

4.3.1.3 Pressure-indicating (Fuji) film. Pressure-indicating film is very useful for evaluating the contact pressure between members in a bolted joint. Pressure-indicating film testing can provide very valuable information when developing boundary conditions for a component analysis procedure. The film is available in different pressure ranges depending on the joint to be evaluated (head gasket cooling passage seals verses head gasket combustion seal for instance). While the film itself only provides a measure of pressure in varying shades of red from imprint, some suppliers of pressure-indicating film have developed digital analysis techniques to provided detailed color contour plots (both two and three dimensional) of the measured pressure distributions. Figure 13 shows an example of using pressure-indicating film to evaluate a multilayered steel head gasket joint. From left to right: raw film after test, 2D digital analysis, and 3D digital analysis.

4.3.1.4 Hydraulic fatigue test. Hydraulic fatigue testing is typically the first step in validation of a design. Before calibrations are available, a system can be assembled to cycle the block and head assembly under simulated cylinder pressure using hydraulic fluid. This test provides very early feedback to the design process and validation of the analysis results for cylinder head, block, and main cap stresses as well as head gasket integrity and main cap fretting. The test is typically performed at some predetermined overload pressure from the target maximum

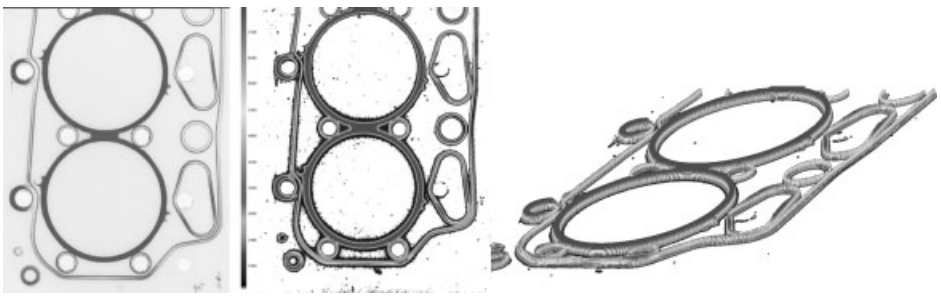


Figure 13. Example of pressure-indicating film test on multi-layer steel head gasket joint.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

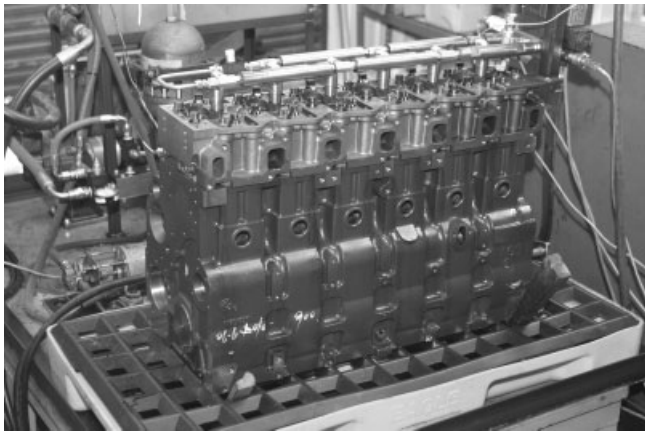


Figure 14. Block/head assembly hydraulic fatigue test.

engine operating peak cylinder pressure. Figure 14 shows an image of a hydraulic fatigue test for a block and head assembly.

4.3.1.5 Overload engine durability testing. Overload durability testing represents the final hurdle to production of a new engine design. As such, this step also represents validation of the analysis results. Manufacturers typically have a specified set of tests at various conditions, which must be run in order to release a new design to production. As with the relative analysis process, it is critical to use consistent procedures and test methods for durability testing in order to establish a historical baseline for pass/fail criteria. This baseline, in addition to any field history data, is the primary piece of data necessary to develop component-based fatigue limits for the analysis procedures. On completion of either engine or rig testing, it is imperative that the component be thoroughly inspected. While secondary failure symptoms such as coolant pressurization or leakage can indicate complete propagation of a crack to failure, the most robust designs and procedures would address crack initiations as well. In a complex block or head casting, these initiations can only be identified through destructive testing. The block or head must be sectioned through each cavity (water jacket, port, oil drain back, etc.), so all surfaces are visible. Crack detection is then accomplished through Magnaglo or Magnaflux (ferrous components) or dye penetrant (nonferrous components). Figure 15 shows an example of a sectioned cylinder head under Magnaglo inspection.

5 GEOMETRY CONSIDERATIONS

5.1 Geometry issues created by CAD

Even with the advances in modern CAD programs, issues still exist in the conversion of CAD geometry to good FEA or CFD meshes. The complex geometries required for cylinder head and block water jackets can lead to generation of very poor geometric features such as thin slivers or slices within the CAD model (Figure 16). These features in turn can lead to poor mesh quality and elements with high calculation errors. To the extent possible, areas with geometry issues should be corrected, and the analysts should verify an adequate mesh quality for valid results.

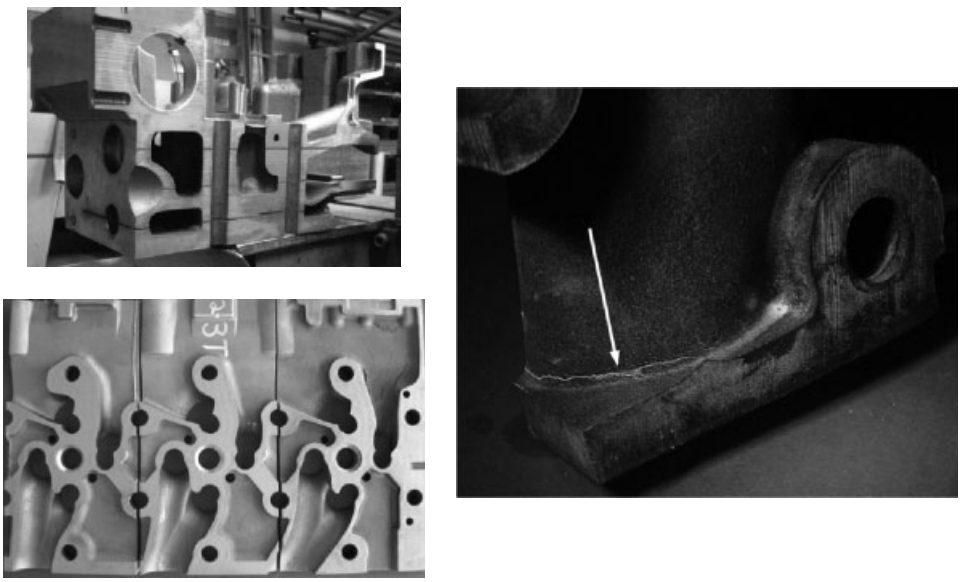


Figure 15. Example of Magnaglo inspection of sectioned cylinder head.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

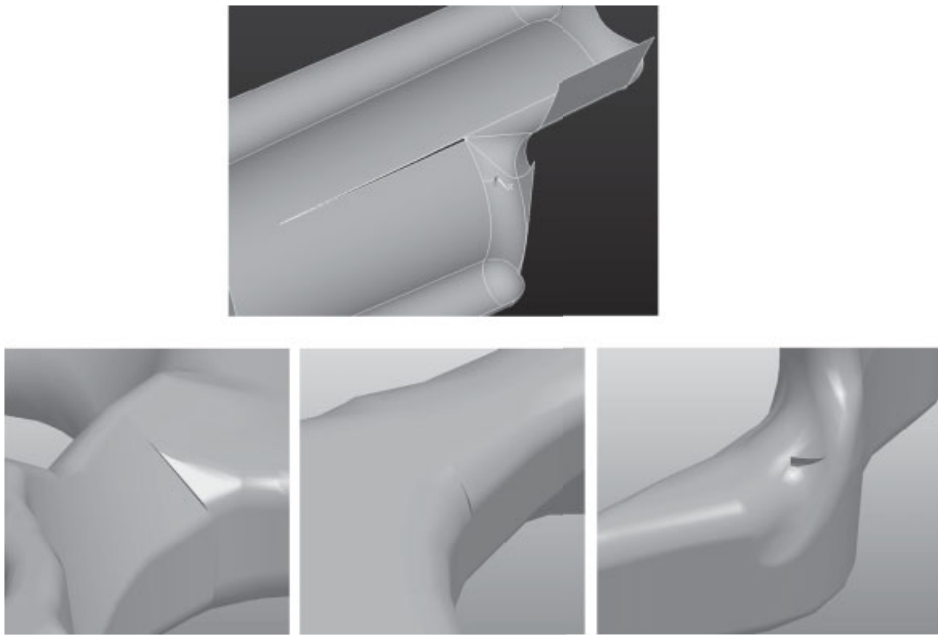


Figure 16. Example of CAD geometry issues which lead to poor element quality.

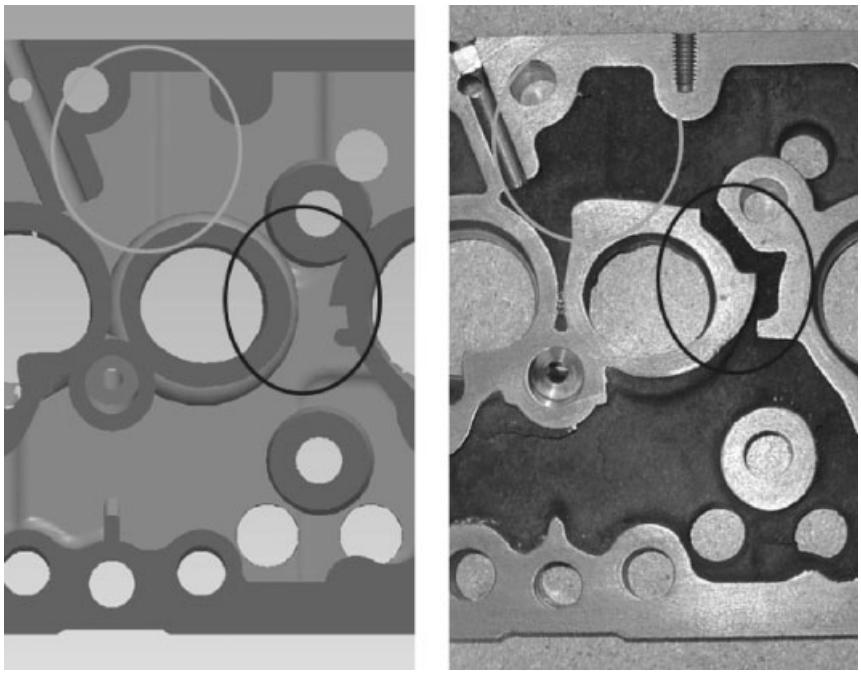


Figure 17. Potential differences between CAD geometry and actual cast production parts.

5.2 Inaccurate geometry

The issue of inaccurate geometry is especially prevalent when analyzing production or legacy designs to develop analysis procedures and component-based fatigue limits or address field issues, which may occur during production. CAD systems continually progress in capability with every release. Early versions of CAD systems could not handle the complex geometry and rounds required for cylinder head and block water jackets. Therefore, legacy CAD

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

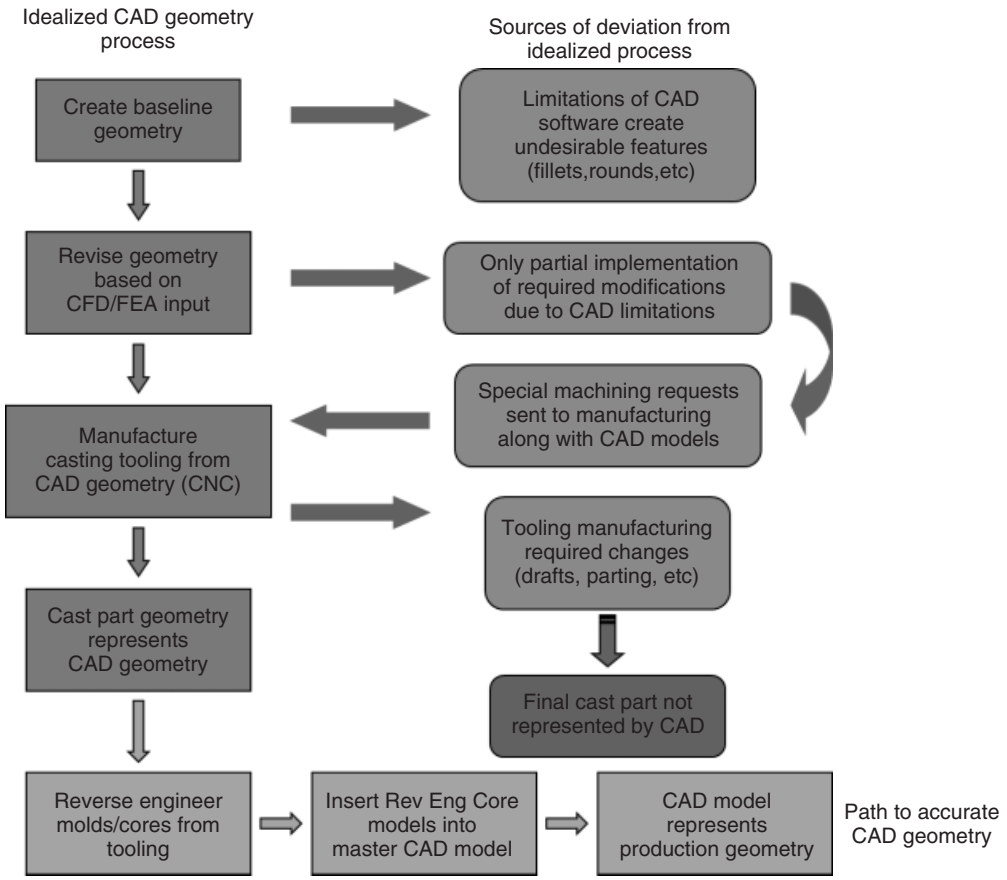


Figure 18. Actual CAD geometry process from design to production.

models typically have significant geometry differences from the actual production part.

This problem is not relegated to legacy products however. Once the CAD geometry leaves the designer and progresses to manufacturing, the potential exists for multiple sources of modification compared to the geometry originally analyzed during the design process. This is especially true for the manufacture of complex castings such as blocks or heads. Figure 17 shows an example of differences between CAD and actual cast geometry. Figure 18 shows the idealized process for progressing CAD geometry from design to production and highlights the potential sources of modification to a given block or head geometry and the process most often followed to take a new design to production.

ACKNOWLEDGMENTS

The author would like to acknowledge the work of Mark Tussing as one of the first people to introduce the concept of the four classifications of analysis and how to effectively apply CAE analysis to engine design.

RELATED ARTICLES

Engine Configurations
Critical Layout Dimensions
Gaskets and Sealing
Cooling Systems
Diesel Fuel Injection Systems

REFERENCES

Bannantine, J.A., Comer, J.J., and Handrock, J.L. (1990) *Fundamentals of Metal Fatigue Analysis*, Prentice Hall, Upper Saddle River.

Shigley, J.E. and Mischke, C.R. (1989) Mechanical Engineering Design, 5th edn, McGraw-Hill, Inc, New York.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto136
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

NVH Considerations in Engine Development

Thomas E. Reinhart
Southwest Research Institute, San Antonio, TX, USA

1 Introduction 1
2 Vibration 2
3 Engine Radiated Noise Generation 4
4 Noise Transmission Path 7
5 Engine Exterior Surfaces 7
6 Radiation Efficiency 8
7 Vibration Modeling 9
8 Noise Modeling 10
9 Low-Noise Engine Design Guidelines 11
References 12

1 INTRODUCTION

Engines are one of the primary noise and vibration sources in most vehicles and machines. Under certain operating conditions, engine-radiated noise can dominate both interior and exterior noise levels, and under some conditions, engine-generated vibration dominates the vibration perceived by the driver and passengers.

1.1 Engine noise

Engines generate noise in a number of ways. Engine-radiated noise comes directly from vibrating surfaces of the engine. Engine surfaces function as loudspeakers, broadcasting the effects of internal engine forces as audible noise. There are a number of approaches available to the designer to reduce engine-radiated noise, and some of these approaches are explored in this chapter. Engines also put pressure pulsations into the intake and exhaust systems. These pulsations can propagate through the intake and exhaust systems, resulting in audible noise from the air inlet and exhaust outlet. Intake and exhaust pressure pulsations can also cause vibration of the intake and exhaust systems, which lead to noise being radiated from the surfaces of these systems. There is generally little that an engine designer can do to reduce intake and exhaust pressure excitations, without having a significant effect on the engine's performance. Therefore, intake and exhaust pressure pulsations are generally accepted as a given. The intake and exhaust systems are designed to achieve the desired levels and frequency content of noise, given the pressure inputs provided by the engine. The design of intake and exhaust systems is covered in Gas–Breathing and Air Management.

Over a range of engine size and type ranging from small gasoline passenger car engines to heavy-duty truck diesel engines, noise levels at the rated condition (maximum speed and load) tend to be similar. In general, engine-radiated noise is a significant issue from 300 or 400 Hz up through 20,000 Hz, the maximum frequency of human hearing. Intake and exhaust noises are often a lower frequency phenomenon. For example, a four-cylinder four-cycle engine running at 2000 rpm has a firing frequency of 67 Hz. Intake and exhaust noises can be expected at this frequency, and at several harmonics (multiples) of this fundamental frequency. Intake and exhaust noises are not limited to low frequencies, however. Compression brakes, common on heavy-duty truck engines in some markets, produce high exhaust noise levels over a wide frequency range (Reinhart and Wahl, 1997). Turbochargers can produce high frequency tones related to once per revolution frequency of the turbocharger shaft, and to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

blade pass frequencies. These very high frequency sounds (sometimes beyond the range of human hearing) will propagate through the intake and exhaust systems. Overall, engines can generate significant noise over the entire range of human hearing (20–20 kHz).

1.2 Engine vibration

Engines vibrate as a result of internal forces generated by the operation of the engine. This vibration can reach the chassis or cabin through the engine mounting system, or via any other linkage between the engine (or powertrain) and another vehicle component. Thus, coolant and hydraulic hoses, the intake and exhaust systems, wiring harnesses, cables, and any other attachment may provide a transmission path allowing engine vibration to reach the vehicle.

Once engine vibration has been transmitted into the vehicle structure, it may become apparent as either tactile vibration or audible sound. Audible sound results when engine vibration causes some vehicle component to vibrate in a way that generates sound. Tactile vibration appears in forms such as steering wheel or seat track movement, mirror vibration, and many other common vibration responses in a vehicle or machine.

Tactile vibration is generally a low frequency problem, as human sensitivity to vibration rolls off rapidly as frequency increases. When vibration frequency increases to 200 Hz, it must have very high acceleration amplitude for a human to sense it (Brüel & Kjær, Inc., 1998). The most common source of tactile vibration is engine firing frequency. For example, a four-cylinder, four-cycle engine will have a firing frequency of 20 Hz at 600 rpm idle speed. It is very difficult to fully isolate a frequency this low from the chassis; therefore, it is common to feel some vibration in the vehicle as a result. As the engine speed increases, however, the mounting system is more effective at isolating higher frequency vibration, and as the human body is less sensitive to higher frequency vibration, the perception of vibration fades away. Often, unfortunately, the tactile vibration is replaced by an audible response generated by the cabin from the remaining vibration input from the engine. In the case of our four-cylinder example, this audible noise is often referred to as *four-cylinder boom*.

The lowest frequency vibration issue that is common to an engine occurs at half order (once per two engine revolutions), which is the firing frequency of any given cylinder in a four-cycle engine. This half-order vibration is the result of variation in combustion from cylinder to cylinder. The worst case example of this is complete misfire in one cylinder.

2 VIBRATION

Vibration can be split into two categories. Rigid body vibration occurs when the engine (or the complete powertrain, consisting of all components rigidly attached to the engine) moves as a rigid body. Nonrigid vibration occurs when the engine or powertrain deflects as it vibrates. Typically, nonrigid vibration occurs at or near resonant modes of the engine or powertrain. An example of a resonant mode is a powertrain bending mode, where the engine and transmission move out of phase with each other.

Vibration is caused by forces generated inside the engine. These forces are caused by events such as cylinder pressure fluctuations, impacts between engine components, and unbalanced reciprocating forces and moments.

2.1 Unbalanced forces and moments

As the crankshaft rotates, the piston and rod experience alternating accelerations as they move up and down the bore. Most crankshafts have counterweights in an effort to balance these forces. In a single-cylinder engine, it is impossible for crankshaft counterweights to fully cancel the forces involved in the reciprocating motion of the piston and rod because the counterweights develop a rotating force rather than a reciprocating force. When the engine has more than one cylinder, the combination of reciprocating and rotating forces may add up in such a way as to cause a moment. This is a force that causes the engine to rotate about the roll axis, or another axis perpendicular to the roll axis. The roll axis is the axis of least inertia for an engine or powertrain. In multicylinder engines, there are certain configurations where it is possible to fully cancel both reciprocating forces and moments. Inline six-cylinder engines and V-12 configurations with the proper bank angle (60° or a multiple of 60) are examples of this. Other common engine configurations have some remaining unbalanced force and/or moment.

In some cases, balance shafts can be added to cancel forces or moments that cannot be dealt with in the crank train design. For example, inline four-cylinder engines often use two counterrotating balance shafts to eliminate the second-order free force generated by this configuration. The Bosch Automotive Handbook has a very good section on engine balancing, including a table showing equations for the unbalanced free forces and moments for many common engine configurations (Robert Bosch GmbH, 2000).

Because diesel engines typically use higher cylinder pressures than spark-ignition (SI) engines, it is common for diesel engines to have a higher piston and rod mass for a given engine displacement. This makes the unbalanced forces and moments of diesel engines larger, and thus more of a challenge, than those of an SI engine.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.2 Gas pressure reaction forces

As a piston goes up the compression stroke, work is put into compressing the charge in the cylinder. This causes a negative torque on the crankshaft, which has the effect of slowing down the crank. As the piston goes down the power stroke, positive torque is applied to the crankshaft, speeding it up. Thus, the crankshaft experiences speed fluctuations from the alternating torques applied by the pistons and rods. The alternating torque that causes these crankshaft speed fluctuations reacts against the engine block. As the crankshaft accelerates in one direction, the engine block accelerates in the opposite direction. The result of this is that the engine rocks back and forth in response to gas pressure forces. This rocking motion is called the *roll response* of the engine, where the engine moves about the "roll axis."

The roll response does not necessarily occur about the crankshaft axis. It occurs about the axis of least inertia of the powertrain—the roll axis. This axis typically starts at the front of the engine above the crankshaft centerline and ends at the rear of the engine somewhere near the crankshaft centerline. The exact location of the roll axis varies with the mass distribution of the engine and all rigidly attached components, such as the transmission and transfer case.

The amplitude of vibration about the roll axis depends on a few simple parameters:

- powertrain mass
- cylinder pressure
- cylinder displacement.

Vibration amplitudes owing to gas pressure reaction forces increase with cylinder pressure and displacement and decrease with mass. This does not give the engine designer much room to work with. Reducing the displacement or cylinder pressure will decrease the performance, while increasing the mass is detrimental to the overall vehicle design. One way to decrease gas pressure reaction forces for a given power level is to increase the number of cylinders in the engine, thus decreasing the displacement of each cylinder. While this is an attractive approach from a smoothness viewpoint, it tends to decrease engine efficiency and increase cost.

The gas pressure reaction forces of diesel engines pose a larger challenge than those of SI engines. There are two reasons for this. First, diesel engines typically have a higher compression ratio than SI engines. The higher compression ratio creates higher cylinder pressures, which in turn lead to higher gas forces. Second, diesel engines typically operate unthrottled, except where throttling is used for emission control reasons, such as to drive exhaust gas recirculation (EGR) flow or to regenerate emissions after treatment devices. The air/fuel ratio of a diesel engine varies widely, becoming very high at light load. In most SI engines, the air/fuel ratio is maintained at or close to the stoichiometric value by throttling. Thus, at light load, an SI engine has very little air in the cylinder, leading to low gas pressure reaction forces. A diesel engine operating at light load, on the other hand, typically has a large quantity of air in the cylinder, giving higher gas forces. Gas force reactions can be a particular issue for diesel engines operating at low speed and light load, where the driver expects the engine to be smooth. Some diesel engines use an intake throttle during shutdown, to reduce gas forces as the engine firing frequency passes through powertrain mounting resonance frequencies at speeds between idle and zero.

2.3 Minimizing vibration transmission to the cabin

The engine designer has some control over two issues that cause vibration issues in a vehicle. The engine should be designed to achieve the highest possible powertrain bending natural frequency. For an inline four-cylinder engine, the powertrain bending natural frequency should be well above the second order at maximum engine speed. For an inline six-cylinder engine, the powertrain bending frequency needs to be well above the maximum driveshaft speed in the highest (typically overdrive) gear. If the engine operates at the speed where the powertrain bending frequency is excited, the vibration of the powertrain is substantially amplified by the resonant response, leading to high vibration excitation of the vehicle.

The other issue that is at least partly under the control of the engine designer is the natural frequency of accessories mounted on the engine. Where possible, the accessories should have a mounted natural frequency that is significantly above the firing frequency at maximum engine speeds. This goal is absolutely critical for inline four-cylinder engines, which have very high excitation amplitudes of the second order from unbalanced inertial forces as well as gas reaction forces.

Engine mounting systems need to be designed to isolate the engine vibration at low speed. This can be a challenge because excessively soft mounts lead to a loss of control of powertrain position when the vehicle is subjected to road input (such as a large bump). Engine mounts may be designed with nonlinear stiffness to allow relatively free engine motion near idle while still providing control of engine position under high loads and during road input events. Hydraulic damping is also sometimes used to improve mount performance. In a few cases, actively

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

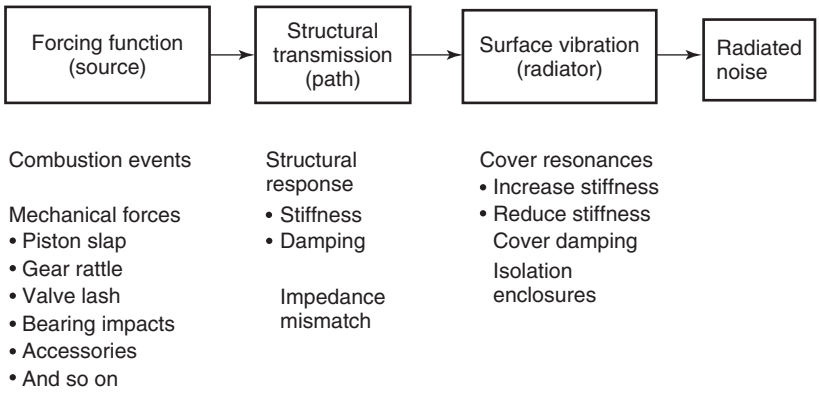


Figure 1. Source–path–radiator model of engine noise generation.

controlled mounts are used. This relatively expensive approach involves generating forces that offset inputs from the engine, or actively varying hydraulic damping based on operating conditions.

Vehicle designers also have a responsibility to help isolate engine vibration from the vehicle. In addition to mounting systems, it is important to avoid a situation called *modal alignment*. This occurs when vibration modes of the vehicle line up with a common excitation force from the engine, such as firing frequency at idle. Another source of trouble is when more than one vehicle component or substructure has resonances at the same frequency as another vehicle component. This issue is called *modal alignment*.

Driven by both regulatory and customer requirements for increased fuel efficiency, engines are tending toward lower cylinder count and higher cylinder pressure (higher brake mean effective pressure, BMEP) for a given application. These trends make the job of an NVH (noise, vibration, and harshness) engineer more difficult, as a lower cylinder count typically involves larger free forces and moments. These larger forces occur at lower frequencies, which make them harder to isolate. The higher cylinder pressures also create larger gas reaction forces—another challenge for the NVH engineer.

3 ENGINE RADIATED NOISE GENERATION

The sources of noise include combustion pressure and a range of mechanical events. These sources do not radiate noise directly because they generally occur deep within the engine structure. However, they do apply large forces to internal engine components. These forces excite vibration in the engine structure (block, cylinder head, crankshaft, etc.). When the vibration reaches exterior surfaces, noise is radiated. The process can be viewed as a source (forcing function), path (structural vibration transmission), and radiator (vibrating surface) model, as shown in Figure 1.

3.1 Combustion excitation in diesel engines

Diesel and SI engines have fundamentally different approaches to controlling combustion, differing in their combustion excitation. In a diesel engine, air is compressed in the cylinder until it is hot enough to cause the fuel to burn without any ignition source. Combustion is controlled by the fuel injection system. Injection timing, rate, and pressure are three parameters that are typically controlled by the injection system. Once fuel enters the combustion chamber, it takes some time to evaporate and warm up to the temperature that enables combustion to start. The time between the start of fuel injection and the start of combustion is called *ignition delay*. Depending on the operating condition, injection may continue well after the start of combustion. Figure 2 shows a typical cylinder pressure trace for a diesel engine operating at high load, but with low boost pressure. This condition is typical of a naturally aspirated engine or a turbocharged engine under transient acceleration.

At the start of combustion, virtually all of the fuel injected during ignition delay ignites and burns very rapidly, causing a spike in cylinder pressure. This event is called the *premix burn spike*, and this spike in cylinder pressure is what gives diesel engines their characteristic sound. The step increase in cylinder pressure acts similarly to an impact force, which is applied to both the piston crown and the cylinder head. The size of the premix burn spike determines the rate of pressure rise in the cylinder and thus the amplitude of the combustion excitation applied to the engine structure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

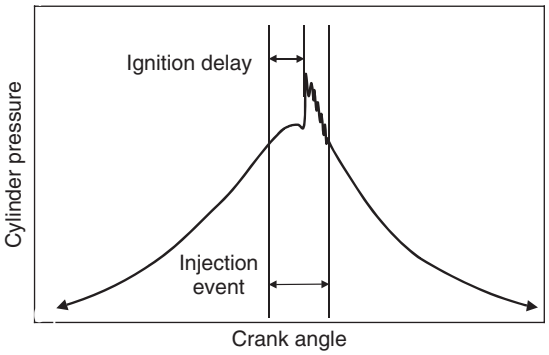


Figure 2. Cylinder pressure trace of a diesel engine operating at high load, but with low boost pressure.

Several parameters influence the size of the premix burn spike and thus the level of combustion noise. These include the following:

- *Intake Manifold Temperature*. A higher temperature shortens ignition delay, reducing the excitation.
- *Compression Ratio*. A higher compression ratio provides a hotter, denser charge at the start of injection, reducing ignition delay and combustion excitation.
- *Injection Timing*. Timing near top dead center (TDC) produces the shortest ignition delay, whereas very advanced or retarded timing causes a longer ignition delay.
- *Boost Pressure*. Turbocharging can greatly reduce ignition delay because of the higher temperature and pressure in the cylinder. At full load, many turbocharged engines have an ignition delay of only 2 or 3 crankshaft degrees, and the start of combustion becomes invisible on the cylinder pressure trace. In general, combustion noise tends to become insignificant with boost pressures over 70 kPa.
- *Injection Characteristics*. The use of a pilot injection (a small quantity of fuel injected ahead of the main event) can greatly reduce combustion excitation by establishing combustion with a very small amount of fuel, and then injecting the remaining charge into the already burning mixture. Injection pressure is also an important parameter. Lower injection pressure leads to a slower rate of heat release once combustion begins because droplets of fuel are less fully atomized. This reduces the rate of pressure rise.
- *Cetane*. This measures the willingness of the fuel to auto-ignite. Higher cetane values give a shorter ignition delay, and thus lower combustion noise. Unfortunately, cetane values vary significantly in markets around the world and are outside of the engine maker's control
- *EGR*. The use of EGR slows combustion, reducing the excitation from a given level of ignition delay. EGR also tends to increase ignition delay; therefore, it is not always beneficial for noise.

With stringent emission regulations, some diesel engines use more than one combustion mode. There may be step changes in the number of injection events, air/fuel ratio, EGR rate, and so on. Engine designers must find ways of minimizing the potential for step changes in combustion noise that can result from these combustion mode changes (Reinhart, 2009).

3.2 Combustion excitations in SI engines

SI engines typically draw a mixture of air and fuel into the cylinder during the intake stroke, or, if they use direct injection, the injection event is not directly the cause of combustion. In an SI engine, it is the spark timing that determines the start of combustion, and combustion propagates from the ignition location across the combustion chamber. Overall rates of combustion (rates of heat release) tend to be lower in an SI engine than in a CI engine. This typically results in much lower rates of pressure rise than are found in a diesel engine, and thus lower combustion noise.

Combustion noise can still be a challenge in SI engines, depending on the combustion strategy chosen by the designers. Combustion systems that are developed to maximize the rate of combustion typically have the most issues with noise. Increasing the combustion rate is often desired for emissions or efficiency reasons. In general, any measures taken to increase charge motion in the cylinder, such as increased tumble or squish, tend to increase the rate of combustion and thus combustion noise. The use of dual spark plugs is another approach that tends to increase combustion rates and noise by providing two flame fronts, each of which can consume a significant portion of the available air/fuel mixture simultaneously, before the two flame fronts merge.

3.3 Mechanical excitations

Engines can radiate noise because of forces from a range of mechanical sources. Some of the most common sources are listed:

- Gear backlash impacts, driven by
  - crankshaft torsional vibration
  - fuel system torque inputs
  - valve train and other sources of torque input.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- Axial shaft motion and impacts
  - often driven by a combination of gear backlash impacts and helical gears.
- Gear meshing frequency pure tones
- Meshing impacts from chain or belt timing drive systems
- Turbocharger blade pass, once per revolution, and other pure tones
- Piston slap
- Valve and injector train clearance impacts
- Bearing impacts, as clearances are taken up
- Fuel system hydraulic and mechanical dynamics (diesel and direct injection gasoline)
- Crankshaft bending, driven by cylinder pressure and crankshaft distortion in response to cylinder pressure and inertial forces
- Forces generated by accessories.

At higher speeds and loads (with the exception of naturally aspirated diesels), most engines are dominated by mechanical noise sources, and some engines are mechanically dominated across the full operating range. Most diesel engines are combustion dominated at idle and under transient operation where boost pressure is low.

3.4 Separating combustion and mechanical sources

Many mechanical and combustion noise sources sound alike: impact-like events that repeat at firing frequency. Impact-like events are broadband, which means they excite a wide range of frequencies. This allows them to excite many engine and component resonances. In the case of a highly resonant structure, the frequency spectrum of the vibration or noise response will show distinct peaks, even though there may not be distinct peaks in the excitation force spectrum. Because all of these impact-like events share similar characteristics, they are very hard to separate from each other, either by listening or by analyzing the results of a noise measurement.

A few mechanical noise sources, such as gear whine, turbocharger noise, and certain types of accessory noise, can be identified and quantified using frequency analysis. Knowledge of the engine speed and number of gears in a mesh, for example, can be used to quantify gear meshing whine amplitude. (Note that for a single plane gear train, all meshes in the train have the same meshing frequency. As a result, it is possible to identify and quantify gear whine, but it is not possible to assign responsibility for the noise to any specific mesh in a train.)

Three basic approaches are available to help separate combustion and mechanical noise sources. One approach is to intentionally modify the combustion noise. In a diesel engine, this can be done by modifying injection timing, by changing fuel cetane, by changing intake air temperature, or by turning pilot injection on and off. Changing injection timing has the drawback that it may also substantially change a mechanical noise source such as gear rattle, in addition to changing combustion noise. The other methods are less prone to this error, but all approaches come with the risk of unintentionally changing mechanical noise. In SI engines, combustion noise can be changed by modifying tumble or squish, or by disconnecting one spark plug in a dual-plug system. Again, changes made in an effort to modify only combustion noise sometimes have an unintended effect on mechanical noise.

Another method for evaluating the amplitude of combustion noise is to measure cylinder pressure and feed the signal into a combustion noise meter. The meter will provide an estimate of the sound pressure level emitted by the engine, based only on combustion excitation. The combustion meter level can then be compared with the measured noise level to determine the portion of noise related to combustion. A limitation of this approach is that the meter uses an assumed transfer function between the cylinder pressure spectrum and radiated noise. In practice, engine designs vary in their transfer functions, and hence this method is prone to some error.

A third method for separating combustion and mechanical noise sources is to eliminate some particular mechanical source. A gear train can be replaced with a belt drive for test purposes, or teflon-padded pistons can be used to eliminate the possibility of piston slap. Accessories not essential to engine operation can be simply removed or remotely operated for a test. These tests can give reliable measurements of the contribution from a particular noise source.

3.5 Gear rattle

Gear rattle is of particular concern in medium- and heavy-duty diesel engines (Zhao and Reinhart, 1999). These engines typically have fairly large timing drive systems with several gear meshes. The methods available to reduce gear rattle noise include reducing the torsional input to the gears or suppressing the rattle with devices that eliminate lash. Torsional input to the gear train can be reduced by increasing the capacity of the crankshaft damper or by adding damping to gears in the timing drive. Some unit injector engines use viscous dampers on the camshaft gear to reduce fuel system torsional input to the gear train. Other engines use scissors gears to reduce or eliminate

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the last in gear meshes. Putting the gear train at the flywheel end of the engine is another effective noise reduction technique. Crankshaft torsional vibration tends to be lower at the flywheel because of the high inertia at that point. This effect is most pronounced at high speeds and high loads, where crankshaft deflections are largest.

4 NOISE TRANSMISSION PATH

Engine noise is not typically radiated directly from the points where forcing functions are applied to the structure. In most cases, these points are buried within the engine structure. The forces induce deflections and vibration in the structure. This vibration propagates out to exterior surfaces, which in turn radiate noise.

One goal of engine structural design is to minimize the vibration of exterior surfaces in response to a given forcing function input. The most effective way to achieve this is to provide an impedance mismatch (step change in stiffness) between the exciting force and the exterior surface. Ideally, the structure should be very stiff where the force is applied, then relatively compliant, and then stiff again at or just below the exterior surface. While this approach can be very effective, it is also very difficult to realize in a cast component such as a cylinder block or head. As a result, the degree of improvement that can be achieved by structural optimization is limited compared to what can be achieved with control of input forces or the design of engine covers.

The design tools used for structural optimization are finite element model (FEM) and boundary element model (BEM). FEM is used to determine the dynamic response of a structure to a given force input. The input force may be measured on a running engine, or it may be an assumed force. The model is iterated to reduce the vibration response at selected points because of force inputs at selected locations. For example, a block model may be modified to reduce vibration at the oil pan mounting flange or on the block skirt. If a direct calculation of radiated noise is desired, the FEM results can be fed into a BEM to determine the radiated noise that results from a given force input and structural characteristics.

5 ENGINE EXTERIOR SURFACES

Exterior surfaces vibrate in response to the movement of the underlying structure. As the surface vibrates, it creates pressure fluctuations in air that propagate as audible noise. There are a number of techniques available to minimize noise radiation from exterior surfaces.

5.1 Damping

Highly resonant covers will amplify the input from the underlying structure at resonance frequencies. Adding damping to a resonant cover will reduce the amplitude of these resonant peaks, bringing down the overall noise level radiated by the cover. Damping can be added by changing the material used to make the cover, by modifying the bolted joint between the cover and the structure, or by adding damping features such as constrained layer treatments.

A constrained layer damping treatment consists of the cover wall, a thin layer of polymer adhesive, and a secondary wall (the constraining layer). As the cover vibrates, there is relative motion between the cover wall and the secondary wall. This relative motion imposes strain on the polymer layer, and the viscoelastic properties of the polymer layer enable it to absorb some of the vibration energy in the component. Adding an effective damping treatment to a highly resonant cover can achieve a noise reduction of up to 3 dB.

Covers made of cast aluminum, cast magnesium, cast iron, or stamped steel tend to be lightly damped, and hence they often benefit from the addition of damping treatments. Composite covers with a high level of reinforcing fiber may also be lightly damped. Composites with low fiber content often enjoy high inherent damping levels, depending on the material properties of the material used.

5.2 Stiffness increase

Many designers respond almost instinctively to a noise issue by trying to increase the stiffness of the responsible cover. The goal of a stiffening exercise should be to increase the first resonance of the cover to a frequency beyond the range of significant excitation. Unfortunately, in most SI engines, there is significant excitation up to 3 or 4 kHz, whereas in diesel engines, there can be significant excitation at even higher frequencies. As a result, increasing stiffness tends to be successful only when applied to relatively small components. It is fairly easy to achieve a first resonance of 5 kHz in a component that is <100 mm long, but it is impossible to do this on a 1 m $\times$ 200 mm valve cover on a large engine.

5.3 Stiffness decrease

Reducing the stiffness of an engine cover that has high noise levels is counterintuitive, but for large covers on larger engines (especially engines of 8 L or more in displacement), it can be a very effective noise reduction

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

approach. Vibration of a cover does not translate into sound on a 1 : 1 basis. Depending on the mode shape (vibration pattern) of the component, only part of the vibration may translate into radiated sound. The relationship between vibration amplitude and radiated sound is called *radiation efficiency*, and this will be described in more detail in Section 6.

In general, thin, flat surfaces achieve the lowest radiation efficiencies. Therefore, large, flat stamped steel components tend to be quieter than the same component made in thicker, stiffer cast aluminum. It is generally impossible to do a large casting with a wall thickness small enough to achieve low radiation efficiency.

5.4 Isolation

Isolating a cover from the underlying structure can provide a large noise reduction. The goal of isolation is to separate the cover from the underlying structure with a soft isolation gasket and grommets. This provides a large impedance mismatch (Section 4), and can greatly reduce the amount of vibration transmitted to the cover. Noise reductions of up to 7 or 8 dB can be achieved with a very effective isolation system. Many engine isolation gasket systems are designed to achieve vibration isolation above 200 or 300 Hz. The choice of isolation frequency represents a trade-off between noise reduction performance and other desired characteristics such as accurate location of the cover, prevention of oil leaks, and durability of the isolation system. Typical examples of isolated covers on engines include valve covers and oil pans.

One of the challenges of isolation design is to avoid leaks. Leaks are typically at a minimum when there is very high unit loading between the gasket and the mounting flange. Unfortunately, high unit loads tend to go along with high overall joint stiffness. Small beads on the gasket can achieve high unit loads with relatively low overall stiffness, thus minimizing leaks while achieving good isolation performance.

5.5 Enclosures

Noise enclosures are typically a last resort. In a sense, they are an admission of failure: we could not fix the noise, so we had to cover it up. Enclosures also tend to be expensive and difficult to package. On the other hand, enclosures can be very effective in achieving a given noise requirement quickly.

Enclosures work on the principle of isolation (impedance mismatch). They are mounted on the engine in such a way as to minimize vibration transmission to the enclosure. Most acoustic enclosures are designed to have high internal damping, to reduce the tendency for resonances to increase noise radiation. Enclosures may also have a noise-absorbing material on the side facing the engine to absorb the noise radiated by the underlying engine components and to prevent the noise from leaking out from gaps between the enclosure and the underlying engine.

6 RADIATION EFFICIENCY

Radiation efficiency is a measure of how much of a surface's vibration gets translated into radiated sound. The efficiency is measured in percentage, such that a value of 100% indicates that all vibration is translated into sound. Radiation efficiency is a very important concept for engine NVH engineers to understand and make use of, particularly for large covers on larger size engines. Because radiation efficiency is counterintuitive, engineers employing techniques to reduce radiation efficiency may need some salesmanship in order to get their ideas approved for implementation (or even for consideration).

6.1 Explanation of radiation efficiency

Consider the example of a loudspeaker. The speaker is designed so that the cone moves as a rigid body in the frequency range of interest. This results in the entire cone moving uniformly. As the cone moves, it pushes on the surrounding air, creating a pressure fluctuation in the air. At very low frequency, as the cone moves, the air can easily move out of the way of the cone. There is a local acoustic pressure near the cone, but this pressure decays rapidly to zero as you move away from the cone. As the frequency increases, some of the air cannot move back and forth to get out of the cone's way; therefore, compression waves are formed, which radiate as sound. Higher the frequency of cone movement, lesser the air that gets out of the way, and more the sound that is radiated. At the critical frequency, where the wavelength of sound in air matches the diameter of the cone, 100% radiation efficiency is achieved. Further increases in frequency have no significant effect on radiation efficiency.

The wavelength of sound at 10,000 Hz in air is around 25 mm, so that a 25-mm speaker will achieve 100% radiation efficiency at 10 kHz. On the other hand, at 100 Hz, the wavelength of sound in air is around 3 m. Thus, a huge speaker cone would be required to achieve 100% radiation efficiency at 100 Hz. This effect explains why tweeters are small and woofers are large, and why bigger is generally better for woofers. The same situation applies to the exterior

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

surface of an engine. If a large surface area moves in a uniform manner, radiation efficiency is high. If the large surface area has a very complex mode shape (vibration pattern), radiation efficiency can be very low.

6.2 Radiation efficiency examples

Figure 3 shows a comparison of oil pan surface velocities for a thin-wall stamped steel oil pan and a thicker, ribbed cast aluminum oil pan. Both oil pans were tested on an 11-L heavy-duty diesel engine. In comparing the two oil pans, the steel pan has higher vibration levels up to about 1000 Hz and again from 2000 to 3000 Hz. On the basis of measured vibration levels, we would expect a higher noise level. However, Figure 4 shows the measured sound power levels for the two oil pans. The steel pan is quieter than the aluminum pan at all frequencies, and above 1000 Hz, the advantage is around 10 dB.

The steel pan suffers somewhat higher vibration levels than the aluminum one because it is much less stiff. However, this lower stiffness translates into a large advantage in terms of lower radiation efficiency and thus lower radiated noise.

6.3 Parameters that control radiation efficiency

Several parameters influence radiation efficiency; one is size. A larger cover has the potential to achieve higher radiation efficiency, especially at low frequencies. Shape is another important factor. Adding shape to a surface adds a lot of stiffness, increasing radiation efficiency. Large, flat surfaces lend themselves to a low radiation efficiency design. Material modulus and density both play a role. Higher modulus increases stiffness, and thus radiation

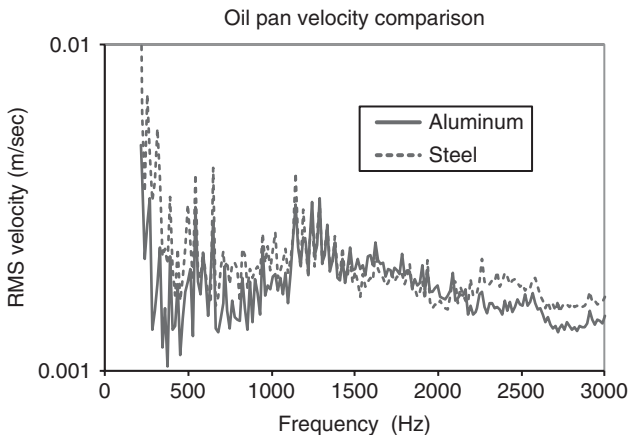


Figure 3. Comparison of stamped steel and aluminum oil pan mean square velocities on an 11-L HD diesel engine.

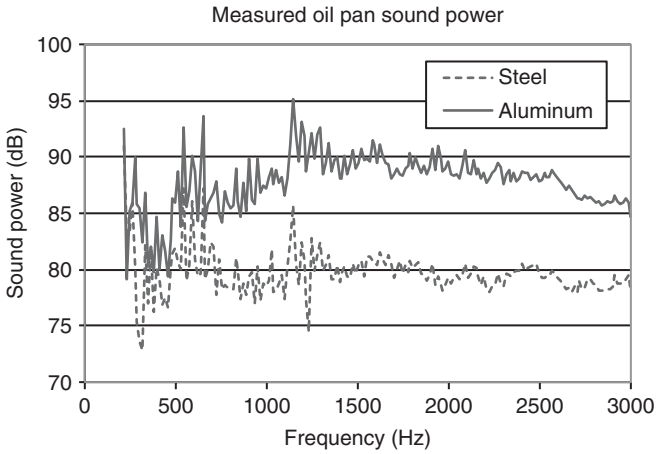


Figure 4. Comparison of stamped steel and aluminum oil pan radiated sound power on an 11-L HD diesel engine.

efficiency. Higher density reduces natural frequency by adding mass to the component. In the case of steel and aluminum, the two factors roughly cancel each other out. Steel has both higher modulus and higher density; therefore, a component with identical dimensions will have a similar radiation efficiency in either material. The advantage of steel is that components can often be made much thinner, which results in lower natural frequencies, more complex mode shapes at a given frequency, and thus lower radiation efficiency. In the case of the two oil pans in the earlier example, the steel pan had mostly flat surfaces of 1.5 mm thickness, whereas the aluminum pan had stiffening ribs and a wall thickness of 5 mm.

6.4 Designing for low radiation efficiency

Designing large covers with a thin, flat surface is the easiest way of obtaining low radiation efficiency. Care must be taken to avoid introducing other problems when using this approach. For example, components with a low natural frequency can be excited by engine firing frequency, resulting in high vibration levels and high stress. FEM can be used to tweak a design in ways that limit stress, even when a vibration mode is excited by firing frequency. The cover design must also meet other structural requirements of the engine, which can be a challenge in a low stiffness design.

7 VIBRATION MODELING

Tactile vibration is a low frequency issue. At low frequencies, the engine (and assembled powertrain) typically moves as a rigid body. Thus, for purposes of engine mount

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

development, simplified rigid body models of the powertrain can often be used. In the simplest case, the chassis is assumed to be infinitely stiff. This approach often leads to significant errors; therefore, the model is usually set up to include the stiffness and dynamics of the chassis or frame that the engine mounts on.

Many powertrains can suffer from powertrain bending frequencies that occur within the operating range of the engine. In larger powertrains, these resonances may occur at frequencies where the vibration can be felt by the operator. Powertrain bending resonances can also lead to fatigue and failure of components such as flywheel housings and transmission bell housings. To deal with this, a simplified FEM of the entire powertrain structure can be used. The model only needs sufficient detail to calculate the first few mode shapes and natural frequencies of the overall powertrain so that the models can be much less detailed than those used for stress or thermal calculations.

8 NOISE MODELING

More detailed FEMs are used to model vibration that will result in radiated noise. These models need to cover all the mode shapes in the frequency range of interest. The size of finite elements should be selected to have at least five or six elements over the length of the highest frequency mode shape of interest. Consider a flat rectangular valve cover that is 200 mm wide and 600 mm long. To define the first mode shape, only 5 or 6 elements across are required, along with 15–18 elements down the length of the cover (to keep elements roughly square). To handle the $n = 6$ mode shape across the cover, 30–36 elements across would be required, and three times that number down the length. This is a far simpler model than that required for any sort of stress analysis. Most models that are actually used are far finer than this example would suggest, but unlike in the case of a stress model, additional resolution does little to improve the accuracy of the result.

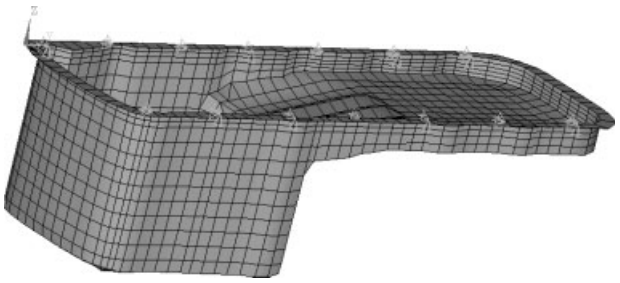


Figure 5. Finite element model of stamped steel oil pan.

Noise models can be done at the component level, at the engine level, or at the full powertrain level. Modeling individual components has advantages. It is far easier to determine the input to an engine cover, for example, than to the block or cylinder head. Vibration at the mounting flange of a cover can be directly measured and fed into the model as an input. Another advantage of small models is run time. With faster run times, it is easy to explore many design options. A very large model with a very long run time may limit the analyst to only two or three runs. This greatly reduces the scope for design optimization.

8.1 Noise modeling examples

Figure 5 shows a very simple FEM of the stamped steel oil pan that produced the results shown in Figures 3 and 4. Figure 6 shows the vibration response of this model at 240 Hz and the calculated radiated sound power of the oil pan at 240 Hz. Figure 7 compares the results of the analytical model to measured sound power data obtained by using the acoustic intensity method. Despite the very simple structural model used, excellent agreement between the analysis and experiment was obtained. In fact, at frequency where some discrepancies can be found, it is likely that the discrepancy is caused by experimental error.

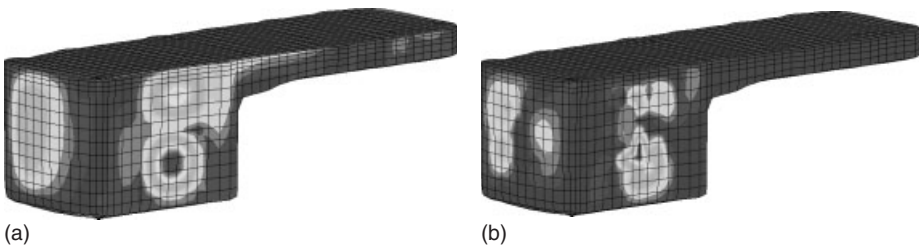


Figure 6. Vibration response (a) and radiated sound power (b) at 240 Hz.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

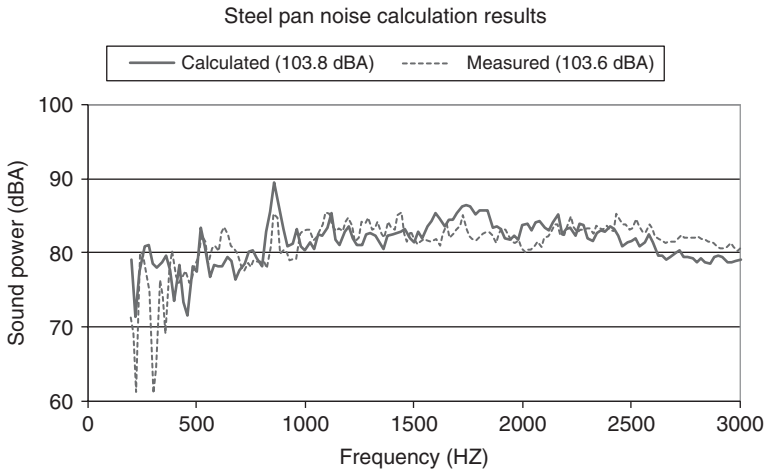


Figure 7. Comparison of calculated and measured sound power for the oil pan shown in Figures 5 and 6.

9 LOW-NOISE ENGINE DESIGN GUIDELINES

The engine block structure should be as stiff as possible. One way to achieve this is to split the block at the crankshaft centerline. The lower section, called the *bedplate*, carries the main bearing caps. A low-cost way of achieving high block stiffness is to use a long skirt block with a simple flat plate tying the oil pan flanges together. A flat plate is very stiff in shear, and when the block tries to bend or twist, it will have to deflect the plate in shear. A long skirt is needed with this approach, as the stiffening plate needs to be below the main caps so that the plate does not need holes to accommodate the caps. The plate will still need holes to accommodate the rod envelope.

The stiffness of main bearing bulkheads is particularly important. During firing, the crankshaft deflects and applies forces to the bulkheads that act along the crank axis. As the bulkhead deflects, it causes deflection in the skirt, which will directly radiate noise as well as pass vibration into the oil pan. If the natural frequency of the bulkhead can be increased as much as possible in the fore–aft direction, this will reduce deflection and thus noise. The natural frequency target for heavy-duty engine bulkheads should be 800 Hz. Smaller engines need a higher target value.

Another alternative for stiffening the bulkheads is to tie the main bearing caps together. This should be done without tying the caps to the block skirt. Cross-bolting of main caps, which ties them directly to the block skirt, tends to be counterproductive from a noise point of view. In highly loaded V-type engines, it may be required for structural integrity, however.

Large, flat cast panels should be avoided. These tend to be lightly damped, and have resonances within the range where significant forcing functions are present. Adding shape is a good way to add stiffness to a cast panel without adding too much weight.

The mounting flange for the flywheel housing or transmission bell housing should be as tall, wide, and stiff as possible. This helps ensure a high powertrain bending natural frequency. A good target for powertrain bending natural frequency is equal to high idle rpm divided by 24, for a four-cylinder engine. This puts the resonance at 2.5 order at top speed, providing margin to the large second-order force input.

Large engine covers should be designed with a low first natural frequency of 100–150 Hz. If this is not feasible, the large cover should be stiffened and isolated. Smaller covers should be designed for a first resonance of 5 kHz or more. If this is not feasible, isolation should be considered.

If the engine is likely to be used in noise-sensitive applications, it is a good practice to design in space to allow for close-fitting noise enclosures.

Combustion noise targets should be set early in the design phase. Achieving a good combination of performance, emissions, and noise requires close collaboration between the engineers responsible for these often competing characteristics.

For diesel engines, the selection of the fuel system and how it is driven has a huge effect on the noise of the engine. A low cyclic torque high pressure pump will reduce torsional inputs to the timing drive, helping to reduce noise. This is especially important in engines that use gears for the timing drive. Flexibility on the number of injection events, injection timing, and injection pressure are keys to controlling combustion noise.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Engines that use gears for the timing drive should focus on minimizing the size of the gear train and the torsional inputs to the gear train. This suggests, for example, putting the cam in the block rather than in the head, and selecting a low cyclic torque fuel system. More gear meshes provide more opportunity for gear rattle, and thus more noise.

REFERENCES

Brüel & Kjær, Inc. (1998) Human Vibration, Publication BR0456. www.cyut.edu.tw/~hcchen/downdata/human%20vibration.doc.

Reinhart, T.E. (2009) Diesel combustion mode switching—a substantial NVH challenge. SAE Paper 2009-01-2080.

Reinhart, T.E. and Wahl, T.J. (1997) *Characteristics of compression brake noise*. 5th International Congress on Sound & Vibration, December 1997, pp. 2643–2650. See also Reinhart, T.E. and Wahl, T.J. (1997) Developing a test procedure for compression brake noise. SAE Paper 972038.

Robert Bosch GmbH (2000) *Bosch Automotive Handbook*, 5th edn, Robert Bosch GmbH, Stuttgart, pp. 399–403.

Zhao, H. and Reinhart, T.E. (1999) The influence of diesel engine architecture on noise levels. SAE Paper 1999-01-1747.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto137
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Exhaust Emission Control Considerations for Diesel Engines

Timothy V. Johnson

Corning Incorporated, Corning, NY, USA

1 Introduction 1
2 Diesel Particulate Filters (DPF) 1
3 Lean NOx Control Technologies 5
4 Diesel Oxidation Catalysts (DOC) 9
5 Emission Control System Design 11
6 Summary/Conclusions 14
References 15

1 INTRODUCTION

Advanced emission control technologies have been utilized to clean engine exhaust since 1999 for light-duty (LD) diesels and since 2005 for heavy-duty (HD) diesels. Criteria pollutant tailpipe regulations are tightening quite significantly, and all modern diesel engines now have some form of exhaust emissions control.

This chapter focuses on key developments related to emissions and technologies for diesel engines in the automotive and HD markets: diesel PM (particulate matter) control, lean NO_x control, diesel oxidation catalysts (DOCs), and comments on system design.

This chapter is not intended to be all-encompassing and comprehensive. Representative papers and presentations were chosen here that provide examples of new, key developments and direction.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 DIESEL PARTICULATE FILTERS (DPF)

Diesel particulate filters (DPFs) are the most effective means of reducing PM from diesel exhaust. They can remove more than 99% of the hazardous soot particles. Although DPFs have been in commercial production for first-fit applications since 1999, there is still much optimization activity in the field. Work is continuing on DPF substrate design, DPF regeneration, and ash characteristics and management.

2.1 Substrate characteristics

Boger *et al*. (2011a, b) reported on a new substrate for reducing back pressure in DPFs. They tightened the pore size distribution and decreased porosity in the next-generation aluminum titanate filter to provide either a 2–3 g/L increase in soot mass limit or a 20–25% back pressure reduction, depending on cell geometry. Catalyzed samples of the low pressure-drop version has 20–30% lower back pressure with no soot on the filter and 15–20% lower back pressure with 6 g/L soot loading (Figure 1). The soot mass limit was similar to that of an SiC (silicon carbide) filter with the same cell geometry, but the SiC version has 50% higher back pressure. As a result of lower thermal conductivity, the regeneration efficiency of the new filter in a standard drop-to-idle test at 575°C is 6% higher than the earlier version, and 16% higher than the SiC comparison.

To facilitate the addition of deNO_x catalyst to the DPF, Warkins *et al*. (2011) increased the porosity in a new aluminum titanate high porosity (AT-HP) filter. Figure 2 shows that with a heavy catalyst coating and high soot loading, the pressure drop of the new DPF is 25%

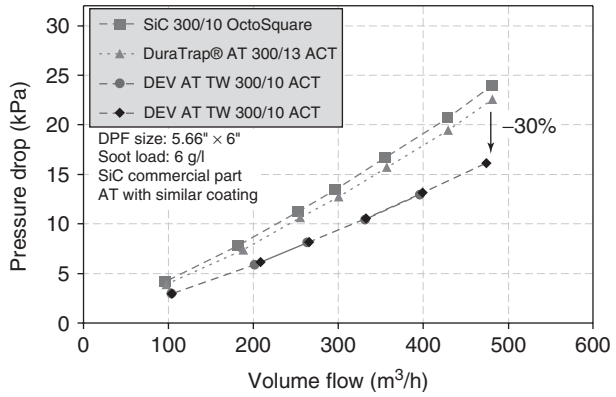


Figure 1. Pressure drop comparisons for the next generation aluminum titanate filter (DEV AT), in the low pressure-drop thinwall (TW) version. 300/10 refers to 300 cells/inch^2 with 10 mil wall thickness (Boger *et al*., 2011a, b). (From Boger *et al*., 2011. Copyright © 2011 SAE International. Reprinted with permission.)

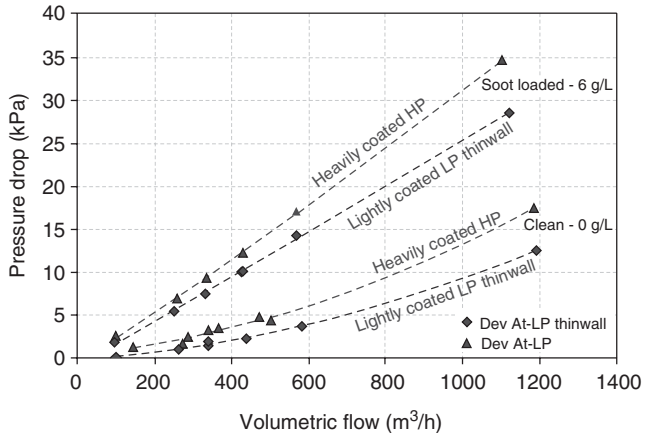


Figure 2. A new AT high porosity filter (AT-HP) with high catalyst loading has slightly higher back pressure than a lightly coated LP low back pressure filter. However, the overall system back pressure is reduced 20% if a separate SCR catalyst is eliminated (Warkins *et al*., 2011). (Reproduced with permission from Warkins *et al*., 2011. © J. Warkins *et al*., Corning, Inc.)

higher than a lightly coated low porosity (LP) thinwall version designed for low pressure drop (−25% vs current). However, if the AT-HP filter does not require a separate SCR (selective catalytic reduction) catalyst, overall system back pressure is reduced upwards of 20%. The new AT-HP filter also has 2 g/L higher soot mass limit than the current commercialized AT filter. Cycle-averaged deNO$_x$ efficiency with SCR catalyst on the HP filter was 62% (urea injection starts at 400 s), versus 23% (injection at 1100 s), when a separate SCR catalyst is placed behind the uncoated DPF.

Taking advantage of improvements in SCR technology, future HD engines will be calibrated to higher NO$_x$ and lower PM to save fuel. This will result in favorable conditions for passive oxidation of soot by NO$_2$ and dramatically decrease the need for active regeneration of the DPF at high soot loads. Less thermal mass will be needed in the DPF to provide a buffer against uncontrolled active regenerations. Boger *et al*. (2011b) reported on the next-generation thinwall cordierite filter to address this trend. Relative to the current offering, the pore size distribution was tightened and made nominally smaller, and the porosity was increased to ∼55%. Wall thickness was reduced 33% in the 200-csi (cells/square inch) geometry. To enable this, the inherent strength of the cordierite was increased. As with membrane technology, this redesigned porosity allows little, if any, soot penetration into the wall that causes rapid build-up of back pressure. Moreover, coated and uncoated filters have little back pressure differences. The result is that soot-laden filters have 40–50% lower back pressure than their 2010 predecessors under a variety of conditions. Interestingly, because of the reduced thermal mass, skin temperatures are higher but centerline temperatures are the same during active regeneration, reducing the thermal stress in the part. Although the authors made no mention of soot mass limit impacts, the filter survives worst-case drop-to-idle testing at 3.5 g/L soot. The lower thermal mass of the DPF allows faster heat-up of a downstream SCR catalyst, resulting in 10% more time for urea injection in the US certification test cycle. This can result in 15% lower cumulative NO$_x$ emissions in the cold-start test (Heibel, 2010).

Inorganic membranes can be added to filter surfaces to enhance filtration and reduce back pressure (Iwasaki *et al*., 2011). In vehicle testing, pressure drop was reduced 30–40% depending on speed and soot load, relative to the same filter without a membrane. This membrane benefit was also demonstrated on SCR-coated DPFs in engine dynamometer testing. Alternatively, in engine tests, the investigators demonstrated that the membrane can be used to increase the soot mass limit of a cordierite DPF about 2 g/L without a back pressure penalty by applying it to a lower porosity substrate.

When high catalyst loadings are added to DPFs, such as when applying deNO$_x$ catalysts, particle number (PN) emissions can counter-intuitively increase if excessive catalyst is applied (Kuwajima *et al*., 2009). The effect is illustrated in Figure 3. PN emissions are correlated to porosity: emissions go down as 10–20 μm porosity increases, but they go up if pores >30 μm increase in volume fraction. As coating load increases, large pore volume goes down but 10–20 μm pore volume remains flat. The net result is decreased PN emission. However, if coating load increases further, the opposite occurs: coating decreases 10–20 μm porosity but

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

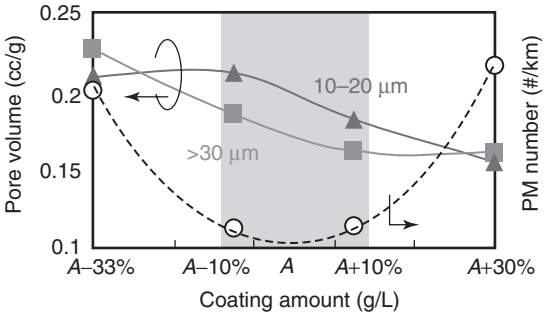


Figure 3. PN emissions (dashed line) reach a minimum with catalyst coating amount, because of the counteracting impact of loading on 10–20 μm and >30 μm pore volume (Kuwajima *et al*., 2009). (From Kuwajima *et al*., 2009. Copyright © 2009 SAE International. Reprinted with permission.)

large pores remain largely unaffected, giving the net result of increased PN emission.

Of a more fundamental nature, Rakovec, Viswanathan, and Foster (2011) used wafers machined out of commercial DPFs to look at how filter wall permeability varies with gas flow rate, PM loading, and PM type. Most of the wall permeability impacts are due to the substrate. Normalized wall permeability starts out similarly for high face velocity, independent of PN levels in the exhaust. But later, exhaust with low PN levels produces higher permeability because of more particles going into wall and forming a thinner cake. Low flow, high PN allows low cake density and higher ending permeability, but with a thicker cake. High nucleation-mode particle loading allows particle penetration into the wall and early bridging, resulting in a fast drop in permeability and low final permeability. The results should be useful in refining filter back pressure models that are used to manage filter regeneration.

2.2 DPF regeneration methods

More than 10 years ago, in the first wide-scale application of DPFs for particulate control on LD diesels, Peugeot chose a ceria-based fuel borne catalyst (FBC) to facilitate the regeneration of the DPF. A new generation of FBC is based on iron, and further improves DPF regeneration characteristics with or without platinum group metals (PGM) catalyst on the DPF (Rocher *et al*., 2011). Compared to the original of 30 ppm Ce and 10 ppm Ce/Fe in the previous version, the new formulation uses only 5 ppm Fe with similar performance, resulting in half the ash load on the DPF. The authors estimate that for a car with a fuel consumption of 7 L/100 km (33 miles/gallon), the DPF ash cleaning interval is 300,000–400,000 km, depending on filter design. The new FBC drops the DPF regenerating start temperature of a stock PGM-catalyzed DPF (CSF, catalyzed soot filter) from 410 to 360°C, and increases the total soot burn from 12% in the baseline ramp-up test (to 500°C) to 75% with the FBC–CSF combination. The improved regeneration efficiency and decreased temperature will reduce thermal exposure of the SCR catalyst in Euro 6 systems, as well as reduce DPF regeneration fuel penalty when the SCR system is located upstream of the DPF.

Warner, Dobson, and Cavataio (2010) investigated the current DPF regeneration dynamics. Active regeneration efficiency, wherein exhaust temperature is increased to ~600°C and the soot is burned by oxygen, is not strongly dependent on oxygen content at levels >2% nor on whether the filter contains precious metal, although the precious metal does oxidize the resultant CO to form CO_2. However, the efficiency is strongly dependent on soot loading because of the build-up of heat. Passive regeneration, wherein the soot is oxidized by NO_2, is much more effective (>3×) at 370°C than at 485°C, as the decomposition of NO_2 back to NO at the higher temperatures overwhelms the faster soot oxidation rate at these temperatures. A DPF with Cu-zeolite behaves similarly to the uncoated filter, and has minimal impact on DPF regeneration. This indicates that NO_2 prefers to oxidize soot rather than be reduced on the zeolite. HNCO (isocyanic acid) is a byproduct of active regeneration without catalyst, and needs to be considered in the mass balance when examining regeneration effectiveness. Active regeneration "costs" about 2–3% fuel consumption, and passive regeneration strategies can drop this penalty by about 20%.

Soeger *et al*. (2005) analyzed the balance between exhaust flow rate, temperature, and soot burning kinetics, as they relate to passive regeneration. The amount of NO_2 available for soot oxidation depends on engine out NO_x, NO_2 conversion, and flow rate (RPM, engine revolutions per minute). High RPM with 350°C exhaust gives highest NO_2 production. However, at temperatures <380°C, the burn rate is controlled by soot oxidation kinetics, so, NO_2 availability generally does not limit soot oxidation rate with properly designed catalyst systems.

Direct oxidation catalysts have been of interest in the field for more than 5 years. These catalysts use oxygen conducting materials (such as ceria, zirconia, or manganate) to burn the soot at the soot–catalyst interface, rather than by oxygen in the gas phase. Southward, Basso, and Pfeifer (2010) showed that a complex ceria material can begin oxidizing soot with model gas at 160°C with completion at 220°C using no or very little precious metal. The exotherm from CO or hydrocarbon (HC) oxidation initiates the reaction. Once started, the exotherm causes soot not in contact with the catalyst to oxidize via the gas phase oxygen. When tested using vehicle exhaust, the balance

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

point temperature (BPT; soot accumulation rate is the same as the oxidation rate) is 20°C lower than with a commercial filter coated with 0.2 g/L of 1 : 1 platinum and palladium (BPT = 420°C). The regeneration efficiency is much better than for a lightly catalyzed filter at 550°C, but similar at 620°C. Iretskaya *et al*. (2010) showed that a rare-earth base metal oxide catalyst has a BPT of 350°C in the absence of NO_2.

Soot can also be burned by adsorbed oxygen on the surface of an SiC fine-particle membrane. Karin and Hanamura (2010) showed much lower activation energy for a DPF with the SiC membrane compared to one without (80 kJ/mol vs 130 kJ/mol), indicating a shift in reaction mechanism. In the paper, they showed anecdotal evidence of the surface oxidation phenomenon, but the presentation had new data confirming this mechanism using thermal desorption spectroscopy. In another study, it was shown that membranes cause the soot to deposit as a layer rather than being dispersed through the porous wall of the filter (Mizutani *et al*., 2010). This creates a more localized exotherm that improves regeneration efficiency upwards of 10–15%.

Biodiesel can impact soot loading and DPF regeneration properties. Vertin, He, and Heibel (2009) conducted a comprehensive investigation of B20 impacts on cordierite DPFs. They blended soy-based methyl ester biodiesel with ultra-low sulfur diesel fuel, and ran dynamometer tests to generate results. PM emissions were reduced 20% with B20 in transient tests, but were similar in steady state tests, indicating that PM differences are cycle dependent. Back pressure relationships are the same for B20 as for standard fuels. Regarding regeneration, at 300°C, B20 does not burn as effectively in the DOC, requiring more fuel to regenerate the DPF. The excess fuel likely collects on the soot, resulting in a larger exotherm during uncontrolled regenerations. At temperatures >300°C, the fuel burns better. There is a minor improvement in passive NO_2 regeneration with B20. No deterioration in catalyst performance was observed after ~120 active regenerations. Soot formed from burning fuel containing 20% biodiesel burns 3× faster than soot from fuel without (Austin *et al*., 2010). The increased reactivity is due to increased particle surface area (Strzelec *et al*., 2010).

2.3 Ash behavior and management

MECA (Manufacturers of Emission Controls Association) published a paper on the state of ash management (MECA, 2005). The reader is directed there for a much more comprehensive review of the topic than can be covered here. The report covers sources and composition of ash, filter designs, filter maintenance intervals, and cleaning procedures.

New filter designs enable extended ash clean intervals. Young *et al*. (2004) showed how increasing the inlet cell size of DPFs relative to the exit cell can increase the ash storage capacity by 50%.

Fujii (2011) looked at how DPF cell geometry and porosity affect filtration efficiency and back pressure as ash (from lube oil and wear) is collected in the filter. He studied filters with 200- and 300-csi and 12 mil (0.3 mm) wall thickness, containing 50% porosity with 15 μm average pore size and 65% porosity with 20 μm pore size. As early ash loadings prevent soot from penetrating into the wall, the lowest back pressure with high filtration efficiency appears with an ash loading of 4–10 g/L. Low back pressure sensitivity to soot and ash loading depends on DPF designs and materials: larger open frontal area gives lower sensitivity, as does higher porosity. Interestingly, filters that were continuously regenerated (like with NO_2) achieve high filtration efficiency (>97%) only after 20 g/L ash is accumulated. Filters managed by periodic regeneration need 140 g/L ash to achieve the same level of efficiency, because the ash generally collects in the back of the filter without forming much of a filtration membrane.

Sappok and Wong (2009a, b) showed in more detailed analyses that an ash membrane is formed before it migrates to the end of the DPF cell to form a plug at the exit side of the DPF. Zinc has a lower packing density than calcium or CJ-4 lube oil mixed ash, so it forms a membrane first (Sappok and Wong, 2009b). Back pressure is generally linear with ash and soot accumulation, as shown by region I in Figure 4. However, after about 12.5 g/L of ash accumulates (~50,000 miles), back pressure rapidly increases with

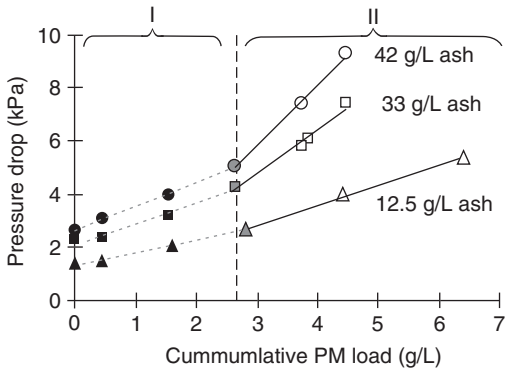


Figure 4. Pressure drop increases faster with increasing soot and ash loading beyond a certain threshold (region II) because of compressing of the membrane from higher back pressure (Sappok and Wong, 2009b). (Reproduced from Sappok and Wong, 2009b. © ASME.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

increasing soot beyond about 3 g/L, zone II. This effect is thought to be caused by the compression of the soot/ash membrane caused by increasing back pressure.

2.4 PM sensors for on-board diagnostics (OBD)

PM on-board diagnostic (OBD) regulations are expected to require a post-DPF soot sensor. Five general types have been reported. The most common type collects soot between two electrodes and measures changes in electrical conductivity (Weigl, Roduner, and Lauer, 2010). It is periodically heated to remove soot, the frequency of which indicates PM level over the period. Resolution is acceptable, but the early 2009 version could not detect DPF failure within the timeframe of the new European drive cycle (NEDC) test, as the regulation requires (Finch, Hnilicka, and Sindano, 2010). The second type also collects soot, but between two parallel plate electrodes (Kondo *et al*., 2010). The change in capacitance indicates the amount of soot that is collected, and hence, the soot PM level in the exhaust. The soot is also periodically burned out. The investigators show acceptable resolution and the ability to detect a failure within the NEDC test. The third type is a real-time sensor and measures soot carrying an electrical charge as it passes between two electrodes (Besch *et al*., 2010). This is the only type that does not accumulate soot and is apparently capable of measuring PN and PM. The fourth type is different from the others, as it is not based on electrical properties of soot (Zidat, 2010). A slip stream contains a small monitoring DPF that begins filling with leaked soot. This results in reduced flow in the slip stream, which is detected with a thermocouple. The device has the required resolution and has potential to take a reading within the NEDC. The fifth device uses radio frequency to measure soot loading on the filter.

3 LEAN NO_X CONTROL TECHNOLOGIES

Lean NO_x control (lean $deNO_x$) technologies will be integral to meet the emerging criteria pollutant regulations for diesel engines. Minimum removal efficiencies on the order of 85% will be needed, but levels up to 97–98% are desired to allow HD engines to operate in high NO_x low fuel consumption regimes. For LD applications, the efficiency is as important, but light-off or low temperature (LT) performance characteristics are even more so.

Two broad approaches to lean $deNO_x$ control are SCR using ammonia and HC-based approaches (HC-$deNO_x$) primarily using lean NO_x traps, but also lean NO_x catalysts (LNCs) (or HC-SCR).

3.1 Selective catalytic reduction

SCR technology is entering its third or fourth generation since commercial introduction in Europe in 2003. Then, systems were removing upwards of 75% NO_x over the European HD transient cycle to meet Euro IV regulations. To meet the emerging Euro VI regulations in 2013, cycle-average $deNO_x$ efficiencies approaching 95% may be realized. Work is continuing in the United States to go even higher in efficiency to meet the current and emerging LD NO_x regulations.

Currently, good SCR performance is limited by urea injection issues (evaporation and hydrolysis) at temperatures $<200^{\circ}$C. Improved mixers allow urea injections at temperatures as low as 180°C and thus drops NO_x ~30% over the United States cold HD transient cycle relative to no mixer (Zhan *et al*., 2010). Alano *et al*. (2011) described a compact mixer that needs only 75 mm of urea mixing length, compared to 350 mm in some commercial LD SCR systems, enabling the SCR catalyst to be placed closer to the engine for faster heat-up. The mixer achieves a urea mixing index of 0.95 (all cross-section NH_3 measurements are within 5% of one another) over a range of gas flows, with a maximum increase in back pressure of 0.4 kPa (4 mbar) during accelerations relative to a conventional system. In the closer position, in tests, the SCR catalyst was up to 25°C hotter and achieved 67% $deNO_x$ efficiency on the NEDC versus 37% for a catalyst place further back. The mixer could be useful if SCR catalyst is placed on a DPF for faster light-off or better DPF regeneration versus two separate systems (DPF–SCR or SCR–DPF).

To accomplish the same objective, urea hydrolysis catalysts are emerging. Kröcher *et al*. (2010) showed that upwards of eight different decomposition products are emitted from urea upon heating, but with a titania decomposition catalyst, as shown in Figure 5, ammonia is produced at temperatures as low as 150–160°C in model gas with no other unexpected decomposition products.

Although the urea infrastructure is well developed in Europe, Japan, and the United States, finding alternative sources for ammonia is still of significant interest to enable SCR catalysts to function better at low exhaust temperatures, decrease the size and cost of the system, and enable the use of the system at very low ambient temperatures. Johannessen (2011) updated the developments on a gaseous ammonia system using chloride-based adsorbents. Both HD and LD systems were described, showing 100× dosing ranges within 5% accuracy and <1.5% deviation in set-point under a range of exhaust conditions. Start-up units are used that initially draw 550 W in HD and 250 W in LD applications, but go down to 100 W range during normal operation. Safety and durability issues

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

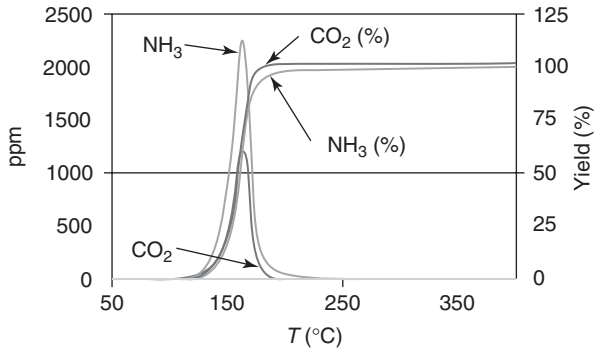


Figure 5. Using a titania decomposition catalyst, urea forms only ammonia at 150–160°C, with no undesirable byproducts in model gas experiments (Kröcher *et al*., 2010). (Reproduced with permission of AVL and the authors from Proceedings of the 6th International Exhaust Gas and Particulate Forum. See Kröcher *et al*., 2010.)

appear addressed, and system optimization through testing and simulation is continuing. Jackson (2011) described an alternative approach that utilizes ammonium carbamate (NH_2COONH_4). It is available as pellets that release ammonia upon heating with auxiliary hot water. Ammonia salts dissolve in the water and depress the freezing point to −30°C. Development issues include faster start-up and better, more-efficient heating. Thomas and Highfield (2011) described some early performance data with an ammonium formate and urea system containing 54% water, versus 67.5% for standard urea solutions. Advantages include reduced freezing point (−30°C), better high temperature (HT) storage stability, lower hydrolysis temperature, no polymerization like with urea (fewer or no deposits), and they demonstrated full "drop-in" capability in a urea system on a new diesel pick-up truck with and without SCR system.

SCR catalyst formulations and design are improving both low and high temperature performance, as well as sensitivity to HC and sulfur poisoning. Vanadia SCR catalysts are used in Europe and emerging markets, but not in the United States or Japan because of durability issues related to thermal exposure when DPFs are regenerated. Advances are now reported (Chapman *et al*., 2010) on vanadia SCR catalysts that have no volatility up to 750°C or higher, versus 550–600°C for some commercial catalysts, giving them similar HT durability to zeolites. $DeNO_x$ performance at 250–350°C is 5–10 points better after HT aging (>700°C) than for a benchmarked commercial catalyst, but less-severely aged catalysts have lower efficiencies than the base catalyst. Walker (2010) reported that new Cu-zeolite formulations now sustain aging to 900°C and form less N_2O.

Reichert (2011) and Narula *et al*. (2011) showed some new advancement in the zeolite SCR catalyst activity. Reichert showed a new zeolite material that exhibits the same LT performance of copper-zeolites and the same HT performance of iron zeolites. Results are shown in Figure 6. Narula showed that it is possible to modify zeolite structures systematically to influence the electron density at metal

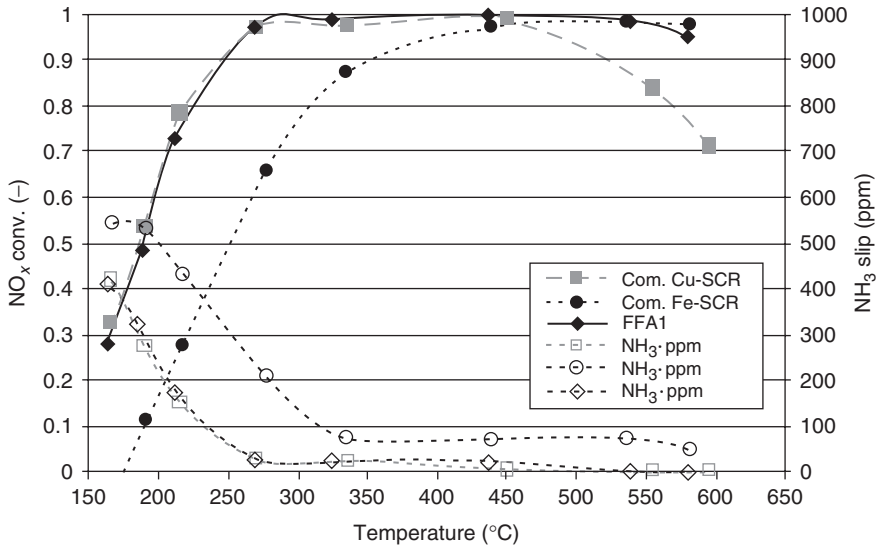


Figure 6. Performance characteristics of Cu- and Fe-zeolite SCR catalysts. "FFA1" is a new Cu-zeolite that has the same low temperature performance of standard Cu-zeolites, but with comparable high temperature performance to Fe-zeolites (Reichert, 2011). (Reproduced by permission of the Car Training Institute. © Martina Reichert.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

centers and to provide ammonia bonding sites in the vicinity of the metal centers. They replaced alumina in the structure with several trivalent cations. In another contribution, Yang and Narula (2010) showed that chemical mixtures of copper and iron zeolites can improve LT performance over than copper alone, and when lanthanum is added to the binary formulation, performance is improved further.

Han *et al*. (2011) showed that LT performance and reduced HC effects can be achieved if a ceria oxygen storage catalyst is layered on top of an iron zeolite catalyst. The catalyst helps urea decomposition, thus improving the LT deNO_x capability from 32% to 58% at 200°C, an LD steady-state test. After 8 h of exposure to high HC levels from a burner, the layered catalyst maintained a deNO_x efficiency of 80% at 240°C, whereas the original version was only at 60% under the same conditions because of HC poisoning.

On sulfur poisoning, Tang, Huang, and Kumar (2011) used SO_2 levels equal to those obtained with ultra-low sulfur diesel fuel (<15 ppm sulfur), the copper zeolite catalyst started losing deNO_x efficiency after about 400 h of operation at temperatures of 200–300°C. Through 1300 h of operation, the catalyst had deteriorated continuously from 98% deNO_x efficiency to 60% efficiency. Moreover, the $NO_2 : NO_x$ ratio from the filter deteriorated from 0.60 to 0.30 during the first 600 h, but then remained the same. They found that most of the sulfur was in the top layer of washcoat in the first third of the catalyst. Most of the poisoning was attributed to ammonium sulfate, which comes off at 400–500°C, and to a much lesser extent, copper sulfate, which comes off at 500–850°C. When heated to 500°C, the SCR catalyst performance recovered, and this was done every 700 h of operation at the lower temperatures.

For US LD diesels, removing cold-start NO_x emissions is key to meet the tailpipe emissions regulations. A new combination NO_x adsorber and SCR catalyst configuration was shown by Henry *et al*. (2011a). Figure 7 shows some performance characteristics. The system consists of an upstream passive NO_x adsorber (PNA) that might capture 65% of the NO_x at temperatures <150°C, and then passively releases it at temperatures greater the 150°C. At these temperatures, a copper zeolite is just becoming active and can reduce some of this released NO_x.

Nowadays, SCR substrates are generally have 300 or 400-csi. Heibel (2010) showed that in the mass transfer controlled regime (230–350°C), 600-csi substrates react 35% faster than 400-csi catalysts.

Work is continuing on the combined SCR + DPF system, wherein SCR catalyst is coated onto the DPF. This allows SCR catalyst to be placed on the vehicle without using an added component, and can get the SCR catalyst closer

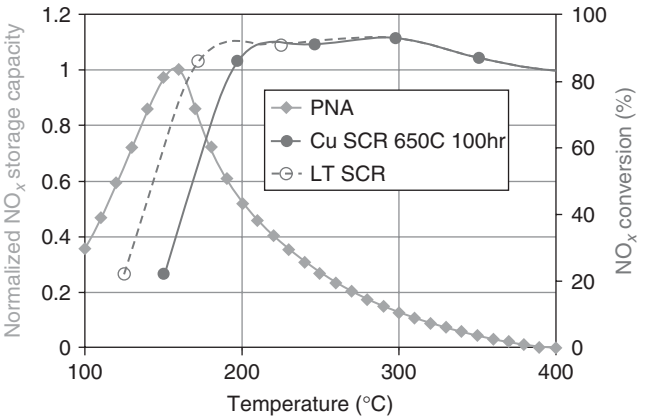


Figure 7. An upstream passive NO_x adsorber (PNA) captures NO_x generated at $T < 150$°C. An LT urea SCR catalyst can then convert this NO_x upon release at $T > 150$°C (Henry *et al*., 2011a). (Reproduced from Henry *et al*., 2011a.)

to the engine for faster light-off. Numerous reports dating to 2008 showed that total NO_x removal efficiency is thus improved, with little compromise in DPF regeneration. Tan, Solbrig, and Schmieg (2011) showed a new issue when soot is accumulated on the DPF + SCR: ammonia storage capacity decreases for fresh samples at all temperatures and soot loadings tested (200–350°C, 1.0–2.5 g/L), but is not affected by soot loading for aged samples (except at 200°C). Loss of ammonia storage capacity impacts SCR performance at 200°C, but not at 300°C at a soot load of 2 g/L. The researchers also showed that DPF regeneration calibration needs to be adjusted to longer times or higher temperatures to get the same cleaning performance as the base DPF system.

The US Environmental Protection Agency (EPA) capped nitrous oxide (N_2O) emissions in the HD greenhouse gas (GHG) rule, and is proposing a cap in the LD GHG rule. Kamasamudram, Henry, and Yezerets (2011), showed that N_2O is very stable, and forms by three mechanisms in an SCR catalyst:

LT ($T < 250$°C) decomposition of ammonium nitrate by the reaction

$$NH_4NO_3 \rightarrow N_2O + 2H_2O \quad (1)$$

HT oxidation of ammonia by copper zeolites by the reaction

$$2NH_3 + 2O_2 \rightarrow N_2O + 3H_2O \quad (2)$$

Reaction of excess NO_2 (>50% of NO_x) to form ammonium nitrate by the reaction

$$2NH_3 + 2NO_2 \rightarrow NH_4NO_3 \downarrow + N_2 + H_2O \quad (3)$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

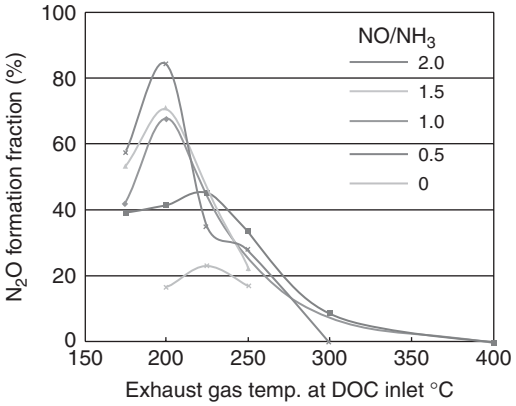


Figure 8. Nitrous oxide formation in ammonia slip catalysts is promoted by high NO : NH_3 ratios coming out of the SCR catalyst (Matsui *et al*., 2011). (Reproduced with permission from Matsui *et al*., 2011. © Society of Automotive Engineers Japan.)

Ammonium nitrate then decomposes by reaction 1.

SCR catalyst improvements can decrease N_2O formation by the first two mechanisms, and better DOC design and control can prevent the third mechanism. Kamasamudram, Henry, and Yezerets (2011) showed that it is possible to reduce N_2O to nitrogen, but these reactions occur at much higher temperatures than those at which they are formed.

For high efficiency SCR deNO_x, excess urea injection is needed, perhaps up to 20% more. Ammonia slip catalysts are needed to prevent ammonia release, but these catalysts can also form N_2O. Matsui *et al*. (2011) showed (Figure 8) that ammonia up to 80% going into the slip catalyst can convert to N_2O if there is also a relatively high amount of NO (2× vs NH_3); the reaction is

$$4NH_3 + 4NO + 3O_2 \rightarrow 4N_2O + 6H_2O \qquad (4)$$

A high NO : NH_3 ratio coming out of the SCR catalyst can occur, for example, if there is poor urea mixing before entering the SCR catalyst and urea is injected at less than stoichiometric requirements, or if ammonia is partially oxidized (by Cu-zeolite, for example) to form NO. Kamasamudram, Henry, and Yezerets (2011) showed that slip catalysts with lower precious metal content minimize N_2O formation.

3.2 HC-based NO_x control

3.2.1 Lean NO_x traps (LNT)

The lean NO_x trap (LNT) is currently the leading deNO_x concept for the smaller lean-burn (diesel and direct-injection gasoline) passenger cars, and is of interest in applications with limited space or in which urea usage is difficult. The deNO_x efficiency is nominally 70%, much lower than that of the next-generation SCR system at 95%, and the precious metal usage is high (∼10–12 g for a 2-L engine). As a result, efforts are focused on improving efficiency while reducing precious metal usage. One of the leading concepts is to use the LNT to generate ammonia during the periodic rich regeneration part of the cycle, and then to store and use this ammonia in a downstream SCR catalyst.

The first LD diesel sold in the United States to meet recent tailpipe regulations had the BlueTec™ 1 emission control system, utilizing an LNT followed by an SCR catalyst. The unique system used the rich cycle of the LNT to generate ammonia, which was captured and used by the downstream SCR for lean NO_x reduction. Weibel *et al*. (2009) reported that ammonia selectivity increases with aging and rich period, and decreases with increasing the air/fuel ratio (λ). Under conditions of $\lambda = 0.88$ and rich periods of 5 s (180 s lean), ammonia selectivity is >70% in the temperature range of 225–350°C for all aging tests temperatures greater that 600°C and 50,000 miles. The SCR adds about 20% deNO_x efficiency over an LNT only configuration.

Theis, Dearth, and McCabe (2011) reported on an interesting study whereby they alternated LNT and SCR slices in one can to check the effect of NO_x, ammonia, and HC distributions on deNO_x performance. The system performance improved as the number of alternating slices of the LNT and SCR increased, keeping the total volume constant. The deNO_x efficiency for the eight-segment system (four pairs of LNT and SCR catalysts) was 81% in a reference test at 275°C, whereas 78% for four segments and 60% for two segments. The reference single LNT with no SCR catalyst had only 30% deNO_x efficiency. The authors also show reduced N_2O, NH_3, HC, and CO emissions with the segmented systems. Various dynamics are operative, but the segmented systems tend to better-match the NO and ammonia concentrations in the SCR, and alternating SCR slices better-adsorb HCs for enhanced utility.

Xu *et al*. (2011) reported vehicle and laboratory testing on a second-generation LNT + SCR system. The DOC + LNT + SCR + DPF system was installed on a prototype F-150 pick-up truck (2610 kg, 4.4 L V8, turbodiesel). The aged system (64 h, 750°C) reduced NO_x by 96% to 13.5 mg/mi, and HC emissions were 14 mg/mi (−99%), bringing the vehicle to within the emerging California LEVIII (low emission vehicle 3) limit values (30 mg/mi HC + NO_x) on the standard certification test cycle. The laboratory work focused on HC reductions from the system. The SCR component reduced HCs about 75%, mainly by adsorption under rich conditions and oxidation under lean conditions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Cavataio *et al*. (2011) compared this capability to that of a urea SCR system for meeting the US EPA Tier 2 Bin 2 (or California LEVIII fleet average) standards. Although the LNT + SCR system is 18% smaller, it had met the target emissions while the SCR system fell short. Further, the LNT + SCR system is estimated to be slightly cheaper, but has most of the cost tied up in precious metal (with its inherent price volatility). On the downside, the fuel penalty was high at 10%, versus 2% for the SCR system. Moreover, sulfur management of the LNT + SCR system was not considered.

LNTs are sensitive to sulfur poisoning, wherein a stable alkali-earth sulfate forms displacing the desired nitrate. They need to be periodically desulfated, even in low sulfur fuel. Chen *et al*. (2009) added PO_x (partial oxidation) catalyst to partially reduce HCs to hydrogen and CO. The result is a 150°C drop in desulfation temperature with double the amount of sulfur coming off. In addition to this benefit, LT performance slightly improves, and expensive rhodium could be replaced with cheaper palladium, reducing precious metal costs.

LNT durability has improved significantly over the years, mainly because of better materials and to the need for lower desulfation temperatures. Ottinger *et al*. (2009a) developed a rapid thermal aging test protocol, and then determined (Ottinger *et al*., 2009b) the thermal aging mechanisms of the three unit operations in an LNT: NO oxidation, NO_2 adsorption, and NO_2 reduction. NO oxidation is hampered by the loss of precious metal surface area. The adsorption of NO_2 is largely affected by the loss of alumina surface area, especially at the lower adsorption temperatures. As baria is relatively unaffected at the higher temperatures and alumina loses surface area at ~900°C, this overall effect is not as significant to the LNT function as the loss of precious metal oxidation kinetics. Aging actually helps the reduction unit process, because the larger baria grains release the nitrate slower, reducing NO_x slip.

In what might be a newly discovered reaction phenomenon, the temperature range of the LNT was extended from 350°C to well over 600°C by managing it differently. Some results are shown in Figure 9. Bisaiji *et al*. (2011) at Toyota oscillated the air–fuel ratio between 16 and 24 depending on conditions, but all within the lean regime using an auxiliary exhaust injector. They proposed a mechanism involving partially oxidized HC intermediaries (observed) reacting with chemisorbed nitrate species, in a type of HC-SCR reaction. The frequency of fuel injection is in 0.5 Hz range and the $deNO_x$ efficiency increases with amplitude of the oscillation, up to about three air : fuel ratio points. Fuel penalties are on the order of 1.5%–3.0% at medium to high load, and running at about 80% $deNO_x$ efficiency (Inoue *et al*., 2011). The

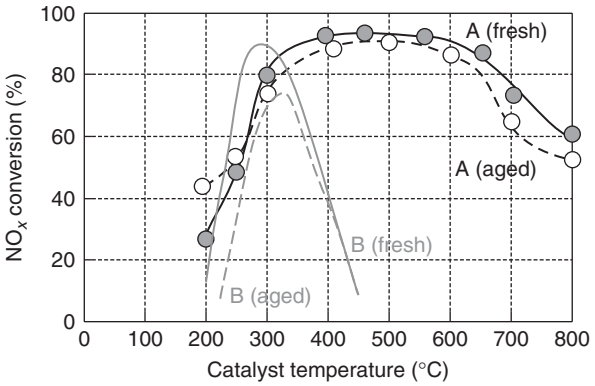


Figure 9. NO_x reduction curves for a standard LNT operation (B) and for the same LNT run using the new method (A) of lean air : fuel oscillations at ~0.5 Hz (Bisaiji *et al*., 2011). (From Bisaiji *et al*., 2011. Copyright © 2011 SAE International. Reprinted with permission)

method is sensitive to sulfur poisoning, but can take loadings 3.0× higher than a standard LNT before dropping off below 80% $deNO_x$ efficiency. At inlet NO_x levels of 100 ppm and temperatures of 370–420°C, the method delivers >80% $deNO_x$ efficiency at a space velocity (SV) of 125,000/h with a 2% fuel penalty. The new method is best operated at >50% load (lower loads at higher RPM), and complements standard LNT operation at the lower loads.

Finally, to wrap up the representative studies on $deNO_x$, Jackson (2011) updated the industry on using an LNC (silver-alumina) with E85 reductant (15% gasoline and 85% ethanol). Converse to urea, E85 does not freeze and leave deposits when injected at LTs. The system performs well (>90% efficiency at 350–450°C, $SV = 38{,}000^{-1}$) after 500 h of aging at 650°C and 100 h at 800°C. The catalyst was coated onto a DPF and demonstrated 60% $deNO_x$ efficiency at 350°C with a C : N level of 3 : 1. When E100 is used, reductant consumption is 25%–35% less than for urea SCR at similar levels of performance: >90% $deNO_x$ efficiency in the 275–375°C range ($SV = 38{,}000\,h^{-1}$).

4 DIESEL OXIDATION CATALYSTS (DOC)

DOCs play two primary roles in commercial emission control systems:

1. Oxidize HCs and CO, either to reduce emissions coming from the engine or to create exothermic heat used to regenerate a DPF.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2. Oxidize NO to NO_2, which is used for continuously oxidizing soot on a DPF, and/or for enhancing the SCR deNO_x reactions, particularly at LTs.

Henry *et al*. (2011b) looked at the interplay of these two functions using a series of iterative reaction-decoupling experiments to explain interactions between HC and NO oxidation. They showed that NO oxidation is inhibited on Pt/Pd because of the reduction reaction with NO_2 by HCs. Long-chain alkanes had a more adverse effect than short-chain alkenes because of slower oxidation rate with oxygen. Decreasing SV was shown to help NO_2 formation in the presence of HCs. Prestoring HCs on the DOC improved NO oxidation up to 300°C. The interplay of CO and HC removal and NO_2 generation takes another tack as well. Spurk *et al*. (2010) investigated NO_2 coming from a catalyzed DPF for use in a downstream SCR system. Surprisingly, they found the NO_2 coming out of the DOC and going into the DPF is not as important as the HCs coming from the DOC. Essentially, the HCs going into the DPF can interfere with the NO_2 formation in the DPF. The Pt/Pd ratio is much more important to NO_2 formation than precious metal loading on the DPF.

A critical HC and CO issues has largely gone unresolved. Premixed combustion strategies offer advantages in low load operating points because of significantly reduced NO_x and PM emissions. However, these advantages come at the price of greatly increased HC and CO emissions, lower exhaust temperature, and lower oxygen content, intuitively presenting a significant challenge to DOCs. Indeed, Sumiya *et al*. (2009) showed that the light-off temperature of DOCs increases with decreasing oxygen and increasing HC and CO. They enhanced a catalyst formulation using materials to supply oxygen, applying a CO adsorption suppressant, and creating a plurality of active sites for multiple functions. As a result, the light-off temperature (T50, temperature of 50% conversion) for aged catalysts dropped from 260 to 180°C for exhaust with 500 ppm HC, 500 ppm CO, and 2% oxygen. The values for 5000 ppm HC and CO levels were 225°C from 280°C previously.

Jen *et al*. (2009) showed that platinum can migrate from DOCs (or presumable DPFs) to SCR catalysts if they are exposed to temperature >670°C for extended periods of time (16 h). SCR efficiencies can decrease, especially if the DOC is exposed to temperatures >750°C, as even minute quantities of platinum can cause oxidation of ammonia. Later, Cavataio *et al*. (2009) showed that if palladium replaces some of the platinum in the DOC, less migration can occur. Although the 2 : 1 Pt : Pd mixture shows some deterioration in SCR deNO_x efficiency, it is much worse for the Pt-only formulation. Washcoat formulation and/or processing can make a difference, and NO, HC, or CO oxidation is unaffected or enhanced with the Pd additions.

Kim *et al*. (2011) did a systematic study on the effects of varying the Pt : Pd ratio on DOC HC and NO oxidation and durability in a variety of conditions. All bimetallic Pt–Pd catalysts show better HC light-off activity and thermal stability than the Pt- or Pd-only catalyst. NO oxidation to NO_2 was found to always depend directly on platinum content, with similar durability trends as with HCs. Figure 10 shows a schematic representation of these findings. They found that HC–CO mixtures synergistically have ~20°C lower light-off temperatures than either one alone.

Glover and coworkers (2011) also did a study on Pt : Pd effects on DOC properties, adding N_2O formation and looking more at fundamentals. CO plays a key role on the overall catalyst performance by its positive effect on propylene oxidation which, in turn, is responsible for NO reduction to N_2O and the onset of NO_2 formation. On the Pt : Pd = 4 : 1 catalyst, propylene partially reduces NO to form N_2O at about 200°C, but this temperature shifts to 250°C when is CO added. The effect of higher Pd concentration on NO conversion is detrimental for NO oxidation to NO_2, but is positive for producing less N_2O, especially at high oxygen concentrations. NO_x storage and release may play an important role in NO_2 formation over the lightly loaded full Pt DOC formulation studied. A 40 g/ft^3 (1.4 g/L) bimetal formulation (Pt : Pd = 4 : 1) is comparable on CO and HC light-off to a 113 g/ft^3 Pt formulation. Closing on N_2O formation, Kamasamudram, Henry, and Yezerets (2011) showed that propylene forms much more N_2O than dodecane ($C_{12}H_{26}$).

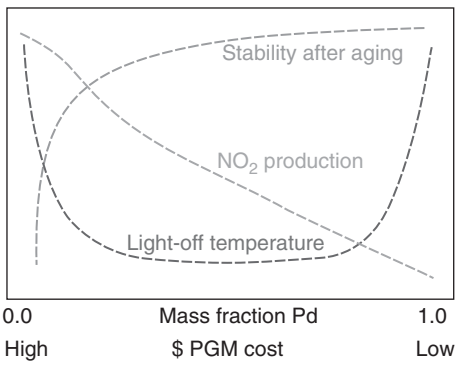


Figure 10. Conceptual impact of increasing Pd content at the expense of Pt in DOCs. Moderate substitutions improve durability and HC oxidation, without significant deterioration of NO oxidation (Kim *et al*., 2011). (From Kim *et al*., 2011. Copyright © 2011 SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Potential adverse effects of biodiesel ash on DOCs and other emissions control components were described by a large research group led by the National Renewable Energy Laboratory (Williams *et al*., 2011). The group reported that after a simulated 150,000 miles of durability testing, HC slip increased nominally 20–25% over the range of temperatures in steady-state tests (240–390°C) as a result of alkali exposure from the biodiesel ash. NO_2 formation declined from 35% to 20%. In addition, the thermal shock parameter of the DPF, as indicated by mechanical property measurements, declined 69% after simulated exposure of 435,000 miles (700,000 km), again because of alkali attack of the cordierite substrate. NO_x emissions from the SCR increased about 50%, but more work was needed to determine if this was due to alkali attack of the zeolite catalyst. The group concluded that operating with fuel at the maximum alkali ash specification will significantly deteriorate emission control system performance.

5 EMISSION CONTROL SYSTEM DESIGN

By 2017, most vehicular diesel engines in the United States, Japan, and Europe will have both DPF and SCR technologies. Exceptions are the smaller cars in Europe, some non-road applications in the United States and Europe, and some HD trucks in the United States.

For LD applications, Holderbaum and Kwee (2009) evaluated the placement of the SCR relative to the DPF. Considering the added fuel needed to heat the SCR system for cold start, and to regenerate the DPF with different frequencies because of changes in passive NO_2 regeneration, the authors conclude that for 1800 kg car with a 2-L engine, if cold starts occur more frequently than once every 60 km it is better to place the SCR in front of the DPF. Figure 11 shows some results. Note that 60 km threshold is greater than the distance used in certification cycles, wherein placing the SCR behind the DPF incurs a 2% fuel penalty versus a front placement. The forward SCR placement aids certification for both CO_2 and NO_x emissions.

Casarella (2011) showed the layout of a LD diesel emission control system with the DPF system in the front Figure 12. In the HD applications, the DOC and DPF are placed in front of the SCR and system is quite similar, except there may be a return line for the urea (called DEF, or diesel emissions fluid, here).

The DPF and the SCR subsystems both require their own control strategies. These can be quite complex. Looking at the filter management portion first, one needs to consider the amount of passive soot burn by NO_2. If this is not enough, then active regeneration is needed. Figure 13 shows a DPF regeneration map for an LD vehicle, wherein the light greyregion represents passive regeneration regimes (Boretto *et al*., 2004).

If the engine is run extensively in the active regeneration regime, the soot will build up on the filter and it will eventually need to be regenerated by active heating with fuel. In LD applications, this is usually done with a post injection in the cylinder. For HD applications, auxiliary fuel is injected into a DOC or, less commonly, a burner system is used in the exhaust. Active regeneration requires a control strategy. An example (Schommers, 2004) of which is shown in Figure 14. Key component models

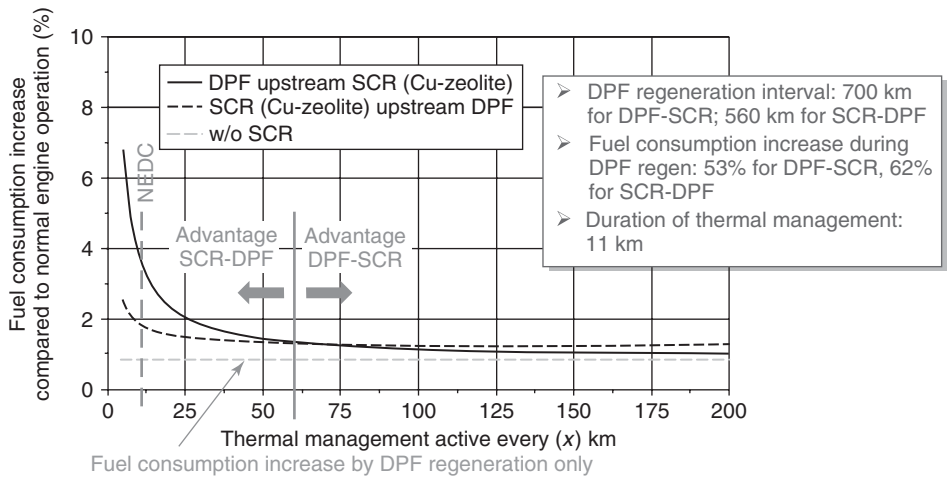


Figure 11. Considering fuel consumption for heating and regenerating the system, placing the SCR in front of the DPF in light-duty applications is beneficial if there is <60 km between cold starts (Holderbaum and Kwee, 2009). (Reproduced by permission of the Car Training Institute. © B. Holderbaum and Kwee.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

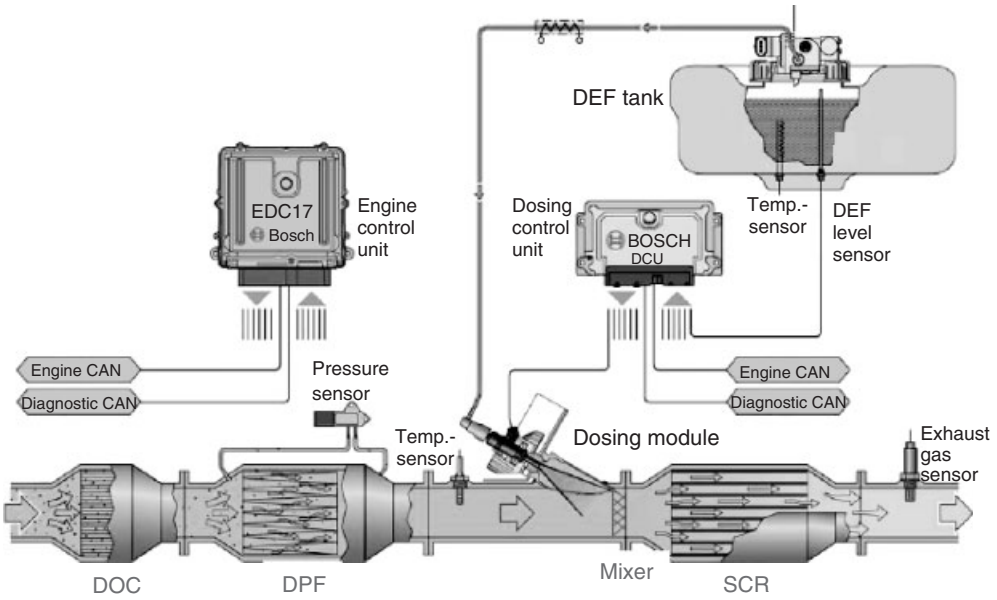


Figure 12. Layout of a DPF + SCR system (Casarella, 2011). (From Casarella, 2011. Reproduced by permission of Bosch.)

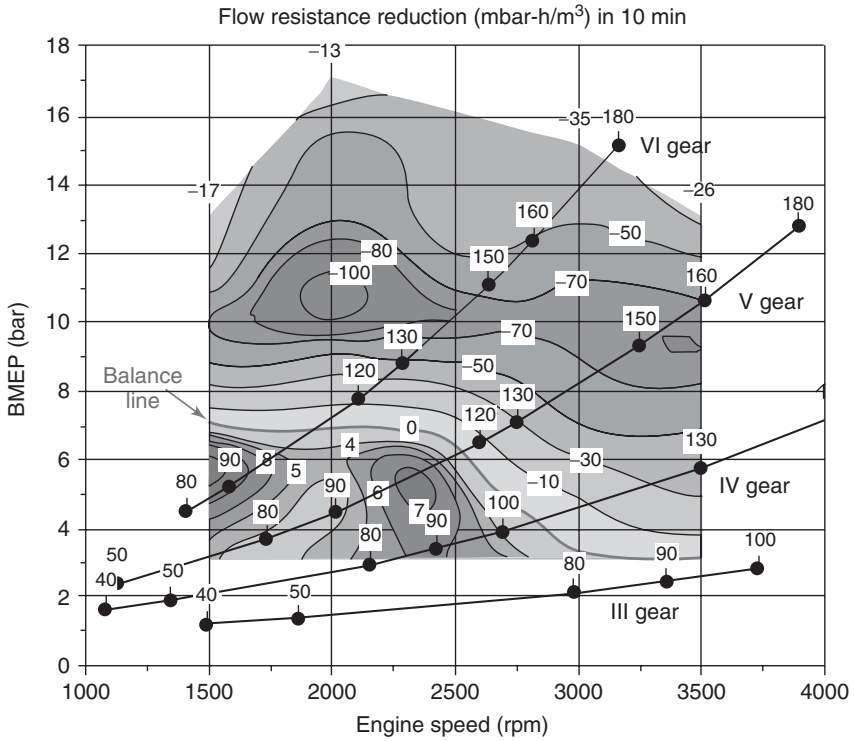


Figure 13. DPF regeneration map for a light-duty diesel car. Above the "balance line" is the passive regeneration regime using NO_2. Below the line is active regeneration. 1.9 L CR DI engine, D-Class vehicle, 10 min backpressure changes at 10 g/L soot (Boretto *et al.*, 2004). (Reproduced from Boretto *et al.*, 2004. © Boretto *et al.*)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

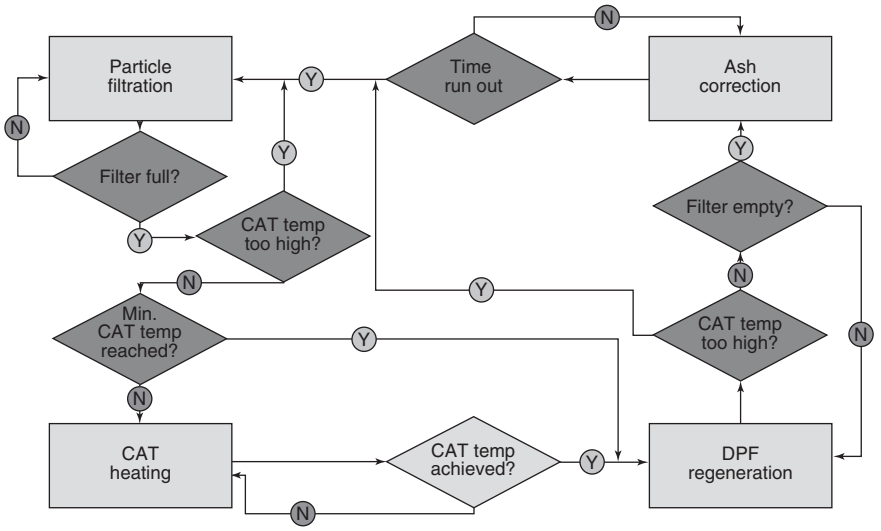


Figure 14. An example of a DPF regeneration control strategy (Schommers, 2004). (Reproduced from Schommers, 2004. © ÖVK.)

are shown in the grey diamond decision boxes, with the filter soot loading model being the most important ("Filter Full?"). Konstandopoulos *et al*. (2000) introduced many of the concepts that are presently used in many DPF soot models. The transient filtration algorithm, a deep-bed-to-cake-model, side-stream reactor technology, and NO_2 turnover/recycling are first described here. In a follow-up paper, Konstandopoulos *et al*. (2005) provide an overview of the multiscalar/multitemporal approach to DPF modeling. The paper addresses simulation from the micro-flow level with three-dimensional computer reconstruction of filter materials, including asymmetric channel designs and ash transport dynamics.

If too much soot builds on the filter and it is burned too quickly, filter damage can occur. Craig *et al*. (2005) described the various regeneration properties of cordierite filters. As shown in Figure 15, peak filter temperatures and gradients and completeness of regeneration are a function of DPF inlet temperature. Flow rate, soot loading, catalysts, and filter thermal mass also have significant impacts. The authors suggest that to increase soot loadings and improve regeneration efficiency, initial inlet temperatures of 550°C might be used, which are then increased as soot loading decreases.

If frequent active regenerations are contemplated, proximity of the DPF to the turbocharger can impact control strategy. For example, by moving the filter system from an under floor position to a close coupled position, the amount of regeneration fuel goes down because the system is inherently hotter. As such, the optimum soot loading, which balances fuel consumption because of increased pressure drop with regeneration frequency, drops from 10–11 g/L

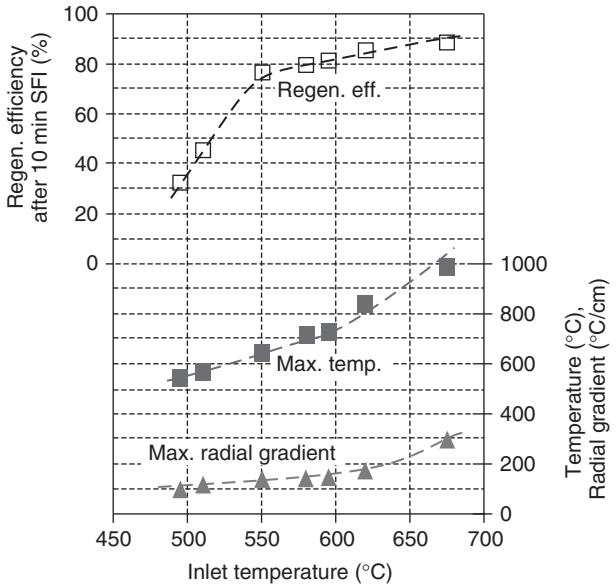


Figure 15. DPF inlet temperatures of >55°C are needed to efficiently regenerate catalyzed cordierite filters with 4 g/L soot loading. To further enhance regeneration, the inlet temperature can be increased later in the process (Craig *et al*., 2005). (Reproduced with permission from Craig *et al*., 2005. © A. Craig *et al*., Corning, Inc.)

down to 6–8 g/L for SiC filters (Oya *et al*., 2005). Such a placement may open up more filter material options.

Much of the work on DPFs is focused on improving regeneration. Chilumukuru *et al*. (2009) concluded that to minimize DPF regeneration fuel penalty, partial regenerations at high soot loadings are preferred. Figure 16 shows

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

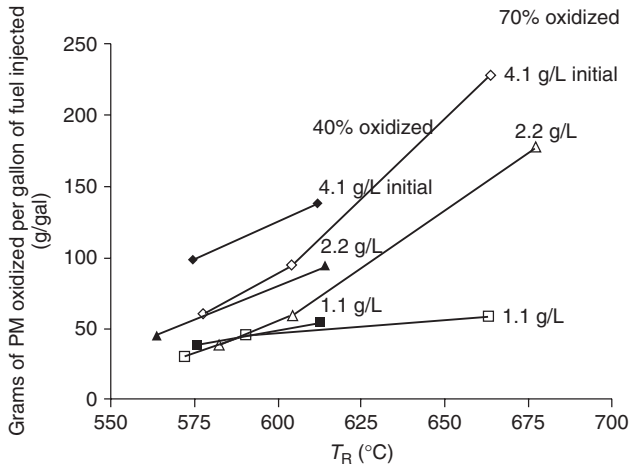


Figure 16. More soot is burned per unit of fuel for high soot loadings and for up to 70% regeneration efficiency (Chilumukuru *et al*., 2009). (From Chilumukuru *et al*., 2009. Copyright © 2009 SAE International. Reprinted with permission.)

the results leading to the conclusion, wherein more soot is burned per unit of fuel at the higher soot loadings, and up to 70% regeneration, the point of diminishing oxidation and increased fuel consumption. Leaving some soot membrane on the filter improves back pressure by keeping soot from penetrating into walls (no ash membrane). It should be noted, however, that one should be very careful in leaving soot on the DPF for too long, as it can age (graphitize or become poisoned) and become very difficult too oxidize, but may occasionally burn unexpectedly.

Post injections are not used for regenerating DPFs in HD application because cylinder wall wetting by fuel can dilute the lube oil. As these engines are expected to run 5–10× longer than an LD engine, this can compromise long term durability. Krüger *et al*. (2004) reported that oil dilution can be minimized or eliminated if several small post injections are done instead of a few larger ones.

SCR system control is becoming more important, as system deNO_x efficiency demands increase to 95%. Urea injection parameters are determined by NO_x quantity in the exhaust (concentration and flow rate), temperature, and the amount of ammonia stored in the catalyst. There is normally closed-loop feedback control using an NO_x sensor at the SCR exit, and in many applications, an NO_x sensor is used upstream to determine inlet NO_x levels.

As urea generally cannot be injected at temperatures <180°C because of evaporation and hydrolysis kinetics, it is important to properly manage ammonia storage in the catalyst for low load applications. Murata *et al*. (2008) showed that SCR efficiency at temperatures <265°C is strongly dependent on the amount of ammonia that is stored in the catalyst. They developed an algorithm that kept stored urea within control limits, resulting in improving deNO_x efficiency from nominally 50–75% in the Japanese HD transient cycle with an average temperature of only 160°C.

6 SUMMARY/CONCLUSIONS

6.1 Diesel particulate filters

Filter technology is advancing to provide systems with incrementally 20–30% lower back pressure, similar or higher soot mass limit to improve DPF management, and better filtration efficiency. DPF designs can help deNO_x performance through reduced thermal mass, allowing faster heat-up of the downstream deNO_x catalyst or by allowing incorporation of the catalyst on the filter, allowing faster light-off with reduced system back pressure. Fundamental knowledge on ash and soot membranes will allow the trend to continue.

6.2 Lean NO_x treatment

Lean deNO_x control is the leading area of interest in the field of vehicular emissions for good reason—NO_x and GHG regulations are tightening, and deNO_x translates to "deCO_2" using diesel or lean gasoline strategies. Work is continuing on alternative reductant forms involving gaseous and solid reagents, as well as fuels. Urea-SCR is accomplishing deNO_x efficiencies of 95% with reasonable systems and temperature ranges. SCR catalysts are evolving with improvements at both the low and HT regimes. HC-SCR approaches are also improving. New system designs for LNT + SCR (*in situ* ammonia) improve performance further, and the potential to meet the tightest NO_x regulations is demonstrated. When run in novel ways, the applicable temperature range of a standard LNT can be significantly expanded (to 650°C), with modest deNO_x efficiencies (80%) at high space velocities (125,000/h), and modest fuel penalties (<3%). Traditional HC-SCR (LNC) methods are getting renewed attention in commercial applications and with E85 reductant.

6.3 Diesel oxidation catalysts

Much of the new reports on DOCs concerned the interplay of precious metal formulations on HC oxidation, NO oxidation to NO_2, and the formation of N_2O. HC and CO oxidation is promoted by replacement of platinum with palladium, but NO_2 formation is compromised. NO_2 cannot form if HCs are present, as the HCs will reduce any NO_2

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

back to NO. HCs are also instrumental in reducing NO to N_2O, particularly at ~200–250°C if CO is not present.

6.4 Emission control system design

Most modern diesel engines will have DOC + DPF + SCR. Placement of the SCR in front or behind the DPF depends on application. If cold start emissions are important, like for LD certification requirements, the SCR is placed up front. The control strategies for the system can be complex, and is generally segregated by the DPF and the SCR. DPF control requires an understanding of passive and active regeneration regimes. Active regeneration is done by heating the filter to soot burning temperatures, the control of which is very important to avoid filter damage. SCR system control depends on proper urea injection strategies, in low load conditions, this will depend on knowing the amount of ammonia stored in the catalyst.

REFERENCES

Alano, E., Jean, E., Perrot, Y., *et al*. (2011) Compact SCR for passenger cars. SAE paper 2011-01-1318.

Austin, G., Naber, J., Johnson, J., *et al*. (2010) Effects of biodiesel blends on particulate matter oxidation in a catalyzed particulate filter during active regeneration. SAE paper 2010-01-0557.

Besch, M.C., Thiruvengadam, A., Carder, D., *et al*. (2010) *In-line, real-time exhaust PM emissions sensor—OBD and DPF control system applications*. 20th CRC On-Road Vehicle Emissions Workshop, March 2010, San Diego, CA.

Bisaiji, Y., Yoshida, K., Inoue, M., *et al*. (2011) Development of di-air—a new diesel deNO_x system by adsorbed intermediate reductants. SAE paper 2011-01-2089, JSAE paper 20119272.

Boger, T., Jamison, J., Warkins, J., *et al*. (2011a) Next generation aluminum titanate filter for light duty diesel applications. SAE paper 2011-01-0816.

Boger, T., He, S., Collins, T., *et al*. (2011b) A next generation cordierite diesel particle filter with significantly reduced pressure drop. SAE paper 2011-01-0813.

Boretto, G., Imarisio, R., Rellecati, P., *et al*. (2004) *Serial application of a catalyzed particulate filter on common rail DI diesel engines for passenger cars*. FISITA Conference, Barcelona, May 2004, Paper F2004V068.

Casarella, M. (2011) *Urea dosing systems and controls for light duty diesel applications*. SAE Light-Duty Diesel Emissions Symposium, Ypsilanti, MI, November 2011.

Cavataio, G., Guo, K., Xu, L., *et al*. (2011) *Comparing urea SCR to in-situ LNT + SCR aftertreatment systems for light duty vehicles*. SAE Light-Duty Diesel Emissions Symposium, November 2011, Ypsilanti, MI.

Cavataio, G., Jen, H.-W., Girard, J.W., *et al*. (2009) Impact and prevention of ultra-low contamination of platinum group metals on SCR catalysts due to DOC design. SAE paper 2009-01-0627.

Chapman, D.M., Fu, G., Augustine S., *et al*. (2010) New titania materials with improved stability and activity for vanadia-based selective catalytic reduction of NO_x. SAE 2010-01-1179.

Chen, H-Y, Mulla, S., Konduru, M., *et al*. (2009) NO_x adsorber catalysts with improved desulfation properties and enhanced low-temperature activity. SAE paper 2009-01-0283.

Chilumukuru, K.P., Arasappa, R., Johnson, J.H., *et al*. (2009) An experimental study of particulate thermal oxidation in a catalyzed filter during active regeneration. SAE paper 2009-01-1474.

Craig, A., Schelling, P., Tao, T., *et al*. (2005) *Performance aspects of cordierite diesel particulate filters in HD applications*. SAE Commercial Vehicle Conference, Chicago, October 2005.

Finch, S., Hnilicka, B. and Sindano, H. (2010) EURO6 light-duty vehicle obd project—evaluation and assessment of proposed EOBD emission thresholds. ACEA-Commissioned Report RD.10/320201.4, November 2010.

Fujii, S. (2011) *DPF design optimization from view point of the life-time performance with ash loading*. Emissions 2011 Conference, June 2011, Ann Arbor, MI.

Glover, L., Douglas, R., McCullough, G., *et al*. (2011) Performance characterization of a range of diesel oxidation catalysts: effect of Pt:Pd ratio on light off behavior and nitrogen species formation. SAE paper 2011-24-0193.

Han, J., Kim, E., Lee, T., *et al*. (2011) Urea-SCR catalysts with improved low temperature activity. SAE paper 2011-01-1315.

Henry, C., Langenderfer, D., Yezerets, A., *et al*. (2011a) *Passive catalytic approach to low temperature NO_x emission abatement*. US Department of Energy, Directions in Engine Efficiency and Emissions Research (DEER) Conference, October 3–6, 2011, Detroit.

Henry, C., Currier, N., Ottinger, N., *et al*. (2011b) Decoupling the interactions of hydrocarbons and oxides of nitrogen over diesel oxidation catalysts. SAE paper 2011-01-1137.

Heibel, A. (2010) *Advances in substrate technology*. SAE Heavy-Duty Diesel Emissions Control Symposium, Gothenburg, September 2010.

Holderbaum, B. and Kwee, H. (2009) *Integration of DPF and SCR—interfaces and interactions*. Presentation at the 5th International CTI Forum, SCR Systems, Fellbach, Germany, April 2009.

Inoue, M., Fukuma, T., Bisaiji, Y., *et al*. (2011) *Di-air: the new deNO_x system for future emission compliance*. Aachen Colloquium, October 2011.

Iretskaya, S., Golden, S., Tadrous, T., *et al*. (2010) PM control with low NO_2 tailpipe emissions by systems with non-PGM catalyzed DPF for passive soot regeneration. SAE paper 2010-01-0563.

Iwasaki, S., Mizutani, T., Miyairi, Y., *et al*. (2011) New design concept for diesel particulate filter. SAE paper 2011-01-0603.

Jackson, T. (2011) *Alternatives to liquid urea SCR for NO_x abatement HC LNC and solid SCR*. CTI Emissions Conference, May 2011, Detroit.

Jen, H.-W., Girard, J.W., Cavataio, G., *et al*. (2009) Detection, origin and effect of ultra-low platinum contamination on diesel-SCR catalysts. SAE paper 2008-01-2488.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Johannessen, T. (2011) *Next generation SCR system for fuel-efficient NO_x reduction*. SAE Light-Duty Diesel Emissions Symposium, Detroit, November 2011.

Kamasamudram, K., Henry, C. and Yezerets, A., (2011) *N_2O emissions from 2010 SCR systems*. US Department of Energy, Directions in Engine Efficiency and Emissions Research (DEER) Conference, October 3–6, 2011, Detroit.

Karin, P. and Hanamura, K. (2010) Particulate matter trapping and oxidation on a catalyst membrane. SAE paper 2010-01-0808.

Kim, C.H., Schmid, M., Schmieg, S.J., *et al*. (2011) The effect of Pt-Pd ratio on oxidation catalysts under simulated diesel exhaust. SAE paper 2011-01-1134.

Kondo, A., Sakuma, T., Sakurai, T., *et al*. (2010) *New particulate matter sensor for on-board diagnostics*. Aachen Colloquium, October 2010.

Konstandopoulos, A.G., Kostoglou, M., Skaperdas, E., *et al*. (2000) Fundamental studies of diesel particulate filters: transient loading, regeneration and aging. SAE paper 2000-01-1016.

Konstandopoulos, A.G., Kostoglou, M., Vlachos, N., *et al*. (2005) Progress in diesel particulate filter simulation. SAE paper 2005-01-0946.

Kröcher, O., Elsener, M., Mehring, M., *et al*. (2010) *Highly-developed thermal analysis methods for the characterization of soot and deposits in urea SCR systems*. AVL 6th International Exhaust Gas and Particulate Emissions Forum, March 2010, Ludwigsburg, Germany.

Kuwajima, M., Okawara, S., Tsuzuki, M., *et al*. (2009) Analysis of sophisticated DPNR catalyst, focused on PM particle number emissions. SAE paper 2009-01-0290.

Krüger, M., Wiartalla, A., Scholz, V., *et al*. (2004) *Diesel engine regeneration mode and engine long term effects*. Aachen Colloquium, October 2004.

Matsui, W., Suzuki, T., Ohta, Y., *et al*. (2011) A study on the improvement of NO_x reduction efficiency for a urea SCR system (Sixth Report)—clarifying N_2O formation mechanism. JSAE paper 20115720, October 2011.

MECA (2005) *Diesel Particulate Filter Maintenance: Current Practices and Experience*, http://www.MECA.org, June 2005.

Mizutani, T., Iwasaki, S., Miyairi, Y., *et al*. (2010) Performance verification of next generation diesel particulate filter. SAE paper 2010-01-0531.

Murata, Y., Tokui, S., Watanabe, S., *et al*. (2008) Improvement of NO_x reduction rate of urea-SCR system by NH_3 adsorption quantity control. SAE paper 2008-01-2498.

Narula, C., Yang, X., Bonnesen, P., *et al*. (2011) High performance NH_3 SCR zeolite catalysts for treatment of NO_x in emissions from off-road diesel engine. SAE paper 2011-01-1330.

Ottinger, N.A., Nguyen, K., Bunting, B.J., *et al*. (2009a) Effects of rapid high temperature cyclic aging on a fully-formulated lean NO_x trap catalyst. SAE paper 2009-01-0634.

Ottinger, N.A., Nguyen, K., Bunting, B.J., *et al*. (2009b) *Effect of thermal aging on NO oxidation and NO_x storage in a fully-formulated lean NO_x Trap*. Presentation at US Department of Energy Directions in Engine Efficiency and Emissions Research (DEER) Conference, Dearborn, Michigan, August 2009.

Oya, T., Yamayose, K., Ogyu, K., *et al*. (2005) Performance evaluation of SiC-DPF sintered with sintering additive. SAE paper 2005-01-0579.

Rakovec, N., Viswanathan, S. and Foster, D. (2011) Micro-scale study of DPF permeability as a function of PM loading. SAE paper 2011-01-0815.

Reichert, M. (2011) *Experience with SCR systems and expectations on future developments*. 7th International CTI Conference: SCR-Systems, July 2011, Stuttgart.

Rocher, L., Seguelong, T., Harle, V., *et al*. (2011) New generation fuel borne catalyst for reliable DPF operation in globally diverse fuels. SAE paper 2011-01-0297.

Sappok, A., Wong, V. (2009a) *Characteristics and effects of lubricant additive chemistry and exhaust conditions on diesel particulate filter service life and vehicle fuel economy*. Presentation at US Department of Energy Directions in Engine Efficiency and Emissions Research (DEER) Conference, Dearborn, Michigan, August 2009.

Sappok, A. and Wong, V. (2009b) Lubricant derived ash properties and their effects on diesel particulate filter pressure drop performance. *Journal of Engineering for Gas Turbines and Power*, **133**, 032805-1–032805-12. DOI: 10.1115/1.4001944 March 2011

Schommers, J. (2004) Das neue Mercedes Benz Dieselpartikelfilterkonzept fuer PKW in Verbindung mit der Abgasstufe EU4. 25 Wiener Motorensymposium, April 2004

Soeger, N., Mussman, L., Sesselmann, R., *et al*. (2005) Impact of aging and NO_x/soot ratio on the performance of a catalyzed particulate filter for heavy-duty diesel applications. SAE paper 2005-01-0663.

Southward, B.W.L., Basso, S., and Pfeifer, M. (2010) On the development of low PGM content direct soot combustion catalysts for diesel particulate filters. SAE paper 2010-01-0558.

Spurk, P., Frantz, S., Schütze, F.W., *et al*. (2010) *NO_2 formation on the DOC/DPF system—a system thought*. AVL 6th International Exhaust Gas and Particulate Emissions Forum, March 2010, Ludwigsburg, Germany.

Strzelec, A., Toops, T., Daw, S., *et al*. (2010) *Particulate matter oxidation kinetics: surface area dependence*. DOE CLEERS Conference, 2010.

Sumiya, S., Oyamada, H., Fujita, T., *et al*. (2009) Highly robust diesel oxidation catalyst for dual mode combustion system. SAE paper 2009-01-0280.

Tan, J., Solbrig, C., and Schmieg, S.J. (2011) The development of advanced two-way SCR/DPF systems to meet future heavy-duty diesel emissions. SAE paper 2011-01-1140.

Tang, W., Huang, X. and Kumar, S. (2011) *Sulfur effect and performance recovery of a DOC + CSF + Cu-zeolite SCR system*. US Department of Energy, Directions in Engine Efficiency and Emissions Research (DEER) Conference, October 3–6, 2011, Detroit.

Theis, J.R., Dearth, M., and McCabe, R. (2011) LNT + SCR catalyst systems optimized for NO_x conversion on diesel applications. SAE paper 2011-01-0305.

Thomas, D. and Highfield, T. (2011) *Ammonium formate/urea based diesel exhaust fluid for superior low temperature SCR performance*. SAE Light-Duty Diesel Emissions Symposium, Detroit, November 2011.

Vertin, K., He, S., and Heibel, A. (2009) Impacts of B20 biodiesel on cordierite diesel particulate filter performance. SAE paper 2009-01-2736.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Walker, A. (2010) *Optimising future catalyst systems*. SAE Heavy-Duty Diesel Emissions Control Conference, Gothenburg, September 2010.

Warkins, J., Heibel, A., George, S., *et al*. (2011) *Light duty filters*. SAE LD Diesel Emissions Control Symposium, November 2011, Ypsilanti, MI.

Warner, J.R., Dobson, D., and Cavataio, G. (2010) A study of active and passive regeneration using laboratory generated soot on a variety of SiC diesel particulate filter formulations. SAE 2010-01-0533.

Weibel, W., Waldbüßer, N., Wunsch, R., *et al*. (2009) *A novel approach to catalysis for NO_x reduction in diesel exhaust gas*. 8th International Catalysis for Automotive Pollution Control, Brussels, April 2009.

Weigl, M., Roduner, C., and Lauer, T. (2010) *Particle-filter-onboard-diagnosis by means of a soot-sensor downstream of the particle-filter*. AVL 6th International Exhaust Gas and Particulate Emissions Forum, March 2010, Ludwigsburg, Germany

Williams, A., McCormick, R., Luecke, J., *et al*. (2011) Impact of biodiesel impurities on the performance and durability of DOC, DPF and SCR technologies. SAE paper 2011-01-1136.

Xu, L., McCabe, R., Tennison, P., *et al*. (2011) Laboratory and vehicle demonstration of "2nd-Generation" LNT + in-situ SCR diesel emission control systems. SAE paper 2011-01-0308.

Yang, X. and Narula C. (2010) *Simple approach to tuning catalytic activity of MFI-zeolites for low temperature SCR of NO_x*. Poster at US Department of Energy Directions in Engine Efficiency and Emissions Research (DEER) Conference, September 27–30, 2010, Detroit.

Young, D., Hickman, D., Bhatia, G., *et al*. (2004) Catalyzed soot filters in close-coupled position for passenger vehicles. SAE paper 2004-01-0948.

Zhan, R., Li, W., Eakle S.T., *et al*. (2010) Development of a novel device to improve urea evaporation, mixing and distribution to enhance scr performance. SAE paper 2010-01-1185.

Zidat, S. (2010) *Diesel particulate filters on board diagnostic*. Poster at US Department of Energy Directions in Engine Efficiency and Emissions Research (DEER) Conference, September 27–30, Detroit.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto138
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Communication of Electric Vehicles

Christian Rehtanz, Willi Horenkamp, and Johannes Rolink

TU Dortmund University, Dortmund, Germany

1 Introduction 1
2 Communications between Charging Station and Vehicle 1
3 Example of a Charge Management Implementation 8
4 Communications in a Smart Grid 9
5 Integrating DC-FAST Charge Charging Stations into the Grid 14
6 Summary 15
7 Outlook 15
References 15

1 INTRODUCTION

The filling station system does not readily lend itself to electric vehicle (EV) charging because today's vehicle batteries have such long charging times. Instead, a new widespread charging infrastructure is needed with charging points at home, at work, and in parking garages, shopping centers, and similar locations. Not only will charging power have to be provided from the grid, but the vehicles will also have to communicate with the grid operator or energy supplier.

As described in the previous chapters, bidirectional charge management is an essential precondition for integrating large numbers of EVs into the power grid and providing ancillary services with EVs. "Bidirectional charge management" refers to two-way communications between EVs and the grid operator, energy supplier, or service center via the charging stations.

This chapter describes various ways to enable data communications between the grid operator, energy supplier, charging station, and EVs. First, though, the charging infrastructure must be efficient and fit for everyday use. The charging process must be as user-friendly as possible, especially when deployed in the mass market. Users would ideally only have to plug in the charging cable. Once the cable is connected, the charging process should automatically start and configure itself based on the battery's state of charge, user behavior, supply of electricity, and current load on the grid. Billing information would have to be securely and reliably transferred between the charging station/vehicle and the grid operator/energy supplier. In addition, value-added services could be integrated into the future communications infrastructure, such as traffic congestion forecasts, map updates for navigation systems, or the remote control of air-conditioning (AC) systems or independent car heaters.

Suitable communications structures are needed to serve the mass market. They will have to be standardized not only nationally but also internationally wherever possible. In addition, the charging infrastructure should also be built cost-effectively. It could use the kind of smart metering technology currently being tested by grid operators in pilot projects on automatic meter reading. This would allow the information technology (IT) requirements to be integrated into a future smart grid.

2 COMMUNICATIONS BETWEEN CHARGING STATION AND VEHICLE

This section describes ways to extend standardized communications processes between the charging station and the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

vehicle charge controller in order to support the integration of EVs into a smart grid. It focuses on internet protocol (IP) communications between the charging station and the vehicle using power line communications (PLC) technology.

2.1 Communications in accordance with IEC 61851

IEC 61851/SAE J1772 describes the technical requirements for connecting EVs to the electricity supply system. These standards differentiate between four charging modes:

- Mode 1: single-phase or three-phase connection, standardized plug-and-socket connections. Requires ground fault circuit interrupter (GFCI) and overcurrent protection.
- Mode 2: same as charging mode 1, but uses a control pilot between the charging station and the vehicle.
- Mode 3: uses a type-specific charging station.
- Mode 4: indirect connection, uses an external charger (DC charging).

Mode 1 is used whenever an EV is plugged into a normal power socket at home. The grid and the vehicle do not need to communicate with each other. Modes 2 and 3, by contrast, cover home-based or public charging stations with charging currents >16 A. These situations require an additional controller and, consequently, a charging station with a control unit. The control unit has three main functions: verifying that the EV is connected properly, constantly monitoring the protective earth conductor, and turning the charging socket on and off. However, it can optionally detect and set the available charging current and lock and unlock the connector. This is done via a control pilot that is integrated in the charging cable alongside the live conductors (Figure 1).

The charging station contains a 1 kHz signal generator with an adjustable duty cycle between 10% and 80%. The duty cycle (pulse-width modulation, PWM signal) communicates the charging station's maximum possible charging current to the vehicle charge controller. The PWM signal is applied across various resistors that indicate the vehicle's operating state to the charging station. When it receives the signal, the charging station can detect the operating state based on the PWM voltage as described later (Figure 1) and control the charging process accordingly:

- State A (12 V): connector not plugged in/vehicle not connected.
- State B (9 V): connector plugged in/vehicle connected, not ready for charging.
- State C (6 V): vehicle ready for charging.
- State D (3 V): vehicle ready for charging, external ventilation required.

IEC 61851 describes the exact charging sequence over the control pilot, so there is no need to detail it in this chapter.

2.2 Communications in accordance with IEC 61851 with extended transfer of parameters

IEC 61851 allows additional parameters to be transmitted before or after charging. They can include the customer/vehicle identification (ID) or the state of charge before or after charging. Parameter transmission over the control pilot is restricted during charging, however, so as not to interfere with the control pilot's protective function. The following section describes the typical process for transmitting a customer's contract number (customer ID) before charging. This requires more hardware and software (ID/PWM control unit) in the charging station and the vehicle than the standard IEC 61851 charging process

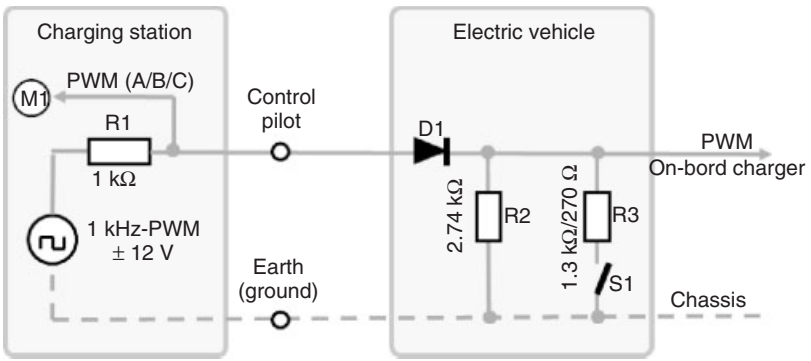


Figure 1. Equivalent circuit diagram of the pilot circuit (according to IEC 61851).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

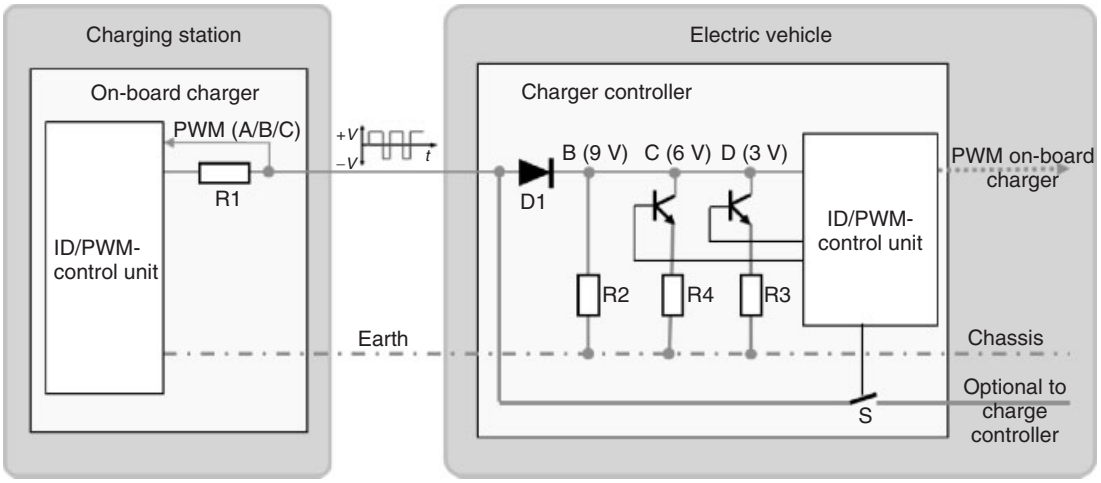


Figure 2. Block diagram of ID transmission over the control pilot.

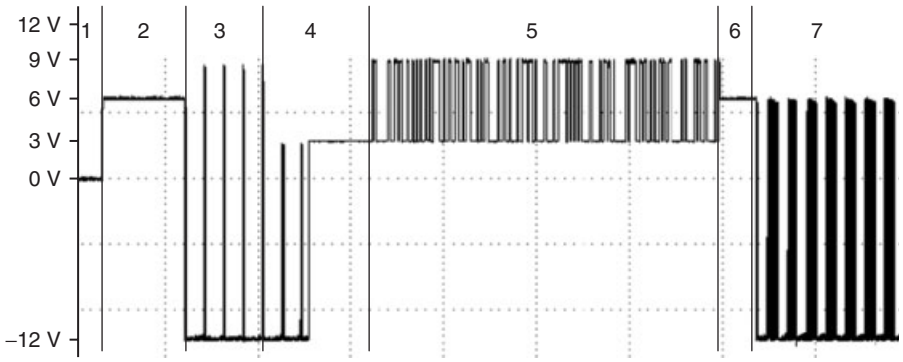


Figure 3. Waveform of an ID transmission.

(Figure 2). The customer or contract ID must either be stored in the charging station or be retrieved by the service center, grid operator, or energy supplier.

Figure 3 shows the waveform of the PWM signal during the transmission of a customer's contract number over the control pilot.

The PWM signal is not sent until the charging station detects that a vehicle is connected. This ensures that the charging station is only turned on after a vehicle has been connected. Table 1 describes the sequence of the ID transmission over the control pilot.

If the charging station does not recognize the ID as valid, the vehicle resends the ID. If another error occurs, the user has to disconnect the vehicle from the charging station and restart data communications. As the data transmissions are encrypted, both the ID module and the charging station must have the encryption key. The key is recalculated for each new ID transmission.

Table 1. Sequence of ID transmission over the control pilot.

State	Function
1	Vehicle is properly connected to the charging station, the charging station subsequently turns on PWM voltage
2	PWM signal 100/0, vehicle sends state C (6 V)
3	Charging station generates PWM duty cycle of 5/95 or 90/10, the vehicle then sends state B for 500 ms
4	Vehicle sends 500 ms state D (3 V) in response
5	Vehicle waits for PWM 100/0, then ID transmission is started (amplitude modulation of the PWM signal between 9 and 3 V)
6	ID transmission complete, vehicle sends state C PWM signal (6 V)
7	Charging station sends maximum charging current and turns on the charging voltage

If a PWM circuit is integrated in the vehicle, the PWM signal can be sent through the ID module logic (cf. Figure 2, Relay S). In this case, the ID/PWM module only identifies

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

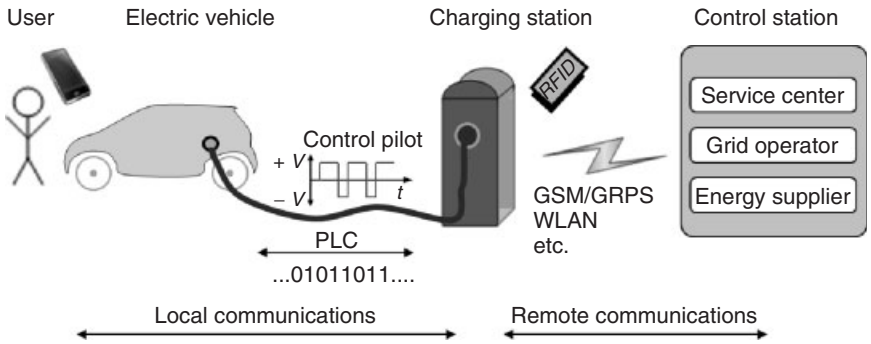


Figure 4. Stakeholders directly involved in the charging process.

the contract ID. All other communications take place in accordance with IEC 61851.

To minimize the ID module's energy consumption in battery mode, the PWM signal sent by the charging station turns the ID/PWM module on and off. This allows the ID/PWM module to run on an integrated battery without requiring a connection to the on-board power supply. The ID/PWM module can, therefore, also be used to activate the charging station for vehicles that only support low charging outputs, such as electric scooters.

2.3 Alternative communication links

Development has focused not only on the actual charging process but also on integrating value-added services such as checking the state of charge during charging. Smart charge management can ensure that the vehicle's battery is charged during periods when there is an abundance of renewable energy or when the load on the grid is low. If the charging station can communicate directly with the vehicle, it can also provide other services such as on-board internet, traffic congestion forecasts, map updates for navigation systems, or the synchronization of e-content (music, videos, and news) with the vehicle's entertainment system. Figure 4 shows communications flows between the various stakeholders during charging. There are two basic types of communications: remote and local. In remote communications, the charging station communicates with a control center belonging to the grid operator, energy supplier, or service center. The control center can then deliver additional services to the user. In local communications, the charging station communicates with the vehicle. The following section examines local communications.

Charging station solutions can be grouped into five different categories of local communications (Table 2):

1. Communications between charging station and vehicle in accordance with IEC 61851.

Table 2. Comparison of local communications categories.

Category	1	2	3	4	5
Transmission of contract information	Not possible	Yes, before and after charging	Specific to the card/person	Specific to the device/person	Specific to the vehicle
Several variable charging profiles	Possible (PWM)	Possible (PWM)	Possible (PWM)	Possible (PWM)	Possible (PLC)
Status messages during charging	Not possible	Not possible	Only via display on charging station	On site via the device	In the vehicle
Dynamic vehicle data, for example, state of charge	Not possible	Not possible	Not possible	Not possible	Possible, regular updates
Charge management in a smart grid	Not possible	Partially possible	Only possible offline	Only possible offline	Possible
Value-added services for the vehicle	Not possible	Not possible	Not possible	Not possible	Possible

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2. Communications with extended transfer of parameters in accordance with IEC 61851.
3. Charging stations with a static user interface but no data interface to the vehicle. This includes smart cards, radio frequency identifications (RFIDs), or similar solutions. User can authenticate himself or herself using the data stored on the card and then start the charging process.
4. Charging stations with an interactive user interface but no data interface to the vehicle. For example, charging stations can be operated from a display or a separate device such as a smart phone.
5. Charging stations with a bidirectional data interface to the vehicle. The data of the vehicle or vehicle owner is sent over an IP-based protocol.

In all categories except category 1, an internet platform aggregates all the information that the charging station receives about the charging process and associates it with the charging station ID. If users can access the platform (e.g., from a smart phone), they can look up status messages before or during charging, or choose from different charging profiles or tariffs.

The following section describes one approach for implementing end-to-end communications between the grid operator, energy supplier, and vehicle/customer using PLC between the charging station and the vehicle. Not only does it enable effective demand-side/charge management for the grid, but also it supports internet value-added services for the vehicle.

2.4 Power line communications

The smart charge communications (SCCs) protocol specified by RWE and Daimler is in the process of being standardized (ISO/IEC 15118; SAE J1772; RWE, Daimler, Insys Microelectronics, and Emsycon, 2010a, b). It specifies an IP communications format for providing the information needed for identification, billing, and charge management and exchanging this information between the vehicle and the charging station. All data is transferred using PLC over the phase and neutral conductors, so there is no need to integrate another communications contact in the charging cable. Communications are implemented by control/communications units integrated in the vehicle and the charging station. The units set up the PLC connection over the charging cable (Figure 5). The charging station communicates with a higher level data center over a wide area network (WAN) to verify the customer/contract data.

PLC between the vehicle and the charging station enables complete automation of the charging process, which are:

- unambiguous identification of the vehicle, vehicle owner, or vehicle operator;
- billing based on kilowatt hour, person, or vehicle;
- integration of value-added/internet services during charging, for example, managed charging;
- updating the navigation system with traffic data supplied by internet traffic services;
- internet-based remote control and/or remote monitoring of the charging process, for example, checking the state

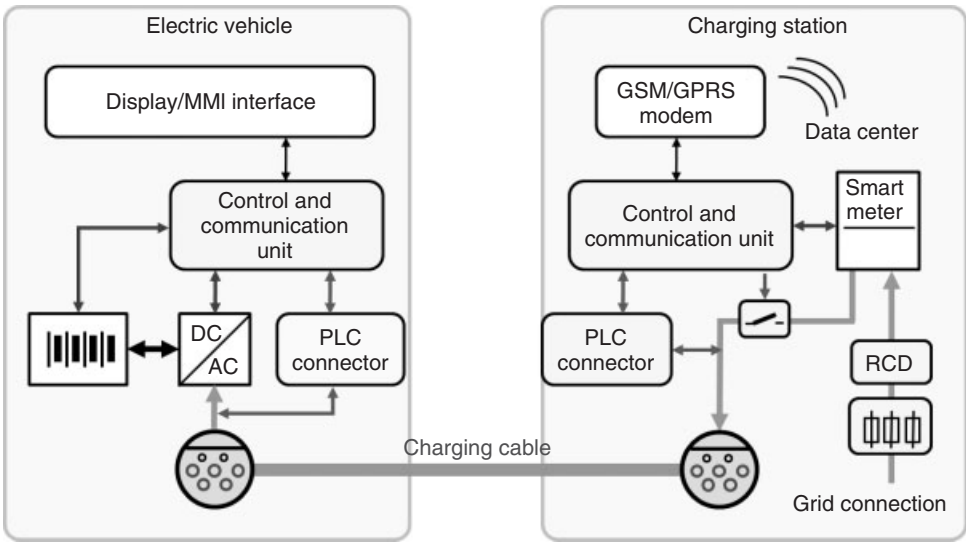


Figure 5. Vehicle and charging station system components.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of the battery, turning on independent car heaters/AC systems.

The specified functions provide a high degree of transparency during and after charging for vehicle operators, grid operators, and electricity suppliers.

2.4.1 Smart charge communications protocol

The SCCs protocol defines bidirectional communications between the vehicle and the charging station. The current communications medium is the latest version of the PLC physical layer, HomePlug1.0 Turbo, which supports a maximum bandwidth of 100 Mbps. The network layer is based on IP Version 4 or 6. To protect against potential attacks, the connection is encrypted at the transport layer using Transport Layer Security Version 1.2. At the beginning of each connection, the encryption key is exchanged using the Diffie–Hellman key exchange method (Rescorla, 1999), which offers strong security against eavesdropping. The session layer consists of the SCC session protocol, which allows a connection to be uniquely assigned to a charging session. The application layer and the presentation layer define the content and presentation of the messages. Figure 6 shows the protocol stack for the SCCs protocol.

2.4.2 Smart message language

The smart message language (SML) protocol was developed specifically for smart metering. The specification's current version, 1.03, is freely available on the internet (Forum network technology/network operation in the VDE, 2008b). There is also a reference implementation based on Java. SML messages can be highly compressed using byte encoding and so require very little bandwidth during transmission. Individual messages can also be streamed, so the sender and the receiver can process messages directly without needing to buffer them. This can reduce controller size and thus cut costs, especially for embedded systems, which is an important consideration in the automotive sector.

SML defines several data types such as strings, (unsigned) integers, and Booleans as well as data types specifically suited to transfer meter readings. Complex data structures can be defined and transferred as messages using a generic container. This allows the protocol to be used for EVs in addition to smart metering.

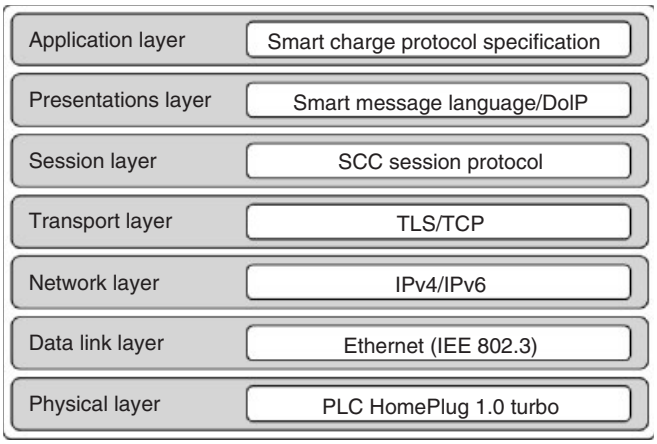


Figure 6. Protocol stack of the smart charge protocol.

2.4.3 Alternative protocols

Extensible markup language (XML) and/or JavaScript Object Notation (JSON) are also possible protocols. These data representation formats have the advantage of being much more common than SML, especially internationally. Moreover, XML and JSON can draw on a large pool of development tools and reference implementations. JSON offers an advantage over XML; JSON messages can also be highly compressed using byte encoding and transmitted much more efficiently. Figure 7 shows a section of the identification message in SML and XML as a comparison. As JSON represents data much like SML in terms of presentation and size, it will not be described in any further detail.

XML uses ASCII (American Standard Code for Information Interchange) encoding, which needs one byte per character. As a result, over six times, as many bytes have to be transmitted with XML (282 bytes) than SML (45 bytes). Table 3 shows the summary and evaluation of the protocols.

2.5 PLC over the control pilot

One way to transmit additional information over the control pilot during charging is by modulating the amplitude of the PWM signal (Section 2.2). However, this process should not breach the upper or lower limits of the voltage range as described in IEC 61851. To prevent this, the system can use a variable base voltage with a predetermined offset. If the voltage generated by the charging station's PWM generator is near the lower limit of the range, the circuit logic will have to ensure that only positively amplitude-modulated signals are applied to the control pilot. The opposite applies if the voltage is near the upper limit. The disadvantage of this approach is a relatively low data

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

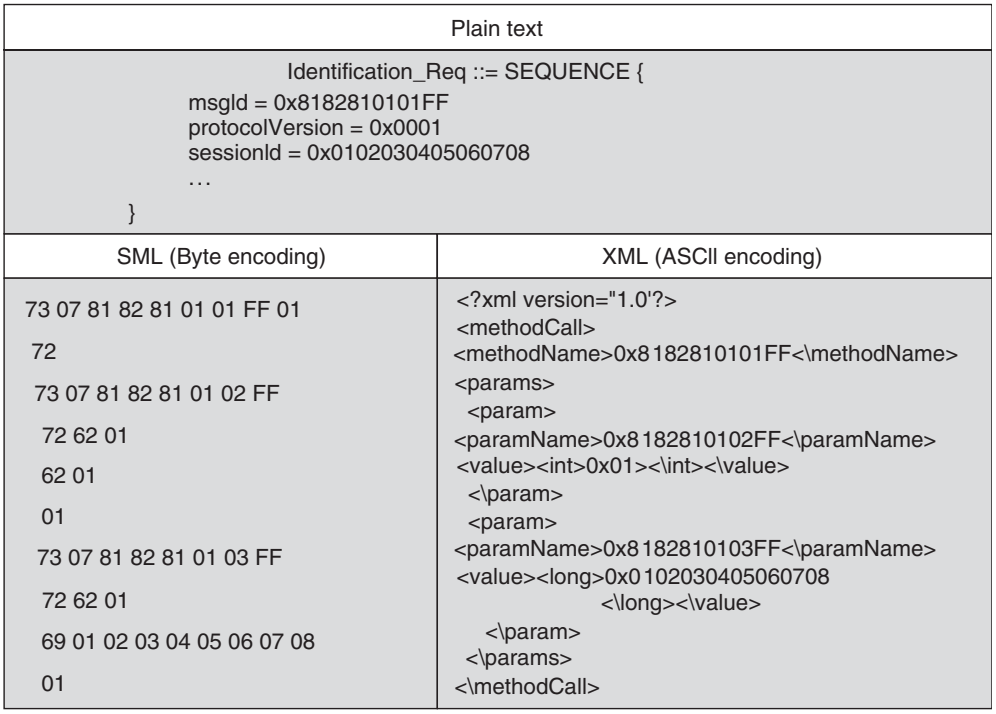


Figure 7. Comparison of SML and XML encodings.

Table 3. Pros and cons of alternative data protocols.

Criterion	SML	XML	JSON
Compact presentation	+++	−	++
Widespread acceptance/ standardization	+	+++	++
Availability of reference implementations	−	++	++
Availability of tools	−	++	++
Suitable for use in embedded systems	+++	+	++

+++, Excellent; ++, Very good; +, Good; −, Sufficient.

rate and considerable extra effort, especially if bilateral communications are required between the vehicle and the charging station.

PLC technology can also be used. In this case, data is sent and received over the control pilot instead of a live conductor. This approach enables much higher data rates in bidirectional communications than amplitude modulation. Figure 8 (Area A) shows a PLC signal at −12 V and during a PWM pulse (Area B). The same data packet was sent and received at (A) and (B). The PWM signal edge interferes with the PLC signal. Appropriate error handling can be used to ensure that data transmissions are error-free. In this case, the question is whether the data rate reduction at time (B) is acceptable. As PLC requires the charging station to constantly transmit a 5/95 PWM signal, it may be possible to only allow PLC in the periods when the PWM signal voltage is constant. Data could then be transmitted without interference 95% of the time, or 950 μs. Additional hardware and software would, however, be needed to synchronize the PLC signal with the PWM signal.

As with the previous approach, the additional data transmissions over the control pilot should not violate the limits of the voltage range or rise and fall times of the PWM signal during charging. This can be done by configuring the control pilot circuit upstream from the charge controller or by optimizing the interface circuit for the PLC signal. Optimizing the input power of the PLC signal with respect to the maximum charging cable length will also reduce electromagnetic compatibility (EMC) disturbances.

Transmitting data over the control pilot with PLC technology offers one key benefit: the control pilot is much shorter than a live conductor and so is much less susceptible to disturbances. As a result, a PLC signal can be transmitted with much less power over a control pilot than over the grid. In addition, the PLC signal does not propagate throughout the power grid, which makes it much easier to comply with EMC limits.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

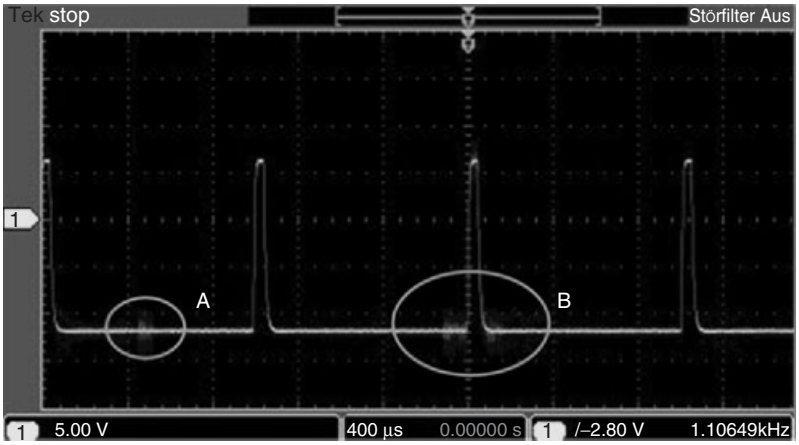


Figure 8. Sending a PLC signal over the control pilot.

3 EXAMPLE OF A CHARGE MANAGEMENT IMPLEMENTATION

Universal charge management can be implemented by combining PLC with the control pilot. The charging station controller (CSC) handles PWM signal generation, management, and PLC modem control. The meter readings are captured from the optical port of an electronic power meter and processed in the CSC. The implementation can use either the SML protocol or the D0 interface specified in DIN EN 62056-21 as the data protocol. Remote communications with the control center use a WAN connection (e.g., cellular connection).

The vehicle communications unit is designed so that the charging station can be activated even if PLC is not used. In this case, the charging station can be activated by ID communication or a service center.

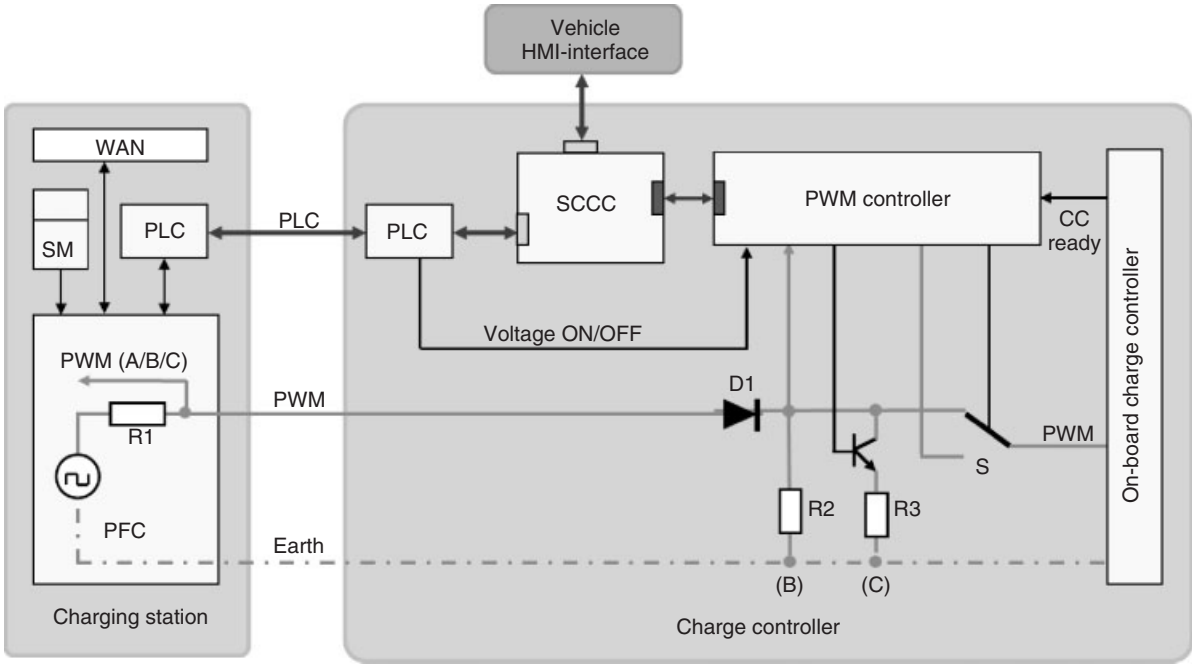
3.1 The charging process

Once the charging station and the vehicle have been physically connected, the charging station applies a PWM signal across resistor R2 (state B). If the PWM controller detects a PWM signal between 10% and 80%, this means the charging station does not support PLC. The PWM signal will instead be used to specify the current for the charge controller. The charge controller tells the PWM controller to apply power across resistor R3 (state C). This starts the charging process and activates the socket in the charging station (Figure 9).

If the vehicle's PWM controller detects a 5/95 PWM signal, the charging process is managed via PLC (Figure 10). State B is indicated using R2. The charging station then instructs the vehicle to use IP communications and activates the Dynamic Host Configuration Protocol (DHCP) server. The vehicle obtains an ID address via DHCP and initiates the communication process by sending the vehicle ID to the charging station. The charging station responds with the charging point ID, among other things. This is followed by the service selection: the vehicle looks up for the available services, and the charging station responds with a list of available services and optional parameters. After selecting the service, the grid/tariff parameters for the charge are negotiated. The vehicle sends the contract ID, the energy supplier, and the estimated energy requirements. The charging station responds with the maximum charging current, the nominal voltage, and a list with tariff information linked to the preferred charge curve (for load management). Next, the vehicle sends out a message to lock the connector (line lock) followed by the initial meter reading lookup request (metering request and metering receipt). The vehicle then prompts the charging station to turn on the voltage. The charging station checks the status word sent by the vehicle and the pilot signal for state C (ready for charging), and then turns on the voltage. During charging, the vehicle constantly sends data packet requests via metering request and metering receipt messages. These include not only the meter reading but also the current tariff, the maximum permitted charging output, and the status word of the vehicle and/or the charging station. The vehicle checks the meter reading and the charge parameters and, to confirm the amount of energy charged, sends a signed copy of the parameters received.

If the maximum charging output changes because the load on the grid has changed, the charging station may prompt the vehicle to renegotiate the grid parameters (power discovery). To end the charging process, the vehicle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Shortcuts:

PLC:	Power line communications connector	PFC:	Pilot function controller
SCCC:	Smart charge communication controller	SM:	Smart meter
PWM:	Pulse width modulation	PWMC:	PWM controller
HMI:	Human machine interface	CC:	Charge controller

Figure 9. Block diagram of PWM communications, without PLC.

sends a power delivery message with the request to turn off the voltage and unlock the connector (line lock).

4 COMMUNICATIONS IN A SMART GRID

This section describes the integration of vehicles into a smart grid. It presents various approaches, from managed charging to bidirectional communications using smart metering technologies.

In future, charging infrastructure will likely consist of various charging station designs and functions:

1. Charging socket at home, charging output up to 3.7 kW single phase or 11 kW three phase, standard socket.
2. Charging at work, charging output up to 3.7 kW single phase or 11 kW three phase, possibly with energy consumption metering.
3. Parking lots and shopping centers, charging outputs >3.7 kW, DC charging stations.
4. Park and ride parking lots, low charging outputs per vehicle because of the long timeframes, with energy consumption metering.
5. Parking spaces with short vehicle-standing times, DC charging stations.

The combination of managed charging and a smart grid is ideal, especially if many vehicles are parked for long periods of time and the energy stored in the vehicle batteries is fed back into the grid. Figure 11 shows what parameters affect charge management. Charging output depends on two parameters: power supply and current grid load. However, it is also indirectly affected by the current state of the grid and the connection and disconnection conditions. Vehicle charge times, by contrast, depend on the current state of charge and the distance traveled by the vehicle operator.

The Section 4.1 describes several options in connection with smart grid applications, which are managed charging and grid and energy management.

4.1 Managed charging using the control pilot

Managed EV charging can play a significant role in load balancing—particularly in low voltage supply systems—as long as the vehicles are charged at home or work and the charging time window is long enough (Figure 12).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

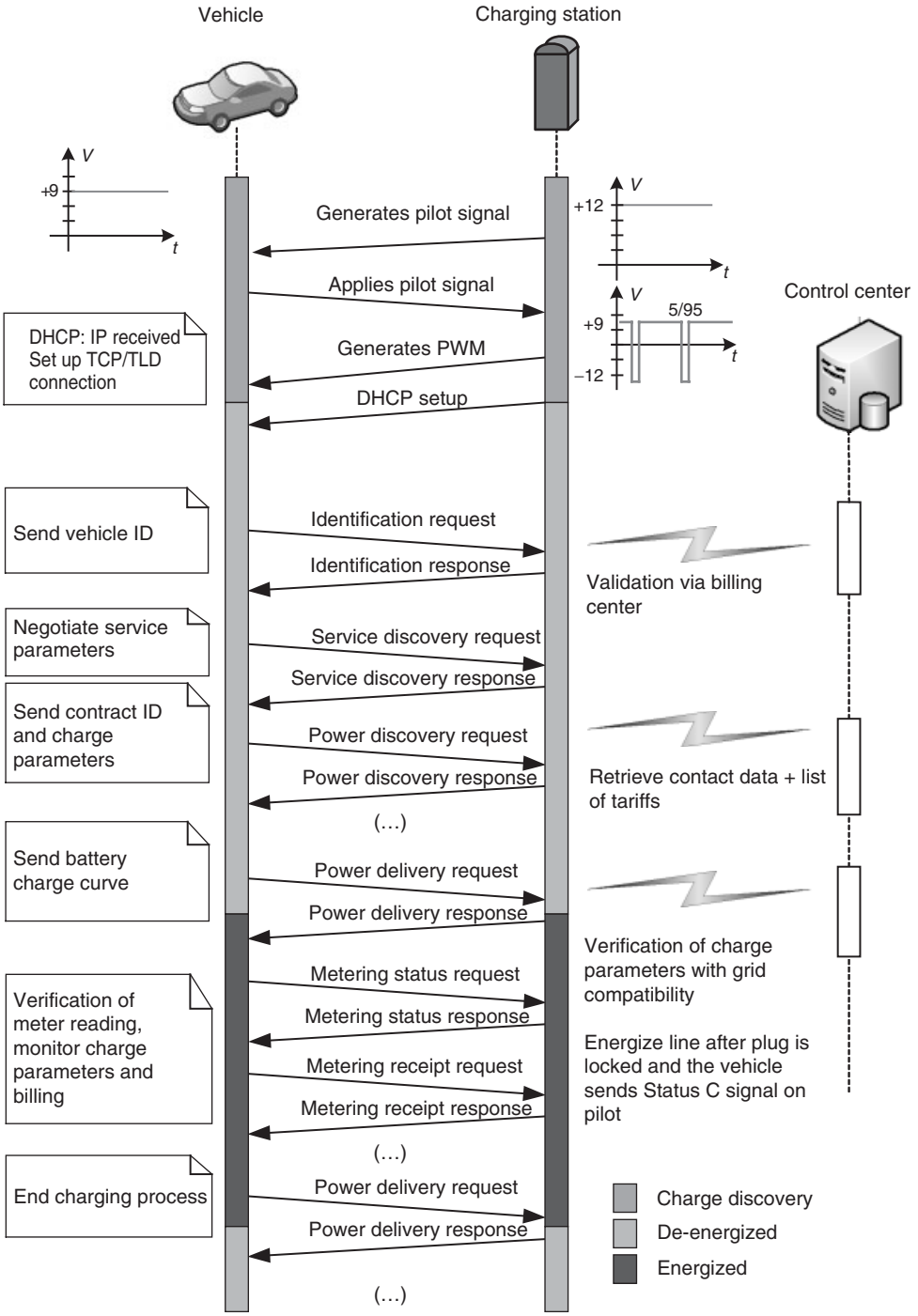


Figure 10. Smart charge communication protocol (developed by RWE and Daimler).

The shaded area shows charging time windows specified by a grid operator along with the maximum charging current.

The charging current is specified by the charging station via the pulse duty cycle (PWM signal) and processed by the charge controller in the vehicle. The PWM signal can be generated by a microcontroller (PWM generator) that can be controlled by an external signal (Figure 13).

In the simplest case, the charging process (charging time and charging current) can be controlled by a time switch. Time switches with multiple switching contacts can be used

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

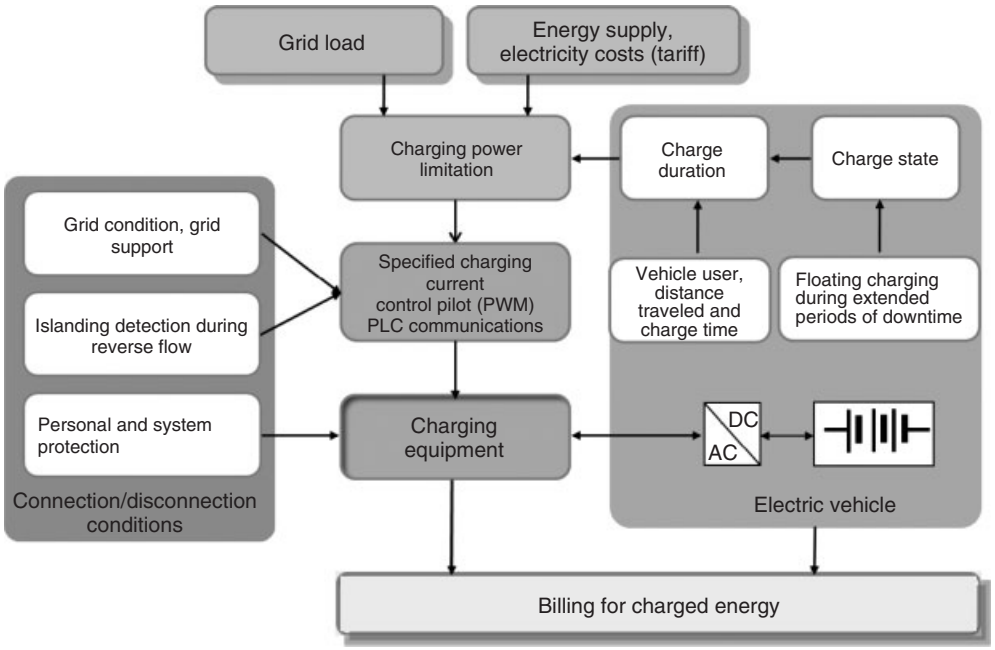


Figure 11. Parameters influencing charge management.

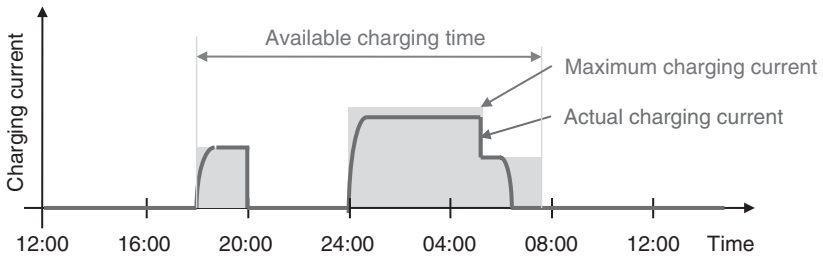


Figure 12. Charging current depending on the charging time.

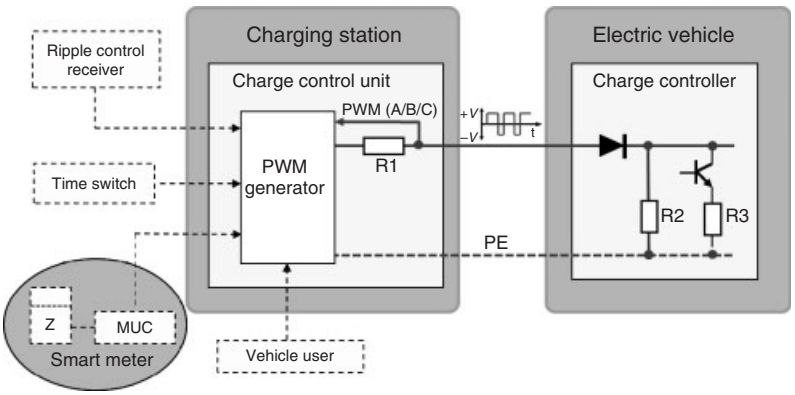


Figure 13. Specifying the charging current over the control pilot.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to specify variable charging times and varying charging currents. Another possibility is to turn loads on and off with a ripple control signal. Many modern-day low voltage supply systems support the use of ripple control to turn loads on and off or to switch to certain tariffs. Ripple control signals can also set charging times and maximum charging currents for PWM signal generation. The advantage of this approach is that it does not require any additional communications structures. Once smart metering has been widely deployed, the smart meter can take over this function, possibly in connection with the multiutility communications (MUCs) controller. The charging station, grid operator, and energy supplier can communicate over PLC, wireless (global system for mobile communications, GSM and general packet radio service, GPRS), or internet (digital subscriber line, DSL) (Rehtanz, Horenkamp, and Ruthe, 2011).

4.2 Active grid and energy management with smart metering technologies

Active grid/charge management refers to the bidirectional transfer of energy between the grid and vehicle batteries. Energy stored in vehicle batteries can not only power the vehicle but also be fed back into the grid. However, this requires additional charging cycles, which can reduce battery's operating life with current technologies. Users also expect batteries to be charged by a particular time. As power flows in both directions, these charge time expectations can only be met if the vehicle charging controller and the charging station (service connection point) are configured accordingly. IT and grid requirements are even higher if EV batteries are also used to provide control power.

The universal availability of smart metering, in the medium term, will enable the integration of distributed energy generation units into a data network and active EV-based grid and energy management. Smart metering technology represents a key pillar in this strategy, particularly when it comes to smart charging of EVs.

4.2.1 *Grid and charge management with a smart metering system*

This section describes how a charging control unit can be connected to a smart metering system made by Echelon Corporation (2010a,b,c). The smart meters have to be equipped with an MEP (multipurpose expansion port) interface®. This is a serial interface that enables communications between the server and an external MEP device. Data can be exchanged bidirectionally between a central computer and the end-user device. The MEP device can also access the meter directly. Data communications between the server and the MEP device are prioritized. High priority data is transmitted as quickly as possible and immediately relayed to a control center by the meter. The MEP protocol is a vendor-specific, bidirectional single layer protocol. Messages can optionally be encrypted with 128-bit RC4 encryption to prevent unauthorized users from accessing the data. Figure 14 shows how the system has been implemented with a grid operator. It is currently being trialed to determine performance parameters such as latency time and communications reliability.

The smart meter captures billing-related data and establishes a narrow band PLC connection to the secondary

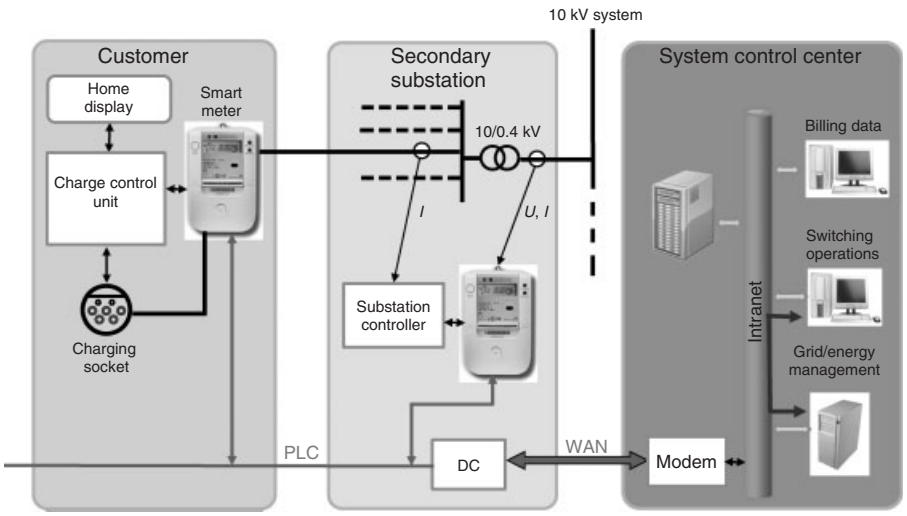


Figure 14. Charging control unit with an Echelon smart metering system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

substation. The charging control unit communicates with the meter over the MEP interface, generates the PWM signals, and verifies that the charging cable is properly connected to the vehicle. A separate serial or Ethernet port can be connected to a home display to enable the customer to control charging output based on considerations such as the current tariff or planned trip distance.

The secondary substation has a data concentrator in addition to the smart meter. The data concentrator transmits data to the control center over a WAN. No additional meters are needed to track current flowing through individual low voltage feeders. Current levels on individual feeder conductors are captured by current transformers and transmitted to the smart meter over the MEP interface by the substation controller. This allows feeder current levels in the substation to be transmitted to or retrieved by the control center in response to certain events.

The secondary substation and the control center communicate over a WAN (PLC, cellular, optical fiber, etc.). The control center can, for example, provide the customer's charging control unit with tariffs that reflect the current load on the grid or current power supply levels. Charging can then either be performed automatically or initiated manually by the customer.

4.2.2 Grid and charge management using MUC

MUC offers another way to use smart metering technology for grid, energy, and charge management (Forum network technology/network operation in the VDE, 2008a). One of the goals of MUC's developers was to separate the electric meter from the data collector used for communications between users (customer, grid operator, etc.). The MUC Requirements Specification V1.0 does not, however, explicitly support bidirectional communications for managing EV charging. Figure 15 shows how MUC can be used in more advanced applications.

Unlike the previous approach (Section 4.2.1), in which all communications went through a control center, this approach enables local and central grids and energy management systems to communicate directly with one another. Energy and grid management is optimized at four levels:

1. Local, autonomous optimization of the building's energy consumption. This is done in conjunction with local controllable loads, such as vehicle batteries or heat pumps together with distributed generation (photovoltaics, PV, and micro combined heat and power, CHP). The actual optimization method varies depending on parameters such as the current supply of electricity and the current tariff. Customers can also set charging output as needed on home displays or similar devices.
2. Energy requirements are also coordinated at the secondary substation level in connection with grid management. If only certain secondary substations are overloaded, the load only needs to be limited on

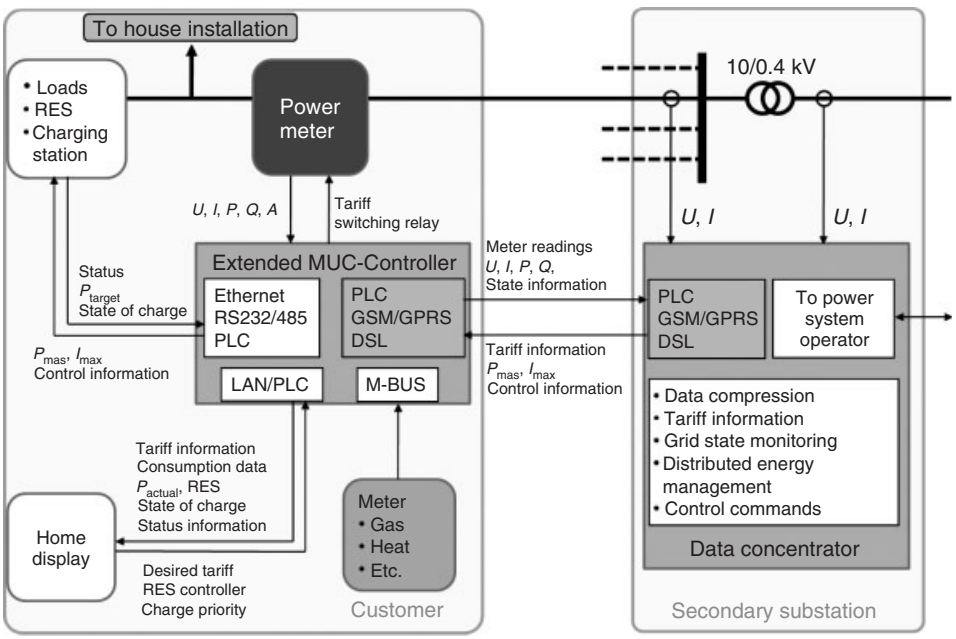


Figure 15. Grid and charge management using MUC.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

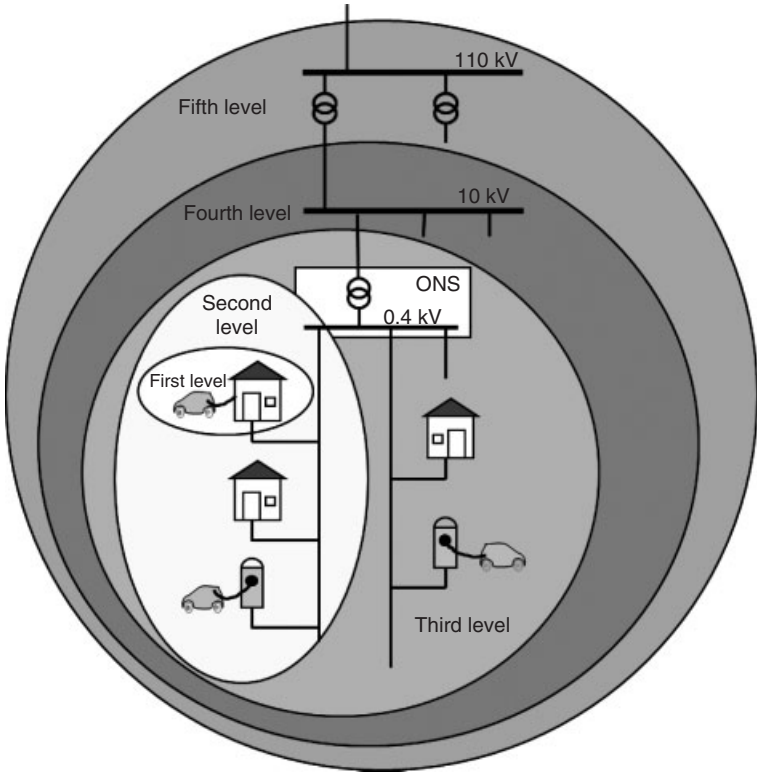


Figure 16. Energy and grid management for various voltage levels.

the overloaded line. If the load includes EVs that are being charged, a PWM control unit can limit the charging output.

3. Once the secondary distribution transformer or high voltage conductor is overloaded, the grid operator will need to be notified, so it can limit charging power or take other steps.
4. This level involves the coordination and optimization of energy and grid management for secondary substations connected to the medium voltage network.
5. Grid and energy management in connection with large power.
6. Plants such as wind farms on the transmission network.

Figure 16 shows the various levels of multilevel grid and energy management.

5 INTEGRATING DC-FAST CHARGE CHARGING STATIONS INTO THE GRID

DC fast charging stations are also being discussed. In most cases, they would be installed at public parking lots. DC charge times currently range from 20 to 30 min. Contemporary battery technologies allow fast charging of up to 80% of battery capacity.

Pros:

- The vehicle only needs a small charger.
- Fewer fast charging stations than AC charging stations.

Cons:

- Stations are only connected to vehicles during charging and then vacated for the next customer to use.
- Service point requires relatively high supply voltage.
- Charge parameters have to be adapted to the battery.
- Relatively heavy plugs and charging cables.

Figure 17 shows one way to integrate DC charging stations into a power grid. This setup only requires an AC/DC converter to the grid. PV systems can feed power directly onto an unregulated DC bus bar via a DC/DC converter. Charge parameters have to be tailored to the vehicle battery using a step-up or step-down converter, as battery voltages and battery charging characteristics will probably not be standardized in the foreseeable future.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

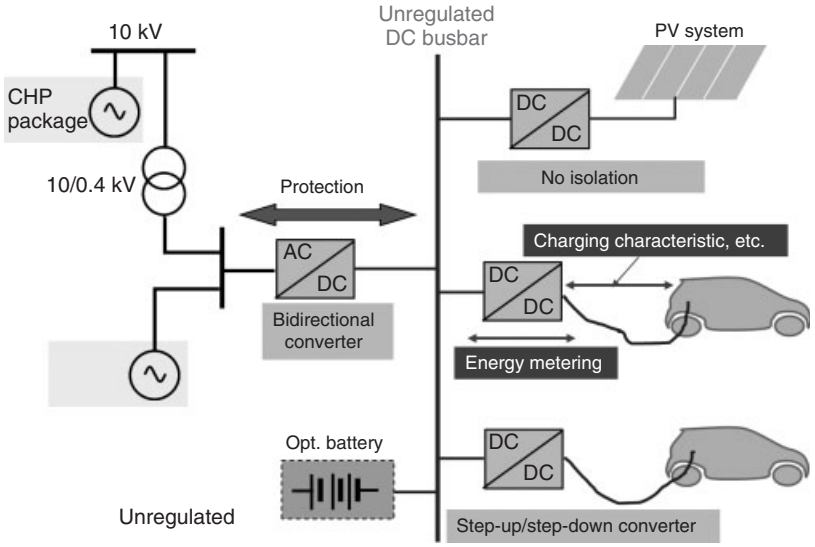


Figure 17. Integration of a DC charging station into the grid.

Stationary battery storage can optionally be integrated in the station, too.

This scenario requires bidirectional communications between the charging station and the vehicle in order to transmit the vehicle's battery charge parameters. The CAN bus is generally used for communications, but some experts have also considered using PLC technology based on IP-based communications.

6 SUMMARY

This chapter described how to extend data communications over the control pilot using standardized procedures for vehicle-to-charging station communications (IEC 61851/SAE J1772), illustrated using the transmission of a customer contract number. It then discussed alternative structures for local and remote communications between the customer, grid operator, and energy supplier. These included not only smart card/RFID solutions but also IP-based communication using PLC technology. The pros and cons of various protocols (SML/XML) were evaluated, with the focus on the SCCs protocol specified by RWE and Daimler.

The discussion touched on the use of live and control pilots to transmit data between the vehicle and the charging station using PLC technology. Several approaches to integrate EVs into the smart grid were examined. Bidirectional charge management, that is, using smart metering technologies to manage vehicle charging, was presented alongside the integration of DC charging stations into the grid.

7 OUTLOOK

A smart grid requires bidirectional communications between customers, grid operators, energy suppliers, and, possibly, service centers. Many different communications structures and approaches are currently being developed and field-tested. Researchers have begun to examine the real-world applicability of various login procedures, transmission media, and data transfer protocols. An ideal first step would be to implement managed vehicle charging. While this can theoretically be done with standardized processes and systems and existing communications structures, it would be restricted to a very small scale. Mass EV deployment would require fully automated charging that can immediately adapt to changes in electricity supply and grid conditions. Bidirectional charge management is essential in these cases, especially if EVs are used to provide ancillary services. If smart metering becomes widely adopted, the communications infrastructure built for smart metering could also be used for smart, fully automated charge management.

REFERENCES

Echelon Corporation (2010a) IEC Electric Meter v3.1 User's Guide, Version: 078-0371-01A, San Jose.

Echelon Corporation (2010b) NES System Software 4.1 Programmer's Guide, Version: 078- 0320-01F, San Jose.

Echelon Corporation (2010c) MEP Device Developer's Guide, Version 078-0372-01G, San Jose.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Forum network technology/network operation in the VDE (2008a) Requirements Specification MUC Multi Utility Communication Version 1.0.

Forum network technology/network operation in the VDE (2008b) Smart Message Language, Version 1.03.

Rehtanz, C., Horenkamp, W., and Ruthe, S. (2011) Integration in Smart Grid erfordert bidirektionale Kommunikation in *etz-Elektrotechnik und Automation*, VDE-Verlag, Offenbach, Germany, pp. 75–77.

Rescorla, E. (1999) *Diffie-Hellman Key Agreement Method*, http://tools.ietf.org/html/rfc2631 (accessed 24 July 2012).

RWE, Daimler, Insys Microelectronics, and Emsycon (2010a) Project e-Mobility: Smart Charge Protocol Specification Part A & B.

RWE, Daimler, Insys Microelectronics, and Emsycon (2010b) Specification—Communication Protocol between Electric Vehicles and Charging Units, Version 0.8.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto139
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Ultracapacitors in Hybrid and Plug-in Electric Vehicles

Andrew Burke
University of California, Davis, CA, USA

1 Introduction 1
2 Ultracapacitor Concepts and Performance 2
3 Ultracapacitor Cost and Life 8
4 Simulations of Hybrid (HEV) and Plug-in (PHEV) Vehicles Using Ultracapacitors 13
5 Summary and Conclusions 22
References 22

1 INTRODUCTION

This chapter is concerned with ultracapacitors (ECCs) and their applications in electric drive vehicles in place of or in combination with batteries. The electric drive vehicles considered are battery-powered electric vehicles and hybrid vehicles (HEVs and PHEVs). The first major section of the chapter deals with ultracapacitor concepts and performance, including a description of the assembly and construction of devices and materials used in them; testing methods; and comparison of the performance of ultracapacitors and batteries. The second major section is concerned with the present and future costs of ultracapacitors, their calendar and cycle life, and a comparison of the cost and cycle life of ultracapacitors and batteries. The final section presents vehicle simulation results for electric and hybrid vehicles that incorporate ultracapacitors in the electric driveline. Special attention is given to sizing the ultracapacitor unit to minimize volume and cost and the control strategies that take advantage of the high efficiency and charge acceptance of ultracapacitors compared to batteries. Simulation results are also given for vehicles using combinations of ultracapacitors and advanced batteries having high energy density (300 Wh/kg).

The most common electrical energy storage device used in vehicles is the battery. Batteries have been the technology of choice for most applications, because they can store large amounts of energy in a relatively small volume and weight and provide suitable levels of power for many applications. Shelf and cycle life have been a problem/concern with most types of batteries, but people have learned to tolerate this shortcoming due to the lack of an alternative. In recent times, the power requirements in a number of applications have increased markedly and have exceeded the capability of batteries of standard design. This has led to the design of special high power, pulse batteries often with the sacrifice of energy density and cycle life. ECCs are being developed as an alternative to pulse batteries. To be an attractive alternative, capacitors must have much higher power and much longer shelf and cycle life than batteries. By "much" is meant at least one order of magnitude higher. ECCs have much lower energy density than batteries and their low energy density is in most cases the factor that determines the feasibility of their use in a particular high power application.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 ULTRACAPACITOR CONCEPTS AND PERFORMANCE

2.1 Cell construction, energy storage, and materials

An ultracapacitor, often referred to as an *ECC*, is an electrical energy storage device that is constructed much like a battery (Figure 1), in that it has two electrodes immersed in an electrolyte with a separator between the electrodes. The electrodes are fabricated from a high surface area, porous material having pore diameters in the nanometer (nm) range. The surface area of the electrode materials used in an ECC is much greater than that used in battery electrodes being $500-2000\,m^2/g$. Charge is stored in the micropores at or near the interface between the solid electrode material and the electrolyte. The energy stored in the capacitor is given by $\frac{1}{2}\,CV^2$, where C is its capacitance (Farads) and V is the voltage between the terminals. The maximum or rated voltage of the capacitor is dependent on the characteristics of the electrolyte used in the device. The charge Q (coulombs) stored in the capacitor is given by CV. The charge and energy stored in the ECC are calculated using the same expressions as for a simple dielectric capacitor. However, calculation of the capacitance of the ECC is much more difficult as it depends on complex phenomena occurring in the micropores of the electrodes.

It is convenient to discuss the mechanisms for energy storage in ECCs in terms of double-layer and pseudocapacitance processes separately. The physics and chemistry of these processes as they apply to ECCs are explained in great detail in Conway (1999) and Béguin and Frąckowiak (2013). In the following sections, the mechanisms are discussed in terms of how they relate to the properties of the electrode materials and electrolyte.

2.1.1 Double-Layer capacitors

Energy is stored in the double-layer capacitor as charge separation in the double layer formed at the interface between the solid electrode material and the liquid electrolyte in the micropores of the electrodes (Figure 1).

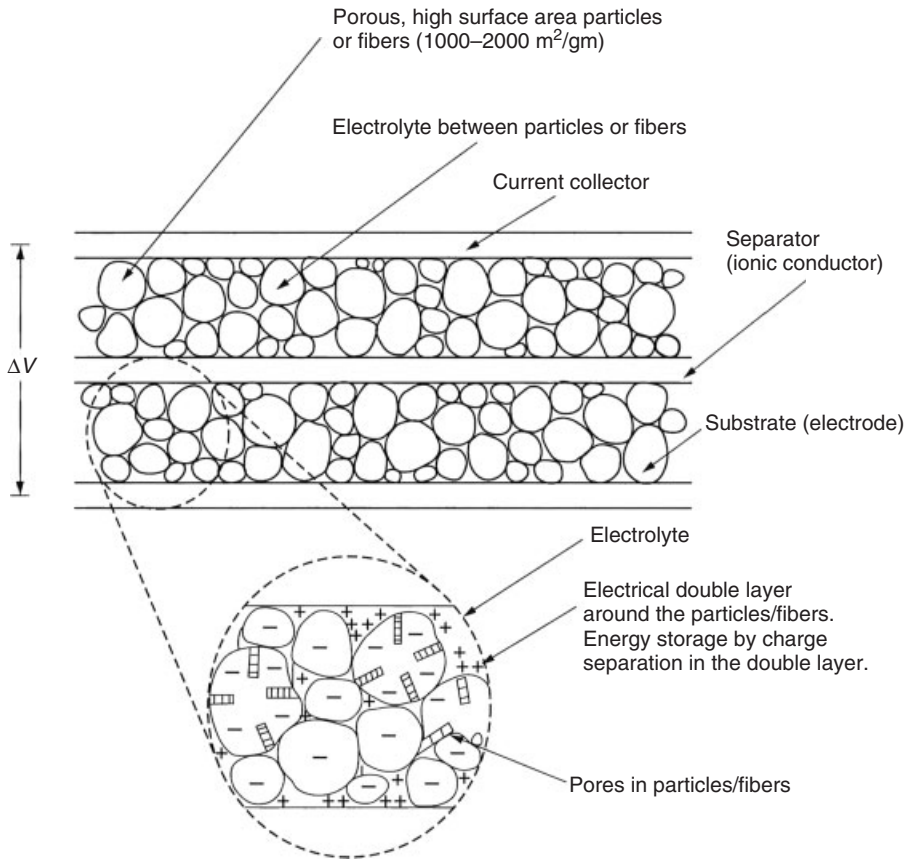


Figure 1. Schematic of an electrochemical capacitor. (Reproduced with permission from A.F. Burke, Lindel's Handbook of Batteries (Burke, 2010) © The McGraw-Hill Companies, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The ions displaced in forming the double layers in the pores are transferred between the electrodes by diffusion through the electrolyte. The capacitance is dependent primarily on the characteristics of the electrode material (surface area and pore size distribution). The specific capacitance of an electrode material can be written as

$$\frac{C}{\mathrm{g}} = \left(\frac{F}{\mathrm{cm}^2}\right)_{\mathrm{act}} \times \left(\frac{\mathrm{cm}^2}{\mathrm{g}}\right)_{\mathrm{act}}$$

where the surface area referred to is the active area in the pores on which the double layer is formed. In simplest terms, the capacitance per unit of active area is given by

$$\left(\frac{F}{\mathrm{cm}^2}\right)_{\mathrm{act}} = \left(\frac{K_{\mathrm{eff}}}{\text{thickness of the double-layer}}\right)_{\mathrm{eff}}$$

where K_{eff} is the dielectric constant of the electrolyte.

Determination of the effective dielectric constant K_{eff} of the electrolyte and the thickness of the double layer formed at the interface is complex and not well understood (Conway, 1999; Béguin and Frąckowiak, 2013). The thickness of the double layer is very small (a fraction of a nanometer in liquid electrolytes) resulting in a high value for the specific capacitance of 15–30 μF/cm^2. For a surface area of 1000 m^2/g, this would result in a potential capacitance of 150–300 F/g of electrode material. As indicated in Table 1, the measured specific capacitances of the carbon materials being used in ultracapacitors are much less than these high values, being in the range of 75–175 F/g for aqueous electrolytes and 40–120 F/g for organic electrolytes. This is the case, because for most carbon materials, a relatively large fraction of the surface area is in pores that cannot be accessed by the ions in the electrolyte. This is especially true for the organic electrolytes for which the size of ions is much larger than in an aqueous electrolyte. Porous carbons for use in ultracapacitors should have a large fraction of their pore volume in pores of diameter 1–5 nm. Materials with small pores (<1 nm) show a large falloff in capacitance at discharge currents >100 mA/cm^2 especially using organic electrolytes. Materials with the larger pore diameters can be discharged at current densities of >500 mA/cm^2 with a minimal decrease in capacitance.

Table 1. Specific capacitance for various electrode materials.

Material	Density (g/cm^3)	Electrolyte	F/g	F/cm^3
Carbon cloth	0.35	KOH	200	70
		Organic	100	35
Activated carbon	0.7	KOH	160	112
		Organic	100	70
Aerogel carbon	0.6	KOH	140	84
Particulate carbon from SiC	0.7	KOH	175	122
		Organic	100	70
Particulate carbon from TiC	0.5	KOH	220	110
		Organic	120	60
Advanced graphitic carbon	0.7	Organic	180	126
Anhydrous RuO_2	2.7	Sulfuric acid	150	405
Hydrous RuO_2	2.0	Sulfuric acid	650	1300
Doped conducting polymer	0.7	Organic	450	315

The cell voltage of the ultracapacitor is dependent on the electrolyte used. For aqueous electrolytes, the cell voltage is about 1 V and for organic electrolytes, the cell voltage is 2.5–3.5 V depending on the carbon being used.

2.1.2 *Electrochemical capacitors utilizing pseudo-capacitance*

For an ideal double-layer capacitor, the charge is transferred into the double layer and there are no Faradaic reactions among the solid material, the ions, and the electrolyte. In that case, the capacitance ($\mathrm{d}Q/\mathrm{d}V$) is a constant and independent of voltage. For devices that utilize pseudo-capacitance, most of the charge is transferred at the surface or in the bulk near the surface of the solid electrode material. Hence, in this case, the interaction between the solid material and the ions/electrolyte involves Faradaic reactions that in most instances can be described as charge transfer reactions. The charge transferred in these reactions is voltage dependent resulting in the pseudo-capacitance ($C = \mathrm{d}Q/\mathrm{d}V$) also being voltage dependent. Three types of electrochemical processes have been utilized in the development of ultracapacitors using pseudo-capacitance. These are (i) surface adsorption of ions from the electrolyte, (ii) redox reactions involving ions from the electrolyte (Long, 2008), and (iii) the doping and undoping of an active conducting polymer material in the electrode (Chandrasekhar, 1999). The first two processes are primarily surface mechanisms and are hence dependent on the surface area of the electrode material. The third process involving the conducting polymer material is more of a bulk process and the specific capacitance of the material is much less dependent on its surface area although relatively high surface area with micropores is required to distribute the ions to and from the electrodes in a cell. In all cases, the electrodes must have high electronic conductivity to distribute and collect the electron current. An understanding of the charge transfer mechanism can be inferred from $C(V)$, which is often determined using cyclic voltammetry (Barsoukov and Macdonald, 2005).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

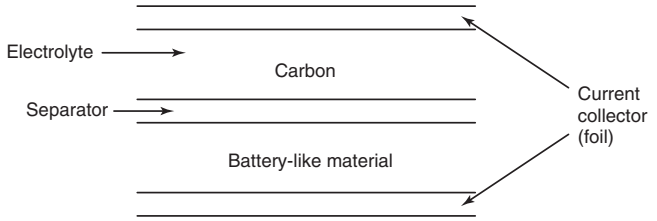


Figure 2. Schematic of a hybrid electrochemical capacitor.

For assessing the characteristics of devices, it is convenient to use the average capacitance (C_{av}) calculated from

$$C_{av} = \frac{Q_{tot}}{V_{tot}}$$

where the Q_{tot} and V_{tot} are the total charge and voltage change for a charge or discharge of the electrode. This permits a determination of the specific capacitance (C_{av}/g) of the material for the electrolyte of interest. As shown in Table 1, the specific capacitance of pseudo-capacitance materials is much higher than that of carbon materials. It is thus expected that the energy density of devices developed using the pseudo-capacitance materials will be higher.

2.1.3 Hybrid capacitors

ECCs can be fabricated with one electrode being of a double-layer (carbon) material and the other electrode being of a battery-like material (Figure 2). Such devices are often referred to as *hybrid capacitors*. Some of the hybrid capacitors developed to date have used metal oxides (e.g.,, lead or nickel oxide) as the battery-like material in the positive electrode and an aqueous electrolyte. Others have used graphite, lithium doped graphite, or lithium titanate oxide in the negative electrode and an organic electrolyte. The energy density of these devices can be significantly higher than that for double-layer capacitors (Table 2), but as shown in Figure 3b, their charge/discharge characteristics (V vs Q) can be very nonideal (nonlinear). Hybrid capacitors can also be assembled using two nonsimilar mixed metal oxides or doped conducting polymer materials (Chandrasekhar, 1999).

2.2 Characterization of ultracapacitor cells

The characterization of the performance of ultracapacitors requires testing of devices over a wide range of test conditions. The test procedures for evaluating ultracapacitors are discussed in detail in Burke and Miller (2010) and FreedomCar (2004). The device characteristics to be determined from the tests are summarized in Table 3. The testing includes constant current, constant power, and pulsed discharges of the devices. The capacitance and resistance of the devices are determined from constant current discharges and the energy density from the constant power tests. The resistance for both charge and discharge pulses are determined from the pulse tests. A typical set of test data for a carbon/carbon double-layer ultracapacitor device is given in Table 4.

Many commercially available and advanced prototype devices have been tested at the University of California, Davis (Burke and Miller, 2009; Burke, 2009). A summary of the test results are given in Table 5. All the devices (except those designated HC) use carbon in both electrodes. The energy densities vary in a relatively narrow range

Table 3. Performance characteristics of electrochemical capacitors.

1	Energy density (Wh/kg vs W/kg)
2	Cell voltage (V) and capacitance (F)
3	Series and parallel resistance (Ω and Ω-cm^2)
4	Power density (W/kg) for a charge/discharge at 95% efficiency
5	Temperature dependence of resistance and capacitance especially at low temperatures (20 °C)
6	Cycle life for full discharge
7	Self-discharge at various voltages and temperatures
8	Calendar life (hours) at fixed voltage and high temperature (40–60 °C)

Table 2. Summary of performance characteristics of hybrid ultracapacitors.

Device	Voltage	Weight (g)	Capacitance (F)	Resistance (Ω)	Wh/kg (400 W/kg)	W/kg at 95% efficiency	W/kg Matched Impedance
Telcordia Lithium titanate/carbon	2.8	43	500	0.011	10.5	466	4173
IMRA Metal oxides	3.35	5.4	56	0.082	13.4	712	6338
UC Davis Carbon/PbO_2	2.25	0.54	13	0.200	9.7	1318	11700
Power systems AC/graphitic C	3.3	0.15	1800	3.0	8.0	486	4320
Futji AC/graphitic C	3.8	0.23	1800	1.5	9.2	1025	10375
JM Energy AC/graphitic C	3.8	0.206	2000	1.9	12.1	1038	9223
Yunasko AC-Faradaic	2.7	0.119	8100	1.0	25.9 (915 W/kg)	2910	15315

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

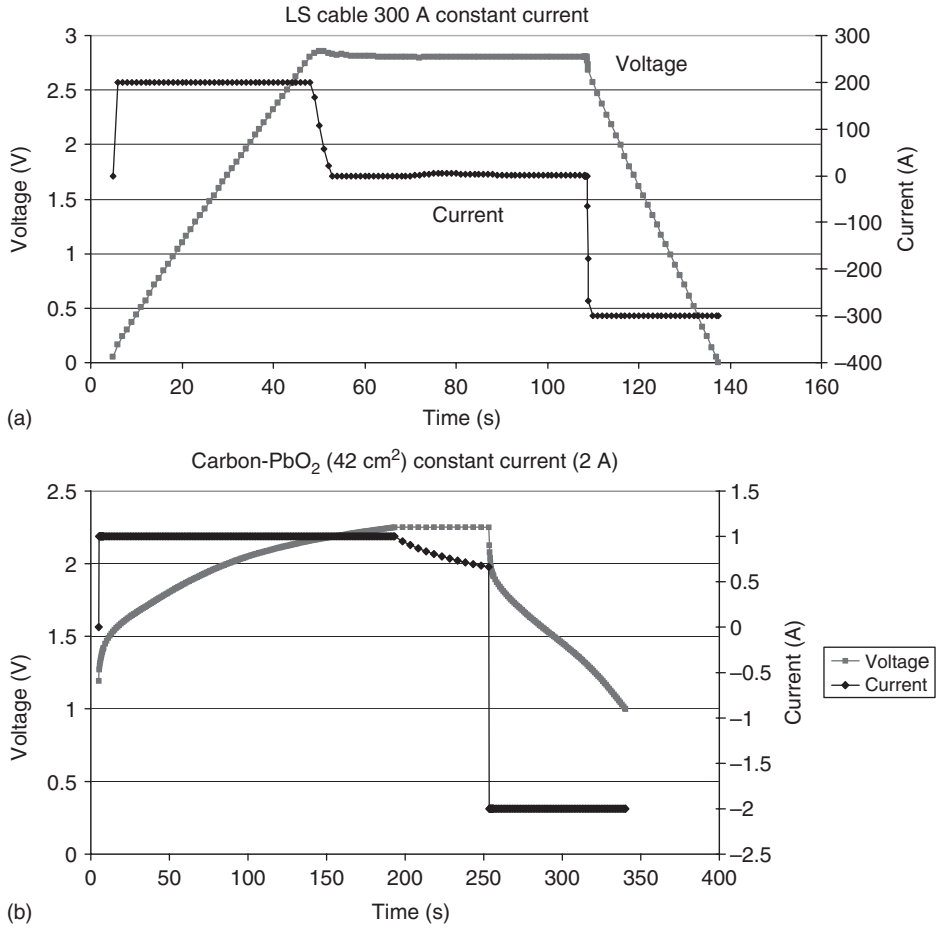


Figure 3. (a) Voltage versus current trace for a double-layer capacitor (C/C). (b) Voltage versus current trace for a hybrid capacitor (C/PbO_2).

(4–6 Wh/kg), but the power capabilities vary over a much larger range (900–10000 W/kg for 95% efficient pulses). Hence, there is much greater uncertainty concerning the power capability of a device than its energy density. As result, it is advisable to test a particular device to determine its power capability before using it in a vehicle design.

Hybrid devices (HC) have a higher energy density than the carbon/carbon devices and are in an early stage of development Figure 3 (Burke, Kershaw, and Miller, 2003; Long, 2008; Ruch, Kotz, and Wokaun, 2009; Naoi *et al*., 2012). The energy densities of the hybrid devices vary between 8 and 30 Wh/kg. In general, the power capabilities $(W/k)_{95\%}$ of the hybrid devices are lower than that of the carbon/carbon devices, which would make it difficult to take full advantage of their increased energy density in some vehicle applications. It seems likely that further development of the hybrid devices will result in improved power capability. Future development of the lithium doped graphite and lithium titanate oxide/carbon devices at cell voltages up to 3.8 V could result in energy densities in the 30–40 Wh/kg.

2.3 Characterization of ultracapacitor modules and packs

The ultracapacitor pack in vehicle applications will consist of many cells in series (80–200) in order to meet the relatively high voltage (200–400 V) of the electric driveline. It is likely that the cells will packaged in modules having a voltage of 16–125 V. The electrical characteristics (C and R) of the modules follow directly from those of the cells. However, the weight and volume of the module will be significantly greater than the sum of the weights and volumes of the cells (see Table 6 for a 48 V module). The weight and volume packaging factors are likely to be in the range of 0.6 and 0.5, respectively, when the thermal and cell management functions are included in the modules (Burke, 2010). However, the energy storage and power capability

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Test data for a carbon/carbon ultracapacitor device

Constant current discharge data 2.7V - 0			
Current (A)	Time (s)	Capacitance (F)	Resistance (m)
10	120.5	450	-----
20	60.3	453	-----
40	30	453	-----
80	14.7	452	1.4
120	9.6	455	1.4
160	7.1	456	1.3

Constant power discharges data 2.7 – 1.35V				
Power (W)	W/kg[a]	Time (s)	Wh	Wh/kg
12.5	219	95.5	0.332	5.82
22	385	54.9	0.336	5.89
41.5	728	28.8	0.332	5.82
80.5	1412	14.6	0.326	5.72
120	2105	9.1	0.303	5.31

[a] weight of device — 57 gm as tested.

of the modules can be calculated from the characteristics of the cells with confidence. Large ultracapacitor packs have been developed for bus and train applications storing up to 5 kWh (Hooks, 2011); therefore, developing packs storing 50–150 Wh for automobile applications should not be difficult.

2.4 Comparisons of the power capabilities of ultracapacitors and batteries

There has been considerable discussion in the literature (Chu and Braatz, 2002; Burke and Miller, 2011; Burke, 2007) comparing the power capability of lithium-ion batteries and ultracapacitors. The conclusions vary from statements that lithium batteries have power capability equal to that of ultracapacitors to statements that ultracapacitors have an order of magnitude higher power capability than lithium batteries. The pulse power capabilities of ultracapacitors and batteries can be calculated from the following expressions which are derived in Burke and Miller (2011):

Ultracapacitors

$$P = \frac{9}{16}\ (1 - \text{Eff})\ \frac{{V_{\text{rated}}}^2}{R}$$

where the pulse is at $V = 3/4 V_{\text{rated}}$

Eff is the pulse efficiency, $\text{Eff} = 1 - (IR/V_{\text{pulse}})$

Batteries

$$P = \text{Eff}\ (1 - \text{Eff})\ \frac{{V_{\text{oc}}}^2}{R}$$

where Eff is the pulse efficiency, $\text{Eff} = V_{\text{pulse}}/V_{\text{oc}}$

Table 5. Summary of ultracapacitor device characteristics (Burke and Miller, 2009; Burke, 2009).

Device	V_{rate}	C (F)	R (mΩ)[c]	RC (s)	Wh/kg[a]	W/kg (95%)[b]	W/kg Matched Impedance	Weight (kg)	Volume (L)
Maxwell	2.7	2885	0.375	1.1	4.2	994	8836	0.55	0.414
Maxwell	2.7	605	0.90	0.55	2.35	1139	9597	0.20	0.211
Vinatech	2.7	336	3.5	1.2	4.5	1085	9656	0.054	0.057
Vinatech	3.0	342	6.6	2.25	5.6	710	6321	0.054	0.057
Ioxus	2.7	3000	0.45	1.4	4.0	828	7364	0.55	0.49
Ioxus	2.7	2000	0.54	1.1	4.0	923	8210	0.37	0.346
Skeleton technology	2.85	350	1.2	0.42	4.0	2714	24200	0.07	0.037
Skeleton technology	3.4	840	0.58	0.49	6.7	3846	34364	0.145	0.097
Yunasko[d]	2.7	510	0.9	0.46	5.0	2919	25962	0.078	0.055
Yunasko[d]	2.75	480	0.25	0.12	4.45	10241	91115	0.060	0.044
Yunasko[d]	2.75	1275	0.11	0.13	4.55	8791	78125	0.22	0.15
Yunasko[d] (HC)	2.7	7200	1.4	10	26	1230	10947	0.119	0.065
Yunasko[d] (HC)	2.7	5200	1.5	7.8	30	3395	30200	0.068	0.038
Ness	2.7	1800	0.55	1.0	3.6	975	8674	0.38	0.277
Ness	2.7	3640	0.30	1.1	4.2	928	8010	0.65	0.514
Ness (cyl.)	2.7	3160	0.4	1.3	4.4	982	8728	0.522	0.379
LS cable	2.8	3200	0.25	0.80	3.7	1400	12400	0.63	0.47
BatScap	2.7	2680	0.20	0.54	4.2	2050	18225	0.50	0.572
JM energy (HC)	3.8	1100	1.15	1.21	10	2450	21880	0.144	0.077
		2300 (plastic case)	0.77	1.6	7.6	1366	12200	0.387	0.214

[a] Energy density at 400 W/kg constant power, $V_{\text{rated}} - 1/2\ V_{\text{rated}}$.

[b] Power based on $P = 9/16 \times (1 - \text{EF}) \times V2/R$, EF = efficiency of discharge.

[c] Steady-state resistance including pore resistance.

[d] All devices except those with [d] are packaged in metal/plastic containers.

Note: Those with [d] are laminated pouched packaged.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 6. Ness Module 48 V 111 F (18 × 2000 F cells)

Constant current discharge data 48.6-24V			
Current (A)[b]	Time (s)	Capacitance (F)	Resistance mΩ[(d)]
50	58.3	119.8	----
100	28.8	119.2	4.4
200	14.1	118.1	4.27
300	9.3	118.7	4.26
400	6.9	119.7	3.9

Constant power discharge data 48.6−24 V					
Power (W)	W/kg[(a)] Based on cell (wgt.)	Time (s)	Wh	Wh/kg Based on cell (wgt.)	Δ T °C
2000	285	52.4	29.1	4.15	0.81
3000	427	34.5	28.8	4.10	0.86
5000	712	20.3	28.2	4.02	0.87
7000	997	14.3	27.8	3.96	1.0
9500	1353	10.3	27.2	3.87	1.13

[a] weight and volume of module − 12.2 kg, 12.5 L
weight and volume of cells − 7.02 kg, 5.53 L
[b] charging done at discharge current up to 200 A and then at 200 A
[c] temperature increase for three discharge cycles
[d] resistance from Δ V at .1 s

R and V_{oc} are taken for the state-of-charge (SOC) of the pulse

In the case of the ultracapacitors, a significant fraction of the energy stored in the device can be in the pulse; for the battery, only a small fraction of the energy is in the pulse.

Detailed comparisons of the power capability of ultracapacitors and batteries taken from Burke and Miller (2011) are shown in Table 7. As indicated in the table, neither of the extreme statements concerning the relative power capabilities of ultracapacitors and batteries is valid in general and comparisons should be made between specific devices for specific applications. Comparisons are often made based on the matched impedance power of the two types of devices. These comparisons indicate that most ultracapacitors have a power capability (W/kg) of 3–6 times that of lithium batteries. However, for vehicle applications, the matched impedance power is not appropriate and should not be the basis for comparison.

For charge sustaining HEV applications, a good basis of comparison is the W/kg at 95% efficiency at the SOC at which the devices will be used in the vehicle. On this basis, there are lithium batteries with the same power capability as some carbon/carbon ultracapacitors, but there are some ultracapacitors with power capability 4–8 times that of the highest power lithium batteries presently available for vehicle applications. In other words, it is not possible to make general statements that are applicable to all devices of either type. The issue is further complicated when one notes that the density of the lithium batteries is about twice that of carbon/carbon ultracapacitors (2.2 g/cm^3 for the batteries and 1.2 g/cm^3 for the ultracapacitors). Hence, on a volume basis W/L at 95% efficiency, the differences between the batteries and ultracapacitors are often quite small. Hence, for HEVs, batteries alone and ultracapacitors alone can be an option with the decision being based on cycle life and cost in addition to relative power capability (Burke and Miller, 2010; Burke, Miller, and Van Gelder, 2007).

For plug-in hybrid and battery electric vehicle applications, the maximum useable power density from the lithium-ion battery can be higher than in an HEV because the peak power of the driveline is used less frequently and

Table 7. Comparisons of the power capabilities of various devices for HEV and PHEVs using the different methods for calculation (Burke and Miller, 2011).

Device	Matched Impedance	USABC Min/max	Efficient Pulse EF = 95%	Efficient Pulse EF = 80%
Lithium batteries 60% SOC				
Kokam NCM 30 Ah	2893	2502	550	1848
Enerdel HEV NCM 15 Ah	5491	4750	1044	3507
Enerdel EV NCM 15 Ah	2988	2584	568	1908
EIG NCM 20 Ah	2688	2325	511	1721
EIG Fe Phosph. 15 Ah	2141	2035	458	1540
Altairnano LiTiO 11 Ah	1841	1750	350	1180
Altairnano LiTiO 3.8 Ah	4613	4385	992	3341
Ultracapacitors $V_0 = 3/4V_{rated}$				
Maxwell 2890 F	8836	4413	994	
Nesscap 3100 F	8730	4360	982	
Batscap 2700 F	18224	9102	2050	
Skeleton technology 840 F	39175	17076	3846	
Yunasko 1275 F	78125	38680	8791	
LSCable 3200 F	12446	4609	1038	
JSR 2000 F	9228	6216	1400	

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

consequently charge/discharge efficiently is less important. For example, a pulse power efficiency of 80% is probably sufficient and most of the lithium batteries have a power capability of >1000 W/kg, 2200 W/L for that efficiency. In addition, the battery is larger (heavier) in these vehicles, and as a result, the power density requirement is less demanding. For PHEVs and EVs, the best application of ultracapacitors is likely to be in combination with batteries designed for high energy density, long cycle life, and low cost. In those cases, as discussed in a later section of the article, the ultracapacitors greatly reduce the peak currents and dynamic stress on the batteries and thus extend their cycle life.

3 ULTRACAPACITOR COST AND LIFE

3.1 Present and future costs

Reducing the present high cost/price of ultracapacitors is a key issue in achieving high market penetration in the future especially of midsize and large devices. There are many applications for which ultracapacitors are presently precluded or even seriously considered because they remain too expensive even though their selling price has decreased significantly in recent years. The cost of manufacture of any product is closely tied to production volume with the cost decreasing rapidly with increased volume up to relatively high production rates. Sales of capacitors in the many millions of units per year are necessary to reduce the unit costs to levels at which the large markets are likely to develop. Semi-automated production facilities presently exist at a number of companies for ultracapacitors of all sizes. In fact, production capabilities exceed sales volumes for most devices and that is the reason the price of devices has decreased markedly in recent years. It is common to speak of the price of devices in terms of cents per Farad (cents/F) or $/Wh stored. It is easier to interpret the price information on the cents/F basis as it does not concern the cell voltage or what fraction of energy stored can be used in a particular application. For example, for a 10 F device, if the price is quoted as 10 cents/F, the device cost would be $1. Similarly, a 2500 F device would cost $25 at 1 cent/F.

The cost to manufacture a carbon/carbon device depends on the material and production costs. At the present time, material costs are high. The cost of carbon suitable for use in ultracapacitors can be as high as $100/kg with the average price being in the $30–$50/kg range. The cost of the electrolyte solvent is also high in the range of $5–10 per liter for both propylene carbonate and acetonitrile. The ionic salts that dissociate in the solvent into the positive and negative ions that move into and out of the double layer in the microporous carbon to store the energy are also expensive being $50–$100/kg. As the analysis of ultracapacitors is straightforward, material costs can be calculated (Burke, Kershaw, and Miller, 2003; Burke, 2010) with good accuracy. The result of a typical costing exercise is shown in Table 8. Note the strong dependency of the cents/F and $/Wh unit costs for the device on the unit material costs. Presently, the price of ultracapacitors is high because both the material and the manufacturing costs are high. With more automated production and reduced material costs, it is anticipated that the price of ECCs in high volume can be in the range of 1–2 cents/F for small devices and 0.25–0.5 cents/F for large devices such as those needed for vehicle applications.

Ultracapacitors cannot compete with batteries in terms of $/Wh, but they can compete in terms of $/kW and $/unit to satisfy a particular vehicle application. Both energy storage technologies must provide the same power and cycle life and sufficient energy (Wh) for the application. The weight of the battery is usually set by the system power requirement and cycle life and not by the minimum energy storage requirement. Satisfying only the minimum energy storage requirement would result in a much smaller, lighter battery than is needed to meet the other requirements. On the other hand, the weight of the ultracapacitor is determined by the minimum energy storage requirement. The power and cycle life requirements are usually easily satisfied. Hence, the ultracapacitor unit can be a more optimum solution for many applications

Table 8. Material costs for a 2.7 V, 3500 F capacitor[a].

Carbon		Electrolyte ACN			Device	Unit Costs			
F/g	gC/dev.	$/kg	$/L	$/kg salt	Total mat. $	$/kg	$/Wh	$/kW	Cents/F
75	187	50	10	125	17.0	29	6.4	29	0.48
120	117	100	10	125	15.5	26	6	26	0.44
75	187	5	2	50	3.6	6.0	1.3	6	0.10
120	117	10	2	50	2.5	4.2	0.93	4.2	0.070

[a]4.5 Wh/kg, 1000 W/kg at 95% efficiency

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 9. Relationships between ultracapacitor and battery costs.

Batteries		Ultracapacitors			
Cost \$/kWh	Cost[a] \$/kW	Related cost cents/F $V_{cap} = 2.6$	Related cost cents/F $V_{cap} = 3.0$	Cost[b] \$/kWh $V_{cap} = 3.0$	Cost \$/kW $V_{cap} = 3.0$
300	30	0.25	0.34	3626	7.3
400	40	0.34	0.45	4800	9.6
500	50	0.42	0.56	5973	11.9
700	70	0.59	0.78	8320	16.6
900	90	0.76	1.0	10667	21.3
1000	100	0.84	1.12	11947	23.9

[a] Battery 100 Wh/kg, 1000 W/kg;
[b] capacitor 5 Wh/kg, 2500 W/kg.

and its weight can be less than that of the battery even though its energy density is less than one-tenth that of the battery.

Consider the example of a charge sustaining hybrid such as the Prius. If the energy stored in the capacitor unit is 125 Wh and that in the battery unit is 1500 Wh, the unit costs of the capacitors and battery are related by

$$\left(\frac{\$}{\text{Wh}}\right)_{\text{cap}} = 0.012 \left(\frac{\$}{\text{kWh}}\right)_{\text{bat}}$$

The corresponding capacitor costs in terms of cents/F and \$/kWh are given by

$$\left(\frac{\text{cents}}{\text{F}}\right)_{\text{cap}} = 0.125 \times 10^{-3} \times \left(\frac{\$}{\text{kWh}}\right)_{\text{bat}} \times V^2{}_{\text{cap}}$$

$$\left(\frac{\$}{\text{kWh}}\right)_{\text{cap}} = 9.6 \times 10^4 \left(\frac{\text{cents}}{\text{F}}\right)_{\text{cap}} / V_{\text{r}}{}^2$$

Table 9 shows the ultracapacitor costs calculated using the above equations for a range of battery costs (\$/kWh).

The results indicate that for the charge sustaining hybrid application, ultracapacitor costs of 0.3–1.0 cent/F will be competitive with lithium battery costs of \$300–\$1000/kWh. Present (2010) ultracapacitor costs are about 1 cent/F and future costs are expected to be significantly lower. Note also that the \$/kW costs of the capacitor units are about one-fourth those of the batteries.

3.2 Factors affecting calendar and cycle life

One of the advantages of ultracapacitors relative to batteries is their long cycle life, which may be as long as one million (10^6) deep discharge cycles for cells at room temperature and less than rated cell voltage. The calendar life or the time over which the cycling takes place is also highly dependent on cell voltage and temperature. In most life testing of ultracapacitors, it is assumed that cycling (even dynamic cycling at high rates) is not the principal cause of degradation and the life testing is done with the devices held at constant voltage and temperature for long periods of time (Miller, Butler, and Goltser, 2006; Miller and Butler, 2006). These are often referred to as *float* tests. Analysis of ultracapacitor life (years) for these conditions is considered in this section.

Unfortunately, estimating the lifetime (years) of a pack of capacitor cells in a particular application is much more complicated than simply testing cells at room temperature. One primary reason for this difficulty is that in most applications (vehicles in particular), many cells are connected in series to attain the required system voltage. In addition, the temperature varies across the pack even with cooling and the cells spend some time at voltages approaching their maximum rated voltage even with cell balancing circuitry. These factors significantly reduce the pack lifetime from that expected based on single cell testing.

The estimation of cell and pack lifetime is considered in detail analytically in Miller, Butler, and Goltser (2006); Miller and Butler (2006). The analysis is based on the assumption that the cell lifetime statistics can be expressed in terms of a Weibull distribution.

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right], F = \text{the fraction of cell failures}$$

where t = time, α = characteristic life, and β = shape factor

Cell lifetime testing must be done using test conditions that result in the same aging mechanisms (cell failures) as in the application of interest. For ultracapacitors, the testing could be either cycling at specified power levels, voltage ranges, and temperatures or float at specified temperatures and voltages. As noted previously, for vehicle applications, it is likely that lifetime data (Miller, Butler, and Goltser, 2006; Miller and Butler, 2006) from the cell float tests are the most appropriate. Such data can be curve fit to obtain values for α and β for the single cells. Both parameters can vary over wide ranges depending on the cell voltage and temperature of the tests. In the case of activated carbon cells, the characteristic time (α) can vary from > 10,000 h at room temperature and 2.3 V to < 500 h at 60°C and 2.8 V. The shape factor (β) can vary between about 4 for room temperature and 2.3 V to about 15 for 60°C and 2.8 V. A low value of β means that cell failures occur gradually over time, whereas a high value of β indicates that nearly all the cells fail together over a short period of time.

The lifetime characteristics of a pack of ultracapacitors depend strongly on the number (M) of the cells connected in series. Assuming the pack failure statistics are also Weibull, the shape factor of the pack is the same as that of the cells

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

in the pack, and each cell failure is independent of other cells, the pack failure function F_{pack} can be written as

$$F_{\text{pack}} = \left\{1 - \exp\left[-\left(\frac{t}{\alpha_c}\right)^{\beta}\right]\right\}^{M}$$

$$F_{\text{pack}} = M\left(\frac{t}{\alpha_c}\right)^{\beta} \text{ for a small fraction of failed cells}$$

If we express F_{pack} as

$$F_{\text{pack}} = 1 - \exp\left[-\left(\frac{t}{\alpha_p}\right)^{\beta}\right], \alpha_p$$

$$= \text{characteristic time of the pack}$$

One finds

$$\alpha_p = \frac{\alpha_c}{M^{1/\beta}}$$

Hence, the characteristic time of the pack is much less than that of the cells. For example, for a pack with 200 cells in series, the characteristic time is reduced by a factor of 3.76 if β is 4 and by a factor of 1.42 if β is 15. For most packs, the characteristic time would likely be 1/3 to 1/2 that of the cells.

The next factor to consider in the estimation of the lifetime of a pack is the effect of nonuniformities in voltage and temperature of the cells. Even with cooling and cell balancing circuits, there will be nonuniformities in the pack. This will especially be the case for applications in which the pack provides dynamic, high power as in vehicles. The effects of nonuniformities are considered analytically in Miller, Butler, and Goltser (2006); Miller and Butler (2006). The analysis is based on the following assumptions concerning the effect of temperature and voltage on the failure of single cells: in the case of temperature, a 10°C decrease in temperature doubles the cell life; in the case of voltage, a 0.1 V decrease in voltage doubles the cell life. The analytical forms of the temperature (T °K) and voltage (V) effects on the cell characteristic life time τ are

$$\tau = a\ \exp\left(\frac{b}{T}\right), \frac{\tau}{\tau_0} = \exp\left[-6155\ \left(\frac{T - T_0}{{T_0}^2}\right)\right]$$

$$\tau = A\ \exp(-BV), \frac{\tau}{\tau_0} = \exp[-6.93(V - V_0)]$$

These relationships project a reduction in characteristic time of about $1/\sqrt{2}$ for variations of 0.5°C in temperature and 0.05 V in voltage. This corresponds to a maximum temperature difference of 10°C with an average difference of 5°C for an average temperature of 30°C. Similarly, the maximum voltage difference is 0.1 V with an average difference of 0.05 V.

Applying these relationships to a particular application requires detailed knowledge of the application and specification of the cell operating conditions and tolerable cell failure rates. Consider the following example of a capacitor pack in a hybrid passenger car. The pack failure requirements are 98% reliability for 5 years and 80% reliability for 12 years. The corresponding mileage values are 50,000 and 120,000 miles. If the average speed is 25 mph, the operating time requirements are 2000 and 4800 h for the pack. The pack voltage is 300 V with 125 cells in series. The question to be answered is what cell characteristic time (hours) is needed for the cells if the shape factor of their failure distribution is 10. For a Weibull distribution, the relationship among the proportion (P) of failures, time to failure (t_F), and distribution characteristics (α_{pack}, β) is

$$t_F = \alpha_{\text{pack}}[-\ln(1 - P)]^{1/\beta}$$

For the times to failure (2000 and 4800 h), the corresponding α_{pack} values are 2954 and 5581 h. Taking the maximum value of 5581, the cell α_c is 9041 h for $M = 125$ and $\beta = 10$. Assuming the average temperature variation is 5°C and the average voltage variation is 0.05 V/cell in the pack from the base values of 30°C and 2.5 V/cell, the cell statistics on a float type test at the base values of temperature and voltage should exhibit a α_c of 18082 and a $\beta = 10$. This corresponds to a float time of about 2 years at 30°C and 2.5 V/cell.

Next consider cell balancing to reduce cell-to-cell variations in voltage (Burke and Miller, 2005; Jung, 2002). The capacitor pack will consist of many cells (20–200) connected in series. If each of the cells were identical having exactly the same capacitance, series resistance, and parallel resistance, the voltage of all the cells would be same at all times and be equal to the average cell voltage (V/number of cells). There would not be concern that the voltage of some of the cells would exceed at times a maximum specified by the cell manufacturer. This maximum voltage is often referred as the working (continuous) voltage limit of the cell and experiencing voltages above that limit for times longer than a few seconds can significantly reduce the life time of the cell. This is not a safety issue per se as the cells can withstand considerably higher voltages without the pressure relief vent being activated. Limiting the maximum cell voltage in the pack is then primarily an issue of cell life, the maximum useable energy of the unit, and system efficiency in high power cycling. Controlling the variability of the voltage between the cells is referred to as cell balancing. The objectives of the cell balancing are primarily to minimize the differences

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

in the cell voltages and to restrict the maximum voltage of any cell to less than the continuous working voltage. It is desirable that the cell balancing system provide a means of monitoring the voltages of the cells in order to ensure that it is maintaining the voltages in the proper range. Monitoring the voltage and temperature of the cells is needed to ensure long cycle life (10 years or longer).

The complexity of the balancing approach required and the absolute need for cell balancing depends to a large extent on the magnitude of the differences of the characteristics of the cells (capacitance and resistances) (Kotz *et al*., 2007). These differences are dependent on the uniformity of the materials used in the cells and quality control in the manufacture of the cells. The specification of ±15% or 20% variability for capacitance and resistance often seen on spec sheets is grossly greater than tolerable for high voltage strings of ultracapacitors. Experience (available test data (Rafik *et al*., 2006)) has shown that the variability in the capacitance can be quite small for relatively large cells with the variation between the maximum and minimum being 1–1.5% and the standard deviation of the distribution of the capacitance being about 0.5%. Experience has shown larger percentage variations in the resistance of low resistance (fraction of a mΩ) cells, but the standard deviation is close to the accuracy of the resistance measurement (0.01 mΩ). Experience for the self-discharge voltage of a batch of cells after 1 h has shown a maximum variation of <5 mV and a standard deviation of <1 mV (Rafik *et al*., 2006). Hence, present manufacturing practice for ultracapacitor cells seems to be reasonably good and improving so that the prospects for use of relatively simple approaches to cell balancing seem to be better than was the case several years ago.

Relating the variability in cell characteristics to the cell-to-cell voltage variability for complex discharge/charge cycles is not a simple matter (Miller, Butler, and Goltser, 2006; Miller and Butler, 2006). Of primary interest are the magnitude of the maximum cell variations, especially those in the direction of high voltage, and whether the magnitudes of the voltage variations tend to increase for long-term cycling without cell balancing and/or equalization. Available data (Kotz *et al*., 2007) seem to indicate that the magnitude of cell-to-cell voltage variations do not increase with cycling, but rather seem to stabilize even without cell balancing. This seems to be the case even as the cells age.

Variations in cell capacitance will lead to the largest cell-to-cell voltage variations, but those variations will not tend to increase with extended charge/discharge cycling because the capacitance differences have self-compensating effects for charge and discharge pulses. Variations in cell resistances (series and parallel) lead to smaller cell-to-cell voltage variations, and those variations also tend to be self-compensating for cycles that consist of sequential charge/discharge pulses. Variations in self-discharge (parallel resistance) can lead to significant differences in cell voltages during extended rest periods without cell balancing. This could lead to large differences in cell voltages when charge/discharge cycling is resumed. For vehicle applications involving charge/discharge cycling, the effect of variations in cell characteristics on cell-to-cell voltage variations can be determined from cycle testing of the capacitor unit for relatively short times. The cycling time should be long enough to reach thermal equilibrium.

All the ultracapacitor manufacturers are developing cell balancing circuits for use with long series strings of their cells. Modules are supplied with balancing circuits installed. There are a number of approaches to cell balancing. These approaches (Figure 4) range in complexity from placing a simple resistor in parallel with each cell to connecting an active circuit with a power source to charge or discharge each cell separately as needed to reduce cell-to-cell voltage variations. The simple approaches are likely to be adequate for most applications if the variability of the cell characteristics is relatively small (a few percent); the more complex approaches will be needed if the cell variability is relatively large due to poor manufacturing quality control, large temperature gradients in the capacitor pack, and/or the effects of aging. It should be recognized at the outset that the currents involved with the balancing of the cells are small being 1A or less so that they are in most applications much smaller than the pulse currents in/out of the cells. This means that during the high power portions of the charge/discharge, the effects of the cell balancing on the cell voltages is small. Cell balancing has the largest effect during periods of low power demand or rest. Hence, regardless of the cell balancing approach used, the variability of the characteristics of the cells must be relatively small if cell-to-cell voltage variations are to be tolerable.

As discussed previously, the life time (time to significant failures) in the cells decreases by about a factor of 2 for a 0.1 V increase in cell voltage if that voltage increase is experienced in a significant fraction of the time. In general, the life time of the cells decreases markedly as the maximum set voltage for the cell balancing circuit is increased beyond about 2.4 V/cell for the present (2009) cell technology. Hence, it appears that the cell voltages should be limited to an average of about 2.4 V/cell with a cell-to-cell variation of < 0.1 V/cell. This should result in long float and cycle life times of >500,000 cycles and 15,000 h on float at near room temperature (25–30°C).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

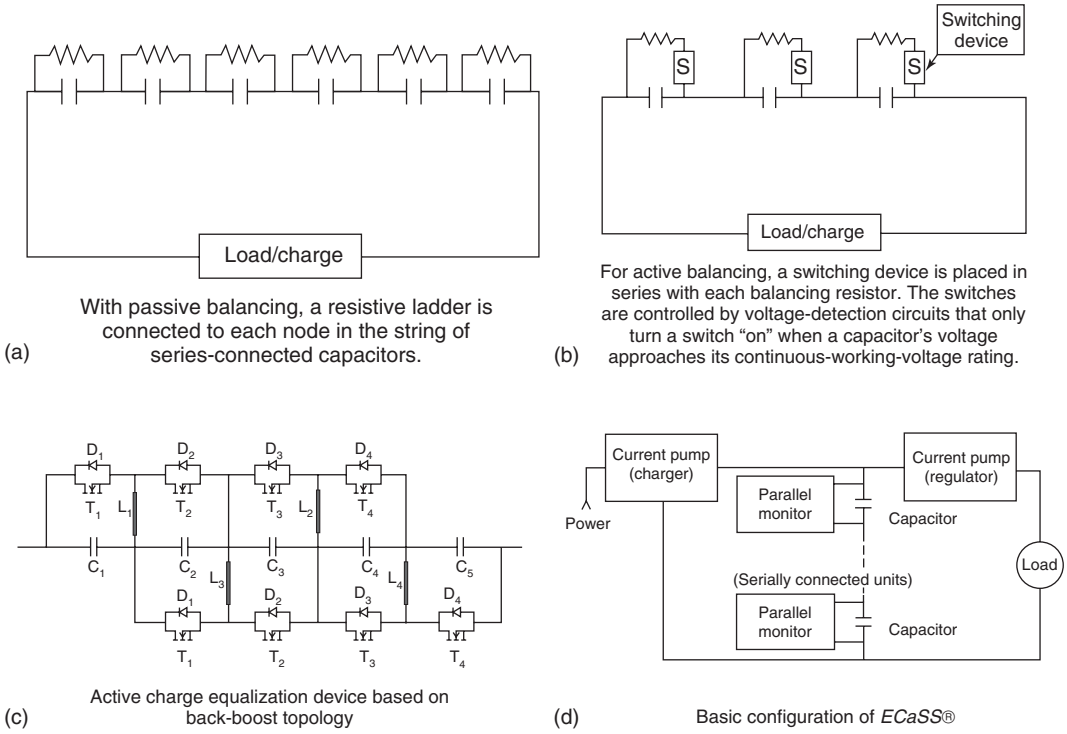


Figure 4. Cell balancing circuits.

3.3 Comparisons of the characteristics of ultracapacitors and batteries

In general terms, ultracapacitors have long life and high first cost and batteries have lower first cost and shorter life. Neither ultracapacitors nor lithium batteries are mature technologies in terms of high volume production and consequently cost. It is expected that the cost of both ultracapacitors and lithium batteries will decrease significantly in the next 5–10 years if they are used in mass marketed vehicles. Comparisons of the cost and life characteristics of ultracapacitors and lithium batteries for HEVs and PHEVs are shown in Table 10. The physical and performance characteristics of both technologies reflect their status in 2010. In the case of the ultracapacitors, the values shown in Table 10 follow from the results

Table 10. Comparisons of the cost and life characteristics of carbon/carbon ultracapacitors and lithium batteries (2010).

Carbon/Carbon Ultracapacitor for a HEV							
6 Wh/kg, 3000 W/kg, 17 kg	Cents/F	$/unit	$/kW	$/kg	Calendar life (Years)	Cycle life (Shallow cycles)	Cycle life (Deep cycles)
100 Wh 50 kW	1.0	1320	26	78	15	Unlimited	$0.5–1.0 \times 10^6$
	0.5	660	13	39	15	Unlimited	$0.5–1.0 \times 10^6$
Lithium batteries for a HEV							
80 Wh/kg, 1200 W/kg 12.5 kg	$/kWh	$/unit	$/kW	$/kg	Calendar life (Years)	Cycle life (Shallow cycles)	
1000 Wh 50 kW	1000	1000	20	80	10	300×10^3	
	600	600	12	48	10	300×10^3	
Lithium batteries for a PHEV							
130 Wh/kg, 600 W/kg, 77 kg	$/kWh	$/unit	$/kW	$/kg	Calendar life (Years)	Cycle life (Shallow cycles)	Cycle life (Deep cycles)
10 kWh 50 kW	700	7000	140	91	10	300×10^3	$2–3 \times 10^3$
	400	4000	80	52	10	300×10^3	$2–3 \times 10^3$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

shown in Tables 4, 8, and 9. In the case of the lithium batteries, the values in Table 10 are based on the results in Nelson (2008); Nelson, Santini, and Barnes (2009). The performance of ultracapacitors in HEVs is discussed in the next section. The information in Table 10 indicate that first costs of carbon/carbon ultracapacitors and lithium batteries for use in HEVs will be the same if the cost of the capacitors is 0.5 cents/F and the cost of the batteries is $600/kWh.

However, it is likely that the cycle life of the ultracapacitors would exceed significantly that of the lithium batteries. The weight and size of the lithium batteries would be smaller than that of the carbon/carbon ultracapacitors, but that of hybrid capacitors could be smaller than the lithium batteries if their cycle life proves to be sufficiently long for the HEV application. Information for lithium batteries for use in a PHEV are also given in Table 10. The batteries in a PHEV must have good cycle life for both deep cycle discharges and shallow cycling in the HEV mode of operation of the vehicle. The PHEV lithium batteries will be more expensive than the HEV batteries simply because they store much more energy. Ultracapacitors will have application in PHEVs only if the lithium or other advanced batteries do not have sufficient power capability to meet the power (kW) and energy storage (kWh) requirements of the vehicle without oversizing the battery. The application of ultracapacitors with advanced batteries having high energy density (>300 Wh/kg) is discussed in the next section.

4 SIMULATIONS OF HYBRID (HEV) AND PLUG-IN (PHEV) VEHICLES USING ULTRACAPACITORS

4.1 Powertrains and control strategies using ultracapacitors

Schematics of hybrid powertrains using ultracapacitors are shown in Figure 5. When the ultracapacitor is used alone, there is no need for special electronics if the motor electronics can handle the increased voltage swing of the ultracapacitors. When the ultracapacitors are used in combination with batteries, there is a need for additional electronics to control the power from either the battery or the ultracapacitors.

When ultracapacitors are used alone as the energy storage unit in a charge sustaining hybrid (HEV), the objective of the control strategy is to permit the engine to operate near its maximum efficiency. As shown in Burke, Zhao, and Van Gelder (2009); Burke (2007); Burke and Zhao (2010), this can be done by operating the hybrid vehicle on the electric drive when the power demand is less than the power capability of the electric motor; when the vehicle power demand exceeds that of the electric motor, the engine is operated to meet the vehicle power demand plus to provide the power to recharge the ultracapacitor unit. In this mode, the electric machine is used as a generator and the engine operating point is selected along its maximum efficiency line (torque vs RPM). The recharging power is limited by the power of the electric machine because

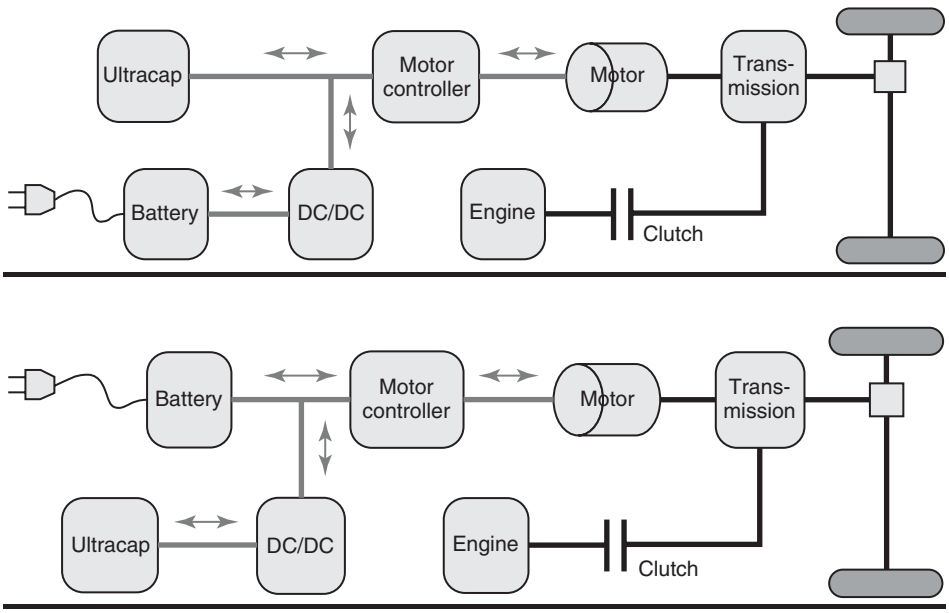


Figure 5. Schematics of powertrains using ultracapacitors and batteries.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ultracapacitors can have pulse power efficiency >95% for W/kg values of over 2000 W/kg (Table 7). This control strategy is referred to as the *sawtooth strategy* because a plot of the ultracapacitor SOC has the form of a saw blade. This strategy can result in frequent on–off operation of the engine and its effect on vehicle drivability needs to be evaluated in actual vehicles.

4.2 Hybrid-Electric vehicles (HEV) using ultracapacitors in place of batteries

Simulations of midsize passenger cars using the ultracapacitors in micro-hybrid and charge sustaining hybrid (HEV) powertrain designs are discussed in this section. The simulations were performed using the Advisor vehicle simulation program modified with special routines at UC Davis (Burke, Zhao, and Van Gelder, 2009; Burke, 2007; Burke and Zhao, 2010). All the powertrains were in the same vehicle having the following characteristics: test weight 1660 kg, $C_d = 0.3$, $A_F = 2.25\,m^2$, and RRCF = 0.009. The engine map used in the simulations was for a Ford Focus 2L, four-cylinder engine. The rated engine power was 120 kW for the conventional internal combustion engine (ICE) vehicle and the micro-hybrid and 110 kW for the charge sustaining hybrids. All the hybrids use the single-shaft arrangement similar to the Honda Civic hybrid. The same electric motor map was used for all the vehicle simulations. The simulation results are summarized in Table 11 for a conventional ICE vehicle and each of the hybrid designs. It is clear from Table 11 that large improvements in fuel usage are predicted for all the hybrid powertrains using ultracapacitors for energy storage. The simulation results will be discussed separately for each hybrid design.

The results for the micro-hybrids indicate that significant improvements (10–25%) in fuel economy can be achieved using a small electric motor (4 kW) and small ultracapacitor units (5–10 kg of cells). In the micro-hybrid designs, the rated engine power used was the same as that in the conventional ICE vehicle in order that the performance of the hybrid vehicle when the energy storage in the ultracapacitors is depleted would be the same as the conventional vehicle. The ultracapacitors were used to improve fuel economy with only a minimal change in vehicle acceleration performance.

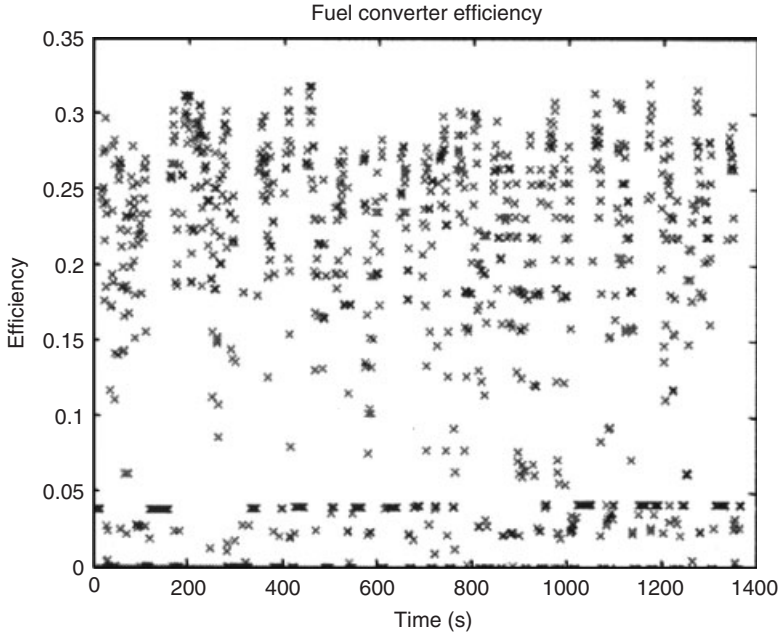
The fuel economy simulation results for charge sustaining hybrids are also shown in Table 11 using carbon/carbon and hybrid ultracapacitors. The fuel economy improvements range from 70% on the Federal Urban Drive Schedule (FUDS) to 22% on the US06 driving cycles. The prime advantage of the high power electric driveline and the larger energy storage possible with the hybrid ultracapacitors is that the larger fuel economy improvements can be sustained over a wide range of driving conditions. All the advanced ultracapacitors have high power capability and thus can be used with the higher power electric motor used in charge sustaining hybrid drivelines. The hybrid ultracapacitor technologies give the vehicle designer more latitude in powertrain design and in the selection of the control strategies for on/off operation of the engine even if the improvement in

Table 11. Mild-HEV and micro-HEV advisor simulation results using carbon/carbon and hybrid ultracapacitors and a midsize passenger car (weight = 1660 kg, $C_d = 0.3$, $A_f = 2.2\,m^2$, fr = 0.009).

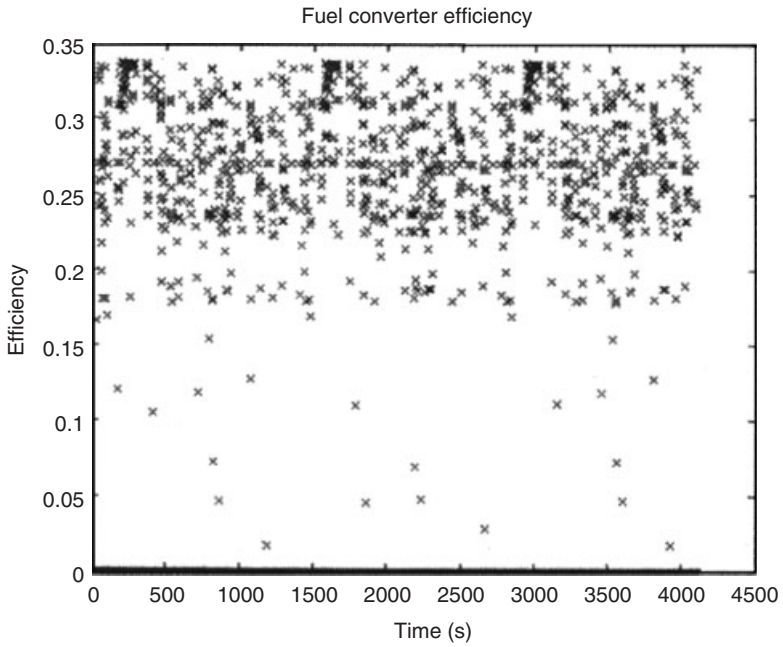
Energy Storage System	Weight of the Ultracapacitors (kg)	Energy Stored	mpg FUDS[a]	mpg FEDHW	mpg US06
Mild HEV		20 kW electric motor			
Yunasko hybrid	10	300 Wh	45.1	48.0	33.2
	5	150 Wh	43.6	46.2	33.2
JM Energy hybrid	10	100 Wh	43.6	46.2	33.0
Yunasko C/C	21	100 Wh	45.4	47.7	34.4
Maxwell C/C	25	100 Wh	44.3	47.1	33.6
ICE Ford Focus engine 120 kW			25.5	36.8	26.8
Fuel economy improvement			72%	25%	22%
Micro start stop HEV	Ultracapacitor with a lead-acid battery	4 kW electric motor			
Yunasko hybrid	5 kg	150 Wh	32.4	41.4	28.9
	3 kg	75 Wh	32.1	41.2	28.5
Yunasko C/C	11 kg	50 Wh	32.2	41.2	28.6
Maxwell C/C	12 kg	50 Wh	32.3	41.3	28.3
Fuel economy improvement			26%	12%	7%

[a] L/100 km = 238/mpg for gasoline engines.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) Engine operating efficiency for the ICE vehicle-average engine efficiency



(b) Engine operating efficiency for the micro-hybrid-average engine efficiency

Figure 6. A comparison of engine efficiencies for a conventional ICE vehicle and a hybrid on the FUDS cycle using carbon/carbon ultracapacitor. (Reproduced with permission from A.F. Burke, Lindel's Handbook of Batteries (Burke, 2010) © The McGraw-Hill Companies, Inc.)

fuel economy is about the same as using carbon/carbon devices.

The control strategy used was the "sawtooth" strategy discussed in the previous paragraph. As shown in Figure 6, this results in a large improvement in average engine efficiency from 19% in the ICE vehicle to 30% in the hybrid vehicle even though the electric motor had a peak power of only 6 kW.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 12. Characteristics of present and future battery cell technologies for EVs and PHEVs.

Chemistry Anode/ Cathode	Cell Voltage Max/nom.	Ah	Weight (kg)	R (mΩ)	EV (Wh/kg)	HEV (W/kg) 95%	EV (W/kg) 75%	Cycle Life (deep)	Thermal Stability
Present technology batteries									
Graphite/$NiCoMnO_2$	4.2/3.6	30	.787	1.5	140	521	2060	2000–3000	Fairly stable
Graphite/Mn spinel	4.0/3.6	15	0.424	2.7	127	540	2120	1500	Fairly stable
Future technology batteries									
Graphite/composite MnO_2	4/3.6	5	0.09	20	200	250	1350	—	Fairly stable
Silicone carbon composites/ composite MnO_2	4/3.6	20	0.24	4.5	295	621	2250	—	Fairly stable
Rechargeable Zinc-air	1.3/1.15	20	60	6.6	385	156	616	—	Very stable
Present Technology power devices									
Supercapacitor Activated carbon/activated carbon	2.7/1.35	500F	.068	1.3	5.5	2320	11600	500 K	Very stable
Power battery Lithium titanate oxide	2.8/2.5	4	.23	1.15	40	1310	5170	20–50 K	Very stable

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Additional computer simulations were made for higher motor power (up to 12 kW) and larger ultracapacitor energy storage (up to 50 Wh). It was found that the improvements in fuel economy were only marginally greater. Using a motor power of 3 kW reduced the fuel economy improvement on the FUDS by more than 50%.

4.3 Plug-in hybrid-electric vehicles (PHEV) using ultracapacitors in combination with atteries

For PHEVs, ultracapacitors could be used in combination with batteries if the power capability of the batteries was insufficient to meet the power requirement of the vehicle. Powertrains of this type are shown in Figure 5. The control strategy in the charge depleting mode is to limit the power from the battery to the average power needed by the vehicle with the ultracapacitor providing the additional power during vehicle accelerations (Miller, Bohn, and Deshpande, 2008; Dixon, Nakashima, and Ortuzar, 2010). The ultracapacitors also accept all the energy recovered during regenerative braking. If engine operation is needed, the "sawtooth" strategy is used with the ultracapacitors being recharged using engine power. In the charge sustaining mode of operation of the PHEV, the electric drive is operated using only the ultracapacitors like that previously described for the HEV.

A detailed study of plug-in hybrids using advanced batteries is presented in Burke and Miller (2010); Burke (2011). The characteristics of the advanced batteries used in the simulations are given in Table 12.

The battery and ultracapacitor units used in the simulations are given in Table 13. Simulations were performed with the batteries alone and in combination with the ultracapacitors. The nominal energy storage unit voltage was 240 V (approximately) in all cases with the maximum currents limited to about 300 A even in the cases of the batteries alone. In all cases, the batteries were depleted to 30% SOC from 100% SOC.

All the PHEV simulations were performed using the following vehicle inputs:

$C_D = 0.27$, $A_F = 2.2\,m^2$, $f_r = 0.008$, test weight = 1650 kg
Engine: Honda 1.3L, iVTEC engine map, scaled to 90 kW
Electric motor: Honda hybrid Civic AC PM 2006 efficiency map, scaled to 70 kW
DC/DC inverter: constant efficiency 0.96
Transmission: 5-speed manual (3.11, 2.11, 1.55, 1.0, 0.71, FD = 3.95), automatically shifted in the model

4.3.1 PHEV fuel and electricity use characteristics

The simulation results are summarized in Tables 14 and 15 showing results for the charge depleting and charge sustaining modes of the PHEV.

With the batteries in combination with the ultracapacitors, the PHEVs were able to operate in the all-electric mode until the battery SOC = 30% on the FUDS and HW highway driving cycles. In all cases for the US06 driving cycle, the vehicle had blended operation (engine and electric drive both needed) in the charge depleting mode. The use of the ultracapacitors with the batteries permits all-electric operation of the vehicle over a wide range of driving conditions with higher Wh/mi for all the driving cycles. Hence, in the charge depleting mode, the fuel economy (mpg) is higher by 50–100% using the ultracapacitors for all the batteries. The fuel economy in the charge sustaining mode is also higher for all the driving cycles using the ultracapacitors. The acceleration times of the vehicle were lower using the ultracapacitors than for the batteries alone. With the ultracapacitors, the acceleration times were 2.7 s for 0–30 mph and 6.9 s for 0–60 mph. For the batteries alone, the acceleration times varied somewhat with the battery used ranging from 2.9–3.2 s for 0–30 mph and 8.6–9.8 s for 0–60 mph. Hence, in all respects, vehicle performance was improved using the ultracapacitors for all the batteries studied.

Table 13. The advanced battery and ultracapacitor units used in the simulations.

Lithium-ion batteries				
$NiCoMnO_2$	30 kg	20 Ah cells	60 in series	3.5 kWh usable
Composite MnO_2	32 kg	40 Ah cells	60 in series	4.6 kWh usable
Silicone composite	22 kg	30 Ah cells	60 in series	4.6 kWh usable
Zinc-air batteries				
Zn-air	32 kg	60 Ah cells	180 in series	9.5 kWh usable
Carbon/carbon ultracapacitors				
Symmetric C/C	20 kg	1350F cells	110 in series	100 Wh usable

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 14. Simulation results for the advanced batteries with ultracapacitors.

Battery Type[a]	Cycle	Range (mi)	kW max. Control	kW max Batteries	Efficiency Batteries	kW max. Cap.	Efficiency Cap.	Wh/mi Batteries	Mode of operation	mpg 20 mi	mpg 40 mi	Charege Sustaining HEV (mpg)
Composite MnO_2	FUD	22	40	18	0.94	40	0.97	215	AE	None	97	52.8
32 kgbat	HW	20	45	18	0.91	45	0.96	227	AE	None	109	56.3
20 kgcap	US06	30	68	21	0.91	68	0.94	180	Blended	71.9	56	38.3
Si Carbon/ composite MnO_2	FUD	20	40	18	0.94	40	0.97	220	AE	None	99	52.8
22 kgbat	HW	20	45	19	0.91	45	0.97	225	AE	None	110	56.8
20 kgcap	US06	30	68	21	0.91	68	0.94	190	Blended	71.1	52	38.4
Recharg. Zn-air	FUD	40	45	19	0.87	45	0.97	228	AE	None	None	54.5
32 kgbat	HW	38	45	19	0.81	45	0.97	242	AE	None	None	57.7
20 kgcap	US06	66	68	21	0.82	68	94	149	Blended	62.4	60	38.8

[a] Weight of cells only.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 15. Simulation results for the batteries alone.

Battery Type[a]	Cycle	Range (mi)	kW max. Control	kW max. batteries	Efficiency batteries	Wh/mi batteries	Mode of operation	mpg 20 mi	mpg 40 mi	mpg Charge sustaining HEV
EIG										
NiCoMn	FUD	27	30	30	0.94	125	Blended	134	85	47
30 kg	HW	24	20	20	0.93	137	Blended	110	87	47
	US06	57	58	58	0.88	102	Blended	48	45	37
Composite MnO_2	FUD	36	30	30	0.92	135	Blended	134	104	46.9
32 kg bat	HW	31	20	20	0.91	147	Blended	167	113	46.6
	US06	64	58	58	0.87	92	Blended	48	48	34.1
Si Carbon/composite MnO_2	FUD	35	30	30	0.93	138	Blended	138	106	46.9
22 kgbat	HW	32	20	20	0.92	148	Blended	169	114	46.9
	US06	64	58	58	0.88	87	Blended	48	48	35.7
Recharg. Zn-air	FUD	66	30	30	0.84	139	Blended	139	137	39.4
32 kgbat	HW	63	20	20	0.83	156	Blended	169	169	41.1
	US06	93	36	36	0.72	101	Blended	48.5	48.5	30.1

[a] Weight of cells only.

Encyclopedia of Automotive Engineering. Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

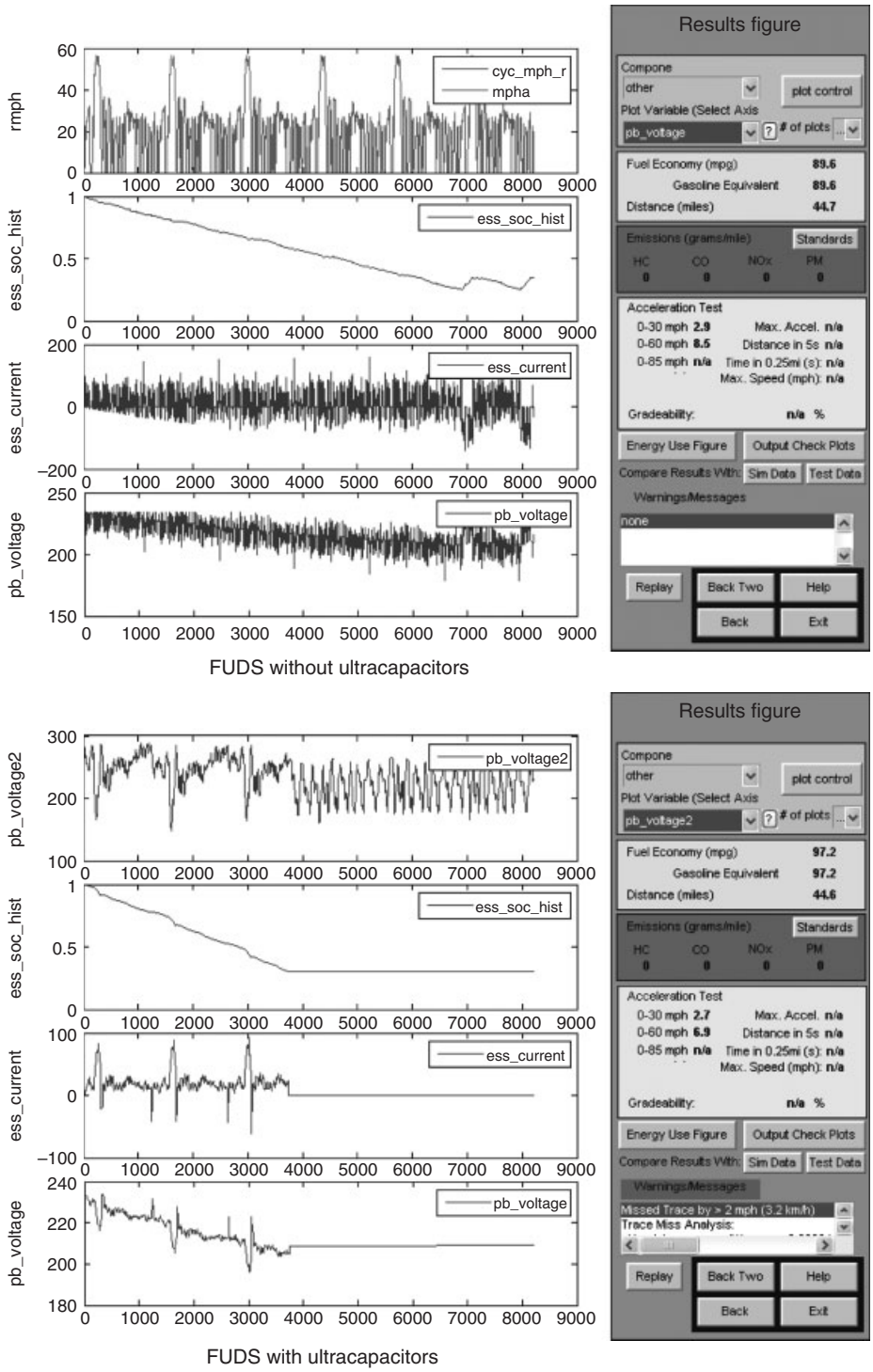


Figure 7. The silicone carbon lithium battery on the FUDS with and without ultracapacitors.

4.3.2 Reductions in battery stress using ultracapacitors

Simulation results for a passenger car using silicone carbon lithium-ion batteries with and without the ultracapacitors are shown in Figures 7 and 8 for the FUDS and US06 driving cycles. The battery SOC, voltage, and current are shown as a function of time in the figures. The effects of the load leveling of the power demand from the batteries using

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

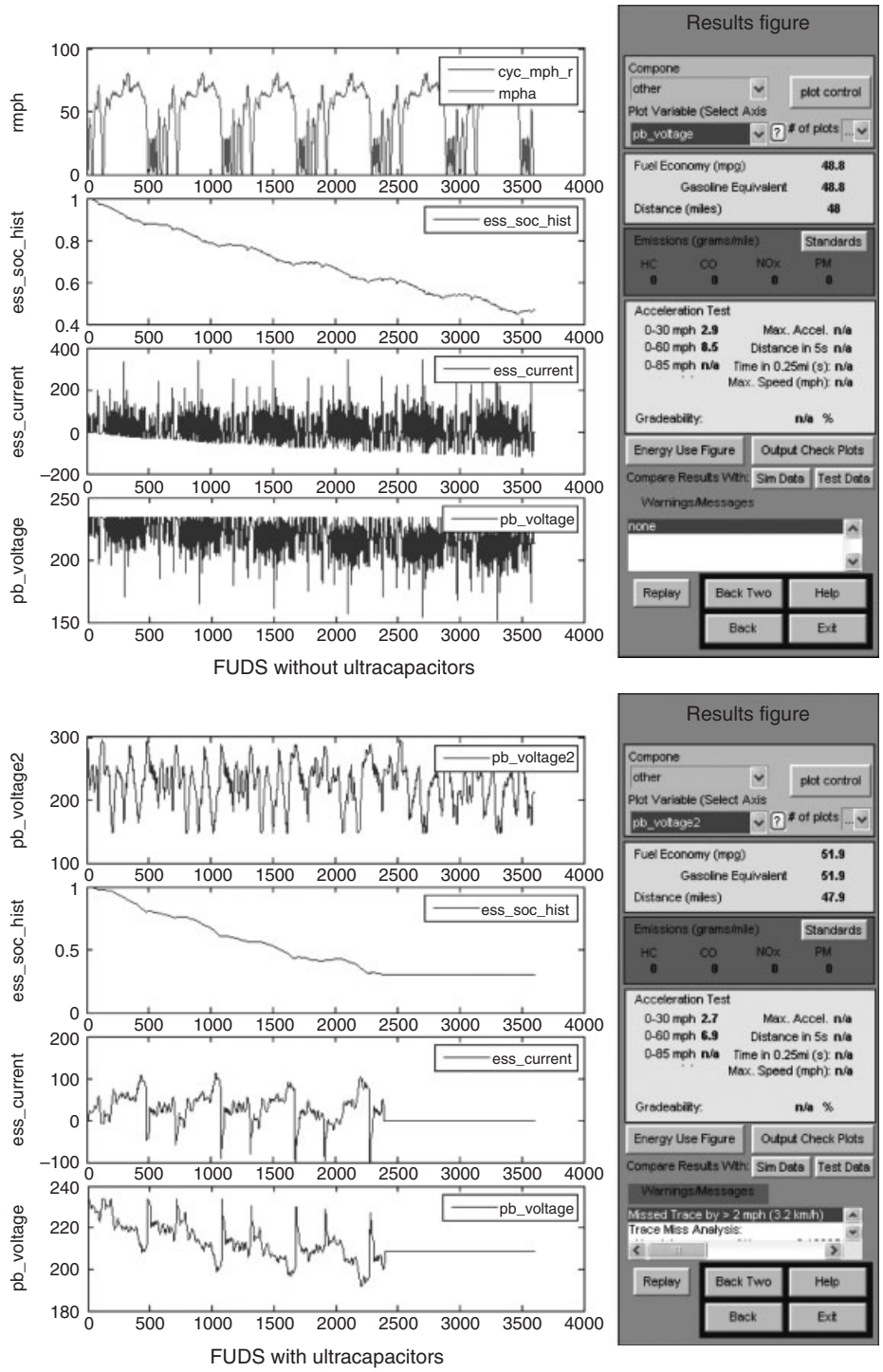


Figure 8. The silicone carbon lithium battery on the US06 with and without ultracapacitors.

the ultracapacitors are evident in the figures. The batteries provide the average current and the ultracapacitors meet the current fluctuations. The peak currents from the batteries are lower by a factor of 2–3 using the ultracapacitors. The minimum voltages of the batteries are significantly higher using the capacitors, and the voltage dynamics (fluctuations) are dramatically reduced. Hence, the stress on the battery and resultant heating are much reduced. The simulation results in Figures 7 and 8 also show that the ultracapacitors are utilized over a wide voltage range,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

indicating that a large fraction of their usable energy storage (100 Wh) is being used to load level the batteries. This is only possible using a DC/DC converter between the battery and the DC bus.

The simulation results also indicate that using ultracapacitors, batteries with a wide range of power characteristics can be used in PHEVs and also EVs without sacrificing vehicle performance and subjecting the batteries to high stress and resultant shorter cycle life. This could be especially important in the future as high energy density batteries such as zinc-air and possibly lithium-air are developed. It is likely that these battery types will not have commensurate increases in useable power density, and without ultracapacitors, the battery unit in PHEVs and EVs would be sized by the maximum power requirement (kW) rather than the range (mi)/energy requirement (kWh). This would significantly increase weight, volume, and the cost of the battery unit. It is also unlikely that the air electrode will have charge acceptance capability and thus regenerative braking performance approaching that of ultracapacitors or even lithium-ion batteries. This is another advantage of the use of ultracapacitors with the air-electrode batteries.

5 SUMMARY AND CONCLUSIONS

In this chapter, ultracapacitor concepts and performance are reviewed. Test data are presented for carbon/carbon and hybrid ultracapacitors based on testing at the University of California, Davis. It was found that carbon/carbon devices are commercially available with energy densities of 4.5–5 Wh/kg and power densities of 1–2 kW/kg with a pulse efficiency of 95%. Hybrid ultracapacitors using carbon, graphite, and metal oxides were also tested. The hybrid devices had energy densities between 12 and 30 Wh/kg and power densities of 1–3 kW/kg. Comparisons of the power densities of ultracapacitors and lithium batteries indicated that the best ultracapacitors had power densities a factor of 2–3 higher than the high power batteries, but there are some batteries with higher power capability than the lower power capacitors. For use in hybrid vehicles, high power density is very important for ultracapacitors.

The cycle life and cost of ultracapacitors were also considered. It was found that the life (hours) characteristics could be expressed as a Weibull distribution and that the life of an ultracapacitor pack is dependent on the cell life characteristics, voltage and temperature variations, and the number of devices in series. A cost analysis indicated that the material costs depended primarily on the cost of carbon and that a cell cost of about 0.5 cents/F cell or less seemed likely for volume production of ultracapacitors. This could make ultracapacitors cost competitive with high power lithium batteries for HEV applications.

The application of ultracapacitors in hybrid-electric vehicles is considered in detail utilizing vehicle simulation results. The use of ultracapacitors alone is analyzed for micro-hybrids and mild/full charge sustaining hybrids and in combination with high energy density batteries for plug-in hybrids. It was found that ultracapacitors could be used in place of batteries in the charge sustaining hybrids, resulting in fuel economy improvements of 25–50% even for relatively small electric drive systems of 6–10 kW. The ultracapacitor units only stored 50–100 Wh of useable energy and their efficiency was high being 95–98%.

Vehicle simulations were also performed for plug-in hybrid vehicles that used ultracapacitors combined with advanced, high energy density batteries. The batteries had energy densities of 200–400 Wh/kg, but only moderate power capability of 300–1000 W/kg for 90% efficient pulses. The simulation results indicated that the advanced batteries combined with ultracapacitors could provide all-electric operation on the FUDS and Federal Highway cycles in the charge depleting mode and excellent fuel economy in the charge sustaining mode. Vehicle operation with the advanced batteries without the ultracapacitors required blended operation on all driving cycles. Comparisons of the current/voltage/power profiles of the batteries with and without the ultracapacitors indicated that the peak currents and thus the stress on the batteries were reduced by about a factor of three using the ultracapacitors. This reduction could lead to a large increase in battery cycle life.

REFERENCES

Barsoukov, E. and Macdonald, J.R. (2005) *Impedance Spectroscopy*, Wiley-Interscience.

Béguin, F. and Frąckowiak, E. (eds) (2013) *Supercapacitors - Materials, Systems, and Applications*, Wiley-VCH, Weinheim, Germany.

Burke, A.F. (2007), Comparisons of lithium-ion batteries and ultracapacitors in hybrid-electric vehicle applications. Paper presented at *EET-2007 European Ele-Drive Conference*, Brussels, Belgium, June 1, (paper on CD of proceedings).

Burke, A.F. (2007) Batteries and ultracapacitors for electric, hybrid, and fuel cell vehicles. *Proceedings of the IEEE*, **95**, Special issue on Electric Powertrains, April, IEEE, USA.

Burke, A.F. (2009) Ultracapacitor technologies and applications in hybrid and electric vehicles. *International Journal of Energy Research*, **34**(2), John Wiley & Sons Ltd., Chichester, UK.

Burke, A.F. (2010) Electrochemical capacitors in *Lindel's* Handbook of Batteries, 4th edn, McGraw-Hill, USA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Burke, A.F. (2011) Ultracapacitors alone and in combination with batteries in hybrid-electric vehicles: system considerations and performance in *Encyclopedia of Sustainability Science and Technology*, Springer.

Burke, A.F., Kershaw, T., and Miller, M. (2003), Development of advanced electrochemical capacitors using carbon and lead-oxide electrodes for hybrid vehicle applications, Institute of Transportation Studies Report UCD-ITS-RR-03-2, June, USA.

Burke, A.F. and Miller, M. (2005) Cell Balancing Considerations for Long Series Strings of Ultracapacitors in Vehicle Applications. *Proceedings of the Advanced Capacitor World Summit*, San Diego, California, USA, July 11–13.

Burke, A.F., Miller, M., and Van Gelder, E. (2007) Ultracapacitors and Batteries for Hybrid Vehicle Applications. *23rd Electric Vehicle Symposium*, Anaheim, California, December (paper on CD of proceedings).

Burke, A.F. and Miller, M. (2009) Electrochemical capacitors as energy storage in hybrid-electric vehicles: present status and future prospects, EVS-24, Stavanger, Norway, May (paper on the CD of the meeting).

Burke, A.F. and Miller, M. (2010) Testing of electrochemical capacitors: capacitance, resistance, energy density, and power capability. *Electrochimica Acta*, **55**, 7538–7548.

Burke, A. and Miller, M. (2010) Lithium batteries and ultracapacitors alone and in combination in hybrid vehicles: fuel economy and battery stress reduction advantages. Paper presented at the *Electric Vehicle Symposium 25*, Shenzhen, China, November.

Burke, A.F. and Miller, M. (2011) The power capability of ultracapacitors and lithium batteries for electric and hybrid vehicle applications. *Journal of the Power Sources*, **196** (1, January), 514–522, Elsevier, UK.

Burke, A.F., Zhao, H., and Van Gelder, E. (2009) Simulated performance of alternative hybrid-electric powertrains in vehicles on various driving cycles. EVS-24, Stavanger, Norway, May (paper on the CD of the meeting).

Burke, A.F. and Zhao, H. (2010) Projected fuel consumption characteristics of hybrid and fuel cell vehicles for 2015-2045. Paper presented at the Electric Vehicle Symposium 25, Shenzhen, China, November.

Chandrasekhar, P. (1999) *Conducting Polymers, Fundamentals and Applications*, Kluwer Academic Publishers.

Chu, A. and Braatz, P. (2002) Comparison of commercial supercapacitors and high-power lithium-ion batteries for power assist applications in hybrid electric vehicles: 1. Initial characterization. *Journal of Power Sources*, **112**(1), 236–246, October, Elsevier, UK.

Conway, B.E. (1999) *Electrochemical Capacitors: Scientific Fundamentals and Technological Applications*, Kluwer Academic/Plenum.

Dixon, J., Nakashima, I., and Ortuzar, M. (2010) Electric vehicle using a combination of ultracapacitors and ZEBRA battery. *IEEE Transactions on Industrial Electronics*, **57**(3), 943–949, March, IEEE, USA.

FreedomCar (2004) Ultracapacitor test manual. Idaho National Engineering Laboratory Report DOE/NE-ID-11173, September 21, USA.

Hooks, N. (2011) Electric tramway applications of large supercapacitor energy storage units: designs and in-revenue demonstration. Presentation at the Advanced Energy Solutions Conference, San Diego, California, September.

Jung, D.Y. (2002) Shield ultracapacitor strings from overvoltage yet maintain efficiency. *Electronic Design*, **50**, 81–86, May 27,

Kotz, R., Sauter, J.C., Ruch, P. *et al*. (2007) Voltage Balancing of a 250V Supercapacitor Module for a Hybrid Fuel Cell ehicle. *Proceedings of the 16th International Seminar on Double-layer Capacitors and Hybrid Energy Storage Devices*, Deerfield Beach, Florida, USA, December.

Long, J.W. (2008) Electrochemical capacitors emPowering the 21st century. *Interface*, **17**(1, Spring), Electrochemical Society, New Jersey, USA.

Miller, J.M., Bohn, T., and Deshpande, U. (2008) DC-DC converter buffered ultracapacitor in active parallel combination with lithium battery for plug-in hybrid electric vehicle energy storage. SAE Paper 2008-01-1003, SAE World Congress, Detroit, Mich., April, Society of Automotive Engineerings, PA, USA.

Miller, J.R., Butler, S.M., and Goltser, I. (2006) Electrochemical Capacitor Life Predictions Using Accelerated Test Methods. *Proceedings of the 42nd Power Sources Conference*, paper 24.6, p. 581, Philadelphia. Pa., USA, June.

Miller, J.R. and Butler, S.B. (2006) Capacitor System Life Reduction Caused by Cell Temperature Variation. *Proceedings of the Advanced Capacitor World Summit*, San Diego, California, USA, July.

Naoi, K., Ishimoto, S., Miyanmoto, J., and Naoi, W. (2012) Second generation 'nanohybrid supercapacitors': evolution of capacitive energy storage devices. *Energy & Environmental Science*, **5**, 9363–9373, Royal Society of Chemistry, UK

Nelson, P.A. (2008) Interim report on the cost study for plug-in hybrid vehicle batteries. Argonne National Laboratory Report, April.

Nelson, P.A., Santini, D.J., and Barnes, J. (2009) Factors determining the manufacturing costs of lithium-ion batteries for PHEVs. EVS-24, Stavanger, Norway, May (paper on the CD of the meeting).

Rafik, F., Gualous, H., Callay, R., Crausaz, A., and Berthon, A. (2006) Supercapacitors characterization for hybrid vehicle applications. Paper presented at *ESSCAP 2006*, Lausanne, Switzerland (available on web).

Ruch, P., Kotz, W., and Wokaun, A. (2009) Electrochemical characterization of single-wall carbon nanotubes for electrochemical double-layer capacitors using non-aqueous electrolyte. *Electrochemica Acta*, **54**, 4451–4458, Elsevier, UK.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto140
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Thermodynamic Analysis

Jeffrey D. Naber, Jaclyn E. Johnson, and Charles H. Margraves

Michigan Technological University, Houghton, MI, USA

1 Introduction 1
2 Thermochemistry 10
3 Second Law Discussion 13
4 Energy/Loss Analysis 13
5 Simplified Engine Cycles 13
Endnotes 15
References 15

1 INTRODUCTION

In this chapter, we will layout and summarize the thermodynamic fundamentals for analysis of internal combustion (IC) engines including the constitutive relations, thermochemistry, and first and second law analysis. Many excellent texts on the subject exist for both outlining the fundamentals of thermodynamics with applications (e.g., Van Wylen and Sonntag, 1986; Wark, 1988; Myers, 2007) and IC engines analysis through thermodynamics (Taylor, 1985a, 1985b; Obert, 1973; Heywood, 1988). This chapter provides a concise summary with a consistent framework for analysis and introduction of advanced topics in the remaining chapters.

First, a brief discussion on units is presented. Most analysis and published work is now done in SI units (s, m, kgmol, kg, N, W, K/°C); however, there is a persistence of English units (s, ft, lbmol, lbm, lbf, hp, °R/°F) in application of IC engines, and centimeter–gram–second (CGS) units (s, cm, gmol, g, dyne, erg or cal, K) in application of chemistry and kinetics. Even in application of SI units, there is a need to rationalize and thoroughly check units. A mechanism as applied by Myers (2007) is used and illustrated in this example for determination of power (P) from torque (T) and engine speed (N) (which can be defined in terms of angular speed ω) for an engine producing 350 ft-lbf of torque at 3500 rpm. Power is the product of rotational speed and torque, as given by

$$\begin{aligned} P &= \omega \cdot T \\ &= 2\pi \cdot N \cdot T \\ &= \frac{3500\,\text{rev}}{\text{min}} \left| \frac{350\ \text{ft-lbf}}{-} \right\| \frac{2\pi\,\text{rad}}{\text{rev}} \left| \frac{1\ \text{min}}{60\,\text{s}} \right| \frac{0.3048\,\text{m}}{1\,\text{ft}} \\ &\quad \left| \frac{1\,\text{N}}{0.2248\,\text{lbf}} \right| \frac{1\,\text{kW s}}{1000\,\text{N m}} \\ &= 174\ \text{kW} \end{aligned} \tag{1}$$

In this formulation, the engineering values for the parameters are first evaluated with units written in a ratiometric form and separated by single vertical lines (|). This is followed by the application of the necessary conversion factors (e.g., 1 min = 60 s) written as ratios (1 min/60 s = 1). A double line (||) separates the applied parameters in the equation and conversion factors. In this formulation, cancellation of units is easily analyzed and checked. In addition to the standard conversion factors, those for rotational position are important for a number of engine characteristic factors including engine speed, torque and work, as outlined here for a four-stroke engine:

$$1\ \text{rev} = 360^\circ\ \text{crank angle} = 2\pi\ \text{rad}$$

$$1\ \text{cycle} = 2\ \text{rev} = 720^\circ\ \text{crank angle} = 4\pi\ \text{rad}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Additionally, care is required in the use of the mole, which is confusing for those having initially studied chemistry using a gram-mole (gmol and often just given as mol) and then thermodynamics where kilogram mole (kgmol) is often used.[1] The mole equivalencies are given as

$$1\ \text{kgmol} = 1000\ \text{gmol} = 2.2\ \text{lbmol}$$

with the corresponding changes in Avogadro's constant that defines the number of elementary entities or constituent number of particles (CNP) per mole of a substance

$$\begin{aligned} &6.02214 \times 10^{23}\ \text{CNP/gmol} \\ &= 6.023 \times 10^{26}\ \text{CNP/kgmol} \\ &= 2.732 \times 10^{26}\ \text{CNP/lbmol} \end{aligned}$$

1.1 Constitutive relationships

For thermodynamic analysis and simulation, one must develop an equation set that comprises an accurate constitutive relationship between the knowns and unknowns to provide a closed solution to the problem. These can be set up to solve for state-to-state transitions as in ideal cycle analysis or to solve for the continuous or discrete time-varying dynamics of the problem. The primary equations used in the area of IC engines for this purpose are as follows:

- equations of state (EOSs) including mixture rules
- intrinsic property relationships between T, P, v, u, h, s, g, and so on
- conservation of mass
- conservation of atomic elements (O, N, C, H, etc.)
- species molar balances
- conservation of linear and angular momentum
- conservation of energy
- entropy/availability/exergy
- rate and flux relationships for flow, mass transfer, heat transfer, and kinetics.

The conservation equations may be closed (e.g., fixed mass) or open and include rate terms for the fluxes of mass, momentum, and energy along with generation and consumption/destruction terms. Development of a constitutive set of relations requires first the definition of an appropriate control volume or control mass and the related control surface. The remainder of this section covers the primary equations above.

1.1.1 Equation of state

EOSs provide a means to represent pressure, volume, and temperature relationships of different fluids, with varying levels of complexity depending upon the desired accuracy of the representation. These range from the simplified ideal gas law to relationships with numerous parameters to improve the accuracy of fluid property representation with a trade-off of increased computational demand (Martin, 1979). EOSs are represented either in the standard absolute pressure (P), molar volume (ratio of volume to mole) ($\overline{v}$), and absolute temperature (T) form, or through the use of compressibility Z, which is defined as $Z = P\overline{v}/R_u T$, where R_u is the universal gas constant (8.3145 kJ/kgmol K).

The simplest EOS is the ideal gas EOS, which will be discussed in the next section. Complexity increases with two-constant EOS including the Redlich–Kwong EOS, which is valid at higher pressures, and the Van Der Waals EOS, which considers that gas molecules occupy a finite volume and that there are attractive forces between molecules (Moran and Shapiro, 2008). Under simplifying assumptions, these EOSs reduce to the standard ideal gas law. Further EOSs exist such as the Beattie–Bridgeman EOS, which includes five constants determined based on curve fitting to experimental data of pressure, volume, and temperature, and the eight-constant Benedict–Webb–Rubin EOS valid for light hydrocarbons (Moran and Shapiro, 2008). Additional EOSs exist, with different applicability, complexity, and number of constants, including cubic EOSs, and other modified EOSs (Martin, 1979; Slavinskaya, Zizin, and Aigner, 2010; Wei and Sadus, 2000).

1.1.1.1 Ideal gas. Much of the analysis in IC engines is based upon the evaluation of the working gas, which can be assumed for many problems with minimal error as a perfect or ideal gas and with a compressibility of 1. This approximation is typically valid for combustion as these systems are usually at high temperatures and low densities (Turns, 2000). Under these perfect gas conditions, the relationship between pressure, temperature, and volume is defined as the *ideal gas law* (Equation 2):

$$\begin{aligned} PV &= NR_u T \\ P &= cR_u T \\ PV &= mRT \\ Pv &= RT \\ P &= \rho RT \end{aligned} \tag{2}$$

where V is volume, N is moles of the gas, c is the molar concentration[2], m is mass of the gas, R is the gas constant

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

for the specific gas, v is the specific volume (ratio of volume to mass), and ρ is density. The universal and specific gas constants are related by

$$R \equiv \frac{R_u}{M} \tag{3}$$

where M is the molecular weight[3] of the gas (e.g., 28.01 kg/kgmol for N_2). As an example, the molar concentration and density of air ($M_{air} = 28.97$ kg/kgmol) at standard conditions ($P = 101.325$ kPa, $T = 25^\circ C = 298.15$ K) is

$$\begin{aligned} c &= \frac{P}{R_u T} = \frac{101.325\ \text{kPa}}{-} \left| \frac{\text{kgmol K}}{8.3145\ \text{kJ}} \right| \frac{-}{298.15\ \text{K}} \\ &\quad \left\| \frac{1000\ \text{N}}{\text{kPa m}^2} \right| \frac{1\ \text{kJ}}{1000\ \text{N m}} \\ &= 0.0409\ \frac{\text{kgmol}}{\text{m}^3} = 0.0409\ \frac{\text{kgmol}}{\text{m}^3} \\ &\quad \left\| \frac{6.023 \times 10^{26}\ \text{molecules}}{\text{kgmol}} = 2.46 \times 10^{25}\ \frac{\text{molecules}}{\text{m}^3} \right. \\ \rho &= c \cdot M = \frac{0.0409\ \text{kgmol}}{\text{m}^3}\ \frac{28.97\ \text{kg}}{\text{kgmol}} \Bigg| \\ &= 1.18\ \frac{\text{kg}}{\text{m}^3} \end{aligned} \tag{4}$$

A useful form for the ideal gas EOS when developing analysis for closed systems with varying volume is in differential form

$$\frac{dP}{P} + \frac{dV}{V} = -\frac{dM}{M} + \frac{dm}{m} + \frac{dT}{T} \tag{5}$$

In a closed system where either there are no reactions or where the products and reactants have similar molecular weights, which is often a good first approximation for hydrocarbon fuels, this simplifies to

$$\frac{dT}{T} = \frac{dP}{P} + \frac{dV}{V} \tag{6}$$

In the case of mixtures of ideal gases, a partial pressure mixing rule is applied. Each constituent i in the mixture has its own partial pressure P_i, and the total mixture pressure is the sum of the partial pressures multiplied by the mole fraction for each species (which is known as *Dalton's Law*). To apply Dalton's law requires the assumption that each component can be treated as an ideal gas at the mixture temperature and volume.

Correspondingly, a partial volume rule exists, known as *Amagat's model*, which assumes that each mixture component can be treated as an ideal gas at the mixture pressure and temperature. Therefore, the partial volume V_i of the component in the overall mixture is determined based on the product of the constituent's mole fraction with the total mixture volume.

Similar relationships exist for other thermodynamic properties of ideal gases, including internal energy, enthalpy, entropy, and specific heats. Here, the representing mixture properties are the sum of the product of the constituent mole fraction with its individual constituent thermodynamic property for each of the constituents in the mixture (Moran and Shapiro, 2008). In a manner similar to the molar basis, a mass basis can be used to determine mixture properties. Note that the above discussion applies to ideal gas mixtures that are already formed. As an alternative case, ideal gases can form a mixture with gases initially being separate, which is an *irreversible* mixing process that produces entropy. This will involve an energy balance and, for simplification, typically requires assumptions of no work and no heat transfer (adiabatic process), and application of a closed system or an open system (control volume) energy balance, depending on the process.

1.1.1.2 Liquids. Under many circumstances, liquids can be treated as incompressible, meaning their density is constant, thereby simplifying the analysis. One exception, however, is for high pressure injection systems where compressibility of the fuel is important and should be considered.

Liquid fluid compressibility is characterized by the bulk modulus of elasticity E, which defines for the given fluid how easily a unit volume of fluid (or density) can be changed when the working pressure is changed (Munson *et al*., 2009), that is,

$$E = -\frac{dP}{dV/V} = \frac{dP}{d\rho/\rho} \tag{7}$$

An increase in pressure will yield a volume reduction, and therefore E is always positive. A large bulk modulus means a large pressure change is needed to yield a small volume change, and is representative of an essentially incompressible fluid.

An understanding of fluid compressibility also comes into consideration for the speed of sound, an intensive property (intensive properties are scale-invariant, meaning independent of system size or amount of material) defined based on the medium through which sound (pressure pulse) is moving. The speed of sound defines the velocity at which small disturbances propagate through the fluid. The speed of sound c is defined as

$$c = \sqrt{\frac{dp}{d\rho}} = \sqrt{\frac{E}{\rho}} \tag{8}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

1.1.1.3 Real fluids.

1.1.1.3.1 Applicability. Under the high pressures that occur in engines and injection, there is sometimes a need to consider real fluids and their behavior. Real fluids remove the assumptions of incompressible flow (constant density) and inviscid flow (viscosity is zero, friction effects are negligible), which are applied to ideal fluids.

1.1.1.3.2 General compressibility. Compressibility is a dimensionless ratio of the product of pressure and volume to the product of the ideal gas constant and temperature. For a compressibility of 1, the ideal gas law is valid. As the compressibility deviates from unity, different EOSs or corrections are needed to define species pressure, volume, and temperature data. To understand conditions for which compressibility may approach 1, critical pressure and temperature can be considered. The critical pressure (P_c) is the maximum pressure at which the liquid and vapor phases can simultaneously exist, with the critical temperature (T_c) being the maximum temperature at which the liquid and vapor phases can simultaneously exist. When the pressure of a substance is small in comparison to the critical pressure, or temperature is large relative to the critical temperature, meaning at low pressures and high temperatures compressibility is near unity and the ideal gas equation is valid (Moran and Shaprio, 2008).

Generalized compressibility charts exist for gases, which provide compressibility as a function of reduced pressures and temperatures to assist in specifying pressure–volume–temperature states for a gas (Moran and Shapiro, 2008). Reduced pressure (P_r) is the ratio of the pressure to the critical pressure of the substance, and reduced temperature (T_r) is the ratio of the temperature to the critical temperature of the substance. The generalized compressibility chart for gases provides a reasonably accurate substitute for tabulated pressure–volume–temperature data for a given substance (Moran and Shapiro, 2008). These charts are generalized, meaning they are valid for a range of gases, within reasonable accuracy.

For more exact and fluid-specific dependent relationships, EOSs are expressed as pressure–volume–temperature relationships, but also can be expressed in terms of compressibility. For example, EOSs can be cubic in compressibility, such as the Van der Waals or Redlich–Kwong EOS (Moran and Shapiro, 2008).

1.1.1.3.3 EOS example. All EOSs exhibit advantages and disadvantages in regards to applicability, complexity, and accuracy for the given fluid. Therefore, careful consideration is needed when choosing an EOS, weighing all the important parameters. These more complicated EOSs possess numerous constants, which must be analytically or experimentally determined. For example, in cubic EOSs, being cubic in compressibility, the EOS constants are determined in terms of critical properties of the fluid of interest, with the first two pressure–volume derivatives evaluated at constant temperature equaling zero when evaluated at the fluid's critical point (Martin, 1979).

Various nonideal gas EOSs exist, with one example being the Peng–Robinson cubic EOS, which is defined by Equation 9, and is relevant and used in diesel fuel modeling (Nesbitt, 2011; Desantes *et al*., 2007).

$$
\begin{aligned}
P &= \frac{R_u T}{\overline{v} - b} - \frac{a\alpha}{\overline{v}^2 + 2b\overline{v} - b^2} \\
\alpha &= 1 + \kappa(1 - T_r^{0.5})^2 \\
\kappa &= 0.37464 + 1.54226\omega - 0.26992\omega^2 \\
a &= 0.45724\frac{R_u^2 T_c^2}{P_c} \\
b &= 0.07780\frac{R_u T_c}{P_c}
\end{aligned}
\tag{9}
$$

where a, b, and κ are EOS constants and vary depending on the species considered, α is an EOS parameter which is a function of temperature, and ω is a species-acentric factor, a species property. If the volume and temperature data are known, along with species properties (critical properties and acentric factor), the pressure of the species can be determined at these defined conditions using Equation 9. Other EOSs exist, all with different accuracies and applicability (Reid, Prausnitz, and Poling, 1987).

1.1.1.4 Intensive properties (u, h, s, g).

In addition to pressure, volume, and temperature, other properties of interest are used to define species. These properties can be classified as *extensive* or *intensive*. *Extensive* properties depend on the size, or mass, of the systems, with examples including mass, volume, energy, and others. *Intensive* properties are independent of the size of a system, with examples being specific volume, pressure, and temperature. Important thermodynamic properties include internal energy (u), enthalpy (h), entropy (s), and Gibbs free energy (g), all being intensive (independent of mass). These can also be referred to as *specific properties*, for example, specific internal energy. Some properties are related, as will be further discussed. For example, enthalpy h is the combination of internal energy (u) and the product of pressure (P) and specific volume (v). Properties denoted with a lower case letter are defined as specific properties; when written in upper case, these are extensive properties with the exceptions of pressure and temperature.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

1.1.1.5 Mixture relations. Thermodynamic properties, such as those just discussed, are available for single-component species. However, fuels and other constituents are composed of several components, and therefore methods to determine mixture properties are imperative in the study of fuels and combustion as they relate to IC engine analysis. Previously discussed were mixture rules for ideal gases; however, to uphold accuracy, mixtures cannot always be treated as ideal gases and therefore alternative approaches for mixture relations are needed.

In order to apply EOSs or methods to determine thermophysical properties of mixtures, the mixture parameters must be known, such as composition, mass fraction, and mole fraction. Mixture parameters can be specified if the independent properties of the components are known. Mixture compositions can be related on a mass or molar basis using mass fractions (mass of species divided by the total mass of the mixture), or mole fractions (moles of species divided by the total number of moles in the mixture). These two fractions will each, independently, sum to 1 and can be related to each other by molecular weights. Once the composition is specified, intensive properties can be used to fully define the mixture state, with two intensive properties needed for a complete definition, such as pressure and temperature, or pressure and volume.

In order to evaluate pressure–temperature–volume properties of gas mixtures, one method is to use mole fraction relationships to evaluate coefficients for the EOS, being combination rules for constants in the EOSs (Moran and Shapiro, 2008). Another mixture relationship is *Kay's Rule*, which is a principle of corresponding states, enabling calculation of mixture critical properties such as temperature and pressure. Mixture critical properties are just the mole-fraction-weighted sum of the individual constituents (Moran and Shapiro, 2008). Additionally, there exist empirical mixing rules, such as the additive pressure and additive volume rules, which just state that the total pressure (volume) is a sum of each of the component pressures (volumes) at a given temperature and volume (pressure). The properties of pressure (volume) for each component would be independently determined from an EOS or from tabulated data (Moran and Shapiro, 2008).

1.1.1.6 Two-phase mixtures. Another set of mixture relations involves a fluid that has two phases, such as liquid and vapor phases. For a single-component liquid, a two-phase liquid–vapor mixture at equilibrium is defined as being in the *saturation state* above which both vapor and liquid phases exist. Between the saturated liquid and saturated vapor states, the liquid and vapor coexist to form a saturated mixture, which has a given percentage in the vapor phase and the remaining percentage in the liquid phase. In the state of saturation, the pressure and temperature are dependent and this dependence is a line in a pressure–temperature (P–T) diagram, as shown in Figure 1 for *n*-heptane. This line defines the saturation curve, or vapor pressures at the corresponding temperatures, effectively separating the liquid and vapor states. The last point (with regard to highest pressure and temperature) defines the critical point of the fluid (*n*-heptane, $T_c = 540.0\,\text{K}$, $P_c = 27.4\,\text{bar}$), with conditions above this being defined as supercritical.

The saturation dome separating the compressed liquid from the superheated vapor can be shown in a pressure–volume (P–v) or pressure–enthalpy (P–h) diagram with the illustration of a P–h diagram for *n*-heptane shown in Figure 2. As P and T are dependent in the saturated state (Figure 1), a second independent intrinsic property must be used to define the state. The quality (x) is thus used, which is defined as the mass fraction of the mixture that is vapor (calculated from vapor

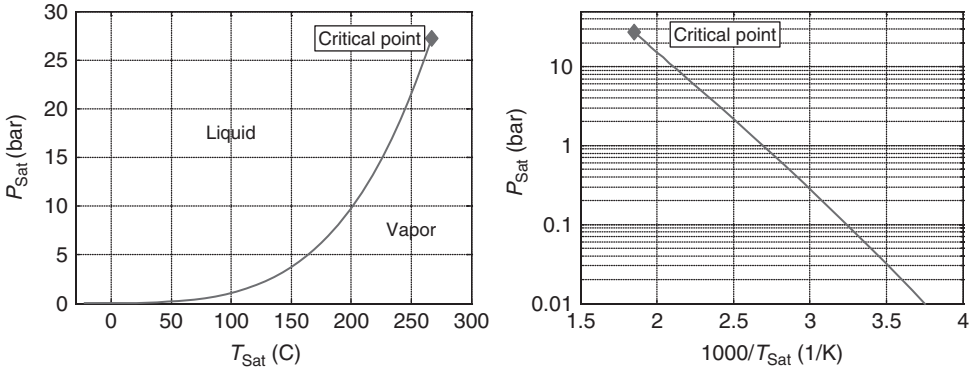


Figure 1. Pressure–temperature diagram for *n*-heptane showing the saturation curve and dependency of pressure and temperature.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

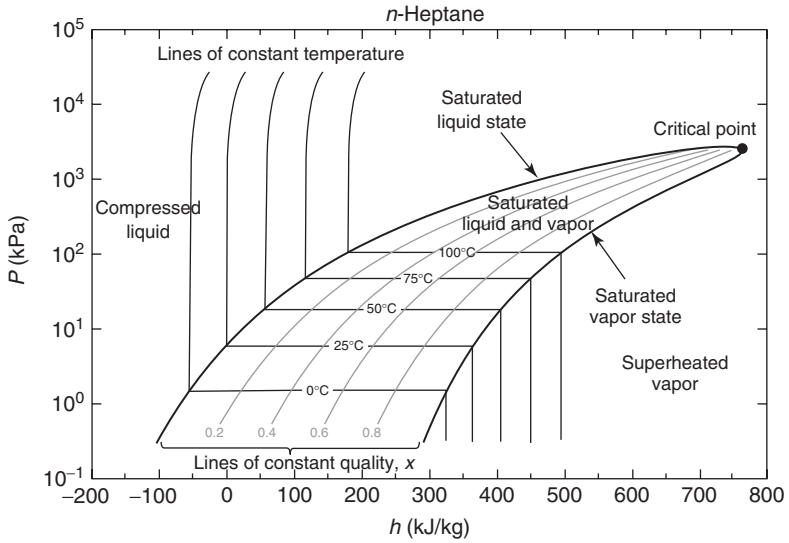


Figure 2. Pressure–enthalpy diagram for *n*-heptane fuel, showing the vapor dome, along with saturation states and quality lines.

mass m_{vapor} and liquid mass m_{liquid}):

$$x = \frac{m_{\text{vapor}}}{m_{\text{vapor}} + m_{\text{liquid}}} \tag{10}$$

Lines of constant quality are shown underneath the saturation dome in Figure 2. The left side of the dome represents the saturated liquid state conditions, and the right side of the dome defines the saturated vapor conditions. Temperature is constant horizontally underneath the dome, with the change being in the quality of the mixture between the vapor and liquid fractions. The specific energy required to transition from liquid ($x = 0$) to vapor ($x = 1$) is the internal energy (u_{fg}) or enthalpy of vaporization (h_{fg}). The enthalpy of vaporization is often referred to as the *latent heat of vaporization*.

The dependence between P and T for saturation can be estimated by the Clausius–Clapeyron equation

$$\frac{\mathrm{d}P}{P} = \frac{h_{\text{fg}}}{R_{\text{u}}T^2}\mathrm{d}T \tag{11}$$

With small changes of P and T, where h_{fg} remains nearly constant, the equation can be integrated to yield

$$\ln(P) = -\frac{h_{\text{fg}}}{R_{\text{u}}}\left(\frac{1}{T}\right) + C \tag{12}$$

This discussion on saturation and two-phase mixtures is increasingly important in IC engines. As diesel engines are primarily direct-injection and spark-ignition(SI)engines are transitioning to direct-injection, understanding the fuel injection process in terms of thermodynamics is imperative to understand the overall process. During this injection process, there exists a two-phase mixture, with liquid fuel during injection transitioning to vapor during mixing and penetration across the combustion chamber. A pressure–enthalpy diagram (Figure 2) can provide the path of the fuel states through the injection and vaporization process. Fuel is injected at an elevated pressure through an injector nozzle in a constant enthalpy process. The pressure is decreased during injection for this constant enthalpy process down to the charge gas pressure in the combustion chamber. Once the fuel is injected into the chamber, it will mix with the typically hotter charge gas, and the enthalpy of the fuel portion will increase. As a result, the mixture will transition from a liquid fuel to a saturated fuel–air mixture as it traverses the vapor dome.

1.1.2 Control volume/surfaces

Another set of important constitutive relationships are conservation equations. Applying conservation laws of mass, momentum, and energy to define thermodynamic processes requires definition of the system under consideration. A *system* defines what is being studied, and the shape or volume of the system may vary. There are two types of systems that will be considered, a *closed* system and an *open* system.

1.1.2.1 Closed system with discrete states. One type of system is a *closed* system, which is a fixed amount of matter, also known as a *control* mass, meaning there is no mass transfer across the closed system's boundary. Although a closed system does not have mass flow, there

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

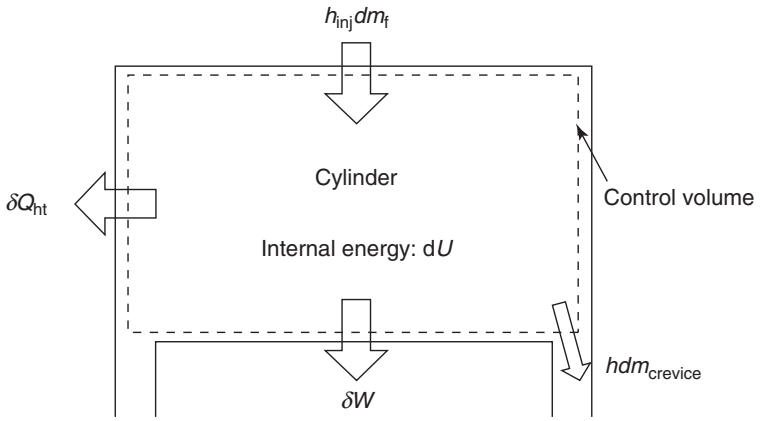


Figure 3. Open system (control volume) thermodynamic processes for an internal combustion engine. (Adapted from Gatowski *et al.* 1984).

still may be work done on or by the system, heat transfer across the system boundary, and reactions occurring within the system, which change the composition (mole and mass fractions of the species) but not the total mass.

1.1.2.2 Open dynamic system. An *open* system, synonymous with *control volume*, is defined as a region in space through which mass, or matter, can flow. A *control surface* is the boundary of the defined control volume, typically represented as a dashed line (Figure 3) around the system it is representing. Open systems can therefore exchange mass, work, and heat.

1.1.2.3 General conservation. General conservation can be applied to a control volume or an open system with flows across the control surface. This general conservation defines that the change in a system is equal to "in" minus "out," plus "generated" minus "destroyed," for any given type of property such as mass or energy. As another definition, general conservation can be inflow plus what is produced, equal to the outflow plus what is stored and what is destroyed.

1.1.3 Conservation relationships

Various conservation laws or relationships exist that govern the laws of nature. Conservation laws include conservation of mass, linear and angular momentum, and energy. These laws can be applied using the previously discussed systems.

1.1.3.1 Conservation of mass. The law of conservation of mass states that, for a system, the mass is neither created nor destroyed. Conservation of mass can be represented through the continuity equation when using control volume analysis. Applying the conservation of mass to a control volume of a system, it is determined that the time rate of change of mass within the control volume (m_{cv}) is equal to the sum of all mass flows into the control volume ($\dot{m}_{in}$) minus the sum of all mass flows out of the control volume ($\dot{m}_{out}$) as expressed by

$$\frac{dm_{cv}}{dt} = \sum_{in} \dot{m}_{in} - \sum_{out} \dot{m}_{out} \tag{13}$$

where the subscript "in" denotes the inlet, the subscript "out" denotes the exit of the control volume, and the subscript "cv" denotes the control volume.

1.1.3.2 Conservation of momentum. The conservation of momentum for a system is Newton's second law, which states that the time rate of change of linear momentum of the system (linear momentum being mass times velocity) is equal to the sum of the external forces acting on the system. The linear momentum equation is expressed by

$$\frac{\partial}{\partial t}\int_{cv} \text{Vel} \times \rho dV + \int_{cs} \text{Vel} \times \rho \times \text{Vel} \cdot \hat{n} dA = \sum F_{\text{contents of the cv}} \tag{14}$$

where the subscript "cs" denotes the control surface, the quantity $\text{Vel} \times \rho dV$, which (velocity multiplied by mass) is the momentum of a particle, $\hat{n}$ is the direction of the outward normal of the control volume, dA is the infinitesimal area of the control surface, and F is the external forces acting on the system.

1.1.3.3 Conservation of energy. The conservation of energy, also known as the *first law of thermodynamics*, states that for a system the time rate of increase of the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

total energy (E) stored in the system is equal to the energy transfer by heat (Q) and work (W), and the difference in energy transfer rates into and out of the control volume boundary (across the control surface) due to kinetic energy ($\mathrm{Vel}^2/2$, where Vel is velocity), potential energy (gz, where g is acceleration due to gravity and z is height relative to reference location), and specific enthalpy h. This is expressed as

$$\frac{\mathrm{d}E_{\mathrm{cv}}}{\mathrm{d}t} = \dot{Q}_{\mathrm{cv}} - \dot{W}_{\mathrm{cv}} + \sum_{\mathrm{in}} \dot{m}_{\mathrm{in}} \left(h_{\mathrm{in}} + \frac{\mathrm{Vel}_{\mathrm{in}}^2}{2} + g z_{\mathrm{in}} \right) - \sum_{\mathrm{out}} \dot{m}_{\mathrm{out}} \left(h_{\mathrm{out}} + \frac{\mathrm{Vel}_{\mathrm{out}}^2}{2} + g z_{\mathrm{out}} \right) \tag{15}$$

As an example for the application of the conservation of energy, we can consider a combustion chamber of an engine that can have mass flows in and out of the valves, fuel injection into the cylinder, and flows from crevice regions in the chamber. Mass flows are typically defined as positive when the flow of mass is out of the control volume. Once the fuel is burning, either via a compression-ignition (CI) or an SI event, there is energy conversion from chemical energy stored in molecular bonds to sensible (thermal) energy (temperature change), with this being treated as an internal energy term. In addition, there are also heat transfer losses to the chamber walls from the hot combustion products and work done, which is the useful work to push the piston down and provide energy to rotate the shaft and produce work. Depending on the desired analysis, certain flows can be neglected, for example, flows through valves when interest is on the portion of the cycle occurring when valves are closed (i.e., compression and expansion strokes as opposed to a full cycle analysis). This conservation process is shown schematically in Figure 3.

The overall conservation relationship in terms of energy (first law of thermodynamics) is defined as

$$\mathrm{d}U = \delta Q_{\mathrm{ht}} - \delta W - \sum_i \mathrm{d}m_i h_i \tag{16}$$

which considers change in internal energy $\mathrm{d}U$, heat transfer to the system δQ_{ht} work transfer from the system $\delta W = p\ \mathrm{d}V_{\mathrm{CV}}$, and enthalpy exchange across the control volume boundary for fuel injection into the cylinder and flows from crevice regions, $h_{\mathrm{inj}}\mathrm{d}m_{\mathrm{f}}$ and $h\mathrm{d}m_{\mathrm{crevice}}$, respectively.

1.1.4 *Second law of thermodynamics (entropy, exergy)*

While the first law of thermodynamics discusses the movement and storage of energy, which is a conserved property that cannot be created or destroyed, the second law provides information on the direction in which a process will naturally occur (i.e., energy flows from high temperature to low temperature). The property used to determine this direction is known as *entropy* (S), which can be defined, using statistical thermodynamics, as:

$$S = k \ln \Omega \tag{17}$$

where k is the Boltzmann constant and Ω is the thermodynamic probability (Schroeder, 2000). From this approach, entropy can be viewed as a measure of the disorder of the molecules in a system.

For most system analyses, the change in entropy is what is desired rather than the actual value. Because entropy is not a conserved value like energy (it can be generated but not destroyed), tracking the changes requires a balanced rather than a conserved equation:

$$\frac{\mathrm{d}S_{\mathrm{cv}}}{\mathrm{d}t} = \dot{S}_{\mathrm{in}} - \dot{S}_{\mathrm{out}} + \dot{S}_{\mathrm{gen}} \tag{18}$$

The subscripts "cv," "gen," "in," and "out" stand for control volume, generated, in, and out, respectively for the time rate of change of entropy $\dot{S}$. Using the Clausius inequality, it can be shown that

$$\mathrm{d}S \geq \frac{\delta Q}{T} \tag{19}$$

When the process is reversible, in which case no entropy is generated, the equal sign is enforced. When the process is irreversible (through mechanisms such as friction, heat transfer, and chemical reactions), the inequality is applied and entropy will be generated. It is this generation term that leads to less efficient systems, which will be shown in the discussion of exergy.

Changes in entropy between two states (denoted as subscripts "1" and "2" in the following equations), like any other thermodynamic property in a simple compressible system, can be calculated using two other measured properties such as temperature and pressure. For ideal gases, the change in specific entropy s can be calculated as follows:

$$s_2 - s_1 = \int_1^2 c_{\mathrm{p}} \frac{\mathrm{d}T}{T} - R_{\mathrm{u}} \ln \frac{P_2}{P_1} \tag{20}$$

$$s_2 - s_1 = \int_1^2 c_{\mathrm{v}} \frac{\mathrm{d}T}{T} + R_{\mathrm{u}} \ln \frac{v_2}{v_1} \tag{21}$$

where c_{p} and c_{v} are the specific heat at constant pressure (CP) and volume, respectively, and v is the specific volume.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The specific heat for most ideal gases is a function of temperature only (see Pressure and Heat Release Analysis for a discussion on the dependence of specific heat on pressure). If this relationship is linear, or, in the case of many monatomic gasses, constant, the change in specific entropy can be calculated using the following equations:

$$s_2 - s_1 = c_{p,avg} \ln \frac{T_2}{T_1} - R_u \ln \frac{P_2}{P_1} \tag{22}$$

$$s_2 - s_1 = c_{v,avg} \ln \frac{T_2}{T_1} + R_u \ln \frac{v_2}{v_1} \tag{23}$$

where $c_{p,avg}$ and $c_{v,avg}$ are determined by taking the specific heat at the average of the two temperatures.

If the relationship between the specific heat and temperature is highly nonlinear, then a relationship must be determined, often using fourth-order polynomials, and the first integral shown in Equations 20 and 21 must be completed to accurately determine the change in entropy.

Variable specific heats at CP and constant volume are tabulated in various reference books or have defined equations for approximating values as a function of temperature, for common gas and liquid working fluids. Refer to Moran and Shapiro (2008), Turns (2000), Reid, Prausnitz, and Poling (1987), and Reynolds (1979), as examples.

While entropy can be somewhat vague and difficult to understand, an alternative approach is to examine a property of a system known as *exergy* or availability. This property is used to determine the maximum potential work a system has at a given state relative to the so-called dead state. The dead state, often chosen to match the local environmental conditions, serves as a common ground to measure changes in exergy within a system. Exergy, like entropy, is not a conserved property but, unlike entropy, exergy cannot be created but can be destroyed.

The exergy ψ, of a system at any point can be determined through the following equation:

$$\psi = (h - h_o) - T_o(s - s_o) + \frac{(\text{Vel} - \text{Vel}_o)^2}{2} + g(z - z_o) \tag{24}$$

All values with subscript "o" represent the values at the dead state.

An exergy balance is shown below:

$$\frac{d\psi_{cv}}{dt} = \dot{\psi}_{in} - \dot{\psi}_{out} - \dot{\psi}_{des} \tag{25}$$

where the subscript "des" denotes destroyed. A change in exergy between any two states is given as

$$\psi = (h_2 - h_1) - T_o(s_2 - s_1) + \frac{V_2^2 - V_1^2}{2} + g(z_2 - z_1) \tag{26}$$

Now a link can be established between the entropy generated in the system and the exergy destroyed, as

$$\dot{\psi}_{des} = T_o S_{gen} \tag{27}$$

This equation indicates that, as more entropy is generated within a system, the potential for work decreases. Therefore, in order to make a system more efficient (i.e., more work is possible through the same amount of energy input), the reduction of the entropy being generated is important.

1.1.5 Isentropic processes of compression, expansion, and work

In order to determine the efficiency of work-producing or work-consuming devices, it is often preferable to use isentropic devices as an idealized model with which to compare the actual devices. Generally, the inlet conditions and outlet pressure for the isentropic device are set to match those of the actual device while all other outlet conditions of the isentropic device must be calculated. This is done by using relationships between pressure, temperature, and volume established from the entropy change relations provided earlier.

If a constant specific heat approximation is appropriate, then Equation 22 can be used to determine a relationship between pressure and temperature across the device being considered. The left-hand side of the equation will be equal to zero (for an isentropic, or constant entropy process) leaving the following:

$$0 = c_{p,avg} \ln \frac{T_2}{T_1} - R_u \ln \frac{P_2}{P_1} \tag{28}$$

where 1 represents the inlet and 2 the outlet. Using the relationships between the ratio of specific heats ($k = c_p/c_v$) and the gas constant ($R_u = c_p - c_v$), the following equality can be established:

$$\left(\frac{T_{out}}{T_{in}}\right) = \left(\frac{P_{out}}{P_{in}}\right)^{\frac{k-1}{k}} \tag{29}$$

A similar relationship can be established between the pressure and volume:

$$\left(\frac{P_{out}}{P_{in}}\right) = \left(\frac{V_{in}}{V_{out}}\right)^{k} \tag{30}$$

Now, using the exit conditions of the isentropic device, the work to (compression) or from (expansion) in the device W_{actual} can be calculated using a first law analysis. The efficiencies can then be defined as either work-producing

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

efficiency ($\eta_{\text{work-producing}}$) or work-consuming efficiency ($\eta_{\text{work-consuming}}$), as defined below:

$$\eta_{\text{work-producing}} = \frac{W_{\text{isentropic}}}{W_{\text{actual}}} \tag{31}$$

$$\eta_{\text{work-consuming}} = \frac{W_{\text{actual}}}{W_{\text{isentropic}}} \tag{32}$$

For variable specific heat considerations, the relationship for specific heat and temperature must be determined to complete the integral of the temperature term as was discussed.

1.1.6 Compressible and incompressible flow

Different flows will occur, compressible and incompressible, depending on the forces and processes in the flow. For example, there are instances where the compressibility may be unity (ideal gas law is valid) but the flow may still be compressible (varying density). The definition of flow as compressible or incompressible must consider the *process* of the flow, not solely the type of fluid (gas or liquid) that is flowing (Balachandran, 2006). This relates to the consideration of the process and force transmission, which occurs essentially instantaneously for incompressible flow but takes some finite time for compressible flow.

One metric used to understand whether flow is compressible or incompressible is to determine the Mach number of the flow (Ma), which is defined as the ratio of the speed of the flow to the speed of sound. For gases, when this ratio is <0.3, the flow can be considered as incompressible, and when it is >0.3 the flow should be considered compressible.

Of more relevance to an engine, considering, for example, flow through a throttle or in turbochargers, is compressible and choked flow. For compressible flow, with a Mach number of 1, this yields conditions of choked flow, meaning the mass flow rate of the flow through an orifice or fixed area cannot increase further unless the upstream pressure is increased or the upstream temperature is reduced. Choked flow is encountered when the critical pressure ratio of static to stagnation pressure is 0.53 at a specific heat ratio of 1.4 for air.

2 THERMOCHEMISTRY

In an engine, fuel, which is typically a mixture of hydrocarbons, mixes with air to produce a combustible fuel–air mixture that is ignited by a plasma discharge (SI engines including gasoline) or through autoignition (CI engines including diesel). This process depends on the mixture composition of the charge gas, which includes residuals from the combustion of prior cycles, combustion chamber conditions, fuel properties, and so on. To characterize this process, an analysis of the composition of the charge gas, fuel, and their mixtures from the standpoint of thermochemistry is required.

2.1 Air and humidity standard

Air is composed of a number of gases, and the properties of a standard dry air are given in Table 1. In this table are the mole and mass fractions and the molar ratio based upon 1 mol of oxygen (O_2), which will be used in the thermochemistry balance equation with fuel. Mole fraction is defined as the ratio of the moles of the given species to the total moles, with mass fraction being the ratio of the mass of the given species to the total mass. Overall, there are 3.775 (3.728 + 0.045 + 0.002) moles of other gases per mole of oxygen, in 1 mol of air.

In most cases, the combustion air is not dry and contains water vapor in significant amounts. The quantity of water vapor is most often given by relative humidity (RH) or humidity ratio. RH is the partial pressure of water vapor (p_{vap}) in air to the saturation pressure of the vapor (p_{sat}) at the same temperature (T):

$$\text{RH} = \frac{p_{\text{vap}}(T)}{p_{\text{sat}}(T)} \tag{33}$$

The humidity ratio (ω) is the ratio of the mass of water vapor ($m_{\text{H}_2\text{O,Vap}}$) to mass of dry air ($m_{\text{Air,Dry}}$), which can be related to the ratio of the partial pressure of the water vapor to the partial pressure of the dry air ($p_{\text{Air,Dry}} = p - p_{\text{vap}}$) by the molecular weights (M).

$$\begin{aligned}\omega &= \frac{m_{\text{H}_2\text{O,Vap}}}{m_{\text{Air,Dry}}} = \frac{p_{\text{vap}}}{p - p_{\text{vap}}} \cdot \frac{M_{\text{H}_2\text{O}}}{M_{\text{Air,Dry}}} \\ &= \frac{M_{\text{H}_2\text{O}}}{M_{\text{Air,Dry}}} \cdot \frac{\text{N}_{\text{H}_2\text{O}}}{N_{\text{Air,Dry}}} = \frac{1}{1.608} \cdot \frac{\text{N}_{\text{H}_2\text{O}}}{N_{\text{Air,Dry}}}\end{aligned} \tag{34}$$

Table 1. Composition of dry air.

Gas	Molar mass	Volume (ppm)	Molar ratio	Mass (%)
N_2	28.013	780,740	3.728	75.51
O_2	31.998	209,430	1.000	23.14
Ar	39.948	9340	0.045	1.29
CO_2	44.010	390	0.002	0.06
Others[a]	19.700	100	0.000	0.01
Air	28.965	1,000,000	—	100.00

[a] Includes neon, helium, methane, krypton, hydrogen, and so on.
(Adapted from Moran and Shapiro (2008), updated to include current atmospheric carbon dioxide levels (National Oceanic & Atmospheric Administration, 2011).)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The molar humidity ratio is then $M_{\text{Air,Dry}}/M_{\text{H}_2\text{O}} \cdot \omega = 1.608\,\omega$, where 1.608 is the ratio of molecular weight of dry air and water. The number of moles of H_2O from humidity per mole of oxygen in air is then $1.608/y_{\text{O}_2,\text{Air,Dry}} = 7.677$ where $y_{\text{O}_2,\text{Air,Dry}} = 0.2094$. For our thermochemisty representation of humid air, we can then write air on the basis of one mole of oxygen as

$$\text{O}_2 + 3.728\ \text{N}_2 + 0.045\ \text{Ar} + 0.002\ \text{CO}_2 + 7.677\,\omega\,\text{H}_2\text{O} \tag{35}$$

2.2 Thermochemistry of fuel mixtures with air and residuals

For most fuels, the composition is comprised of carbon, hydrogen, oxygen, nitrogen, and sulfur, with trace amounts of other substances. Typical fuels such as gasoline and diesel consist of individual compounds that are unknown, but testing can provide atomic mass or mole fraction compositions of these fuels. Additionally, in some cases an average molecular weight can be determined such that the fuel can be represented by the molecular average formula $C_XH_YO_ZN_WS_V$. Sulfur and nitrogen levels are usually low, and for this analysis, they will be neglected. The ideal reaction of this fuel in air, including humidity, to combustion products for a lean ideal reaction is given by

$$\begin{aligned} C_XH_YO_Z + \lambda R\,(&\text{O}_2 + 3.728\ \text{N}_2 + 0.045\ \text{Ar} \\ &+ 0.002\ \text{CO}_2 + 7.677\,\omega\,\text{H}_2\text{O}) \rightarrow X\,\text{CO}_2 \\ &+ (Y/2 + 7.677\lambda R\,\omega)\ \text{H}_2\text{O} + \lambda R(3.728\ \text{N}_2 \\ &+ 0.045\ \text{Ar} + 0.002\ \text{CO}_2) + (\lambda - 1)RO_2 \end{aligned} \tag{36a}$$

where λ is the excess air ratio ($\lambda = 1$ is the stoichiometric condition), and

$$R = X + \frac{Y}{4} - \frac{Z}{2} \tag{36b}$$

In this relationship, which is valid for $\lambda \geq 1$, the fuel $C_XH_YO_Z$ reacts with air to produce the ideal combustion products (CO_2 and H_2O).

The inverse of λ is the equivalence ratio Φ. Equivalence ratios >1 ($\lambda < 1$) signify rich combustion, meaning more fuel than required for complete combustion of air. Equivalence ratios <1 ($\lambda > 1$) define lean combustion, meaning excess air is present, or there is more air than needed for complete combustion of the fuel. A λ or equivalence ratio of 1 represents stoichiometric combustion, meaning the correct amount of fuel and air for complete combustion of all fuel. This corresponds to a stoichiometric fuel-to-air ratio (FAR) or stoichiometric air-to-fuel ratio (AFR), as tabulated in several texts, including Heywood (1988). With the stoichiometric values known, the excess air ratio or equivalence ratio can be determined based on the actual FAR or actual AFR as defined in Equation 37. The FAR and AFR are mass-based quantities, meaning they are the ratio of mass of fuel to mass of air, or vice versa.

$$\phi = \frac{1}{\lambda} = \frac{\text{FAR}_{\text{actual}}}{\text{FAR}_{\text{stoich}}} = \frac{\text{AFR}_{\text{stoich}}}{\text{AFR}_{\text{actual}}} \tag{37}$$

2.3 Enthalpies of formation

Fuels have different energy contents, which can influence combustion and emissions, as well as fuel economy. This energy content is represented through the heating value of the fuel; the larger the heating value, the more the energy available in the fuel for useful work, as will be discussed. To determine these energy contents, the thermodynamics of the intensive properties including chemical energy are required.

Enthalpy of formation $\overline{h}_{\text{f}}^{\circ}$ of a chemical is defined as the enthalpy increase from a reaction forming one mole of the given chemical, from its elements, for a standard state at the given temperature. As discussed, enthalpies require a definition of the reference state, with the most commonly used reference state being a temperature T_{ref} of 25°C (298.15 K) and pressure of 1 atmosphere (101.325 kPa). Elements at their reference state, which is the stable state of the element (e.g., O_2, N_2, H_2, He), are defined as having an enthalpy of formation of zero. These enthalpies of formation are tabulated or can be calculated with the available property data (Heywood, 1988; Moran and Shapiro, 2008, etc.). As an example application, enthalpies of formation can be used in the calculation of enthalpy of combustion (heating values).

Enthalpies of formation are used to define the enthalpy of a fluid at some defined condition. For the case of mixtures of ideal gases, their enthalpy can be calculated at the desired temperature, on either a mass or molar basis depending on the available enthalpy data. For the case of mole fraction x (as denoted by the overbar on the variables), the total mixture enthalpy is the weighted sum of the product of the mole fraction and enthalpy of each component species i at the desired mixture temperature, that is,

$$h = \sum_i x_i \bar{h}_i \tag{38}$$

Mole fraction is defined as the ratio of the number of moles of the species i to the total number of moles. To determine the molar-based enthalpy of each species, the enthalpy of formation and sensible enthalpy change $\Delta\overline{h}_{\text{s},i}(T)$, must be known (tabulated or functionally) to enable calculation of the enthalpy at the desired temperature, and is shown

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

below for a molar basis:

$$\bar{h}_i(T) = \bar{h}^{\circ}_{f,i}(T_{ref}) + \Delta\bar{h}_{s,i}(T) \quad (39)$$

The sensible enthalpy change is evaluated based on an enthalpy change from the defined reference temperature (e.g., 298.15 K) to the temperature of interest (T). These molar-based enthalpies are evaluated for each species of interest to enable the ideal gas mixture enthalpy to be calculated.

2.4 Heating values

The *heating value* or *heat of combustion* of the fuel provides a measure of the energy in the fuel, defined as the heat of reaction at either CP or constant volume (CV) at standard temperature (typically 298.15 K) for complete combustion of a unit mass of fuel (Heywood, 1988). That at CP is the most common heating value used in combustion; differences between heating values at CP and CV are small (Heywood, 1988).

Also of concern is whether the water in the products from combustion is in the liquid or gaseous state. If the heating value is defined using water in the liquid phase, this is defined as the *higher heating value* or *gross heating value*, Q_{HHV}. The *lower heating value* or *net heating value* (Q_{LHV}) defines the heating value when the water from combustion is in the vapor phase. These heating values are directly related, with the lower heating value more commonly used in combustion work. Heating values for fuels are tabulated (see Table 2 for the lower heating values of some typical fuels and hydrocarbons) and can be directly measured using calorimeters.

2.5 Maximum work

An IC engine can be treated as an open system (control volume) in which heat and work are exchanged with surroundings, and there is enthalpy flow of reactants and products into and out of the system. Hence, the first and second law of thermodynamics can be applied to the system to determine the *maximum useful work* that can be provided by the engine (Heywood, 1988).

Applying the first law of thermodynamics results in

$$\Delta Q - \Delta W_u = \Delta H \quad (40)$$

where ΔW_u is the usable work. The second law of thermodynamics tells us that

$$\frac{\Delta Q}{T_a} \leq \Delta S \quad (41)$$

As heat transfer occurs at atmospheric temperature (T_a), combining the two equations enables the definition of a maximum work relationship:

$$\Delta W_u \leq -(\Delta H - T_a \Delta S) \quad (42)$$

To achieve maximum work, the pressure and temperature of the reactants must reach the atmospheric values. For these conditions, the change in enthalpy and entropy can be calculated, which enables us to define the change in Gibbs free energy G at atmospheric temperature and pressure (T_a, P_a):

$$\Delta W_u \leq -(\Delta G)_{T_a,P_a} \quad (43)$$

The maximum work $\Delta W_{u,max}$ can then be defined as

$$\Delta W_{u,max} = -(\Delta G)_{T_a,P_a} \quad (44)$$

which is the difference in Gibbs free energy between the reactants and products at atmospheric pressure and temperature for complete combustion. This maximum work can be further related to other parameters such as availability conversion efficiency, the ratio of actual work delivered compared to the maximum work, and others. Refer to Heywood (1988) for details and further discussion. Also of note is that enthalpies and Gibbs free energies of combustion reactions for common hydrocarbons are very similar at the standard temperature, as shown in Table 2. As the Gibbs free energies are difficult to determine for typical hydrocarbon fuels, the heating value, which can be measured, can be used in the above equations for a close approximation.

Table 2. Enthalpies and Gibbs free energies of combustion reactions and lower heating values for typical fuels and hydrocarbons.

Hydrocarbon—combustion reaction	$\Delta\overline{h}^{\circ}_{298}$ (MJ/kmol) (H_2O in products is gas)	$\Delta\overline{g}^{\circ}_{298}$ (MJ/kmol)	LHV (MJ/kg)
Propane (C_3H_8)	−2044.0	−2074.1	46.4
Benzene (C_6H_6)	−3135.2	−3175.1	40.2
iso-Octane (C_8H_{18})	−5074.6	−5219.9	44.3
Gasoline	N/A	N/A	44.0
Ethanol	N/A	N/A	26.9
Diesel	N/A	N/A	42.5

(Modified from Heywood (1988).)

2.6 Equilibrium

In an IC engine, combustion does not yield the ideal results discussed above and requires the calculation of a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

complex and detailed set of chemical reactions and associated kinetics of a large number of species. One simplifying assumption can be made that the products of combustion are in chemical equilibrium, meaning the products of combustion are shifting during expansion to maintain equilibrium. Chemical equilibrium can be determined by the first and second law of thermodynamics, leading to

$$(\Delta G)_{\mathrm{P,T}} = 0 \tag{45}$$

as a definition of equilibrium for a CP and temperature process. The Gibbs free energy change can then be related to chemical potential and moles of species, which is further related via the ideal gas law to define equilibrium constants. Refer to Heywood (1988) for derivation and examples.

This equilibrium will result in differing combustion product compositions as a function of equivalence ratio, temperature, and pressure based on species and temperature-dependent equilibrium constants. As the temperature increases, products begin to dissociate into other products at more significant levels, such as to OH, O, and H (Heywood, 1988). For lean combustion, N_2, CO_2, H_2O, and O_2 remain the major species, while for rich mixtures CO and H_2 increase rapidly. In addition to these species, products can also include unreacted or partially reacted hydrocarbons due to quenching and thermal formation of oxides of nitrogen. These species require additional nonequilibrium analyses.

3 SECOND LAW DISCUSSION

In order to determine the entropy change across a reacting system, the importance of which will be discussed in the next section, it is necessary to calculate the entropy of all constituents in the fuel, oxidizer, and reactants. This requires setting a common reference state for both the pressure and temperature. In general, the common reference pressure P_{ref} is taken to be atmospheric, while the common reference temperature is taken at absolute zero, where the entropy of a substance is also zero. For an individual species i, the molar entropy is determined by

$$\overline{s}_i(T_{\mathrm{m}}, P_i) = \overline{s_i^{\mathrm{o}}}(T_{\mathrm{m}}, P_{\mathrm{ref}}) - \frac{R_{\mathrm{u}} \ln(x_i P_{\mathrm{m}})}{P_{\mathrm{ref}}} \tag{46}$$

where m represents the mixture, the subscript "ref" represents the reference atmospheric condition, the superscript "o" represents the entropy measured with respect to absolute zero temperature, and x_i is the mole fraction of the constituent. In order to find the total entropy of a mixture, the entropy of each individual constituent is multiplied by its mole fraction, and then all values are summed. It should be noted that, for real gas effects, modifications would have to be made to each individual constituent. Cengel and Boles (2010) provide a correction for using entropy departure charts.

4 ENERGY/LOSS ANALYSIS

In order to determine losses of potential work in a system, it is necessary to perform a second law analysis using either entropy or exergy concepts. Performing an entropy balance across a reacting system yields the following:

$$\sum S_{\mathrm{r}} - \sum S_{\mathrm{p}} + S_{\mathrm{gen}} + \sum_k \frac{Q_k}{T_k} = \Delta S_{\mathrm{sys}} \tag{47}$$

where r represents the reactants, p the products, "gen" the entropy generation term, and "sys" the system entropy change term. If the process is proceeding at steady state, the right-hand side will go to zero. Solving for S_{gen} and relating it to the exergy destroyed enables the determination of the loss of potential work through the system.

5 SIMPLIFIED ENGINE CYCLES

Ideal simplified models can be used to approximate engine cycles that rely on the developed and previously discussed thermodynamic concepts. These models consist of idealizing processes to approximate real engines, including intake, compression, combustion, expansion, and exhaust. The standard, simplified, ideal gas engine cycles include constant-volume, constant-pressure, and limited-pressure cycles, based on how chemical energy conversion to sensible energy (combustion) is occurring. The overall cycle includes an adiabatic intake stroke, adiabatic and reversible (isentropic) compression stroke, followed by combustion, which can occur at *CV, CP*, or part CV and part CP (*limited pressure*, *LP*), with assumptions of adiabatic and complete combustion. The remainder of the cycle includes an expansion process (adiabatic and reversible—isentropic) and is followed by an adiabatic exhaust stroke.

Although these simplified engine cycle analyses are ideal, they can approximate actual engine combustion strategies including fast combustion at top dead center (infinitely fast combustion—CV, spark ignition or Otto cycle), slow and late combustion (CP, diesel cycle), or combustion strategies in between (limited-pressure cycle). Each of the strategies can be independently analyzed to characterize the cycle processes including efficiencies, indicated mean effective

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

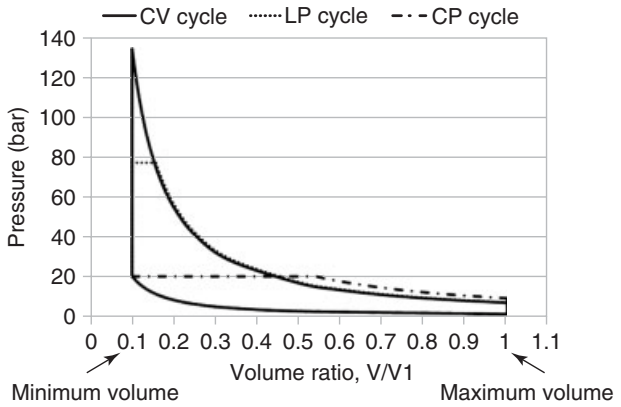


Figure 4. Constant volume, constant pressure, and limited pressure pressure–volume diagram for an ideal cycle analysis.

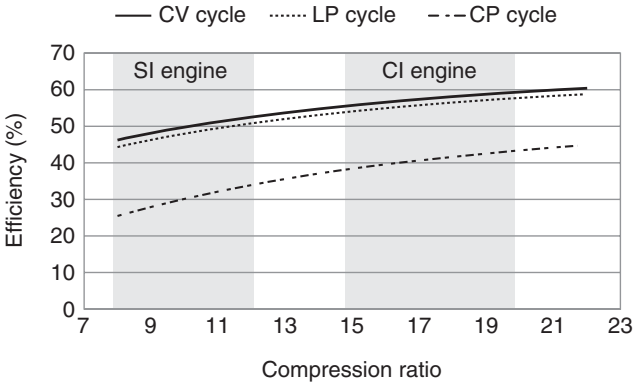


Figure 5. Constant volume, constant pressure, and limited pressure fuel-conversion efficiency as a function of compression ratio for an ideal cycle analysis.

pressure (IMEP), and values of maximum pressure ratios. The cycles will be summarized and efficiencies compared; however, full cycle derivations, which involve application of the first and second law of thermodynamics, as well as other combustion properties such as IMEP and maximum pressure ratios, are outside the current scope. Refer to Heywood (1988) or Stone (2002) for further details. The cycle assumes that the working fluid is an ideal gas with constant specific heats.

Shown in Figure 4 is a pressure–volume diagram for the three ideal cycles: CP, CV, and LP. These cycle diagrams are produced assuming that the engine is operating at a compression ratio of 10 (typical for SI engines), inlet pressure of 1 bar, and inlet temperature of 300 K. The engine is operating on gasoline, combusting in air, with a specific heat ratio of 1.3. The LP cycle assumes that 50% of the fuel is burnt at constant pressure, and 50% at CV.

The CV cycle, which has combustion occurring infinitely fast at CV, historically having been linked to SI engines, results in the largest peak pressure, as observed in Figure 4. The lowest peak pressure is observed with the constant pressure cycle, which has historically been associated with the diesel engine cycle, with combustion occurring at constant pressure, and hence the volume is allowed to expand to keep the pressure constant. Note that neither the CV nor the constant pressure cycle accurately represents current SI or CI diesel engines, but they represent the limits. The LP cycle, which most closely approximates an actual engine operating cycle, results in a peak pressure falling between these two values, as it combines a CV and constant pressure cycle. For this case, it is assumed that 50% of the fuel is burnt at constant pressure and 50% at CV; however, the split of fuel combustion between these stages can vary. As a larger fraction of the fuel is burnt at CV, the peak pressure will further increase, approaching the CV peak pressure. These different operating cycles also yield different fuel conversion efficiencies, as shown in Figure 5. Not only do the cycles yield different efficiencies at a set compression ratio but the compression ratio also directly influences the efficiency.

As the compression ratio is increased, the ideal fuel conversion efficiency increases, as shown. There are, however, limits, mechanically, on the maximum compression ratio based on increased pressures achieved and what the engine can physically withstand. Other limitations include engine knock, prevalent in SI engines, and NOx emissions. At a given compression ratio, the efficiency is largest for the CV cycle, followed by the LP, and finally the constant pressure cycle. Depending on the distribution of the fraction of fuel burnt at CV versus constant pressure for the LP cycle, the placement of the LP cycle with regard to efficiency between the CV and CP cycles changes.

It is typical in real engine operation that CI or diesel engines have a higher efficiency than SI engines. This is attributed in part to the differences in compression ratios, with diesels operating at higher compression ratios, typically in the 15–20 range, compared to gasoline engines in 8–12 range.

Alternative cycles are possible, such as the *Atkinson cycle* or *overexpanded cycle*, which involves an expansion stroke greater than the compression stroke. The advantages of this cycle are increases in work and in efficiency when compared to the standard CV cycle (Stone, 2002; Heywood, 1988). This engine operational strategy can be achieved with variable valve timing, and is prevalent in engines for hybrid electric vehicles. In addition to alternative cycles, modifications can be made in the analysis of the aforementioned cycles. For example, as opposed to an ideal

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

gas analysis, a fuel–air cycle can be considered, which has an unburned working fluid frozen in composition, and the burned gas mixture in equilibrium to provide a slightly improved approximation to actual engine cycles (Heywood, 1988). Actual engine cycles have lower efficiencies (about 80% less) compared to ideal cycles along with lower pressures. These differences can be attributed to heat transfer, the time required to burn the charge (noninstantaneous combustion), blow-down losses from the exhaust based on valve timing, gas flow and leakage into crevices in the combustion chamber, and incomplete combustion (Heywood, 1988).

In addition to these ideal cycle analysis, more complex cycle analysis can be applied using *one-zone* and *two-zone combustion modeling* to approximate realistically IC engine operating processes. To simplify, zero-dimensional combustion models are usually developed, which neglect engine flows and are classified into one-, two-, or multi-zone models. As the number of zones increases, the model complexity increases. In single-, or one-zone, models, the fluid (fuel, air, products) is a complete thermodynamic system and the first law of thermodynamics is applied to the system as a whole. For two-zone models, the engine fluid is treated in two states: a burned gas mixture (combustion products) and unburned gas mixture (fuel and air). Instead of just having one thermodynamic system, now there are two thermodynamic systems, with the first law applied to both zones. Additionally, there is energy and mass transfer between the two systems and the surroundings (e.g., the cylinder walls). Both models (one- and two-zone) enable the determination or calculation of in-cylinder pressure histories, energy release, or mass fraction burning rates, depending on known or assumed values including, for example, specific heat ratio, engine geometrical data, thermodynamic properties, and initial cylinder conditions of temperature. Greater detail on these types of models, including equation development, is provided in (see Pressure and Heat Release Analysis and Zero- and One-Dimensional Methodologies and Tools) and in Yeliana *et al*. (2008) and Yeliana (2010).

ENDNOTES

1. Two revolutions (rev) for a four-stroke engine. One rev for a two-stroke engine. This equates to n_R as utilized in Heywood (1988).
2. Molar concentration has units of moles or molecule per unit volume, the conversion factor being the Avogadro number.
3. Also referred to as the molar mass.

REFERENCES

Balachandran, P. (2006) *Fundamentals of Compressible Fluid Dynamics*, Prentice Hall, New Delhi.

Cengel, Y. and Boles, M. (2010) *Thermodynamics: An Engineering Approach*, 7th edn, McGraw Hill, New York.

Desantes, J.M., Lopez, J.J., Garcia, J.M., and Pastor, J.M. (2007) Evaporative diesel spray modeling. *Atomization and Sprays*, **17**, 193–231.

Gatowski, J.A., Balles, E.N., Chun, K.M., *et al.* (1984) Heat release analysis of engine pressure data. SAE Technical Paper 841359.

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*, McGraw Hill, New York.

Martin, C.J. (1979) Review: cubic equations of state—which?. *Industrial and Engineering Chemistry Fundamentals*, **18** (2), 81–97.

Moran, M.J. and Shapiro, H.N. (2008) *Fundamentals of Engineering Thermodynamics*, 6th edn, John Wiley & Sons, Inc., Hoboken.

Munson, B.R., Young, D.F., Okiishi, T.H., and Huebsch, W.W. (2009) *Fundamentals of Fluid Mechanics*, 6th edn, John Wiley & Sons, Inc., Hoboken.

Myers, G.E. (2007) *Engineering Thermodynamics*, 2nd edn, AMCHT Publications, Madison.

National Oceanic & Atmospheric Administration, Trends in Atmospheric Carbon Dioxide. Available at: http://www.esrl.noaa.gov/gmd/ccgg/trends/, 2011 (accessed 25 July 2013).

Nesbitt, J.E. (2011) Diesel spray mixing limited vaporization with non-ideal and multi-component fuel thermophysical property effects. PhD Thesis. Michigan Technological University.

Obert, E.F. (1973) *Internal Combustion Engines and Air Pollution*, Harper & Row, New York.

Reid, R.C., Prausnitz, J.M., and Poling, B.E. (1987) *The Properties of Gases and Liquids*, 4th edn, McGraw-Hill, New York.

Reynolds, W.C. (1979) *Thermodynamic Properties in SI*, Stanford University, Stanford.

Schroeder, D.V. (2000) *Introduction to Thermal Physics*, Addison Wesley Longman, New York.

Slavinskaya, N.A., Zizin, A., and Aigner, M. (2010) On model design of surrogate fuel formulation. *Journal of Engineering for Gas Turbines and Power*, **132**, 11501-1–11501-11.

Stone, R. (2002) *Introduction to Internal Combustion Engines*, Macmillan Press Ltd, London.

Taylor, C.F. (1985a) *Internal Combustion Engine in Theory and Practice*, *Thermodynamics, Fluid Flow, Performance*, vol. 1, 2nd edn, MIT Press, USA.

Taylor, C.F. (1985b) *Internal Combustion Engine in Theory and Practice*, *Combustion, Fuels, Materials, Design*, vol. 2, 2nd edn, MIT Press, USA.

Turns, S.R. (2000) *An Introduction to Combustion: Concepts and Applications*, 2nd edn, McGraw Hill, New York.

Van Wylen, G.J. and Sonntag, R.E. (1986) *Fundamentals of Classical Thermodynamics*, 3rd edn, John Wiley & Sons, New York.

Wark, K. (1988) *Thermodynamics*, 5th edn, McGraw-Hill, New York.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Wei, Y.S. and Sadus, R.J. (2000) Journal review: equations of state for the calculation of fluid-phase equilibria. *AIChE Journal*, **46** (1), 169–196.

Yeliana, Y. (2010) Parametric combustion modeling for ethanol-gasoline fueled spark ignition engines. PhD Thesis. Michigan Technological University.

Yeliana, Y., Cooney, C., Worm, J., and Naber, J. (2008) The calculation of mass fraction burn of ethanol-gasoline blended fuels using single and two-zone models. SAE Technical Paper 2008-01-0320.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto141
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Exhaust Emissions

M. Matti Maricq

Ford Motor Company, Dearborn, MI, USA

1 Introduction 1
2 Sampling Engine Exhaust 2
3 Emissions Measurement Methods 5
4 Advanced Methods of Emissions Measurement 9
5 Conclusions and Outlook for the Future 12
References 13

1 INTRODUCTION

Engine exhaust emissions are increasingly under regulatory scrutiny and, therefore, an increasingly important activity in motor vehicle development. Under the Clean Air Act (most recently amended in 1990), the US Environmental Protection Agency (USEPA, 2011a) set National Ambient Air Quality Standards (NAAQS) for six "criteria" pollutants: O_3, CO, NO_2, SO_2, lead, and particulate matter (PM). Regions that experience more than an allowed number of annual exceedances in daily or hourly averages (ranging from 1 to 8 h) of any of these pollutants are designated as "nonattainment" and subject to developing State Implementation Plans aimed at bringing the region into attainment. Four of the criteria pollutants include undesired byproducts of combustion: CO and PM arise from rich combustion when insufficient air is present (see UHC and CO Formation and Models and Particulate Formation and Models). NO_x ($NO + NO_2$) is produced at combustion temperatures by reactions between nitrogen and oxygen (see NO_x Formation and Models). SO_2 forms from the oxidation of organic sulfur compounds present in fossil fuels. Lead emissions occurred in the past because of the use of tetraethyl lead as an antiknock agent in gasoline; however, this practice ceased in the United States in the 1970s and has been, or is being, phased out in the rest of the world as well. Ozone is not directly emitted by combustion sources; rather, its ambient levels depend intimately on the tropospheric concentrations of hydrocarbons (HCs) and NO_x, which are emitted in engine exhaust among other sources.

Motor vehicle emissions are regulated by the US Environmental Protection Agency (USEPA, 2011b), California Air Resources Board (CARB, 2011a), the European Union (UNECE, 2011), Japan, India, and a growing number of other countries. The "standard" pollutants include CO, HCs, NO_x, and PM, but this list is growing. Additional regulations are beginning to appear for compounds such as formaldehyde and ammonia and for particle number counting. Although not traditionally thought of as pollutants, CO_2, N_2O, and CH_4 are coming under regulation owing to their global warming potential (USEPA, 2011c). Emissions standards are predominantly written for room temperature and sea level conditions. However, additional standards cover vehicle operation at cold temperature and altitude.

Besides a growing list of species to monitor, the emissions standards are tightening. The US EPA and CARB are about to promulgate their Tier 3 and LEV III standards, and the European Union is beginning to address its Stage 7 limits. The standards are complex in that they can vary with vehicle weight, vehicle type (passenger vs commercial), engine technology (diesel vs gasoline), and so on. The standards are often comprised of a number of "bins" to allow vehicle manufacturers some flexibility, that is, if they certify vehicles in higher emissions bins, they must offset this by sales of lower emitting vehicles. Vehicles must be certified for a full useful life, which increases

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Table 1. LEV II emission standards for light duty vehicles (<8500 lbs).

Category	120,000 miles/11 years				
	NMOG (g/mi)	CO (g/mi)	NO_x (g/mi)	PM (g/mi)	HCHO (g/mi)
LEV	0.090	4.2	0.07	0.01	0.018
ULEV	0.055	2.1	0.07	0.01	0.011
SULEV	0.010	1.0	0.02	0.01	0.004

to 150,000 miles under upcoming US regulations. Table 1 provides a simplified list of California's LEV II emissions standards. The categories low emitting vehicle (LEV), ultra low emitting vehicle (ULEV), and super ultra low emitting vehicle (SULEV) represent different emissions bins, where ULEV is the average required under LEV II and SULEV under LEV III.

Measuring emissions from modern vehicles, especially downstream of aftertreatment systems, poses a significant analytical challenge, as the species of interest, NO_x, HCs, PM, and so on, are typically present at parts per million (ppm) levels in a high temperature exhaust gas mixture comprised largely of CO_2 and water vapor, along with nitrogen and any uncombusted oxygen. Because high temperatures and water condensation can introduce substantial errors or damage-sensitive measurement equipment, correct sampling of the exhaust is at least as important as the measurements themselves. Regulatory methods stipulate measurement of the net pollutant emissions during a prescribed test or drive cycle (USEPA, 2011b; UNECE, 2011). According to regulations, the exhaust species are collected during the test, typically into Teflon bags, and measured afterwards. Engineering development, on the other hand, is better served by understanding the time-dependent history of each pollutant's emissions during the test or drive. Relating these two approaches to each other requires knowledge of the exhaust flow, which becomes an important auxiliary measurement.

The techniques to measure each species range from optical to electrical to gravimetric, and often multiple techniques exist. This chapter begins with the sampling methods used in emissions testing and then describes many of the typically used instruments. A number of new techniques are seeing increased application in emissions testing and two of these are treated in detail: Fourier transform infrared (FTIR) spectroscopy for gaseous emissions and real-time aerosol instrumentation to record PM emissions. The dynamometers employed to absorb the energy produced by engine operation and to mimic vehicle driving are also an important aspect of emissions testing. For descriptions of these devices and their use in engine testing, the reader is referred to Engine Performance.

2 SAMPLING ENGINE EXHAUST

2.1 Exhaust gas characteristics

Engine exhaust flow rate depends on displacement, speed, and load. It can vary from about 7 m^3/h (9 kg/h) at idle for a small 1.0-L automobile engine to 600 m^3/h (720 kg/h) for moderately high speed/load operation of a 6.5-L light-duty truck, and to higher rates for heavy-duty engines. The bulk composition of this exhaust is a consequence of combustion chemistry:

$$CH_{1.8} + 6.9\ \text{Air} \rightarrow CO_2 + 0.9\ H_2O + 5.45\ N_2 \quad (1)$$

Equation 1 applies to gasoline engines that operate at a stoichiometric air/fuel ratio so that little oxygen remains in the exhaust. $CH_{1.8}$ represents the average composition of petroleum-based fuels, predominantly a mixture of aliphatic and aromatic HCs. As a result, the exhaust is approximately 14% carbon dioxide and 12% water vapor by volume. The water fraction increases with the use of biofuels because of the oxygen in the fuel molecules. Combustion of E85 fuel (85% ethanol/15% gasoline), a highly oxygenated fuel blend, is described by

$$0.15\ CH_{1.8} + 0.85\ C_2H_5OH + 13.2\ \text{Air} \rightarrow 1.85\ CO_2 + 2.69\ H_2O + 10.4\ N_2 \quad (2)$$

Here, the water content in the exhaust climbs to 18%, whereas CO_2 falls to 12% by volume. On the other hand, water vapor and CO_2 fractions both decrease in diesel engine exhaust, which runs under lean conditions. At a typical diesel air/fuel ratio of 20 : 1, the water and CO_2 contents drop to about 9% and 10%, respectively.

The other major characteristic of engine exhaust is its temperature. As with flow, this depends on engine type, speed, and load. Typical gasoline engine-out exhaust varies in the range of 400–600°C and cools as it flows through the exhaust system. Diesel engines run somewhat cooler because of better thermal efficiency and higher air/fuel ratio, with exhaust temperatures normally between 200 and 400°C. However, for diesel vehicles equipped with particulate filters (diesel particulate filter, DPF), the exhaust temperature exceeds 600°C during active filter regeneration.

2.2 Raw sampling

There are two general strategies to sampling engine exhaust: (i) raw sampling and (ii) dilution sampling. Many commercially available emissions benches employ the former for

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

second-by-second emissions monitoring. Regulatory procedures for emissions certification, however, generally stipulate the second approach. There are advantages and disadvantages to each, as described in the following paragraphs.

The main advantage of raw sampling is that it preserves pollutant concentrations, a not inconsequential benefit under increasingly tighter regulatory standards. The primary disadvantage is the necessity to deal with hot humid samples, which can lead to losses, measurement interferences, and potentially, instrument damage. A straightforward approach is to employ heated transfer lines and detectors that prevent water condensation throughout the sampling and measurement processes, the so-called "wet" method. Heated lines offer the additional benefit of reducing losses during transport, such as losses of heavy HCs, ammonia, and PM by condensation and thermophoresis onto transfer line walls. Typical temperatures are 150°C for gasoline engine emissions measurements and 191°C for diesel measurements owing to their higher propensity for heavy HCs.

Raw sampling works well with methods such as flame ionization detection (FID) of HCs, as the detector operates at high temperature and is not sensitive to water vapor. However, it is less suitable in situations that are susceptible to interference from water vapor, for example, nondispersive infrared (NDIR) detection. This is a favored method for monitoring CO and CO_2; however, it has difficulty in distinguishing IR absorption by these compounds from that by water vapor. In such applications, one can "drop" out water by condensation leading to a "dry" measurement of emissions. However, it is important to correct the measured pollutant concentrations to account for the missing water, as this amounts to 9–18% of the exhaust, depending on fuel and engine types. The relationship between wet and dry concentrations of species x referenced to a common temperature (e.g., 25°C) is given by

$$c_{x\,\mathrm{wet}} = \frac{1 - \chi_{\mathrm{H_2O\,wet}}}{1 - \chi_{\mathrm{H_2O\,dry}}} c_{x\,\mathrm{dry}} \tag{3}$$

where $\chi_{\mathrm{H_2O\,wet}}$ and $\chi_{\mathrm{H_2O\,dry}}$ represent the mole fractions of water vapor in the exhaust and downstream of the condenser, respectively. Further details about wet and dry raw emissions calculations can be found in CFR 40 Part 89.412–89.418 (eCFR, 2011).

2.3 Dilution sampling

In contrast to the raw approach, dilution sampling cools the exhaust sample by mixing it with clean dry air or nitrogen (Hildemann, Cass, and Markowski, 1989). Heated lines are still required to avoid losses if there is any sample transport before dilution. Sufficient diluent is then mixed with the sample to ensure that it remains above the dew point at room temperature. Intermediate approaches are also possible where heated dilution air is used to cool an exhaust sample, but to a temperature higher than room temperature. This allows the use of lower dilution ratios and is therefore of interest when making very low level emissions measurements.

2.3.1 Direct exhaust sampling

There are a number of variations to dilution sampling. The simplest is to draw and dilute a constant volume of exhaust, as illustrated by the ejector pump system in Figure 1. The ejector pump is supplied with pressurized clean air or nitrogen at 2 atm overpressure, which expands past a nozzle creating a pressure drop that draws in exhaust through a short heated line. Heaters are used for the diluent and the first-stage diluter to avoid condensation and thermophoretic clogging by PM. A second stage (not shown) dilutes this mixture with air at room temperature to cool the sample without condensation. This method is relatively simple and

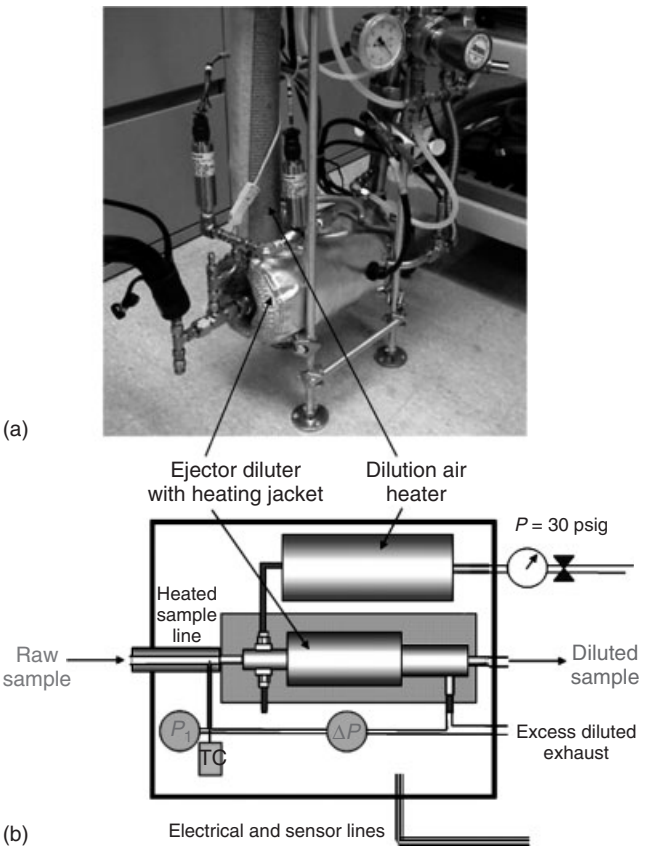


Figure 1. Dilution unit for direct tailpipe sampling. (a) Photo of an ejector pump diluter. (b) Schematic diagram of first stage of ejector pump dilution system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

can be made portable, as illustrated. It is particularly useful for engine-out measurements, where probe locations are often difficult to access and concentrations are relatively high; pre-DPF measurement of PM is an example.

2.3.2 Constant volume sampling

Constant volume sampling (CVS) takes a different tact, one usually applied to full exhaust flow. Figure 2 illustrates this approach for chassis dynamometer testing. Two variants are visible: (A) "remote mix tee" (RMT) and (B) dilution tunnel. The RMT connects to the vehicle tailpipe via a short, ∼1 m, stainless steel pipe. It is important to make this connection with metal and to avoid the use of silicone couplers or similar materials that can outgas HCs and shed particles leading to erroneous data. Temperature- and humidity-preconditioned dilution air (38 and -9°C dew point) enters the top of the RMT. However, the diluent flow is not kept constant; it rather varies and the total flow of exhaust plus diluent is kept constant. In the regulatory method, the diluted exhaust is then sampled into Teflon bags (marked "C" in Figure 2) to record integrated emissions over a drive cycle. The dilution air is simultaneously sampled into "ambient bags" to account for HCs, CO, and NO_x present in the dilution air.

The CVS approach draws diluent plus exhaust through a choked venturi, thereby creating a constant total flow. Consequently, the dilution ratio varies inversely with engine exhaust flow so that species concentrations in the CVS are proportional to the *fluxes* of emissions exiting the tailpipe. Raw sampling, in contrast, yields the *concentrations* of emissions in the exhaust. This is an important distinction because the CVS approach allows a simple calculation of test averaged emission rates directly from their CVS concentrations collected into Teflon bags,

$$\dot{M}_{x,\text{test}} = \frac{(\overline{C}_{x,\text{CVS}} \cdot V_{\text{CVS}} - \overline{C}_{x,\text{A}} \cdot V_{\text{A}})}{d} \tag{4}$$

where $\dot{M}_{x,\text{test}}$ is the test averaged mass emission rate, d is the distance traveled (in heavy duty testing, it is the energy utilized), $\overline{C}_{x,\text{CVS}}$ is the averaged mass concentration of species x in the sample bag, $\overline{C}_{x,\text{A}}$ is its concentration in the ambient bag, and V_{A} and V_{CVS} are the total volumes of diluent and exhaust plus diluent over the test. When $\overline{C}_{x,\text{CVS}} \gg \overline{C}_{x,\text{A}}$, the diluent volume can be approximated as $V_{\text{A}} = V_{\text{CVS}}\left(1 - \frac{1}{DF}\right)$, where DF is the average dilution factor. At very low emissions levels, *proportional ambient sampling* is used to improve the approximate ambient correction in Equation 4 (Silvis and Chase, 1999).

Dilution tunnels for PM measurement operate in the same manner, except that the engine exhaust must be transported to the upstream end of the tunnel where it mixes with dilution air. This transfer is conventionally done through a stainless steel hose that is insulated and heated to keep water from condensing and corrugated to allow flexibility to accommodate different size vehicles. The transfer hose in most facilities is many meters in length for logistical reasons, which presents a problem for low level emissions measurements because of storage/release artifacts. These arise from material deposited on the transfer hose walls in one test and then released by the exhaust heat of a subsequent vehicle test (Maricq *et al*., 1999). To minimize such artifacts, it is recommended to use an RMT approach also

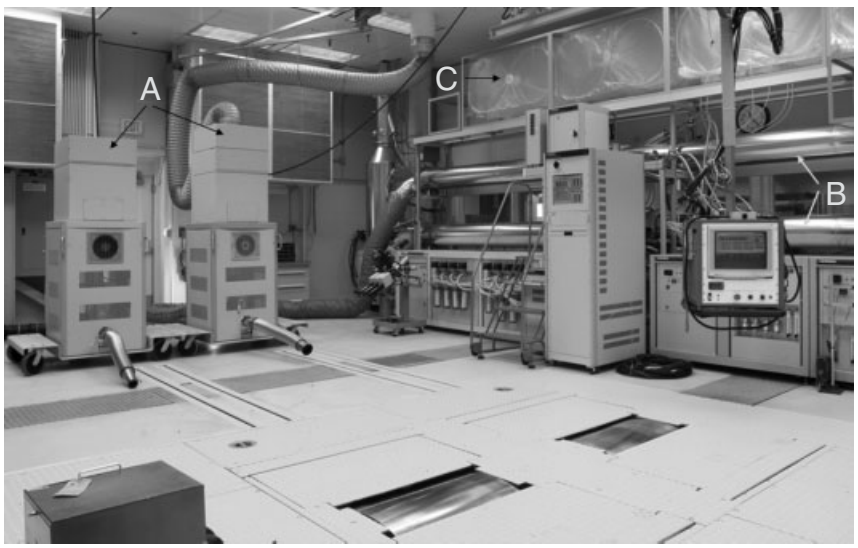


Figure 2. Chassis dynamometer test cell. "A" marks the remote mix "tees" for gasoline and diesel vehicles. "B" denotes the dilution tunnels that run along right-hand wall. And "C" shows the Teflon bags for integrated sampling.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

for dilution tunnel sampling; that is, dilute at the tailpipe with an RMT and deliver diluted exhaust to the tunnel.

As emissions levels continue to fall, it has become progressively more difficult to use the CVS approach as outlined earlier owing to the presence of pollutants in the dilution air. Attention is turning to a modification termed the *bag mini-diluter* (Guenther *et al*., 2000; Sun *et al*., 2005). In this approach, a small portion of the exhaust is sampled, diluted, and proportionally collected into Teflon bags. Proportional sampling implies that bag concentrations of HCs, CO, and NO_x are proportional to their emissions rates. Taking a small sample allows the use of "zero air" as a diluent, eliminating the need of a background correction as in Equation 4 for the ambient bag. A similar approach is feasible for PM measurement, but with the sample collected onto filters instead of into bags (Khalek *et al*., 2002).

2.4 Exhaust flow measurement

Exhaust flow is an important quantity in emissions measurement because it relates the mass per second of emissions exiting the tailpipe to their instantaneous mass concentrations,

$$M_x(t) = C_x(t)\, F_{\text{exh}}(t) \tag{5}$$

where x indicates HC, CO, NO_x, PM, or other species of interest and $F_{\text{exh}}(t)$ is the exhaust volume flow rate. There are two important considerations with respect to Equation 5: first, the concentration and flow measurements must be accurately aligned in time (<1 s) and second, the time response of the instruments recording species x concentration and exhaust flow rate must be comparable. A situation where this becomes problematic is the case of a long transfer hose because the delay time to the dilution tunnel then varies with exhaust flow rate.

Exhaust flow can be measured in a number of ways. Exhaust mass flow can be calculated from measured fuel and air intake rates based on conservation of mass, and the volume flow deduced from the stoichiometry of combustion. This method is often applied in engine dynamometer testing where fuel and air rates are recorded, but is less convenient in vehicle testing. A second approach is via CO_2 tracer. Combustion chemistry, for example, Equation 1, is used to relate CO_2 in the exhaust to the total flow including water vapor and remaining air. This method works well for gasoline engines where stoichiometry maintains a constant ratio between air and fuel. It becomes more problematic for diesel engines and gasoline engines with fuel shutoff, where the air and fuel rates vary independently. A third approach is followed in the CVS method. Here, a subsonic venturi is employed to record the diluent flow, and the difference between this and the constant total flow is equated to the vehicle exhaust flow. This method is preferred for vehicle testing, but requires some refinements for plug-in hybrid testing where the exhaust flow is near zero for extended times (Nevius and Rooney, 2010).

3 EMISSIONS MEASUREMENT METHODS

After emissions are sampled, raw or diluted, they are delivered to various instruments for concentration measurement. Multiple techniques exist for each species, but those used in the majority of emissions test cells include FID for HCs, chemiluminescent detection of NO_x, and NDIR absorption for CO and CO_2 (USEPA, 2011b; CARB, 2011b). Raw measurements require fast time response (1 or 0.1 Hz) and robustness against temperature and contamination. Dilute measurements need higher sensitivity but can trade off lower time resolution for regulatory applications.

3.1 Carbon monoxide and dioxide

Both CO and CO_2 exhibit strong absorptions of mid-infrared (IR) radiation owing to their molecular vibrations. Therefore, light penetrating a sample of these compounds experiences a decrease in intensity characterized by Beer's law,

$$I(\lambda) = I_0 \exp(-C_x \sigma_x(\lambda)\, L) \tag{6}$$

where C_x is the concentration of species x, $\sigma_x(\lambda)$ is its wavelength-dependent absorption cross section, and L is the path length through the sample cell. The optical cross section is a property of the molecule, but varies with pressure and temperature. These conditions and path length are fixed by instrument design, whereby concentration is calculated from light attenuation.

Figure 3 presents a schematic diagram of NDIR detection. Broadband IR radiation is produced by Globar, a small electrically heated SiC rod. The light is chopped

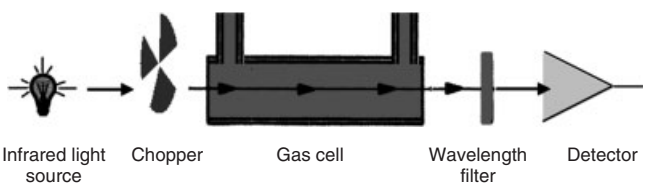


Figure 3. Schematic diagram of nondispersive infrared monitor. Various fixed wavelength filters are chosen to monitor CO, CO_2, N_2O, and other infrared active gases.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

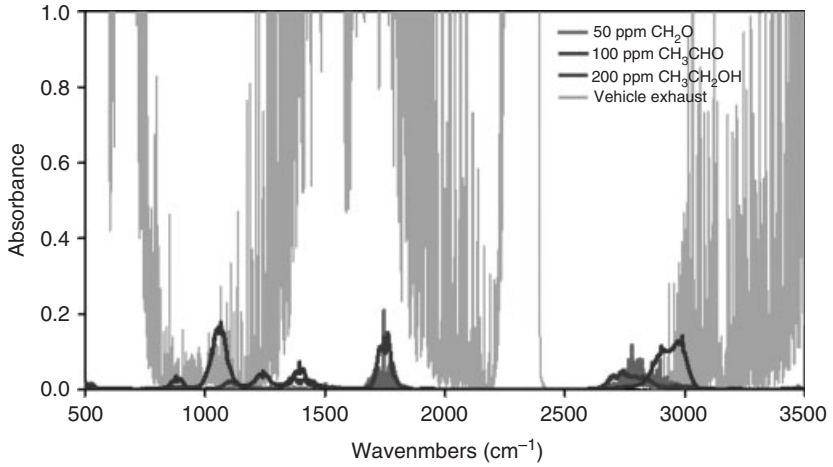


Figure 4. Infrared spectrum of a vehicle exhaust sample compared to reference spectra of ethanol, formaldehyde, and acetaldehyde, at concentrations expected for engine operation with ethanol-blended gasoline.

to take advantage of the lower electronic noise associated with AC detection. It passes through the sample cell and a wavelength-dependent filter and impinges on a pyroelectric or semiconductor detector. The wavelength-selective filter is chosen to optimize absorption by the compound of interest while minimizing interference from other compounds, CO_2, H_2O, and other emissions.

NDIR is generally limited to CO and CO_2 in vehicle emissions because of their relatively high concentrations and strong IR absorptions. Figure 4 illustrates why. It displays the IR spectra of formaldehyde, acetaldehyde, and ethanol at levels typical for gasoline engines using ethanol-blended fuels. Not only is IR absorption by these compounds swamped by CO_2 and water vapor, but they interfere with each other as well. In Section 4, we discuss the use of FTIR spectroscopy for emissions measurement, a dispersive method that provides substantially superior selectivity.

3.2 Hydrocarbons

Flame ionization provides a sensitive means to detect organic molecules (Sternberg, Gallaway, and Jones, 1962). The principle is illustrated in Figure 5. The sample is introduced into a hydrogen flame. Of the many high temperature reactions that occur as the material in the sample combusts, some produce ions and electrons (e.g., $CH + O \rightarrow CHO^+ + e^-$). These are collected in an electric field and recorded by a picoammeter.

The sensitivity of the FID to various organic compounds varies. For HCs, it acts similarly to a carbon counter; the relative response factor scales with carbon number, $RF = \alpha N_C$. As a result, HC concentrations are often expressed

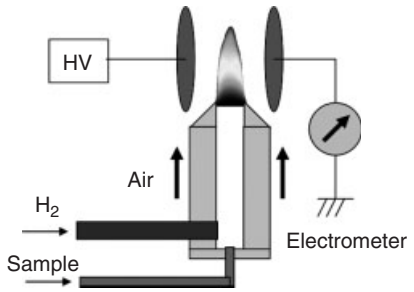


Figure 5. Schematic illustration of a flame ionization detector.

as parts per million carbon (ppmC), but sometimes they are reported as parts per million propane equivalent. At high concentrations, the FID can underestimate HC content as a result of the loss of signal due to ion recombination reactions. However, good design can achieve linearity over four orders of magnitude in concentration as well as millisecond time response (Cheng, Summers, and Collings, 1998).

Oxygenated compounds generally display a reduced FID response. FID response is negligible in the case of CO and CO_2, which is a benefit for engine exhaust applications, and also for carbonyl carbons in aldehydes, ketones, and esters. In the case of alcohols, the carbon with the OH group registers with a relative response of 40–75%, depending on the type of alcohol. For FID of specific organic compounds, such as in gas chromatography, it is possible to define effective carbon numbers (Scanion and Willis, 1985) and to calculate individual response factors from molecular structure and combustion enthalpies (Jorgensen, Picel, and Stamoudis, 1990; de Saint Laumer *et al.*, 2010). This does not apply to exhaust emissions measurements, which typically do not speciate HCs; thus,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

either additional measurements or empirical corrections are needed when testing with oxygenated fuels.

The presence of oxygen influences flame temperature and chemistry and, therefore, ion formation. Use of a hydrogen/helium mixture, usually 40/60, reduces the interference. It can be further compensated for through calibration. Spanning the FID against a known concentration of propane in nitrogen accounts for the lack of oxygen in gasoline vehicle exhaust, whereas for diesel emissions one can calibrate against propane in air.

Regulations call for measurement of non-methane hydrocarbons (NMHCs) and non-methane organic gases (NMOG). The former is accomplished by splitting the exhaust sample. Part is introduced directly into the FID to determine total HCs. The other portion first passes over a heated metal oxide catalyst that oxidizes the NMHCs, but not the relatively unreactive methane. The NMHC concentration is then found by subtraction. If oxygenated compounds are present, for example, gasoline engines run with ethanol/gasoline blends, they are underestimated by FID. For gasoline and low level blends (0–20%), the effect is small and can be accounted for by a correction factor, NMOG $= \alpha$ NHMC, where $\alpha = 1.03$ for gasoline and 1.1 for 10% ethanol.

3.3 Nitrogen oxides

3.3.1 Chemiluminescence analyzer

Chemiluminescence provides a very sensitive means to detect NO. It is based on the fact that when NO reacts with ozone to produce NO_2, a fraction of this product is electronically excited (indicated by the asterisk) (Clough and Thrush, 1967),

$$NO + O_3 \rightarrow NO_2 + O_2 \tag{7a}$$

$$\rightarrow NO_2{}^* + O_2 \tag{7b}$$

Some of these excited molecules are quenched by collisions with other molecules (Equation 8a); however, the rest emit light in the 600- to 3000-nm wavelength range (Equation 8b). To reduce possible interferences, the detected light is generally restricted with a filter to the 300- to 900-nm range.

$$NO_2{}^* + M \rightarrow NO_2 + M \tag{8a}$$

$$NO_2{}^* \rightarrow NO_2 + h\nu \tag{8b}$$

In practice, the linearity of chemiluminescent detection depends on a number of factors (Zabielski, Seery, and Dodge, 1984). First, the light intensity is a function of the production rate of $NO_2{}^*$ in Equation 7b. This is handled by using an excess of O_3 to create pseudo-first-order kinetic conditions, and using a capillary, critical orifice, or mass flow controller to regulate its flow rate. Second, the intensity of emitted light results from a competition between emission and quenching (Equations 8a and 8b) that depends on both the number density and chemical nature of the species present in the detector (Matthews, Sawyer, and Schefer, 1977). CO_2 and H_2O are more efficient quenchers than the N_2 and O_2 in air; therefore, an increasing presence of these in the exhaust sample can lead to underestimation of NO_x. These factors are mitigated by design and compensated for by calibration.

Other nitrogen compounds, NO_2 and NH_3, can also be monitored by chemiluminescence after first converting them to NO. It is possible to do this thermally (Sigsby *et al*., 1973), but usually a metal reactant is included to reduce the necessary operating temperature; for example,

$$3NO_2 + Mo \rightarrow 3NO + MoO_3 \tag{9}$$

Pretreatement of an exhaust sample via Equation 9 thus yields the NO_x concentration, whereas the NO_2 level is found by subtraction of the NO concentration measured without pretreatment. While ammonia can be measured by first oxidizing it to NO, its presence in engine exhaust interferes with the NO_x measurements. This interference can be eliminated by scrubbing NH_3 from the exhaust sample using filters treated with phosphoric acid (Shah *et al*., 2007).

3.3.2 Nondispersive UV

Recently, there has been an increased interest in nondispersive ultraviolet (UV) detection of NO_x owing to the inconvenience of chemiluminescence detectors for portable emissions measurement systems (PEMSs). The method follows the same principle as illustrated for NDIR in Figure 3, except for the use of a UV lamp and a detector. Both NO and NO_2 absorb in the UV, but NO has relatively sharp absorption lines. This allows it to be distinguished from NO_2 as well as from aromatic and unsaturated HCs that also exhibit UV absorption.

3.4 Particulate matter

3.4.1 Smoke meter

PM has traditionally been recorded by smoke meter in engine testing applications and gravimetrically in the regulatory method for vehicle exhaust measurement. The smoke

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

meter is a type of opacity meter that operates on the principle that soot decreases the transmission and reflection of light. It therefore measures the soot component of PM as opposed to the total PM. A filter-based smoke meter operates by drawing exhaust through a paper filter. As soot collects, it darkens the filter causing a reduction in its reflectivity. This reduction is converted into a smoke number (0–10 scale) by

$$SN = 10\left(1 - \frac{I}{I_{\text{clean}}}\right) \tag{10}$$

where I is the reflected light intensity from the blackened filter and I_{clean} is the reflected intensity from a clean filter. While this describes the basic principle, a variety of implementations exist and there are numerous conventions for defining *SN*. Converting *SN* to soot mass emissions is not straightforward as it depends on the nature of the soot, in particular, its optical properties. Further information on smoke number conventions and conversions is available from Homan (1985). Its slow time response and low sensitivity limits the use of the present-day smoke meter for steady-state engine-out measurements. However, its robustness against high soot levels makes the smoke meter a good choice in applications with high PM emissions.

3.4.2 Gravimetric mass

Regulatory methods stipulate a mass-based measurement of PM. The current practice is to direct the vehicle's exhaust (or engine's) into a CVS dilution tunnel, sample a fraction of the diluted exhaust through a filter (47 mm diameter, 2 μm pore size; Teflon membrane recommended over Teflon-coated glass fiber), and record the mass increase from the difference in pre and post-test filter weight. PM is a mixture of solid (soot and ash) and semivolatile materials (heavy HCs and sulfuric acid) that can vary depending on the engine, vehicle, and fuel. There is no calibration standard for PM; rather, the regulatory method operationally defines PM.

The lowering of US PM emissions standards in 2007 to 10 mg/mi for light-duty vehicles and 10 mg/hphr for heavy-duty engines, and an equivalent tightening of the mass standards in Europe and Japan, have necessitated substantial refinement of the gravimetric method. At these low emissions rates, less than 100 μg PM collects on the filter, which amounts to a few hundred parts per million change in filter weight. Regulations require the use of a 0.1-μg resolution balance to weigh the filters. At this sensitivity, temperature, humidity, air flow, ambient pressure, and static charge significantly impact balance performance. To reduce these impacts, weighing must be performed in a temperature- and humidity-controlled enclosure (e.g., $22.8 \pm 0.6^{\circ}$C, $50 \pm 1\%$ relative humidity, and $4\frac{1}{2}$ changes of HEPA-filtered low velocity air). The weighing is buoyancy corrected to account for any pre to post-test atmospheric pressure change. Moreover, the filters must be neutralized (using radioactive or corona bipolar ionizers) to reduce static charge, a particularly important step for Teflon filters (Chase *et al*., 2005).

As partitioning of semivolatile material between gas and particle phases depends on sampling conditions (temperature, dilution, ratio, and dilution rate), these are now specified in recent revisions to the regulatory PM method (USEPA, 2011b) in efforts to reduce PM measurement variability; for example, filter temperature is confined to $47 \pm 5^{\circ}$C. At the upcoming 3 mg/mi LEV III and Tier 3 standards, the sampling and weighing improvements remain inadequate. The amounts of PM collected are now so low that gaseous adsorption onto the filter media presents a sizable interference (Chase *et al*., 2004). This is illustrated in Figure 6 by post-DPF light-duty diesel exhaust PM recorded onto back-to-back Teflon filters. The primary and secondary filters show comparable weight gains, although they should differ by a factor of 100, as by design filter efficiency is >99%. Therefore, the secondary filter weight gain is attributed to gaseous adsorption. This implies a comparable gaseous adsorption onto the primary filter; thus, in Figure 6, the net PM in the exhaust is essentially zero. There are ongoing efforts to improve this situation, but applicability of the present gravimetric filter method becomes progressively more limited below 3 mg/mi (or 3 mg/hphr).

3.5 Greenhouse gas emissions

Owing to global warming concerns, greenhouse gas emissions from engines and vehicles have recently come under scrutiny (USEPA, 2011c). Compounds of significance to combustion engines include CO_2, CH_4, and N_2O. Water vapor is not included because its contribution to the greenhouse effect is already saturated. CO_2 and CH_4 are of interest for fuel economy and HC emissions reasons; therefore, their measurement is standard test cell procedure, as described earlier. N_2O is not produced during combustion; rather, it is a byproduct of incomplete NO_x reduction in catalyst aftertreatment systems. For model year 2012 and beyond, light-duty vehicle CH_4 and N_2O emissions will be capped at 0.030 and 0.010 g/mi, respectively (USEPA, 2011c).

There are a number of approaches to N_2O measurement, all variants based on the absorption of IR radiation. The standard method employs NDIR, as illustrated in Figure 3 and described earlier, for detection of CO and CO_2. This

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

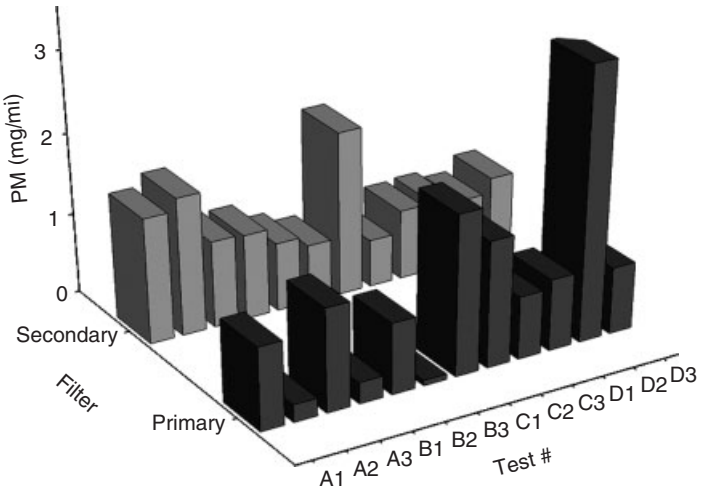


Figure 6. Demonstration of gaseous artifact at low level PM emissions. Back-to-back Teflon filters show comparable weight gains in post-DPF measurement of diesel PM emissions, even though filters are >99% efficient at collection PM.

method applies in principle to N_2O as well, but its exhaust concentrations are lower and spectral interferences are more problematic than for CO and CO_2. A version of this approach has recently been introduced, which uses photoacoustic detection. Absorption of a chopped IR beam by N_2O causes modulated heating of the exhaust sample, and this creates pressure waves that are detected by a microphone. This provides a sensitive means of detection, but it remains limited by the same spectral interferences as NDIR. Two better approaches, FTIR and laser spectroscopy, are described in Section 4.

4 ADVANCED METHODS OF EMISSIONS MEASUREMENT

The instruments described earlier have been improved and optimized for combustion applications over the years to where they are now standard equipment in emissions test cells. They are aimed at regulated emissions, whereas engine exhaust is, in reality, more complex. For research purposes, when developing new combustion concepts or aftertreatment systems, the ability to characterize the exhaust in more detail is valuable. This section describes some spectroscopic and aerosol tools that can provide this detail and are finding increased use in emissions testing.

4.1 Gaseous emissions

4.1.1 Fourier transform infrared spectroscopy

NDIR is a nondispersive optical technique that does not make full use of the wavelength-dependent nature of molecular absorption. In contrast, FTIR spectroscopy does; it records the entire mid-IR spectrum of radiation absorption, usually from 500 to 3500 cm^{-1} (inverse wavelength). Butler *et al.* (1981) first took advantage of this approach to measure engine emissions 30 years ago. Since then, there have been many instrumental improvements so that presently available spectrometers can sample raw exhaust and record emissions concentrations at 5 Hz time resolution.

A schematic diagram of FTIR operation is presented in Figure 7. It is similar to the NDIR approach, except that it replaces the fixed wavelength filter with a Michelson interferometer. IR radiation passes through the sample cell and is split into two beams: one reflects from a fixed mirror and the other from a movable mirror. When these beams are recombined, they interfere with each other in a way that depends on wavelength and path length difference.

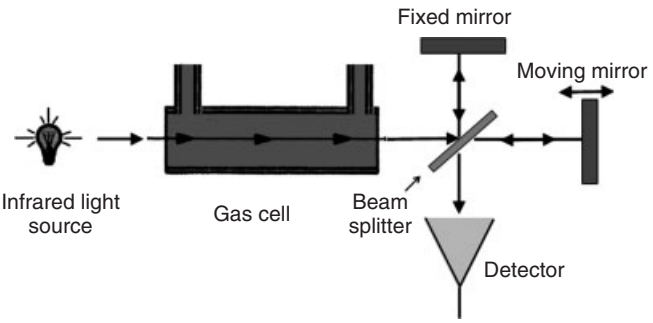


Figure 7. Schematic diagram of Fourier transform infrared spectrometer. Scanning the movable mirror generates an interferogram, the Fourier transform of which yields the infrared spectrum.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

This interference is recorded as a function of path length difference to yield an interferogram. Fourier transform of the interferogram gives the frequency spectrum of radiation transmitted through the cell, $I(\lambda)$. Comparing this spectrum with that of the empty cell, $I_0(\lambda)$, and using Beer's law, Equation 6, yields the IR absorbance of the sample, $A(\lambda) = \ln[I_0(\lambda)/I(\lambda)]$; an example is illustrated in Figure 4.

Atomic composition and bond strengths vary from one molecular species to another; thus, molecules exhibit characteristic vibrational frequencies that give each species its own IR "fingerprint." The idea then is to take the IR absorbance spectrum of an unknown mixture, such as an engine exhaust sample, and deconstruct it into a superposition of individual molecular fingerprints [i.e., their IR absorption cross sections $\sigma_i(\lambda)$],

$$A(\lambda, t) = L \sum_i C_i(t)\, \sigma_i(\lambda) \tag{11}$$

where L is the optical path length. An obvious benefit of this approach is the ability to determine from a single measurement the concentrations of a great number of exhaust constituents, $C_i(t)$. A second, less obvious, benefit is that the spectroscopic method does not need calibration or spanning. Because absorbance is determined as the ratio of two light intensities, neither the strength of the light source nor the sensitivity of the detector enters the calculation. In effect, the only "calibration" necessary is the wavelength-dependent IR cross section of each species, which is a property of the molecule and not the experimental apparatus.

The apparent simplicity of Equation 11 belies a number of technical details that must be carefully addressed to make reliable measurements by FTIR (Gierczak *et al*., 1991). First, it is necessary to collect a library of reference spectra, $\sigma_i(\lambda)$, relevant to engine exhaust. This includes H_2O, CO_2, CO, NO, NO_2, N_2O, HNO_3, NH_3, SO_2, SO_3, H_2SO_4, H_2S, HCHO, CH_4, and a host of other organic compounds. The list depends on engine technology (diesel, gasoline, and two-stroke) and on fuel (gasoline, ethanol blends, diesel, biodiesel, and natural gas) because of differences in exhaust composition. Furthermore, the reference spectra of these molecules depend on temperature and pressure. Temperature affects the extent to which the molecules rotate, which changes the vibrational lineshapes. Pressure determines the rate of molecular collision which in turn causes line broadening. Therefore, the reference spectra must be recorded at the same temperature and pressure as used for the exhaust measurements.

Fitting the entire IR spectrum of an exhaust sample to the library of exhaust species is computationally intensive and it is difficult to ensure that the fitting procedure finds a global best fit instead of a local minimum. In practice, one selects a set of compounds to be measured and identifies for each compound a spectral region, or regions, where it can be best monitored with the least inference from other exhaust species, a process called *masking*. The choice of regions depends on engine technology and fuel; a region free from interferences in one case may not remain so in another. As an example, the regions near 1050 and 2800 cm^{-1} in Figure 4 can be used to monitor ethanol, formaldehyde, and acetaldehyde in the emissions from gasoline engines run with ethanol-blended fuels. Much in the art of FTIR emissions measurements lies in choosing these regions and recognizing anomalous results that contradict the choice when these occur. When this is done successfully, FTIR provides a powerful tool for engine exhaust analysis.

4.1.2 *Laser spectroscopy*

The advent of tunable IR lasers (Pb-salt and quantum cascade laser) opens up the possibility of highly selective detection of CO, CH_4, HCHO, NO, NO_2, N_2O, NH_3, and numerous other small molecules. These lasers have very high spectral resolution and sufficient sensitivity to allow operation of the sample cell at subambient pressure. Lower pressure reduces broadening of the IR absorption lines and enables identification of absorption features specific to one compound and free from interference by other exhaust species. Unlike the FTIR, these laser-based methods do not record a broad spectrum, so that multiple lasers, or lasers that can rapidly hop from one spectral line to another ($<$0.1 s), are needed for simultaneous detection of multiple pollutants.

4.2 Aerosol methods for PM

4.2.1 *PM characterization*

Mass measurement of PM, as required by regulations, represents a one-dimensional metric of what is a complex substance. PM is a heterogeneous material both in terms of its physical (solid and liquid) and chemical (soot, ash, sulfate, and organic) properties. The constituents of motor vehicle PM have various origins. Soot and ash form as byproducts during combustion. Sulfate is converted from SO_2 by oxidation in the catalytic converter, whereas HCs are mostly removed by the catalyst. Some remaining HCs and sulfate condense on the internally mixed soot and ash particles as they cool in the exhaust system. Further cooling as the exhaust exits the tailpipe and mixes in the atmosphere can lead to nucleation of new particles, often

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

promoted by the presence of sulfate (Keskinen and Rönkkö, 2010).

A variety of aerosol tools exist to interrogate the nature of exhaust PM. Particle size is the most common attribute examined. The definition of size is clear for spherical droplets, but not so straightforward for agglomerates such as soot. In this case, size is based on the concept of equivalent diameter; that is, if a particle follows the same behavior as a spherical particle of diameter d, then it is assigned that diameter. There are many possible choices for equivalent diameter, but two common ones are aerodynamic and mobility. Aerodynamic diameter, d_a, is a measure of a particle's ability to follow flow streamlines. This depends on both the particle's size and mass, and so is referenced to unit density spheres. Mobility diameter, d_m, relates to a particle's diffusion rate, and is mass independent. Aerodynamic size distributions can be measured using cascade impactors; the electrical low pressure impactor (ELPI) combines this with electrical detection to provide transient measurement capability (Marjamäki *et al*., 2000). The scanning mobility particle sizer (SMPS) represents a preeminent tool to record mobility equivalent size distributions. Particles are first brought to a well-defined bipolar Boltzmann charge distribution and then passed through a cylindrical electric field. Those within a narrow range of electrical mobility exit into a condensation particle counter (CPC). In this case, size distributions are recorded by scanning the electric field; therefore, this method applies to steady-state engine operation. Transient capability has been achieved by arranging the cylindrical electric field as a stack of rings and recording the currents that arise on each disk as particles segregated by the field deposit (Reavell, Hands, and Collings, 2002).

Size distribution measurements often show that combustion exhaust PM is bimodal. The mode of larger particles (accumulation mode) is associated with soot formed in the engine, perhaps with some condensed semivolatile material. It is lognormal in shape, with a geometric mean mobility diameter typically between 50 and 80 nm and a universal geometric standard deviation of 1.6–1.8. The smaller (nucleation) mode generally lies below 20 nm, but at high concentrations can appear at larger size owing to growth by coagulation. If the diluted exhaust sample is heated in a thermodenuder before size measurement, the nucleation mode disappears in the majority of cases, indicating that these particles are semivolatile (organic and sulfate).

In this method, particles are distinguished as soot plus condensed HCs versus nucleated organics plus sulfate. They can instead be analyzed for elemental carbon (EC) versus organic carbon (OC). To accomplish this, PM is collected onto quartz filters. But instead of weighing, the filters are thermally analyzed, first in an inert atmosphere to desorb organic material, and then in an oxygen atmosphere to oxidize soot (Chow *et al*., 2001). Both the desorbed organic material and soot are converted to CO_2 for NDIR detection, or further to CH_4 and recorded by FID.

Tandem measurements open a wealth of possibilities for deeper investigation into particle properties (Park *et al*., 2008). Examples include particle volatility and effective density. In each case, a differential mobility analyzer selects particles of a certain mobility diameter. Particle volatility can be examined by passing these through a heat pipe and remeasuring their size (Sakurai *et al*., 2003). In this way, one learns how the EC/OC composition varies with particle size. Particle effective density can be determined by taking the mobility diameter-selected particles and measuring their aerodynamic diameter (Maricq and Xu, 2004) or mass (Park *et al*., 2003). The effective density differs from the intrinsic material density in that it averages both the material and the voids that exist when one describes fractal-like particle morphologies in terms of equivalent spheres. For soot, it varies approximately as

$$\rho_e = \rho_0 \left(\frac{d_m}{d_0} \right)^{(D_f - 3)} \tag{12}$$

where $\rho_0 = \sim 1.8\ \mathrm{g/cm^3}$ is the soot intrinsic material density, $d_0 = \sim 20\ \mathrm{nm}$ is the primary particle diameter, and $D_f = \sim 2.3$ is the mobility fractal dimension (Maricq and Xu, 2004).

4.2.2 *Solid particle counting*

The European Union has recently introduced a solid particle number standard as a regulatory requirement in addition to particulate mass (UNECE, 2011). The main impetus for this was to enforce adoption of DPF technology to reduce soot emissions. It differs from the conventional PM mass regulation in two important ways: replacement of mass with a number metric and the stipulation of solid as opposed to total PM. The latter is by necessity because particle count is very sensitive to variations in nucleation mode particles from test to test and test site to test site. But it is a positive step, as this decision recognizes that all PM is not alike.

The solid particle counting method is illustrated in Figure 8 and described in the review by Giechaskiel *et al*. (2012). As with PM mass measurement, vehicle exhaust is directed into a dilution tunnel. A sample from the dilution tunnel is further diluted in a heated diluter and then passed through a heated residence chamber ($\sim 350^\circ$C), the volatile particle remover (VPR). Besides eliminating liquid particles, this also evaporates condensed semivolatile

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

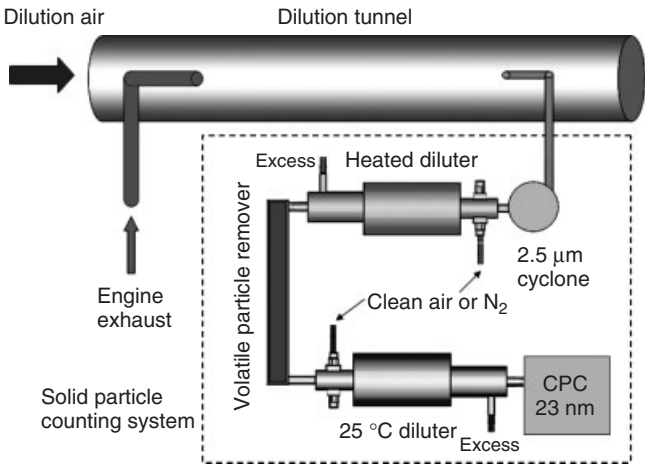


Figure 8. European Union PMP (Particle Measurement Program) method for solid particle counting.

material from soot particles, but this evaporation does not affect solid particle number. The sample is then cooled by dilution to room temperature. Gaseous concentrations at this point are too low for renucleation to occur, so a CPC registers solid particles.

This approach is very sensitive; it can record PM emissions at mass rates equivalent to <0.1 mg/mi. However, it requires significant care. The CPC needs to be calibrated against an electrometer to account for its size-dependent counting efficiency. Corrections must be made for size-dependent particle losses through the VPR. It is, however, not a unique means to record solid exhaust PM. The photoacoustic soot sensor (Schindler *et al*., 2004) also records the solid PM component of engine exhaust PM. The principle is the same as for the detection of gases. Modulated light is absorbed by soot (but not by semivolatile droplets) and converted to heat, which generates pressure waves detected by a microphone. Interestingly, photoacoustic measurements of motor vehicle soot mass correlate well with solid particle counts (Kirchner, Vogt, and Maricq, 2010) owing to the limited variability in the mean sizes and widths of soot size distributions observed in motor vehicle exhaust.

5 CONCLUSIONS AND OUTLOOK FOR THE FUTURE

Owing to regulatory requirements and reasons of corporate stewardship, engine exhaust emissions have joined performance, fuel economy, and safety as a central attribute in engine and vehicle design. Over the years, the scope has broadened considerably. Emissions regulations now apply to light-duty vehicles, heavy-duty on-road vehicles, off-road machinery, locomotives, ships, aircraft, snow mobiles, and small engines such as those found in lawn mowers and weed whackers. The principle species of CO_2, CO, HCs, and NO_x have been joined by HCHO, NH_3, N_2O, particle number, and more generally by the list of compounds associated with air toxics. Heavy-duty engine testing has evolved to include transient operation in addition to the conventional steady-state test points. Light-duty testing has seen additions of aggressive driving cycles and air conditioner cycles. Test procedures have been evolving to keep pace with tightening emissions stringencies.

The basic methods used to measure the standard emissions have largely remained the same: NDIR for CO and CO_2, FID for HCs and CH_4, and chemiluminescence for NO_x. However, the sensitivity of these instruments has increased 100-fold. In addition, many new measurement methods have been developed. One is the use of FTIR to simultaneously measure a variety of exhaust gas emissions. Second is the use of highly selective and sensitive laser-based optical instruments to measure NO, NO_2, NH_3, N_2O, and a growing list of other compounds. Third is the introduction of instruments such as the Dekati Mass Monitor and photoacoustic soot sensor to measure second-by-second PM emissions. More important, but perhaps less appreciated, are the changes that have occurred in sampling. As emissions levels have fallen, more care has been required in the sampling and dilution processes. Proportional ambient sampling was introduced to better account for pollutants present in the dilution air. New filter handling and weigh room procedures were introduced to improve PM measurement capability. At even lower emissions levels, one will need to utilize bag mini-diluters or partial flow dilution to record accurate data.

The challenges in emissions measurement are not abating. Impending next generation LEV III and Tier 3 regulations in the United States and Stage VI and VII regulations in the European Union are adding additional challenges to efforts aimed at fuel economy and greenhouse gas improvements. As other countries in Asia-Pacific, South America, and elsewhere are adopting emissions regulations, discussions have begun on world harmonized test procedures. The desire to understand real-world emissions and ensure that regulated emissions levels are maintained under real-world conditions has driven the development of on-board PEMSs, so that emissions testing is no longer confined to test cells. These and other activities ensure that emissions measurement will remain a vibrant and important field for some time to come.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

REFERENCES

Butler, J.W., Maker, P.D., Korniski, T.J., and Haack, L.P. (1981) On-line characterization of vehicle emissions by FT-IR and mass spectrometry. SAE Technical Paper 810429.

California Air Resources Board (2011a) Amendments to the Low-Emission Vehicle Program—LEV III, http://www.arb.ca.gov/msprog/levprog/leviii/leviii.htm (accessed 30 November 2012).

California Air Resources Board (2011b) Test Methods for Vehicle Exhaust Hydrocarbon Speciation, http://www.arb.ca.gov/testmeth/slb/exhaust.htm (accessed 8 July 2013).

Chase, R.E., Duszkiewicz, G.J., Lewis, D., and Podsiadlik, D.H. (2005) Reducing PM measurement variability by controlling static charge. SAE Technical Paper 2005-01-0193.

Chase, R.E., Duszkiewicz, G.J., Richert, J.F.O., *et al*. (2004) PM measurement artifact: organic vapor deposition on different filter media. SAE Technical Paper 2004-01-0967.

Cheng, W.K., Summers, T., and Collings, N. (1998) The fast-response flame ionization detector. *Progress in Energy and Combustion Science*, **24**, 89–124.

Chow, J.C., Watson, J.G., Crow, D., *et al*. (2001) Comparison of IMPROVE and NIOSH carbon measurements. *Aerosol Science and Technology*, **34**, 23–34.

Clough, P.N. and Thrush, B.A. (1967) Mechanism of chemiluminescent reaction between nitric oxide and ozone. *Transactions of the Faraday Society*, **63**, 915–925.

de Saint Laumer, J.Y., Cicchetti, E., Merle, P., *et al*. (2010) Quantification in gas chromatography: prediction of flame ionization detector response factors from combustion enthalpies and molecular structures. *Analytical Chemistry*, **82**, 6457–6462.

eCFR (2011) Electronic Code of Federal Regulations. Title 40: Protection of the Environment, http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&rgn=div5&view=text&node=40:20.0.1.1.3&idno=40#40:20.0.1.1.3.5.1.12 (accessed 15 August 2013).

Giechaskiel, B., Mamakos, A., Andersson, J., *et al*. (2012) Number emissions within the European legislative framework: a review. *Aerosol Science and Technology*, **46**, 719–749.

Gierczak, C.A., Andino, J.M., Butler, J.W., *et al*. (1991) FTIR: fundamentals and applications in the analysis of dilute vehicle exhaust. *Proceedings of SPIE*, **1433**, 315–328.

Guenther, M., Henney, T., Silvis, W.M., *et al*. (2000) Improved bag mini-diluter system for ultra-low level vehicle exhaust emissions. SAE Technical Paper 2000-01-0792.

Hildemann, L.M., Cass, G.R., and Markowski, G.R. (1989) A dilution stack sampler for collection of organic aerosol emissions: design, characterization and field tests. *Aerosol Science and Technology*, **10**, 193–204.

Homan, H.S. (1985) Conversion factors among smoke measurements. SAE Technical Paper 850267.

Jorgensen, A.D., Picel, K.C., and Stamoudis, V.C. (1990) Prediction of gas chromatography flame ionization detector response factors from molecular structures. *Analytical Chemistry*, **62**, 683–689.

Keskinen, J. and Rönkkö, T. (2010) Can real-world diesel exhaust particle size distribution be reproduce in the laboratory? A critical review. *Journal of the Air and Waste Management Association*, **60**, 1245–1255.

Khalek, I.A., Ullman, T.L., Shimpi, S.A., *et al*. (2002) Performance of partial flow sampling systems relative to full flow CVS for determination of particulate emissions under steady-state and transient diesel engine operation. SAE Technical Paper 2002-01-1718.

Kirchner, U., Vogt, R., and Maricq, M. (2010) Investigation of EURO-5/6 level particle number emissions of European diesel light duty vehicles. SAE Technical Paper 2010-01-0789.

Maricq, M.M., Chase, R.E., Podsiadlik, D.H., and Vogt, R. (1999) Vehicle exhaust particle size distributions: a comparison of tailpipe and dilution tunnel measurements. SAE Technical Paper 1999-01-1461.

Maricq, M.M. and Xu, N. (2004) The effective density and fractal dimension of soot particles from premixed flames and motor vehicle exhaust. *Journal of Aerosol Science*, **35**, 1251–1274.

Marjamäki, M., Keskinen, J., Chen, D.R., and Pui, D.Y.H. (2000) Performance evaluation of the Electrical Low-Pressure Impactor (ELPI). *Journal of Aerosol Science*, **31**, 249–261.

Matthews, R.D., Sawyer, R.F., and Schefer, R.W. (1977) Interferences in chemiluminescent measurement of NO and NO2 emissions from combustion systems. *Environmental Science and Technology*, **11**, 1092–1096.

Nevius, T.A. and Rooney, R.T. (2010) Improved PHEV emissions measurements in a chassis dynamometer test cell. SAE Technical Paper 2010-01-1295.

Park, K., Cao, F., Kittelson, D.B., and McMurry, P.H. (2003) Relationship between particle mass and mobility for diesel exhaust particles. *Environmental Science and Technology*, **37**, 577–583.

Park, K., Dutcher, D., Emery, M., *et al*. (2008) Tandem measurements of aerosol properties—a review of mobility techniques with extensions. *Aerosol Science and Technology*, **42**, 801–816.

Reavell, K., Hands, T., and Collings, N. (2002) A fast response particulate spectrometer for combustion aerosols. SAE Technical Paper 2002-01-2714.

Sakurai, H., Tobias, H.J., Park, K., *et al*. (2003) On-line measurements of diesel nanoparticle composition and volatility. *Atmospheric Environment*, **37**, 1199–1210.

Scanion, J.T. and Willis, D.E. (1985) Calculation of flame ionization detector relative response factors using the effective carbon number concept. *Journal of Chromatographic Science*, **23**, 333–340.

Schindler, W., Haisch, C., Beck, H.A., *et al*. (2004) A photoacoustic sensor system for time resolved quantification of diesel soot emissions. SAE Technical Paper 2004-01-0968.

Shah, S.D., Mauti, A., Richert, J.F.O., *et al*. (2007) Measuring NOx in the presence of ammonia. SAE Technical Paper 2007-01-0331.

Sigsby, J.E., Jr, Black, F.M., Bellar, T.A., and Klosterman, D.L. (1973) Chemiluminescent method for analysis of nitrogen compounds in mobile source emissions nitric oxide, nitrogen dioxide, and ammonia. *Environmental Science and Technology*, **7**, 51–54.

Silvis, W.M. and Chase, R.E. (1999) Proportional ambient sampling: a CVS improvement for ULEV and lean engine operation. SAE Technical Paper 1999-01-0154.

Sternberg, J.C., Gallaway, W.S., and Jones, D.T. (1962) The Mechanism of Response of Flame Ionization Detectors in *Gas*

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Chromatography (eds N. Brenner, J.E. Callen, and M.D. Weiss), Academic Press, New York, pp. 231–267.

Sun, E.I., McMahon, W.N., Peterson, D., *et al*. (2005) Evaluation of an enhanced constant volume sampling system and bag mini diluter for near zero exhaust emission testing. SAE Technical Paper 2005-01-0684.

UNECE (2011) Regulations No. 83 Uniform Provisions Concerning the Approval of Vehicles with Regard to the Emission of Pollutants According to Engine Fuel Requirements, http://www.unece.org/fileadmin/DAM/trans/main/wp29/wp29regs/r083r4e.pdf (accessed 26 April 2011).

United States Environmental Protection Agency (2011a) National Ambient Air Quality Standards, http://www.epa.gov/air/criteria.html (accessed 14 December 2012).

United States Environmental Protection Agency (2011b), Code of Federal Regulations, Title 40—Protection of the Environment, Part 1065—Engine Testing Procedures, http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&tpl=/ecfrbrowse/Title40/40cfr1065_main_02.tpl (accessed 15 August 2013).

United States Environmental Protection Agency (2011c), Transportation and Climate, Regulations and Standards, http://www.epa.gov/otaq/climate/regulations.htm#noticeI (accessed 14 May 2013)

Zabielski, M.F., Seery, D.J., and Dodge, L.G. (1984) Influence of mass transport and quenching on nitric oxide chemiluminescent analysis. *Environmental Science and Technology*, **18**, 88–92.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto142
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Trends—Spark Ignition

Stefan Pischinger[1], Martin Nijs[1], Philipp Adomeit[2], Dieter Seebach[1], Bastian Lehrheuer[1], Thomas Dünschede[2], Adrien Brassat[1], Karsten Wittek[2], Tolga Uhlmann[1], Gregor Schürmann[1], Björn Höpke[1], Moritz Bähr[1], Joao Serpa[1], and Axel Kuhlmann[1]

[1]*RWTH Aachen University, Aachen, Germany*
[2]*FEV GmbH, Aachen, Germany*

1	Introduction	1
2	Downsizing and Downspeeding	4
3	Boosting	6
4	Variable Valve Trains	9
5	Potential of Variable Valve Trains	11
6	SI for Hybrids	13
7	Advanced Combustion Concepts	18
8	Variable Compression Ratio	19
9	Novel Engine Design	21
10	Thermal Management	23
References		24

1 INTRODUCTION

Current and future combustion engine development targets a considerable reduction in fuel consumption and emissions to meet strict emission legislation despite steady or increasing customer comfort and performance demands. Additionally, reduced time to market is required despite increasing variations in customer demands and fuel qualities in worldwide markets. It can also be observed that the fuel consumption, or equivalently CO_2, has already been reduced significantly over the last decade (Figure 1).

In the European Union (EU), a vehicle-weight-dependent CO_2 emission limit of 130 g per km at a base vehicle weight of 1372 kg has been set for 2012 with increasing penalties until 95 € per g CO_2 per km. For 2020, the intention is to reduce this limit to 95 g per km (European Union, 2009a, b). A comparison of the limit for 2012 in the EU with the CO_2 emissions in the new European driving cycle (NEDC) for gasoline and gasoline hybrid passenger cars licensed in 2011 shows that only the best in class cars in the small and the lower middle sized segment meet the targets (Figure 2). For 2020, further fuel consumption reductions of >30% are necessary to fulfill the requirements. In the United States, a 12% fuel consumption reduction is required between 2012 and 2016 (Environmental Protection Agency and Department of Transportation, 2010), which follows a similar trend as Europe.

To achieve these limits, gasoline engines have made progress in efficiency through the introduction of technologies such as direct injection (DI) in combination with downsizing and turbo- or supercharging, variable valve trains, and a variety of additional measures to reduce friction losses such as electrification of accessories.

The high specific power of modern engines indicates the rising degree of downsizing (Figure 3). Without any additional measures, downsizing would lead to lower maximum power output contrary to customer's demands. Consequently, reasonable downsizing concepts provide high power density and correspondingly high specific power.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

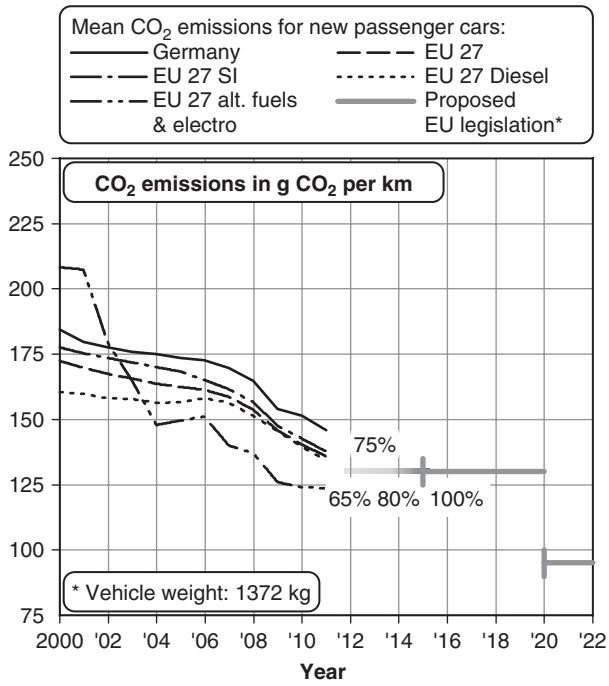


Figure 1. CO_2 emissions in the EU. (Reproduced from Central Data Repository, 2011. © European Environment Agency.)

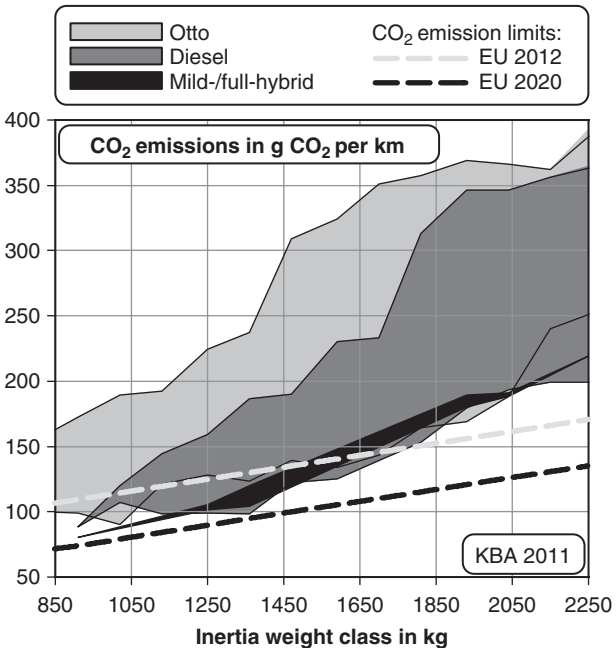


Figure 2. CO_2 Emissions in the NEDC 2011. (Reproduced from KBA, 2011. © Federal Motor Transport Authority.)

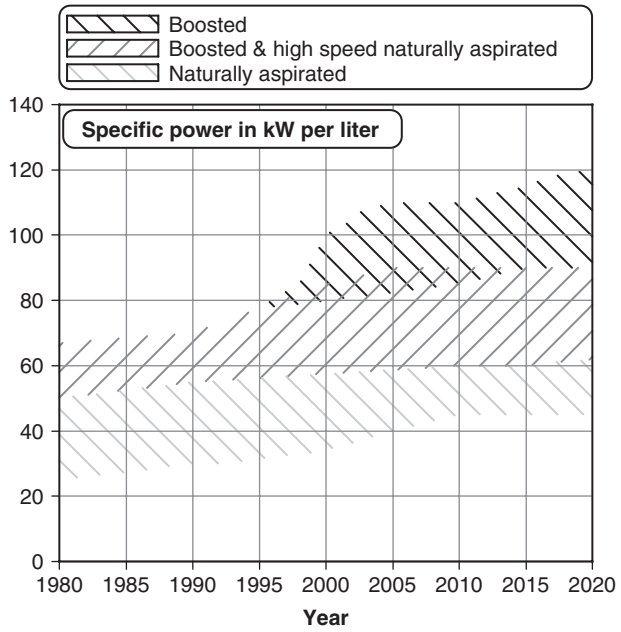


Figure 3. Trend specific power (FEV GmbH, 2011). (Reproduced by permission of FEV GmbH.)

Higher power output can be achieved by either increased engine speed or higher load. While increased engine speed would lead to rising friction losses, current and future downsizing concepts use boosting or electrification to ensure the required reserve torque. The potential fuel benefits of highly boosted downsized engines compared to natural aspirated engines with equal power output amount to 10–30% (Golloch, 2005). The rising model share of boosted gasoline engines in Europe substantiates the boosting/downsizing trend (Figure 4). It can be assumed that this trend will be followed by the United States and other markets as well.

The higher the degree of downsizing, the higher is the negative effect of turbocharging on engine dynamics during transient operation (turbo lag, see Intake Boosting). Current engine development shows different ways to improve the transient response, as there are advanced charging technologies such as twin-scroll, two-stage turbocharging, or variable turbine geometries. Turbocharging in combination with variable valve timing (VVT) or lift can also improve the engine's transient behavior. Despite the higher costs, these technologies are already in use in all vehicle classes. Supercharging in combination with variable valve lift and DI enables scavenging and thereby an improved low-end torque (Figure 5). Current variable valve lift systems reach fuel consumption benefits of up to 10% by reducing throttling losses and controlling residual gas fraction. For engines with six or more cylinders, cylinder deactivation

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

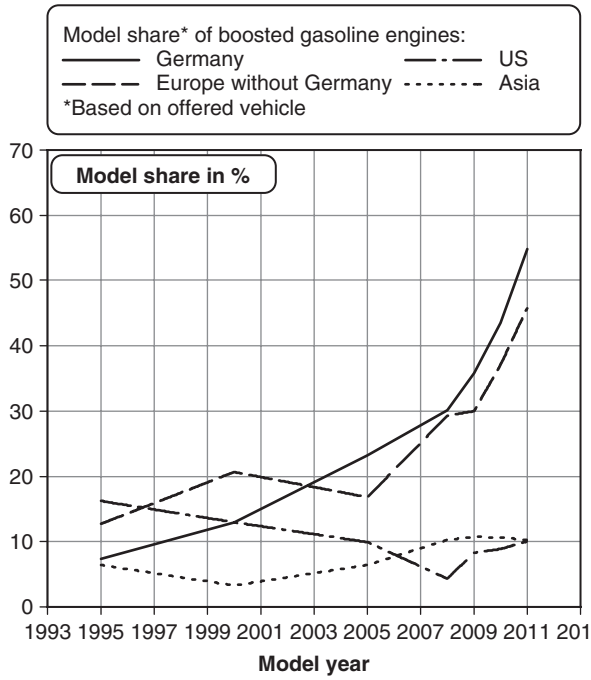


Figure 4. Trend boosted gasoline engines (FEV GmbH, 2011) (Reproduced by permission of FEV GmbH.)

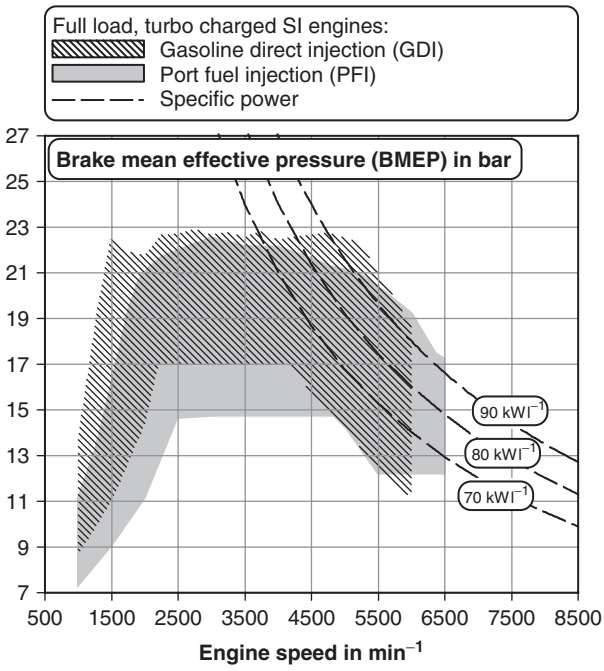


Figure 5. GDI compared to PFI (33 engines, FEV GmbH, 2011). (Reproduced by permission of FEV GmbH.)

can reduce fuel consumption by up to 18% in part load operation.

High supercharging leads to an increased propensity for preignition and irregular combustion (knocking). Owing to that, the degree of downsizing is limited. However, a high compression ratio (CR) is necessary for best efficiency. The adaption of DI on gasoline engines has helped lower the knock tendency and increase the CR by 1 to 1.5 units, thereby reducing the fuel consumption by up to 4% at part load operating points (Figure 6). The high CR of current gasoline engines indicates the rising share of DI (Figure 7).

Moreover, new concepts with variable CR promise further improvement regarding knock tendency and an increased fuel efficiency. Hereby, the investigations of variable CR focus especially on reducing complexity and costs while increasing usability and durability.

Another advantage of DI is the lean engine operation with stratified charge. Wall- or air-guided injection has already been used in series production vehicles since 1997. In 2006, spray-guided combustion systems with high pressure injection (200 bar) and multihole injector tips for fuel metering went into production, which improved stratification and thereby gaining fuel consumption benefits.

So far, lean engine operation is limited by reduced ignitability of the diluted air fuel mixture. Conventional ignition systems fail at local air fuel ratios higher than 22 ($\lambda > 1.5$). One solution to raise this limit and simultaneously decrease cyclic combustion variations is provided by a scavenged prechamber ignition or the corona ignition system. Contrary to the conventional local ignition, the spark corona ignition systems generate a significantly larger, intense plasma ignition source, which spreads throughout the combustion chamber.

For conventional stoichiometric-operated combustion engines, the three-way catalyst (see Gas Aftertreatment Systems) is the most commonly used for the simultaneous reduction of HC, CO, and NO_x. In lean engine operation, HC and CO are still converted by the three-way catalyst, whereas for the reduction of NO_x, additional aftertreatment technologies, such as SCR systems or NO_x storage catalysts (see Gas Aftertreatment Systems), have to be applied. With the upcoming legislation for exhaust emissions (EU6), the particulate number will be limited for DI gasoline engines. Current series production engines show that the first phase-in limit of 6×10^{12} # per km in 2015 can be achieved by optimized combustion systems. To meet the particle number limit of 6×10^{11} # per km starting 2018, DI engines might need a gasoline particulate filter (GPF) (European Union, 2011).

A further measure to reduce CO_2 and other emissions are alternative fuels using, for example, alcohol blending. Ethanol is most likely the short-term solution for gasoline engines. For 2020, the target for the overall use of renewable energies within the transportation sector is set to 10% in the European Directive 2009/28/EC.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

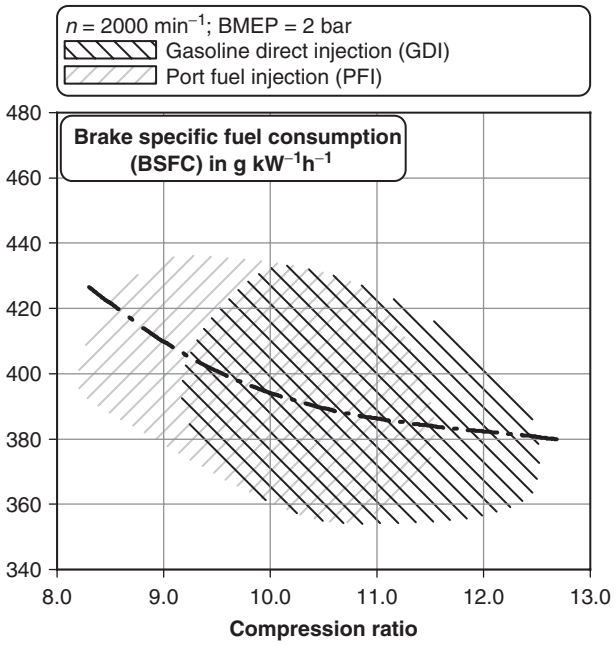


Figure 6. Influence of compression ratio on fuel consumption (127 engines, FEV GmbH, 2011). (Reproduced by permission of FEV GmbH.)

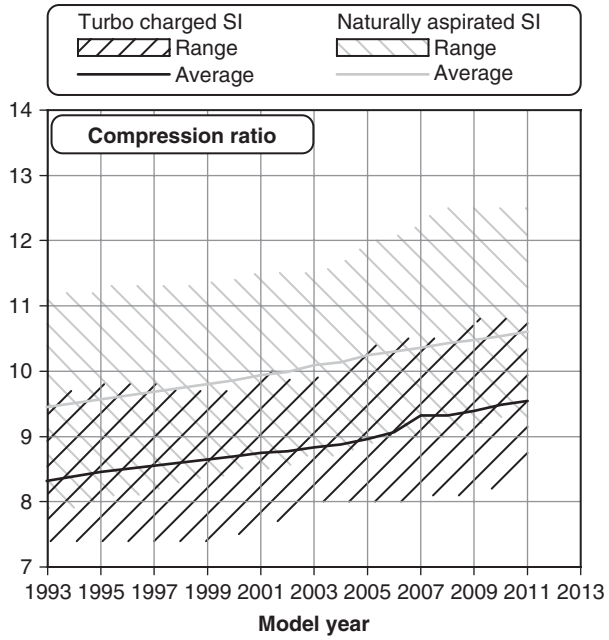


Figure 7. Trend compression ratio (FEV GmbH, 2011). (Reproduced by permission of FEV GmbH.)

In addition to these technologies, a further improvement in fuel consumption of up to 10% is possible through reduced friction losses (see Lubrication and Friction) and using advanced thermal management (see Engine Thermal Management). High potential to reduce friction losses is assigned to variable auxiliary drives such as variable oil or water pumps. The engine itself can be optimized regarding friction either by design or by material. Additionally, current research examines new coatings or fuel additives for further friction reductions.

The following example gives an impression of how a reasonable combination of the beneficial measures mentioned above can reduce CO_2 emissions and improve fuel efficiency. Assuming a 1.6-L I4 natural aspirating port fuel injection (PFI) engine with intake and exhaust cam phasing as a baseline (with relative CO_2 emissions defined as 100%), applying DI can already gain 2% CO_2 reduction (Figure 8). In combination with variable valve lift, a further 6.5% reduction is possible. Replacing this engine by a 1.0-L I3 boosted engine with equal performance gains a 12% fuel consumption benefit. Using additional measures such as external exhaust gas recirculation (EGR), variable CR, optimized thermal management, and friction reduction, a total CO_2 reduction of 27% can be achieved. Optimized automatic transmissions (see Engine/Transmission Matching) and hybridization (see Overview of Electric, Hybrid and Fuel Cell Vehicles) are necessary to achieve a reduction of >30% and meet the future emission targets mentioned above. As shown in Figure 8, measures to reduce CO_2 emissions come along with additional costs relative to the baseline, which have to be considered in the context of possible penalty payments for not meeting CO_2 fleet emissions.

2 DOWNSIZING AND DOWNSPEEDING

Currently, downsized gasoline engine concepts combined with boosting, gasoline DI, and VVT provide the major directions for meeting the desired fleet fuel economy targets. By increasing the engine load, downsized engines are operated with reduced specific fuel consumption. Thus, fuel efficiency can be improved by up to 15% for the "NEDC" (Weinowski *et al*., 2010). In addition to downsizing, lower piston velocities resulting from downspeeding lead to lower friction losses and thereby lower specific fuel consumption.

To maintain the power output equal to current natural aspirated engines, boosting technologies are applied to downsized engines (Alt *et al*., 2001; Brinkmann, Pingen, and Walder, 2003; Bormann *et al*., 2009). When the same power output is demanded, the engine load of the downsized engine is increased. Thus, the engine operation point is shifted to more efficient engine operating conditions. The de-throttling allowed by downsizing leads to reduced gas

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

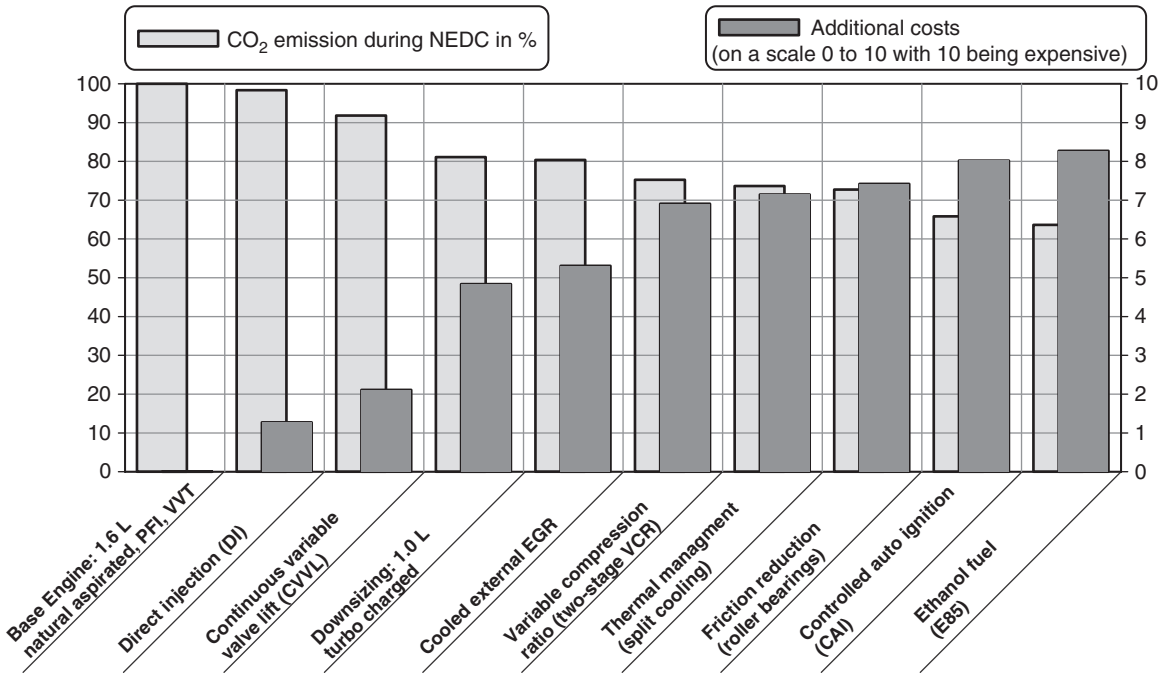


Figure 8. Measures to reduce CO_2 emissions contrary to additional costs.

exchange losses and reduced friction. Therefore, the overall engine efficiency is improved. The comparison of a conventional 1.4 L, NA engine, and a 0.7 L turbocharged engine, featuring gasoline DI and independent intake and exhaust cam phasing for NEDC is shown in Figure 9.

Downsized engine concepts demand a lower engine displacement, which can be achieved with a reduced number of cylinders. For the small vehicle class (800–1200 kg), the current state-of-the-art four-cylinder engine will be substituted in the future by three or even two cylinder engines, as shown in Figure 10.

Possible measures for a further displacement reduction are bore diameter and stroke. Decreasing the stroke alone is not preferred for downsized engine concepts, because it significantly determines the overall engine efficiency. Rather, an increased piston stroke is used, which results in a decreased surface to volume ratio. This in turn results in a more compact combustion chamber layout and thereby reduced heat losses to the wall (Gand, 1986; Bick, 1990). Furthermore, the elongated piston stroke enables an improved and intensified in-cylinder charge motion during intake stroke (Pischinger *et al*., 2001). Regarding the thermal load of internal combustion engines, small bore diameters are advantageous as well. Thus, the difference in wall temperature between the coolant and the combustion chamber and thereby the mechanical stresses due to thermal load is decreased. A smaller bore diameter also leads to reduced flame zones and thereby reduced burn delay and burn duration (Sterlepper, 1992). The losses due to incomplete combustion can also be diminished (Gand, 1986; Bick, 1990; Pischinger *et al*., 2001). While decreasing diameter has the above-mentioned benefits, challenges such as a rising oil dilution do arise when significantly reduced bore diameters below ~70 mm are considered for future downsizing engine concepts featuring gasoline DI.

The arrangement and the dimensions of gas exchange valves for downsized engines remains challenging. Particularly, the coolant capabilities of the water core and the positioning of spark plug and injector have to be considered within the overall combustion chamber layout. When using a standard central mounted spark plug, for example, intake and exhaust valve diameters can be increased by approximately 8% with a lateral injector positioning (Dünschede and Pischinger, 2010; Korte *et al*., 2011). However, the injector layout and targeting has to be addressed carefully to avoid excessive wall wetting. The injector position and the injector target of DI gasoline engines can help improve engine start-up and catalyst heating capabilities by allowing significantly retarded spark timings. The increased exhaust gas heat flux allows a fast and reliable catalyst light-off to fulfill future emission legislation. Furthermore, during engine part and full load at low engine speeds and during high speed engine operation, the injector layout has to support charge homogenization and stable injection patterns with reduced cyclic fluctuations in a short time span.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

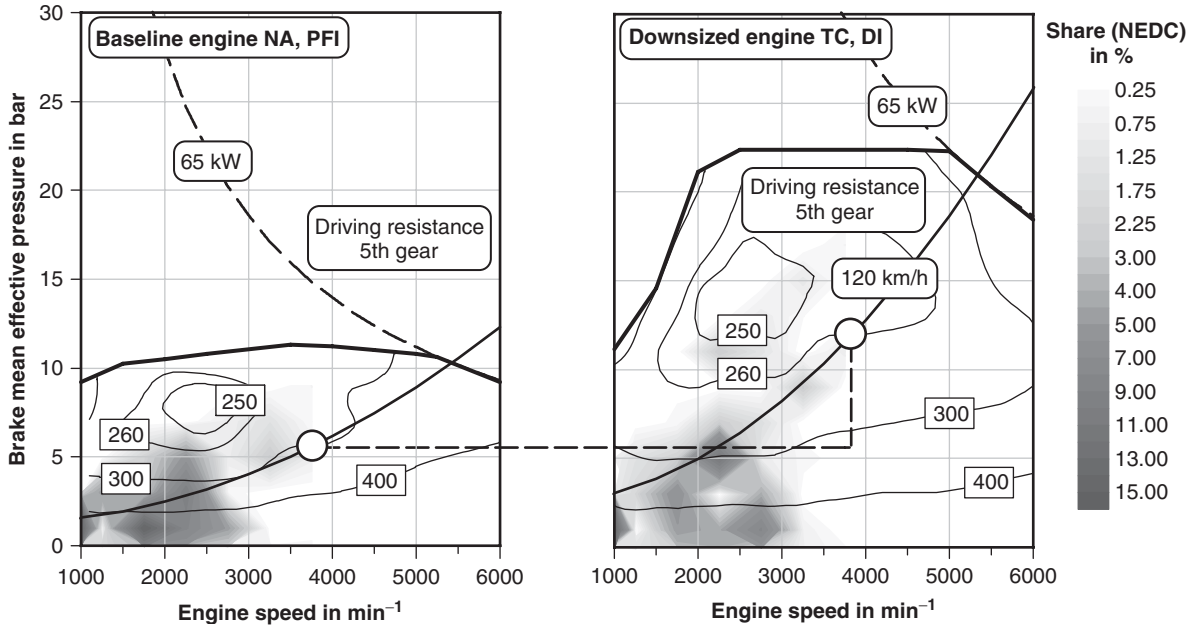


Figure 9. Load point shift for downsized engines.

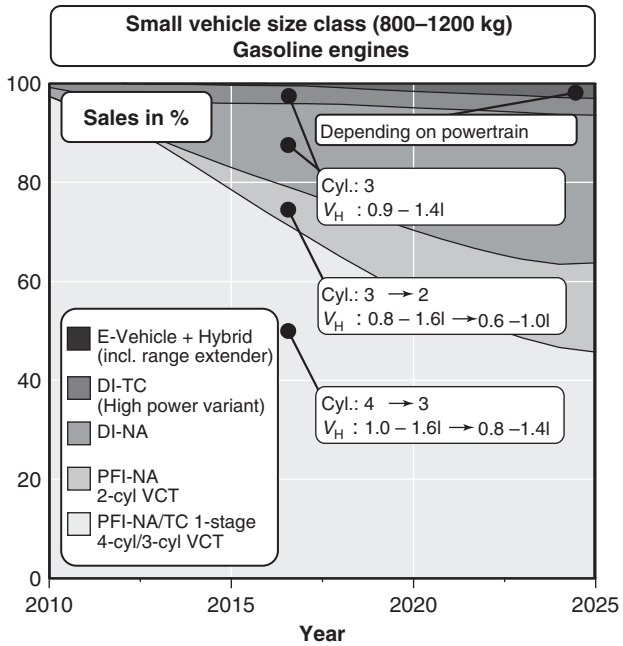


Figure 10. Roadmap engines for small vehicles, Europe.

Owing to the reduced engine displacement of the downsized engine compared to a natural aspirated engine with the same power output, the engines low-end torque is reduced. In addition, a boosted engine featuring an exhaust gas turbocharger has reduced transient response. Regarding a further increased market share for downsized engine concepts, improved low-end torque and enhanced transient response is mandatory to fulfill customer demands. The significantly increased boost pressure levels of downsized engines demand advanced boosting concepts to overcome these challenges. Thus, two-stage boosting of gasoline engines will become common for solving the current trade-off between acceptable low-end torque and demanded peak power output (Adomeit *et al*., 2010). Moreover, increased leakage and frictional losses of current turbochargers limits the available boosting technology for engine displacements below 1.0 L (Watson and Janota, 1982; Scharf *et al*., 2010). The combination of mechanical supercharging and exhaust gas turbocharging enables improved transient response of downsized engine concepts.

As a final note on downsized engines, they will be operated with increasing time spent at higher specific engine loads under real-world driving conditions. Accordingly, the risk of unwanted combustion events such as knocking or preignition rises. Knocking can be avoided by spark retard leading to higher exhaust temperatures. To stay within the specific temperature limit, fuel enrichment is necessary in terms of cooling. Alternative fuels using alcohol blending with increased heat of vaporization and knock resistance can help avoid fuel enrichment and maintain fuel efficiency.

3 BOOSTING

Turbocharger development focuses on enlarging the turbocharger operating range as well as increasing the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

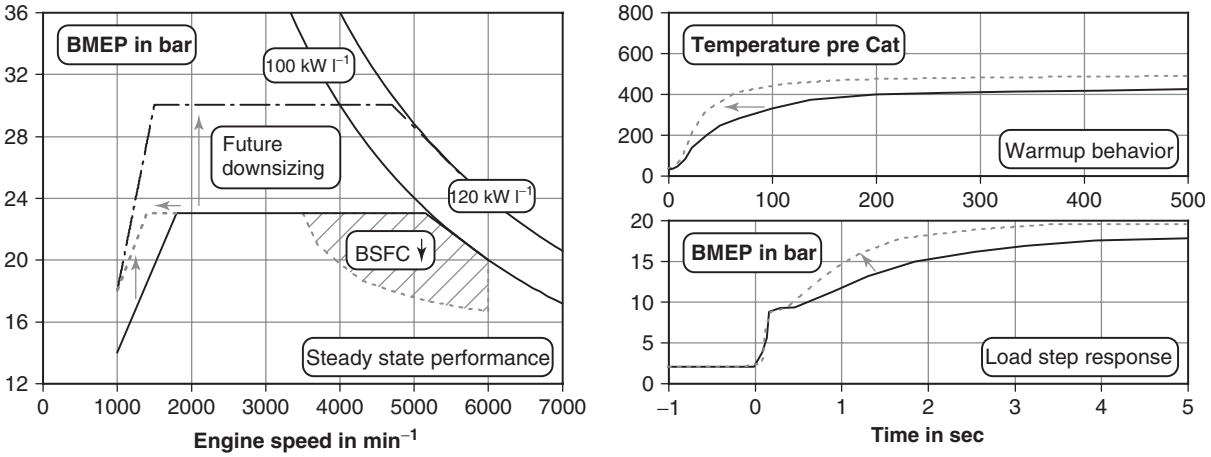


Figure 11. Boosting trends to improve steady-state and transient engine performance.

aerodynamic and mechanical turbocharger efficiency. All efforts have to be dedicated to benefiting the interaction of engine and turbocharger. Downsized engines have to comply with special requirements regarding performance and emissions, which are significantly affected by the boosting system. Figure 11 schematically illustrates the effect of boosting system optimization on engine performance in different operating domains. For a given specific power target, the aim is to improve stationary brake mean effective pressure (BMEP) at low speeds (low-end torque performance). If increased specific power is required, the level of BMEP has to be increased by enhancing boost pressure levels. At high speed and load, brake specific fuel consumption (BSFC) is to be optimized by extending the stoichiometric operating range and reducing the amount of enrichment. Improved warm-up behavior of exhaust manifold and turbine reduce the time to reach the catalyst light-off temperature, which significantly impacts cold-start emissions. The transient response of the engine-turbocharger system is determined by the ability to build up boost pressure after a load step and can be improved by using specific boosting technologies. Subsequently, the trends in boosting system technology to address the challenges associated with downsizing applications are presented. Details and fundamentals of turbocharging are discussed in Intake Boosting.

Single-stage compressors can already provide a sufficiently wide operating range for current engine concepts using a casing treatment such as ported shroud. However, single-stage operation is likely to exhibit limitations in terms of flow capacity and boost pressure demanded by future, enhanced downsizing. One possible solution for enlarging the compressor operating range is using two compressor wheels on one shaft driven by a single turbine. Here, the two wheels can either operate in parallel increasing the flow capacity or alternatively in a serial arrangement providing enhanced boost pressure. These arrangements can also contribute to improving the trade-off between operating range and transient response, because they usually feature reduced rotational inertia for a given flow and boost range.

Real multistage boosting systems offer even larger potential for increasing the specific power required for highly downsized engine concepts. A variety of different configurations exist, combining either two turbochargers or one turbocharger and one supercharger. Additional degrees of freedom arise from the thermodynamic coupling between the charging devices and the engine. A promising solution is based on exhaust port separation, which was made possible recently through application of variable exhaust valve control. The exhaust valves of each individual cylinder discharge into two separated manifolds and can thus act on different turbines. The two turbochargers can then be connected in a parallel-sequential arrangement or work in series-sequential mode. These systems yield a high degree of complexity regarding package design and control of the transition between single- and two-stage operation (Freisinger *et al*., 2010). Since the charging group represents a considerable heat sink, a key challenge is to minimize the thermal heat capacity of the charging group to ensure fast catalyst light-off and consequently comply with current and future emissions regulations (Schmuck-Soldan, Koenigstein, and Westin, 2011). Despite these drawbacks, two-stage charging systems allow a significant increase in SI engine low-end torque while enhancing the dynamic response at low engine speeds simultaneously. Furthermore, two-stage systems provide sufficient resources for the application of low pressure, high load EGR.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

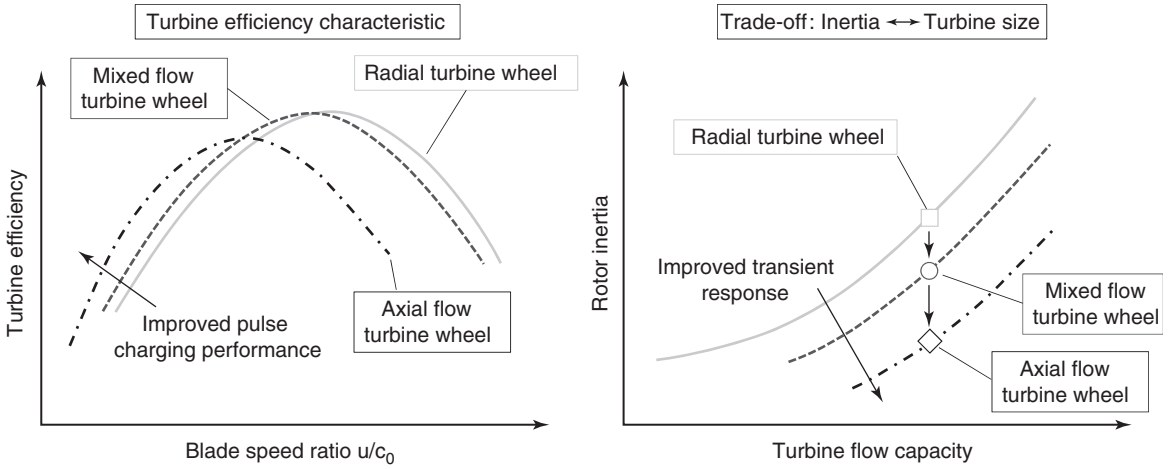


Figure 12. Comparison of radial, mixed flow, and axial turbine design characteristics.

The established design concept of automotive turbochargers consists of a radial turbine combined with a centrifugal compressor. On the turbine side, the radial design is widely accepted to be superior to the axial design for small turbine sizes employed in automotive applications, because axial turbines suffer more from rotor tip clearance losses. It has been suggested recently, however, that moving from radial to axial designs might be feasible. A less radical approach already commonly employed is the intermediate mixed flow turbine. The motivation for this consideration arises from the pulsating conditions present in the exhaust gas of an internal combustion engine. In order to extract as much energy from a pressure pulse as possible, a turbine has to be designed for high efficiency at high pressure ratios or low blade speed ratios (BSRs), respectively. Low BSR turbine efficiency benefits both transient and steady-state performance of the engine-turbocharger system. The only viable way to achieve this is shifting the efficiency peak to lower BSR. This, in turn, requires a nonzero inlet blade angle which can only be attained with a mixed flow or axial turbine design (Lotterman *et al*., 2011). It should be noted, however, that radial turbines offer higher peak efficiencies that are available over a wider BSR range. The swallowing capacity of the turbine increases when moving from a radial via a mixed flow until an axial design. That means a specified flow capacity can be achieved with a smaller dimensioned turbine and thus a reduced rotational inertia of the rotor (Rajoo and Martinez-Botas, 2008). The trade-off between the conflicting turbine performance requirements is schematically shown in Figure 12 for the different design options. Peak efficiency for the axial turbine is assumed to be marginally lower due to the stronger impact of tip clearance losses as mentioned before.

An alternative path in reducing the inertia without compromising efficiency is using alternative turbine wheel materials such as intermetallic titanium aluminide (γTiAl). Current state-of-the-art turbine wheel materials are Inconel for temperatures up to 950°C and MAR for temperatures up to 1050°C (Sonner *et al*., 2010). In comparison, γTiAl features significantly lower density, thereby reducing the rotational inertia of the rotor by approximately 30%. Increasing the temperature limit and ensuring long-term reliability at those temperatures contribute to the additional costs as well as the complicate manufacturing process (Lotterman *et al*., 2011; Sonner *et al*., 2010).

The degree of pulse charging is ever increasing in modern SI engines. In order to achieve optimum utilization of the kinetic energy and decoupling of the gas exchange of the individual cylinders, multi-scroll turbine housings have been developed. Scroll separation can be achieved by either a twin-scroll (twin-entry) or dual-volute (dual-entry) design. In current series application, the twin-scroll design is the preferred solution, because the dividing wall is more resistant and the wheel blades are less exposed to vibration excitation. Synergy effects may arise by combining scroll separation with variable geometry turbine (VGT) design (Sauerstein *et al*., 2009).

VGTs can be applied to SI engines in order to enhance the steady-state and dynamic torque build-up at low engine speeds. Additionally, efficiency benefits can emerge at wide open throttle compared to a conventional wastegate turbocharger as the complete exhaust gas enthalpy is utilized in the expansion process and not partially bypassed. Special emphasis has to be put on the control strategy of the VGT actuator during transient operation. Drawbacks

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

in the transient performance might emerge, if scavenging gas exchange is not being enabled in the initial phase of the transient (Sonner *et al*., 2010; Neußer *et al*., 2011). A key challenge in applying a VGT to SI engines is the development of an actuating mechanism able to withstand the considerably higher exhaust gas temperatures compared to CI engines. Smooth operation of the actuating mechanism has to be ensured without increasing the gap between housing and moving parts too much, which is required due to the different thermal expansion of the materials (Breitbach *et al*., 2007). This can only be achieved by using expensive nickel-based alloys. Even so, the temperature resistance of current series application is limited to 980°C. Promising design concepts considered for series production involve a variable nozzle ring (Neußer *et al*., 2011) or a variable slider mechanism (Breitbach *et al*., 2007).

Instantaneous load management is crucial in today's SI engine operation, which requires fast and precise actuation of the wastegate valve. State of the art is the actuation by indirectly controlled pneumatic systems. Recently, electrical actuators were developed offering advantages in terms of dynamics. High closing forces make it possible to keep the wastegate completely closed under pulsating conditions. This reduces the leakage through the wastegate and thus improves the transient response of the engine (Schwerdel *et al*., 2009). Additionally, the wastegate can be opened completely in non-boosted operation. This reduces exhaust back pressure and enables faster catalyst light-off due to bypassing of the turbocharger acting as a heat sink (Hagelstein *et al*., 2009). A higher market penetration of electrical actuators can be expected with decreasing costs.

Several design attempts toward highly integrated exhaust systems have been made recently. The primary motivation is the reduction in expensive nickel-based alloys as turbine housing material for withstanding exhaust temperatures of 1050°C. One approach is to integrate the exhaust manifold into the cylinder head. This allows the cylinder-out gas temperatures to be increased without increasing the effective turbine inlet temperature, which is usually limited by the temperature resistance of the material (Bäumel *et al*., 2011). This eliminates the need for fuel enrichment in large areas of the operating map and results in a considerable benefit in high load fuel consumption due to stoichiometric operation. Furthermore, this concept yields improved cold-start behavior: First, engine warm-up is accelerated due to the added heat source which reduces parasitic friction losses. Second, faster catalyst light-off can be achieved due to a reduction in surface area of the exhaust duct (Ernst *et al*., 2011). The same advantages apply to water-cooled aluminum turbine housings. Synergy effects regarding package and cooling demand can be used when combining a water-cooled turbine housing with an integrated exhaust manifold. All water-cooled exhaust system designs act as a significant heat sink. The dissipated enthalpy potential directly affects engine performance under closed wastegate operation, because any loss in enthalpy cannot be compensated here. This aspect has to be considered as well as when designing the exhaust aftertreatment system (Prevedel *et al*., 2011).

A different approach to conventionally casted exhaust system components is using a sheet metal design for exhaust manifold as well as turbine housing. The insulating air gap between inner and outer sheet metal layer leads to lower heat rejection into the engine bay. Furthermore, this concept offers significantly lower thermal inertia, which improves the engine's transient response and catalyst light-off time without adding substantial production complexity (Björnsson, Johansson, and Kunde, 2010). It does not solve, however, the problem of fuel enrichment for component protection.

4 VARIABLE VALVE TRAINS

Within the last 30 years, variable valve lift systems have spread into almost every SI engine segment with increasing importance as a measure for reducing fuel consumption and optimizing performance, as well as efficiency. Those systems enable the variation of valve opening timing and valve lift depending on the operating point. The achievable improvement depends on the systems degree of variability. The following section provides an overview for variable valve lift systems seen in current production or development. They are divided into cam-actuated and direct-actuated systems.

Cam-actuated systems vary in timing, duration, and lift of the valve opening by adjusting the operating mechanism. Valve timing variation is realized by phasing the camshaft relative to the crankshaft. The first camshaft actuator was introduced in 1983 by Alfa Romeo. Besides two-position cam phasing, continuous cam phasing can be identified as state of the art (Schmidt *et al*., 1998). Continuous cam phasing can be realized with helical toothing moved by hydraulics or inverted use of the vane pump principle (Figure 13).

The first E-motor operated intake camshaft adjuster called the VVT-iE was implemented in the 4.6-L V8 32-V double overhead camshaft (DOHC) engine for Lexus 460. Today's production cam phasing systems enable phasing angles of until 80° crank angle (40° camshaft angle). Such a system applied to engines with DOHCs enables controlling size and position of the valve overlap.

There is also a possibility of using transmission elements (compensation elements) between the cam profile and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

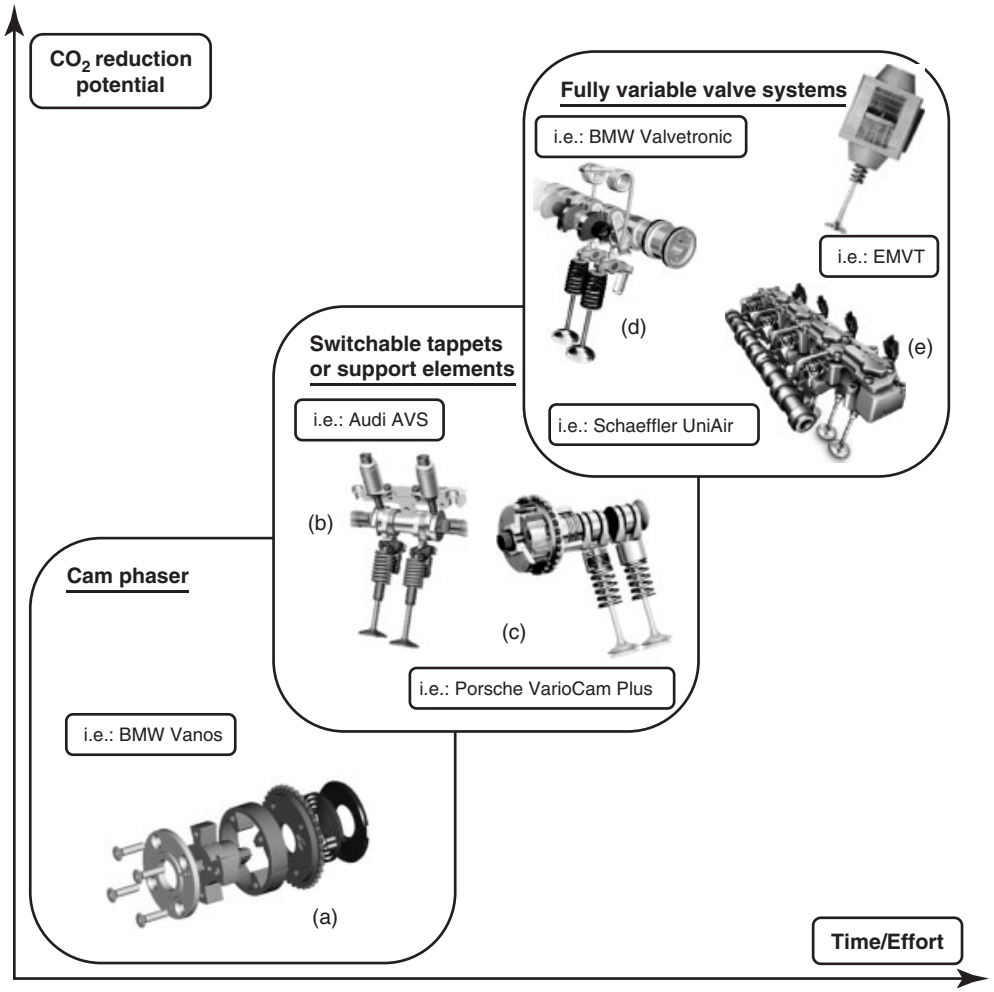


Figure 13. CO_2 reduction potential of variable valve trains; figures: (a): (From Landerl, Klauer, and Klüting, 2004. Reproduced by permission of BMW AG.); (b): (From Wurms *et al*., 2006. Reproduced by permission of Audi AG.); (c): Dr. Ing. h.c. F. Porsche AG., . (Reproduced by permission of Porsche AG.); (d): (From Borgmann *et al*., 2004. Reproduced by permission of BMW AG.); (e): Schaeffler Technologies AG & Co. KG., (Reproduced by permission of Schaeffler Technologies AG & Co. KG.)

valve tappet. Switchable tappets or support elements allow the disconnection of individual valves. The alternating operation with two different cam profiles has also been realized in which the camshaft for each valve has two distinct cam lobes that actuate concentric surface areas of a switchable tappet. Therefore, a part and full load optimized valve lift can be realized, which mitigates the otherwise necessary compromise in valve timing and valve lift. An example for this technology is the VarioCam Plus (Eichler and Heiselbetz, 2001), which uses control bucket tappets combined with a camshaft phaser and is installed in various high performance engines produced by Porsche AG. Another example produced by Mitsubishi in series is the MIVEC engine, in which the switching mechanism is integrated in the rocker arm. This way, valve lifts for low and high engine speeds and a complete deactivation of individual valves for cylinder shut-off is realized. Honda produces a similar solution, which is used for various engines in series under the name of the VTEC system.

The Audi Valvelift System (AVS) first used in the 2.8-L V6 FSI engine switches between two valve lift curves. Sliding cam units with two different cam profiles are attached to the base camshaft on evolvent splines. The switching of the respective cam profiles in the base circle phase is made possible by the hydraulic actuated pins in combination with two switching grooves. The transfer to the valve via the roller cam followers and cam shaft bearings are basically adopted by the conventional valve train (Wurms *et al*., 2006). The AVS was also installed on the exhaust camshaft of the TFSI 2.0-L system to ensure

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the firing order separation, which prevents the exhaust gas backflow into the suction at maximum valve overlaps.

In addition to the systems switching between two valve lift curves, there are other technologies that enable variable valve lift. UniAir is a hydraulic system that is used in several mass production vehicles by Alfa Romeo and Fiat. It is based on the principle of lost motion in which the cam lift is transmitted hydraulically to the valve so that the closing of the valve is triggered prematurely by redirecting the hydraulic pressure. As a result, the lift and closing time of the valve are decoupled from the cam profile. Hydraulic systems are used to dampen the motion of the valve as it approaches the valve seat.

To also change the valve lift continuously, mechanical valve controls were developed. Most of those concepts change the transmission ratio of the rocker arm via an eccentric shaft. In 2001, a fully variable valve control system was introduced to the market by BMW. This system enables an engine operation without throttle by continuously adjusting intake valve timing and lift. A similar system called Univalve was developed by enTec and is currently marketed by Kolbenschmidt Pierburg AG.

Since 2008, Toyota produces their variable valve timing system called the Valvematic. The system uses a special valve control lever (follower) for the continuous adjustment of the timing and the valve lift, which is arranged between the camshaft and conventional valve lever. An electronic actuator controls the valve lift and the valve opening duration. Additionally, a hydraulic variable cam phaser is used on both intake and exhaust. Latest investigations aim at an increased variability with lower complexity such as by the combination of a sliding cam unit with a full variable cam profile (Nitz *et al*., 2010).

Electromechanical valve actuators use electromagnets in combination with valve springs forming a spring/mass oscillator. The valve motion is mainly forced by spring forces and the electromagnets are used to initiate and support the opening and closing process. As a result, the necessary energy is reduced significantly compared to other electromagnetic valve controls. However, the energy demand using electromechanical valve control still has to be taken into account in comparison with conventional mechanical valve operation. Suggesting a generator efficiency of 80%, electromechanical systems have a part load energy demand comparable to current low friction valve trains (Salber, 1998). Again, the large design effort and difficulties in dynamic in-vehicle operation currently limits the use of electromechanical valve actuators. For future valve lift systems, it is most important to gain a high variability at low cost and design effort. Therefore, less complex mechanical or hydraulic solutions need to be developed.

5 POTENTIAL OF VARIABLE VALVE TRAINS

Variable valve trains enable improved fuel economy by the optimization of the gas exchange process. The cam layout and therefore the valve timing of standard, fixed valve trains reflect a compromise of acceptable engine stability at idle conditions for reduced engine operating speeds, improved part load fuel efficiency, and low emissions and enhanced volumetric efficiency at full load engine operation.

The valve timing for "intake valve closing" (IVC) has to maintain high levels of engine torque at low engine speeds and high peak power output at rated engine speed. During full load operation, maximum volumetric efficiency has to be maintained. Exhaust valve opening (EVO) has to consider the expansion work of the exhaust gases, whereas exhaust losses at high engine speeds and high engine loads have to be avoided. The trade-off between these contrasting requirements can be solved partially by the application of variable valve systems. By the application of variable valve timings, the opening and closing of valves can be adapted to the engine speed and load accordingly.

Further adaptation to a tailored gas exchange can be enabled by the application of a variable valve lift system. Variable valve lift systems can be classified by the valve timing, the valve opening period, and the valve lift. These functionalities can be combined to a fully variable valve lift (Unger, 2004; Haas and Rauch, 2010). The application of a fully variable valve train enables reduced gas exchange losses due to de-throttling of the intake air system.

The valve overlap is determined by the "intake valve opening" (IVO) and "exhaust valve closure" (EVC) and controls the amount of exhaust gas residuals. The recirculation of exhaust gas enables a partly de-throttled engine operation at part load engine operation. Thus, fuel efficiency can be improved and the level of raw emission for a stoichiometric-operated gasoline engine can be reduced. On the other hand, the combustion stability is deteriorated by the amount of recirculated exhaust gas, thus burn delay and burn duration increases. Misfire can occur and the HC emission level increases significantly (Tuttle, 1980, 1982; Hara, Nakajima and Nagumo, 1985; Wichart, 1987). The de-throttled engine operating range achieved by using exhaust gas recirculation can be enlarged by combustion systems featuring increased in-cylinder charge motion, which can enhance combustion stability and the amount of recirculated exhaust gas that can be used. (Wurms, 1994; Löbbert, 2006).

At increased engine speed and engine load, the fuel efficiency potential of de-throttling is reduced as the amount of throttling required for load control declines. However,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

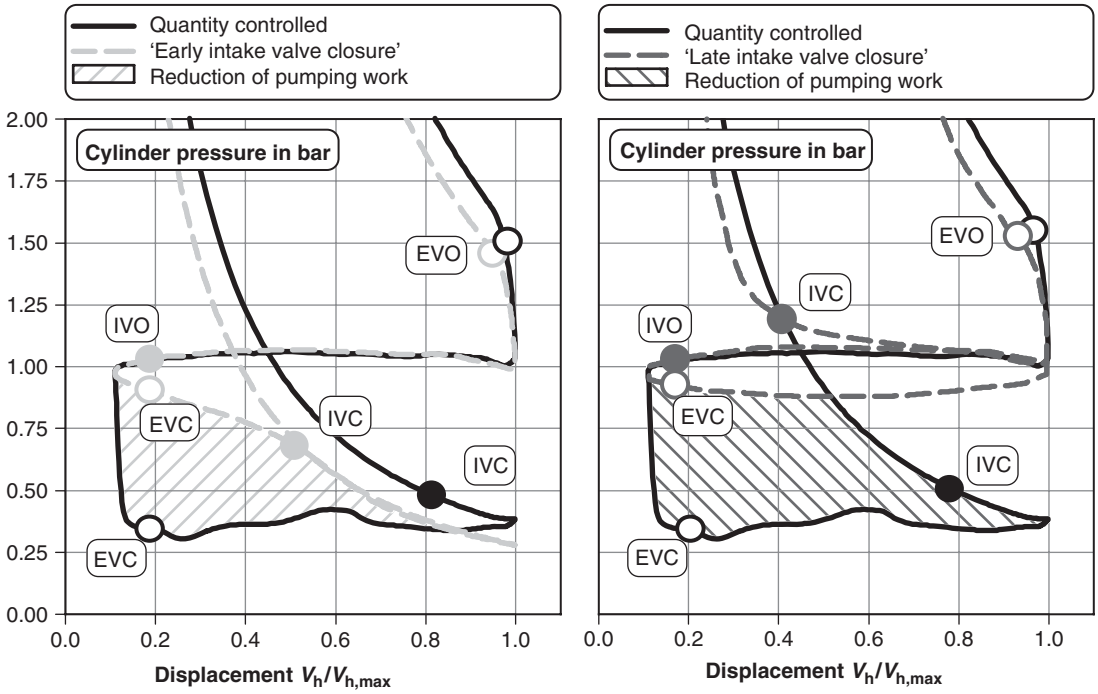


Figure 14. Load control by means of EIVC and LIVC for $n = 2000/\text{min}^{-1}$; IMEP = 2.8 bar.

further limitations arise related to increasing charge temperatures and knock sensitivity. Variable valve timing in the form of late intake valve closure (LIVC) can be used to reduce the effective CR. Owing to the upward moving piston stroke, a part of the cylinder charge is pushed back to the intake system. Thus, the final compression pressure and temperature decline. This reduces knock sensitivity and spark timing can be advanced, enabling improved engine efficiency. With early intake valve closure (EIVC), the gas exchange ends during the intake stroke before bottom dead center (BDC), as soon as the desired charge air mass is introduced. The cylinder charge is expanded after IVC and will be compressed afterward. The load control by EIVC or LIVC compared to conventional intake charge quantity control of internal combustion engines is shown in Figure 14.

When combustion and gross indicated efficiencies are constant, EIVC and LIVC can reduce gas exchange losses by approximately 35% for the engine operation point shown ($n = 2000/\text{min}^{-1}$, indicated mean effective pressure (IMEP) = 2.8 bar). During NEDC, a fuel consumption potential of approximately 8% was shown by the application of a fully variable valve train (Flierl *et al*., 2011).

The application of a switchable valve lift system enables adequate potential for future engine concepts, with reduced complexity compared to a fully variable valve lift system (Wurms *et al*., 2006; Luttermann, Schünemann, and Klauer, 2008). Different cam profiles for maximum torque, rated speed, and zero valve lift enable diversified potential. A reduced cam event length allows improved scavenging capabilities for current four-cylinder engines by the separation of the engine firing order (Budack *et al*., 2009). Furthermore, reduced valve lift combined with reduced event length can be used for de-throttled engine operation, using "EIVC" in a predefined engine operation area. The reduced valve lift can be combined with single or dual intake-port shrouding to intensify intake charge motion. Thus, combustion stability during low load engine operation can be improved and fuel efficiency in this engine operation area is enhanced (Gottschalk and Tschöke, 2000; Dilthey, 2004).

The cam profile and valve timing at rated speed enables lowered gas exchange losses at maximum cylinder charge with reduced residual gas content after gas exchange. Thus, charge temperature is reduced, knock sensitivity is lowered, and the spark timing can be advanced, thereby engine efficiency can be improved. The zero valve lift also enables cylinder deactivation which can be used to increase engine efficiency during low load engine operation (Indra, 2011; Middendorf *et al*., 2012).

Additional potential can be gained by different phasing of the intake valves. Different cam profiles on the intake side will lead to enhanced swirl motion. Thus, charge homogenization and residual gas tolerance can be improved, and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

emissions and fuel consumption can be lowered (Wyatt *et al*., 2011).

The application of different valve timings/events enables the usage of the exhaust gas pulse for modulated boost concepts. The immediate exhaust gas pulse at EVO is used for pulse turbocharging, whereas the rest of the exhaust stroke is bypassed and not used for turbocharging. Thus, the gas exchange is improved due to the reduced gas exchange losses by lowered exhaust gas pressure (Wyatt *et al*., 2011; Ross and Zellbeck, 2010).

Further CO_2 and pollution emission legislation will be one of the main drivers for future engines. Valve train variability will play a dominant role in tailoring gas exchange of internal combustion engines to help meet these future demands.

6 SI FOR HYBRIDS

The development targets of the combustion engine in a conventional powertrain differ quite significantly from the requirements for hybrid powertrains. One main development focus for the conventional combustion engine is the improvement of the part load fuel consumption. This target can be achieved by different engine technologies, for example, downsizing, downspeeding, variable valve lift, and cylinder deactivation, as already discussed. Hybrid powertrains on the other hand offer the possibility to limit the operation range and avoid part load engine operation. This can be achieved by shifting the engine load point to higher engine loads or by substituting engine operation points with high fuel consumption at very low loads by electric driving (Figure 15). Thus, technologies that aim to achieve improved fuel consumption at low engine loads have a lower impact in hybrid powertrains.

Several investigations have been conducted to investigate the optimal layout of the combustion engines in hybrid powertrains. (Seibel and Pischinger, 2007; Balazs and Pischinger, 2011). These investigations are based on combinations of design of experiments, vehicle simulations, and numerical optimization methods.

Figure 16 displays the fuel consumption results for a parallel hybrid powertrain in several vehicle classes (Seibel and Pischinger, 2007). Additionally, the fuel consumption of a standard engine (naturally aspirated engine with variable valve timing) in conventional powertrains is shown.

Comparing the standard engine with VVT in a conventional and a hybrid powertrain, a fuel consumption benefit of 20–23% is achieved. With the turbocharged DISI engine, a fuel consumption benefit of 21–28% can be achieved compared to the standard engine with VVT in conventional vehicles. The fuel consumption benefit with hybridization

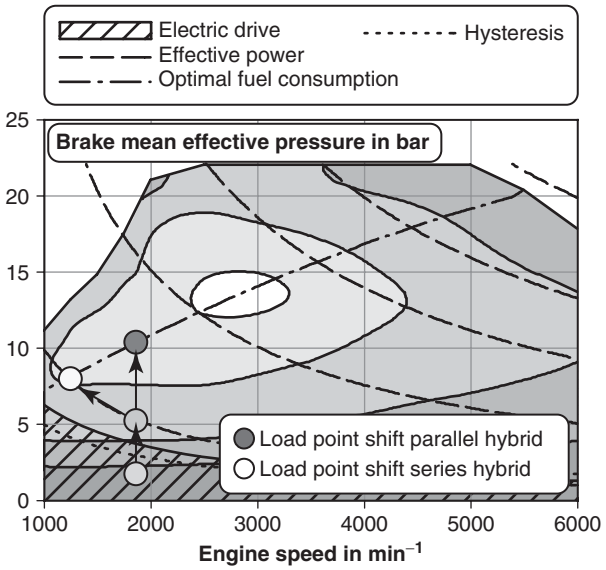


Figure 15. Engine operation range in hybrid powertrain.

increases with increasing vehicle and engine size. The same trend can be seen when comparing the fuel consumption differences of the different engine concepts within one vehicle class. For the compact class vehicle, the differences between the engine concepts are very small. However with increasing power to weight ratio in bigger vehicle classes, the fuel consumption benefits of advanced engine technologies, such as engine downsizing, increase significantly.

Typically, in larger vehicle classes, the ratio of engine power to vehicle weight increases significantly. Therefore, the engine is operated at lower loads. Shifting the engine load point to higher loads (by engine downsizing or hybridization) or substituting the engine operation at low loads by electric driving results in a bigger fuel consumption benefit.

In Figure 17, the CO_2 emissions of a hybrid vehicle with naturally aspirated engine, operating based on the Atkinson combustion process, and with a turbocharged DI engine are shown. The results were obtained from vehicle simulations performed by Balazs and Pischinger, 2011. The layout and scaling of the components in the parallel hybrid powertrain and the operation strategy have been optimized for five different driving cycles. This was done with equal vehicle performance targets, such as maximum vehicle speed, acceleration performance, elasticity, and gradeability.

The Atkinson engine represents a simple engine featuring an intake camshaft with long intake event (late IVC) and optimized engine friction. This results in low fuel consumption values but a low specific full load torque. In contrast to this, the turbocharged DI engine achieves 22-bar BMEP and 100 kW/L. The possibility of supporting the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

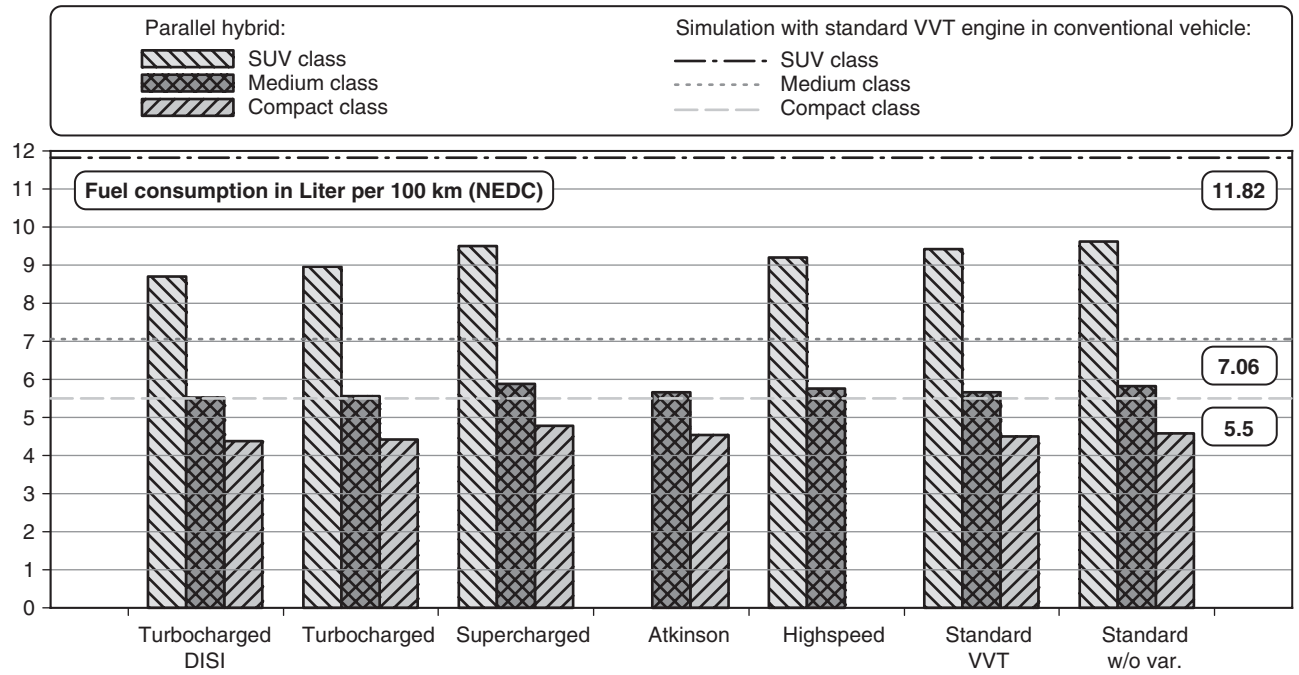


Figure 16. Fuel consumption of different engine concepts in a parallel hybrid vehicle. (Reproduced from Seibel and Pischinger, 2007. © FVV.)

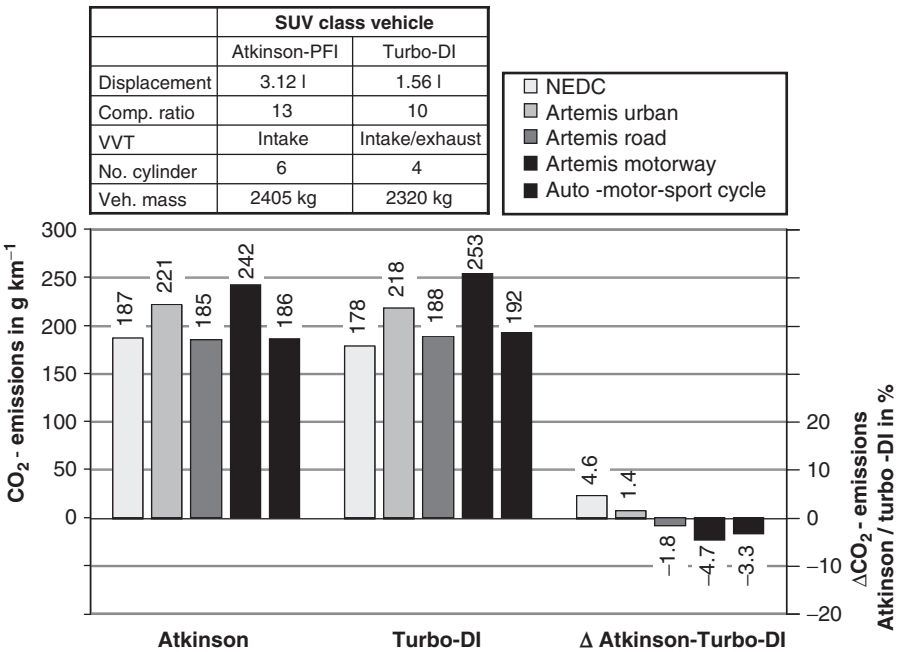


Figure 17. CO_2 emissions of a parallel hybrid vehicle with automatic dual clutch transmission in different driving cycles. (Reproduced from Balazs and Pischinger, 2011. © FVV.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

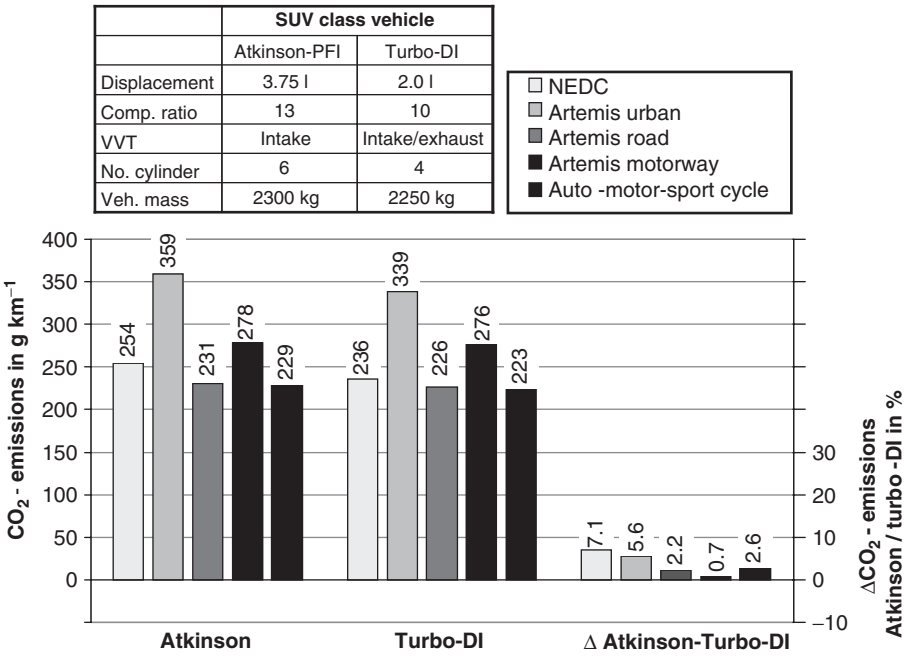


Figure 18. CO_2 emissions of a conventional vehicle with automatic dual clutch transmission in different driving cycles.

engine at low engine speeds by boosting with the electric machines enables a layout for high specific power rather than focusing on achieving a high low-end torque. The engine displacements have been adapted in order to achieve the same maximum power.

In Figure 17, it can be seen that the downsized engine has benefits when the engine is operated at low to medium engine loads (NEDC, Artemis Urban), whereas the Atkinson engine provides benefits at higher loads. Fuel consumption drawbacks at low engine speeds can be compensated by hybrid-specific features (electric driving, load point shift). In conventional powertrains, the benefit of the downsized engine increases (Figure 18).

In plug-in hybrid vehicles and electric vehicles, the electric motor is the prime mover. With increasing battery capacity and electric machine power, the electric driving range increases. This results in a lower time share of combustion engine operation. Figure 19 shows a typical strategy for a plug-in series hybrid vehicle. At the beginning of the driving cycle, the battery is fully charged. The vehicle is powered by the battery and the electric machine (charge depletion mode). After approximately 1500 s, the combustion engine begins operating when a certain vehicle speed is exceeded in order to power the electric traction motor and the battery. With this strategy, the state of charge (SOC) is maintained at 20% (charge sustaining mode). With increasing electric drive range, the importance of criteria for achieving low fuel consumption decreases, whereas the importance of criteria such as noise, vibrations, and harshness (NVH), packaging, and production cost increases.

In summary, the main development focus for combustion engines in electrified powertrains can be assumed as follows.

In the cost-sensitive small vehicle class, many vehicles will be equipped with simple combustion engines (e.g., naturally aspirated, port fuel injection) including limited technologies for engine improvement (e.g., friction reduction, Atkinson process)

In large vehicle classes, the diversity of different engine concepts will be higher for hybrid vehicles. Benefits by engine downsizing increase. In addition, combustion engines with cost-intensive technologies will be available in hybrid vehicles.

With increasing degree of electrification, the combustion engine will be simplified. In particular, for urban utilization, the share of electric drive increases and the combustion engine remains as a so-called "range extender" extending the electric drive range for long distances and exceptional cases. Criteria such as NVH and packaging gain importance while the impact of fuel consumption on the total energy consumption of the vehicle decreases.

6.1 Motivation and function of range extenders

Battery electric vehicles (BEVs) are facing drawbacks including short all electric ranges and the lack of heating

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

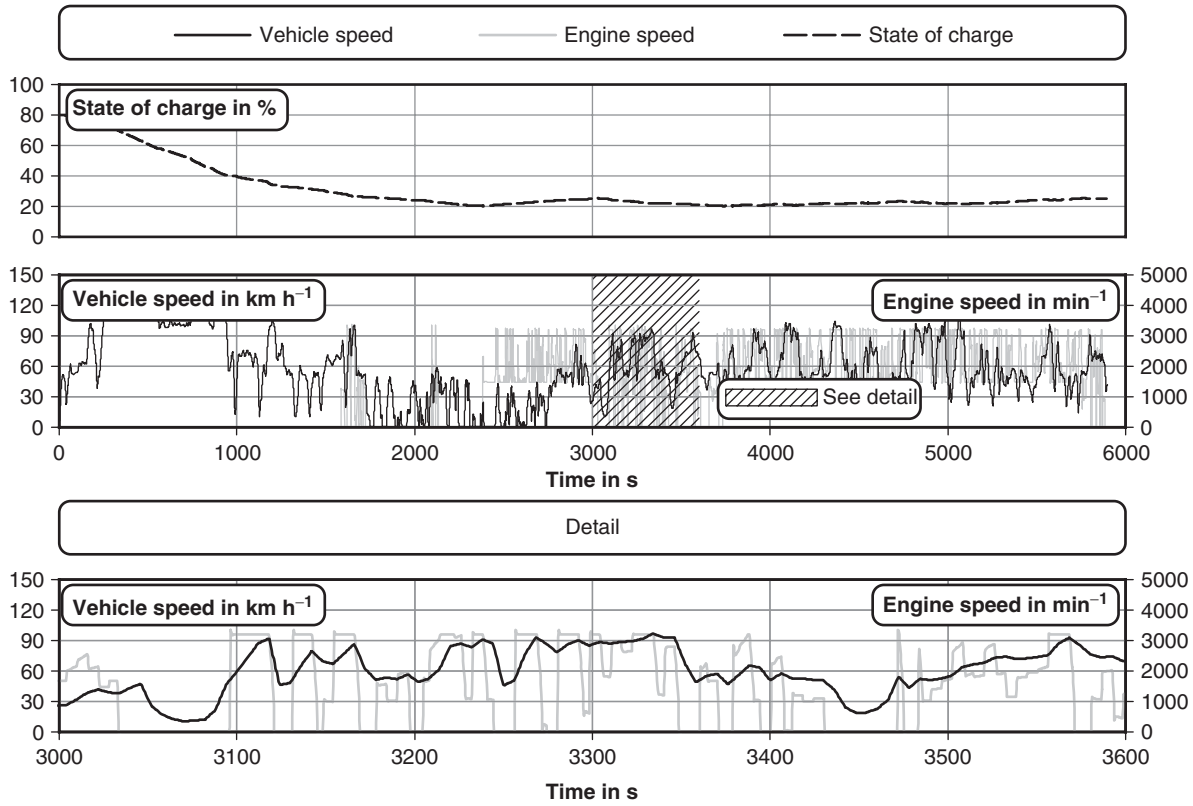


Figure 19. Operation strategy of plug-in series hybrid. (Reproduced by permission of FEV GmbH.)

energy from the combustion engine. Today's batteries (Li-ion) still have low energy densities in combination with high costs; so, they are still not suitable for longer travelling distances. In order to overcome this drawback, range extenders can be used to produce electric energy when needed for longer distances. Hence, the batteries can be dimensioned even smaller, providing energy just for daily driving distances.

Furthermore, the lack of heating energy can be overcome if the range extender cooling circuit is linked to the vehicle's heat exchanger. It can be further improved if the exhaust gas enthalpy is recovered by an intercooler. At low ambient temperatures, instead of using electric energy from the battery to heat up the cabin, the range extender can be operated at an efficient operating point to provide the needed heating energy and simultaneously recharge the battery for an extended electric range.

The actual range extender integration strongly depends on the given vehicle category, which mainly defines requirements on development targets such as usability, driving performance, or convenience aspects (Fraidl, 2010). Since range extender modules (REMs) only operate when needed, start-up and response time or life expectancy become less critical specifications, whereas required space or NVH gain in importance. These new aligned requirements provide the opportunity to consider such potential technologies as external combustion or low-vibration space saving rotary piston. In Figure 20, important REM characteristics are compared for exemplary technologies.

As shown in Figure 21a, the range extender is driven as a serial hybrid. As a consequence, the range extender package has maximum freedom degrees. However, multiple energy conversions result in significant efficiency losses. The range extender in Figure 21b can be driven as serial or combined hybrid. By the use of a mechanical interface to the electric powertrain, its installation location is already fixed. The lower number of energy conversions results in higher efficiencies. In Figure 21c, the electric traction motor and the range extender are driving different axels. If the range extender should be usable also at low vehicle speeds (in case of low battery state of charge), a multistage gearbox is mandatory.

One of the major development targets for range extenders is a good NVH behavior. The excitations from the range extender onto the vehicle body can be distinguished between airborne sound and vibrations at the engine mount positions. Airborne sound is radiated by the engine surfaces

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

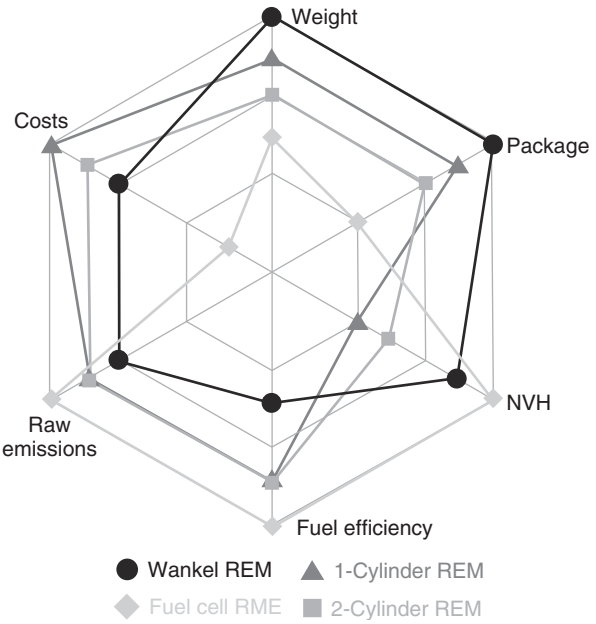


Figure 20. Comparison of the characteristics of different range extender modules.

and the intake and exhaust orifices (Genender, Speckens, and Schürmann, 2011).

Vibrations at the engine mount positions mainly result from free oscillating mass forces, which strongly depend on engine architecture. Without considering any mass balancing systems, to some extent two-cylinder engines, the inline three-cylinder engine, as well as the Wankel engine might be suitable concepts (Fraidl, 2010; Pischinger *et al*., 2011a; Warth *et al*., 2011).

The discontinuous combustion in multicylinder engines induces an oscillating torque with high amplitudes, especially at low engine speeds, and is most noticeable with a small number of cylinders connected to the crankshaft. The result is a rolling torque. Especially during start and stop, highly noticeable vibrations occur at the engine mounts. Intelligent layout of engine mounts or a special range extender layout for compensation of the rolling torque should be taken into account (Pischinger *et al*., 2011a).

Airborne sound of the intake and exhaust orifices is affected by the firing interval. High orifice sound levels are especially produced in engines with only a few cylinders (single cylinder, two-cylinder) and uneven firing intervals. The exhaust side with its until 20 dB higher pressure pulsations is more critical than the intake orifice (Genender, Speckens, and Schürmann, 2011). Therefore, high volumes and small runner diameters are needed to minimize the orifice sound. For a two-cylinder engine, intake and exhaust muffler volumes should be around 15 L.

For optimizing NVH, the operation strategy of the range extender offers another degree of freedom. Wind and road noise are the major interior noise components in an electric vehicle at higher driving speeds. However, during standstill and at low vehicle speeds, there is no significant masking noise such as wind and road noise. As a result, range extender operation is inappropriate under these conditions, unless absolutely necessary. In general, depending on the interior noise level of the electric vehicle, the range extender should not be used below vehicle speeds of 30 –50 km/h (Figure 22).

At higher vehicle speeds, the range extender speed should be linked to vehicle speed. Its load should vary with the driver's torque demand and its energy produced should directly be used for propulsion without buffering within the battery. At highway speed, the range extender can be driven steady at its rated power output point, because the driving noise is sufficient to mask the range extender interior noise.

Another key to successful market introduction are costs as the range extender technology is meant to enable reasonable priced electric vehicles. Naturally aspirated gasoline engines with port fuel injection deliver sufficient specific power output and allow exhaust gas aftertreatment systems to be kept simple and therefore cost-effective. Compared to DI combined with charging technologies, NVH behavior is also better. As efficiency and highly transient engine operation are minor targets, state-of-the-art technologies offering variable operation (e.g., cam phasing or variable

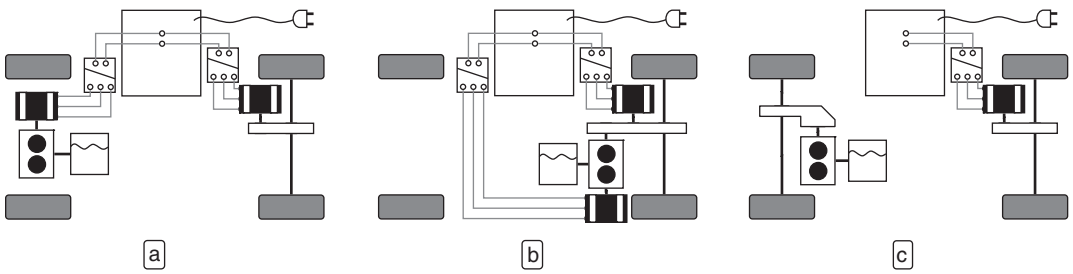


Figure 21. (a–c) Range extender topologies. (Reproduced by permission of FEV GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

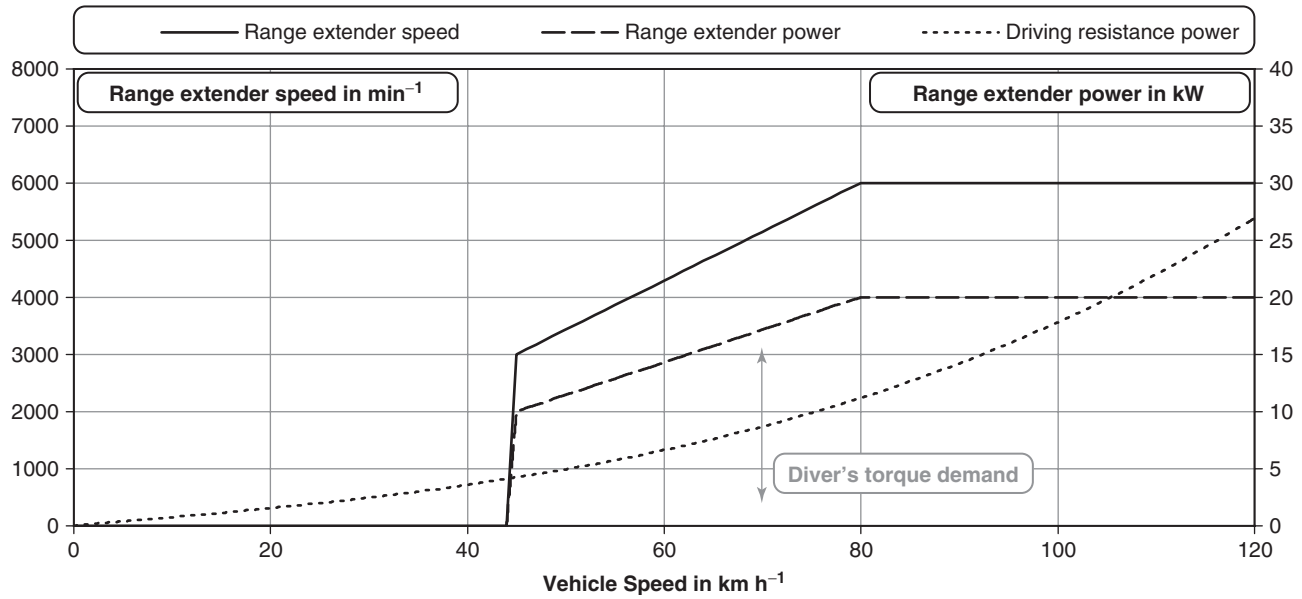


Figure 22. NVH optimized range extender operation strategy.

intake manifold) can be dropped. Moreover, a specific power output of 35–40 kW/L might already be sufficient, and a pushrod-driven OHV valve train with two valves per cylinder as well as rather low nominal engine speeds (≤4000 rpm) can be realized. Additionally, a high amount of carryover parts and geometries from existing engine families (bore diameter, bore pitch, engine architecture) enable integration into existing manufacturing processes.

7 ADVANCED COMBUSTION CONCEPTS

Compression ignition of gasoline under lean or dilute conditions has proven good potential to significantly cut down CO_2 emissions by up to 30% and reduce NO_x emissions by up to 99% in contrast to conventional throttled stoichiometric SI engine operation at part load. Gasoline compression ignition (GCI) has been implemented in various ways and referred to by various names in the literature (e.g., controlled autoignition (CAI), homogenous charge compression ignition (HCCI), etc.). The reduction in fuel consumption from GCI is mainly driven by the de-throttling of the intake system, the optimized charge properties due to the lean mixture, and the low temperature combustion and the comparatively quick combustion process. The distinct reduction in NO_x emissions due to the low temperature combustion (see NO_x Formation and Models and Advanced Compression-Ignition Combustion for Ultra-Low NO_x and Soot) provides the fuel consumption benefits of stratified lean, spark-ignition combustion without the drawback of additional NO_x aftertreatment (see Emission Control Systems—Oxides of nitrogen, Advanced Compression-Ignition Combustion for Ultra-Low NO_x and Soot, and Gas Aftertreatment Systems). Moreover, recent investigations are showing pathways to boosted, high load, low emission operation of advanced GCI strategies as discussed in Advanced Compression-Ignition Combustion for Ultra-Low NO_x and Soot.

High amounts of trapped residuals from the previous combustion cycle help initiate the autoignition. The two most promising internal EGR strategies are exhaust port recirculation (EPR) and combustion chamber recirculation (CCR). EPR is characterized by a backflow of residuals from the exhaust port by opening the exhaust valve during the suction stroke, whereas CCR is characterized by an early exhaust closing and a short exhaust and intake event length (large negative valve overlap).

The autoignition process of hydrocarbons is controlled by chemical reaction kinetics and, according to Westbrook (2000), triggered by prompt decomposition of hydrogen peroxide (H_2O_2) into hydroxyl radicals (OH•) during the compression stroke at approximately 1000 K (see Fundamental Chemical Kinetics). Measures that can be taken to generate this temperature lead to earlier autoignition and vice versa. In general, GCI combustion can be influenced by:

- the global pressure and temperature in the combustion chamber (e.g., determined by the trapped amount of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

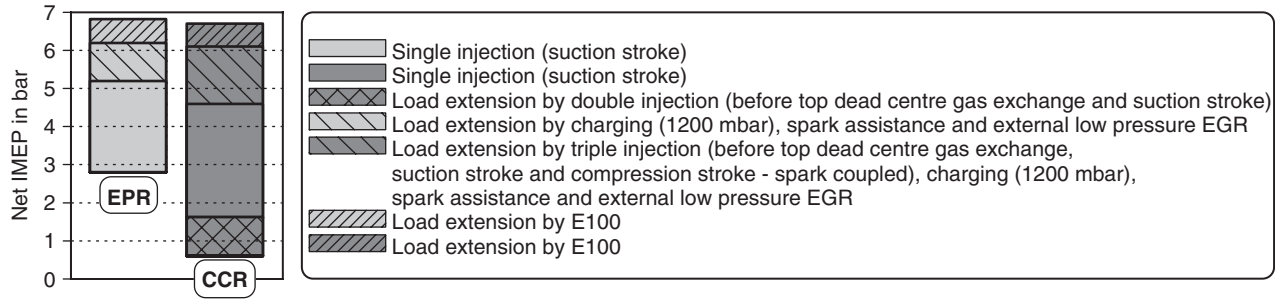


Figure 23. Strategies for load extension measured at a single cylinder engine with compression ratio of CR=12 at an engine speed of 2000/min^{-1}. Boundary conditions for GCI: a standard deviation of net IMEP of >0.15 bar, a maximal cylinder pressure gradient of <5 bar per °CA, or a relative air/fuel ratio <1.

residuals, the CR, or the thermodynamic conditions in the intake system),

- the local stratification of residuals, temperature as well as air/fuel ratio (e.g., depending on the calibrated injection strategy or the exhaust gas strategy used), and
- the fuel properties (e.g., primarily the autoignition characteristics, the enthalpy of vaporization, and the volatility).

Nevertheless, GCI currently reveals disadvantages such as a limited operation range. Toward low engine loads, the decreasing exhaust enthalpy leads to a deteriorated combustion and finally to misfiring. Toward higher engine loads, the maximum pressure gradient is significantly increasing.

Figure 23 illustrates possible strategies to enlarge the operational range starting from a single injection strategy during the suction stroke for both EPR and CCR. The early closing of the exhaust valves for CCR enables the possibility to inject a distinct fuel amount shortly before TDC gas exchange into a regime of compressed hot residuals that causes a major impact on combustion phasing. It leads—at sufficiently lean conditions and constant valve timings—to an advanced combustion, due to the partial conversion and particularly reformation of the fuel. This strategy can be used in combination with a spark-assisted operation to expand the load range significantly.

As the maximum load range is limited by a steeply increasing maximal pressure gradient, it is necessary to adopt strategies that slow down the reaction progress. This can be achieved by a further dilution through boosting or the usage of external EGR. In addition, the previously described injection before TDC gas exchange can be used to further lean out the mixture, as less residuals are necessary to achieve a certain combustion phasing.

To further extend the operational range toward low as well as high loads, it is possible to adopt a late third injection that is coupled to a supporting spark and leads to a stratified combustion in the vicinity of the spark plug. Moreover, the usage of alternative fuels such as ethanol with high RON/MON enables a further broadening of the maximum load range as the maximum pressure rise rate can be reduced by combustion duration extension.

As the NO_x emissions rise with increasing engine loads, it might be necessary to switch from lean to stoichiometric conditions at higher engine loads, in order to allow the usage of the three-way catalyst technology. Fully de-throttled operation can be maintained by using high amounts of external EGR.

8 VARIABLE COMPRESSION RATIO

The application of downsizing to reduce fuel consumption in context with boosting presents a disadvantage because of the higher knock sensitivity at higher engine loads. At a given boost level and fuel quality, this can only be compensated for by either spark retardation or adjustment of the CR. Despite the fact that the majority of the boosted engines in contrast to the naturally aspirated engines are equipped with direct fuel injection, today's boosted engines are designed for RON 95 gasoline with a CR of approximately 9.6. In comparison, naturally aspirated engines have a higher CR by 1–1.5 units.

While the lower CR at full load results in lower pressures and temperatures at the end of compression in concert with a beneficial location of the center of combustion offering improved efficiency, a higher CR at part load without knock limitation is always beneficial in terms of thermal efficiency and thus fuel consumption (Figure 24). The application of a variable CR mechanism lends itself to resolving this conflict.

Over the years, in order to achieve variable compression, a variety of different concepts have been published (Schwaderlapp *et al.*, 2001; Drangel and Bergsten, 2000).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

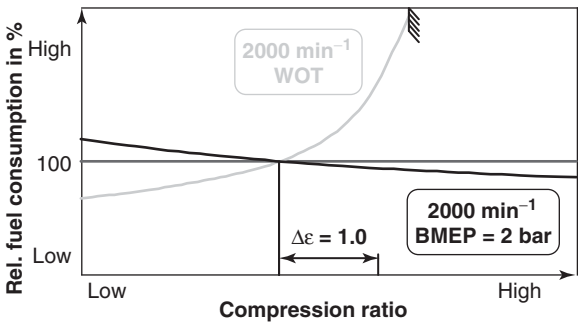


Figure 24. Fuel consumption depending on CR under part load and full load (WOT: wide open throttle) operation conditions.

Some of these concepts have already been realized in test vehicles (Pischinger *et al*., 2003). All of these systems entail a variation of the compression volume via a variable piston position at TDC. Figure 25 illustrates the classification of the systems according to type of TDC position variation mechanism.

The systems in the first group apply to unconventional cranktrains. There are different kinematic concepts of such multicomponent configurations using one moving pivot or support point. In general, the moving masses increase and more friction is added. However, as stated by Aoyama *et al*. (2006), an optimized kinematic design might result in an actual reduction of the friction.

The systems in the second group achieve the change in compression by varying the distance between the crankshaft and the cylinder head. The FEV system (Schwaderlapp *et al*., 2001) uses a crank in eccentric main bearings while the Saab system (Drangel and Bergsten, 2000) tilts the cylinder head relative to the crankcase. Both of these unconventional cranktrain systems that vary in the distance between the crankshaft and cylinder head require intense design modifications to existing engine architecture. The system can therefore only be applied to a new engine family.

The systems with variable kinematic lengths entail variable powertrain components that can be used instead of the conventional components and thus only require minor modifications in existing engine architectures. The FEV variable length connecting rod belongs to this group of VCR systems and has been continuously developed over the past years (Wittek, Tiemann, and Pischinger, 2009a, 2009b).

Investigations so far have shown that the fuel consumption improvement potential for a continuous CR adjustment over the European NEDC cycle is between 6% and 8% (Schwaderlapp *et al*., 2001; Pischinger *et al*., 2003). In this case, the CR can be optimized for each operating point. However, less complex approaches such as a two-stage VCR can result in close to the same fuel economy improvement. The fuel consumption improvement potential of a two-stage VCR system for a modern downsizing gasoline engine is between 5% and 7% depending on the driving profile and the degree of downsizing, as cycle simulations have shown (Weinowski *et al*., 2012). A two-stage VCR system has the following advantages: easily adaptable, low system complexity, and significantly lower cost. On the other hand, due to the knock limitation occurring now at part load caused by the optimized higher CR, the CR may need to be switched back to the lower value as designed for full load operation earlier.

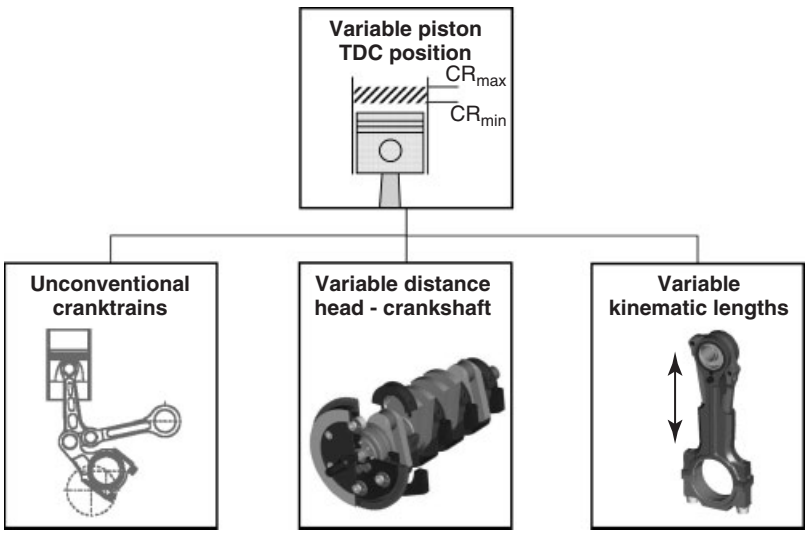


Figure 25. Classification of VCR systems with variable piston TDC position.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Furthermore, VCR systems can be used to exploit the benefit of higher octane rating of alternative fuels, such as CNG or ethanol, in bi-fuel or flex-fuel concepts. A fixed CR has to be designed according to the lowest expected octane rating, whereas VCR enables an efficient operation for different octane ratings. Thereby, VCR also allows compensating for worldwide gasoline octane rating fluctuation.

9 NOVEL ENGINE DESIGN

The ever increasing limitations of fuel consumption require a continuous improvement in engine efficiency, which is creating significant engine design challenges. The recent advanced developments of the combustion process and thermodynamics demand novel component designs, for example, variable valve timing and variable CR. Recent development trends show a tendency of increasing in-cylinder pressures and temperatures, which lead to greater mechanical and thermal loads on the engine parts. In addition, the market demands for lightweight designs and weight reduction increase even more the challenge for engine development. Current trends show several approaches to cope with these challenges.

One of the newest design trends targeting the reduction in friction forces on the cranktrain is the application of an offset between the crankshaft axis and the cylinder axis for inline engines. To achieve this reduction in the friction forces, the crankshaft has to be positioned away from the cylinder axis in a direction such that the piston arrives at TDC at a crankshaft angle $>0^\circ$. With this characteristic, the crank angle distance from TDC to BDC is also $>180^\circ$, resulting in an asymmetry of the cranktrain kinematics. Increasing the offset actually increases the friction forces during three of the four strokes of the engine cycle, as the side forces during these strokes are increased. Only during the power stroke is there a positive effect of the offset on the friction forces, as the overall distance between piston and crankshaft axis is larger than that in a conventional engine without an offset. This reduced friction during the power stroke increases the available torque on the crankshaft (Figure 26). A trade-off needs to be met between the reduced friction forces during the power stroke and the increased friction forces during the rest of the engine cycle, where usually the optimum offset is too large to be applied due to packaging constraints.

Another design trend is aimed at reducing moving and rotating masses in the engine, crankshaft, and camshafts by using hollow shaft designs. Special casting processes allow the production of hollow shafts. In addition, the application of assembled camshafts with hollow shafts has become a common approach to reduce masses in the valve train (Artur *et al*., 2010)

To avoid high thermal load on exhaust components such as the turbocharger turbine and the catalyst, latest engine designs employ integrated exhaust manifolds (Figure 27). The cylinder head and the exhaust manifold are realized as one aluminum cast part. In this configuration, the exhaust

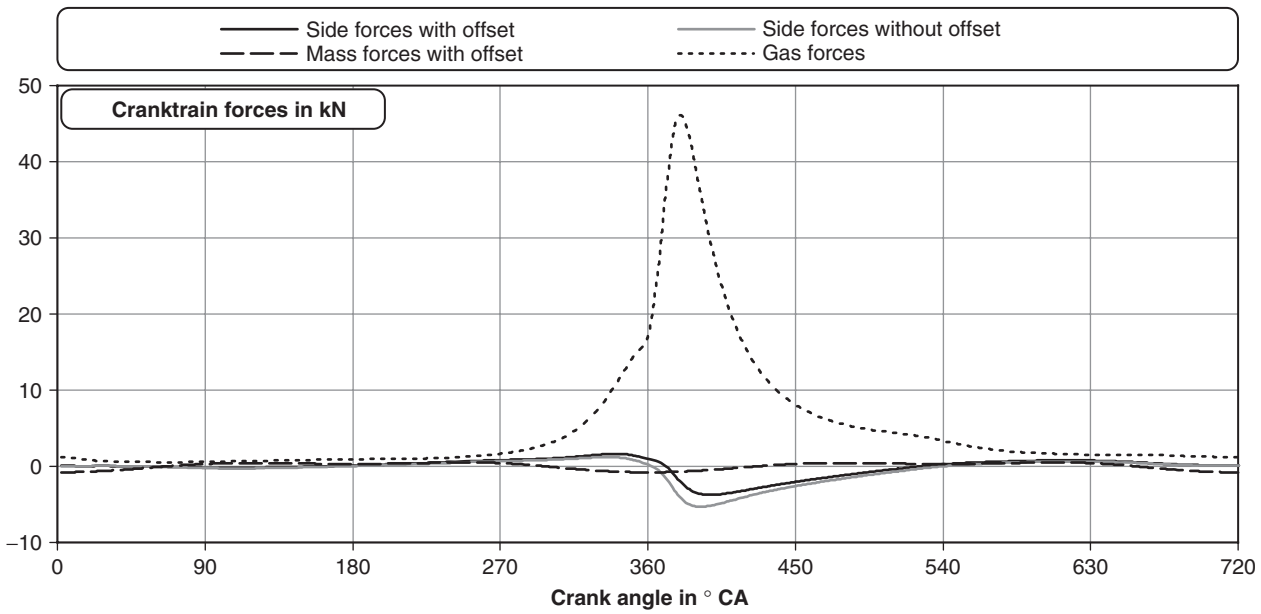


Figure 26. Cranktrain forces over the engine cycle; side forces are proportional to friction forces.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

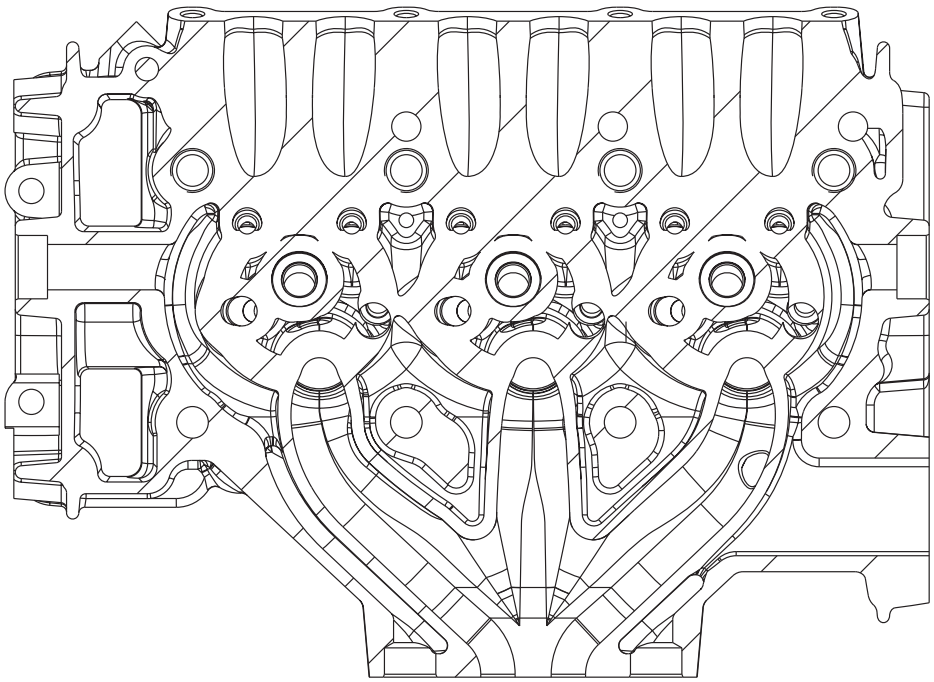


Figure 27. Cylinder head with integrated exhaust manifold (three-cylinder DI engine; three valves; exhaust: bottom).

ports can be cooled by the main water jacket of the cylinder head (Kirsch *et al*., 2011). As a consequence, the maximum temperature of the exhaust gas flow is reduced more rapidly, which reduces mixture enrichment requirements at full load operation. Furthermore, engine and catalyst warm-up can be achieved much faster, which results in further fuel consumption reductions (Ernst *et al*., 2011). Additionally, a beneficial outcome of the integrated exhaust manifold is the reduction of expensive heat-resistant Ni-alloys, which are used for conventional exhaust manifolds.

The reduction in internal friction is another major development target. Roller bearings show high potential for friction reduction. Plain bearings are still the prevalent type of bearings in the engine. Roller bearings in the cranktrain are yet to prove feasible for mass production applications. Nevertheless, roller bearings on assembled camshafts are already in use in current applications, bringing a further decrease in the friction forces on the valve train (Artur *et al*., 2010). Mass balance shafts with roller bearings have become as well state-of-the-art technology.

To achieve further improvements in fuel consumption, original equipment manufacturers (OEMs) are currently introducing engines with aluminum crankcases with spray-coated cylinder bore surfaces. This allows elimination of the cylinder liners, resulting in lighter engines as well as optimized water jackets between cylinder bores. Before honing the cylinder bore surface, with this process, a coating of a material which can be chosen according to the piston material is sprayed on a previously prepared cylinder bore surface. This has the advantage that an optimal material matching between piston and bore surface is achieved. With this optimal material matching, lower friction at the piston running surface is achieved as well as lower oil consumption (Pischinger, Ring, and Nijs, 2011b). Compared to other bore surface treatment processes, such as Nikasil, Alusil, and Lokasil, this process is cheaper, is easier to machine, has lower environmental impacts (compared to Nikasil), and better material matching is achieved.

Furthermore, the application of variable oil pumps shows high potential for the elimination of parasitic losses in the engine. In conventional oil pumps, oil flow and pressure rise with increasing engine speed. At high engine speeds, oil pressure and oil flow are higher than the required amounts. This results in unnecessary power consumption by the oil pump. With variable oil pumps, pressure or volumetric flow can be adjusted according to the engine demand, reducing parasitic power losses. There are several different concepts to achieve this, depending on the pump type.

Similar to the potential for reduction of parasitic losses through the application of variable oil pumps, comparable effects are achieved by the application of variable water pumps. However, the application of variable water pumps also allows an optimization of the thermal management, which leads to significant additional fuel savings.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

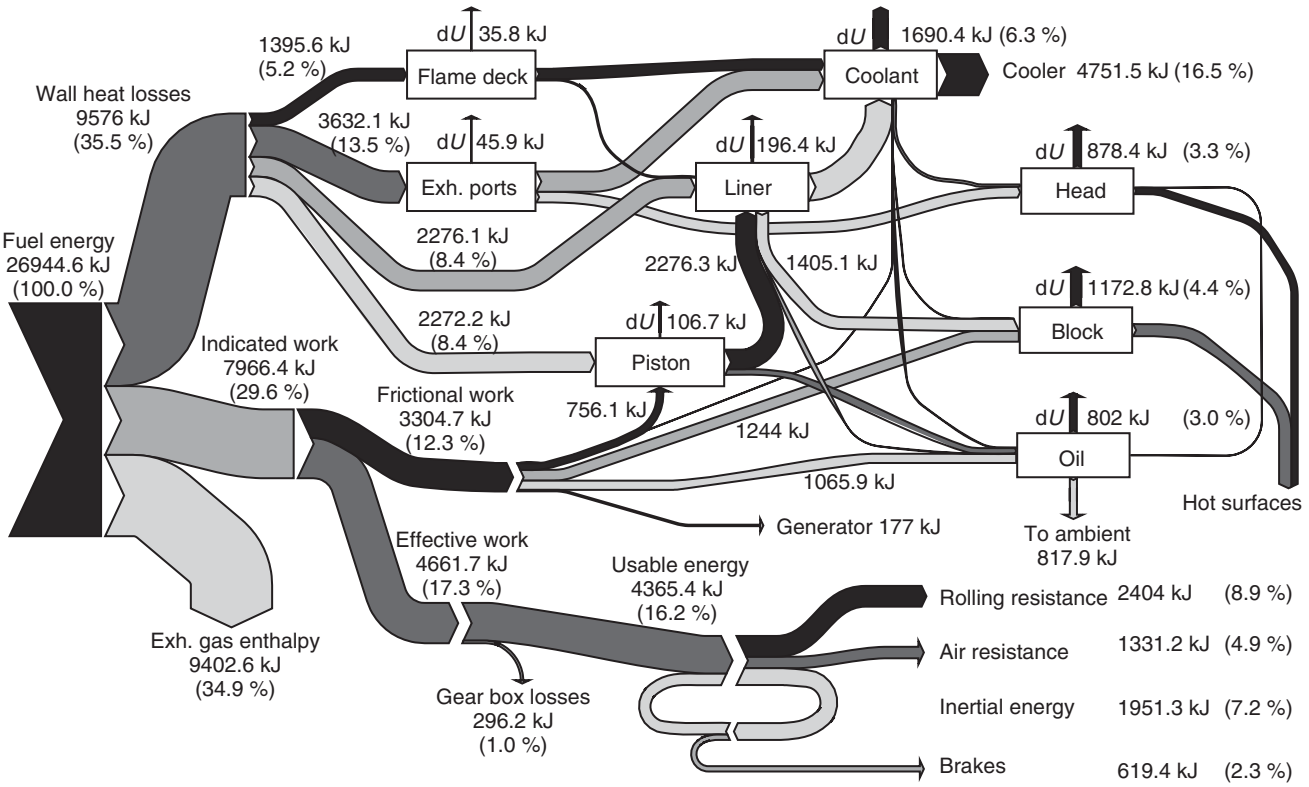


Figure 28. Exemplary cumulated energy distribution in NEDC.

10 THERMAL MANAGEMENT

Thermal management recently has moved more and more into the focus of new engine developments and makes up an important part of reaching lower fuel consumption rates. Figure 28 shows simulated cumulated energy flows of an upper middle class vehicle equipped with a four-cylinder 1.8-L gasoline engine when driving the NEDC. The diagram is a result of a thermal management simulation covering all relevant energy flows needed to describe a powertrain warm-up. Thermal inertias interact with heat inputs and outputs. The biggest heat source originates from the combustion process itself indicated as being wall heat losses which are input into thermal inertias close to the combustion chamber discretized as finite elements. Engine friction mainly is transferred into heat flowing into engine oil and engine structure and therewith has to be taken into account when simulating warm-up behavior.

The cumulated engine frictional work makes up a big portion of the fuel energy provided and thus yields the possibility for numerous fuel consumption improvements. Thermal management measures for reducing friction aim at a quick engine warm-up and at implementing advantageous engine operating conditions such as high system temperatures at low loads. The knowledge and manipulation of the heat distribution is the core business of thermal management. The effect of variable pumps for oil or coolant, different cooling concepts, artificial heat inputs, or any other variable parameter such as controllable piston cooling jets can be evaluated, as all relevant influences of any change in the system are covered.

As examples, thermal management analysis shows that the benefit of an electrical coolant pump strongly depends on its operating strategy, which should consider maximum allowable material temperatures in the cylinder head to guarantee structural strength and to avoid oil carbonization on the liner surface. In addition, split cooling concepts that separate the crankcase coolant jacket from the remaining cooling system to increase the liner temperature for lower piston friction show a measureable benefit in engine warm-up. Overall, depending on the operating strategy and system setup, benefits through split cooling, an electrical coolant pump, or a combination of both show advantages between 0.5% and 2% fuel consumption reduction in NEDC.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

REFERENCES

Adomeit, P., Sehr, A., Glück, S., *et al*. (2010) *Zweistufige Turboaufladung - Konzept für hochaufgeladene Ottomotoren Motortechnische Zeitschrift*, **71**, 5.

Alt, M., Schaffner, P., Krebs, W. *et al*. (2001) *Benzindirekteinspritzung in Kombination mit Aufladung*. 10th Aachen Automobile and Engine Technology Colloquium, Aachen.

Aoyama, S., Moteki, K., Takahashi, N. *et al*. (2006) *Untersuchung eines variablen Verdichtungsverhältnissystems mit einem Mehrfachverbindungsmechanismus und seine Auswirkung auf die Motorleistung*. 15th Aachen Automobile and Engine Technology Colloquium, Aachen.

Artur, C., Lemaitre, F., Schneider, F., *et al*. (2010) *Camshaft with roller bearings to reduce mechanical losses Motortechnische Zeitschrift*, (03), 28.

Balazs, A. and Pischinger, S. (2011) *Final Report "Untersuchung zur optimierten Auslegung von Hybridantriebsträngen unter realen Fahrbedingungen" FVV Herbsttagung 2011*, Informationstagung Motoren, Heft R 556 2011.

Bäumel, F., Jedro, J., Weber, C. *et al*. (2011) *Der Abgasturbolader für die dritte Generation der AUDI R4-TFSI-Motoren am Beispiel des neuen 1, 8l TFSI*, ATK Dresden.

Bick, W. (1990) *Einflüsse geometrischer Grunddaten auf den Arbeitsprozeß des Ottomotors bei verschiedenen Hub-Bohrungs-Verhältnissen*. Dissertation, RWTH Aachen.

Björnsson, H., Johansson, L. and Kunde, O. (2010) *World first turbo system with integrated sheet metal turbine-manifold for the new 2.0 GTDI engine*, ATK Dresden.

Borgmann, K., Fröhlich, K., Hofmann, R. *et al*. (2004) *Die Bedeutung der Reihensechszylinder Ottomotoren für BMW*. 13th Aachen Automobile and Engine Technology Colloquium, Aachen.

Bormann, D., Pingen, B., Müller, B. *et al*. (2009) *Der Antriebsstrang mit einem kleinen Downsizingmotor – Auslegungsstrategien und Systemkomponenten*. 30. Internationales Wiener Motorensymposium.

Breitbach, H., Vieweger, K., Bürthel, W. *et al*. (2007) *Vergleichende Betrachtung von Aufladesystemen beim Ottomotor*, ATK Dresden.

Brinkmann, F., Pingen, B., Walder, K. (2003) *Benzindirekteinspritzung mit Turboaufladung – ein Brennverfahren für Downsizingkonzepte*, Haus der Technik, München.

Budack, R., Kuhn, M., Trost, W. *et al*. (2009) *Vorteile auslassseitiger Ventiltriebsvariabilitäten beim Turbomotor*, Tagung: Variable Ventilsteuerung, Haus der Technik, Essen.

Central Data Repository (2011), *Monitoring of CO_2 emissions from passenger cars – Regulation 443/2009*, http://www.eea.europa.eu/data-and-maps/data/co2-cars-emission. 1e1fde93627786389bf10ee8372a6ddb (accessed 12 September 2011).

Dilthey, J. (2004) *Möglichkeiten zur Verbesserung der drosselfreien Laststeuerung beim otto-motorischen Prozess*. Dissertation, RWTH Aachen.

Drangel, H. and Bergsten, L. (2000) *Der neue Saab SVC Motor – ein Zusammenspiel zur Verbrauchsreduzierung von variabler Verdichtung, Hochaufladung und Downsizing*. 9th Aachen Automobile and Engine Technology Colloquium, Aachen.

Dünschede, T., Pischinger, S. (2010) *Vorstudie Motor B*, FVV Vorhaben Nr. 1015.

Eichler, F. and Heiselbetz, Ch. (2001) *The New Porsche Carrera Engine with 3.6 Liter Displacement and VarioCam Plus*. 10th Aachen Automobile and Engine Technology Colloquium, Aachen.

Environmental Protection Agency and Department of Transportation (2010) *Light-Duty Vehicle Greenhouse Gas Emission Standards and Corporate Average Fuel Economy Standards; Final Rule*, Washington, DC.

Ernst, R., Friedfeldt, R., Lamb, D. *et al*. (2011) *The New 3 Cylinder 1.0L Gasoline Direct Injection Turbo Engine from Ford*. 20th Aachen Automobile and Engine Technology Colloquium, Aachen.

European Union (2009a), *DIRECTIVE 2009/28/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC*, European Parliament, Brussels.

European Union (2009b), *Verordnung des Europäischen Parlaments und des Rates zur Festsetzung von Emissionsnormen für neue Personenkraftwagen im Rahmen des Gesamtkonzepts der Gemeinschaft zur Verringerung der CO_2-Emissionen von Personenkraftwagen und leichten Nutzfahrzeugen*, European Parliament, Brussels.

European Union (2011), *Amending Regulation (EC) No 715/2007 of the European Parliament and of the Council and Commission Regulation (EC) No 692/2008 as regards emissions from light passenger and commercial vehicles (Euro 6)*, European Commission, Brussels.

FEV GmbH (2011), *Benchmarking*, Aachen.

Flierl, R., Paulov, M., Schmitt, S., *et al*. (2011) Potenziale des vollvariablen Ventiltriebs beim Turbo−Ottomotor mit Direkteinspritzung *Motortechnische Zeitschrift*, **72**, 7−8.

Fraidl, G.K. (2010) *Einfluss der Elektrifizierung auf den Verbrennungsmotor, 4*, VDI Fachtagung Ventiltrieb und Zylinderkopf, Würzburg.

Freisinger, N., Friedrich, J., Karl, G. *et al*. (2010) *Zweistufige Turboaufladung an einem 4-Zylinder Ottomotor*; ATK Dresden.

Gand, B. (1986) *Einfluß des Hub-Bohrungs-Verhältnisses auf den Prozeßverlauf des Ottomotors*. Dissertation, RWTH Aachen.

Genender, P., Speckens, F.W., and Schürmann, G. (2011) Akustikentwicklung von Range-Extendern für Elektrofahrzeuge *MTZ – Motortechnische Zeitschrift*, **72** (3) 192−197.

Golloch, R. (2005) *Downsizing bei Verbrennungsmotoren ein wirkungsvolles Konzept zur Kraftstoffverbrauchssenkung*, Springer, Berlin.

Gottschalk, W. and Tschöke, H. (2000) *Untersuchungen zur Gemischbildung im Bereich des Ventilspaltes bei kleinen Ventilhüben und Drosselregelung*, FVV Vorhaben Nr. 674.

Haas, M. and Rauch, M. (2010) Elektrohydraulischer vollvariabler Ventiltrieb *Motortechnische Zeitschrift*, **71**, 3.

Hagelstein, D., Hentschel, L., Strobel, S. *et al*. (2009) *Die Aufladeentwicklung für den neuen 1.2l TSI Motor von Volkswagen*, ATK Dresden.

Hara, S., Nakajima, Y. and Nagumo, S. (1985) *Effects of Intake-valve closing timing on spark ignition engine combustion*. SAE 8500074.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Hu, P., Seibel, J., Zhang, H. *et al*. (2010) *Strategy of Range Extending Electric Vehicle Based on User's Approval*.

Indra, F. (2011) Zylinderabschaltung für alle Hubkolbenmotoren *Motortechnische Zeitschrift*, **72**, 10.

Kirsch, U., Hadler, J., Szengel, R. *et al*. (2011) *The New 1.0-Litre, 3-Cylinder MPI Engine for the UP!*. 20th Aachen Automobile and Engine Technology Colloquium, Aachen.

Korte, V., Fraser, N., Taylor, J., *et al*. (2011) Effizientes Downsizing für zukünftige Ottomotoren *Motortechnische Zeitschrift*, **72**, 5.

Kraftfahr-Bundesamt (KBA) (2011), *Statistics 2011*, Flensburg.

Landerl, C., Klauer, N. and Klüting, M. (2004) *Die Konzeptmerkmale des neuen BMW Reihensechszylinder Ottomotors*. 13th Aachen Automobile and Engine Technology Colloquium, Aachen.

Löbbert, P. (2006) *Möglichkeiten und Grenzen der Teillaststeuerung von Ottomotoren mit vollvariablem Ventilhub*. Dissertation, Technische Universität Dresden.

Lotterman, J., Schorn, N., Jeckel, D. *et al*. (2011) *New Turbocharger Concept for boosted Gasoline Engines*; ATK Dresden.

Luttermann, C., Schünemann, E., Klauer, N. (2008) *Enhanced valvetronic technology for meeting SULEV emission requirements*. SAE 2006-01-0849.

Middendorf, H., Theobald, J., Lang, L., *et al*. (2012) Der 1.4 l-TSI-Ottomotor mit Zylinderabschaltung *Motortechnische Zeitschrift*, **73**, 3.

Neußer, H.J., Kerkau, M., Schwarzenthal, D. *et al*. (2011) *40 Years of Porsche Turbo Engines The Basis for Future Innovations*. 32 Internationales Wiener Motorensymposium.

Nitz, N., Elendt, H., Ihlmann, A. *et al*. (2010) *INA Schiebenockensystem*. Schaeffler Kolloquium.

Pischinger, F., Bick, W., Hermanns, H.-J., and Peters, B. (2001) *Abschlußbericht des Sonderforschungsbereichs 224 Motorische Verbrennung*, Aachen.

Pischinger, S., Habermann, K., Yapici, K.I., *et al*. (2003) Der Weg zum konsequenten Downsizing *MTZ – Motortechnische Zeitschrift*, **64** (5), 398–405.

Pischinger, M., Wittek, K., Genender, P. *et al*. (2011a) *V2-Range Extender Module with FEVcom – a Barely Noticeable Companion in Your Electric Vehicle*. 20th Aachen Automobile and Engine Technology Colloquium, Aachen.

Pischinger, S., Ring, F. and Nijs, M. (2011b) *Reibungsminimierung der Kolbengruppe – Einfluss verschiedener Lineroberflächen auf die dynamische Kolbengruppenreibkraft*; 2. ATZ-Fachtagung, 29th and 30th November 2011, Esslingen.

Porsche AG, Dr. Ing. h.c. F. (n.d.), *Porsche Technologie Lexikon*, http://www.porsche.com/microsite/technology/default.aspx?pool=germany&ShowSingleTechterm=PTVCP&Category=&Model=&SearchedString=&SelectedVariant=PMTAll (accessed 29 April 2013).

Prevedel, K., Kometter, B., Neubauer, M. *et al*. (2011) *Aluminium statt Stahlguss – Konzept, Auslegung und Ergebnisse mit Diesel-VNT Ladern an Ottomotoren*, ATK Dresden.

Rajoo, S. and Martinez-Botas, R. (2008) Mixed flow turbine research: a review *Journal of Turbomachinery*, **130**, 044001–044011.

Ross, T. and Zellbeck, H. (2010) Neues ATL-Konzept von Vierzylinder-Ottomotoren *Motortechnische Zeitschrift*, **71**, 12.

Salber, W. (1998) *Untersuchungen zur Verbesserung des Kaltstart- und Warmlaufverhaltens von Ottomotoren mit variabler Ventilsteuerung*. Dissertation RWTH Aachen.

Sauerstein, R., Dabrowski, R., Becker, M. *et al*. (2009) *The Dual-Volute-VTG from BorgWarner – A New Boosting Concept for DI-SI-Engines*, ATK Dresden.

Schaeffler Technologies AG & Co. KG (n.d.-a), *UniAir*, http://www.ina.de/content.ina.de/de/press/press-media/pressmedia_detail.jsp?id=3342023 (accessed 13 November 2013).

Schaeffler Technologies AG & Co. KG (n.d.-b), *UniAir*, http://www.ina.de/content.ina.de/de/press/press-media/pressmedia_detail.jsp?id=3342023 (accessed 13 November 2013).

Scharf, J., Schorn, N., Smiljanovski, V. *et al*. (2010) *Methods for Extended Turbo-charger Mapping and Turbocharger Assessment*. 15 Aufladetechnische Konferenz, Dresden.

Schmidt, G., Flierl, R., Hofmann, R. *et al*. (1998) *Die neuen BMW-6-Zylindermotoren*. 19 Internationales Wiener Motorensymposium.

Schmuck-Soldan, S., Koenigstein, A. and Westin, F. (2011) *Two-Stage Boosting of Spark Ignition Engines*. 32 Internationales Wiener Motorensymposium.

Schwaderlapp, M., Pischinger, S., Yapici, K.I. *et al*. (2001) *Variable Verdichtung – eine konstruktive Lösung für Downsizing-Konzepte*. 10th Aachen Automobile and Engine Technology Colloquium, Aachen.

Schwerdel, U., Koch, A., Claus, H. *et al*. (2009) *Neue Turbolader mit elektrischer Waste-Gate Verstellung*, ATK Dresden.

Seibel, J. and Pischinger, S. (2007) *Final Report Optimized Layout of Internal Combustion Engines for Hybrid Powertrains FVV Frühjahrstagung 2007*. Informationstagung Motoren, Heft R 537.

Sonner, M., Wurms, R., Heiduk, T. *et al*. (2010) *Unterschiedliche Bewertung von zukünftigen Aufladekonzepten am stationären Motorprüfstand und im Fahrzeug*. ATK Dresden.

Sterlepper, J. (1992) *HC-Emissionen und Flammenausbreitung im Feuerstegbereich beim Ottomotor*. Dissertation, RWTH Aachen.

Tuttle, J. (1980) *Controlling engine load by means of late intake-valve closing*. SAE 800794.

Tuttle, J. (1982) *Controlling engine load by means of late intake-valve closing*. SAE 820408.

Unger, H. (2004) *Valvetronic - Der Beitrag des Ventiltriebs zur Reduzierung der CO_2-Emission des Ottomotors*. Die Bibliothek der Technik, Band 263, Verlag Moderne Industrie.

Warth, M., Bassett, M., Hall, J. *et al*. (2011) *Design and Development of the MAHLE Range Extender Engine*. 20th Aachen Automobile and Engine Technology Colloquium, Aachen.

Watson, N. and Janota, M.-S. (1982) *Turbocharging the Internal Combustion Engine*, Macmillan Education Ltd, Palgrave Macmillan, Basingstoke.

Weinowski, R., Sehr, A., Dieterich, C. *et al*. (2010) *Auf dem Weg zu 95 g/km CO_2 - Konzept eines aufgeladenen 3-Zylinder DI-SOHC-Ottomotors mit kleinem Bohrungs¬durch¬messer*. 19th Aachen Automobile and Engine Technology Colloquium, Aachen.

Weinowski, R., Wittek, K., Haake, B. *et al*. (2012) *CO_2-potential of a two-stage VCR system in combination with future gasoline powertrains*. 33 Wiener Motorensymposium, Wien.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Westbrook, C.K. (2000) Chemical kinetics of hydrocarbon ignition in practical combustion systems in *Proceedings of the Combustion Institute*, vol. 28 (eds V. Sick and L.P.H. deGoey), pp. 1563–1577.

Wichart, K. (1987) *Möglichkeiten und Maßnahmen zur Verminderung der Ladungswechselverluste beim Ottomotor*. VDI-Fortschrittsberichte Reihe 12, Nr. 91.

Wittek, K., Tiemann, C., and Pischinger, S. (2009a) Zweistufiges variables Verdichtungsverhältnis durch exzentrische Kolbenbolzenlagerung *Motortechnische Zeitschrift*, **70** (02).

Wittek, K., Tiemann, C. and Pischinger, S. (2009b) *Two-stage variable compression ratio with eccentric piston pin and exploitation of crank-train forces*, Paper Number 09PFL-0468, SAE Congress, Detroit.

Wittek, K., Speckens, F.W., Pischinger, M. *et al*. (2011) *Combustion Engines: Enabler for E-Mobility*. 6 MTZ Fachtagung Der Antrieb von morgen, Wolfsburg.

Wurms, R. (1994) *Einfluss einlassseitig erzeugter Ladungsbewegungen auf das Betriebsverhalten von Vierventil- Ottomotoren*. Dissertation, RWTH Aachen.

Wurms, R., Dengler, S., Budack, R. *et al*. (2006) *Audi valvelift system – ein neues innovatives Ventiltriebssystem von Audi*. 15th Aachen Automobile and Engine Technology Colloquium, Aachen.

Wyatt, S., Jörgl, V., Becker, M., *et al*. (2011) Konzentrische Verstellnockenwellen für Otto- und Dieselmotoren *Motortechnische Zeitschrift*, **72**, 10.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto143
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Stoichiometric Exhaust Emission Control

Joseph E. Kubsh

Manufacturers of Emission Controls Association, Arlington, VA, USA

1 Introduction 1
2 Three-Way Catalytic Converters 3
References 14

1 INTRODUCTION

One of the most important technology bases that have emerged from the automotive industry in the past 50 years is the development, introduction, and continued evolution of automotive emission control technology. The centerpiece of this technology base is the three-way catalyst used on gasoline, stoichiometric, spark-ignited vehicles in all major world markets today. The name three-way catalyst was applied to catalytic controls that were capable of reducing all three criteria pollutants: carbon monoxide (CO), oxides of nitrogen (NO*x*), and volatile organic compounds (VOCs) within a narrow range of inlet exhaust gas compositions that corresponded to approximately the stoichiometric air/fuel ratio of the engine. Nowadays, more than 95% of the new gasoline automobiles sold around the world are equipped with catalytic converters that utilize three-way catalysts, adding to more than 750 million vehicles worldwide that have been equipped with catalysts since their first introduction in the United States in 1975.

Automotive catalytic emission controls were pioneered in the United States in response to public health concerns associated with elevated ambient ozone levels stemming, in part, from automotive tailpipe emissions of hydrocarbons (HCs) and oxides of nitrogen. These public health concerns were translated into emission control regulatory programs by both the US federal government and the state of California. On the federal level, the Clean Air Act Amendments of 1970 mandated significant reductions in automobile tailpipe emissions of CO, NO*x*, and VOCs starting in 1975. These federal standards led to the introduction of oxidation catalysts on automobiles starting with the 1975 model year to control CO and VOCs, and the use of three-way catalysts to control CO, NO*x*, and VOC tailpipe emissions starting in 1981. California, with severe smog problems in its large metropolitan areas, was provided with its own authority to set automobile emission standards and has typically led the US federal government and the world with the tightest standards requiring the best available emission control technology for automobiles.

1.1 Technology forcing exhaust emission regulations

Since the mid-1970s, US federal and California light-duty motor vehicle tailpipe emission regulations have been continually pushed to lower levels in response to air quality concerns. At the forefront of these new waves of regulatory programs aimed at significantly reducing emissions from light-duty vehicles are the US Environmental Protection Agency's (EPA) Tier 2 and the California Air Resource Board's (ARB) Low Emission Vehicle II (LEV II) programs. California acted first, adopting their LEV II program in the late 1998, followed by EPA finalizing the Tier 2 regulations in December 1999. Both the ARB LEV II regulations and the EPA Tier 2 regulations began their phase-in with the 2004 model year. In a parallel or slightly delayed time frame relative to these US initiatives, Europe (Euro 3 and Euro 4 regulations), Japan (Japan Low Emission Vehicle regulations), and Korea (Korea Low

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Emission Vehicle regulations) also established new, more severe light-duty emission regulations during the 1990s and established even more stringent light-duty vehicle emission standards in the 2000–2011 time frame (e.g., Euro 5 and Euro 6 regulations). Emission regulations for new vehicles based on the use of three-way catalyst technologies are now being implemented in almost every world market including large emerging markets in Brazil, India, and China. The introduction of catalytic converters in the United States and other world markets also required these countries to introduce unleaded gasoline as vehicle operation on leaded fuel results in dramatic deactivation of the active precious-metal-based catalytic materials (e.g., Pt, Pd, and Rh) present in three-way catalytic converters (TWCs).

All these current US light-duty vehicle emission programs require significant reductions in HC, CO, and NO*x* emissions relative to vehicle emission requirements associated with the regulations that precede each of these new emission programs (e.g., EPA's Tier 1 or California LEV I regulations). The LEV II regulations, for example, maintain tight HC emission levels established in the LEV I program (adopted in 1990; implementation began with the 1994 model year) but significantly reduce NO*x* emission requirements compared to LEV I requirements. The Tier 2 program draws from both the California LEV I and LEV II programs in significantly tightening both HC and NO*x* tailpipe emissions relative to Tier 1 regulations that were first implemented with the 1994 model year. An important input into each of these regulatory processes was the ability of emission control technologies to meet these increasingly tighter tailpipe emission standards in a cost-effective manner. The Manufacturers of Emission Controls Association (MECA) provided important technical inputs into the EPA Tier 2 and California LEV II rulemaking processes by completing a successful test program in the late 1990s that demonstrated that advanced three-way catalysts were capable of significantly reducing exhaust emissions from four different Tier 1-compliant passenger cars and trucks. Details of this test program were reported in a Society of Automotive Engineers (SAE) technical paper published in 1999 (Webb *et al*., 1999). Compared to precontrolled vehicles sold in the United States before 1975, today's Tier 2 and LEV II cars and trucks are meeting emission standards that require reductions of up to 98% with respect to VOCs, 96% for CO, and 98% for NO*x*. MECA completed a second light-duty gasoline vehicle test program in 2006 that demonstrated that advanced TWC systems allow even the heaviest light-duty gasoline trucks (e.g., SUVs and larger pickup trucks) to achieve very low exhaust emissions of HCs and NO*x*. Results of this test program are available in Anthony and Kubsh (2007).

Some additional details concerning the EPA Tier 2 and California LEV 2 emission regulations are provided here as each program includes the tightest light-duty emission certification categories currently in place in the world today. Full useful life tailpipe emission standards for the fully phased-in US EPA Tier 2 and California LEV II programs are summarized in Tables 1 and 2, respectively. Each of these programs provides auto manufacturers with several different certification categories to choose from for their light-duty vehicle fleet. In the case of EPA's Tier 2 program, auto manufacturers may select appropriate vehicle emission certification categories that allow their fleet of new vehicles to achieve a fleet average NO*x* emission limit of 0.07 g/mi (equivalent to a Tier 2, Bin 5 NO*x* fleet average). To comply with California's LEV II requirements, auto manufacturers must select appropriate vehicle emission certification categories that allow them to meet a declining annual fleet average NMOG standard that reached 0.035 g/mi NMOG for passenger cars and 0.043 g/mi NMOG for heavier, light-duty trucks in 2010.

Tailpipe emissions are measured on a chassis dynamometer using the US Federal Test Procedure (FTP, a vehicle speed vs time driving cycle). The concept of multiple certification categories was first introduced with California's LEV I program with Transitional Low Emission Vehicle (TLEV), Low Emission Vehicle (LEV), and Ultra-Low Emission Vehicle (ULEV) certification options that varied by vehicle weight class (e.g., passenger

Table 1. California LEV II 120,000 mile FTP tailpipe emission limits.

Certification level	NMOG (g/mi)	CO (g/mi)	NO*x* (g/mi)
LEV-2	0.090	4.2	0.07
LEV-2/LDT2[a]	0.090	4.2	0.10
ULEV-2	0.055	2.1	0.07
SULEV	0.010	1.0	0.02

[a]The LEV-2/LDT2 certification category is limited to no more than 4% of the LDT2 light-duty truck production for a given manufacturer.

Table 2. US EPA Tier 2 120,000 mile FTP tailpipe emission limits.

Certification level	NMOG (g/mi)	CO (g/mi)	NO*x* (g/mi)
Bin 1	0.0	0.0	0.0
Bin 2	0.010	2.1	0.02
Bin 3	0.055	2.1	0.03
Bin 4	0.070	2.1	0.04
Bin 5	0.090	4.2	0.07
Bin 6	0.090	4.2	0.10
Bin 7	0.090	4.2	0.15
Bin 8	0.125	4.2	0.20

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

car and light-duty truck weight classes). The EPA Tier 1 light-duty emission regulations also had weight class-specific emission regulations but only one set of emission standards for each gasoline vehicle weight class. The Tier 2/LEV II programs have several common features that are also significant changes from either Tier 1 or LEV I requirements: (i) fuel neutral requirements (emission standards are equivalent for gasoline and diesel-fueled vehicles); (ii) 120,000 mile full useful life durability; and (iii) a single set of standards that does not vary with light-duty vehicle weight class [up to 8500 lb. gross vehicle weight for all passenger cars and light-duty trucks and up to 10,000 lb. for medium-duty passenger vehicles (MDPVs)]. Treating passenger cars and light-duty trucks on an equivalent emission basis is an important focus for both the Tier 2 and LEV II programs. Both of these programs place a premium on cold-start emission performance and high emission system efficiencies with respect to NO*x* emissions.

Both EPA and California are in the process of proposing and finalizing a future set of light-duty vehicle emission standards that, once finalized, will further reduce US vehicle emission limits to a fleet average level consistent with California's current SULEV emission limit and EPA's Tier 2, Bin 2 emission limit by 2025. California's LEV III emission standards and EPA's Tier 3 emission standards are both expected to be finalized in 2013. A phased-in compliance schedule for California's LEV III emission limits begins in model year 2015, whereas EPA's Tier 3 program is expected to begin in model year 2017. In addition to tighter emission standards for HCs, CO, and NO*x*, the LEV III and Tier 3 proposals are also expected to tighten emission standards for particulates from future light-duty vehicles.

Reaching the tailpipe emission levels associated with today's Tier 2 and LEV II emission regulations, or future Tier 3 and LEV III emission limits on stoichiometric gasoline vehicles, requires a concerted systems approach that includes the use of advanced spark-ignited engines, advanced engine control strategies, clean fuels, clean lubricants, and advanced emission control technologies. Both ARB and EPA have included the clean fuel component in their LEV II and Tier 2 regulatory programs with respect to gasoline sulfur levels. ARB established a 30 ppm sulfur average for gasoline as a part of their California Phase II reformulated gasoline requirements. This sulfur level was further reduced to an average of 15 ppm sulfur starting in 2004 with the introduction of California Phase III reformulated gasoline regulations and was capped at 20 ppm starting in 2012. Similarly, the EPA included gasoline sulfur level regulations as an integral part of their Tier 2 regulatory package with the phase-in of 30 ppm average S levels started in 2005. EPA has proposed a 10 ppm gasoline sulfur average in conjunction with its Tier 3 light-duty vehicle emissions proposal. A recent publication by workers at Umicore (Ball, Clark, and Moser, 2011) showed that performance degradation of an advanced three-way catalyst system on a PZEV-class four-cylinder passenger car is still found of gasoline sulfur levels at 33 ppm. This publication shows that the lost catalyst performance observed at 33 ppm gasoline sulfur levels can be mitigated by either purging the catalysts of sulfur by operating the vehicle at higher speeds and loads or by reducing gasoline sulfur levels down to 3 ppm. Lubricant constituents such as phosphorus and inorganic elements such as Zn and Ca have also been shown to act as catalyst poisons or catalyst masking agents driving lubricant producers to optimize lubricant formulations to insure adequate engine lubrication characteristics with minimal impacts on catalyst performance and driving engine designers to minimize engine oil consumption characteristics of advanced engines. Clean fuels and lubricants are a necessary prerequisite for maintaining the high performance levels of the advanced engine and emission systems required for Tier 2/LEV II compliance, as well as future compliance with Tier 3/LEV III emission limits.

2 THREE-WAY CATALYTIC CONVERTERS

The TWC has been the primary emission control technology on light-duty gasoline stoichiometric vehicles since the early 1980s. The use of TWCs, in conjunction with oxygen sensor-based, closed-loop fuel delivery system, allows for simultaneous conversion of the three criteria pollutants, HCs, CO, and NO*x*, produced during the stoichiometrically calibrated air/fuel combustion process of an internal combustion, spark-ignited engine. Figures 1 and 2 depict a cut-away drawing and a cut-away photo of typical TWCs, one using a ceramic substrate and the other using a metallic substrate. The active catalytic materials are present as a thin coating of precious metals (e.g., Pt, Pd, and Rh) and oxide-based inorganic promoters and support materials on the internal walls of the honeycomb substrate. The substrate typically provides a large number of parallel flow channels to allow for sufficient contacting area between the exhaust gas and the active catalytic materials without creating excess pressure losses.

Catalytic materials are typically applied by contacting the substrate with a water-based slurry containing the active inorganic catalyst materials. The coated substrate is contained within an outer metal-based shell that facilitates connection of the converter to the vehicle's exhaust system through flanges or welds. The honeycomb-based

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

substrates are typically either ceramic or metal foil-based. Cordierite, a magnesium alumino-silicate compound, is the preferred ceramic substrate material because of its low coefficient of thermal expansion, good mechanical strength characteristics, and good coating adhesion properties. The ceramic substrate is formed as a single body using an extrusion process followed by high temperature firing. Metal-foil-based substrates are made from thin ferritic-based specialty stainless steel foils brazed together to form the parallel flow passages. The ferritic foil alloy provides good oxidation resistance in the exhaust environment, good mechanical strength, and an oxidized surface that promotes good adhesion of the catalytic coating to the foil. In the case of ceramic substrates, a special oxide fiber-based mounting material (typically referred to as a *mat*) is used between the substrate and the metal outer shell to hold the substrate in place, provide thermal insulation, and cushion the ceramic body against the shell. The outer metal shell or mantle is an integral part of the metal substrate production scheme, and no additional mounting materials are generally required. As shown in Figures 1 and 2, in some cases, the converter housing or "can" can be surrounded by a

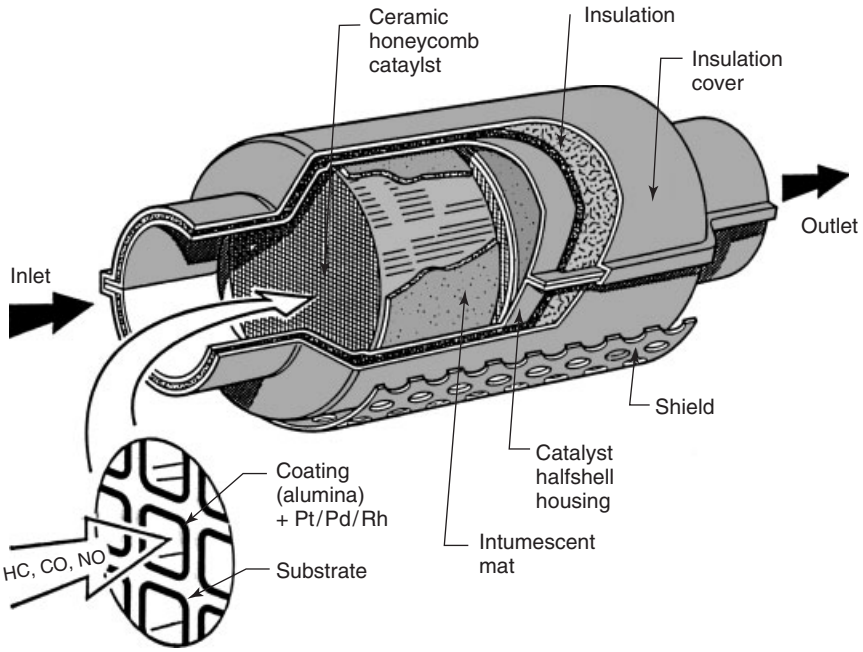


Figure 1. Three-way catalytic converter with ceramic substrates.

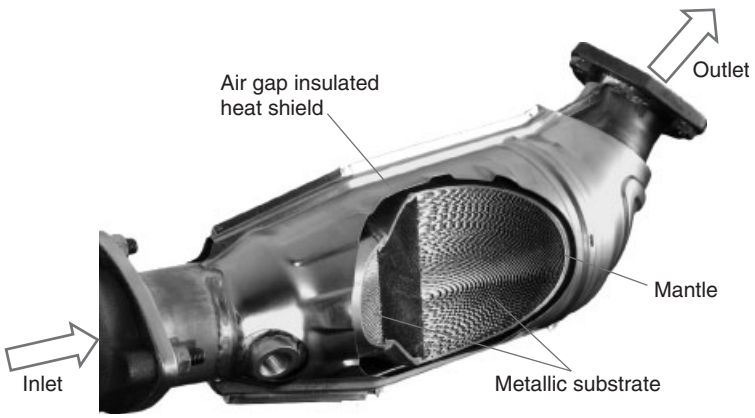


Figure 2. Three-way catalytic converter with metallic substrates. (Reproduced by permission of Emitec Gesellschaft für Emissionstechnologie mbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

second metal shell with an annular gap between these two metal shells. This type of arrangement provides additional heat insulation to the converter. The annular region between the two shells may be left as an air gap or filled with an insulating material such as an inorganic fiber-based material.

Although the primary components and function of a TWC have remained relatively constant during its more than 30 years of use on light-duty gasoline vehicles, each of the primary converter components (catalytic coating, substrate, and mounting materials) has gone through a continuous evolution and redesign process aimed at improving the overall performance of the converter while maintaining a competitive cost-effectiveness of the complete assembly. The performance-based catalytic converter re-engineering effort has had three main focuses: (i) wide application of close-coupled converters mounted near the exhaust manifold of engines for improved performance following a cold engine start; (ii) the development of thin-wall, high cell density substrates for improved contacting efficiency between the exhaust gas and the active catalyst and lowering the thermal mass of the converter; and (iii) the design of advanced, high performance TWCs for both close-coupled and underfloor converter applications that emphasize excellent thermal durability and efficient use of the precious metals platinum, palladium, and rhodium. Each of these three emission control technology platforms is discussed in more detailed in subsequent sections of this chapter.

Advanced TWC formulations often utilize multilayer architectures and/or axial placement of different catalyst materials along the length of the substrate that allow for the optimization of specific catalytic functions (e.g., improved light-off characteristics or improved overall efficiency for reducing HCs, CO, and/or NOx). These advanced catalysts also utilize a variety of advanced materials (in addition to the active precious metals) that promote the oxidation and reduction reactions associated with three-way catalysts and allow these catalysts to maintain activity in severe thermal exhaust environments. Catalyst substrate channel or cell densities as high as 1200 cells/in^2 have been used on production catalytic converters with 600 cells/in^2 substrates used in many late model vehicle applications. A similar re-engineering effort has occurred with other exhaust system components such as exhaust manifolds and exhaust pipes that complement improvements in catalytic converter technology. The focus of these manifold and other exhaust component improvements has been exhaust system thermal management and heat conservation through the use of low thermal mass, air-gap-insulated components, or other heat insulation strategies.

Current state-of-the-art, stoichiometric gasoline emission control systems are defined by SULEV (Super Ultra-low Emission Vehicle) or PZEV (Partial Zero Emission Vehicle) compliant light-duty vehicle sold in the US market. There are a number of recent references (Inoue and Mitsuishi, 2009; Matsuzono *et al*., 2008; Laurell, Dahlgren, and Vaisanen, 2007; Kidokoro *et al*., 2003; Oguma *et al*., 2003) that describe these systems that typically include combinations of close-coupled and underfloor converter systems that utilize high performance three-way catalysts displayed on high cell density substrates. These SULEV/PZEV systems utilize advanced cold-start calibration schemes including cold-start engine spark retard, higher cold-start idle speeds, and/or secondary air injection during the initial engine cold-start to accelerate the warm-up of the close-coupled converter within a few seconds after a cold engine start.

2.1 Close-coupled converters

Achieving high conversion efficiencies for both HC and NOx emissions during normal vehicle operation represented by the FTP driving cycle, for example, has focused attention on cold-start performance of catalytic converters for both US Tier 2 and LEV II light-duty applications. LEV I HC emission requirements introduced by California in 1994 provided the first regulatory driver that placed importance on cold-start emissions. Numerous studies published in the late 1980s and 1990s have discussed the high percentage of FTP driving cycle emissions associated with the early stages of vehicle operation following a cold engine start situation (Hughes and Witte, 2002; Pfalzgraf *et al*., 2001; Nishizawa *et al*., 2001; Brueck *et al*., 2001; Domesle *et al*., 2001; Williamson *et al*., 2000; Lafyatis *et al*., 2000; Holy *et al*., 2000; Moore *et al*. 1999; Webb *et al*., 1999; Ehmann *et al*., 1999; Takahashi *et al*., 1998; Kishi *et al*., 1998). This is especially true for 1990s vintage vehicles sold in the United States designed to comply with less severe Tier 1 emission standards. HC tailpipe emission profiles during FTP testing of Tier 1 vehicles are generally dominated by emissions emitted during the first 1–2 min of operation after the cold-start. This large fraction of cold-start emissions in Tier 1 vehicles stemmed from significant fuel enrichment used by auto manufacturers to facilitate engine start under cold conditions and significant delays in converter warm-up to catalyst operating temperatures required for high conversion efficiencies (e.g., 350°C or higher). Heat-up delays were usually associated with relatively long distances and the associated poor heat transfer between the converter location and the engine exhaust ports. NOx emission profiles also have a component related to cold-start operation but are generally distributed more uniformly through the FTP

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

driving cycle on Tier 1 certified vehicles because of NO*x* emission events associated with vehicle accelerations and decelerations.

To more effectively deal with cold-start emissions, converter volumes have been moved closer to the engine exhaust ports to minimize exhaust system heat losses and accelerate the heat-up of catalysts during the critical time following engine start. Converters located near the engine exhaust valves (e.g., at the exit of the exhaust manifold) are referred to as *close-coupled converters* (or sometimes light-off converters or preconverters). LEV I and ULEV I compliant light-duty vehicles introduced in the mid-1990s were the first significant applications of exhaust systems featuring close-coupled catalytic converters. In some applications (typically smaller displacement engines), a vehicle may have all or a large fraction of the required catalyst volume located close to the engine exhaust manifold. In other applications (typically larger displacement engines), the exhaust system will include smaller volume converters located close to the engine followed by a larger converter volume located further downstream in the exhaust in an underfloor location. In these multiple converter exhaust schemes, the size of the close-coupled converter is balanced between thermal mass (minimal catalyzed substrate mass for faster heat-up), diagnostic (adequate oxygen storage capacity), and durability considerations (sufficient volume to maintain required performance over extended mileage).

In larger engines, dual exhaust system configurations are often used with parallel systems for each cylinder bank (in the case of V-type engine designs) or groups of cylinders (in the case of in-line engine designs). These parallel systems may each incorporate close-coupled and underfloor converters or parallel close-coupled converters that lead into a y-pipe and a single underfloor converter. A schematic of an exhaust system layout featuring dual close-coupled converters flowing into a single underfloor converter is shown in Figure 3. Owing to their close orientation to the engine, the close-coupled converter(s) can reach temperatures required for high conversion efficiencies of HCs, CO, and NO*x* in 30 s or less following engine start, compared to heat-up times of 60 s or more associated with underfloor-only converter systems.

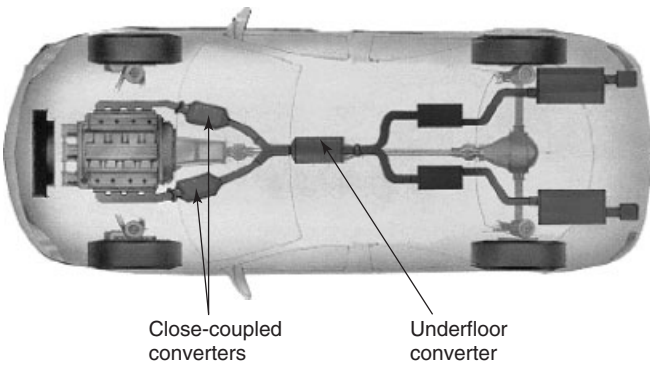


Figure 3. Exhaust system with close-coupled converters.

Fast dynamic converter heat-up, a requirement for low cold-start emissions, is also facilitated by advanced cold-start engine calibration strategies. These strategies include retardation of the engine spark, reduced idle speed, use of secondary air injection, and/or lean start strategies. Numerous examples of these cold-start strategies have been described in the literature (Inoue and Mitsuishi, 2009; Laurell, Dahlgren, and Vaisanen 2007; Kidokoro *et al*., 2003; Oguma *et al*., 2003; Matsuzono *et al*., 2003; Brueck *et al*., 2002) and are a key part of the systems approach required to achieve high conversion efficiencies for HC and NO*x* at the early stages following engine start. Each of these engine start-up strategies seeks to maximize conditions at the close-coupled converter that accelerate its heat-up following engine start (e.g., additional unburned fuel to combust over the catalyst, minimized total exhaust flow during initial engine idle, and slight excess of oxygen to combustibles in the exhaust to promote full oxidation at the catalyst).

Rapid converter heat-up also has placed greater emphasis on exhaust system thermal management. Efficient transfer of heat generated during the combustion process to the catalytic converter with minimal heat losses to the surrounding environment is facilitated by insulated exhaust manifolds and insulated exhaust pipes (Kidokoro *et al*., 2003; Oguma *et al*., 2003; Pfalzgraf *et al*., 2001; Webb *et al*., 1999). The preferred method of insulation is through the use of low thermal mass and air gap components. Insulated manifolds and pipes featuring dual wall construction separated by air gaps have been developed to improve light-duty vehicle cold-start and warm-start emission performances. These air gap components generally make use of a thin, low thermal mass, durable inner wall to facilitate fast heat-up characteristics. An air gap between the thinner inner wall and a thicker outer wall provides insulation to minimize heat losses between the engine and the converter(s). These air gap exhaust components provide significant reductions in converter heat-up during the FTP test protocol, which in turn provides significant reductions in cold-start and warm-start vehicle emissions.

Placement of catalytic converters closer to the engine results in dramatic reductions in cold-start emissions of all criteria pollutants (especially HC and CO emissions that are most associated with cold engine start conditions). The close-coupled converter environment also raises converter maximum operating temperatures relative to underfloor environments. This, in turn, has placed added

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

demands on the thermal durability of catalysts and other converter components used in these more severe close-coupled converter applications. In particular, fiber-based mounting materials and packaging assemblies used with ceramic substrates have been reengineered and optimized to meet these more severe thermomechanical environments, as well as the longer durability requirements associated with the Tier 2 and LEV II emission regulations. Similarly, metal substrate construction methods and brazing schemes have also been optimized for the high mechanical loads and high temperatures encountered in close-coupled applications. A discussion of high temperature catalyst designs is presented in a subsequent section of this chapter.

2.2 High cell density substrates

Tier 1 compliant vehicles generally relied on substrate designs that utilize straight flow channeled monoliths with square cross-sectional channel openings. Channel sizes that equate to 400 channels or cells per square inch of frontal area (designated as 400 cpsi) became an industry standard for many applications in the late 1980s and 1990s. In Tier 1 applications of ceramic substrate designs with 400 cpsi, ceramic substrate wall thickness was typically 0.0065 in or 6.5 mils, with some limited usage of 400 cpsi substrates with 8 mil walls. Limited applications of ceramic monoliths with triangular shaped cells were also used for Tier 1 applications with cell densities of 236 or 300 cpsi (wall thickness of 6.5–11.5 mils). Metal substrates were also introduced with channel densities of up to 400 cpsi but with thinner metal foil walls that were typically 50 μm (approximately 2 mils) in thickness. These standard metal substrate designs typically utilize sinusoidally corrugated metal foils layered between flat foils to produce parallel flow channels.

Interest in automotive emission systems with high conversion efficiencies and improved cold-start performance to meet more severe emission requirements, such as Tier 2/3 and LEV I/II/III standards, encouraged the development of a new generation of both ceramic and metallic substrate designs that offer significantly higher cell densities (more flow channels per cross-sectional area) and thinner walls separating flow channels. These two key substrate characteristics provide increased geometric surface area (GSA) per unit volume of monolith for efficient distribution of the active catalytic coating, relatively small flow channels (or more precisely, relatively small values for the channel hydraulic diameter) for good heat and mass transfer characteristics, and reduced substrate thermal mass for faster heat-up during emission critical cold-start events. Figure 4 provides a comparison of relative specific geometric areas and bulk densities of ceramic substrates with progressively higher cell densities and thinner wall thickness.

As discussed in the many references associated with these high cell density substrates, substrate GSA is an important physical property in heterogeneous catalysis associated with the effective mass transfer of reactants present in the exhaust stream of an engine (e.g., HCs, CO, NOx, H_2O, and O_2) to the solid surfaces that contain the active catalytic sites (references include Mueller-Haas *et al*., 2003; Hughes *et al*., 2003; Hughes and Witte, 2002; Leonhard *et al*., 2002; Brueck *et al*., 2002). Increasing this specific geometric area provides for more efficient contact between the reactants and the active catalytic sites and, in turn, a higher overall conversion efficiency of these reactants in a given volume of catalyzed monolith. Increasing cell density at a constant monolith wall thickness provides increased GSA but results in higher bulk density or thermal mass because of the resulting higher fraction of walls per given cross-sectional area (or, stated in another way, higher cell density at a constant wall thickness lowers the fraction of the frontal area open to the flow of exhaust gas). To compensate for this bulk density effect, substrate manufacturers have successfully developed high cell density products with significantly thinner walls than the "standard products" used primarily in Tier 1 applications. For example, ceramic

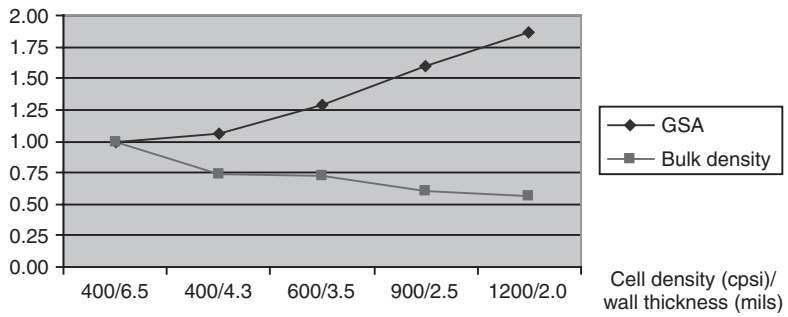


Figure 4. Relative geometric area and bulk density of ceramic substrates.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

substrates with 6.5 mil walls offered in "standard products" have been reduced to wall thickness in the range 1.5–3.5 mils in high cell density substrates. Similarly, metal substrates utilize 50 μm foils in "standard products" with high cell density products typically constructed with foil thickness ranging from 20 to 40 μm (approximately 0.8–1.6 mils). Thinning the monolith walls provides significant reductions in the thermal mass/bulk density of high cell density products. This low thermal mass characteristic enables catalyst-coated substrates to heat-up more quickly than heavier, "standard" wall thickness substrates. Fast dynamic heat-up of converters is key to achieving low tailpipe emissions during the critical cold-start and warm-start periods associated with normal driving operations and required to comply with Tier 2 and LEV II emission regulations. To further illustrate the properties and benefits associated with thin wall, high cell density substrates, results from three technical papers are briefly discussed later.

Hughes and Witte (2002) completed a comprehensive study of the impacts of high cell density substrates on light-duty vehicle emission performance in both the FTP and US06 test cycles. Their study made use of ceramic substrates covering a range of cell densities, including the "standard" ceramic substrate product with 400 cpsi/6.5 mil wall thickness, used in many Tier 1 applications, and high cell density, thin wall ceramic substrates such as 600 cpsi substrates with 3.5 and 4.5 mil wall thickness, and 900 cpsi substrates with 2.5 mil wall thickness. Table 3 summarizes ceramic substrates used in this study along with their accompanying properties including specific GSA and bulk density.

The performance of these substrates was investigated by catalyzing each substrate with an identical advanced Pd/Rh TWC (100 g/ft^3 total precious metal loading with Pd/Rh = 14/1; all substrates coated with a total coating weight of 140 g/L of substrate), aging the converters containing these catalyzed substrates using a Ford accelerated aging protocol, and performing triplicate FTP and US06 drive cycle tests on each aged converter. The Ford accelerated aging protocol was performed on an engine dynamometer and simulated approximately 50,000 miles of actual service life. FTP and US06 chassis dynamometer tests were run using a 2.0 L, four-cylinder, four-valve test vehicle with a single aged converter mounted at the exit of the exhaust manifold in a close-coupled location on the test vehicle. Catalyzed monolith volumes of both 1.0 L (50% of engine swept volume) and 0.5 L (25% of engine swept volume) were evaluated on the test vehicle using both drive cycles.

Figures 5 and 6 summarize NMHCs (non-methane hydrocarbons) and NO*x* average emission performance, respectively, of aged converters evaluated on the test vehicle during FTP evaluations as a function of substrate type (cell density and wall thickness). Emissions data are included in these figures for both the 1.0 L catalyzed volume and the 0.5 L catalyzed volume converters, appropriately weighted for each of the three phases of the FTP driving cycle [cold-start (Bag 1), hot transient (Bag 2), and hot-start (Bag

Table 3. Ceramic substrate properties for standard and high cell density products.

Cell density (cpsi)	400	400	600	600	900
Wall thickness (mils)	6.5	4.5	4.5	3.5	2.5
Open frontal area (%)	75.7	82.8	80.0	83.6	85.6
Geometric surface area (m^2/l)	2.74	2.87	3.45	3.53	4.37
Bulk density (g/l)	401	279	324	267	267

(Based on Hughes and Witte, 2002. Copyright © 2002 SAE International. Reprinted with permission.)

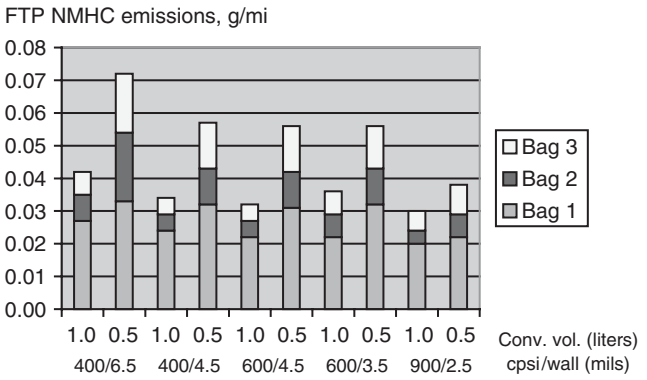


Figure 5. NMHC FTP emissions for substrates with varying cell density and wall thickness. (See Hughes and Witte (2002) for details.)

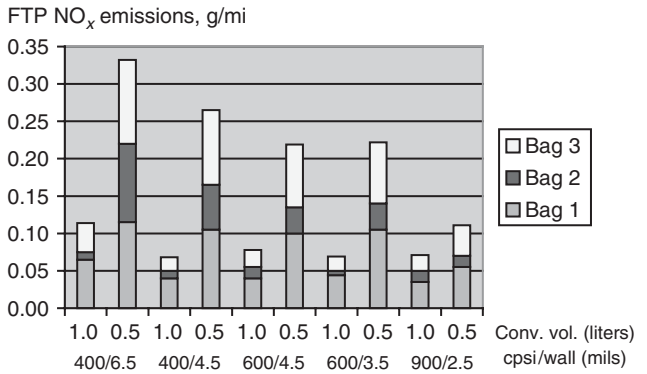


Figure 6. NO*x* FTP emissions for substrates with varying cell density and wall thickness. (See Hughes and Witte (2002) for details.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3)]. These data clearly show the significant decrease in both NMHC and NO*x* emissions that result from the use of high cell density/thin wall substrates relative to the base case 400 cpsi/6.5 mil wall standard. Lower emissions of NMHC and NO*x* are evident in each phase (or "bag") of the FTP drive cycle: cold-start phase (phase 1 or "bag" 1), warmed-up transient phase (phase 2 or "bag" 2), and warm-start phase (phase 3 or "bag" 3). These reduced tailpipe emissions stem from the higher GSA of these advanced substrates, the smaller hydraulic diameter of each coated channel, and the lower thermal mass of the higher cell density substrates. Thermal mass is proportional to the substrate bulk density values shown in Table 3 [thermal mass = (substrate bulk density) × (substrate volume) × (substrate-mass-specific heat capacity)].

Emission results presented by Aoki *et al*. (2002) also detail the performance of advanced high cell density ceramic substrates with respect to FTP NMHC emissions on a late model, four-cylinder test vehicle. This study evaluated engine-aged converters with equivalent volume (substrate dimensions of 106 mm diameter × 114 mm long) and equivalent precious metal loading (150 g/ft^3 advanced trimetal [Pt/Pd/Rh] catalyst) on a vehicle with a 2.3 L engine (vehicle calibrated for ULEV I performance with lean start strategy; converter inlet approximately 1.1 m downstream of the engine's exhaust valves). Converters were aged for 50 h using an accelerated engine aging protocol with a maximum catalyst temperature of 850°C. Aged converters with substrate cell densities from 300 to 1200 cpsi and varying wall thickness were evaluated on the test vehicle using the FTP drive cycle. Figure 7 compares the NMHC FTP emissions measured on the test vehicle for the various aged converters versus the specific GSA of the substrates evaluated by this program. In this figure, each ceramic substrate design is denoted by its cell density (cpsi) and wall thickness in mils (e.g., 600/3.5). The results show a strong relationship between NMHC emissions and substrate GSA with higher substrate GSA contributing to lower NMHC emissions in the FTP test cycle, a result consistent with the results shown in Figures 5 and 6. The results from Aoki *et al*. also show a relatively large benefit in emission performance for 600 cpsi substrates relative to 300 and 400 cpsi substrate designs. Smaller relative emission benefits were achieved in this study for additional increases in cell density beyond 600 cpsi (e.g., 900 and 1200 cpsi substrate designs). The relative magnitudes of the emission benefits shown in Figures 5–7 for different substrate cell density and wall thickness options will be impacted by the vehicle application environment including the number and location of catalysts in the exhaust system and the engine calibration strategy employed on the test vehicle. These optimization parameters again emphasize the overall systems design philosophy that needs to be employed to achieve the required emission performance with the most cost-effective system design.

Results presented by Marsh *et al*. (2001) show similar trends in reducing HC and NO*x* emissions with advanced high cell density metal substrates during FTP emission tests utilizing a 2.4 L, five-cylinder test vehicle. In this study, cell densities as high as 1600 cpsi were evaluated for their impacts on emissions performance. Physical properties for the metallic substrates evaluated in this study are summarized in Table 4, including values of the flow

Table 4. Metallic substrate properties for high cell density products.

Cell density (cpsi)	600	800	1000	1200	1600
Wall thickness (mils)	30	25	20	20	20
Hydraulic diameter (mm)	0.85	0.75	0.66	0.60	0.52
Geometric surface area (m^2/l)	3.77	4.32	4.88	5.36	6.08
Thermal mass (J/K)	689	681	641	680	750

(From Marsh *et al*. (2001. Copyright © 2001 SAE International. Reprinted with permission.)

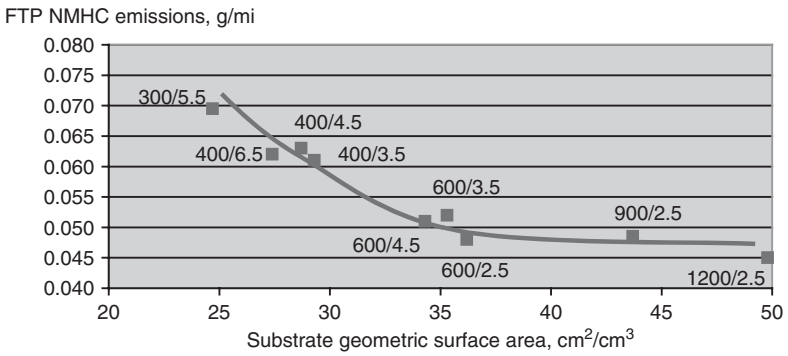


Figure 7. NMHC FTP emissions versus substrate geometric surface area. (See Aoki *et al*. (2002) for details.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

channel hydraulic diameter. As in the study by Hughes and Witte (2002), converters were evaluated on the test vehicle using the same volumetric precious metal and total catalyst loading of an advanced TWC on each metallic substrate. Converters were located near the exit of the exhaust manifold on the five-cylinder engine. FTP HC and NO*x* emissions reported by Marsh *et al*. (2001) for these various high cell density metallic substrate-based catalysts are detailed in Figures 8 and 9, respectively. Similar to the results reported by Hughes and Witte (2002), FTP HC and NO*x* emissions were reduced in this study by utilizing higher cell density, thinner wall metal substrates. Improvements in HC emissions were most strongly impacted by the combined increase in cell density with thinner walls between channels as this substrate design strategy lowers thermal mass and increases geometric area (e.g., moving from 600 cpsi/30 μm wall to 1000 cpsi/20 μm wall), critical properties for maximizing converter heat-up and mass transfer characteristics during the HC intensive cold-start period. Further increases in cell density at constant wall thickness (e.g., 1000, 1200, and 1600 cpsi with 20 μm wall thickness) equate to higher thermal mass substrates with poorer heat-up characteristics during the cold-start phase of the FTP test cycle. The additional geometric area of these highest cell density designs helped to compensate for the higher thermal mass but no net benefit in cold-start HC performance was realized. NO*x* benefits were shown in each case as cell densities increased, largely because of more effective contacting efficiency between the exhaust gas constituents and the active catalyst coating present on the walls of the substrate. Somewhat higher pressure drop of these substrates with increasing cell density may also have contributed to some reductions in engine-out NO*x* levels in certain driving modes because of increased levels of internal exhaust gas recirculation (EGR) within the engine's combustion chambers.

Results like those shown in Figures 5–9 and many other studies aimed at understanding the impacts of advanced substrate properties such as cell density, hydraulic diameter, and thermal mass have allowed researchers and

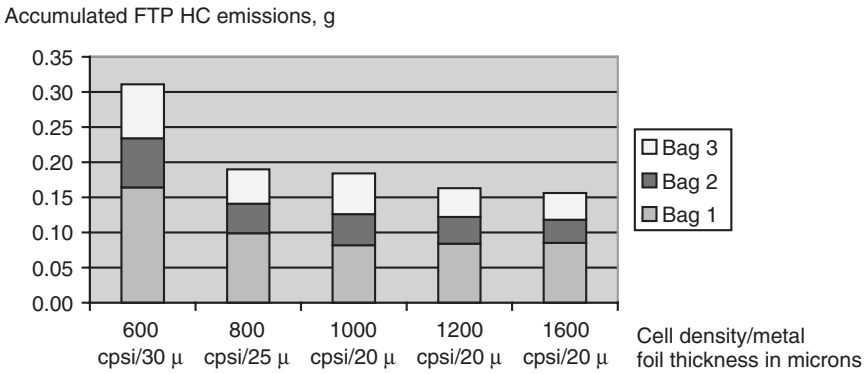


Figure 8. Accumulated FTP HC emissions for a three-way catalyst coated on high cell density metal substrates. (See Marsh *et al*. (2001) for details.)

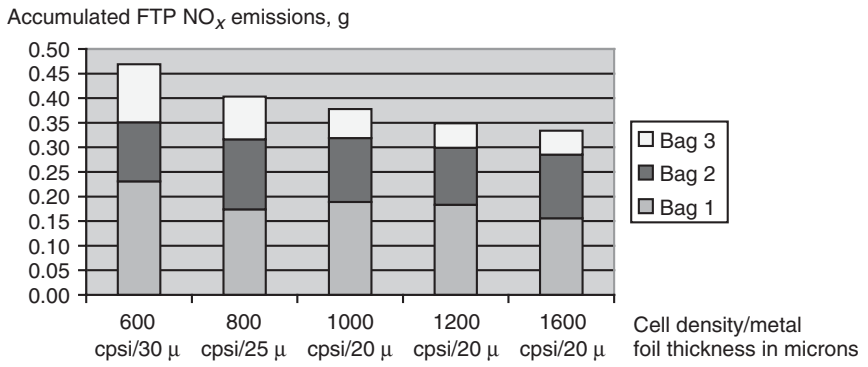


Figure 9. Accumulated FTP NO*x* emissions for a three-way catalyst coated on high cell density metal substrates. (See Marsh *et al*. (2001) for details.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

design engineers to develop sophisticated mathematical models that accurately predict the performance of catalytic converters during vehicle operation including performance during the FTP test protocol (Lafyatis *et al*., 2000; Umehara *et al*., 2000; Becker *et al*., 2001; Marsh *et al*., 2001; Aoki *et al*., 2002; Leonhard *et al*., 2002). These models generally include mathematical descriptions of the heat and mass transfer processes that occur within catalytic converters. Becker *et al*. (2001) used a modeling approach to predict the emission performance of a variety of substrate types and designs. In their work, they report that the catalytic performance of these substrates could be strongly correlated with key substrate physical properties: higher catalytic efficiency was proportional to substrate GSA and inversely proportional to bulk density and substrate channel hydraulic diameter. Large GSA in combination with small channel diameters provide good heat and mass transfer characteristics, whereas low substrate bulk density results in fast dynamic converter heat-up properties. In addition to cell density and wall thickness modifications, metal substrates have been developed and put into production that incorporate structural elements that help to promote turbulence within flow channels that promotes enhanced contacting efficiencies between the gas-phase reactants and the active catalyst components that are coated on the substrate channel surfaces.

The production of high cell density ceramic and metallic substrates is subject to the many quality system requirements of the auto industry. These advanced substrates are manufactured with precise specifications on all key fabrication parameters, resulting in only small variations in the key performance-related physical properties such as bulk density, cell density, and wall thickness. For example, ceramic monolith wall thickness typically varies by ± 0.5 mils or less for nominal wall thickness in the range 2–4 mil. Similarly, metal foil thickness in metal substrates varies by $\pm 0.2–0.3\,\mu m$ for foils in the range $20–50\,\mu m$. Cell densities in ceramic substrates are controlled by the precision die used in the extrusion process and process controls associated with the extrusion and firing operations. Cell densities in metal substrates are controlled by tight specifications on the process used to produce corrugated foils, as well as process controls on other key production operations. As an example, cell densities in metal substrates vary by $\pm 5\%$ for high cell density substrates ranging from 600 to 1600 cpsi. Modifications to traditional canning operations and mounting materials have also been developed for high cell density, thin wall ceramic substrates to ensure a mechanically robust, durable converter package. Similarly, high cell density, thin wall metal substrates have reengineered brazing strategies and matrix/mantle connection methods to maintain required mechanical durability for all light-duty vehicle applications.

2.3 Advanced three-way catalysts

Three-way catalysts have traditionally relied on highly dispersed precious metals (Pt, Pd, and Rh) supported on high surface area aluminum oxide with the addition of a variety of base metal oxide promoters and oxygen storage materials to provide the simultaneous HC and CO oxidation and NO*x* reduction behavior required in automotive emission control applications. Oxygen storage and release behavior of TWCs is an important functionality required to maintain acceptable performance during air/fuel perturbations that occur as a result of the closed-loop air/fuel feedback control algorithm associated with oxygen sensors. Cerium oxide-based materials contained in TWC formulations have been the primary source of this oxygen storage behavior. Catalyst performance criteria associated with meeting the low emission requirements of Tier 2 or LEV II applications include maintaining high conversion efficiencies for all three criteria pollutants during all phases of vehicle operation (e.g., start phases, accelerations, decelerations, and cruise conditions) for extended operational lifetimes (i.e., 120,000 mile durability). These demands for high conversion efficiencies and extended durability have evolved TWC formulations and design strategies significantly in recent 15 years.

The interest in cold-start performance discussed with close-coupled converters previously puts emphasis on TWC light-off characteristics, especially with respect to HCs (Light-off generally refers to the catalyst temperature required to achieve significant conversion activity with respect to the pollutants of interest.). Pd-based TWCs (e.g., Pd-only, Pd/Rh, or Pt/Pd/Rh trimetallic catalyst formulations) became the preferred choice for close-coupled applications because of the inherent good HC light-off performance of Pd relative to Pt or Rh (Waltner *et al*., 1998; Nagashima *et al*., 2000; Williamson *et al*., 2000; Williamson *et al*., 2001; Ohmoto *et al*., 2002; Truex *et al*., 2002). Close-coupled applications also place a premium on catalyst thermal stability/durability as these close-coupled converters expose catalyst materials to significantly higher operating temperatures than temperatures associated with converters located in cooler, underfloor locations. This thermal durability requirement also placed attention on Pd-based close-coupled TWCs because of Pd's superior thermal stability compared to other precious metals. The thermal stability of other catalytic materials used in TWC formulations was equally important in meeting the demands of close-coupled catalysts. This need resulted in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the concerted development of new catalyst materials such as stabilized aluminas (a support material for precious metals), stabilized cerias and zirconias (both are catalytic promoters and oxygen storage materials), and the development of more stable precious metal impregnation strategies that helped to push maximum catalyst operating temperatures from 800°C to over 1000°C over recent 15 years. The longer durability requirements of Tier 2 and LEV II emission regulations, as well as the inclusion of heavier light-duty trucks with their relatively higher exhaust temperatures compared to passenger cars, contributed to this focus on improving TWC thermal stability.

TWCs with high conversion efficiencies and extended durability with respect to NO*x* (HCs and CO, as well) are additional important criteria of Tier 2 and LEV II emission systems. The need for high performance and extended durability catalysts influenced all aspects of catalyst design and the selection of materials used as supports, promoters, and oxygen storage materials. Enhancements in oxygen storage materials, in particular, have been a key element in pushing catalyst performance in advanced TWC formulations. New families of ceria-zirconia materials have been developed that provide higher capacities of thermally stable oxygen storage and release functionalities to TWCs (references include: Hirasawa *et al*., 2009; Rohart *et al*., 2007; Kanazawa *et al*., 2003; Truex *et al*., 2002; Schmidt *et al*., 2001; Williamson *et al*., 2001; Williamson *et al*., 2000). Synergies between these new ceria-zirconia materials and the catalytically active precious metals have led to improvements in intrinsic catalyst light-off characteristics, broader three-way operating windows with respect to the simultaneous oxidation and reduction reactions as a function of inlet air/fuel ratio, and more highly dispersed and thermally stable precious metal activity. Figure 10 provides an example of TWC performance improvements with respect to NO*x* emissions stemming from the use of new materials such as advanced ceria-zirconia-based oxygen storage materials (Williamson *et al*., 2001).

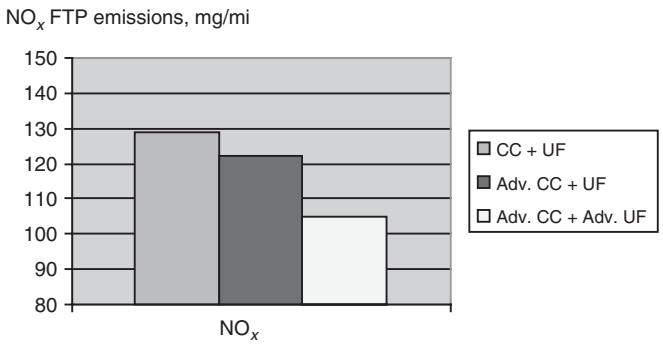


Figure 10. Impact of advanced catalyst formulations on NO*x* FTP emissions from a 5.3 L, V8 test vehicle equipped with close-coupled (50 g/ft^3 Pd-only)+ underfloor converters (30 g/ft^3 Pt/Rh = 3/1). (See Williamson *et al*. (2001) for details.)

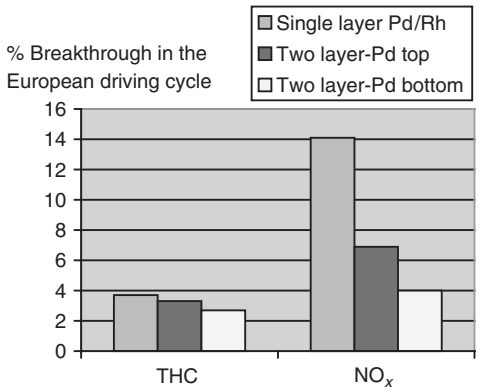


Figure 11. Impact of catalyst coating architecture (single layer vs double layer) on total hydrocarbon (THC) and NO*x* performance of a Pd/Rh three-way catalyst (100 g/ft^3 total precious metal loading with Pd/Rh = 5/1; 100 h aged catalyst on a 1.8 L, 4-cylinder test vehicle. (See Lindner *et al*. (1996) for details.)

New catalyst design strategies have also been developed to better tailor precious metal performance in advanced catalysts. A primary example of these tailored design strategies is the development of multilayer catalyst coating architectures in which preferred precious metal functionalities and oxygen storage performance can be segregated to maximize performance and minimize unwanted negative interactions that may result by co-mingling certain catalyst materials (Punke *et al*., 1995; Lindner *et al*., 1996; Nagashima *et al*., 2000, Schmidt *et al*., 2001; Williamson *et al*., 2001; Ohmoto *et al*., 2002; Schmidt *et al*., 2002; Ball *et al*., 2005; Aoki *et al*., 2009; Aoki *et al*., 2011). For example, undesirable alloying of precious metals due to high temperature sintering phenomenon can be minimized by segregating precious metals in unique chemical layers in a multilayer coating format. An example of the performance improvements achieved by new multilayer catalyst architectures is shown in Figure 11 (Lindner *et al*., 1996). These advanced catalyst materials and catalyst design strategies have cascaded through all TWC formulations (precious metal types including Pd-only, Pd/Rh, Pt/Rh, and trimetal) and applications (close-coupled and underfloor converters) to deliver cost-effective high performance and durable catalysts required to meet the needs of Tier 2 and LEV II light-duty applications.

The need to reduce cold-start HC emissions for Tier 2/LEV II and future Tier 3/LEV III applications has also resulted in the development of HC adsorber functions that can be added to three-way catalyst formulations or stand-alone HC adsorbers that are integrated into the exhaust

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

system of light-duty vehicles that utilize three-way catalysts. HC adsorbers have seen only limited commercial applications on a few selected models of SULEV or PZEV-compliant vehicles thus far. In these applications, materials that can adsorb typical exhaust HCs at ambient temperatures and then desorb the HCs at elevated temperatures are used to capture HC emissions during the cold-start phase of the emissions test cycle. When the HCs are desorbed later in the cycle, as the adsorber materials reaches its HC desorption temperature(s), the desorbed HCs can be oxidized by available three-way catalysts (assuming that the catalyst has reached a temperature that facilitates HC oxidation). Synthetic zeolites have been the HC adsorber material of choice for automotive applications. In some cases, the adsorber can utilize a mixture of zeolites in order to broaden the HC capturing efficiency of the adsorber relative to the range of HCs that are associated with automotive exhaust gas. Zeolite properties such as silica/alumina ratio, crystal structure, the presence and composition of exchanged cations, and zeolite pore size have been shown to impact the specific adsorption capacity and desorption properties of a zeolite relative to the HC species present in the exhaust (Mukai *et al*., 2004; Kanazawa and Sakurai, 2001; Goralski *et al*., 2000). A key design property for maximizing the impact of an HC adsorber on cold-start HC emissions is the overlap of the desorption temperature with the onset of catalyst HC oxidation activity over the whole regulated durability time frame of the emissions system (e.g., 120,000 or 150,000 miles).

A zeolite adsorber layer has been added to commercial three-way catalyst formulations that are displayed in underfloor converters to reduce cold-start HC emissions (Inoue and Mitsuishi, 2009; Lupescu *et al*., 2009; Oguma *et al*., 2003; Ballinger and Andersen, 2002). In this configuration, the ceramic or metallic monolith is coated with successive layers of the zeolite-based adsorber material and the three-way catalyst formulation to form an integrated, multifunctional converter. In another commercial application (Inoue *et al*., 2000), the underfloor HC adsorber material is physically separated from the underfloor three-way catalyst and an exhaust valving arrangement is used to first direct the bulk of the cold exhaust through the adsorber-containing monolith. Once the close-coupled three-way converter has reached catalyst light-off temperatures, the exhaust valve is opened to allow flow through the underfloor three-way converter. As this converter warms-up, heat is transferred to the adsorber function (located in an outer annulus of the underfloor converter) and HCs are desorbed and directed into the underfloor converter for conversion via a catalyzed HC oxidation reaction.

To achieve the full performance benefits of these advanced catalyst/advanced substrate combinations for Tier 2 and LEV II applications, it has also been necessary to develop improved engine operating algorithms that more closely match inlet catalyst conditions with the optimal operating window of the catalyst in order to maximize catalyst efficiency for all three criteria pollutants. The discussion on close-coupled converters included the development of cold-start engine operating strategies that accelerate converter heat-up during the crucial cold-start process. Similarly with respect to NO*x* emissions, tighter air/fuel control strategies during all modes of vehicle operation (especially high NO*x* emission modes associated with vehicle accelerations and decelerations) have been developed to achieve the low NO*x* emission requirements of the Tier 2 and LEV II programs. Precise air/fuel control strategies balance the relative concentration of oxidants and reductants in the exhaust stream within the catalyst's preferred operating window under highly dynamic vehicle operations. Similarly, vehicle calibrators can make use of EGR strategies to minimize engine-out NO*x* levels during some vehicle operating modes and maximize emission system performance. These EGR calibration strategies may involve either internal EGR calibrations through changes in exhaust valve lift or timing characteristics or external EGR calibrations through changes in the duty cycle of an external EGR valve during certain portions of a given driving cycle. The interplay and optimization of engine controls and emission control technology is a necessary part of the overall systems approach and integration required in meeting Tier 2/LEV II low NOx emission goals on light-duty vehicles.

New high performance TWCs have also required the development of new precision substrate coating processes and equipment capable of producing and placing complex coating formulations on the interior walls of ceramic and metallic substrates (both standard and advanced high cell density substrates) in high volume production. These advanced catalyst formulations have also been tailored to be compatible with advanced high cell density substrates (Lafyatis *et al*., 2000; Williamson *et al*., 2000; Domesle *et al*., 2001; Williamson *et al*., 2001; Schmidt *et al*., 2002; Hughes *et al*., 2003). For example, the volume-based catalyst loading on a high cell density substrate must be balanced to provide required performance and durability characteristics without adversely affecting the overall thermal mass (and the resulting dynamic heat-up) and pressure drop characteristics of the coated substrate. Similar to substrate manufacturing processes, catalyst manufacturing processes must also be operated within the rigorous automotive industry quality control requirements. Catalyst formulation and coating specifications on all materials (precious metals, support materials, oxygen storage materials, etc.) minimize physical and chemical variations between production parts and production lots of a given catalyst type.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.4 Gasoline particulate filters

The emerging interest and growing number of regulatory programs associated with reducing greenhouse gas emissions from motor vehicles (and improving vehicle fuel economy) has put significant attention on developing and commercializing gasoline direct injection (GDI) technology. This gasoline engine platform has combustion characteristics that parallel diesel engines that use direct fuel injection strategies and provides lower fuel consumption compared to port injected gasoline engines that have been the dominant stoichiometric, gasoline engine technology on light-duty vehicles for more than a decade. Compared to port fuel injected gasoline engines, the initial wave of commercial GDI engines introduced in the 2005–2011 time frame has been shown to produce higher levels of particulates under typical emission test cycle conditions (e.g., the US EPA FTP cycle). Future US and European light-duty emission standards (e.g., LEV III, Tier 3, Euro 6) are expected to lower PM emission requirements for some or all classes of gasoline vehicles. As a result of this interest in lower PM emission levels, emission control technology developers and auto manufacturers have begun to assess the feasibility of using gasoline particulate filters (GPFs) on future GDI engines. A number of recent references (Chan *et al.*, 2013; Eakle, Zahn, and Weber, 2010; Mikulic *et al*., 2010; Saito *et al*., 2011) are available that describe the development and performance of first-generation GPFs on GDI vehicles. The filter technology base is drawn from the large experience base with diesel particulate filters (DPFs). The filter substrates are high efficiency, wall-flow type filters made of ceramic materials that are essentially identical to wall-flow filters that have been used commercially on light-duty and heavy-duty diesel engines since 1999 (see Exhaust Emission Control Considerations for Diesel Engines on particulate emission control technologies for more information on wall-flow particulate filters). In these recent references, GPFs have been evaluated on GDI vehicles with and without a three-way catalyst coating present on the wall-flow ceramic filter. Similar to DPFs, the wall porosity of the GPFs can be adjusted to modify the pressure drop and catalyst coating capacity of the filter. Similar to DPFs, wall-flow GPFs are capable of large reductions (>90%) of particulate emissions across a very broad range in particle size. Vehicle and engine manufacturers are also evaluating improved fuel injection strategies and other improvements to engine combustion technologies as a way of reducing particulate emissions from GDI engines. The topic of particulate emission from GDI engines is expected to grow in interest in the coming decade as more work is done to understand the health impacts of ultrafine particulates.

REFERENCES

Anthony, J. and Kubsh, J. (2007) The potential for achieving low hydrocarbon and NOx exhaust emissions from large light-duty gasoline vehicles. SAE International Congress, Detroit. SAE paper no. 2007-01-1261, 2007.

Aoki, Y., Sakagami, S., Kawai, M., *et al*. (2011) Development of advanced zone-coated three-way catalysts. SAE International Congress, Detroit. SAE paper no. 2011-01-0296, 2011.

Aoki, Y., Yoshida, T., Tanabe, T., *et al*. (2009) Development of double-layered three-way catalysts. SAE International Congress, Detroit. SAE paper no. 2009-01-1081, 2009.

Aoki, Y., Miyairi, Y., Ichikawa, Y., and Abe, F. (2002) Product design and development of ultra-thin wall ceramic catalytic substrate. SAE International Congress, Detroit. SAE paper no. 2002-01-0350, 2002.

Ball, D., Clark, D., and Moser, D. (2011) Effects of fuel sulfur on FTP NOx from a PZEV 4 cylinder application. SAE International Congress, Detroit. SAE paper no. 2011-01-0300, 2011.

Ball, D., Nunan, J., Blosser, P., *et al.* (2005) Flexmetal catalyst technologies. SAE International Congress, Detroit. SAE paper no. 2005-01-1111, 2005.

Ballinger, T. and Andersen, P. (2002) Vehicle comparison of advanced three-way catalysts and hydrocarbon trap catalysts. SAE 2002 International Congress, Detroit. SAE paper no. 2002-01-0730.

Becker, R., Wilson, R., Brayer, M., *et al*. (2001) Prediction of catalytic performance during light-off phase with different wall thickness, cell density, and cell shape, SAE 2001 International Congress, Detroit. SAE paper no. 2001-01-0930.

Brueck, R., Mueller-Haas, K., Breuer, J., and Webb, C. (2002) Advanced performance of metallic converter systems demonstrated on a production V8 engine. SAE 2002 International Congress, Detroit. SAE paper no. 2002-01-0347.

Brueck, R., Kaiser, F., Konieczny, R., *et al*. (2001) Study of modern application strategies for catalytic aftertreatment demonstrated on a production V6 engine. SAE 2001 International Congress, Detroit. SAE paper no. 2001-01-0925.

Chan, T., Meloche, E., Kubsh, J., *et al.* (2013) Impact of ambient temperature on gaseous particle emissions from a direct injection gasoline vehicle and its implications on particle filtration. SAE 2013 International Congress, Detroit. SAE paper no. 2013-01-0527.

Domesle, R., Lindner, D., Mueller, W., *et al*. (2001) Application of advanced three-way catalyst technologies on high cell density ultra-thin wall ceramic substrates for future emission legislations. SAE 2001 International Congress, Detroit. SAE paper no. 2001-01-0924.

Eakle, S., Zahn, R., and Weber, P. (2010) Simultaneous reduction of PM, HC, CO, and NOx from a GDI engine. SAE 2010 International Congress, Detroit. SAE paper no. 2010-01-0365.

Ehmann, P., Rippert, N., Umehara, K., and Vogt, C. (1999) The development of a BMW catalyst concept for LEV/EU3 legislation for a 8 cylinder engine by using thin wall ceramic substrates. SAE 1999 International Congress, Detroit. SAE paper no. 1999-01-0767.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Goralski, C., Chanko, T., Lupescu, J., and Ganti, G. (2000) Experimental and modeling investigation of catalyzed hydrocarbon trap performance. SAE 2000 International Congress, Detroit. SAE paper no. 2000-01-0654.

Hirasawa, Y., Katoh, K., Yamada, T., and Kohara, A. (2009) Study on new characteristic CeO_2-ZrO_2 based material for advanced TWC. SAE 2009 International Congress, Detroit. SAE paper no. 2009-01-1078.

Holy, G., Brueck, R., and Hirth, P. (2000) Improved catalyst systems for SULEV legislation: first practical experience. SAE 2000 International Congress, Detroit. SAE paper no. 2000-01-0500.

Hughes, K., Schmitz, K., Radke, D., *et al.* (2003) Impact of ultra-thin wall catalyst substrates for TIER 2 emission standards. SAE 2003 International Congress, Detroit. SAE paper no. 2003-01-0658.

Hughes, K. and Witte, W. (2002) Ultra-thin wall substrates—trends for performance in FTP and US06 tests. SAE paper no. 2002-02-0356, SAE 2002 International Congress, Detroit.

Inoue, T., Kusada, M., Kanai, H., *et al*. (2000) Improvement of a highly efficient hybrid vehicle and integrating super low emissions. SAE 2000 Fall Fuels & Lubricants Meeting, Baltimore. SAE paper no. 2000-01-2930.

Inoue, K., and Mitsuishi, S. (2009) Development of atmospheric air-level emission vehicle technology for gasoline engines. SAE 2009 International Congress, Detroit. SAE paper no. 2009-01-1076.

Kanazawa, T., Suzuki, J., Takada, T., *et al;* (2003) Development of three-way catalyst using composite alumina-ceria-zirconia. SAE 2003 International Congress, Detroit. SAE paper no. 2003-01-0811.

Kanazawa, T. and Sakurai, K. (2001) Development of the automotive exhaust hydrocarbon adsorbent. SAE 2001 International Congress, Detroit. SAE paper no. 2001-01-0660.

Kidokoro, T., Hoshi, K., Hiraku, K. *et al.* (2003) Development of PZEV exhaust emission control system. SAE 2003 International Congress, Detroit. SAE paper no. 2003-01-0817.

Kishi, N., Kikuchi, S., Seki, Y., *et al.* (1998) Development of the high performance L4 engine ULEV system. SAE 1998 International Congress, Detroit. SAE paper no. 980415.

Lafyatis, D., Will, N., Martin, A., *et al.* (2000) Use of high cell density substrates and high technology catalysts to significantly reduce vehicle emissions. SAE 2000 International Congress, Detroit. SAE paper no. 2000-01-0502.

Laurell, M., Dahlgren, J., and Vaisanen, J. (2007) The Volvo S40/V50 PZEV MY 2007 with an optimized 2.4l engine. SAE International Congress, Detroit. SAE paper no. 2007-01-1260, 2007.

Leonhard, T., Floerchinger, P., Degen, A., *et al*. (2002) Effect of cell geometry on emissions performance of ceramic catalytic converters. SAE 2002 International Congress, Detroit. SAE paper no. 2002-01-0354.

Lindner, D., Lox, E., van Yperen, R., *et al.* (1996) Reduction of exhaust gas emissions by using Pd-based three-way catalysts. SAE 1996 International Congress, Detroit. SAE paper no. 960802.

Lupescu, J., Chanko, T., Richert, J., and DeVries, J. (2009) Treatment of vehicle emissions from the combustion of E85 and gasoline with catalyzed hydrocarbon traps. SAE 2009 International Congress, Detroit. SAE paper no. 2009-01-1080.

Marsh, P., Acke, F., Konieczny, R., *et al.* (2001) Application guideline to define a catalyst layout for maximum catalytic efficiency. SAE 2001 International Congress, Detroit. SAE paper no. 2001-01-0929.

Matsuzono, Y., Iwamoto, T., Narishige, T., *et al*. (2008) Advanced washcoat technology for PZEV application. SAE 2008 International Congress, Detroit. SAE paper no. 2008-01-0812.

Matsuzono, Y., Sakanushi, M., and Kitagawa, H. (2003) Development of a low precious-metal automotive perovskite catalytic system for LEV-II. SAE 2003 International Congress, Detroit. SAE paper no. 2003-01-0814.

Mikulic, I., Koelman, H., Majkowski, S., and Vosejpka, P. (2010) A study about particle filter application on a state-of-the-art homogeneous turbocharged 2L DI gasoline engine. 2010 FEV Aachen Colloquium, Aachen, Germany.

Moore, W., Richmond, R., Vaneman, G., and Dou, D. (1999) Evaluation of high cell density substrates for advanced catalytic converter emissions control. SAE 1999 Fall Fuels & Lubricants Meeting, Toronto, Canada. SAE paper no. 1999-01-3630.

Mueller-Haas, K., Brueck, R., Rieck, J., *et al.* (2003) FTP and US06 performance of advanced high cell density metallic substrates as a function of varying air/fuel modulation. SAE 2003 International Congress, Detroit. SAE paper no. 2003-01-0819.

Mukai, K., Kanesaka, H., Akama, H., and Ikeda, T. (2004) Adsorption and desorption characteristics of the adsorber to control the HC emission from a gasoline engine. SAE 2004 Powertrain & Fluid Systems Conference, Tampa. SAE paper no. 2004-01-2983.

Nagashima, K., Zhang, G., Hirota, T., and Muraki, H. (2000) The effect of aging temperature on catalyst performance of Pt/Rh and Pd/Rh TWCs. SAE 2000 Spring Fuels & Lubricants Meeting, Paris, France. SAE paper no. 2000-01-1954.

Nishizawa, K., Mitsuishi, S., Mori, K., and Yamamoto, S. (2001) Development of second generation gasoline PZEV technology. SAE 2001 International Congress, Detroit. SAE Paper No. 2001-01-1310.

Oguma, H., Koga, M., Momoshima, S., *et al.* (2003) Development of third generation gasoline PZEV technology. SAE 2003 International Congress, Detroit. SAE paper no. 2003-01-0816.

Ohmoto, H., Kobayashi, T., Ishikawa, K., and Yamada, T. (2002) Catalyst design for meeting stringent LEV-2 NOx regulation. SAE 2002 International Congress, Detroit. SAE paper no. 2002-01-0348.

Pfalzgraf, B., Fitzen, M., Siebler, J., and Erdmann, H. (2001) First ULEV turbo gasoline engine—the Audi 1.8 l, 125 kW 5-valve turbo. SAE 2001 International Congress, Detroit. SAE paper no. 2001-01-1350.

Punke, A., Dahle, U., Tauster, S., *et al.* (1995) Trimetallic three-way catalysts. SAE 1995 International Congress, Detroit. SAE paper no. 950255.

Rohart, E., Verdier, S., Takemori, H., and Suda, E. (2007) High OSC CeO_2/ZrO_2 mixed oxides used as preferred metal carriers for advanced catalysts. 2007 SAE International Congress, Detroit. SAE paper no. 2007-01-1057.

Schmidt, J., Franz, J., Merdes, N., *et al.* (2002) Utilization of advanced three-way catalyst formulations on ceramic ultra-thin

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

wall substrates for future legislation. SAE 2002 International Congress, Detroit. SAE paper no. 2002-01-0349.

Saito, C., Nakatani, T., Miyairi, Y., *et al.* (2011) New particulate filter concept to reduce particle number emissions. SAE 2011 International Congress, Detroit. SAE paper no. 2011-01-0814.

Schmidt, J., Busch, M., Waltner, A., *et al.* (2001) Utilization of advanced Pt/Rh TWC technologies for advanced gasoline applications with different cold start strategies. SAE 2001 International Congress, Detroit. SAE paper no. 2001-01-0927.

Takahashi, H., Ishizuka, Y., Tomita, M., Nishizawa, K., all of Nissan (1998) SAE 980674: engine-out and tailpipe emission reduction technologies of V-6 LEVs. SAE 1998 International Congress, Detroit.

Truex, T., Golden, S., Polli, A., *et al.* (2002) Advanced low platinum group metal three-way catalyst for LEV-II and ULEV-II compliance. SAE 2002 International Congress, Detroit. SAE paper no. 2002-01-0344.

Umehara, K., Makino, M., Brayer, M., *et al.* (2000) Prediction of catalytic performance for ultra-thin wall and high cell density substrates. SAE 2000 International Congress, Detroit. SAE paper no. 2000-01-0494.

Waltner, A., Loose, G., Hirschmann, A., *et al.* (1998) Development of close-coupled catalyst systems for European driving conditions. SAE 1998 International Congress, Detroit. SAE paper no. 980663.

Webb, C., Bykowski, B., Weber, P., and McKinnon, D. (1999) Using advanced emission control systems to demonstrate LEV II ULEV on light-duty gasoline vehicles. SAE 1999 International Congress, Detroit. SAE paper no. 1999-01-0774.

Williamson, B., Richmond, R., Nunan, J., *et al.* (2001) Palladium and platinum/rhodium dual-catalyst NLEV and Tier IIa close-coupled emission solutions. SAE 2001 International Congress, Detroit. SAE paper no. 2001-01-0923.

Williamson, B., Ball, D., Linden, D., *et al.* (2000) Palladium and platinum/rhodium dual-catalyst emission solutions for close-coupled or underfloor applications. SAE 2000 International Congress, Detroit. SAE paper no. 2000-01-0860.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto146
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Spark Ignition Combustion

Ulrich Spicher

Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

1 Basics and Major Features 1
2 Ignition and Combustion 5
3 SI and Advanced DISI Combustion Systems 8
4 Knocking Combustion 14
5 Pollutant Formation 20
References 21

1 BASICS AND MAJOR FEATURES

Spark ignition (SI) engines for passenger cars in the market today work exclusively according to the four-stroke principle. The four strokes of one engine cycle inside the combustion chamber of the SI engine take place during two crankshaft revolutions and are shown in Figure 1 (Bosch, 2003).

During the first stroke (intake stroke), the intake valves, directly controlled by the intake camshaft profile, are open. Owing to the downward motion of the piston, the air–fuel mixture (external mixture preparation) is drawn into the cylinder through the valves. The second stroke (compression stroke) is characterized by the upward motion of the piston and the compression of the unburned gas mixture of air and fuel. During this second stroke and also during the third stroke (expansion stroke), the intake valves are closed. At the end of the compression stroke, the premixed homogeneous air–fuel mixture is ignited by the spark plug. The combustion process propagates outward from the spark plug and combustion continues until the entire air–fuel mixture is consumed during the expansion stroke, which is characterized by the downward motion of the piston. Owing to the combustion, both the gas temperature and the gas pressure inside the combustion chamber (cylinder) increase significantly, which drives the piston down and imparts a downward force on the crankshaft via the connecting rod. At the crankshaft, this force is converted into engine torque. After the end of the combustion, the exhaust gas expands until the exhaust valves open at the end of the expansion stroke; this is controlled by the exhaust camshaft profile. During the fourth stroke (exhaust stroke), the exhaust valves are open and the piston moves upward, pushing the exhaust gas out of the combustion chamber via the open exhaust valves into the exhaust system. The exhaust gas can reach temperatures of higher than 1000°C at full-load operation.

SI engines are traditionally characterized by a premixed and nearly homogeneous mixture of fuel and air, which is ignited by a spark plug located in the cylinder head. The homogeneous mixture can be prepared by one of two different methods of fuel delivery, either via external mixture preparation in the intake system or via internal mixture formation with fuel injection directly into the cylinder.

1.1 External mixture formation

External mixture formation is employed in the majority of SI engines. A more or less homogeneous mixture consisting of air and fuel vapor is generated before it enters the cylinder. Such systems typically make use of low pressure fuel injection into the intake port close to the intake valves, as shown in Figure 2.

The main advantage of this so-called port fuel injection (PFI) is the longer duration available for vaporization and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

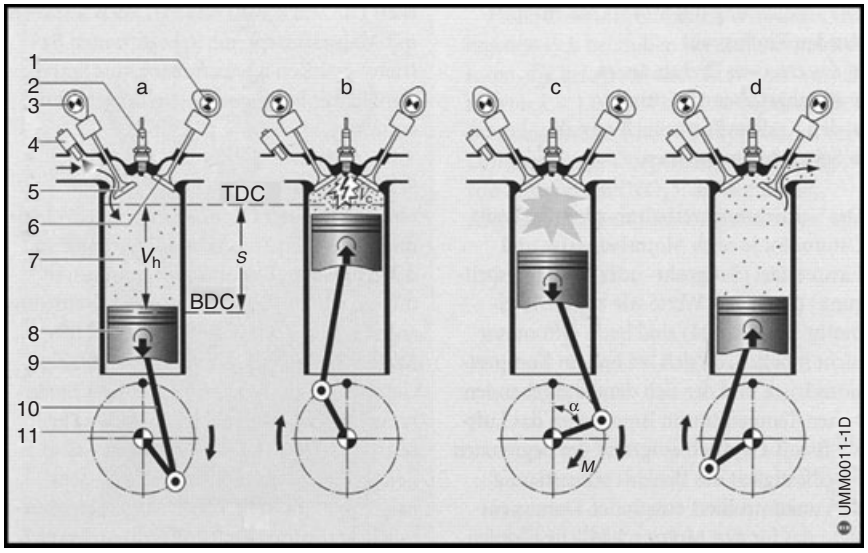


Figure 1. Four-stroke principle of SI engines with external mixture preparation. (Reproduced from Bosch, 2003. With kind permission of Springer Science+Business Media.)

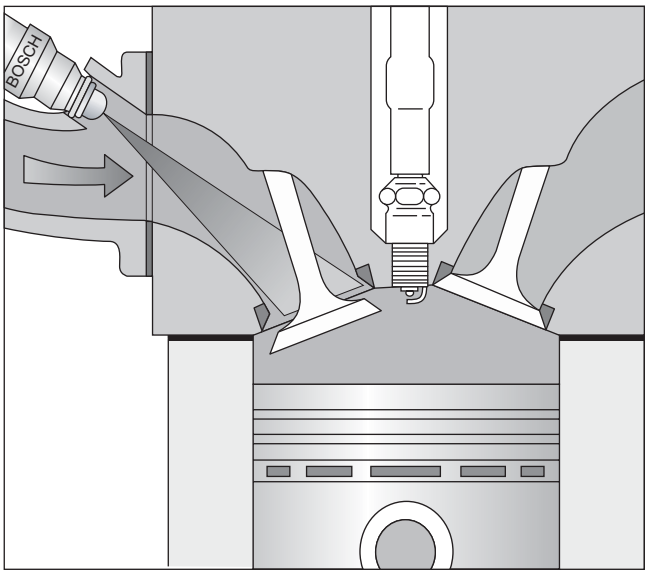


Figure 2. Port fuel injection (PFI) into the intake manifold close to the intake valve.

mixing of the fuel with the intake air during both the intake and the compression strokes until the ignition takes place. The mixture homogeneity is normally sufficient for ignition and combustion. One of the main difficulties is the condensation of fuel on the intake manifold walls during cold start operation, which can lead to an insufficient mixture process and incomplete combustion. This results in high levels of unburned hydrocarbons (UHC) and carbon monoxide (CO) emissions in the exhaust. To avoid these problems, intake manifold injection and special injection strategies (multiple injections) have been developed. The fuel is targeted such that it is injected directly behind the intake valve and partially toward the open intake valve and into the cylinder. In this case, time is sufficient to homogenize the mixture during intake and compression.

1.2 Internal mixture formation

Since Mitsubishi presented the first direct injection spark ignition (DISI) engine with stratified operation in 1995, the development of these engines has made significant progress. Owing to the direct injection of the fuel into the combustion chamber, the mixture formation occurs in the cylinder of the engine. To achieve this direct fuel injection, two different arrangements of the injector and the spark plug have been introduced in the development of DISI engines: the wide spacing arrangement and the narrow spacing arrangement (Figure 3). These arrangements are also often referred to as the *wall-guided and spray-guided arrangements*, respectively. They both can be used to create either homogeneous or stratified mixtures in the cylinder.

With a wide spacing arrangement, the injector is typically located between the intake valves. The fuel is injected downward toward the piston and is then directed by the shape of the piston and/or by the in-cylinder flow to the centrally mounted spark plug. As a result of this relatively long mixture formation process, an optimal mixture stratification cannot be achieved. In addition, wall wetting on the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

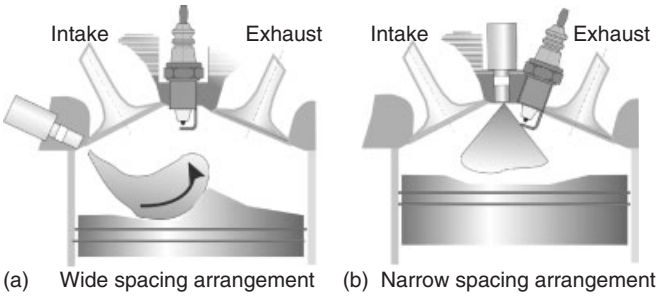


Figure 3. (a,b) DISI with different arrangement of spark plug and injector.

piston surface occurs, which leads to significantly increased UHC emissions.

In concepts with a narrow spacing between the injector and the spark plug, both the injector and the spark plug are located centrally in the cylinder head between the intake valves and the exhaust valves. A primary reason for this configuration is to realize a stratification of the air–fuel mixture during part-load operation of the engine enabling higher fuel efficiency due to overall lean air-fuel mixture. In this case, an extremely rapid vaporization of the liquid fuel is necessary to create an ignitable mixture that reaches the spark plug in a very short time, thus allowing minimum spark advance for best or minimum advance for best torque (MBT). Under optimal conditions, any contact between liquid fuel and the combustion chamber walls should be avoided. At full-load operation, the centrally located fuel injector proves to be very effective for producing a homogeneous charge. Because it is possible to generate a highly stratified air–fuel mixture at part load, an overall very lean mixture can be realized, too. This means that there is more than enough air in the cylinder to allow all of the fuel to burn completely.

Owing to the narrow spacing arrangement and close temporal coupling of fuel injection and spark advance, both mixture formation and stratification are expected to be incomplete at the time of ignition and during the early combustion phase. A further problem caused by this configuration is coking of the injector nozzle, which influences the spray characteristics in a way that can lead to misfiring. Manufacturing tolerances of nozzles, especially those involving nozzle hole geometry, directly affect the spray characteristics and therefore the combustion, too. Furthermore, deterioration of the durability of the spark plug has been observed due to high thermal stress. Another drawback of lean operation is the necessary use of more complicated and expensive NO_x aftertreatment systems in the exhaust to meet emission regulations. As a result of these difficulties, very few SI engines with direct injection and stratified lean mixture preparation are currently in production. However, to achieve a significant reduction in fuel consumption in future SI engines, it is expected that the share of DISI engines with lean combustion and mixture stratification will increase in the future. This requires further development of the injection system, including the fuel pump, the fuel rail, and the fuel injector, which will enable improved and extremely rapid mixture preparation near the end of the compression.

1.3 Operation modes with direct injection

In principle, direct injection in SI engines can be implemented using two different operating modes: homogeneous charge and stratified charge, shown in Figure 4.

In the higher load range, fuel is injected during the intake stroke, thus giving the injected fuel adequate time to mix with the air and form a homogeneous mixture before the start of ignition (Figure 4a). This operation mode shares some similarities with PFI in terms of injection timing and mixture homogeneity. Achieving wide open throttle (WOT) operation at lower load, which would require a lean homogeneous mixture in the combustion chamber, is not possible as the mixture present at the spark plug might be leaner than the lean ignition limit for lighter loads. During part-load operation, the advantages of WOT operation can be realized via operation with charge stratification. Injection is staged to ensure that an ignitable combustible mixture is always present near the spark plug at the time of ignition by creating a stratified air–fuel mixture in the combustion chamber through appropriate mixture preparation processes.

The advantages of operation with stratification of air–fuel mixture can be demonstrated by analyzing the losses during the gas exchange process (Figure 5).

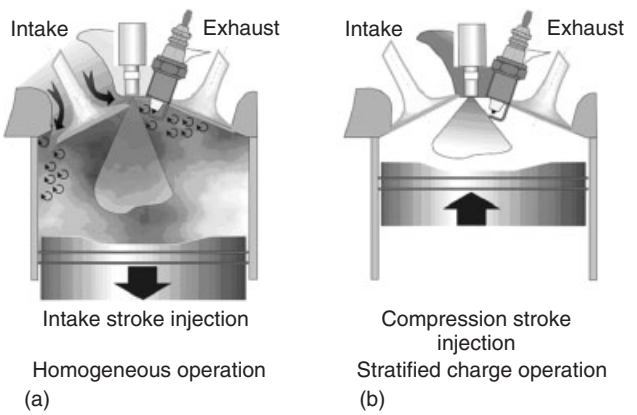


Figure 4. (a,b) Operation modes for DISI engines.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

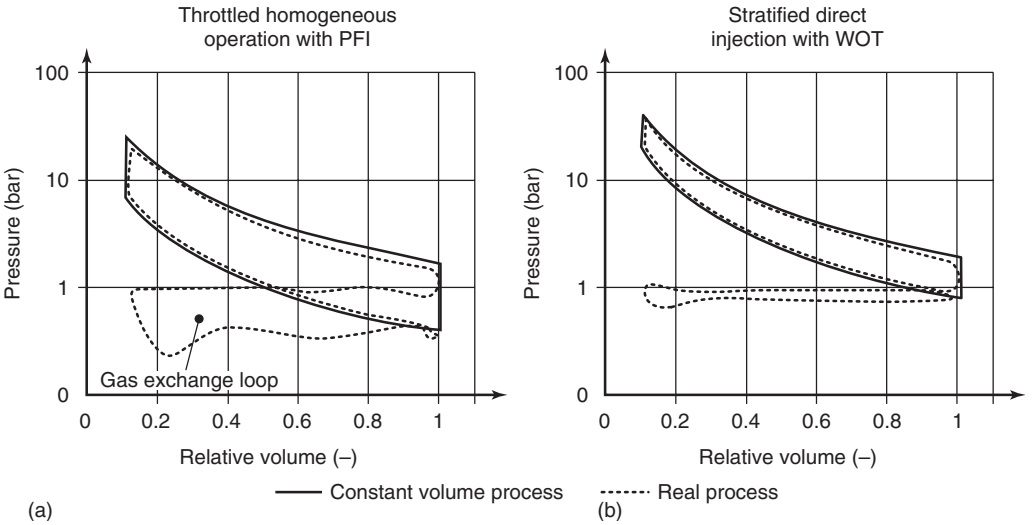


Figure 5. (a,b) Comparison of homogeneous throttled PFI and WOT stratified DISI working cycles at part load (equal engine speed and load). (Reproduced from Lückert *et al*., 2004. With permission from Aachen Colloquium Automobile and Engine Technology, Lückert *et al*., Daimler AG.)

WOT operation with DISI and a stratified charge effectively reduces the pumping losses normally associated with part-load operation of an SI engine. These pumping losses are represented by the area of the gas exchange loop in the $p-V$ diagram. With stratified operation and WOT, this area is significantly smaller than in the case of throttled operation with a homogeneous mixture and PFI. In addition, wall heat losses at stratified operation can be reduced during combustion if the stratified air–fuel mixture and eventual high temperature combustion products are insulated from the wall as much as possible by inducted fresh air and/or residual gas that remain in the cylinder. Overall, the reduced pumping losses will lead to improved engine thermal efficiency.

Stratified operation of a DISI engine is ideally implemented with a WOT. The load is then regulated solely by the quantity of injected fuel, as is the case in a diesel engine. During this process, the fuel is injected directly into the cylinder during the compression stroke shortly before ignition occurs. An ignitable mixture cloud forms in the center of the combustion chamber, surrounded by air that is not required for combustion. This stratification and the layer of insulating air reduce the heat losses through the combustion chamber walls that occur during subsequent combustion. Therefore, the engine is indeed operating with a globally lean mixture, whereby large air–fuel ratio gradients are present in the combustion chamber.

Figure 6 shows the range of the different operation modes for DISI engines in the engine operating map in terms of λ (where λ is defined as the ratio of actual air–fuel ratio to stoichiometric air–fuel ratio). In the low part-load range and the low speed range, the engine should be run with WOT and a stratified air–fuel mixture. This typically requires optimization of the in-cylinder flow and a specially adapted exhaust gas recirculation strategy to improve mixture formation and reduce the formation of nitrogen oxides (NO_x) during combustion. In the mid-load range (labeled $\lambda > 1$), a homogeneous lean air–fuel mixture, which is still within the ignition limit, should be used to avoid an abrupt change from homogeneous mixture formation to stratified mixture formation and vice versa. In the upper load and speed range, homogeneous mixture preparation with an air–fuel ratio of $\lambda = 1$ for operation with three-way catalyst or rich operation ($\lambda < 1$) for cooling at full load is necessary.

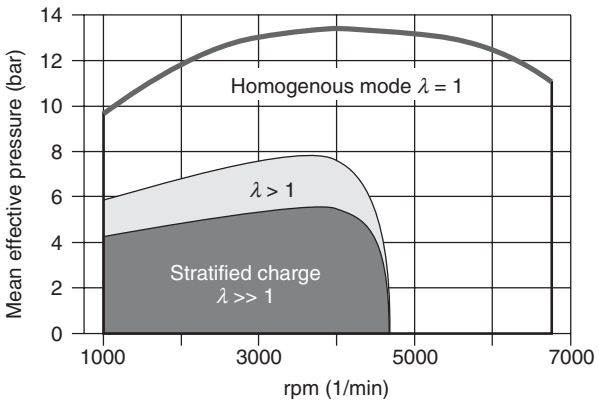


Figure 6. Operation modes in the engine map.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

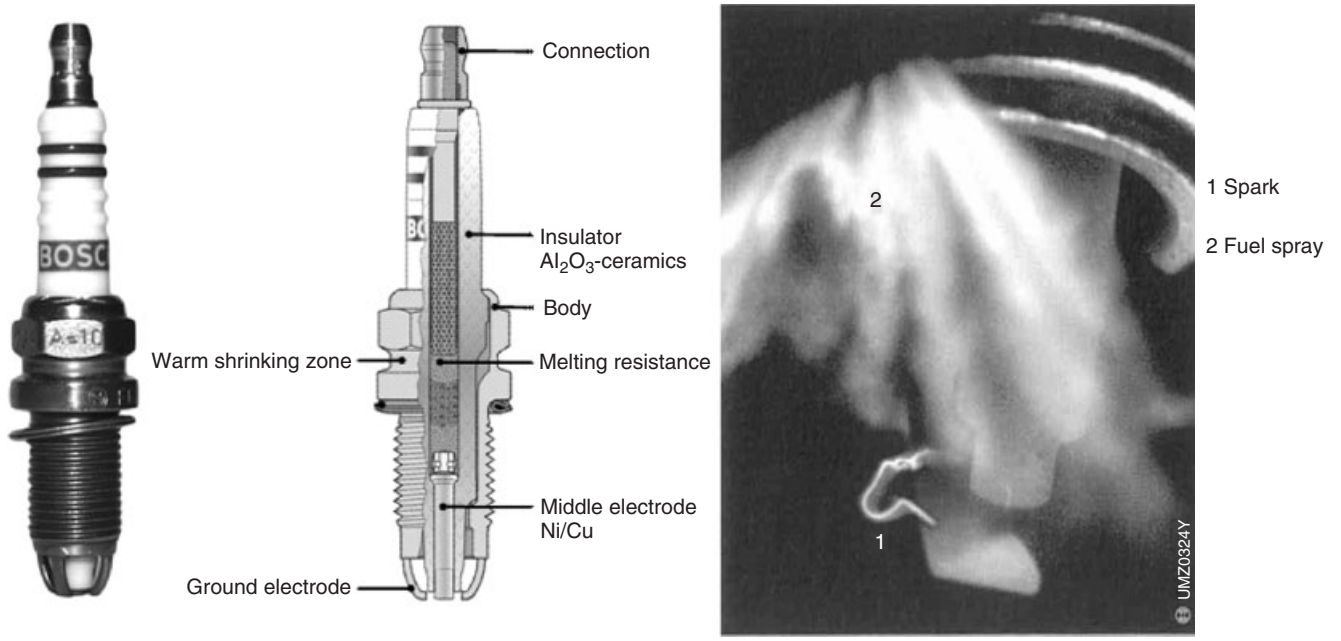


Figure 7. Spark plug and image of spark and liquid fuel spray for a DISI system with a narrow injector and spark plug arrangement. (Reproduced from Bosch, 2003. With kind permission of Springer Science+Business Media.)

2 IGNITION AND COMBUSTION

To initiate combustion, an electric ignition spark is provided by a spark plug. The air–fuel ratio λ around the spark plug must be within the ignition limits, between 0.6 (rich mixture) and 1.5 (lean mixture) for petrol (gasoline) fuel, in order for the mixture to be reliably ignited. The energy required for ignition is provided by an ignition system and supplied to the spark plug of the appropriate cylinder at the required spark advance. A spark plug and an ignition spark are shown in Figure 7 for a DISI system with a narrow injector and spark plug arrangement.

The spark is generated by an arc discharge between the middle and the ground electrodes. The spark plug is subjected to high thermal stress because of cyclical thermal loading with heating by combustion temperatures of up to 3000 K followed by cooling with fresh air–fuel mixture or even liquid fuel from the injection. In addition, the spark plug has to seal the combustion chamber at the maximum cylinder pressures, which can be more than 100 bar at full-load operation.

Ignition timing is an important variable for controlling the combustion of SI engines. It influences engine performance, fuel consumption, and exhaust gas emissions. It takes up to approximately 2 ms from the time of ignition until the time of maximum combustion rate. Therefore, spark advance must be set according to engine speed, engine load, and air–fuel ratio for the current operating conditions. For high combustion efficiency, it is necessary that approximately 50% of the fuel is consumed shortly after top dead center (TDC). In this way, an approximation to the ideal constant volume combustion is achievable.

Independent of the combustion system, which means independent of both fuel injection and mixture formation, spark advance is ideally adjusted to its minimum for best torque (MBT). However, at full load, SI engines are normally restricted in power output because of the occurrence of knocking. Therefore, MBT spark advance cannot safely be achieved. Figure 8 shows the comparison

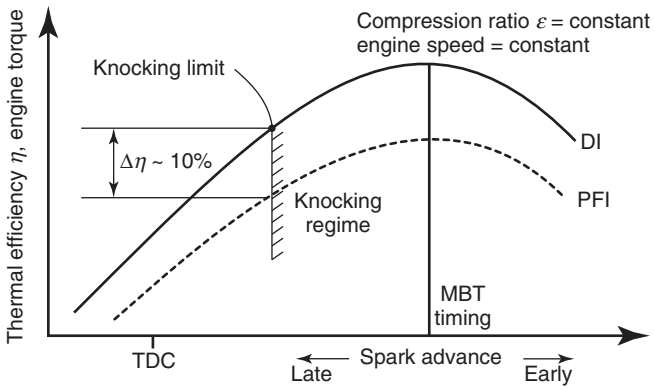


Figure 8. Thermal efficiency and engine torque versus spark advance.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

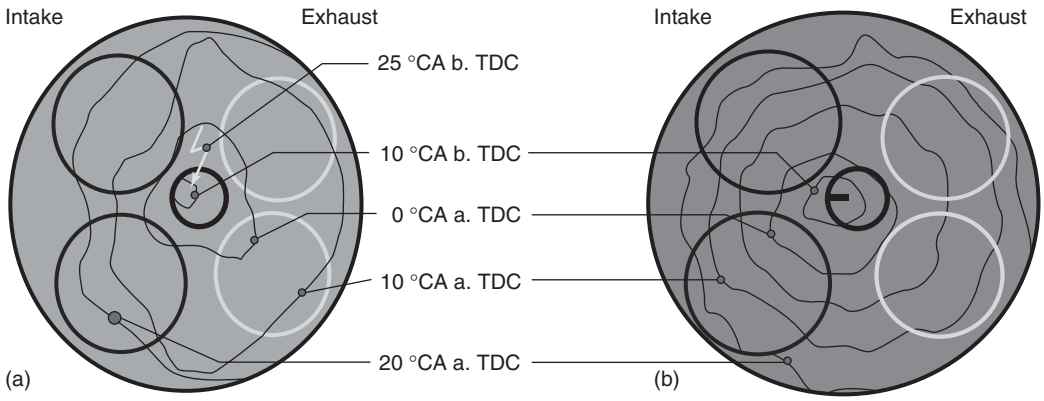


Figure 9. Flame propagation for nonknocking combustion [production engine (a); optimized configuration (b)]. Each contour represents the flame front location at the given °CA before (b) or after (a) TDC.

of engine torque as well as engine efficiency versus spark advance for an engine either with PFI or with direct injection (DISI). Engine speed is the same and constant for both injection strategies.

The flexibility of modern direct injection systems, combined with the charge cooling provided by the accompanying vaporization process inside the cylinder result in a higher volumetric efficiency and improved knocking behavior, which enables an improvement of power output of up to 10% (van Basshuysen *et al*., 2004).

After ignition has been initiated, propagation of the flame front (combustion zone) through the combustion chamber starts. If unrestricted by the combustion chamber walls (piston, cylinder liner, and cylinder head), the flame front would propagate approximately spherically. However, the flame front is distorted by the shape of the combustion chamber and the in-cylinder flow behavior. In addition, the temperature distribution in the combustion chamber as well as the temperatures of the combustion chamber walls and the homogeneity of the air–fuel mixture during combustion influence the contour of the flame front during its propagation. Figure 9a depicts the flame propagation of an individual combustion cycle, measured with the optical fiber technique (Spicher and Krebs, 1990) in one cylinder of a six-cylinder four-valve production engine at WOT operation. The engine speed was 5500 rpm with a spark advance of 25 crank angle degrees (CAD = °CA) before top dead center (BTDC), which represents the knock-limited spark timing.

The flame front (isolines) reaches the areas around the exhaust valves at TDC and has engulfed most of the combustion chamber at 20 CAD after TDC. By varying the in-cylinder flow, redesigning the piston crown and combustion chamber walls, and repositioning the spark plug, combustion can be optimized to produce uniform and approximately spherical flame propagation as shown in Figure 9b. As a result of these modifications, both torque and power output are increased by approximately 3% over the entire full load and speed range when compared to the production engine.

Variance in the processes involved with ignition, mixture composition, and flow turbulences cause statistical fluctuations in successive working cycles. These fluctuations essentially arise from irregularities during the first phase of the combustion sequence (inflammation phase) and invoke characteristic differences in the pressure traces of individual combustion cycles. These manifest themselves as variations in peak pressure values, crank angles of peak pressures, and indicated mean effective pressures (IMEP) (Figure 10). The variations in maximum pressure can be as high as 20% or more, depending on the operating point (load, engine speed, and air–fuel ratio). In addition, the cycle-to-cycle variations can influence the amount of UHC emissions and nitrogen oxide emissions as well as fluctuations in

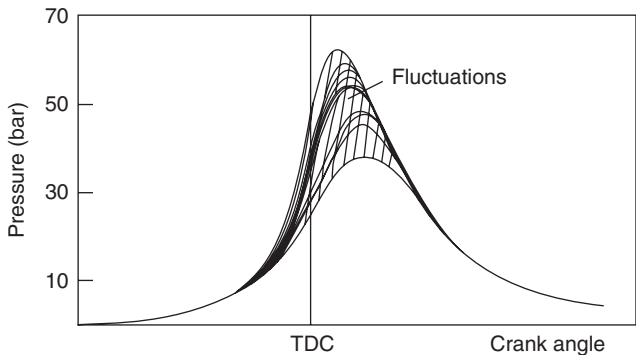


Figure 10. Cyclic variations of combustion as indicated by cylinder pressure traces.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

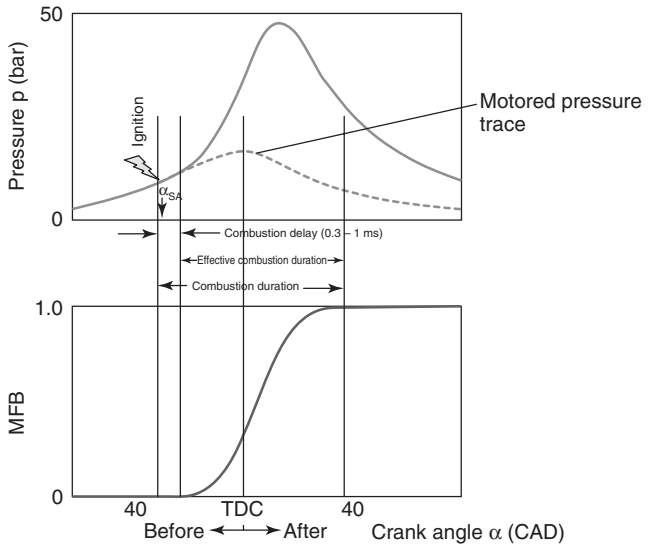


Figure 11. Pressure trace and mass fraction burned (MFB) function.

crankshaft rotational speed, which may be apparent as irregular engine operation.

Combustion behavior can be described by the mass fraction burned (MFB) function, which is calculated from the measured pressure trace using a thermodynamic two-zone model as well as an appropriate heat transfer model (see Pressure and Heat Release Analysis and Zero- and One-Dimensional Methodologies and Tools for related thermodynamic model discussions). The MFB curve has a value of between 0 and 1 (Figure 11).

After ignition, it normally takes a certain time until a measurable difference between the motored (without combustion) and the fired cylinder pressure traces is observed. This time duration of 0.3–1 ms is defined as "combustion delay" and corresponds to an MFB of approximately 2–5%. The total combustion duration from ignition time (MFB = 0) to the end of combustion (MFB = 1) is defined as "combustion duration." It is well known from experience in engine research and development that with homogeneous mixture formation, the point of 50% MFB (MFB50) should be in the range between 6 and 10 CAD after top dead center (ATDC) for maximum thermal efficiency. This means spark advance α_{SA} should be adjusted to achieve such combustion behavior.

A comparison between homogeneous operation with PFI and stratified operation is shown in Figure 12 for an engine speed of 2000 rpm and an IMEP = 3 bar. It is apparent from the p–V diagram on the left that the cylinder pressure at the end of compression is significantly higher for WOT, stratified operation; this is due to the higher pressure at the start of compression. The location of 50% MFB is at its thermodynamic optimum of eight CAD ATDC for the homogeneous operating point, as can be seen in Figure 12b.

The MFB curve for stratified operation indicates a rapid increase in the initial burning rate and a very slow burnout; this results from the mixture stratification. Near the spark plug, air–fuel ratios are rich or stoichiometric, which results in not only high flame speeds and thus rapid rates of combustion but also potentially high particulate emissions. However, the outermost portions of the air–fuel mixture have had a longer time to mix with the combustion air and become lean. These leaner mixture compositions burn much more slowly and are responsible for the long combustion duration observed for the stratified case. These longer combustion durations represent larger deviations from the thermodynamic optimal constant volume process and can lead to increased UHC emissions. To minimize

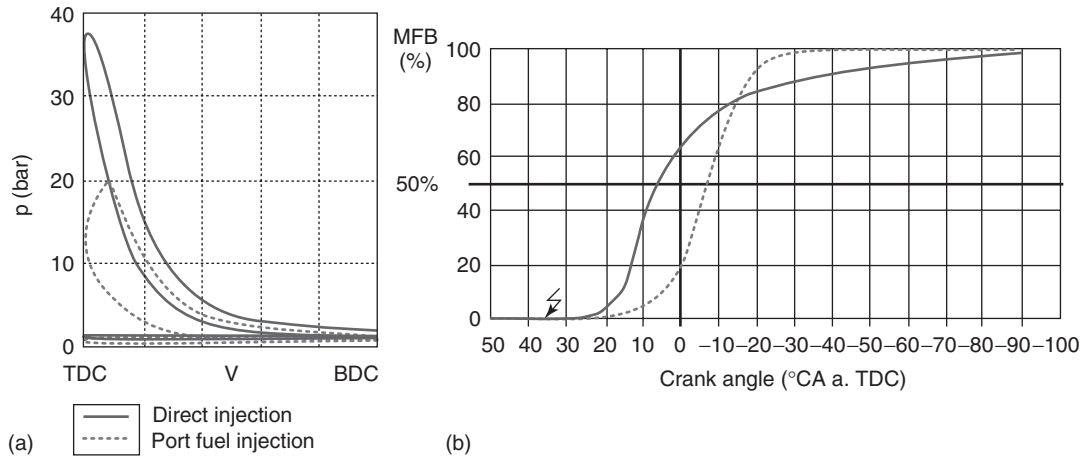


Figure 12. (a,b) Pressure–volume (p–V) diagrams and MFB functions for PFI and DI.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

these pollutant emissions, both the fuel injection and the SI are advanced to ensure that the combustion is as complete as possible at the time of exhaust valve opening. As a result of this early combustion phasing, the point of 50% MFB occurs much earlier than its thermodynamic optimum. The goal of current and future DISI research and development is the mitigation of these problems through optimized mixture formation and stratified combustion processes.

3 SI AND ADVANCED DISI COMBUSTION SYSTEMS

Until the late 1970s, carburetors were the most widely used means of mixture formation. Increasingly, stringent demands for accurate air–fuel ratios led to the introduction and dominance of PFI systems in recent decades. However, with the introduction of the first gasoline direct injection (GDI) engine with stratified operation in 1995, together with the increasing emphasis on reducing fuel consumption, an increasing number of systems with direct injection can be found on the market.

3.1 SI port fuel injection

In combination with catalytic converter technology for aftertreatment of exhaust gases, the use of PFI has displaced carburetors since the mid-1980s (and since the 1970s in California). This trend was driven by the advantages of injection into the intake port in terms of fuel economy, performance, and reduced emissions. This is due to the fact that PFI allows for very accurate metering of fuel to match the range of operating conditions of SI engines.

In SI engines with external mixture formation, the fuel is injected either into the intake manifold (single injection) or into the intake port inside the cylinder head (multipoint injection) (Figure 13). Depending on the induced airflow and intake air mass, detected by an airflow meter in the intake manifold, the proper amount of fuel is injected. In the case of multipoint injection, the total fuel amount is divided among the cylinders. The injectors are responsible for the injection into the port of each cylinder; they inject the fuel over certain injection durations. Both injection timing and injection duration depend on the engine operating conditions and the flow of intake air. The data for injection, ignition, and all other necessary information for engine operation are stored in the engine control unit (ECU), which is responsible for adapting engine operating parameters to fit the desired operating conditions.

3.1.1 Variable valve train (VVT)

One effective method to reduce pumping losses and control the load in an SI engine is to vary the intake valve timing in combination with variable valve lifts. An example of a real world implementation of this concept is shown in Figure 14. On early closing of the intake valve during the intake stroke, the piston continues its downward motion and expands the cylinder contents. In this way, the intake throttle can be completely opened and the intake valve

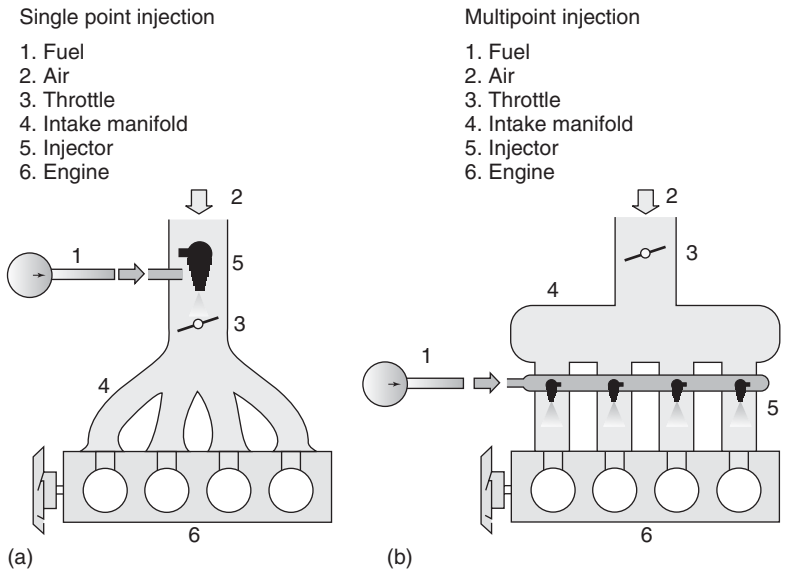


Figure 13. (a,b) Single point (intake manifold) and multipoint injection (intake port). (Reproduced from Bosch, 2003. With kind permission of Springer Science+Business Media.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

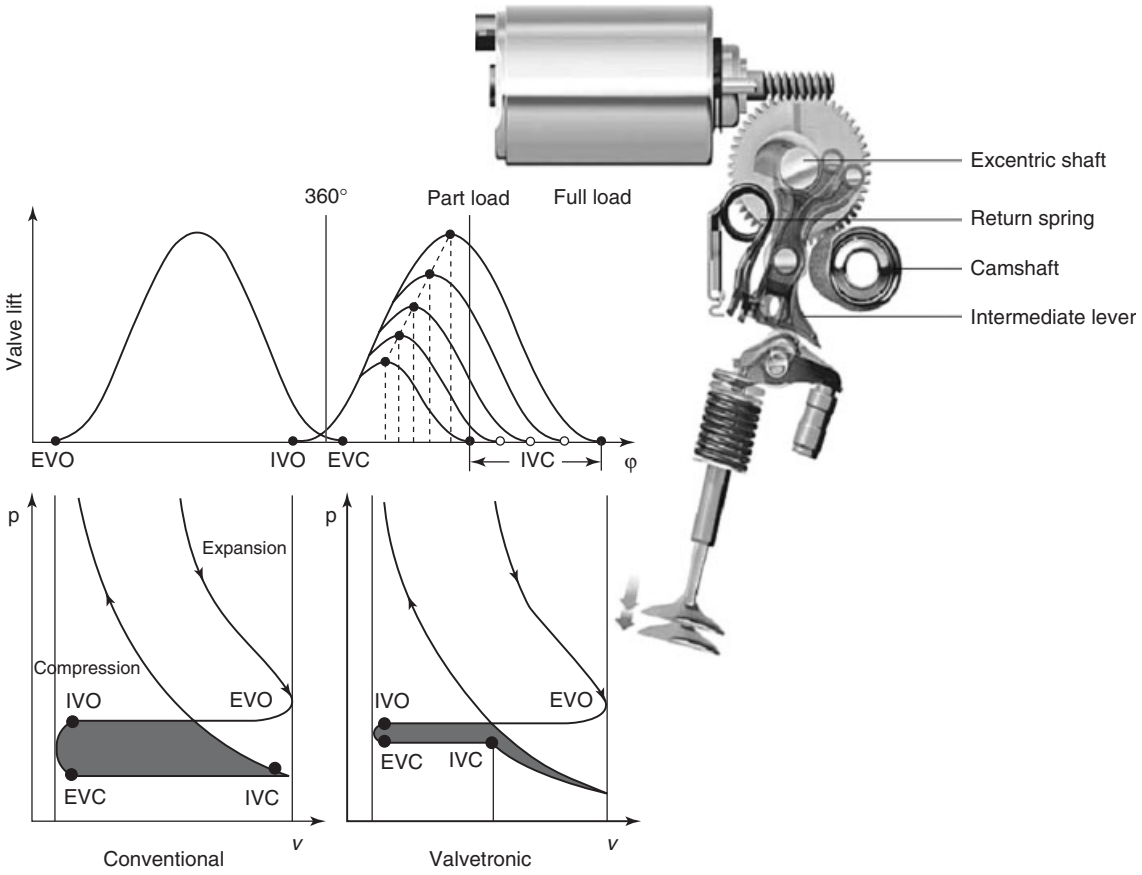


Figure 14. Variable valve train system and early intake valve closing. (Reproduced from Klüting *et al*., 2001. With kind permission of Springer Science+Business Media.)

closing time determines how much air or mixture is drawn into the cylinder. The intake pressure is higher and the duration of the intake process is shorter, so the pumping loop is significantly smaller than for the throttled SI engine (represented by the left diagram in Figure 14). The gas in the cylinder is heated by the warm cylinder walls, so the pressure at a given cylinder volume is slightly higher during the compression stroke than it was during the expansion (Figure 14). The use of this variable valve train (VVT) system can improve fuel economy by up to 10%. This is a result of the lower pumping losses, which is indicated by the smaller shaded area in the lower right plot relative to the lower left plot in Figure 14, as well as the better formation process of the air–fuel mixture because of the higher speed of the intake flow, which is produced by the smaller valve opening area with smaller valve lifts.

3.2 Advanced direct injection (DISI)

The first modern SI engine with direct injection (GDI) was introduced in 1995 in the Japanese market and in improved form in 1997 in the European market. This first generation of stratified charge engines operate according to the wall-guided combustion system with a wide spacing between the injector and the spark plug (Figure 3). The fuel spray is generally guided toward the spark plug by both the bowl-shaped piston crown on the intake side of the combustion chamber and the reverse tumble flow generated by the design of the intake port. The vaporization process and the mixture transportation to the spark plug are influenced by the interaction between the injection spray, the in-cylinder flow, and the piston shape.

In 2000, the first engine with fuel stratified direct injection (FSI) was released onto the European market (a 1.4-L four-cylinder, FSI engine by Volkswagen for the Lupo). As a result of direct injection, the engine's maximum power output could be increased from 74 to 77 kW. The piston was shaped with a bowl on the intake side. A tumble flow was created by a tumble system in the intake ports. The spark plug was centrally located between the intake and the exhaust valves. Owing to the cooling effect of the vaporized fuel, the compression ratio could be increased from 10.5 : 1

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

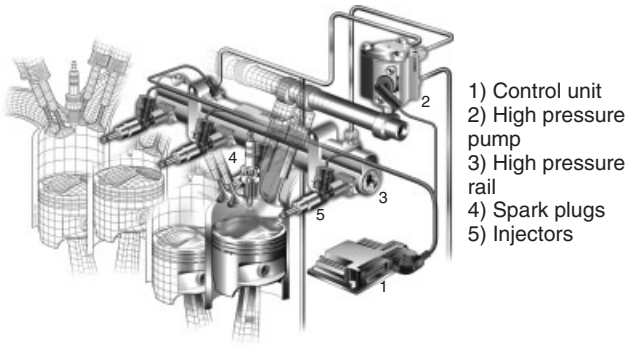


Figure 15. Common rail system in a four-cylinder DISI gasoline engine with lateral injectors. (Reproduced from Bosch, 2003. With kind permission of Springer Science+Business Media.)

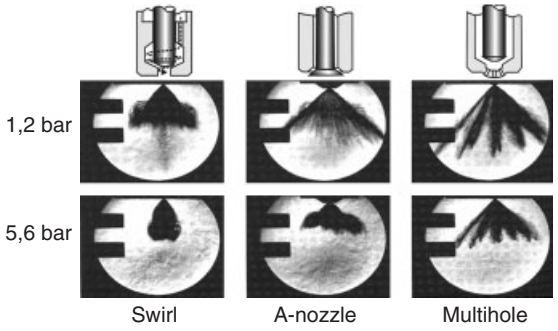


Figure 16. Nozzle types and spray patterns. (Reproduced from Wirth *et al*., 2003. With permission from Aachen Colloquium Automobile and Engine Technology, M. Wirth, D. Zimmermann, R. Friedfeldt, J. Caine, A. Schamel, Ford Werke AG, Köln and A. Storch, K. Ries-Müller, K.-P. Gansert, G. Pilgram, R. Ortmann, G. Würfel, J. Gerhardt, Robert Bosch GmbH, Stuttgart.)

to 12 : 1. However, all current FSI engines are operated with homogeneous mixtures for all engine operation conditions and use of the stratified charge operation is completely avoided, as it is the case with all turbocharged TSI engines until now. The reason for that is the lack of sulfur-free fuel in the worldwide market, which is necessary for suitable aftertreatment systems for lean combustion.

All DISI engines are equipped with common rail injection systems, which have been in production for diesel engines since the beginning of the 1990s; this technology makes modern DISI engines possible. Figure 15 shows a typical common rail system used in a four-cylinder engine for homogeneous mixture formation with a wide spacing between the spark plugs and the injectors. The common rail injection system offers high flexibility to adapt the injection to the engine operating point.

Fuel pressure generation and injection are decoupled in this system. The fuel system of the injection unit consists of a low pressure section (a low pressure pump in the fuel tank) to transport the fuel to the high pressure pump, which supplies the rail with fuel at a pressure of between 100 and 200 bar. The injectors are connected to the rail and thus to a relatively stable source of pressurized fuel. At injection time, the fuel is injected into the combustion chamber as dictated by the ECU for various engine operating conditions (common rail injection systems are also discussed in Trends—Compression Ignition).

Three different types of nozzles are traditionally used for fuel injectors with GDI. The nozzles can be differentiated from one another by the means in which they open and close and the spray patterns that they produce. Figure 16 shows the design of these nozzles as well as their corresponding spray patterns.

The swirl nozzle exhibits very good atomization and high spray flexibility. It also has a relatively low sensitivity to contamination as well as mechanical and thermal interferences. It is relatively affordable because it is produced in large quantities. The fuel spray is cone shaped with spray angles that are influenced by both the fuel pressure and the cylinder pressure. This is a major disadvantage when a stratified mixture is desired, especially with a narrow spacing of the injector and spark plug. Hence, such nozzles are currently favored only for DISI engines operating with purely homogeneous mixtures.

Compared to the swirl nozzle, the outward opening A-nozzle has the distinct advantage of being able to produce a uniform hollow-cone spray. In addition, the spray angle of the A-nozzle is hardly affected by the backpressure. Using this injector type in conjunction with a piezo-actuator, multiple injections during one cycle are possible. Furthermore, the precise control of fuel mass, droplet size, and the shape of the fuel spray makes the piezo-actuated injection system superior to the conventional solenoid-actuated system. With narrow spark plug-injector spacing, the spark plug can be positioned in the recirculation area and also at the boundary of the fuel jet produced by the A-nozzle. The A-nozzle is an attractive choice for lean combustion with stratified mixture preparation because of the stability of its spray profile. The piezo-actuator's capability of multiple fuel injections can be used to improve mixture formation and thus the combustion process. The inflammation and burnout characteristics can be stabilized by optimizing the fuel injection times in relation to the time of ignition.

Owing to the high costs of piezo-actuated injectors, solenoid injectors with multihole nozzles, which are well known from diesel engines, have been introduced in the market in recent years. Sharply defined injection jets are a characteristic of the multihole nozzle. Because of the inadequate atomization quality of the multihole nozzle, it is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

only possible to establish a partially homogeneous mixture in the combustion chamber with the injection jet shape, even when using a variety of configurations. Enriched mixture zones containing excess fuel are directly adjacent to lean mixture zones containing insufficient fuel. Consequently, once an injection jet is ignited—the origin of ignition is determined by the position of the spark plug—the flame front in the combustion chamber does not progress uniformly. Rather, it spreads at varying speeds: the flame front accelerates in enriched zones of individual injection jets but slows down in lean zones between the injection jets. The number of nozzle holes significantly influences the operation characteristics of the engine. To avoid this behavior as well as to create a sufficiently homogeneous mixture, an in-cylinder flow with high turbulence is necessary for good mixture formation. A multihole pattern that results in effective fuel distribution and constant flame spread between the various injection jets produces the best engine operating characteristics. The lean burn capability can be improved by increasing the number of holes and also reducing the individual hole sizes. However, the number of holes is currently limited because of an increasing tendency toward coking. The spray penetrates further with multihole nozzles than with A-nozzles. This can lead to excessive wetting of the piston as well as the cylinder liner during retarded fuel injection operations, thus increasing hydrocarbon and soot emissions.

In order to increase the efficiency of SI engines, it is desirable to operate the engine at higher loads, where pumping losses and hence the brake-specific fuel consumption (BSFC) are lower. This can be accomplished by making the engine smaller, either by reducing the displacement volume or by reducing the number of cylinders; this concept is referred to as *downsizing*. However, in order to meet customers' expectations of increased power and dynamic response, downsizing is very often combined with boosted operation. In 2005, the first downsized engine (four-cylinder, 1.4-L TSI) with boosting was introduced in the market. This engine is equipped with direct injection with multihole injectors, variable tumble flaps, an intake cam phaser, a supercharger, and a turbocharger. The fuel injector is located in the cylinder head below the intake ports and the spark plug is located in the center of the cylinder head between the intake and exhaust valves. Owing to this wide spacing arrangement of injector and spark plug, the mixture formation is homogeneous for the entire engine operating range. Figure 17 shows a schematic of the arrangement of the supercharger (compressor) and the turbocharger of this engine.

The compressor is belt driven by the crankshaft and the amount of boost is adjusted by a special control valve in the intake system. At an engine speed of 2500 rpm, the compressor is disconnected from the crankshaft by a magnetic clutch and the control valve is fully opened.

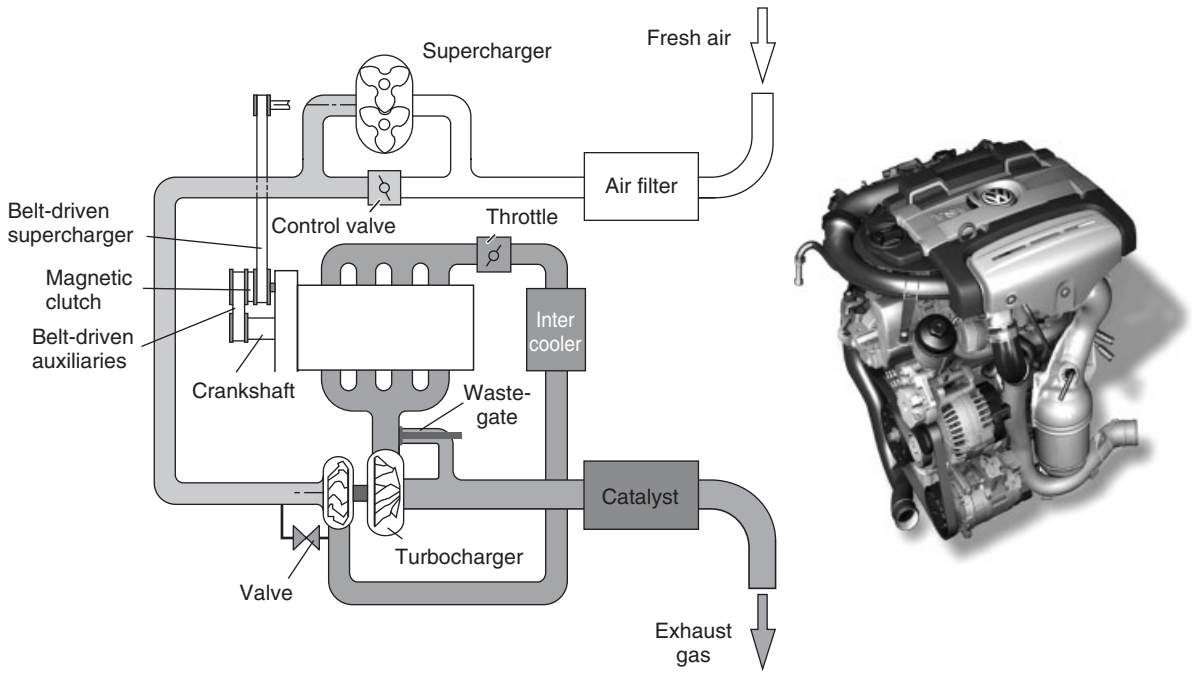


Figure 17. The 1.4-L four-cylinder engine with twin boosting. (Reproduced with permission from Middendorf *et al*., 2005. © RWTH Aachen University.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The charge cooling provided by the vaporizing fuel improves the volumetric efficiency of the engine and decreases its propensity to knock. This allowed the designers to increase the compression ratio of the engine to 10.0 : 1, thus increasing the thermal efficiency. A specially designed piston bowl and the injector design help prevent the injected fuel from wetting the cylinder walls and aids in directing the fuel injected in the compression stroke toward the spark plug during the catalyst heating phase of cold start operation. Injection pressures of up to 150 bar are necessary for the wide range of engine speeds and loads required of this engine. The intake cam phaser and the tumble flaps are used to optimize volumetric efficiency and in-cylinder flow over the entire range of engine operating conditions. The supercharger is used to boost intake pressure at low engine speeds and high loads and increases the available exhaust enthalpy (Figure 18). This significantly improves the dynamic response of the turbocharger, which is used to boost the engine at higher engine speeds. Special materials are used in the design of the turbocharger so that exhaust temperatures of 1050°C are possible, so mixture enrichment for the purpose of cooling at high loads is minimized.

The engine has a peak torque of 240 Nm that is available from 1750 to 4500 rpm, which is a significant contribution in a car that accelerates faster than its predecessors with a fuel consumption improvement of nearly 20%. The fuel consumption during the New European Driving Cycle (NEDC) of the Golf TSI car powered by this engine (125 kW) was 7.2 Le/100 km or 30 mpg. The current engine is equipped with only a turbocharger, which not only results in a reduced power output of 118 kW but also has reduced fuel consumption in the NEDC of 6 L/100 km or 36 mpg.

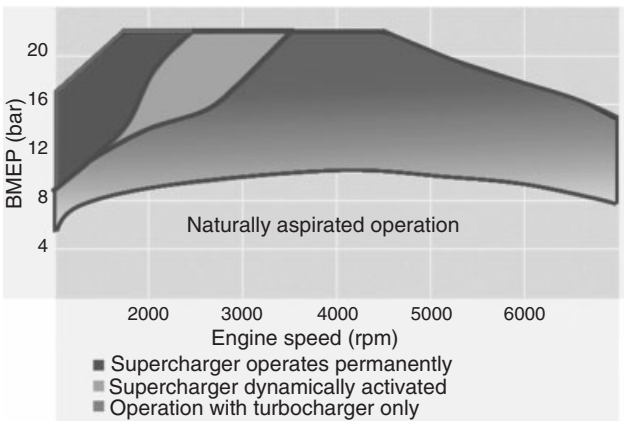


Figure 18. Boosted operation of the 1.4-L TSI. (Reproduced with permission from Middendorf *et al*., 2005. © RWTH Aachen University.)

One new technology is the use of twin-scroll turbochargers. The latest engine to be outfitted with this technology is a four-cylinder 1.6-L engine, which is additionally equipped with fully variable valve lift, intake and exhaust cam phasers, and direct injection, [8]. The charge cooling provided by direct injection makes a compression ratio of 10.5 : 1 possible. In order to significantly reduce pumping losses at part load, the engine is operated at nearly WOT. The continuously variable valve lift and an early intake valve closing are used to control the amount of air that enters the cylinder and, for a homogeneous charge at part load, the load. The valve lifts and timings are also used to increase the residual gas fraction in the cylinder and to generate desirable in-cylinder flow, which improves the combustion characteristics at part load. At higher loads, the in-cylinder flow is created by tumble ducts and the load is controlled by the boost pressure. The twin-scroll turbocharger provides a much improved dynamic response compared to the single-scroll turbocharger.

This technology results in engine torque characteristics that are similar to those of a diesel engine. The maximum engine torque of 250 Nm is available from 1500 up to 4800 rpm (Figure 19). This engine has a displacement volume

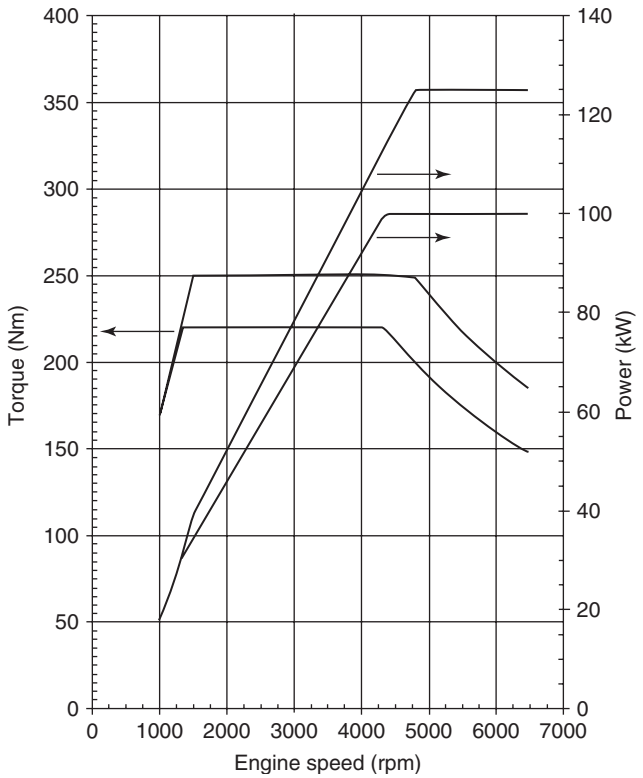


Figure 19. Full-load curves of the 1.6-L TwinPower Turbo Engine. (Reproduced with permission from Schopp *et al*., 2011. © RWTH Aachen University.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

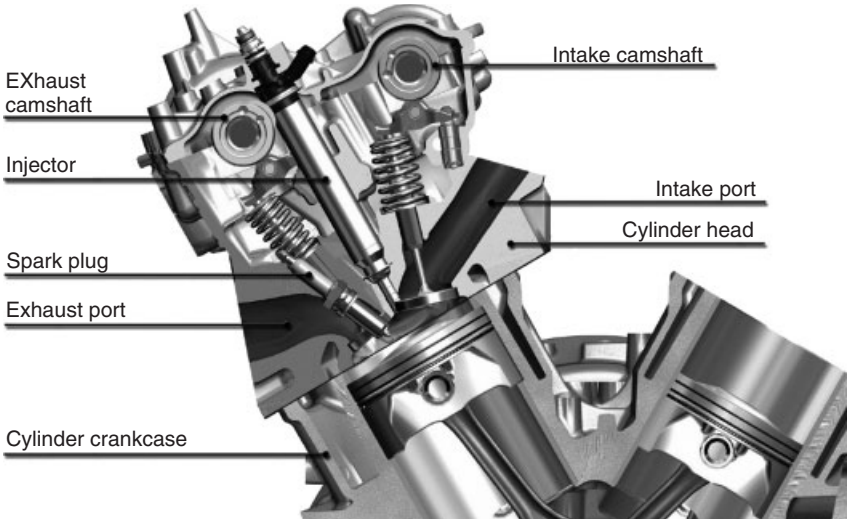


Figure 20. Arrangement of piezo-injector and spark plug for the first production engine with spray-guided combustion system (Altenschmidt *et al*., 2011). (Reproduced by permission of Daimler AG.)

20% smaller than its predecessor and a 30% greater torque output, and the fuel economy is improved by 5% (Schopp *et al*., 2011).

In 2006, the first spray-guided direct injection engine (3.5-L V6 engine) with stratified operation was introduced in the European market (Altenschmidt *et al*., 2011). The fuel injector is an outward opening piezo-actuated A-nozzle and is located together with the spark plug centrally in the four-valve cylinder head (Figure 20). The fuel is injected with fuel pressures of up to 200 bar. Without this arrangement, mixture stratification and combustion for production applications would not be possible. The special characteristics of the piezo-injector, including high nozzle capacity, small spray fluctuations, and the ability to precisely inject the smallest quantities of fuel, made it possible to use special injection strategies for stratified and lean operation.

Figure 21 shows the spray dispersion for an injection at the end of compression. The fuel injection enables a combustible mixture to be positioned at the spark plug in a way that reliable ignition occurs. The spray characteristic forms the basis not only for robust and reliable stratified operation but also for operating modes with critical ignition conditions.

To extend the range of both mixture stratification and lean combustion of a DISI engine in the engine operation map, a special mixture formation strategy was developed, which is shown in Figure 22.

Depending on the engine speed and load, the engine is operated in the most favorable mode of with a homogeneous stoichiometric mixture, a combination homogeneous lean/stratified mixture, or a purely stratified mixture. The corresponding injection strategies are shown as circles with gray-shaded sections representing the injection events. The individual injections take place either during the intake stroke (shown at the right side of the circles) or during the compression stroke (shown at the left side of the circles) to achieve the desired mixture stratification. The best injection strategy with up to three different injections is used for each mode. The ignition timing is adjusted together with the injection strategy to optimize operating behavior. With a triple injection, spark timing typically occurs between the second and third injections.

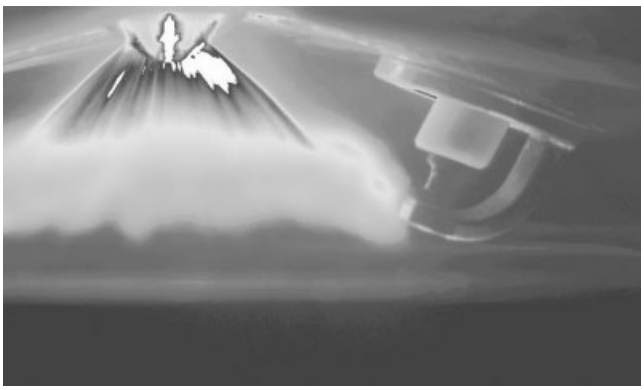


Figure 21. Spray dispersion for injection during compression (Altenschmidt *et al*., 2011). (Reproduced by permission of Daimler AG.)

With purely stratified operation, the air–fuel mixture is located in the center of the combustion chamber, whereas

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

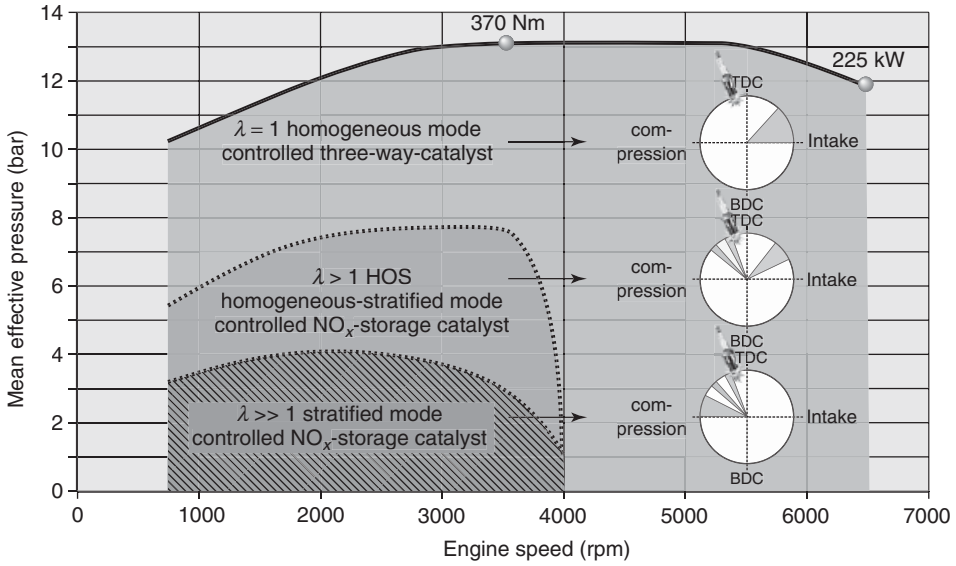


Figure 22. Operating modes in the entire engine map (Altenschmidt *et al*., 2011). (Reproduced by permission of Daimler AG.)

the outer regions of the combustion chamber are surrounded by almost pure air. In homogeneous stratified mode, a homogeneous lean mixture of air and fuel is created with an early fuel injection and is present in the entire combustion chamber. In addition to this lean mixture, a small and stratified charge is induced directly at the spark plug by late-fuel injection during the compression stroke. This predominantly homogeneous distribution is particularly favorable under high load conditions. While the air–fuel mixture can become overly rich in stratified mode, which increases fuel consumption and emissions, these disadvantages are not evident in the homogeneous stratified mode. Using this operating mode, fuel consumption can be further reduced with increasing load (Figure 23).

It can be seen that considerable benefits are achieved for stratified operation up to a brake mean effective pressure (BMEP) of approximately 4 bar. With further increases, BSFC does not continue to improve but actually increases slightly. Fuel consumption with homogeneous stratified mode is more favorable than in stratified mode at a load higher than BMEP = 4.5 bar. Between BMEP of 4.5 and 7 bar, the engine operates in the homogeneous stratified mode. With increased loads of more than 7 bar, the electronic control unit (ECU) switches automatically to the pure homogeneous mode without stratification.

When examining the energy losses for a PFI engine, a PFI engine with VVT, a DISI engine with a wide spacing arrangement, and a DISI engine with a narrow spacing arrangement, the individual losses can be characterized (Figure 24). The energy required to drive the high pressure pump in a DISI engine increases the auxiliary energy demand, and the friction in a DISI engine is slightly higher than for the PFI engine. The degree of dethrottling and therefore the amount of pumping work is at a minimum for a VVT system and for the narrow spacing arrangement. The higher combustion chamber temperatures for the PFI engine tend to increase heat losses, but the lower in-cylinder mass and pressures tend to offset this, so the heat exchange losses do not vary significantly among the various concepts.

Incomplete combustion due to over lean regions of the combustion chamber results in a higher amount of energy loss for the DISI engines. The products of the extremely lean mixtures found in the DISI engines have thermodynamic characteristics superior to those of a stoichiometric mixture found in the PFI engine and the VVT engine, so useful work is more effectively extracted from the expanding combustion products. The higher compression ratio of the narrow spacing DISI engine leads to increased work output and therefore less energy lost to exhaust heat.

4 KNOCKING COMBUSTION

Both the thermal efficiency and the power output can be improved by increasing the compression ratio of SI engines. However, for high specific engine loads, achieved for example by boosting or downsizing SI engines, the increased compression ratio can lead to engine knocking and therewith to severe engine damage. The investigation of engine knocking has been a major research topic since the invention of the SI engine and requires the highest instrumentation accuracy for the analysis of measurement

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

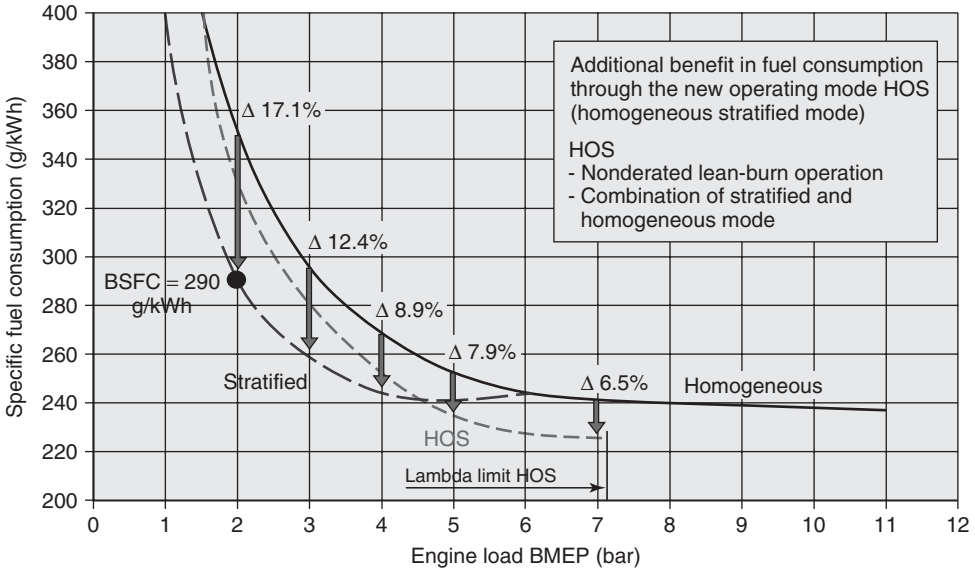


Figure 23. Brake-specific fuel consumption (BSFC) for a homogeneous, stratified, and combined homogeneous/stratified (HOS) mode of operation as a function of engine load at 2000 rpm (Altenschmidt *et al*., 2011). (Reproduced by permission of Daimler AG.)

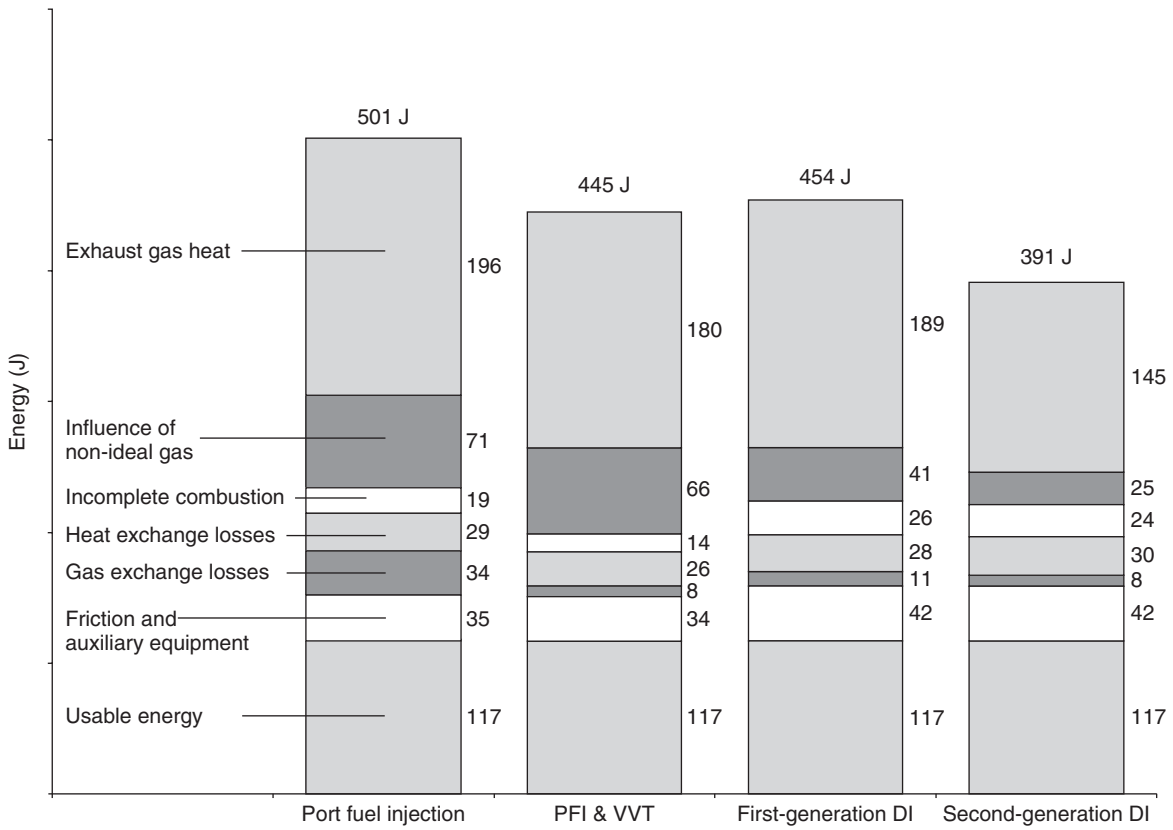


Figure 24. Analysis of energy losses for PFI, VVT, and both first- and second-generation DISI engines for a fixed power output (usable energy). (Reproduced from Lückert *et al*., 2004. With permission from Aachen Colloquium Automobile and Engine Technology, Lückert *et al*., Daimler AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

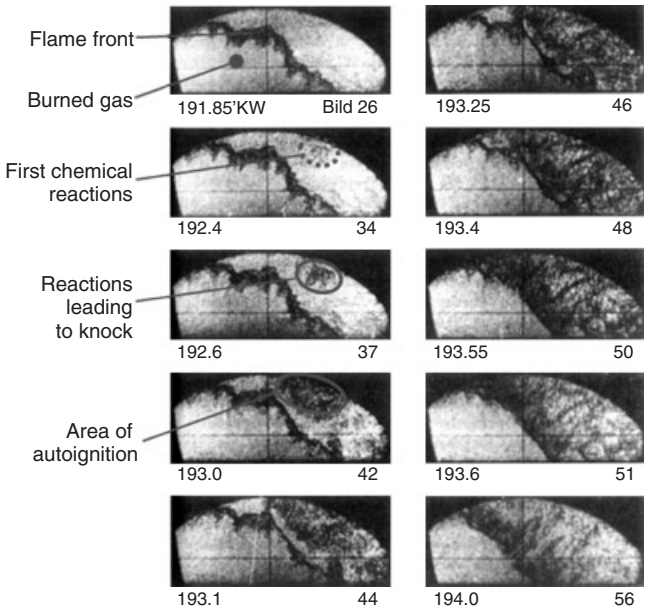


Figure 25. Schlieren photography of knocking combustion. (From Spicher, Kröger and Ganser, 1991. Copyright © 1991 SAE International. Reprinted with permission.)

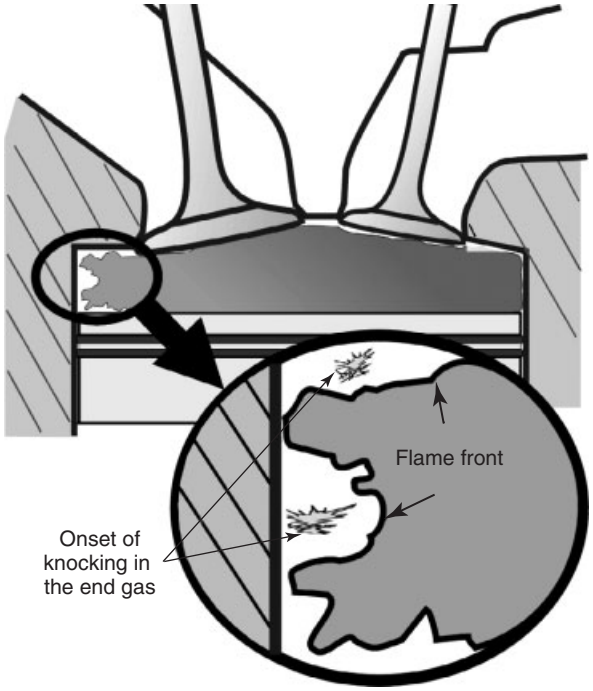


Figure 26. Autoignition in end-gas "hot spots."

results, because the location, the onset, and the intensity of the knocking event must be exactly determined. At present, it is well known that knocking combustion is a result of an autoignition in the region of unburned air–fuel mixture in front of the propagating flame front during the late phase of normal combustion. Many investigations during the past three decades with different optical measurement techniques have confirmed this behavior. Figure 25 represents knocking combustion in the end-gas region, detected by high speed Schlieren cinematography in a single-cylinder engine (Spicher, Kröger and Ganser, 1991). The intensity of knocking during this combustion cycle has been defined as an intermediate level with a zero-to-peak amplitude of 11 bar in pressure oscillations, created by knocking. The first picture (number 26) was taken at 191.85 crank angle degrees (CAD) after bottom dead center (ABDC) or 11.85 CAD ATDC. At this time, only the flame front of the nonknocking combustion can be seen. In picture 34, the first chemical reactions of autoignition are detected. From this autoignition location, a so-called reaction front propagates quickly through the end-gas region from picture 37–44, which seems to press the original flame front back slightly (pictures 46–56).

The pressure and temperature increase in the combustion chamber during spark-ignited flame propagation can lead to critical thermodynamic conditions, thus causing this autoignition either in the end-gas region or in end-gas pockets, shown schematically in Figure 26. These pockets are formed because of a turbulent flame front close to the combustion chamber walls in the last phase of flame propagation.

Some of these pockets with short ignition delays due to nonuniformities in temperature and concentration, the so-called hot spots, are able to autoignite. Their location is strongly dependent on the local temperature–time history in the end gas. For this reason, the phenomenon of regular engine knocking can be detected shortly after the SI and generally occurs in a reproducible manner.

The knock amplitude after a self-ignition depends on the chemical reaction rate. At high reaction rates within the hot spots, pressure compensation by expansion into the surrounding areas is not possible. Pressure waves propagate through the combustion chamber causing mechanical vibrations (knocking) when they reflect on the cylinder walls. The differences of knocking intensity and frequencies result from fluctuations of the end-gas boundary conditions.

Owing to the local autoignition in the end-gas region during the late combustion, pressure waves are induced. The amplitude of the pressure waves depends on the pressure, temperature, mixture composition, and the volume of the pockets in the end gas where autoignition occurs (Rothe *et al*., 2006). The pressure waves propagate through the combustion chamber and are reflected at the combustion chamber walls. The characteristic pressure oscillations can be recognized in the pressure trace shown in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

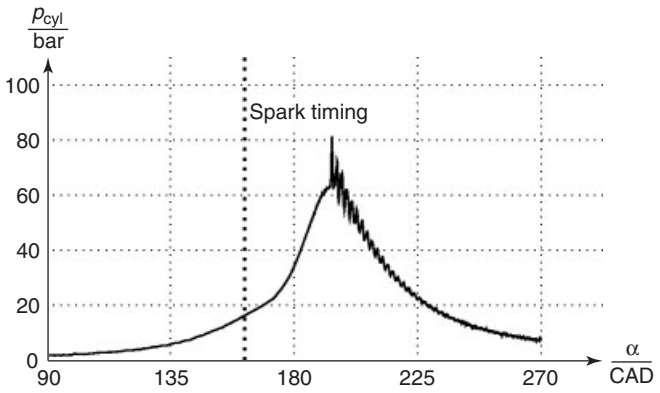


Figure 27. Pressure trace of knocking combustion.

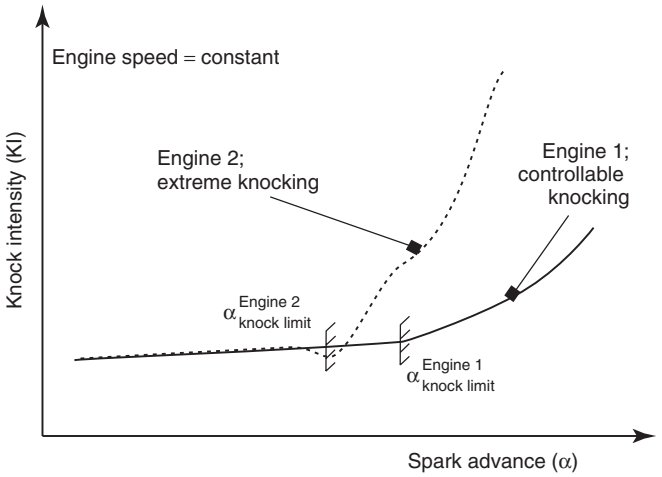


Figure 28. Knock intensity versus spark advance for two engines.

Figure 27. The resulting structural vibrations are responsible for the characteristic knocking sound that gave this phenomenon its name. The pressure wave interferes with the thermal boundary layer at the combustion chamber walls, which leads to locally elevated heat transfer rates. In conjunction with the increased mechanical loading due to the high pressure gradients and vibrations, this can cause severe engine damage. To avoid this, piezoelectric sensors using the inertia of a seismic mass to detect structure-borne sound induced by knock are employed in modern engines to operate them close to the knock limit. This is possible because of the knock limit being rather well defined as a function of ignition timing. The effect is well reproducible and thus controllable. If knock is detected, the spark timing is retarded, which decreases the risk of knock in the subsequent cycles at the cost of thermal efficiency.

In some engines, a more extreme type of knocking combustion can be observed. This phenomenon is referred to as *super-knock*, *megaknock*, or *extreme knock*. It is characterized by pressure amplitudes that are significantly higher than in the case of normal knock. In addition, super-knock occurs stochastically so that it is not controllable in the same manner as normal knock. Figure 28 shows the knock intensity versus spark advance for two different engines (Spicher and Krebs, 1990). While engine 1 shows a slight increase in knock intensity with spark advance, engine 2 shows a steep increase in knock intensity with spark advance. That means engine 1 can be easily controlled by the knock control system. After a certain spark advance when knock intensity exceeds the threshold of knock limit, spark advance is retarded by the knock controller. For engine 2, the knock intensity increases too fast and the threshold of knock limit is exceeded directly after spark advance of knock limit. This behavior leads to extreme knocking combustion, which hardly can be controlled by the knock control system due to its stochastic occurrence nature.

Figure 29 shows the knock characteristics of an engine at two different operating points. Both diagrams show a variation of spark timing. Beginning at 1.6 CAD after knock limit α_{KL}, ignition is advanced in four steps until 1.6 CAD before knock limit.

At 3000 rpm, knock amplitudes Δp increase nearly linearly with spark advance. Therefore, at this operating point, the knock control described earlier works as intended. However, at 5000 rpm, extreme knock events exhibiting significantly higher amplitudes occur stochastically, which cannot be controlled by a knock control unit. Besides these differences, the progress of extreme knock is very similar to that of normal knock.

Figure 30 shows the pressure trace, the flame signal detected with an optical fiber, and the location of knock onset for a combustion cycle with extreme knocking. It can be seen from the pressure trace that the maximum amplitude in knocking pressure oscillation is nearly 70 bar. Knock onset occurs close to the cylinder liner wall at the left intake valve. The reason for that is the structure of the in-cylinder flow generated by the design of the intake manifolds. While the right intake manifold creates a swirl flow, the left intake manifold is designed for high volumetric efficiency. Owing to these two different flow conditions, a weak flow movement with enrichment in air–fuel ratio occurs at the location of knock onset. The CAD difference between the steep increase between oscillations in flame signal and oscillations in pressure trace is 1.2 CAD. On the basis of the engine speed and the distance between the location of the optical fiber and the pressure trace, the speed of the knocking combustion front

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

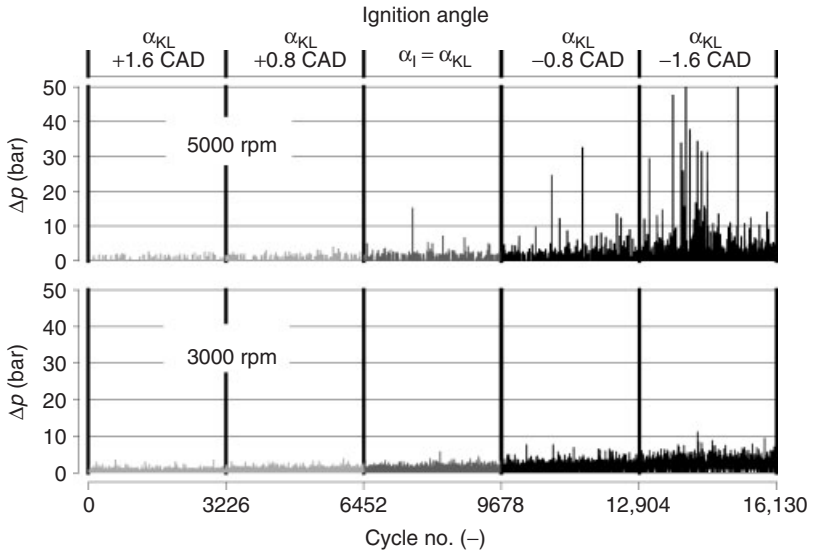


Figure 29. Pressure amplitudes for extreme knock (top) and controllable knock (bottom).

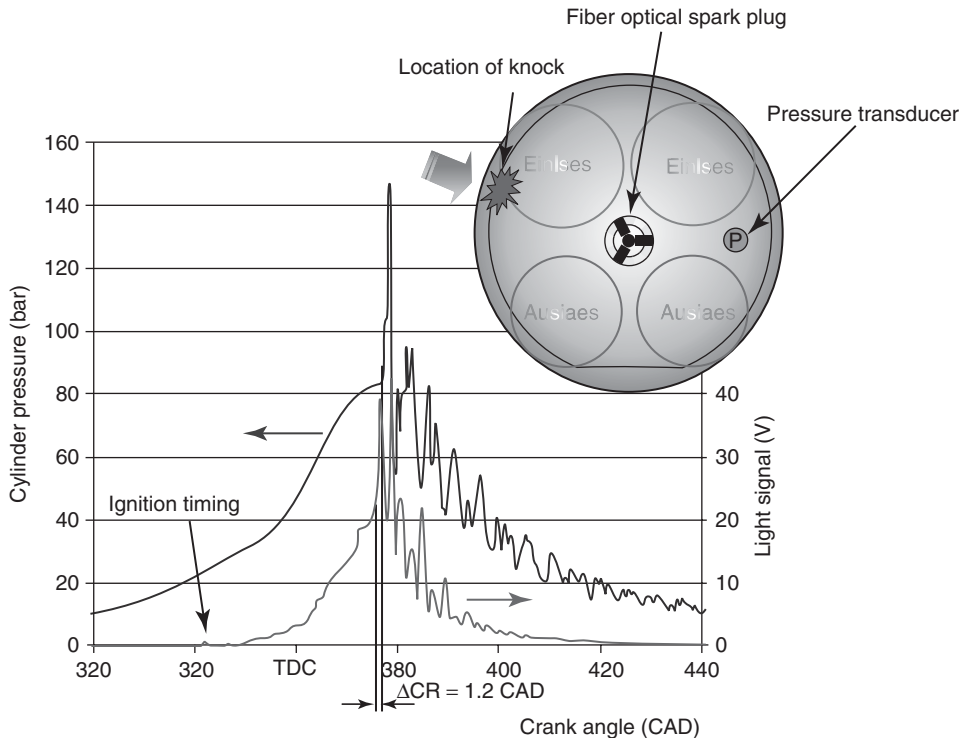


Figure 30. Extreme knocking combustion for a DISI engine.

can be estimated to approximately 1150 m/s, which can be supersonic speed.

Knock as well as extreme knock causes an elevated heat transfer to the walls. Consequently, parts of the combustion chamber can be heated up significantly. This can promote earlier autoignition in the end gas, which results in even stronger knocking and thus further heating. This effect, referred to as *surface ignition*, can be self-preserving or even self-amplifying (Figure 31). The diagram shows the pressure traces of four subsequent engine cycles with

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

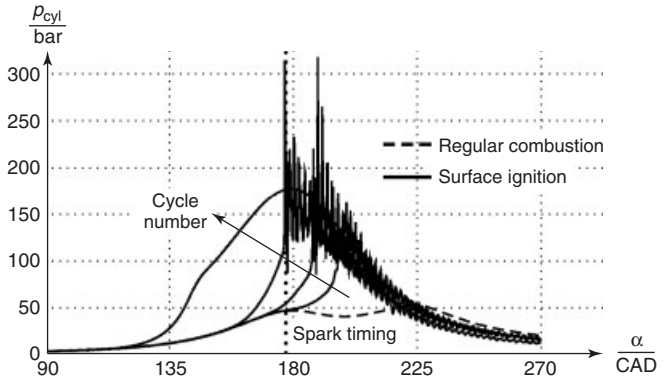


Figure 31. Pressure traces of consecutive cycles after a surface ignition.

surface ignition, and, as a reference, one of a regular combustion event at the same operating point.

It is seen in the figure that in the first cycle with surface ignition combustion starts considerably earlier than in the cycle with regular combustion. The first pressure rise is observable almost simultaneously with spark timing. Owing to this early onset, combustion progresses into very severe knock. In the next two cycles, autoignition timing moves forward because of the heated combustion chamber. In the last cycle plotted in this diagram, combustion starts at an extremely early stage of compression so that the complete cylinder charge is consumed before autoignition can occur in the end gas. Consequently, there is no knock observable in this cycle. However, because of the burned gas being compressed, peak pressure and temperature are much higher than in the case of regular combustion, which can cause immediate engine breakdown. Surface ignition is extremely dangerous but can be avoided reliably by proper dimensioning of the cooling system and the right choice of spark plug.

During recent years, with the appearance of highly supercharged downsizing engines, a new combustion phenomenon has been observed in these engines, especially in full-load operation at low engine speed. This phenomenon, referred to as *premature ignition* or *preignition*, shows some similarities to extreme knock as well as to surface ignition. As in the case of surface ignition, combustion obviously starts before the spark plug fires. However, the phenomenon acts neither self-preserving nor self-amplifying. Instead, preignition occurs stochastically, just like extreme knock. All three phenomena have in common that they are characterized by very high knock amplitudes. Figure 32 shows pressure traces for cycles with preignition and regular combustion at the same operating point.

Owing to the enormous knock amplitudes, preignition is considered one of the most important issues in the development of downsized engines. However as it is the newest and thus least explored of the described phenomena, there is still very little research about it (Palaveev and Spicher, 2010; Dahnz and Spicher, 2010).

Knocking is significantly influenced by the operating conditions, fuel, and the shape of the combustion chamber. In principle, the knocking tendency of an engine increases under the following boundary conditions:

- high pressure and high temperature in the unburned remainder of the air–fuel mixture;
- proximity to the stoichiometric air–fuel ratio ($\lambda = 1$);
- fuel with a low octane number; and
- high compression due to a high compression ratio and/or boosting of the premixed air–fuel mixture.

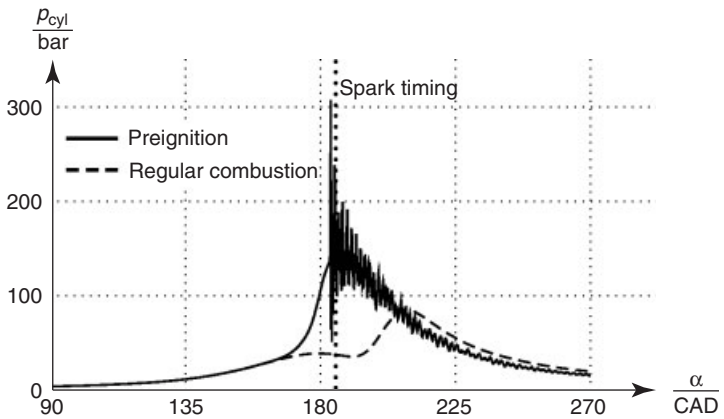


Figure 32. Pressure traces of regular combustion and preignition. (Reproduced by permission of Palaveev and Spicher, 2010. © S. Palaveev and U. Spicher.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

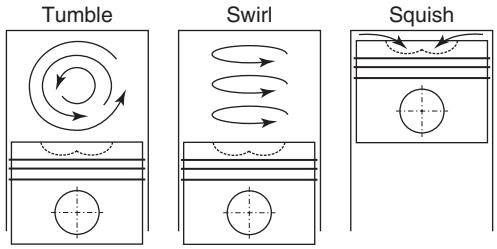


Figure 33. In-cylinder tumble (rotation around a horizontally oriented axis), swirl (rotation around the cylinder axis), and squish flow.

The combustion chamber design can have a considerable influence on combustion and knocking. Combustion chambers with low knocking tendencies (knock-resistant combustion chambers) must meet the following basic requirements:

- Short flame paths: compact combustion chamber, centrally positioned spark plug
- Avoidance of hot spots at the end of the flame path: spark plugs close to the exhaust valve
- High flow velocities: tumble and/or swirl movement and/or squish flow (Figure 33)

5 POLLUTANT FORMATION

The fuels used in SI engines typically consist of hydrocarbons. The combustion of these fuels is the oxidation of these hydrocarbons with the oxygen contained in the intake air. When a hydrocarbon molecule is completely burned, it ideally produces only carbon dioxide (CO_2) and water (H_2O). The following chemical reaction describes generally the conversion of hydrocarbon fuels such as petrol:

$$C_xH_y + \left(x + \frac{y}{4}\right) O_2 \Rightarrow x\,CO_2 + \frac{y}{2} H_2O$$

Owing to incomplete combustion processes, the exhaust gas also contains the following pollutants:

- carbon monoxide (CO);
- unburned hydrocarbons (UHC);
- nitrogen oxides (NO_x);
- particulates including soot, mainly in engines with direct fuel injection; and
- aldehydes (H-C-O compounds), predominantly in engines with pure ethanol or ethanol blends as fuel.

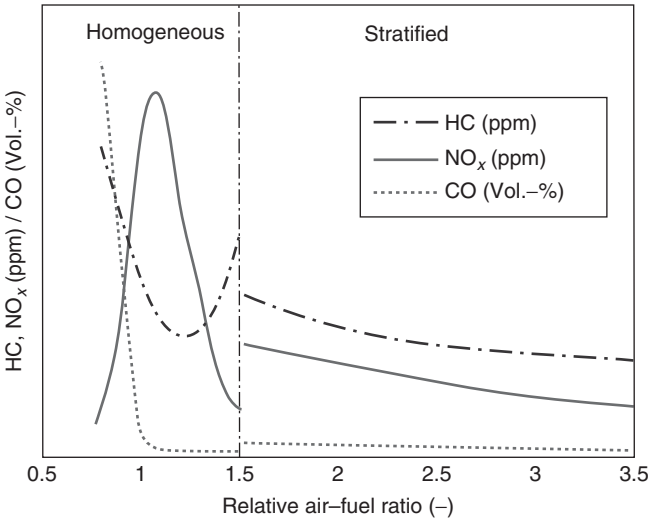


Figure 34. Emissions concentration with homogeneous (PFI and DISI) and stratified (DISI) mixture formation.

With the exception of nitrogen oxides, the various pollutant components are primarily determined by the fuel and its additives. The emission level is significantly determined by the engine operating characteristics. Figure 34 shows the fundamental dependence of pollutant concentrations for CO, NO_x, and UHC (HC) on the air–fuel ratio for an SI engine with PFI as well as with DISI. For an air–fuel ratio between $\lambda = 0.8$ and 1.5, the engine can be operated with homogeneous mixture. The resulting curves of pollutant concentrations therefore match those of an engine with either PFI or DISI. Stratified charge operation for air–fuel ratios greater than 1.5 produces the illustrated characteristics of UHC, nitrogen oxide, and carbon monoxide emissions.

In air-deficient, fuel-rich zones ($\lambda < 1$), carbon monoxide is a product of incomplete combustion. In the range of $\lambda \approx 1$ and with lean combustion ($\lambda > 1$), CO formation essentially results from any inhomogeneous mixture distribution resulting in some rich mixtures or delayed combustion in which the chemical reaction does not reach completion because the temperature of the reacting gases has become too low during expansion and the dissociation of CO_2 into CO and O_2 during combustion.

In addition to CO emissions, incomplete combustion in air-deficient zones leads to the emission of UHCs. UHC emissions also form in combustion chamber regions that have not completely been penetrated by the flame, such as boundary layers near combustion chamber walls where the flame is extinguished by cooling (quench effects) or crevices (e.g., ring land crevices) into which the flame cannot penetrate. Besides these UHC emissions resulting

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

from quench effects, UHC emissions result from crevice volumes, deposits, and desorption from the oil film on the combustion chamber surfaces. During expansion and expulsion in the exhaust manifold, the UHCs formed by various methods are mixed with the hot exhaust gas and oxidized by after-reactions. Both high temperatures and the presence of oxygen are important agents for these after-reactions. The after-reactions are one reason for the minimum level of UHC emissions observed in the presence of only a slight air surplus. An extremely high air surplus during combustion will cause flame quenching that strongly increases the level of UHC and CO concentrations. The combustion temperatures for these very lean conditions are not high enough to maintain flame propagation in the cooler zones of the combustion chamber. The fundamentals of UHC and CO emissions are also discussed in more detail in UHC and CO Formation and Models.

Nitrogen oxide emissions (NO_x) are a combination of the nitrogen monoxide NO (approximately 90%), nitrogen dioxide NO_2 (approximately 9%), and nitrous oxide N_2O (approximately 1%). The maximum combustion temperature and the duration of its effects are the most decisive influences on the NO_x concentration, especially the concentration of NO. While the combustion temperature peaks in the engine are encountered at an air–fuel ratio of $\lambda \approx 0.95$, the maximum NO_x concentration occurs in the lean range between $\lambda \approx 1.05$ and 1.1. This is due to not only the high temperatures but also the availability of oxygen in the burned mixture. NO_x Formation and Models discusses NO_x formation fundamentals in more detail.

First SI direct injection engines with stratified mixture preparation were designed with wide spacing arrangement of fuel injector and spark plug. For injection of the fuel solenoid, actuated injectors and swirl nozzles were used. The fuel pressure for injection was in the range 80–120 bar. The wall-guided strategy with significant wall wetting fuel on the piston surface results in relatively high emissions of both UHCs and soot particles, especially at higher load as well as during cold start and engine warm-up. Owing to these difficulties with particulate emissions, most of the automobile companies changed to pure homogeneous mixture preparation in the entire engine operating range. In addition, to exploit the advantages of direct injection, boosting was introduced in combination with this change in mixture formation. At present, only one automaker is still running SI engines with stratified operation, whereby the injector is centrally mounted in the cylinder head close to the spark plug (spray-guided or narrow arrangement). Instead of solenoid-actuated fuel injectors, these engines are equipped with piezo-actuated injectors and outward opening nozzles (A-nozzle), which allow very flexible multi-injection strategies. Owing to this flexibility of injection timing and numbers of injection pulses, the particulate emissions can be kept to an extremely low level during engine operation over test cycles and also during real world driving (Altenschmidt *et al*., 2011). In addition, during cold start and catalyst heating, a centrally mounted hollow-cone injector (A-nozzle) and multipulse injection strategies with short injection pulses just before ignition timing allows engine operation with low particulate number emissions ($<10^5$ p/cm^3) while keeping the cyclic variations of combustion on a sufficient low level (Piock *et al*., 2011).

Another main parameter to reduce particulate emissions is to increase injection pressures. At present, the fuel pressure with gasoline is limited to 200 bar because of the lack of lubrication from gasoline inside the high pressure fuel pump. Investigations with significantly higher fuel pressures up to 1000 bar have shown that particulate emissions can be kept to an extremely low level during both real world operation and cold start (Buri *et al*., 2009) (Schumann, Kubach and Spicher, 2012). Concluding from research work up to now, with the right injection system and the right injection strategy, an adequate mixture formation at homogeneous and stratified operations can be created, which allows keeping emissions of particulate mass and particulate number significantly below the most stringent emission regulations.

REFERENCES

Altenschmidt, F., Gildein, H., Sauter, W., and Waltner, A., (2011) The Spray-Guided Mercedes-Benz Combustion System—Developed not only for Stratified Mode, SIA Spark Ignition International Conference, Strasbourg, France.

van Basshuysen, R. (editor), (2004) *Internal Combustion Engine, Handbook*, SAE International, R-345.

Bosch (2003) *Gasoline Engine—Management*, Vieweg Publishing Corporation.

Buri, S., Busch, S., Kubach, H., and Spicher, U., (2009) High injection pressures at the upper load limit of stratified operation in a DISI engine. SAE-Paper 2009-01-2657.

Dahnz, C. and Spicher, U. (2010) Irregular combustion in supercharged spark ignition engines—pre-ignition and other phenomena. *International Journal of Engine Research*, **11** (6), 485–498.

Liebl, J., Klüting, M., Achilles, D., and Munk, F., (2001) The new BMW eight-cylinder engine with VALVETRONIC, Motortechnische Zeitschrift, MTZ 10/2001.

Lückert, P., Waltner, A., Schaupp, U., Vent, G., and Rau, E., (2004) Continuation of the gasoline direct injection development at Mercedes-Benz. 13th Aachen Colloquium for Vehicles and Engine Technology.

Middendorf, H., Krebs, R., Szengel, R., Pott, E., Fleiss, M., and Hagelstein, D., (2005) Volkswagen introduces the worlds first double charge air direct injection petrol engine. 14th Aachen Colloquium for Vehicles and Engine Technology.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Palaveev, S. and Spicher, U. (2010) Pre-ignition and knocking combustion in spark ignition engines with direct Injection. JSAE 20105390.

Piock, W., Hoffmann, G., Berndorfer, A., Salemi, P., and Fusshoeller, B., (2011) Strategies towards meeting future particulate matter emission requirements I homogeneous gasoline direct injection engines. *SAE International Journal of Engines*, **4** (1), 1455–1468. DOI: 10.4271/2011-01-1212

Rothe, M., Heidenreich, T., Spicher, U., and Schubert, A., (2006) Knock behavior of SI-engines: thermodynamic analysis of knock onset locations and knock intensities. SAE-Paper 2006-01-0225.

Schopp, J., Kiesgen, G., Kilias, H.-P., Lechner, B., Leistner, M., and Richter, R., (2011) The New 1.6-Litre Turbocharged Engines with Direct Injection and Fully Variable Valve Gear for the New BMW 1 Series Car, 20th Aachen Colloquium for Vehicles and Engine Technology.

Schumann, F., Kubach, H. and Spicher, U. (2012) The Influence of Injection Pressures of up to 800 bar on Catalyst Heating Operation in Gasoline Direct Injected Engines, 8th International Conference on Modeling and Diagnostics for Advanced Engine Systems (COMODIA 2012), pp. 305-310.

Spicher, U. and Krebs, R. (1990) Optical fiber technique as a tool to improve combustion efficiency. SAE-Paper 902138.

Spicher, U., Kröger, H. and Ganser, J. (1991) Detection of knocking combustion using simultaneously high-speed Schlieren cinematography and multi optical fiber technique. SAE-Paper 912312.

Wirth, M., Zimmermann, D., Friedfeldt, R., Caine, J., Schamel, A., Storch, A., Ries-Müller, K., Gansert, K.-P, Pilgram, G., Ortmann, R., Würfel, G., and Gerhardt, J., (2003) The Next Generation of Gasoline Direct Injection: Improved fuel Economy and Optimized System Cost, 12th Aachen Colloquium for Vehicles and Engine Technology.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto147
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Automotive Diesel Engine Development Trends

Dean Tomazic[1], Thomas Koerfer[2], and Stefan Pischinger[2]

[1] *FEV Inc., Auburn Hills, MI, USA*
[2] *FEV GmbH, Aachen, Germany*

1 Introduction 1
2 Historic Review 1
3 Current and Future Technologies 6
Endnotes 14
Further Reading 14

1 INTRODUCTION

Automotive diesel engine applications can be found worldwide because of their low fuel consumption and high level of robustness. With the continuous introduction of stringent emissions regulations, fuel economy targets (the United States), and CO_2 regulations in Europe, the United States and many other countries, the diesel engine development undergoes a permanent change to adapt to its new boundary conditions. Applying new materials, machining/manufacturing processes, design methods, fuel injection systems, aftertreatment components, control algorithms, sensors, and so on, the diesel engine is set to maintain its status as the most efficient automotive powerplant worldwide. This chapter will identify and highlight the corresponding topics relevant to the development of state-of-the-art automotive diesel engines and explain the associated background.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 HISTORIC REVIEW

2.1 Rudolf Diesel

Rudolf Diesel was born in Paris, France in 1858, as the second of three children of Theodor and Elise Diesel. His parents were immigrants from Bavaria, then living in Paris. At the age of 14, he stated his strong interest to become an engineer. After concluding his basic education in 1873 superior, he joined the newly founded industrial school of Augsburg and finished his classes as the best in 1875. Only 2 years later, he received a merit scholarship from the Royal Bavarian Polytechnic of Munich. He graduated in January 1880 with highest academic honors and returned to Paris, where he supported his former professor Carl v. Linde with the building of a new, modern refrigeration plant.

In early 1890, he moved from Paris to Berlin with his family considering management of Linde's R&D department. Initially, he worked in the area of steam processes, but his own research aimed for better fuel utilization. Owing to an accident, he had to stay a couple of months in a hospital, followed by continuous problems regarding general health. In 1887, he published a treatise entitled "Theory and Design of a Rational Heat-Engine to Replace the Steam Engine and Combustion Engines Known Today." Finally, in 1897, he built the first working model of his engine. On September 29, 1913, he entered in Antwerp on board of a ship targeting England. He seemed in good spirits, but was, after he had gone to his cabin, never seen again. The exact circumstances of his death, however, were never clarified up to now.

2.2 Advantages of diesel engines versus gasoline engines

The primary dominating benefit of the diesel engine versus its gasoline counterpart is found in the higher efficiency of the underlying reference cycle, namely the constant-pressure process for the diesel combustion and the constant-volume process for the gasoline combustion. It features theoretically a difference in the utilized compression ratio by the factor of 2 in favor for the diesel engine, offering higher internal combustion efficiencies. In addition, the engine is operated typically with air excess, which describes an additional source of higher fuel efficiency. Furthermore, the load control via delivered fuel quantity, the so-called control of mixture quality, concludes in another advantage against the load control via charge mass, the so-called control of mixture quantity. On the basis of these fundamentals, diesel engines offer overall efficiencies between 55% for large diesel engines (two-stroke for marine or genset applications) and 42% for small, high speed passenger car applications, whereas gasoline engines reach a maximum of 35–37%.

The corresponding internal efficiency differs even more, but because the higher utilized combustion pressures of diesel engines require a more rigid layout of the engine architecture, higher mechanical losses in terms of friction have to be considered, so that the effective numbers are slightly closer to each other.

2.3 Initial challenges

As a student of Professor Carl v. Linde at the University in Munich, Germany, R. Diesel studied the topic of calorific machines. Realizing that the steam engines as the predominant power source at that time were highly inefficient, he focused on the idea of turning the most efficient thermodynamic cycle, the Carnot cycle, into reality. The first fully functional diesel engine with a bore of 250 mm and a stroke of 400 mm delivered about 20 hp at 175 rpm. However, the development of the diesel engine was overshadowed by many challenges. Continuously reducing his expectation of a compression pressure from about 250 bar down to approximately 30 bar to ensure ignition of the induced fuel quickly became an issue. This level of gas compression was rather exceptional at that time. Further challenges were the cooling of the engine. As it was initially uncooled, it allowed only for short periods of operation at a time and thus delayed the development process severely. The application of a liquid cooling system drastically improved the situation and helped to accelerate the development. Another significant issue was the fueling. The introduction of fuel into the high pressure and temperature environment inside the cylinder became a huge problem. In addition, also the precise timing and metering of the fuel quantity turned into an obstacle that was very difficult and time consuming to overcome. As high pressure injection pumps and injection nozzles were not available yet, Rudolf Diesel decided to apply an air-assisted fuel injection design that allowed him to induce the fuel into the combustion chamber. However, despite his realization that the Carnot cycle was not the best cycle to represent his idea, he continued to make further improvements. These improvements led ultimately to the first diesel engine-powered truck in 1927 and the first passenger car in 1936. The main reason for the long development period were not only related to combustion system but also based on the lack of appropriate machinery to manufacture fuel injection pumps and injection nozzles at their relatively small size in larger quantities and at very low tolerances. Precision and lowest possible production tolerances while maintaining a high level of robustness still remain as some of the biggest challenges in diesel engine development even today.

2.4 Passenger car examples

The Mercedes-Benz 260 D in the W 138 series was the world's first series-production diesel passenger car. Seventy-five years ago, in February 1936, 50 years after the invention of the petrol-powered automobile by Carl Benz, Mercedes-Benz presented this revolutionary vehicle at the International Motorcycle and Automobile Exhibition in Berlin.

Its 2.6-L OM 138 four-cylinder engine with the Mercedes-Benz prechamber system and a Bosch injection pump produced 33 kW (45 hp) @ 3200 rpm, and was installed in the chassis of the petrol-powered Mercedes-Benz 200 with a long wheelbase. The Bosch four-plunger injection pump allowed engine speeds of up to 3000 rpm and ensured rapid fuel delivery.

Two years before, in November 1934, after experimenting with various diesel engines in Mercedes-Benz passenger cars, the engineers opted for a modified version of the well-proven six-cylinder in-line engine from the commercial vehicle sector. The result was a four-cylinder unit with a displacement of 2.6 L (bore $\times$ stroke was 90×100 mm). The new engine utilized the base engine's comparably smooth prechamber combustion process. The technical specifications featured OHV (Over-Head Valve) and a five-bearing crankshaft.

Series production of the model 260 D commenced at the end of 1935, and the world's first regular production diesel car was premiered in February 1936 at the International Motorcycle and Automobile Exhibition in Berlin. At an

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

average diesel fuel consumption of 9.5 L, a tank filling was initially sufficient for 400 km, and this increased to not <500 km or more after a model upgrade in 1937. This was not without significance considering the relative scarcity of filling stations at the time.

Even in 1936, the diesel engine in the model 260 D delivered impressive fuel economy: average consumption was slightly above 9 L of diesel per 100 km, considerably bettering the 13 L consumed by the petrol-powered model 200. Moreover, diesel fuel cost only 17 pfennigs per liter for holders of a passenger transport license in 1936: at the time, that was less than half the normal cost of petrol. Taxi drivers in particular immediately opted for this car, which was available in a spacious Pullman version with six seats right from the start.

A new momentum in the development of passenger car diesel brought PSA and VW (Volkswagen). The French OEM (original equipment manufacturer) engineered in 1968 for the first time a diesel engine transversely in a small compact car and VW followed 1976 by following the same approach in the Golf model. Mercedes-Benz implemented a year later for the first time a turbocharged diesel engine in a passenger car. The 300 SD should cut, with its 125 hp I-5 engine, the fuel consumption of Daimler models in the United States.

New developments such as four-valve technology, turbocharging, and intercooling supported the generation of the formerly sluggish aggregate to a more powerful propulsion unit. In the 1980s and 1990s, vehicles such as the VW Golf TD with 51 kW/70 hp allowed enjoying the attractive driving practice with a fuel consumption of 7 L/100 km.

The final breakthrough was achieved by the transition of the direct injection principle from large bore and heavy-duty engines to smaller displacements with higher rated power speeds. In 1988, the 1.9 L Fiat Croma TD with 68 kW/92 hp @ 4200 rpm was the first car that entered the market with this new technology. At about the same time, Ford released a similar engine concept in the light-commercial vehicle segment with 2.5 L displacement. Two years later, Audi introduced with the 2.5 L TDI (turbocharged direct injection) the first direct injection engine of a German manufacturer on the market. The introduced I-5 aggregate delivered initially 88 kW/120 hp. A little later, VW 1.9 L TDI's followed with 66 kW/90 hp and 81 kW/110 hp in various VW and Audi models.

2.5 Diesel combustion system

The diesel combustion system is usually defined by the interaction of the fuel injection sprays considering their corresponding orientation and the piston bowl. In practice, distinction is made between two main combustion system types. These are (a) the indirect-injected (IDI) and (b) the direct-injected (DI) systems as shown in Figure 1.

IDI applications distinguish further between the so-called prechamber and swirl chamber applications. In both cases, the entire combustion chamber is being divided into the main combustion chamber and a smaller separate chamber while both are connected via a smaller port opening. The distribution, shape, and orientation of the corresponding chamber volumes, their surface areas, the length and diameter of the port connecting the two chambers, and the location, orientation, and type of injection nozzle located in

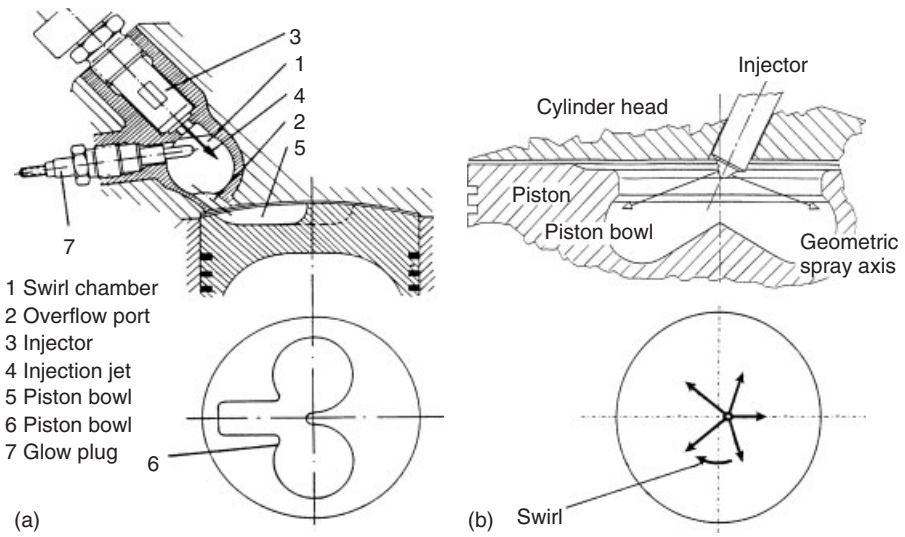


Figure 1. (a) IDI versus (b) DI combustion system (Pischinger, VKA RWTH Aachen, 2004). (Reproduced by permission of FEV Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the smaller chamber and the associated injection pressure (and injection rate shape) have a significant impact on the combustion efficiency of an IDI diesel engine. The airflow and the air velocity present in the small chamber support the combustion process. The volume of the pre or swirl chamber represents usually 25–60% of the entire combustion chamber volume and also houses the injection (pintle) nozzle, which injects fuel in form of a single injection jet at pressures between 200 and 300 bar. The surface temperatures in both the pre and the swirl chambers are high and, therefore, allow for a reduced ignition delay. After fuel injection has occurred, the combustion starts (supported by a glow plug in case of the swirl chamber) rather smooth because of the lack of oxygen (rich) inside these chambers and continues as the hot combustion gases and the remaining unburnt fuel propagates into the main chamber because of thermal expansion allowing the remaining fuel to be oxidized. One major benefit of the IDI combustion system is that the high surface temperatures and the consequential short ignition delay result into a smooth combustion and thus low noise level while being also more tolerant regarding fuel quality (cetane number) fluctuations. Furthermore, the injection pressure is on a relatively low level and thus does not require a high level of work to be generated, while the pump and nozzles are not exposed to excessive mechanical loadings. In addition, these systems allow for wider engine speed ranges and are less prone to nozzle coking because of the self-cleaning function of the pintle nozzle. However, the design requires detailed attention because of the very high thermal load at the exit of the pre/swirl chamber and the area of the piston, which is directly exposed to the hot expanding gas coming from the chamber. Both systems exhibit fairly high geometric compression ratios in the order of 19–22 and lend themselves to come close to a stoichiometric combustion because of the intensive air–fuel mixture formation inside the chamber.

The DI combustion system, which is the predominant process used in the industry nowadays, features only one main combustion chamber. The combustion chamber is defined by the bowl machined in the top of the piston and the fire deck of the cylinder head while a centrally located and preferably vertically oriented injector provides fuel for combustion via evenly spread out injection holes located in the nozzle tip. The mixture preparation, and thus ensuring ignition, is heavily influenced by the charge air motion inside the cylinder and the injection pressure available. Specially designed intake ports allow for a certain swirl level inside each cylinder promoting the ignition and combustion of the fuel injected. Over time, the swirl levels have continuously decreased because of the airflow losses associated to the generation of the swirl resulting ultimately in lower flow coefficients and thus cylinder filling. In that context, distinction is made between swirl and tangential intake ports, which are engineered to provide the desired swirl level at maximum air delivery. Owing to a continuous increase in fuel injection pressure levels from approximately 1350 bar at the beginning to now over 2000 bar, the injection system is able to support the required air–fuel mixture preparation and, therefore, allows for higher cylinder charge filling and ultimately higher engine power output. In parallel to that development, the amount of nozzle holes has increased from approximately 5 up to 8–9 for state-of-the-art high speed direct-injected (HSDI) diesel engines resulting in not only a more efficient combustion process but also lower engine-out emission levels. Usual compression ratios also have declined over time because of increasing fuel injection pressure and have come down from approximately 19–16. This also increases the sensitivity level regarding fuel quality (mainly cetane number) and its impact on engine startability at lower (cold) ambient temperatures and avoiding white smoke formation. Owing to the more compact combustion chamber and avoiding the throttle losses between the main and pre/swirlchambers as well as the heat losses, the DI technology provides a fuel consumption benefit of approximately 5–10% over the IDI technology while exhibiting higher in-cylinder peak pressures (up to 200 bar) and pressure gradients. However, with the ability for multiple injections and precise metering of small injection (pilot) quantities, the noise level of modern DI engines is close to the former IDI engines.

As a result, the state-of-the-art turbocharged and highly efficient and clean HSDI diesel engines have become the prime power source for diesel automotive applications.

2.6 Modern DI and turbo technology

The DI diesel engine always represented the most efficient internal combustion engine used in mobile applications. However, besides the aspect of superior fuel efficiency, also other parameters are relevant for homologation and market acceptance. Consequently, out of continuously strengthened emission standards, the improvement of the emission behavior of diesel engines depicted always the driver for enhanced or new technologies, especially regarding PM and NO_x emissions, as the legal norms were reduced regularly. In addition, the customer-related attributes such as performance and noise, vibration, and harshness (NVH) raised the need for upgraded capabilities of the engine and its main subsystems. With respect to an on-going engineering of the combustion process to meet increasing demands, base engine architecture and main subsystem (e.g., turbocharger and fuel injection system) specifications were part of numerous iterations of improvement and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

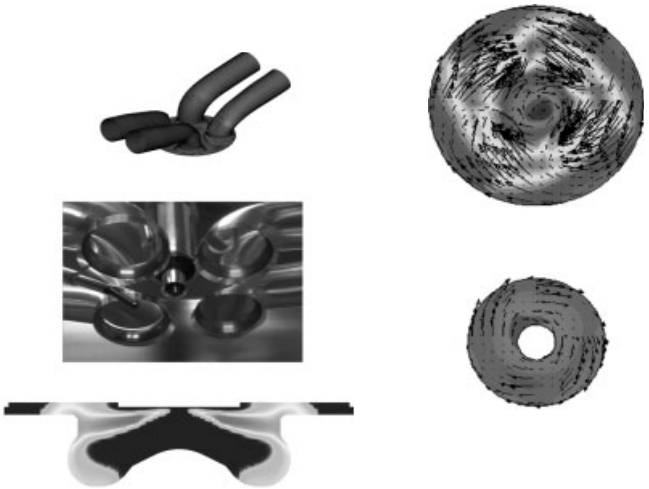


Figure 2. Key elements of modern DI-diesel technology (four-valve layout, bowl geometry, and homogenous flow field). (Reproduced by permission of FEV Inc.)

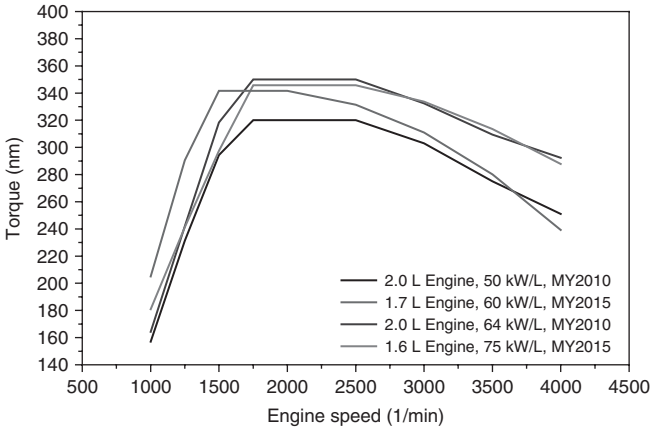


Figure 3. Turbocharger technology as the driver for engine downsizing. (Reproduced by permission of FEV Inc.)

refinement. On the basis of general design aspects such as increased peak firing pressure capability and central, perpendicular installation of injectors, mostly in combination with four-valve technology, a continuous adaptation of the combustion system parameters such as compression ratio and bowl geometry, charge motion, and spray pattern was performed (Figure 2).

A fundamental element in the success story of small HSDI diesel engines represents the charging system. In order to raise the power density and the specific torque, numerous improvements on the turbocharger specifications were realized. Beginning with increased efficiencies for conventional wastegate (WG) turbochargers over to refined VNT (variable nozzle turbocharger) technology to two-stage charging systems, a large bandwidth of charging devices for all engine sizes is meanwhile available to support the actual needs of all kinds of developments. The process of technology upgrade for the elementary diesel technology is continued to support actual trends such as downsizing or other advanced tendencies, as Figure 3 indicates. Furthermore, advanced turbocharger specifications are also assisting in improving part-load engine-out emission performance.

The second important factor for realization of the recent market success is given by the high pressure fuel injection system. Especially in balancing the demands from emission performance, fuel efficiency, and acoustic appearance of a diesel engine, actual common-rail technology delivered a superior contribution. With continuously rising pressure levels up to actually 2000 bar and in conjunction with modern piezo-control, an extremely capable and reliable fuel injection was developed, representing currently the forefront regarding system performance and accuracy (Figure 4).

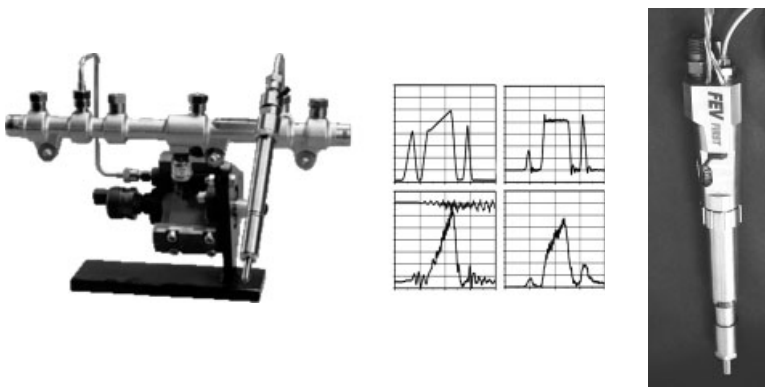


Figure 4. Advanced prototype high pressure common-rail system (FEV FIRST—flexible injection rate shaping tool). (Reproduced by permission of FEV Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Owing to the on-going tightening of the legal and market demands, engineering activities are continued in all areas of DI diesel technology.

3 CURRENT AND FUTURE TECHNOLOGIES

3.1 Current and future emissions legislations (EU/US/JP/BRIC)

The currently valid emission regulation is still related to local and national standards as shown in Figure 5.

In the United States, still exists a tendency toward further minimization of engine-out pollutants, concentrating on lowering PM, NO_x, and HC emissions. Despite the fact that already today's Tier 2 Bin 5 emission limits display the most severe emission norm in the world, a further step of tightening is planned with the introduction of the new LEV 3 (low emission vehicle) legislation with its implementation to be completed by 2025. In Europe, with the actual main market for passenger car diesel engines at an average market penetration of ~50%, a reverse trend is actually seen. While having established already reasonable low emission standards with EU-6 in 2014, a higher focus is actually placed on reduction of greenhouse gases (GHG). Besides a further moderate strengthening of emissions, strict legal demands are formulated for CO_2 emission behavior, as shown in Figure 6. The Japanese emission regulation (JP Post New Long Term, JP PNLT) is strongly related to domestic conditions and remains regarding complexity between European and US standards.

The emerging countries such as China, India, Brazil, and Russia follow, in most cases, the European regulation with a slight temporal delay and small adaptations to local conditions.

For all countries, it is valid that starting with stringent emissions standards such as EURO-5/-6, US Tier 2 Bin 5, or JP PNLT, adequate norms for appropriate fuel quality, especially regarding sulfur content, have to be enacted.

3.2 Design and materials

Since its inception, the combustion engine, independent of gasoline or diesel as its main representatives, has evolved significantly not only in its design but also with respect to the materials used; starting with cast iron (CI) over steel to more advanced materials such as aluminum alloys up to more exotic materials such as titanium and magnesium, as they are primarily used in racing applications. In the following, the main attributes of these materials in context of engine component design will be discussed.

3.2.1 Cylinder block

Being exposed to a variety of mass and gas forces, torsionals, and engine mounting forces, the cylinder block (crankcase) represents one of the main engine components. Housing the cranktrain (crankshaft, rods, and pistons), as

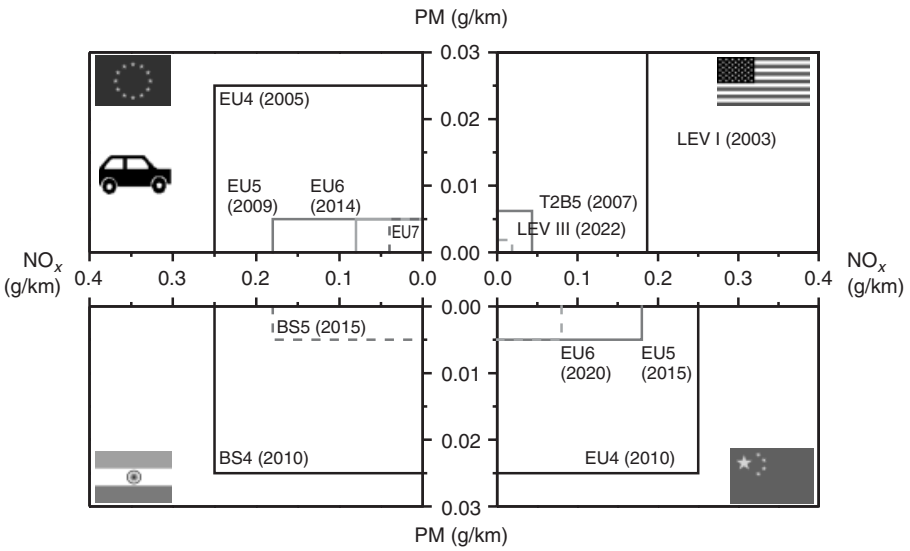


Figure 5. Actual emission regulation in selected major markets and prognosis. (Reproduced by permission of FEV Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

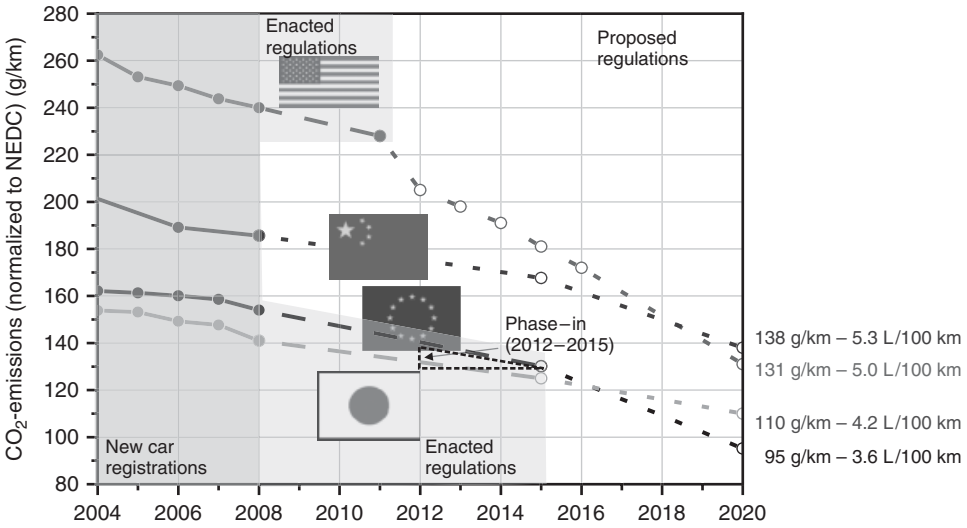


Figure 6. Current global CO_2 regulation in major markets and outlook. (Reproduced by permission of FEV Inc.)

well as coolant, oil, and a variety of accessories as well as interfacing with the cylinder head, oil pan, and transmission, it requires a thoughtful design. As a long-time standard material, CI meanwhile has been surpassed by a variety of different aluminum alloys for gasoline and diesel engine cylinder blocks. Owing to the significantly lower weight, higher heat conductivity, and machining properties, aluminum has become a competitor to CI as the prime block material. Despite its higher price and lower strength, this tendency is continuing even for diesel engines. In order to be able to compete with CI though, design decisions such as open versus closed deck design and deep skirt versus bedplate (girdle) design to stiffen up the block are of significant importance.

Usually, distinction is made between CI and aluminum liners. For CI liners, further distinction is made between pressed-in or slip-fit liners (either dry or wet) and cast-in liners. For aluminum blocks, distinction is made between all-aluminum configurations, depicting either an over-eutectic aluminum alloy or a special liner coating and inserted aluminum liners (slip-fit or cast-in). Owing to cost and durability reasons, the most used configuration in the market is represented by the cast-in CI liner in aluminum blocks.

In addition, the implementation of additional ribs to minimize deformation during operation, as well as for noise reasons, is a must for current production applications.

One additional alternative is represented by compacted graphite iron (CGI). Representing a hybrid between aluminum and CI by depicting good strength at a lower density compared to CI, it has been used for some automotive production applications, where emphasis was placed on lightweight and high specific power. However, owing to difficulties in machining this specific material, high cost has precluded CGI from becoming a standard material. For automotive diesel engines, considering the current threshold of 180 bar in-cylinder peak pressure, one can expect to reach and potentially exceed the 200 bar mark within the next years because of the continuous demand for increased specific power and torque. Using state-of-the-art tools to analytically optimize and validate the structural integrity of an engine block, the main target in the future will be to lower weight while simultaneously increase strength and stiffness.

3.2.2 Cylinder head

Similar to the cylinder block, aluminum has become the standard cylinder head material for gasoline and diesel engines. Only for diesel engine designs exploiting peak pressures exceeding 180 bar, the application of alternative materials such as CI or CGI become an option. While CGI owing to its machining challenges and its material price has become an expensive alternative though, other smart designs using aluminum with an intermediate deck design have potential to comply with future demands.

While the four-valve cross-flow liquid-cooled design has become a standard in the automotive industry for most diesel powertrains, some exceptions using two or three valves can be found as well. Considering further downsizing resulting in smaller bore diameters, some two-valve concepts bear a lot of potential for highly efficient combustion systems at low cost. In addition, with decreasing bore diameters, the accommodation of valves, injectors,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and glow plugs for automotive diesel applications becomes more and more challenging from a thermal and mechanical perspective.

3.2.3 Cranktrain

The cranktrain, consisting of the crankshaft, connecting rods, and pistons, is exposed to gas and mass forces, as well as torsionals in case of the crankshaft. Crankshafts for higher specific power engines are usually forged to ensure high strength and stiffness, whereas cast crankshafts at a lower cost are less robust. Compact design with highest possible pin overlaps in Vee engines and shortest possible cylinder-to-cylinder distance for inline engines are highly desired. To reduce friction, the main and connecting rod bearings need to be minimized in width and diameter to just satisfy all criteria relevant to achieve full engine life.

Similar to crankshafts, connecting rods are also either forged or cast depicting the same trade-off regarding strength, stiffness, and cost. To reduce complexity regarding the machining of the big end, sintered metal connecting rods with fractured bearing caps have been successfully introduced over the past decade. This process eliminates the machining of that part of the components and assures perfect alignment of the rod with the cap because of the individual cracking characteristics. The pistons, for basically all automotive diesel applications, are actually aluminum based because of its excellent heat transfer properties. To cope with the high thermal and mechanical stresses, a steel carrier for the top steel compression ring is incorporated in the top section of the piston while oil jets improve cooling on the bottom side of the piston for higher specific power applications. Anodizing around the rim of the piston bowl helps to mitigate the potential for cracking while being exposed to temperatures up to 350°C. The application of low friction coatings on the skirt of the piston helps to reduce the overall engine friction level, with the piston group being a significant contributor. Latest developments offered some fuel-saving potential with advanced steel piston design, which are currently introduced in first applications.

3.2.4 Valvetrain

The majority of automotive diesel engines employ dual overhead camshaft (DOHC) four-valve applications, while the application of belt versus chain as a driving mechanism is highly depending on the manufacturers' preference. Owing to high cost, gear-driven valvetrains represent an exception for automotive applications. To minimize friction, in most cases, (roller) finger followers or buckets are applied. To reduce maintenance cost, hydraulic valve lash adjusters are incorporated either into the cylinder head or directly into the bucket itself. Ensuring low seating velocities in favor of low NVH requires optimized camshaft profiles considering the kinematics of the entire valvetrain and the components involved. To further optimize the combustion system, variable valve lift on the intake side and potentially cam phasing on the exhaust side to improve catalyst light-off, technologies well known from gasoline engines, will find their application more and more in modern diesel engines. The design, material, and manufacturing of the camshaft depends on the application. While the cast camshafts represent the conventional designs, build camshafts using steel tubing and lobes attached via press fit are a lightweight solution that is found more and more.

3.3 Fuel injection system

In the challenging competition of all alternatives in the late 1990s, ultimately the common-rail injection emerged as the final winner. The dominating merits of this innovation, where the fuel is stored at high pressure in one (for Vee-type engines in two) common supply rail ("common rail") for all cylinders of an engine arrangement, were:

- largely speed-independent injection pressure;
- multiinjection capability at constant nominal injection pressure.

As a function of time, with growing demands from market and legislation, enhanced capabilities and increased functionalities regarding

- higher injection pressure,
- accurate metering, especially concerning small quantities,
- fast opening and closing behaviors,
- increased number of injections,
- short dwell times,
- raised hydraulic efficiency,
- repeatability, and
- long-term stability.

were requested.

Actual premium systems are piezo-controlled and operate with injection pressures of 2000 bar. Future development trends aim at a maximum injection pressure of 2500 bar in the coming years, eventually going to higher values up to approximately 3000 bar at the end of the next decade. In the high volume market, the advantages of the solenoid-controlled layout dominate because of cost and significantly improved, competitive performance values compared to the piezo-controlled injectors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

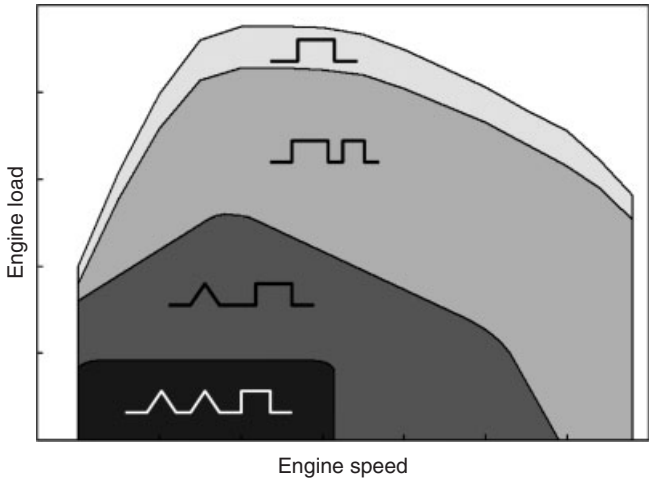


Figure 7. Representative injection characteristics for adsorption mode as function of engine load and speed. (Reproduced by permission of FEV Inc.)

According to the individual needs at various operating points and depending on operational modes such as cold start, warm-up, or regeneration strategies, different injection characteristics are preferred as function of engine speed and load, as shown in Figure 7.

3.4 Combustion system

The selection and optimization of the combustion system represent the heart of modern diesel engines, as all relevant factors such as performance, efficiency, emissions, noise, and reliability strongly depend on the characteristics of the incorporated combustion system. The main objective of the combustion system definition is given by the proper matching of charge motion, spray pattern, and combustion chamber geometry under all operating conditions.

By the proper matching of fuel injection (injection pressure, number of sprays, and spray hole diameter), charge conditions (motion and density), and piston bowl geometry, all fundamental elements of diesel combustion regarding fuel atomization, mixture formation, air utilization, flame propagation, and fuel burn-off are described.

Owing to the needs of continuous emission minimization, significant development effort was performed to lower engine-out emission levels. Main technologies and contributors in the evolution and adaptation of the combustion to today's needs were:

- smaller spray hole orifices in conjunction with increased injection pressure;
- higher number of spray holes and refined spray hole geometries (Figure 8) (e.g., hydro-erosive and conical spray holes → k-factor[1]);
- utilization of multiinjection strategies;
- rate shaping functionalities;
- higher boosting levels due to improved charging systems;
- optimized port geometries regarding flow and swirl characteristics;
- lowered compression ratios and optimized bowl geometries (reduced parasitic volumes, increased ratio of bowl volume to compression volume → higher k-factor[2]);
- EGR (exhaust gas recirculation) (noncooled/cooled HP-EGR; noncooled/cooled LP-EGR);
- valvetrain variabilities.

Owing to conflicting requirements in the entire operational map, for example, between cold start/light load and full load with respect to ignition conditions, best balanced compromises have to be worked out. Owing to limited capabilities of previous fuel injection systems, former engine variants required higher compression ratios to ensure appropriate cold start behavior and good mixture formation. With improved performance and enlarged functionalities of today's common-rail systems, a substantial reduction of compression ratios to values of minimal 14.5 : 1 is achieved in production, resulting in higher degrees of mixture homogenization and controllable peak firing pressures. The trend to lower compression ratios enabled new bowl geometries, which featured better ratios between bowl volume in relation to compression volume, described in the k-factor*. Besides the continuously increasing injection pressures of advanced high pressure fuel injection systems reaching 2500 bar in the near future, also further merits in the hydraulic system technology were introduced. In order to support improved mixture preparation, the number of spray holes raised from five in the early years of DI diesel engines to seven or eight for the majority of engines; meanwhile, even 10 are found in mass production. Furthermore, the introduction of conical spray holes, specified by the k-factor** index, resulted in a significant improvement of spray penetration and fuel atomization. The demand for minimized engine-out NO_x emissions increased EGR rates, which can be recirculated in the ways, upstream of the turbine as high pressure EGR, downstream of the turbine and diesel particulate filter (DPF) as low pressure EGR, or as internal EGR via variable valve timing. The tendency toward higher ratings in conjunction with raised low-end torques affected the design of valvetrain functionalities similar to the needs concerning variable charge motion. Furthermore, the requirement of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

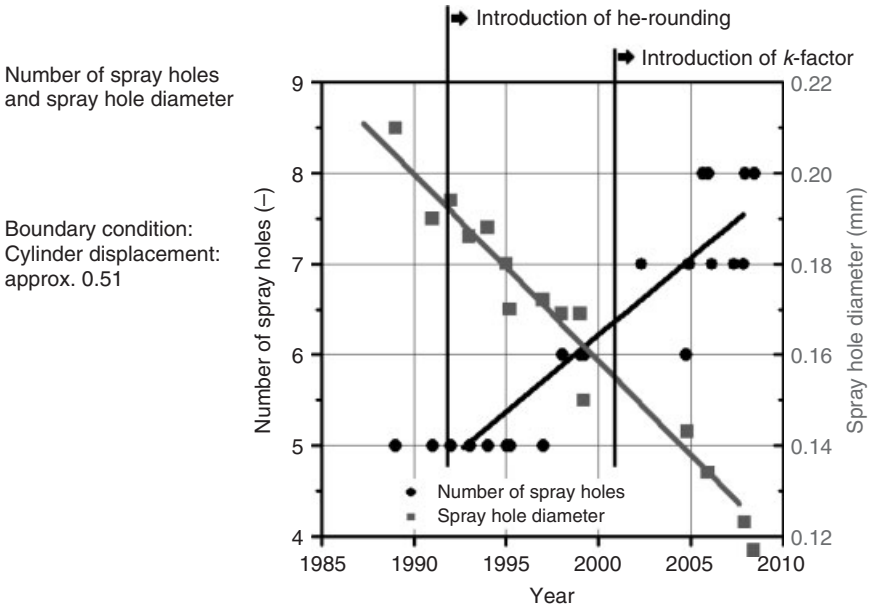


Figure 8. Evolution of spray hole number and spray hole diameter. (Reproduced by permission of FEV Inc.)

adequate temperature levels for good conversion efficiencies of the aftertreatment devices caused additional pressure on increasing valvetrain variabilities. Consequently, various technical solutions concerning valvetrain functionalities are actually launched in the market, at the time being mostly focused on one parameter at a given engine. More powerful, combined solutions are expected soon.

While driving the combustion system to the limits, stabilization measures for production scattering and ambient disturbance have to be implemented. A very powerful technology is represented by the glow plug-integrated pressure sensor. An online sensing of the pressure inside of the combustion chamber with post-processing and automatic correction of the combustion phasing ensures combustion stability with low gaseous emissions (HC and CO) and high efficiency. A further benefit is applied in the field of varying fuel qualities, where the glow plug sensor can adapt the injection scheme according to the corresponding cetane number.

3.5 Air management system

With the introduction of turbochargers, the operating range and power output of the diesel engine could be expanded significantly. Meanwhile, modern turbochargers allowing for precise boost pressure control via either WG or variable turbine geometry (VTG) are standard. To enhance further the engine performance, a two-stage sequential turbocharging has been introduced into the marketplace. These systems, employing different combinations of WG, VGT (variable geometry turbocharger), or fixed geometry turbochargers, enable very high low end torque while providing high power output making the automotive diesel engine very competitive with state-of-the-art gasoline engines. However, such a system requires refined controls and calibration to realize its true benefits. For Vee engines, twin turbo arrangements (one per cylinder bank) are also thinkable—potentially, even with a third larger turbocharger in series allowing for high power output bypassing both high pressure twin turbos.

Owing to continuously more stringent NO_x emissions regulations, high pressure EGR has become a standard feature in all automotive modern diesel engines. Routing a part of the exhaust upstream the turbine back to the intake system also influences the choice of turbine and compressor because of changes in exhaust and airflow across the corresponding impellers. To address most stringent NO_x emission targets such as Tier 2/LEV 2 in the United States, combined high and low pressure EGR systems have been developed and placed into production. In that case, where the high pressure EGR system remains unchanged, an additional low pressure system is added. This low pressure system reroutes exhaust from after the DPF to the inlet of the compressor of the turbocharger as shown in Figure 9.

To minimize thermal throttling in both cases, the exhaust is being routed through an EGR cooler, which is usually cooled by the engine coolant. As a result, the volumetric flow of the EGR is reduced and ultimately, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

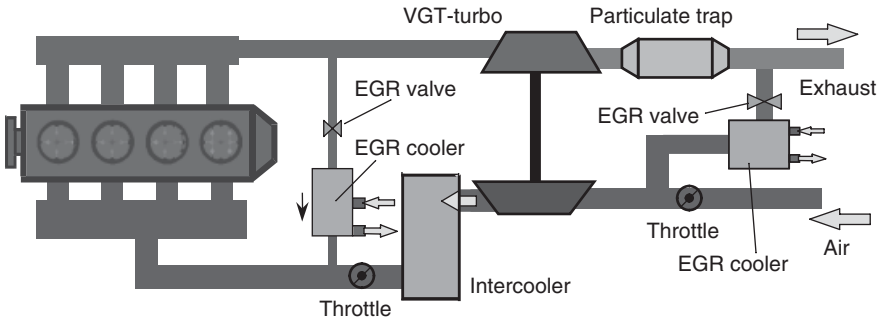


Figure 9. High and low pressure EGR system concepts. (Reproduced by permission of FEV Inc.)

cylinder charge and the amount of oxygen available for the combustion of the fuel increased. As the emissions need to be met over the expected (full useful) lifetime of the vehicle, the minimum cooling power of the EGR cooler has to be maintained. Challenges such as cooler fouling because of condensation of hydrocarbons forming a hard, nonremovable layer on the inside of the cooler leading to increased EGR temperatures are being addressed by the design of the cooler, materials used, cleaning strategies, and bypass operation of the cooler during operating regimes exhibiting higher HC emissions (e.g., cold start and DPF regenerations).

The amount of EGR recirculated is controlled via an EGR valve, which, in most cases, is a poppet-type valve. Control of boost pressure and EGR, which influence each other, has evolved over time. Meanwhile, multivariable controllers (multiple input—multiple output: MIMO) are being applied to accurately control boost and EGR within the entire operating map. Further challenges are primarily related to transient control optimization.

3.6 Exhaust system (DOC, DPF, LNT, SCR, and so on)

In order to meet the increasing demands in emission control, already in the early 1990s, the DOC (diesel oxidation catalyst) or a two-way catalyst was introduced. As the diesel engine typically operates with air excess, only oxidation processes regarding HC and CO can be supported. Main purpose of the initial usage of the DOC was given by the reduction of PM mass because of oxidation of wet elements on the PM, namely accumulated HCs.

In 2005, the installation of DPFs was renewed in Europe by PSA according to Euro-4 norms, after unsatisfactory previous experience in the United Sates with older technologies. Despite the fact that most OEMs targeted Euro-4 compliance without the usage of DPF technology, the overwhelming public response forced all OEMs to follow this direction. While PSA favored a fuel-borne assisted DPF regeneration with lowered temperature levels for soot burn off, all other OEMs concentrated on realizations without additional fluid on board of the vehicle. With enacting the Euro-5 norm and the US Tier 2 standards, the wall-flow DPF technology became mandatory for all applications. In order to improve the emission behavior of in-field vehicles, retrofit systems on the base of open DPF technologies were launched based on fiscal incentives. Depending on base combustion system performance and package boundaries, some companies installed a special fuel injector in the exhaust line to lower the risk of in-cylinder after injection regarding fuel-in-oil consequences and oil dilution. As a function of time, new substrate materials were introduced owing to cost, performance, and robustness aspects. Main substrate types are SiC, AlTi, and cordierite.

With the implementation of wall-flow DPF devices, the tailpipe PM emission level reached levels below ambient air qualities and even future particle number limits can be ensured, so that future emission legislations focused on NO_x emissions. Despite a massive improvement on in-cylinder NO_x-reducing technologies such as amplified EGR rates and homogenized low temperature combustion, future applications require highly sophisticated $DeNO_x$ aftertreatment technology to meet upcoming emission regulations. While SCR (selective catalytic reduction) technology is clearly the selected path for heavy-duty diesel engines, the light-duty diesel engines use two different systems. Besides the SCR technology, the lean NO_x trap (LNT) or NO_x storage catalyst (NSC) also offers substantial benefits regarding individual parameters in the global assessment matrix. The SCR technology is based on a continuous NO_x conversion process in the exhaust line, while applying ammonia derived from an aqueous urea solution once appropriate temperature levels have been realized. In contrast to that, the LNT performs in a discontinuous way. The working principle of the LNT is based on

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a certain adsorption phase, followed by a frequent regeneration phase. The challenge is given by the realization of the regeneration conditions, where the engine has to be operated under rich conditions (relative A/F ratio <1) and the fast transition from normal lean to rich operation and back to lean operation again. Typical intervals are factors of >25 between lean and rich operations. Further challenges are found in the sensitivity of LNT technology regarding sulfur coming from the fuel and oil in the engine. Owing to the pronounced accumulation of sulfur in the LNT, frequent desulfation events have to take place in order to maintain NO_x conversion efficiency. For reliable release of sulfur out of the LNT, quite high temperatures in conjunction with enriched exhaust composition have to be achieved. Therefore, typically, desulfation phases are combined with DPF regeneration events; because, during that phase, elevated thermal boundary conditions are present in the exhaust system.

3.7 Controls and sensors

All of today's automotive diesel engine control systems are torque based, meaning that a torque request indicated by pedal position is turned into a fuel injection quantity, providing that requested torque independent of varying accessory drive loadings caused by, for example, increased alternator current demand and air conditioning. The amount of calibratable parameters for an entire automotive diesel vehicle including the powertrain and transmission can be up to 30,000 parameters, with continuously increasing tendency. To appropriately populate (calibrate) all of these parameters within a reasonable timeframe, satisfying targets in terms of performance, emissions, drivability, fuel consumption, NVH, and so on, very detailed procedures, highly sophisticated tools, and calibration experience are required.

To ease the calibration effort and improve the quality of the calibration and thus engine operation, a variety of sensors have been developed and introduced into production. Cam and crank sensors to determine engine speed and actual position of each cylinder relative to each other, also on an absolute basis, are crucial. Temperature and pressure measurements for boost conditions in the intake manifold are also standard. Sensors related to the aftertreatment systems have also evolved over time. There, temperature, lambda, NO_x, and differential pressure sensors for DPF loading monitoring are standard. For common-rail injection systems, pressure sensors are used to control the rail (injection) pressure in a closed-loop manner compared to commanded rail pressure as calibrated in the rail pressure map. Other sensors monitoring parameters such as coolant temperature for thermostat control and in-cylinder combustion pressures for fuel injection parameter optimization based on, for example, fuel quality variations are also used. Future sensor developments include ammonia (NH_3) sensors to monitor the ammonia break-through downstream the SCR catalyst optimizing the urea injection quantities injected into the exhaust as well as radio frequency (RF) sensor technology to more accurately determine the loading level of a DPF compared to a differential pressure sensor.

To reduce cost and hardware complexity, the development of the so-called virtual sensors has become more and more a part of the engine control system and the associated sensor infrastructure. Using the output of several sensors, another sensor can be modeled virtually providing important information. This becomes especially attractive for sensors that are expensive, difficult to manufacture, have a critical lifetime, or are even suspect to undesired cross-sensitivities. To continuously improve engine operation especially under highly transient conditions, such fast reacting virtual sensors can significantly improve the outcome.

All of these sensors also play an important role regarding on-board diagnostics (OBD), which is a regulatory certification requirement demanding continuous diagnosis of all relevant sensors to ensure proper engine operation (functional monitors) and all subsystems potentially affecting emissions (threshold monitors).

3.8 Fuels and alternative fuels/biofuels

Owing to the continuous decline in fossil fuel reserves, it is mandatory to develop meaningful alternatives to the conventional diesel fuel currently available. On the basis of significant fluctuations of some of the relevant diesel fuel properties such as cetane number and sulfur concentration, each of the relevant world markets have established their own fuel standard/norm that they adhere to. Nevertheless, the desire to develop and market alternative fuels and biofuels has resulted in a variety of products, which to some extent have been introduced while the majority is still under development.

Biodiesel, often also referred to as *FAME* (*fatty acid methyl ester*), using rapeseed (RME, rape methyl ester) or soybeans (SME, soy methyl ester) as a feedstock has become a common fuel for blending in the European Union or the United States, respectively. Representing a share of up to 20% (B20) allows reducing the use of fossil-based fuels. However, as for conventional diesel fuel, the introduction of tight biodiesel fuel standards (e.g., ASTM D6751) to ensure adequate quality is critical, as engine components, mostly fuel injection system related,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

can potentially be damaged. For most engine applications, significant reductions of CO, HC, and particulate matter have been observed while NO_x at the same time have been increasing slightly yet almost linearly as a function of biodiesel concentration.

In order to not use feedstock, which compete with foods consumed by humans, another group of the so-called second-generation biofuels have been established. These fuels use feedstocks that have no direct impact on food cost or availability. One particular example is algae-based fuels. As a part of the production process, algae are grown in a controlled environment feeding off gases such as CO_2 and SO_x under direct sunlight. As a result, while producing the algae, which can then be processed into biofuel, this technology helps to reduce the overall carbon footprint. Research continues to improve the process in terms of algal growth and the associated fuel processing.

Another process utilizes natural gas to develop diesel-like fuel as a part of the gas-to-liquid (GTL). This fuel exhibits significant advantages over conventional diesel fuel, as it has a higher net heat of combustion, practically no sulfur content, very low aromatics, and a drastically higher cetane number. At this point, only cost and production volume prevent a deeper penetration into the market while blending with conventional diesel-like fuels with biodiesel also represents a meaningful approach. Similar attempts are made focusing on biomass-to-liquid (BTL) and coal-to-liquid (CTL).

Other opportunities exist with feedstocks such as jatropha oil, palm oil, canola oil, and other plant or animal-based fats. In addition, some studies are also investigating the use of supplemental fuels such as ethanol, which is already used in gasoline fuels in concentrations of up to 85% (E85). Promising features such as improved in-cylinder cooling resulting in lower NO_x emissions and eventually allowing for even higher engine loads have caused interest despite the need for a separate fuel and injection system.

Beyond biofuels, artificial or the so-called tailor-made fuels from biomass (e.g., 2-methyltetrahydrofurane) exhibiting similar results compared to common diesel fuels are also subject to laboratory investigations worldwide. While promising fuels have already been identified, the cost to produce them in the required quantities is still exceeding the limits making a short- or even mid-term introduction not feasible.

3.9 Noise, vibration, and harshness

The noise level of a combustion engine can be broken down into a variety of different categories as depicted in Figure 10. According to this flowchart, each individual category can be optimized for lowest possible noise levels focusing on the parameters influencing it.

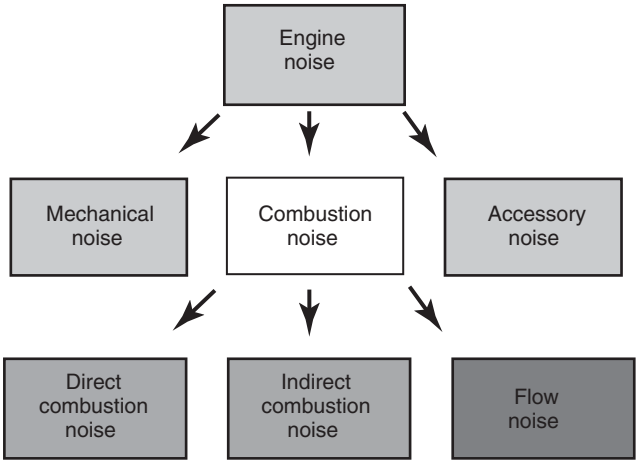


Figure 10. Engine noise categories. (Reproduced by permission of FEV Inc.)

While direct combustion noise contains the total excitation (force and impact) by gas force and all other forces inside the engine, which have a similar course versus time (allowing for a certain cylinder pressure gradient correlation), the indirect combustion noise contains the total excitation by rotary and piston normal forces, which have a different course versus time compared to gas force (e.g., piston slap, crankshaft torsion, and cyclic speed fluctuations caused noise such as gear rattle noise). The flow noise is excited by air mass flow and it is radiated by the intake and exhaust systems.

For a diesel engine, part load noise increase is typically caused by direct combustion noise. The indirect combustion noise and flow noise are increasing with load, caused by the increase of air mass flow and peak pressures. The increase of boost and thus in-cylinder charge pressure typically increases indirect combustion noise because of higher peak pressures, but decreases direct combustion noise based on increasing combustion chamber pressure and temperature level, reduced ignition delay, and pressure gradient.

With the introduction of EGR, the combustion noise can be reduced significantly because of a higher inert gas fraction inside the combustion chamber leading to a reduced premixed charge during the ignition period. The application of EGR in combination with modern fuel injection systems, which allow for one or even more subsequent small pilot injection quantities, results in an even shorter ignition delay leading to an even further reduced direct combustion noise level close to gasoline engine noise levels under warm operating conditions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Further means to reduce noise are reduced moving masses, bearing and piston clearances, stiff cranktrain and engine block/cylinder head architecture (ladder frame or bedplate design), engine downspeeding (reduced maximum engine speed), short ignition delay (soft initiation of combustion), lower valve-seating velocities, noise insulation (via seals or gaskets), and potential encapsulation.

In addition to that, optimizations regarding the intake and exhaust system noise levels attribute to the overall beneficial noise level of a state-of-the-art automotive diesel engine.

3.10 Thermal management

Thermal engine management has become one of the most critical engine management items. Avoiding overheating of the most vulnerable and thus critical engine components such as pistons, cylinder head, valves, and turbochargers mitigating excessive thermal stresses and ultimately engine failure has become a routine. Applying piston oil cooling jets, providing sufficient cooling flow at appropriate temperatures in combination with optimized material selections usually result in thermally very robust engine architectures.

Other challenges, however, are represented by the reduction of exhaust gas temperatures because of increased use of EGR and continuous improvement of thermal efficiency and, therefore, reduced exhaust gas heat losses. In combination with that, vehicle certification cycles such as the FTP-75 (federal test procedure) or NEDC (new European driving cycle) also promote rather low exhaust gas temperatures because of the low vehicle velocities and idle phases especially at the beginning of the test cycles leading to low engine speed and loads. Despite the improvement of catalytic light-off temperatures down to approximately 180–200°C, catalyst heating by increasing the exhaust gas enthalpy during the cold start period of the corresponding certification cycles is required to efficiently reduce CO, HC, and NO_x emissions early on in the test cycle. This catalyst heating is usually achieved by fuel injection timing retardation and can be further complemented by advanced exhaust cam phasing as far as available. Intelligent thermal management also includes taking advantage of the thermal inertia of some of the aftertreatment system components, which can vary drastically based on the difference in size, cell density, porosity, material, and their location within the exhaust system. Monitoring the relevant temperatures of the DOC, LNT/SCR catalyst, and the DPF helps reduce the additional fuel quantities required to achieve the necessary conversion rates of the regulated exhaust constituents.

In addition to that, fast heat-up of the engine coolant and oil is desired as well to help increase exhaust gas temperatures and lower fuel consumption. In that context, applications of technologies such as split cooling and electric coolant pumps are beneficial.

ENDNOTES

1. Nozzle *k*-Factor: (inner diameter−outer diameter)/10 with both diameters (μm).
2. Combustion chamber *k*-factor: (bowl volume*)/compression volume—*bowl volume = compression−parasitic volume.

FURTHER READING

Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals*, McGraw-Hill, Inc., New York, USA.

Mollenhauer, K. (1997) *Handbuch Dieselmotoren*, Springer Verlag, Germany.

Pischinger, S. (2010), Internal Combustion Engines. Lecture Notes, RWTH Aachen, Aachen, Germany.

Stone, R. (1995) *Introduction to Internal Combustion Engines*, SAE, USA.

Taylor, C.F. (1985) *The Internal Combustion Engine in Theory and Practice*, MIT Press, Cambridge, USA.

Van Basshuysen, R. and Schaefer, F. (2004) *Lexikon Motorentechnik*, Vieweg Verlag, Germany.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto148
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Cooling Systems

Stephen F. Bowyer, Haralabos Triantafyllidis, and Mark E. Case
FEV Inc., Auburn Hills, MI, USA

1 Introduction 1
2 Energy Balance 2
3 Fundamentals of Cooling Systems 3
4 Liquid Cooling System Components 4
5 Engine Cooling 10
6 Advanced Cooling Systems 16

1 INTRODUCTION

Heat release from the combustion process will have negative effects on component durability, engine performance, and the combustion process if the engine is not properly cooled. Modern automotive cooling systems have been developed to address several of these issues during the vehicles operation. The following sections outline durability, performance, and emissions characteristics of automotive engines that are addressed by the engine cooling system.

1.1 Mechanical durability

1.1.1 Low cycle fatigue

Combustion gas temperatures inside modern combustion engines significantly degrade the mechanical strength of the components while simultaneously imparting a thermal strain on the components and assemblies. Yielding of the material under high compressive strain caused by thermal stresses can cause catastrophic engine failures by low cycle fatigue. Thermal yielding while in operation and subsequent cooling after engine shutdown imparts significant tensile strain loads in the combustion face of the cylinder head. Each time the engine is heated, further yielding occurs, and, over time, the resultant stresses can become significant enough to cause a crack to form in the base material of the cylinder head.

1.1.2 Thermal distortion

High component temperatures can create sealing issues throughout the engine because of the thermal strains of the components. Combustion gas sealing at the bore interface to the piston is highly dependent on the thermal distortions in the cylinder block. One countermeasure to bore distortion is increased piston ring tension to ensure sufficient contact pressure with the bore wall. This countermeasure, however, is not ideal as increased friction and increased fuel consumption will be observed. A more productive countermeasure is improving the cooling system to provide a more round cylinder bore during operation.

Another critical combustion seal is the valve to the valve seat that is also heavily influenced by the roundness of the valve seat as well as the true position of the valve seat with respect to the projected centerline of the valve guide. Misaligned or out-of-round valve seats are typically caused by poor or inconsistent cooling within the cylinder head.

In addition to the combustion seals already discussed, the engine contains lubrication, cooling, and crankcase gas seals throughout the engine that are affected by the distortion of various engine components. For example, the cylinder head gasket interface typically transfers pressurized oil, low pressure return oil, coolant, and ventilation gases between the cylinder block and the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

cylinder head. Thermal distortion to either of these components can compromise the head gasket's ability to seal the passages resulting in warranty claims and customer dissatisfaction.

1.1.3 Lubrication

The heat from the combustion process is also transferred through the engine components to the lubrication system. Failure to keep the engine's lubricant temperature to an acceptable level will cause a degradation of the lubricant and cause the lubricant to coke or sludge. This will, in turn, lead to further catastrophic failures of moving parts within the engine as the ability of the lubricant to reduce friction and remove heat will be compromised.

1.2 Combustion gas temperatures

As the fresh air charge travels through the engine for the next combustion process, it absorbs heat as it makes its way through the inlet duct, cylinder head, and into the cylinder. The heat absorbed raises the temperature of the final charge and reduces the density of the air captured in the cylinder. This, in turn, reduces the volumetric efficiency of the engine, thereby reducing the engine's power density.

For gasoline engines, a second detriment is the reduced knock margin because of the increased charge temperatures or hot surfaces in the combustion chamber that can serve as knock initiation points. Several countermeasures are possible, such as retarding spark, enrichment, or throttling the engine but all have a negative effect on engine efficiency and should be avoided, as much as possible, by minimizing the heat transferred into the fresh air charge.

For diesel engines, high in-cylinder temperatures contribute to the formation of NO*x* emissions that can be difficult and expensive to manage with aftertreatment systems. Keeping components that make up the combustion chamber cool reduces NO*x* emissions and allows for smaller, less costly, and more efficient exhaust aftertreatment systems.

2 ENERGY BALANCE

Energy from the fuel consumed by the engine can be split into several categories and is highly load and speed dependent with the main categories being shaft power, exhaust heat, heat rejected to the cooling system, and miscellaneous other losses including radiation and remaining chemical energy from incomplete combustion.

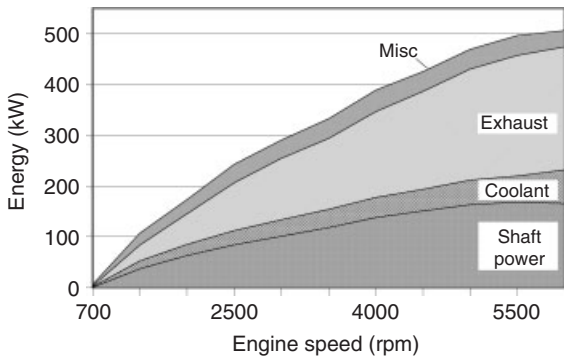


Figure 1. Energy diagram for a turbocharged direct injected gasoline engine at full load. (Reproduced by permission of FEV Inc.)

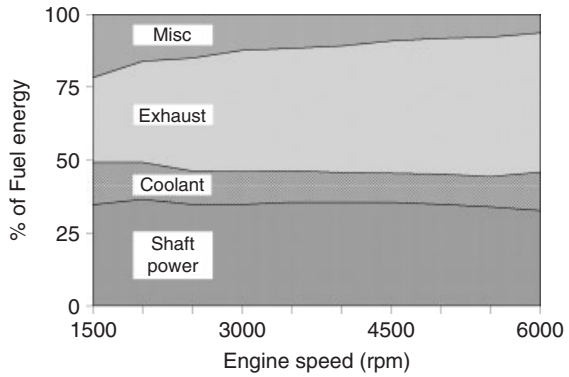


Figure 2. Energy split for a turbocharged direct injected gasoline engine at full load. (Reproduced by permission of FEV Inc.)

Figures 1 and 2 show the heat rejection graphs from a turbocharged direct injection gasoline engine at full load. Here, we can see that the heat rejected to the cooling system is varied within a tight band between 10% and 15% of fuel energy.

Figure 3 shows the heat transferred to the cooling system over various engine speeds and loads. Generally, the heat rejected to the coolant increases as engine power and engine speed increases. Looking only at the low engine speed points (2000 rpm), we see that the heat transferred to the coolant becomes larger as more loading is applied to the engine. Conversely, we see the opposite effect at the high end of the speed range (6000 rpm) and no definable trend for the mid-speed point (4000 rpm). The differences in heat rejection trends for the charted speed ranges are caused by differing boundary conditions. At low speeds, the engine is operated with a near stoichiometric air/fuel ratio for all load cases. As engine speed is increased to the 4000 rpm speed point, the engine is again operated at a near

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

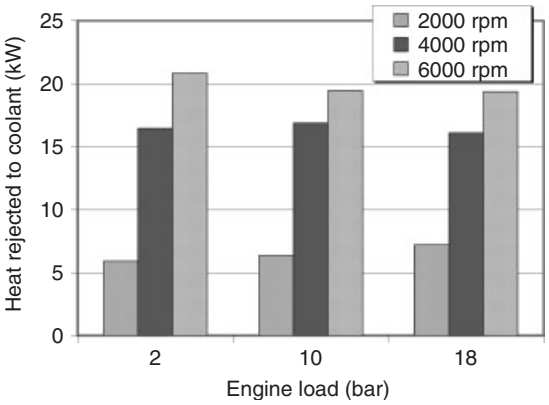


Figure 3. Energy split for a turbocharged direct injected gasoline engine operating at various speed and load points. (Reproduced by permission of FEV Inc.)

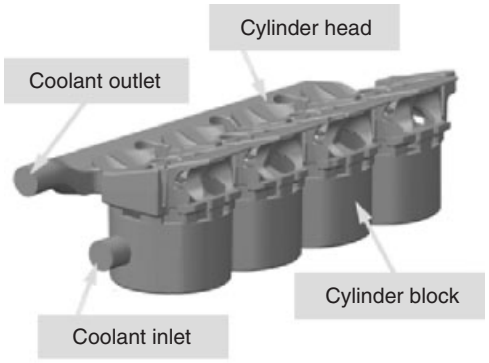


Figure 4. Example liquid cooling jacket. (Reproduced by permission of FEV Inc.)

stoichiometric air/fuel ratio during low loads but enrichment to limit exhaust gas temperatures begins to occur at higher loads causing a reduction in the amount of heat rejected to the coolant. At 6000 rpm, enrichment begins at even lower loads and increases further as engine loading is increased explaining the negative slope shown in Figure 3.

3 FUNDAMENTALS OF COOLING SYSTEMS

3.1 Types of cooling systems

3.1.1 Air cooled

The simplest cooling system available for internal combustion engines is air cooling. In these systems, either vehicle motion or a cooling fan blows ambient air over the engine to provide engine cooling. Cooling is aided by fins that are cast into the cylinder head, cylinder block, and other external components. These systems were used in some early automotive applications because of the systems simplicity and low cost. While this type of cooling system is inexpensive, it is no longer used in automotive applications in developed and emerging markets because of its inability to reject enough heat and also direct cooling to key problem areas in modern, high power density engines. As such, this section is largely devoted to liquid cooling systems.

3.1.2 Liquid cooled

Water- or liquid-based cooling systems have become standard for virtually all automotive markets. Liquid systems operate by pumping a liquid coolant through cast-in passages of the engine that form a jacket around the combustion chamber and cylinder liner (Figure 4). This liquid absorbs heat as it passes through the engine and then exits the engine to a heat exchanger typically mounted at the front of the vehicle where the heat is rejected to the atmosphere.

Keeping the center of the combustion chamber cool is the greatest challenge for the engine's cooling. While the air-cooled system must first transfer the heat from the center of the chamber to the outside surface where it can be rejected to atmosphere, the engineer has the ability to target the cooling of a liquid system to the critical areas that need the most heat transfer. It is for this reason that liquid cooling systems have become standard in modern automotive applications.

3.2 Common liquid cooling schematic

Countless engine cooling systems can be imagined and vary in cooling effectiveness, efficiency, and cost. Vehicle application, thermostat placement, expansion tank style, cabin heating requirement, and the number of external devices requiring cooling can make cooling system layouts differ significantly from one vehicle to another. A reference schematic is shown in Figure 5. This particular cooling system has multiple external devices that require cooling including the center housing of a turbocharger, an engine oil cooler, and an EGR (exhaust gas recirculation) cooler. A two-stage thermostat and pressurized expansion tank are also included in the system. It is clear that this is neither the simplest form of cooling system with the additional add-ons nor as complicated as some systems that incorporate strategies to speed cold start performance. Examples of these systems are discussed in later sections.

Developing the balance of the individual branches of the system requires several iterations of 1D flow analysis.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

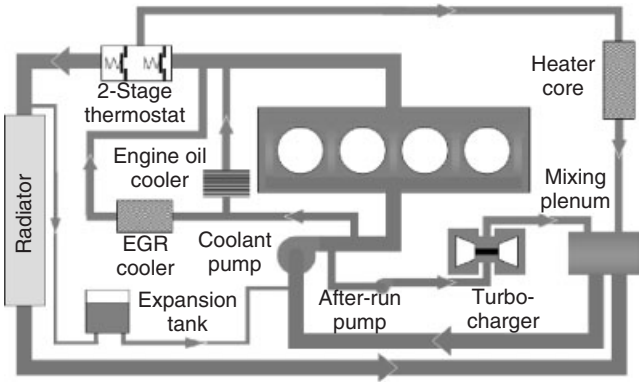


Figure 5. Common cooling system layout for a turbocharged automotive engine. (Reproduced by permission of FEV Inc.)

Table 1. Sample flow rate requirements for common cooling system components on a 175-kW turbocharged four-cylinder.

Component	Requirement (L/min)
Engine	170
Heater core	21
Oil cooler	19
Expansion tank (pressurized)	8
Turbocharger	8
Radiator	120

Careful selection of hoses, pipes, and orifices are selected to provide appropriate flow rates to the components within the system under different boundary conditions. In some cases, pump sizing may be dependent on meeting flow requirements to certain devices such as heater cores rather than cooling requirements for the engine. Table 1 shows some sample flow rates for given components in a typical cooling system.

4 LIQUID COOLING SYSTEM COMPONENTS

4.1 Coolant

Plain water has excellent cooling properties and can be used as an engine coolant. However, there are several significant issues associated with its use that prevent it from being used directly as an engine coolant. The largest issue is that water will freeze near 0°C and can cause severe damage to the engine, radiator, and piping as the frozen water expands. Another significant issue is that water boils near 100°C, whereas the coolant side of the chamber surface frequently ranges in temperature from 120 to 125°C with short-term excursions to 130 or 140°C. Plain water, even under pressure, in contact with surface temperatures this high will begin rapid boiling and the local heat transfer coefficient will be dramatically reduced.

Ethylene glycol is the primary constituent in virtually all automotive coolants. It is an alcohol that, when diluted 50% with water, significantly increases the boiling temperature and simultaneously reduces the freezing temperature of water.

In addition to the ethylene glycol, corrosion resistant additives are mixed with the coolant to protect the internal passages of the cooling system. Three of the most popular additives are based on using silicates, organic acid technology (OAT), or hybrids of the prior two (HOAT, hybrid organic acid technology).

Silicate-based cooling systems were the most popular options for automotive applications and can be identified by their green pigment. The silicates offer great corrosion and pitting protection to metal engine parts. However, the protection is limited to 2 years at which time the silicates can start to fall out of the mixture and begin to build up in the internal passages of the cooling system. This can eventually lead to blocked passages and an ineffective cooling system. The silicates are also abrasive and can limit the longevity of coolant pump seals.

To address the limited change intervals of the silicate-based cooling systems, most major automotive companies have moved to either OAT or HOAT coolant for their factory coolant fill. OAT coolants typically will use carboxylate, sebacate, and 2-EHA as corrosion inhibitors and can be distinguished by their orange pigment. The biggest advantage of the OAT coolants is that there are no silicates to drop out and build up. Thus, a longer service life of the coolant, up to 5 years, can be expected.

HOAT coolant is basically an OAT with benzoate as an inhibitor and a small amount of silicate added in order to obtain the superior corrosion benefits silicate combined with the extended life of the OAT coolant.

4.2 Coolant pump

4.2.1 Mechanical pumps

Most coolant pumps are mechanically driven from either the front accessory drive or the timing drive system. Virtually, all mechanical pumps are centrifugal style with variations in impeller designs and impeller materials used. The most simple and cost effective are simple stamped steel impellers. Pump efficiency can improve significantly with the addition of backing and top plates as shown in Figure 6. These plates are used to seal off the front and back sides of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

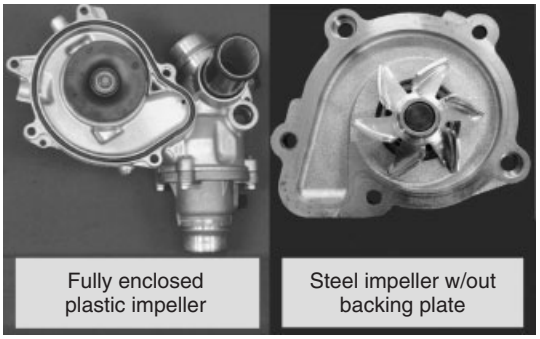


Figure 6. Common coolant pump impeller designs. (Reproduced by permission of FEV Inc.)

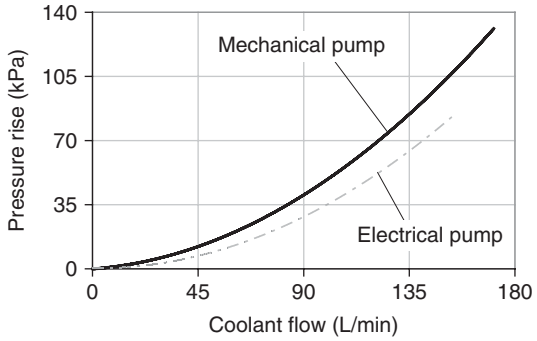


Figure 7. Comparison between mechanical coolant pump for a typical four-cylinder gasoline engine and a 400 W electrical coolant pumps for an automotive application. (Reproduced by permission of FEV Inc.)

the impeller blades to reduce leakage from one blade to the next as the pump rotates. These plates can be either welded to a steel impeller or, more commonly, integrated into a one-piece plastic impeller. Although less effective, some pumps will use only backing plates to improve efficiency. It is important to note that many manufacturers opt for impellers without backing plates and take a penalty in efficiency in order to reduce the hydraulic work at higher speeds when additional coolant flow is not required.

The housing for mechanical pumps can be made from either plastic or aluminum. However, owing to the accessory drive loads and support for appropriately sized bearings, aluminum is the most common housing material. Accessory drive loads and the location of the water pump within the accessory drive must be considered when sizing the bearing for the water pump.

4.2.2 Electrical pumps

The traditional mechanical coolant pump is tied to engine speed and not cooling requirement of the engine. As such, the pump is sized for a worst vehicle operating condition, which is a low speed full load, low airflow, high ambient condition. Thus, the pump is sized for a condition is rarely encountered (less than 5% of vehicle operation), the majority of the time the pump is oversized for the cooling needs of the system. This results in unnecessary parasitic losses and reduced fuel efficiency. In addition, as the mechanical pump is tied to engine operation, it continues to operate during cold start where no coolant flow is necessary resulting in longer warm-up times, increased engine emissions, increased engine friction, and reduced fuel efficiency.

Electric pumps can provide peak cooling when needed but less cooling when it is not due to decoupling cooling flow and engine speed. For hybrid vehicles and engines equipped with stop/start capability, the decoupling of the coolant pump and engine speed also offer better cooling and continued cabin heater performance during temporary engine shutdowns. Finally, with engine downsizing becoming more popular as a means to reduced fuel consumption, electric pumps are becoming more popular. As engine downsizing usually involves replacing a higher displacement naturally aspirated engine with a smaller displacement, turbocharged engine, the electric pump can circulate coolant after engine shutdown resulting in increased turbocharger durability and eliminating the need for a separate after-run pump. Table 2 shows some of the benefits of electrical versus mechanical pumps.

The greatest limitation to the market penetration of electric pumps is the significant cost increase over that of a mechanical pump. However, the cost is a function of the size of the pump, which is directly related to the peak flow capacity, or, more accurately, the size of the electric motor. Reduction in the pressure drop throughout the cooling system can help reduce the electrical pump size and ultimately the cost of the pump. Electric pumps for automotive applications are typically sized at about 400 W. Figure 7 compares a 400 W electric pump to a typical mechanical unit. As can be seen in the chart, the electric pump is not capable of meeting the same flow rate as the mechanical pump at a given pressure. As such, significant effort must be placed on reducing the cooling system pressure drop during the development of the engine.

4.2.3 Clutched and hybrid pumps

A third option is a mechanical pump with either a clutched pulley or a small electric motor. Clutched pumps offer the ability to disconnect the pump from the engine using a vacuum or electrically operated clutch. The pump pulley is allowed to free spin to reduce parasitic losses and speed cold start warm-up while re-engaging to drive the pump

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Benefits of electrical versus mechanical pumps.

Benefits	Parasitic	Controlled Flow	Zero-flow Warmup	Stop/start	Depower	Turbo After Run	Cooling Component Downsizing	Accessory Drive	Packaging
Mechanical	Pump operates throughout the RPM range, consuming engine power	Pump designed to meet worst driving conditions, resulting in excess flow (over cooling) for normal operating conditions (95% of time)	Pump always operates when engine is on	When stopped (engine off), no coolant is supplied to heater core	Certain driving conditions (low RPM/high load) may exceed maximum tempera-ture; engine is depowered to protect the cooling system	Mechanical water pump and after-run pump coexist (cost of two pumps)	Components such as fan and radiator are sized to extreme conditions where pump flow is limited	Pump needs to be driven by accessory belt	Pump needs to be packaged as part of the accessory drive
Electrical	Alternator load increases, but pump consumes energy only as needed	Pump provides exact flow required, no more no less, in every driving condition	Pump can be shut off when not needed (cold start), resulting in faster warmup, reduced emissions, reduced engine friction, and increased fuel efficiency	Pump functions during engine off, providing coolant to heater core to sustain customer comfort and extend stop/start	Pump can output maximum flow regardless of RPM or load, resulting in depower avoidance	Water pump can work as after-run pump to save cost (only one pump needed)	Pump flow is maximized indepen-dent of condition	Pump is driven electrically resulting in a simpler accesory drive	Pump can be packaged anywhere in the under-hood environ-ment

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

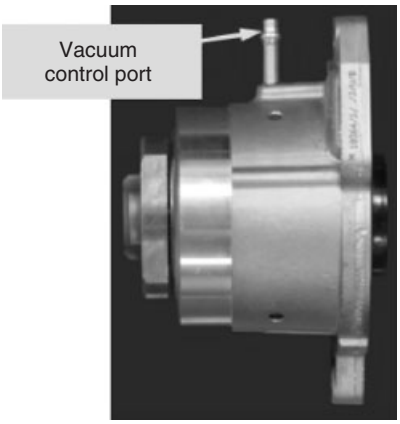


Figure 8. Example of a vacuum-clutched coolant pump. (Reproduced by permission of FEV Inc.)

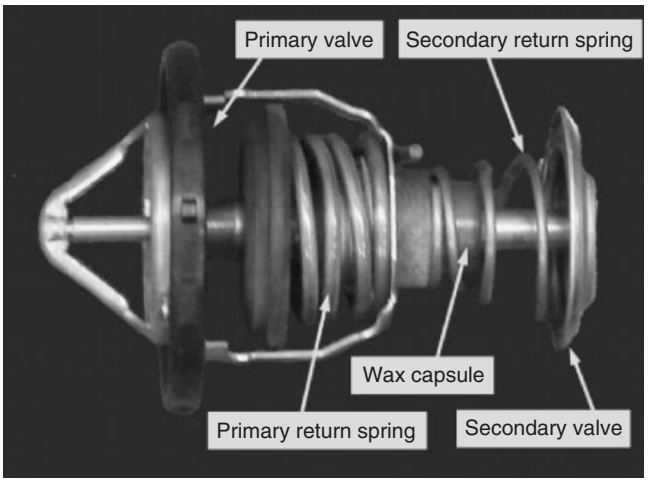


Figure 9. Mechanical thermostat. (Reproduced by permission of FEV Inc.)

when engine cooling is required. An example of a clutched coolant pump is shown in Figure 8.

Hybrid pumps are similar clutched pumps with the addition of a small electric motor capable of spinning the pump for low flow rates. These hybrid pumps, while more expensive than mechanical pumps, not only offer the ability to flow the amount of coolant required for the given engine load but also switch to mechanical operation for high load operations such as towing. At maximum output, mechanical drive efficiency is better than the combined efficiency of an electric pump, which must have its drive power created by an alternator.

4.3 Thermostat

On most engines, the coolant temperature is regulated by the thermostat. In its basic form, a thermostat is a spring-loaded valve used to regulate coolant flow through the vehicle radiator. When the coolant is cold, the thermostat remains closed. As the coolant heats up, wax inside the capsule begins to melt and expand forcing the valve open gradually beginning at a prespecified temperature. As the engine cools, the thermostat then begins to close again. Changes to the thermostat's geometry and blend of wax can be used to calibrate the opening temperature of the valve. In addition, the location of the thermostat is an important factor in the behavior of the overall cooling system. These locations and effects are discussed in the following sections.

In order to speed engine warm-up, some engines use a double valve arrangement. On start-up, the secondary valve opens first that allows the engine coolant to circulate freely through the engine but without flowing through the vehicle's radiator. As the coolant circulates through the engine, the temperature of the coolant continues to rise and the main valve begins to open diverting coolant to the radiator. Figure 9 shows an example of a double valve mechanical thermostat.

4.3.1 Inlet/outlet t-stat

Thermostats are generally placed on the inlet or outlet to the engine. Both configurations have their benefits and disadvantages, and placement is typically driven by the preferences of each OEM.

Outlet thermostats were traditionally the most common configuration. When the coolant heats up and thermostat opens, cold coolant begins to flow through the engine until it reaches the thermostat. This rush of cold coolant causes the thermostat to close and the cycle begins over again. The series of thermal shocks imparted on the engine components during engine warm-up is the main reason that some automobile manufacturers have chosen to move to inlet thermostats.

Conversely, when an inlet thermostat starts to open only a small amount of cold coolant is allowed into the engine but, because it is located on the inlet side, the valve begins to close quickly until the engine heats up the new coolant and the valve opens again. Inlet placed thermostats are not without their own issues, though. The large pressure drop across the valve just before the inlet pump can negatively influence pump efficiency as well as create cavitation issues in the pump.

4.3.2 Mapped thermostat

To obtain more control over the coolant temperature of the cooling system, some thermostats have become electrified.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

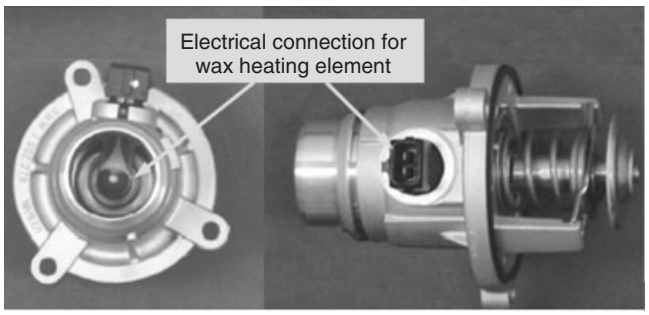


Figure 10. Electric (mapped) thermostat. (Reproduced by permission of FEV Inc.)

The basic function of an electrically assisted thermostat is unchanged in that the valve is opened by expanding the wax within the capsule. However, the opening coolant temperature is set 10–15°C higher than traditional, mechanical only thermostats. This allows the coolant to maintain a higher temperature during part load operation that reduces engine friction through higher oil temperatures and lower viscosity while also improving thermal efficiency by reducing heat transfer into the coolant. During full load operation, a small electric heater within the thermostat is used to increase the temperature of the wax and open the valve further and faster, thereby reducing the overall operating temperature of the coolant and response time of the thermostat to protect the engine structural components and provide better knock margin for gasoline engines at full load. It is worth noting that the electric heater alone cannot open the thermostat when the coolant is below a certain temperature. However, the electric heater does extend the opening temperature range and response time of the thermostat. Figure 10 shows an example of a thermostat with electric assist.

4.3.3 Electric valve

As most OEMs are starting to realize the effects a well-controlled cooling system can have on fuel economy, electric valves are being introduced in the automotive industry as they provide a number of advantages over both traditional and electric thermostats. The two main advantages are precise control of the coolant temperature and much faster response. As these valves can go from fully closed to fully open in about one second, the operating coolant temperature can be set to much higher temperatures than traditional systems reducing engine friction and improving thermal efficiency much like electrical thermostats. However, due to the much faster response, electric valves offer the highest overall engine thermal efficiency improvement as they allow for engine operation very close to the knock limit for gasoline engines. The operation of the valve is controlled by the engine's ECU that can consider ambient air temperature, throttle position, air/fuel ratio, and coolant temperature to determine the best position of the valve. With mechanical systems, the only signal used to determine valve position is the temperature of the coolant at the thermostat. However, with the electric valve, temperature measurements can be made anywhere in the cooling system and the ECU can control the valve. This allows the valve to be positioned at the outlet of the engine to eliminate the pressure drop directly before the coolant pump, whereas the temperature signal can come from the entrance to the engine, thereby eliminating the thermal cycling effect of an outlet thermostat.

4.4 Expansion tank

As the engine coolant increases in temperature, it begins to expand and pressure is built up in the cooling system. This increase in pressure is advantageous in that it increases the boiling point for the coolant reducing the chances for overheating. However, if excessive pressure builds up within the system, permanent damage can be done to engine components and seals. Therefore, a spring-loaded pressure relief valve is placed in the system fill cap, which allows excess coolant and vapor to exit the system and reduce system pressure. There are two main system strategies used to manage this effect as described in the following sections.

4.4.1 Unpressurized

In an unpressurized system, a simple overflow reservoir is used to capture the excess coolant during operation. In this configuration, the system fill cap is generally located at the radiator with a hose that runs from the radiator, downstream of the pressure relief valve, to the expansion tank. As excess coolant bleeds off of the system, it is caught in this reservoir, which remains at atmospheric pressure. As the system cools, the vacuum within the cooling system pulls the excess coolant back into the radiator.

4.4.2 Pressurized

Pressurized reservoirs, sometimes referred to as *hot bottles*, are inherently more expensive but offer some additional advantages. With pressurized reservoirs, a small amount of flow is diverted into the expansion tank. There, the flow velocity is relatively low because of the large volume and low flow rate. This semiquiescent chamber allows for small vapor bubbles to settle out of the coolant before being reintroduced into the system ensuring better cooling properties

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

for the coolant in the engine. The action of removing the vapor bubbles from the coolant is called *deaeration* and is extremely important as very small amounts of vapor can have a detrimental effect on the ability of the coolant to transfer heat.

In addition, some turbocharger applications can be set up to cool the turbocharger after shutdown with the aid of a pressurized expansion tank. This process, known as *thermal siphoning*, is described in more detail in the following section.

4.5 Cooling fan

A cooling fan is used to pull additional air through the vehicles radiator beyond that which is provided through either natural convection or ram air forced through the radiator due to vehicle speed. There are two basic drive methods used for cooling fans, electric and mechanical.

4.5.1 Electric

Most passenger car applications are equipped with thermostatically controlled electric cooling fan(s) for improved vehicle fuel economy and also facilitate transverse mounting of the powertrain. These fans operate only when the radiator temperature exceeds a predetermined limit; fan speed may be continuously controlled as a function of the coolant temperature. The conversion of mechanical energy to electrical energy through the belt to alternator and then back to mechanical work at the electric fan is inherently inefficient. However, despite this inefficiency, electric fans offer superior parasitic loss advantages to most mechanically driven fans because they are decoupled from engine speed and their operation can be controlled based on cooling requirement. This enables the fans to be turned off during low heat rejection load points, cold start, or when available ram air across the radiator is sufficient for engine cooling.

4.5.2 Mechanical

For $^1/_2$- and 1-ton truck applications, the cooling demand may be too high during towing conditions for reasonably sized electric fans. As such, mechanical fans are still commonly used to cover the peak cooling requirements, which are more likely to be a steady-state condition than on a passenger car.

To mitigate the parasitic losses during low loads, mechanical fans are equipped with a thermostatically controlled viscous clutch. This clutch allows the fan to spin at a slower speed than the engine during cold conditions. As the radiator warms up, air surrounding the clutch warms up as well and a thermal spring located on the front of the clutch begins to open a valve within the clutch at around 70–80°C. This valve allows more coupling fluid to be released and brings the fan speed closer to that of the engine.

4.5.3 Hydraulic

Although unusual, there have been applications of hydraulic cooling fans for light-duty vehicles. In one application, the hydraulic power steering pump provides the hydraulic fluid to drive the cooling fan motor. The hydraulic fan motor speed is electronically controlled to match the cooling demands.

4.6 Radiators

The heat absorbed by the coolant as it passes through the engine is released to atmosphere at the radiator. While the operating principle of common automotive radiators has remained unchanged from the early years of the automobile, the construction and materials have changed over time. Modern radiators are constructed using an aluminum core, tube and fins, and composite end tanks, whereas prior art utilized a copper core and brass tanks.

The layout of the automotive radiator has improved from early designs as the industry has converted from vertical or down flow to horizontal or cross flow. This change allows the radiator cap to be placed at the low pressure side of the radiator, beneficial during prolonged high speed operation where the high pressure side of the radiator may exceed the vent pressure of the cap. In addition, moving the tanks to the sides creates more heat transfer area low in the vehicle, where the majority of the ram air used for cooling is available, and also facilitates a lower, more aerodynamic hood line.

The amount of cooling air that travels through the vehicles grill in order to cool the radiator has a significant effect on the vehicles aerodynamics. With increasing concern regarding fuel consumption, active aerodynamic devices are starting to be adopted by some automotive manufacturers. One such device are grill shutters, which allow air to travel through the vehicles grill and provide cooling to the radiator when required but can close off the opening in the grill when additional cooling is not needed. Another device used to reduce aerodynamic drag is a hood extractor, which allows for the air passing through the front fascia to be extracted. This device in addition to aerodynamic benefits offers cooling benefits as it increases the velocity of the air traveling through the radiator, which in turn increases the heat transfer capacity of the radiator.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.7 Oil cooler

For high power density engines and vehicles that will spend large amounts of time under high load, oil temperature can become a concern. To mitigate the risk of oil coking, an oil cooler is frequently used to transfer heat from the oil to the cooling system or directly to ambient air. These oil coolers, with the appropriate control strategy, can also be used to speed the heat-up of the engine oil and reduce oil viscosity during start-up.

4.8 EGR cooler

EGR is commonly used on diesel engines as a method to reduce the in-cylinder creation of NO*x* emissions, whereas some gasoline engines also use EGR as a method of reducing engine knock at high load. However, both of these strategies require the exhaust gases to be cooled before inducted into the engine. As such, EGR coolers are frequently used and plumbed into the engine's cooling system. Depending on the amount of EGR used by the engine, these coolers can account for a significant portion of the engine's cooling requirement. While this requires an increase in the capacity of the cooling system, it can also be beneficial by accelerating coolant warm-up that reduces engine friction and fuel consumption.

4.9 Heater core

The vehicle passenger compartment heater core circuit must be considered when sizing the engines cooling pump. Many customer complaints regarding the cooling system are based on poor heater performance during either warm-up or operation. The fundamental flaw with poor performing systems can be directly related to coolant flow through the heater core. This phenomenon is exaggerated during idle. A simple but not desirable countermeasure is to increase idle speed, which has negative effects on fuel consumption. In some cases, pump sizing can be dictated by the minimum coolant flow rate required to provide adequate cabin heating. Larger vehicles with multiple heater cores are particularly susceptible to this condition. As such, it is extremely important to properly design the heater core circuit to reduce overall pressure drop.

5 ENGINE COOLING

5.1 Cooling jacket design

There are several flow strategies used for the cooling passages within the engine. While there are many configurations not covered here, three of the most commonly used strategies are described in the sections later.

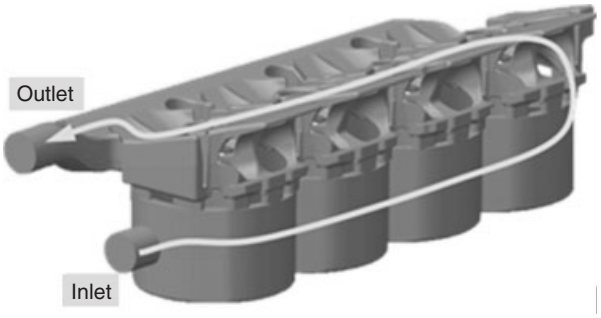


Figure 11. U-flow cooling diagram. (Reproduced by permission of FEV Inc.)

5.1.1 U-flow

The most common cooling system is the U-flow, or sometimes referred to as *axial-flow*, configuration. In this configuration, coolant enters the cylinder block at one end of the engine. The coolant then flows through the cylinder block and then flows up into the cylinder head once it reaches the other end of the block. Once in the cylinder head, the coolant flows back toward the same end of the engine where the coolant entered the block. An illustration of the U-flow cooling jacket is shown in Figure 11. This cooling system, even though more susceptible to thermal bore distortion, is the simplest to design and develop. However, with increasing power densities and ever-increasing fuel consumption targets, other, more advanced cooling strategies are becoming more common.

5.1.2 Split cooling

Split cooling systems are similar to the U-flow configuration in how they flow through the cylinder block and head. However, in this strategy, coolant enters the head and block at the same end of the engine, flow in parallel, and exit at the other end. There is no communication between the head and block as shown in Figure 12. This enables the head and block temperature to be metered by separate thermostats and be controlled to different temperatures. The benefit of this strategy is that the cylinder head can be held at a temperature cool enough to avoid any mechanical failures, whereas the block, which has significantly less heat loading and few delicate features, can be kept at a higher temperature. The higher temperature cylinder block keeps the engine oil temperature higher and viscosity lower, thereby reducing engine friction. This system requires

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

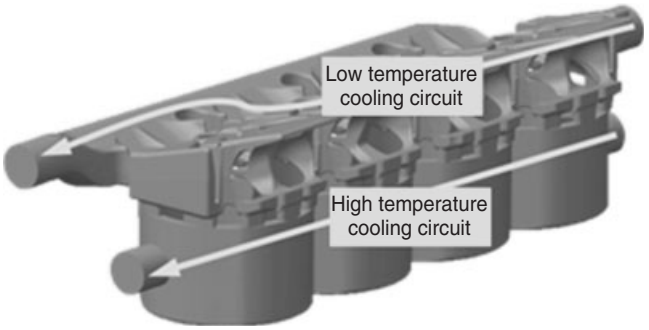


Figure 12. Split cooling diagram. (Reproduced by permission of FEV Inc.)

additional external plumbing and an additional thermostat in comparison to the U-flow design.

5.1.3 Cross flow

The most common cooling jacket design for high power density engines is the cross-flow design that gets its name from the flow path of the coolant across the cylinder head from the exhaust side to the intake side of the jacket. Coolant can enter the engine block from the side or end of the block and then circulates to the opposite side. There, it flows up into the cylinder head, across the cylinder head, and collects on the opposite side. Once there, the coolant must be routed to a location where it can exit the engine. An illustration of the cross-flow cooling jacket is shown in Figure 13.

This cooling system has the best cooling capability because coolant can be directed to specific, high heat flux regions of the cylinder head such as the small bridge between the two exhaust valves in a four-valve-per-cylinder engine. If designed correctly, the cross-flow design can also offer the lowest pressure drop of the three major options. However, development of this system requires more effort and multiple loops of three-dimensional cooling simulation to verify that the flow is balanced across all cylinders, the critical areas receive a proper amount of flow, and that there are no stagnant volumes in the block or head.

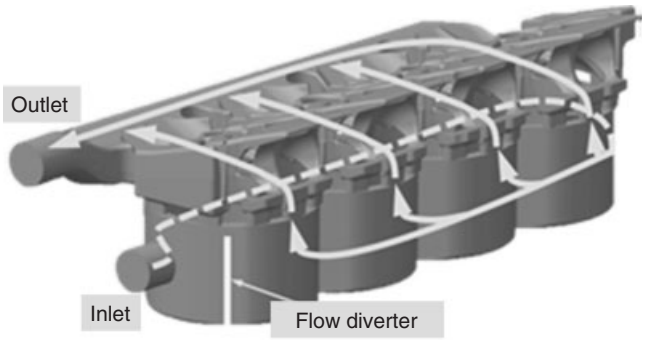


Figure 13. Cross-flow cooling diagram. (Reproduced by permission of FEV Inc.)

One challenge with this strategy is getting uniform cooling flow through the block without stagnation points while simultaneously and providing uniform coolant flow into the cylinder head. A flow diverter can be inserted in the cylinder block that forces the majority of the flow to take the long path through the cylinder block while allowing a small portion of the flow to take a shorter circuit strictly to help balance the cylinder head. The cylinder head gasket is used to meter the coolant between the cylinder block and head. The location and size of the holes can be tuned to provide cooling where it is required in the head and also balance the cylinders.

The coolant entry point and flow distribution in a cross-flow configuration can play a significant role in the uniform cooling of the block. If designed properly, thermal bore distortion can be minimized and coolant temperature difference between block entry and exit points can be increased. This temperature difference increase allows for a reduction in coolant flow, which leads to a water pump downsizing and thus better fuel efficiency because of the reduction in parasitic losses.

5.2 Cylinder head

The cylinder head has the highest heat load in the combustion system and features several delicate, thin bridges between the various valves, spark or glow plugs, and fuel injectors. The cylinder head cooling jacket must be designed to target the highest heat flux areas within the cylinder head. The primary concerns in this respect are the small bridges between the two exhaust valves (for four-valve-per-cylinder engines) and the bridges between the exhaust valves and spark plug or fuel injector as shown in Figure 14. These bridges are highly susceptible to low cycle fatigue and must be kept cool to limit the reduction in material strength and also reduce the thermal growth that causes compressive stress at high load. As such, directing coolant to these critical areas is a primary concern for design engineers. An example flow path for a cooling jacket designed to provide key coolant flow rates to these areas is shown in Figure 15.

Beyond cooling the valve bridges, the chamber is generally receives a sufficient amount of cooling as long as the coolant flow is not stagnant and stays close to the floor of the water jacket.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

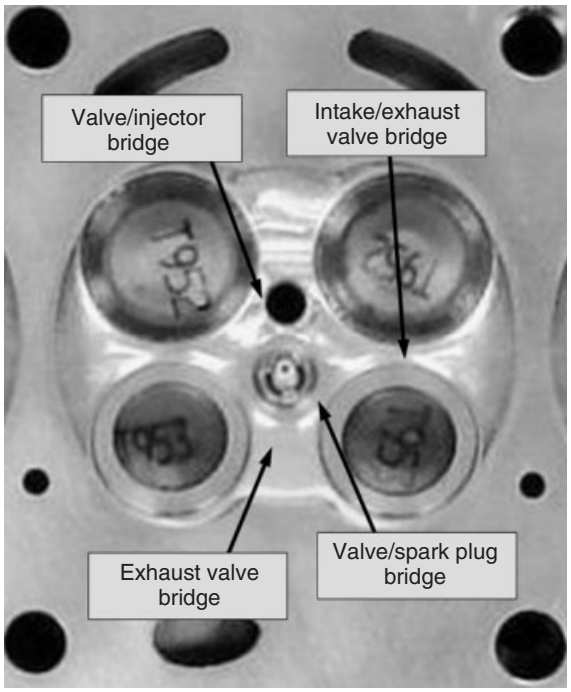


Figure 14. Key bridges in the combustion face of a cylinder head. (Reproduced by permission of FEV Inc.)

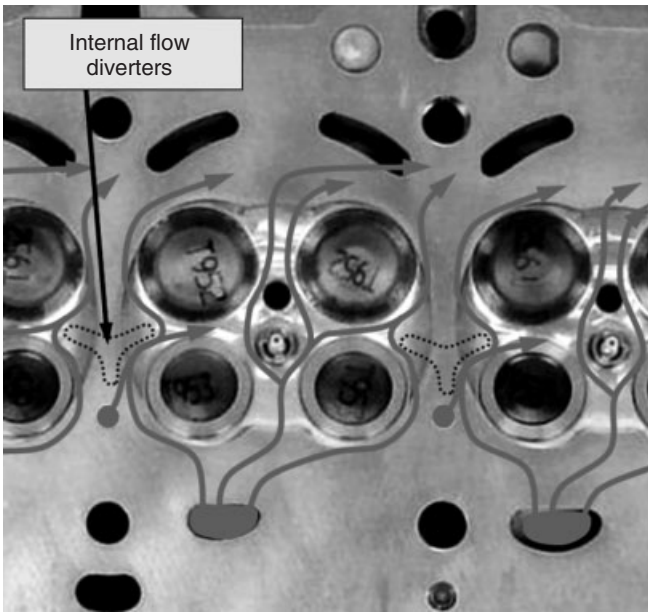


Figure 15. Example cylinder head cooling cross-flow pattern designed to cool all key areas. (Reproduced by permission of FEV Inc.)

5.2.1 Degassing

The cylinder head cooling jacket must be able to purge gas during system filling and vent coolant vapor that may be created during operation. The ceiling of the cooling jacket is largely determined by the oil cavity above it, which has its own requirements including oil drain paths. Regardless, it is important that the cooling jacket maintain its ability to properly vent. It is acceptable, however, to have some localized, slightly negative slopes in the ceiling as these will vent as the engine motion shakes the gas past the trap.

5.2.2 Head gasket

The cylinder head gasket plays a key role in the cooling system. It acts as a metering plate between the cylinder block and head. This enables the engineer to specify the amount of coolant required into each cylinder head core print. The head gasket can also create the highest pressure drop in the engine. As such, the metering passages should be as large as possible while still providing the proper direction for the cooling. Small passages that allow for gas and vapor must be used to transfer gases from the high side of the cylinder block into the cylinder head jacket and eventually out of the engine.

5.2.3 Valvetrain

Cooling the valvetrain must be considered when designing the cylinder head cooling jacket. A significant amount of heat is transferred through the valve head into the cooling jacket through the valve seats. It is critical that the coolant has a moderate flow velocity around the valve seat. Failure to properly cool the valve seats can lead to distorted seat rings and/or valve-to-valve seat adhesion.

Cooling the valve guides, specifically for the exhaust valve, is also critical for the engines durability. Moderate cooling velocity in the jacket surrounding the exhaust valve is required to dissipate the heat released by the valve and prevent galling of the guide. Thermal distortion in the cylinder head between the valve guide and the valve seat can create a misalignment between the projected valve guide centerline and the centerline of the valve seat. Steep temperature gradients within the head can lead to this condition causing high edge loads on the valve and seat and small clearances at the seat when the valve is closed. Both conditions can lead to poor combustion sealing and ultimately failure of the engine.

5.2.4 Two-piece cooling jacket cores

Most automotive diesel engines and some high power density gasoline engines utilize two water jacket cores in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the cylinder head. The primary reason for the additional core is to create an intermediate deck that helps to support the combustion chamber under high cylinder pressure. It is important, though, that the cooling flow pattern be designed properly to keep good cooling characteristics across the combustion chamber.

5.3 Cylinder block

Thermal loading in the cylinder block is significantly lower than in the cylinder head. This is largely due to the very small portion of the cylinder being in direct contact with the combustion gases for the full power and exhaust strokes and the large contact area of the cylinder liner with the cooling system. General cooling jacket design targets should be to cover the top piston ring at top dead center and follow the piston down for at least 80% of the stroke. In applications where high load operation can be expected for extended periods of time, such as towing conditions for light-duty trucks, a cooling jacket that covers 100% of the piston ring travel should be considered.

The main durability concern is at the top of the cylinder liner between adjacent bores. This area is in direct contact with combustion gases for the entire combustion and exhaust strokes with exception of when the piston is at top dead center. Heat is transferred into this small interbore bridge from two cylinders giving it less time to dissipate heat by comparison to other areas about the circumference. The problem is compounded in most aluminum block applications where the cylinder walls are joined together in a Siamese arrangement. While this arrangement allows for a tighter bore spacing and smaller package, there is typically no way to cast in cooling passages between the bores.

Several strategies have been used by the automotive manufacturers to minimize the effect of Siamese bores. Thinner bore walls in the interbore area can be used to drive the coolant toward the pinch point as much as possible. From there, the engine relies on the superior heat transfer characteristics of aluminum to transfer the heat to the coolant.

However, high power density engines typically require better interbore cooling than can be achieved with the directed cooling approach outlined earlier. Instead, various machining operations can be considered to improve cooling in this region. Saw cuts separating the bores at the top 20–30 mm allow coolant to pass from one side of the block to the other has been used frequently. While effective, the two commonly encountered issues with this method are the loss of structural integrity at the top of the liner and the low flow rate generated by the relatively small difference in pressure between the two sides of the block.

Cross drilling from one side of the block to the other is another common practice and can offer a structural advantage to saw cuts as the connection between the top of the cylinder bores is not completely removed. In comparison to saw cuts, which can be made quickly with large diameter circular saw blades plunged into the block, cross drills require very small diameter drills to be used over relatively large distances. The diameter to length ratio creates a high risk of the drill bit wandering and contacting the iron liner. Several variations have been used to try to mitigate this effect. The use of a stepped drill allows for a larger diameter to be used for a portion of the span that creates a stiffer tool that is less prone to walking. Another method is to use the same stepped drill but drill from both sides of the block. These drills effectively form a V configuration and provide the lowest tolerance option for cross drilling.

One final variation of interbore is cross drilling through the block into the cylinder head. This configuration gives the best flow rate of the three most common options, because the pressure difference between the cylinder block and the head is used to develop a consistent flow rate. Examples of all three configurations are shown in Figure 16.

The amount of coolant volume is critical for cabin heating and engine friction attributes of the engine. Large volumes of coolant create a large thermal mass that takes a significant amount of time to bring up to temperature. During this time, passengers go without the ability to heat the cabin of the vehicle while cold oil creates higher engine friction leading to poor fuel economy. As such, it is important that the volume of the cooling jacket be minimized in order to reduce the thermal mass as much as possible. With high pressure die-cast cylinder blocks, the coolant volume is fixed primarily by the engine's stroke and the depth of the cooling jacket. Given the standards for die designs, some manufacturers have opted to install plastic inserts in the block cooling jacket to displace coolant volume and improve cold start conditions. An example of such an insert is shown in Figure 17.

5.3.1 Degassing

One critical parameter that should not be overlooked when developing the water jacket in the cylinder block is the ability to purge air and coolant vapor. Small holes in the head gasket on the high side of the water jacket must be used to transfer gas from the block water jacket to the cylinder head. The hole sizes in the gasket to support degassing of the cylinder block are typically driven by the smallest hole that the head gasket manufacturer can make with reasonable tool life (Figure 18).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

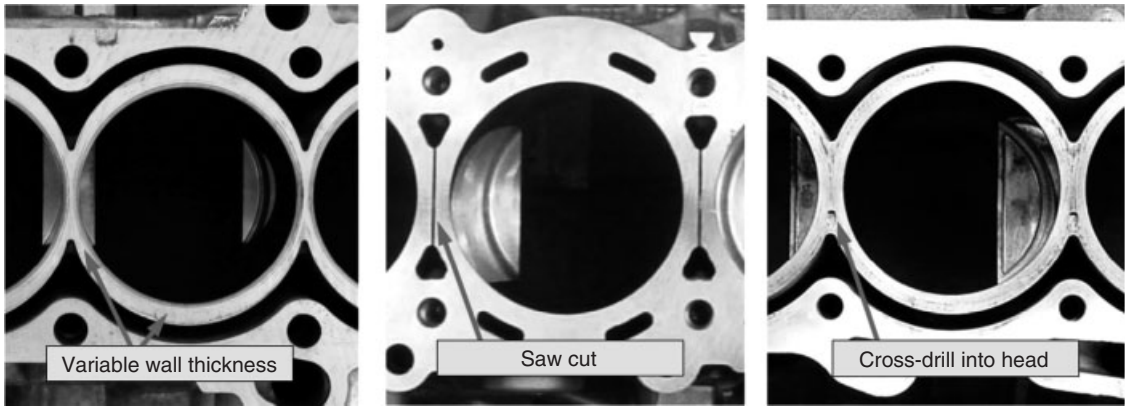


Figure 16. Interbore cooling examples for an engine block with Siamese cylinder bores (a) variable wall thickness, (b) saw cut, and (c) cross drill into the cylinder head. (Reproduced by permission of FEV Inc.)

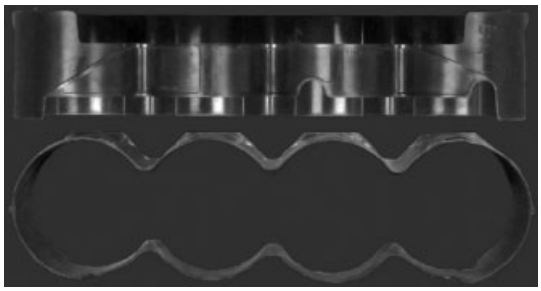


Figure 17. Example of an insert to remove volume from a cylinder block cooling jacket. (Reproduced by permission of FEV Inc.)

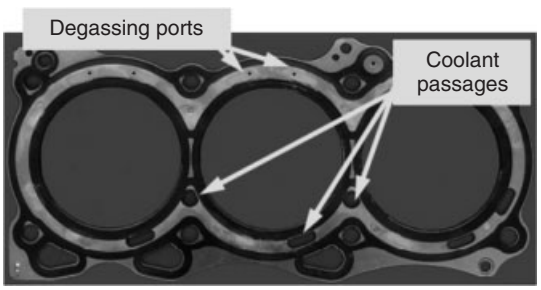


Figure 18. Degassing holes in a V6 cylinder head gasket. (Reproduced by permission of FEV Inc.)

5.4 Piston

Automotive pistons are not cooled by the cooling system directly beyond their heat release to the bore wall. The heat transfer from the piston to the block liner is mainly through the piston rings and a function of the area, materials, and physical connections between rings, piston, and liner. However, it is important to note that piston cooling jets are required for high power density engines in order to cool the piston crown and ring pack. Piston cooling jets are mounted at the bottom of the cylinder and direct a stream of oil onto the bottom side of the piston or into a small, cored annulus cast into the piston crown. Heat from the piston is transferred to the oil where it is then removed from the oil by the cooling system at the oil cooler.

5.5 Turbocharger

Virtually, all automotive turbochargers have a center housing cooled by the engine cooling system. Failure to provide adequate cooling will lead to seizure of the turboshaft and failure of the assembly. General cooling is relatively easy to accomplish as the flow rates required are relatively low, commonly 8 L/min. However, the challenge for the cooling system is typically in the hot-soak condition. Once the engine is shut down, heat in the turbocharger continues to flow into the cooling system. The coolant then vaporizes and, if not replaced with fresh coolant, temperatures in the center bearing increase rapidly and may damage the shaft and bearings as well as engine oil itself whose lubricating properties are degraded when the temperature exceeds its functional limits.

5.5.1 After-run pump

The most robust solution to the hot-soak condition is an after-run pump. Here, a small electric pump is placed in the turbocharger's cooling loop. Once the engine is shut down, the pump continues to circulate coolant through the turbocharger, providing fresh coolant and pushing out any vapor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

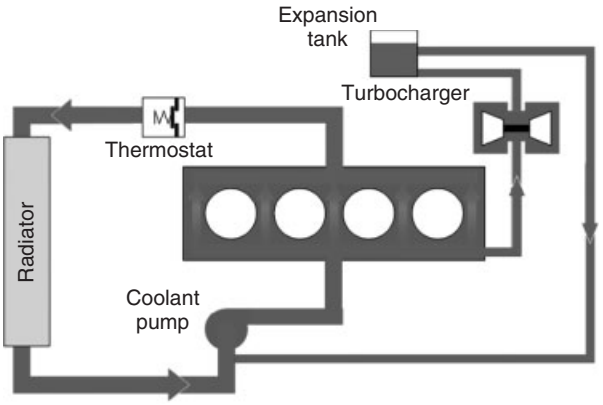


Figure 19. Example layout intended to use thermal siphoning to cool the turbocharger after shutdown. (Reproduced by permission of FEV Inc.)

After-run pumps offer a second advantage. Once the engine is shut down, the pump can also be used to circulate coolant through the heater core. This feature would allow the occupant the ability to stay warm for a period of time with the engine shutoff.

5.5.2 *Thermal siphon*

Thermal siphoning can be a low cost alternative to after-run pumps but is not possible in all applications. The thermal siphoning concept utilizes natural convection in a closed-loop circuit to provide coolant flow to the turbocharger. On shutdown, coolant in the turbocharger is heated and becomes more buoyant than the cooler coolant in the loop. As such, cooler coolant is pushed into the turbocharger as natural convection forces the hotter coolant to rise in the system. For the system to function, the hot coolant must be allowed to rise continuously on its route to the expansion tank, which means that a low mounted turbocharger is required. Any negative slope in the path between the turbocharger and the reservoir will prevent the convection process from completing the circuit. It is also important to point out that the system must be closed loop. This requires that the system incorporate a pressurized expansion tank. A simplified representation of the circuit is shown in Figure 19.

5.6 Charge air cooler

While not part of the traditional engine cooling system, charge air coolers are critical to engine performance in boosted applications. Two types of charge air cooling systems are commonly employed and described later.

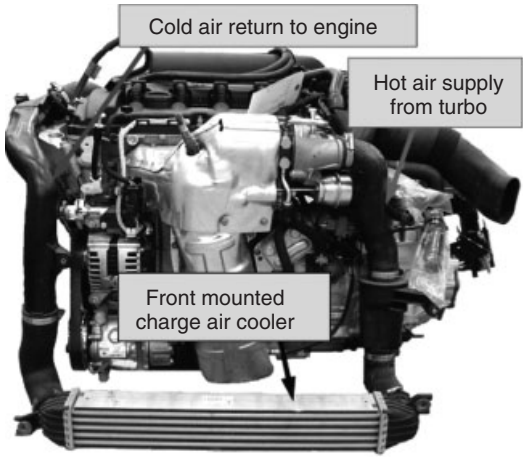


Figure 20. Engine assembly with air-to-air charge air cooler mounted at the front of the vehicle. (Reproduced by permission of FEV Inc.)

5.6.1 *Air-to-air*

Air-to-air charge air coolers are the most common charge air coolers for boosted applications. Once air exits the compressor, it is plumbed to a heat exchanger that is typically located in front of the radiator. As the air passes through the charge air cooler, heat is rejected to a cross flow of ambient air that is forced through the cooler by the vehicle's motion (Figure 20).

This is a relatively simple system with the greatest challenge being fitting an appropriately size heat exchanger in the vehicle and in a location that allows enough ram air to properly cool. The other challenge is the design of the long and complex charge air pipes that navigate through the engine compartment to the heat exchanger and back to the intake while being flexible enough to account for engine roll and thermal growth.

5.6.2 *Liquid-to-air*

Liquid-to-air charge air coolers utilize liquid coolant, typically the same formulation as the engine coolant, to remove heat from the charge air. The liquid cooling circuit is separate from the engine's main cooling circuit in order to keep the temperature of the coolant as close to ambient as possible. The charge air heat exchanger can be placed anywhere between the compressor and the cylinder head. Most often, it is integrated as a plug-in module into the intake manifold. The ability to cool the air in the manifold makes it a popular option for Roots style superchargers. These systems are also becoming more popular options for turbocharged applications for the reasons outlined later. To properly integrate the charge air cooler into the intake, it is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

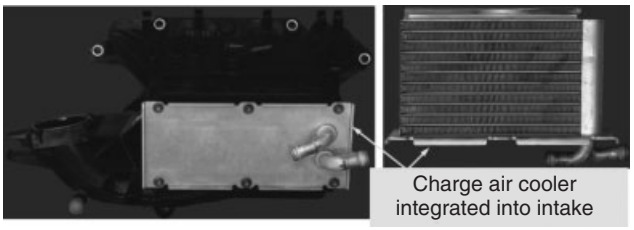


Figure 21. Example of a liquid-to-air charge air cooler with the heat exchanger integrated into the intake manifold. (Reproduced by permission of FEV Inc.)

critical that the engineering team develop the intake manifold to properly distribute air across the exchanger to realize an efficient system. An example of a charge air cooler integrated into an intake manifold is shown in Figure 21.

The cooling loop includes the charge air cooler, low temperature radiator, coolant pump, and associated cooling lines. The low temperature radiator is located in front of the engine's cooling module to have access to the coolest air possible. The coolant pump is typically between 50 and 100 W. Owing to the price difference between the available cooling pumps, significant effort should be made to minimize pressure drops through the entire system in order to minimize the size of the electric pump. A cooling loop schematic is shown in Figure 22. Note that while the system is separated from the engine cooling circuit, they can share a common expansion tank allowing a single coolant fill point for the vehicle plant and end customer.

The liquid-to-air system has several advantages over the air-to-air system. Throttle response is enhanced because of the reduced compressed air volume. The lack of large diameter piping traveling from the compressor to the front of the vehicle and back allows for a compact package with minimal volume. The thermal mass of liquid coolant is also an advantage for drive cycles that frequently operate in boosted mode. In contrast, an air-to-air cooler, with lower thermal mass, heats up quickly and loses effectiveness, resulting in reduced engine power.

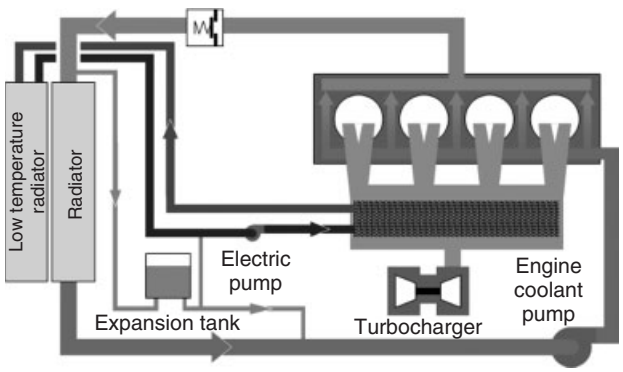


Figure 22. Liquid circuit for liquid-to-air charge air cooling system. (Reproduced by permission of FEV Inc.)

6 ADVANCED COOLING SYSTEMS

6.1 System effects on transient cycles (cold start strategies)

The cooling system effectively removes excess heat from the engine during operation. Unfortunately, the same characteristics that make an effective cooling system also inhibit quick warm-up of the combustion chamber and lubrication system. Traditionally, minimizing the coolant volume in the head and block cooling jackets would suffice. However, with ever-increasing emissions and fuel economy constraints, more effort has been put into cooling system layouts to speed catalyst light-off and improve engine friction during cold start.

There are several system layouts and strategies that can be imagined. One example is shown in Figure 23. In this particular layout, the engine is equipped with an EGR cooler in the cooling circuit. During cold start operation, coolant is allowed to flow through the EGR cooler to speed the heating of the coolant. The hot coolant is then diverted to flow through a transmission fluid heat exchanger and engine oil cooler. Heat is transferred to the transmission fluid and engine oil in order to quickly reduce the viscosity of the fluids and subsequently reduce the frictional losses of the engine. One concern with this strategy is time to catalyst light-off. Care must be taken with respect to how much heat is extracted from the exhaust in front of the catalyst and its effect on time to light-off.

6.2 Effects of hybrid and start/stop operation

One of the benefits of having a liquid cooling system is that the waste heat captured by the cooling system can be used to heat the cabin. If the engine is shut down, the coolant flow through the heater core stops and, during prolonged shutdown, cabin temperatures can drop. A supplemental coolant pump that allows continuous circulation of hot engine coolant through the heater core or an additional electric heater is required for acceptable heater performance in these applications.

During engine shutdown, metal temperatures within the engine increase significantly as heat is further transferred into a stagnant cooling jacket. Under normal engine

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

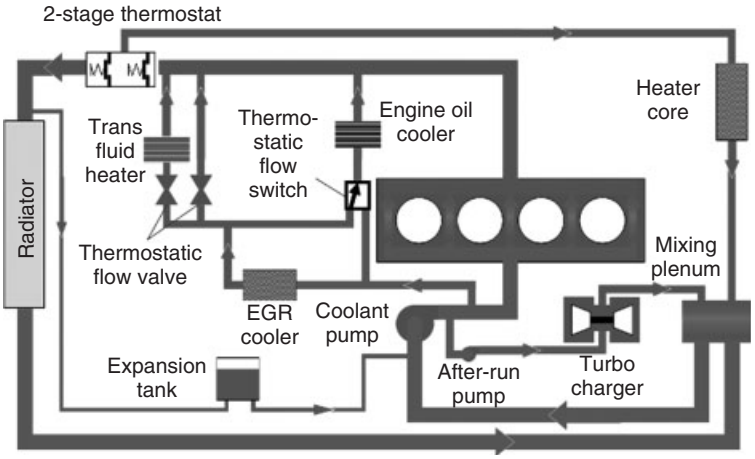


Figure 23. Example cooling circuit utilizing an EGR cooler to speed cold start heating cycles. (Reproduced by permission of FEV Inc.)

operation, these effects can be managed. However, when start/stop operation (engine shutdown during prolonged idle to reduce fuel consumption) is added, the number of cycles dramatically increases through the life of the engine. An electric pump that can be operated while the engine is shut down should be considered for some applications where high power density and start-stop operation are combined. With this configuration, the pump can continue to circulate cool engine coolant through the engine to limit peak hot-soak temperatures.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto149
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Diesel Fuel Injection Systems

Harsha Nanjundaswamy[1] and Hermann Josef Laumen[2]

[1] *FEV Inc., Auburn Hills, MI, USA*
[2] *FEV GmbH, Aachen, Germany*

1 Introduction 1
2 Electronic Diesel Control System 10
References 17
Further Reading 18

1 INTRODUCTION

Diesel engines are used in wide variety of applications such as passenger cars, trucks, locomotives, and marine and power generator applications. Most diesel engines carry fuel injection systems that manage the engine speed and power through controlled fuel injection. Over the past decades, increasing demands on fuel economy, emissions, and performance forced the automotive segment into tremendous improvements in diesel fuel injection system operation in terms of both design and precision. With the introduction of electronics to manage the complex control needs, they have improved the efficiency of fuel injection while making way for flexible fuel injection management such as multiple injections to achieve the comfort and efficiency needs. Section 1.1 reviews the development trends and discusses the state-of-the-art fuel injection system.

1.1 Diesel fuel injection systems

As shown in Figure 1, the diesel fuel injection systems are categorized into three main segments such as pump line nozzle, unit injector type, and common rail fuel injection systems.

In Section 1.1.1, a brief overview of the pump-line-nozzle and unit injector-type fueling systems is discussed followed by the discussion of state-of-the-art common rail fuel systems and advanced electronic diesel control (EDC) system working principles.

1.1.1 Pump-line nozzle

Pump-line-nozzle, consists of three major components to handle the fuel delivery. The pump manages both fuel delivery pressure and quantity using specially machined plungers. The nozzles inside the nozzle holder assembly are preloaded with spring, such that their opening for fuel injection depends on the fuel pressure generated by the pump, thus the pump influences the start of fuel delivery. The high pressure lines connect the pump and the nozzle holder. Shown in Figure 2 is a review of the each type of pump-line-nozzle type of injection system (Robert Bosch GmbH, 2000).

Although the principle of fuel injection remains common between the three categories, the specialization in pump design differentiates them in their operation. The in-line pump consists of a plunger and an element pair for every nozzle holder mounted in-line, where each plunger element pair is spring loaded on a cam, which develops pumping action. All the plunger and the element assemblies are connected by a single control rod, which uses a rack and pinion method to control the position of the plunger's helix with respect to the fuel relief port on the barrel/element side. The position of the helix defines the effective stroke of the plunger for fuel delivery; maximum stroke defines the highest quantity to achieve the full power. The control rod is operated by a governor, which adjusts the control

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

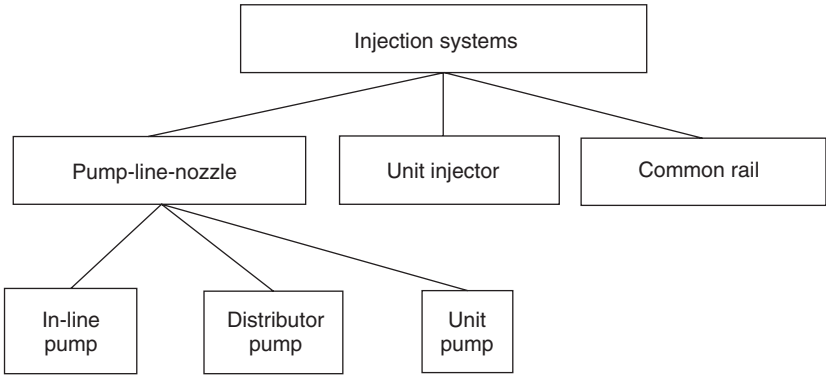


Figure 1. Diesel fuel injection system categorization. (Reproduced by permission of FEV. Inc.)

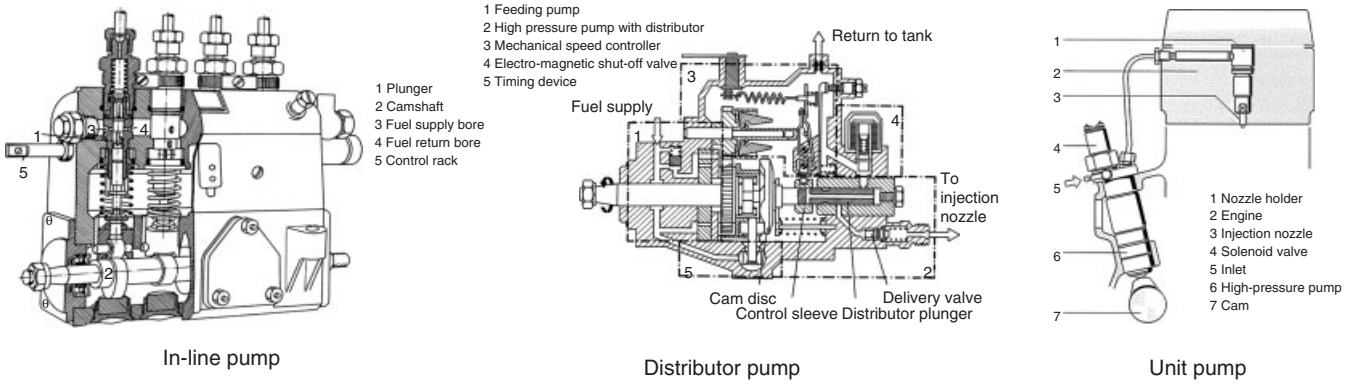


Figure 2. Pump-line-nozzle-type fuel injection systems. (Reproduced by permission of Robert Bosch GmbH.)

rod position to refine the fueling command to meet the idling and safeguard the engine against over-speeding by opening the helix early to cutoff the fuel. The unit pump is a miniature version of the in-line pump, where the pump element for every injector is separated and mounted directly on the engine; the unit pump is operated by a cam on the engine.

The distributor pump has a unique design, where the pump has one plunger with multiple grooves equivalent to the number of injectors connected; the plunger is operated by a single circular cam. A sleeve on the plunger the fuel relief port is controlled by governors that define the fuel quantity. The plunger moves both radially and axially, the axial movement creates pressure, whereas the radial movement creates injection distribution to each nozzle holder. This feature of distribution of high pressure fuel to each injector line by single plunger element coined its name as distributor pump.

As shown in Figure 3, the nozzle holder assembly is the other important part of the pump-line-nozzle fuel injection system. It consists of a nozzle body that is micromachined to carry fuel delivery holes, a needle that is spring loaded to open and close the fuel path to nozzle holes, and a nozzle holder body that houses high pressure fuel and leakage paths. The high pressure fuel enters the injector body through a filtered fuel path into the nozzle body, which houses the needle. When the pressure inside the fuel chamber increases, it lifts the needle against the spring force and opens the nozzle holes to make way for the fuel to inject.

1.1.2 Unit injector

Shown in Figure 4 is a working principle of a unit injector system. The high pressure pump and injector are assembled together in a single housing; the pump is operated by a rocker arm together with a cam on the engine. During the suction stroke, the pump chamber is filled via the open solenoid valve, and during the compression, the solenoid closes to create high pressure and injection. The injection cutoff is regulated by the opening of the solenoid.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

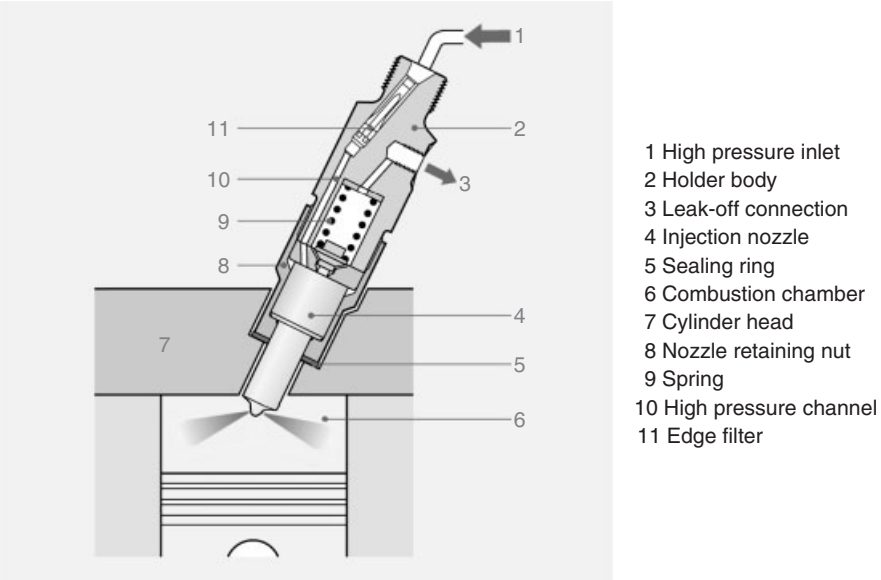


Figure 3. Single spring nozzle holder assembly. (Reproduced by permission of Robert Bosch GmbH.)

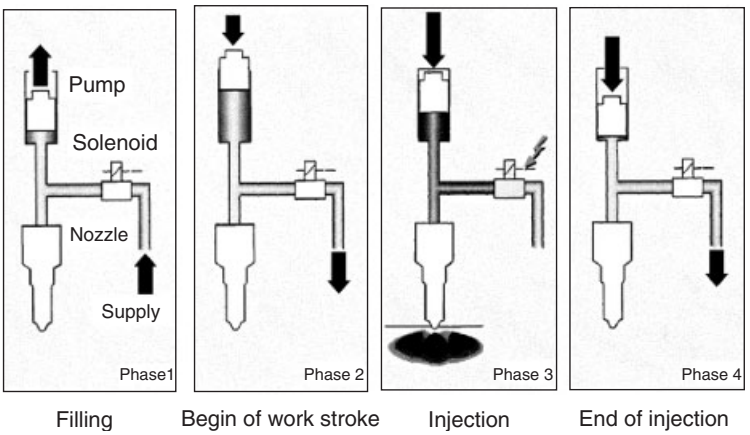


Figure 4. Working principle of the unit injector. (Reproduced by permission of Robert Bosch GmbH.)

1.1.3 Common rail fuel injection system

The key factors of a diesel engine fuel system are shown in Figure 5 (Mahr, 2002). Over the past few decades, the demands placed on diesel engines to operate at low noise and reduced harmful emissions level resulted in the introduction of common rail fuel injection systems. The most known feature of the common rail fuel system is the decoupled rate of injection/fuel injection pressure and start of injection compared to any of its predecessors in diesel fuel injection.

The advantages of a common rail fuel injection system along with advanced electronic control systems paved the way to address the diesel engines efficient combustion demands such as multiple injections, high pressure fuel injection, and broad range for start of fuel injection. Figure 6 shows the common rail fuel injection system hardware components (Robert Bosch GmbH, 1998): the common rail fuel system consists of high pressure fuel pump, common rail fuel reservoir, high pressure fuel lines, and injectors. The constantly available fuel supply enables the highest flexibility regarding injection timing for multiple injection events per cycle. Furthermore, the injection pressure can be adjusted nearly independent from engine speed and load. The electronic control unit (ECU)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

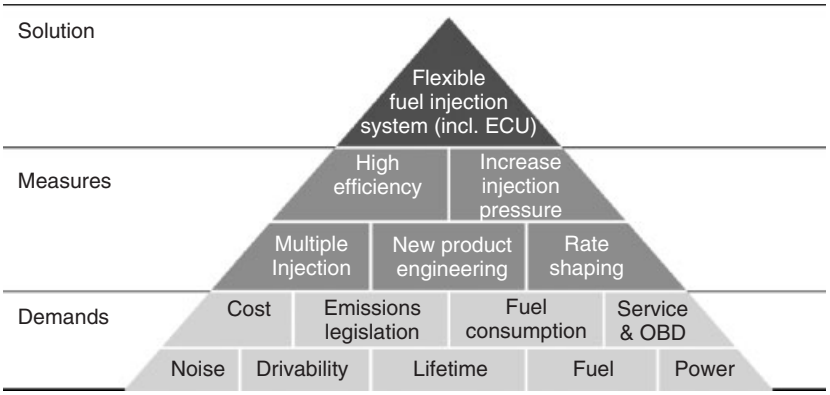


Figure 5. Key factors that influenced the diesel fuel injection systems. (Reproduced from Mahr, 2002. With kind permission from Springer Science+Business Media.)

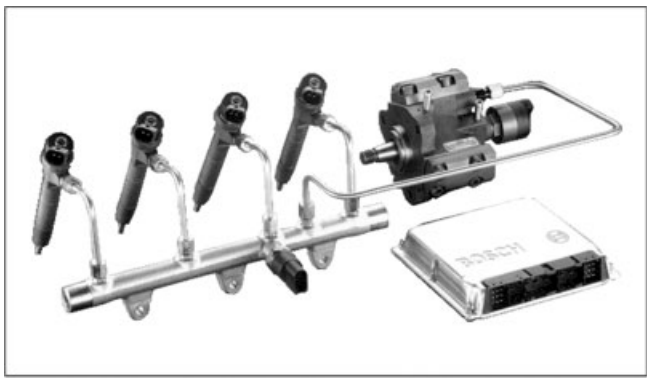


Figure 6. Bosch common rail system. (Reproduced by permission of Robert Bosch GmbH.)

is integrated around the fuel system and engine to ensure controlled operation. The state-of-the-art EDC fuel management by a common rail fuel injection system continues to evolve and expand the horizon of engine management to vehicle management and addresses stringent emissions and fuel economy demands.

Figure 7 shows a schematic diagram of an electronically controlled common rail system for engine management (Robert Bosch GmbH, 1998). The sensors covering both, the engine and the common rail system, send signals to the ECU, whereas the actuators covering both, the engine and the common rail fuel system, receive signals from the ECU. The fuel management portion comprises low and high pressure circuits. The low pressure circuit comprises the fuel tank, fuel filter, presupply pump, and fuel-connecting lines. The components of the high pressure circuit are high pressure pump, fuel accumulator (rail), fuel injectors, and high pressure fuel lines.

The responsibility of the low pressure fuel circuit is to deliver a sufficient volume of fuel to the high pressure circuit and to return the excess fuel from the high pressure circuit back to the fuel tank.

The high pressure fuel circuit has to maintain the fuel in the common rail at desired pressure at all times, which is crucial for fuel delivery and precise control through full engine operation.

The high pressure pump, as shown in Figure 8, is made up of three plungers operated by an eccentric shaft (Robert Bosch GmbH, 1998). For the first-generation common rail system, the fuel pressure is controlled by a rail pressure control valve (PCV). The fuel flow is defined by the pump displacement and the pump speed. The high pressure fuel flow, which is not required by the engine, is spilled by the rail PCV into the low pressure circuit. This kind of rail pressure control has a poor efficiency and generated high fuel temperatures in the fuel return line. Later versions of common rail pumps (shown in Figure 9) are quantity controlled by a metering unit, which delivers only the required amount of high pressure fuel into the rail. A high pressure fuel line connects the pump to the common rail.

Figure 9 shows a later version of a common rail pump, which is designed for future increased rail pressures of up to 2400 bar (Leonhard *et al*., 2008). The high pressure leakage along the plunger is reduced by using a much longer plunger guide. The flat tappet of the older pump version has been replaced by a roller tappet for better lubrication. Furthermore, the speed range of the pump displayed in Figure 9 is larger. This common rail pump is typically driven at crankshaft speed, which improves lubrication.

The fuel accumulator, which is also known as the *common rail*, is shown in Figure 10 (Robert Bosch GmbH, 1998). The common rail consists of a fuel accumulator with high pressure connections, pressure sensor, PCV, and fuel

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

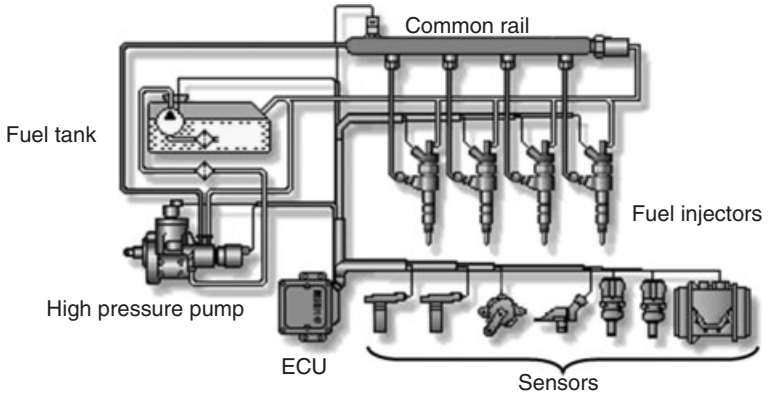


Figure 7. Common rail fuel injection system. (Reproduced by permission of Robert Bosch GmbH.)

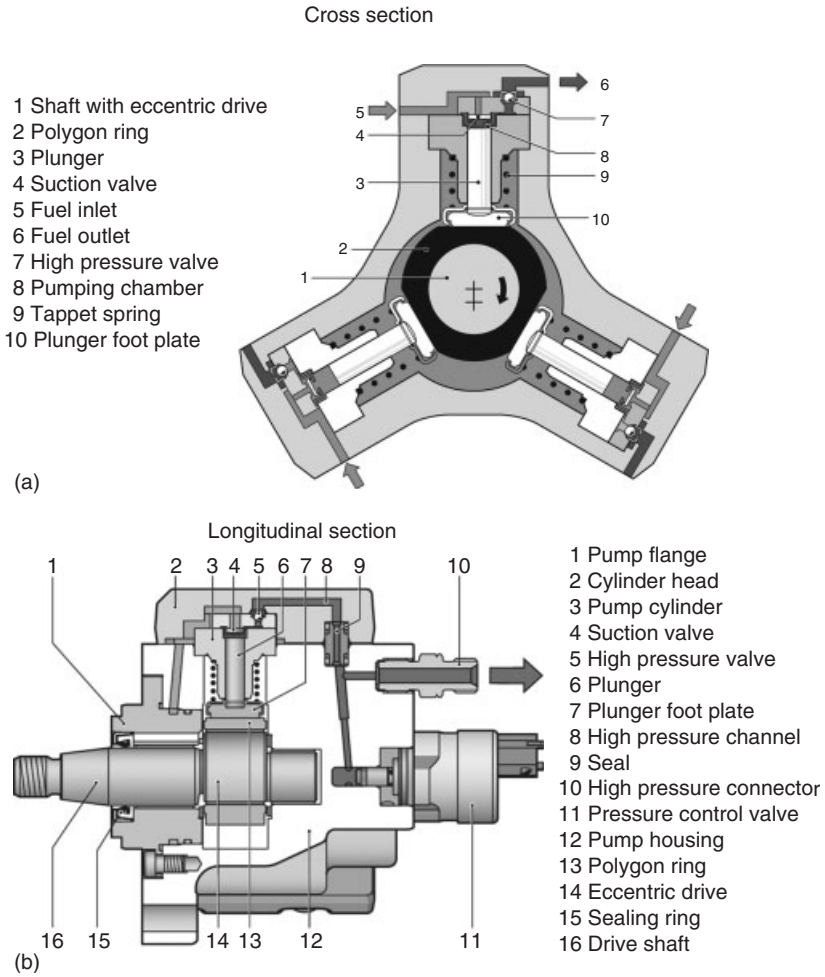


Figure 8. (a, b) High pressure fuel pump with rail pressure control valve. (Reproduced by permission of Robert Bosch GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

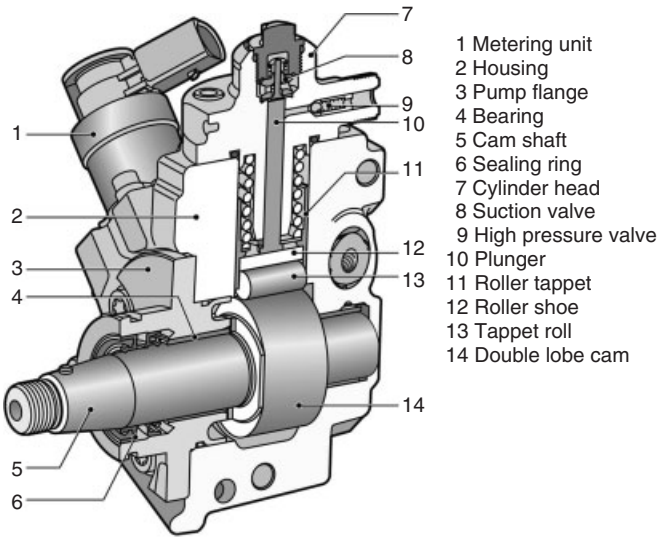


Figure 9. High pressure common rail pump (Leonhard *et al*., 2008). (Reproduced by permission of Robert Bosch GmbH.)

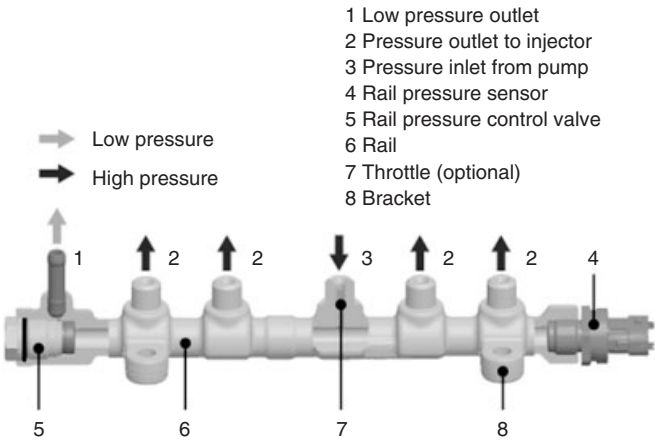


Figure 10. High pressure fuel accumulator—common rail. (Reproduced by permission of Robert Bosch GmbH.)

return connector. The common rail is made of high grade materials; therefore, it can withstand high fuel pressures under all conditions. The pressure sensor provides feedback to the ECU for closed-loop pressure control, whereas the PCV is a pressure regulator on the common rail for pressure management. Usually, both metering unit on the high pressure pump and the PCV are used together to achieve the desired pressure. The metering unit is used for increasing the pressure commands, whereas the PCV is used for decreasing it. However, the logics share the appropriate responsibility with each other to meet the dynamics of pressure management. As the common rail name states, all fuel injectors are connected to a common rail having

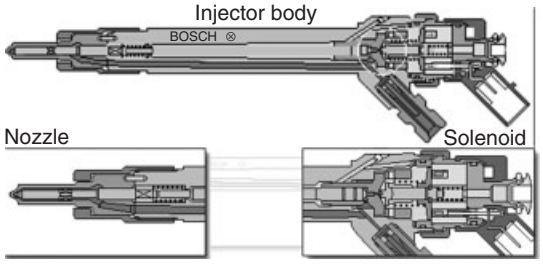


Figure 11. Injector cross section (first and second generations). (Reproduced by permission of Robert Bosch GmbH.)

high pressure fuel available for injection. This enables the fuel injection system to operate all of the injectors at the same precise fuel pressure, which results in a more uniform fuel management among all cylinders. This aspect of the common rail fuel system allows it to provide uniform fuel injection rate control to all cylinders over the entire engine operation range.

Such a hardware design change from all the previous fuel injection equipment enabled smooth engine operation, better drivability, dynamic response, and low emissions and fuel consumption. In addition to achieving high pressure and stable pressure management through the entire engine operation range, one of the most advanced parts of common rail fuel injection system is the fuel injector. Figure 11 shows the three main components of a common rail fuel injector (Robert Bosch GmbH, 1998). The injector body, which is the outer shell of the injector, also carries fuel paths for high pressure supply and low pressure return. The design and construction of the injector body is very robust and machined precisely to meet the packaging needs on the engine and cooling of the injector components by diesel fuel.

The solenoid is electrically operated by the ECU; a high current opens and closes the solenoid valve to actuate the injection. The solenoid valve in turn opens and closes the pressure relief orifice inside the injector body to alter the pressure above the nozzle needle. The nozzle, as shown in Figure 11, comprises a nozzle body, which has fuel injection holes, and a needle that defaults to the closed position by the control rod. On the top of the control rod and at the seat side of the needle, there is a pressure chamber used to move the needle to allow fuel injection. As stated earlier, when the solenoid is operated, it relieves the fuel pressure in the control chamber at the control rod, whereas the lower chamber of the needle pressure remains at the same pressure as the high pressure common rail. This pressure difference across the needle and the control rod moves the needle up, thus opening the fuel holes inside the nozzle body, which allow the fuel to spray into the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

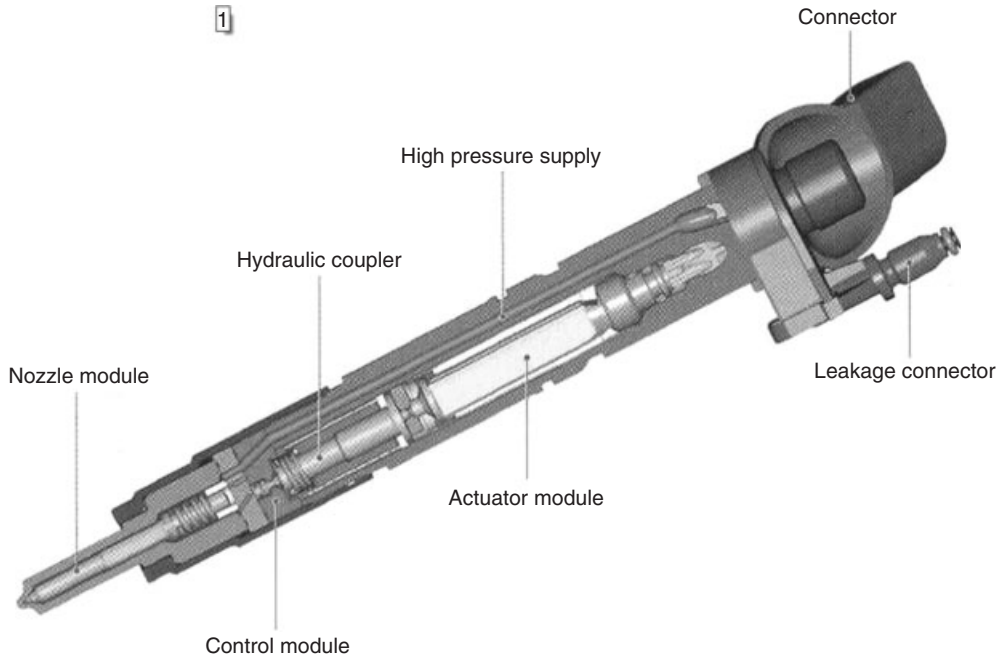


Figure 12. Common rail injector with piezoelectric actuator (Mattes, Boecking, and Kampmann, 2004). (Reproduced by permission of Robert Bosch GmbH.)

combustion chamber. When the solenoid valve closes, it prevents the fuel relief from the control chamber. The increasing pressure on the top end of the control piston, with a slightly larger diameter than the nozzle needle along with the spring force, closes the needle to stop the fuel injection. The drawbacks of this injector are the permanent high pressure leakages around the needle shaft and the control piston. Furthermore, the large combined mass of the moving parts reduces the opening and closing velocities of the nozzle needle.

The third-generation injector, as shown in Figure 12 (Mattes, Boecking, and Kampmann, 2004), uses a piezoelectric actuator. The servo-control chamber is located on the rear side of the nozzle needle. This means that there are no static leakages as with the injector type discussed and shown in Figure 11. Furthermore, the moving mass is significantly reduced, which enables faster needle opening and closing velocities. The smaller needle mass reduces the nozzle seat load, which offers the higher needle dynamic. The piezoelectric actuator provides higher forces and shorter response time between start of energizing and start of injection, which allows for higher fuel metering accuracy.

The nozzle module is shown in more in detail in Figure 13a. The spring presses the sleeve against the top plate and separates the control chamber on the rear side of the nozzle needle from the high pressure area. The needle guide is located in the nozzle shaft region and thereby, closer to the nozzle seat.

The desired injection rate is displayed in Figure 13b (Busch, 2004). A relatively low injection rate should be reached during ignition delay, in order to reduce noise and NO_x emissions. After the start of combustion, a strong increase in the injection rate is important to achieve short injection duration and a fast combustion. At the end of the injection, a fast decrease in the injection rate is important.

The integrated electrical and hydraulic systems designed in the state-of-the-art common rail diesel injector allow for multiple injections in each combustion phase. This is one of the key features of the common rail system designed to address performance, emissions, and NVH (noise, vibration, and harshness) needs as shown in Figure 14 (Mahr, 2002). The early small injections are called *pilot injection*. Pilot injection helps minimize the ignition lag for the main injection, while minimizing the heat-release rates that reduce combustion noise and lower NO_x formation. The main injection is targeted to meet the torque and power demands, whereas postinjections close—coupled to main injection help minimize the particulates by late combustion cycle oxidation. The late post-injection is used for providing fuel into the exhaust to create an exothermic reaction across a DOC (diesel oxidation catalyst) for DPF (diesel particulate filter) regeneration or any exhaust gas treatment device that requires high temperatures to operate.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

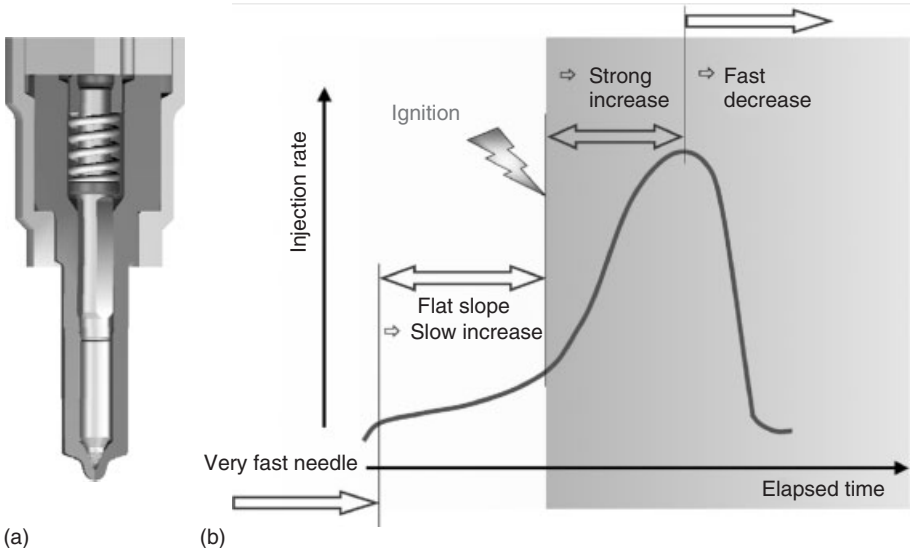


Figure 13. (a) Nozzle cross section and (b) injection rate requirement (Busch, 2004). (Reproduced by permission of Robert Bosch GmbH.)

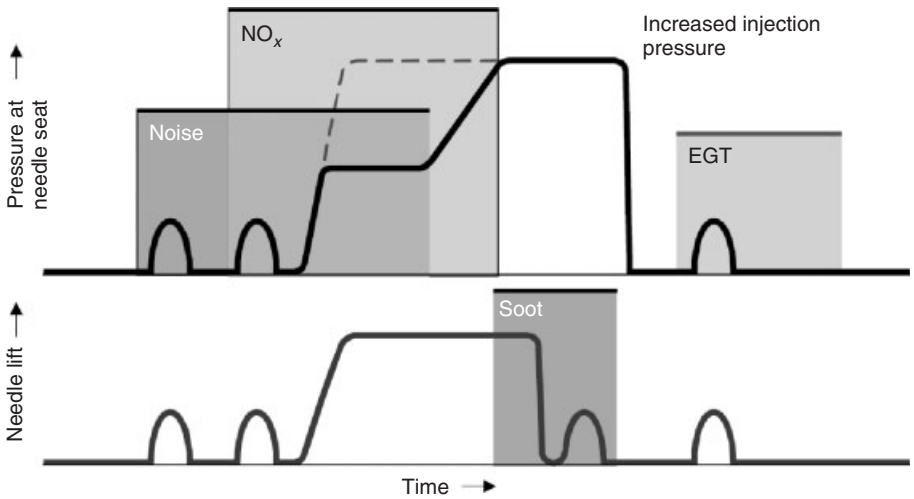


Figure 14. Desire for multiple injection and injector. (Reproduced from Mahr, 2002. With kind permission from Springer Science+Business Media.)

The nozzle is a very important part of the injector designed to meet the emissions, performance and NVH requirements of the combustion process. As shown in Figure 15 (Mahr, 2002), the nozzle geometry has evolved to address some of the diesel fuel injection needs. The needle seat, where the needle rests, has been found to influence combustion noise and fuel leakage. The sac volume, which is the dead volume between needle seat and nozzle hole, influences the hydrocarbon (HC) emissions by allowing fuel to reach the combustion chamber during the exhaust stroke. The nozzle hole geometry was developed to address unique emissions and fuel delivery efficiency needs.

The valve covered orifice (VCO) nozzles were introduced for cam-driven injection systems such as distributor injection pumps. In comparison to the mini sac hole nozzle, the VCO nozzle has a very low parasitic volume below the needle seat, which reduces HC emission. In combination with a common rail injector, which is working with lower opening and closing velocities, the VCO nozzle shows an

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

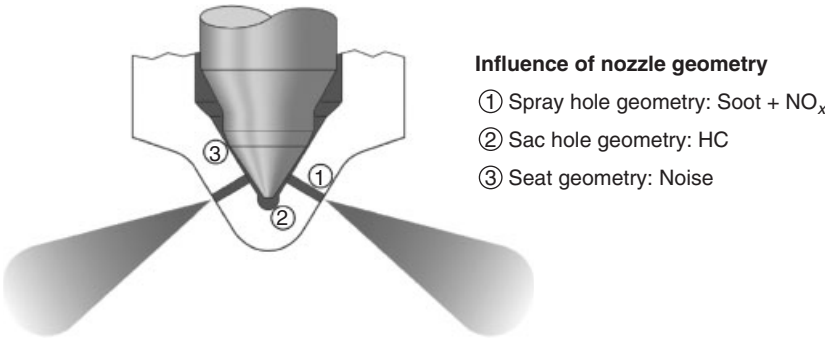


Figure 15. VCO nozzle geometry. (Reproduced from Mahr, 2002. With kind permission from Springer Science+Business Media.)

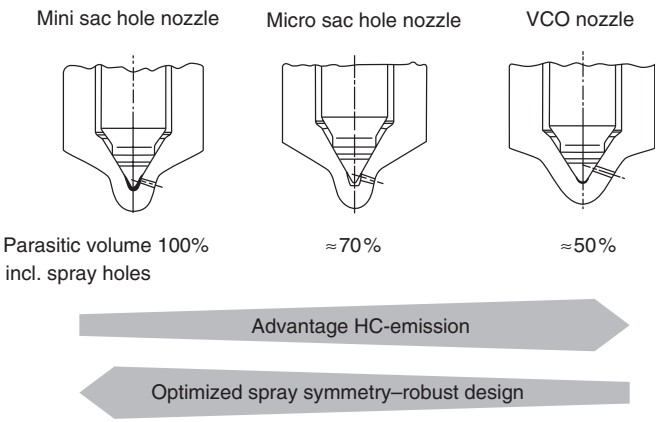


Figure 16. Influence of the nozzle design on spray pattern and HC-emission (Potz *et al*., 2002). (Reproduced by permission of Robert Bosch GmbH.)

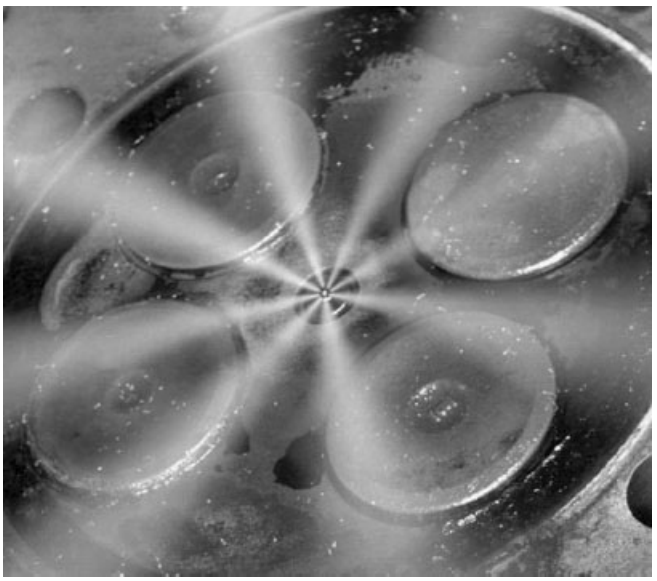


Figure 17. Fuel spray by an eight-hole nozzle inside the combustion chamber. (Reproduced by permission of FEV. Inc.)

asymmetric spray pattern because the needle tip is not always concentric with the nozzle seat. For this reason, the so-called micro sac hole nozzle has been developed, which provides a symmetric spray pattern with a compromise in the parasitic volume. Figure 16 shows examples of various nozzle types (Potz *et al*., 2002).

The state-of-the-art nozzles have anywhere between six and eight holes with a diameter in the range 0.1–0.15 mm for passenger car engines. The diameter of the nozzle holes depends on the minimum-to-maximum fuel injection quantity necessary to meet the power requirement with the given number of holes. However, improvements in nozzle hole geometry have resulted in much smaller holes with improved hydraulic properties that enhance the combustion process while minimizing emissions. Smaller spray hole sizes are possible, yet are difficult to manufacture for series production and are at greater risk of nozzle coking. The spray pattern of a modern eight-hole nozzle mounted in the cylinder head is shown in Figure 17. The fuel is injected at ambient conditions.

Figure 18 shows some examples of new nozzle designs (Mahr, 2002). Hydro-erosive rounding of the spray hole entrance allows a more stable static flow rate by reducing the run-in effect. Furthermore, the tolerances can be reduced because the process of hydro-erosive rounding can be stopped when the target flow rate is reached. The *k*-factor is a measure of conical hole; conical holes were developed to reduce the cavitation phenomena in the spray hole and hence to improve the spray penetration especially when using reduced spray hole diameters. Better spray quality helps improve fuel and air mixture allowing for reduction in particulate matter, which forms during rich combustion phases. The spray distribution among the nozzle holes is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

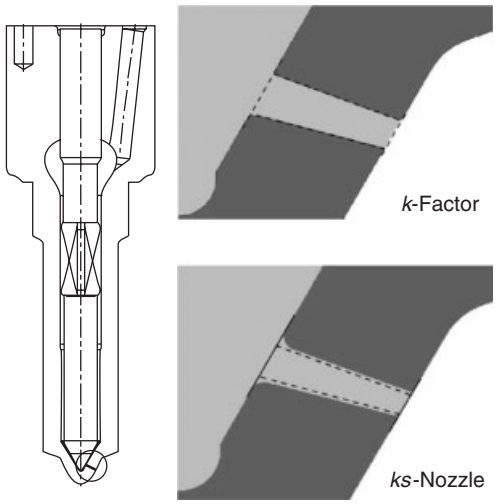


Figure 18. Nozzle hole profile. (Reproduced from Mahr, 2002. With kind permission from Springer Science+Business Media.)

extremely important to address improved air–fuel mixture distribution as well as improved combustion.

The *ks*-nozzles represent an improved conical profile combined with an optimized hydro-erosive grinding, in order to achieve highest discharge coefficients combined with highest spray penetration for optimized air–fuel mixture preparation by avoiding cavitation effects. On the other hand, eliminating spray hole cavitations causes a drawback in terms of deposit formation at the spray hole wall, which may end up in an engine power loss of 5% or even more. To reduce deposit formation in the spray hole, fuel contaminants such as zinc (Zn), sodium (Na), or potassium (K) must be avoided. A Zn concentration of 0.5 mg/kg in the fuel is already critical in terms of engine power loss. For this reason, engine oil-lubricated fuel systems in combination with flow-optimized nozzles may cause a loss of power because of Zn additives in the lube oil.

Continuous development of advanced electronic fuel management, improved fuel injection components, and precision manufacturing enabled more precise fuel management. As shown in Figure 19, the ECU now holds some of the most complex algorithms to improve the fuel delivery and durability (Busch, 2004). Algorithms such as injector quantity adjustment (IQA) are is used to adapt energizing time (ET) of the solenoid for each injector based on the tolerance difference with respect to nominal component at all fuel delivery ranges, allowing for more uniform fuel delivery among all injectors. Fuel balance control (FBC) works on minimizing the torque fluctuations between cylinders and each combustion stroke to make the engine run smoothly. Similarly, pressure wave correction (PWC) is used to adapt for hydraulic wave impact on fuel delivery. There are a variety of algorithms and intrusive tests developed over time to meet the NVH, efficiency, durability, and performance of the fuel injection system.

2 ELECTRONIC DIESEL CONTROL SYSTEM

EDC systems were blended into the mechanically actuated fuel injection systems and introduced to the market during the 1980s. These mild electronic control systems started to address the need for precise and more flexible

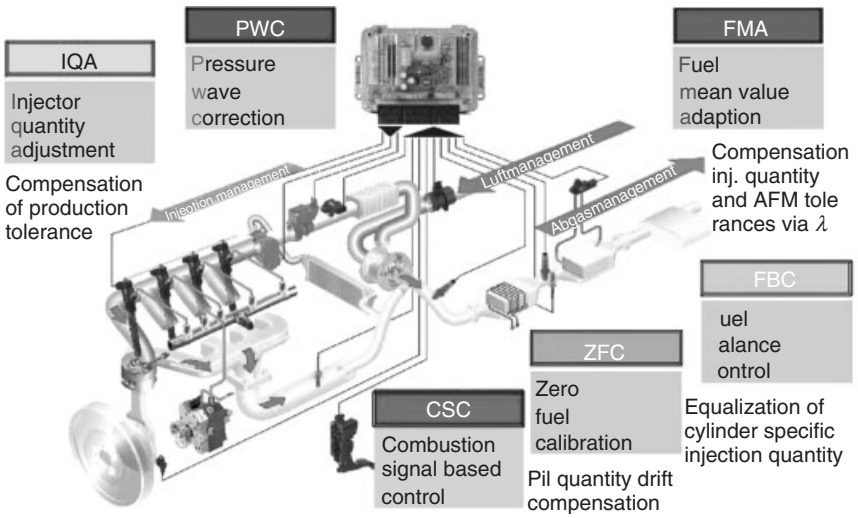


Figure 19. Fuel delivery corrections. (Busch, 2004. Reproduced by permission of Robert Bosch GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

fuel injection. Further on the control system, application widened the scope to full engine and vehicle management systems. The present state-of-the-art diesel engine control system manages a wide spectrum of propulsion management ranging from engine, aftertreatment, driver demand, safety, and comfort. In this section, a brief overview of electronic diesel fuel injection systems is discussed.

2.1 Technical overview of electronic control

Over the past decades, a demand for more fuel-efficient engines with low emissions (such as NO_x, HC, CO, and PM) combined with higher power and torque output to enhance drivability has played a crucial role in bringing electronic control systems into engine management. Electronic control of a diesel engine allows precise and differentiated modulation of fuel injection allowing for control over the engine operation to meet the demands placed on it. As shown in Figure 20, the electronic control system comprises sensors, actuators, and ECU (Robert Bosch GmbH, 1998).

1. The sensors provide the ECU with the operating conditions of the engine such as engine speed and angular position with respect to top-dead center (TDC), amount of airflow, coolant temperature, oil temperature, and exhaust gas condition.
2. The ECU processes the information received from sensors and desired value calculation modules within the ECU through a systematic mathematical calculation sequence (control algorithm). The output of these algorithms results in electrical signals to actuators to develop necessary reactions or controls.
3. The actuators convert the electrical signals into physical reactions. If the ECU commands to open a valve by means of an electrical duty cycle, the actuator in this case, the injector solenoid valve moves the needle in the direction commanded by ECU to inject fuel as desired to meet the torque demands.

2.2 Signal acquisition and processing

Sensors and actuators play an interfacial role between the physical operating systems, in this case, an engine and ECU. The sensors provide the state of the system in the form of an electrical signal. There are several types of sensors used on the engine to determine its state of operation; few are discussed in this section. Shown in Figure 21 is a schematic diagram of signal handling by the ECU (Robert Bosch GmbH, 2004). The electrical signals from sensors are processed inside the signal conditioning unit of the ECU where further signal is transferred to a microcontroller for processing the values inside a logic defined by a software strategy. The ECU's decisions are sent out through driver stages to actuators.

Analog signals: These are voltage-based signals such as temperature sensors, boost pressure sensors, and battery voltage. These assume a voltage range and are processed to physical values inside the signal conditioning unit; further, the values are converted into the digital format for the ECU to process them through the logical algorithms.

Digital signals: These signals assume high [1] or low [0] state, sensors such as switches or digital signal generators,

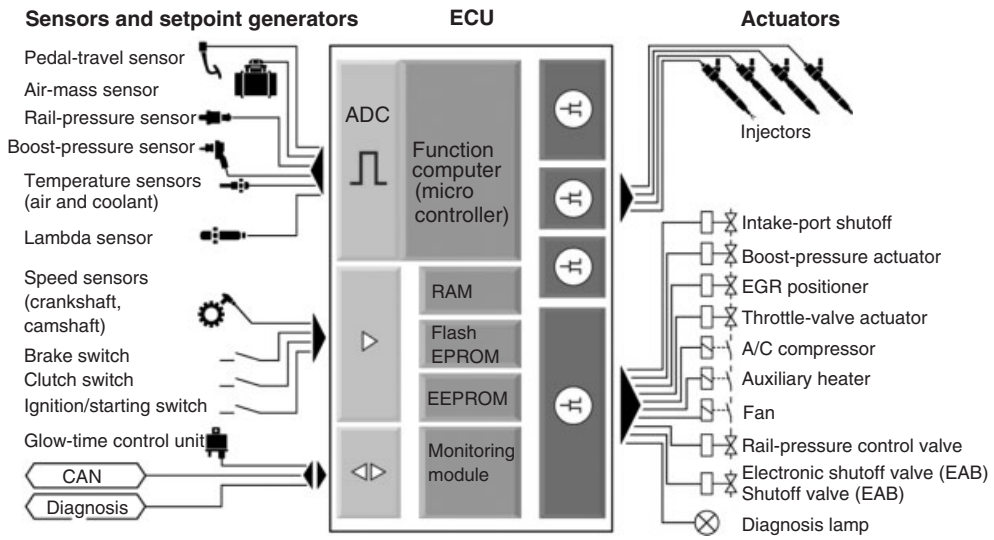


Figure 20. Electronic control system modules. (Reproduced by permission of Robert Bosch GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

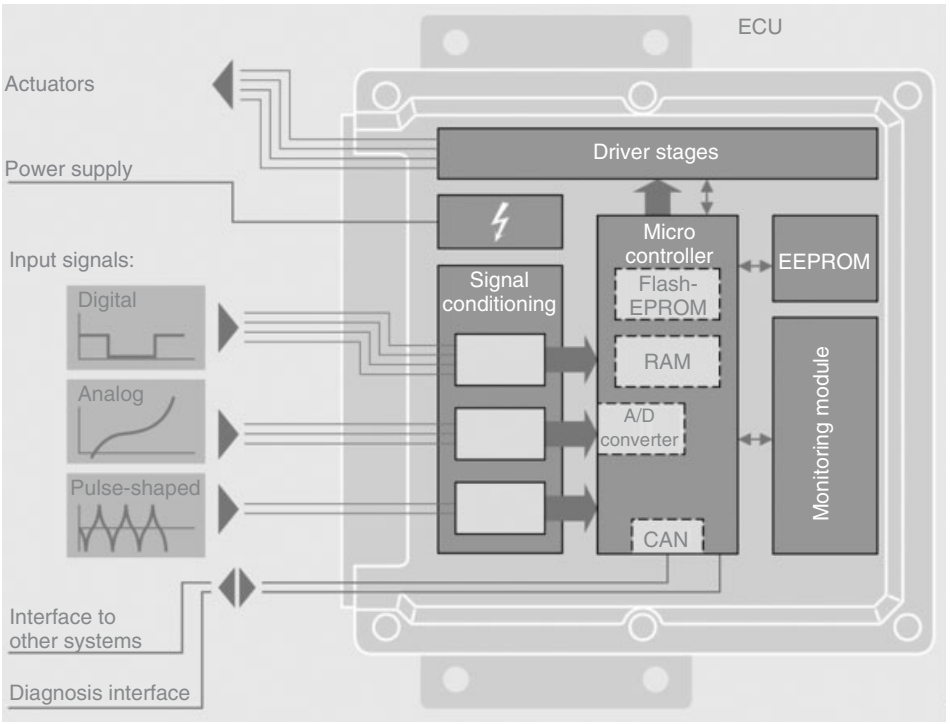


Figure 21. Signal processing inside ECU. (Reproduced from Robert Bosch GmbH, 2004. © John Wiley & Sons.)

such as Hall effect sensors picking up rotational speed will be processed inside the ECU directly.

Pulsed signals: These are inductive sensors such as variable reluctance speed sensors. These signals need processing inside the ECU to convert them into digital signals and then into a physical representation before being processed inside the ECU.

Signal conditioning: Inside the ECU, protective circuitries are used to limit and regulate the sensor voltage range. Filtering techniques are applied and the superimposed interference signals are removed before processing. Sometimes, the signals are amplified to meet the microcontrollers, working range of 0–5 V. For advanced sensor technology such as microelectromechanical systems (MEMS) sensors, the signal conditioning can be performed at the sensor level to reduce the burden on the ECU processing power.

Signal processing: The ECU performs signal processing to execute the tasks necessary to manage engine operation. The open- and closed-loop sequences along with the associated algorithms are performed inside the ECU upon receiving the sensor signals to take appropriate actions. The microcontroller is the ECU's central control unit and operates all ECU functions. The microcontroller consists of RAM (random access memory), ROM (read only memory), central processing unit (CPU), Quartz timer, serial interfaces, and peripheral interfaces. The CPU accesses the software that is stored in EEPROM (electrically erasable programmable read only memory), which defines the sequence of activities that the microcontroller should execute. The working software is downloaded to flash-EEPROM to process the signals received. The computation needs memory space to execute the algorithms. This is done by storing and retrieving the data from RAM.

The output signals from the microcontroller trigger the driver stages, which further handle the actuator signals. The driver stages can handle various types of actuators; pulse-width modulation (PWM) is one of the most used methods for controlling valves such as EGR (exhaust gas recirculation), intake air mass control throttle, and boost pressure modulation by waste gating on turbochargers. Advancements in actuators carrying unique ECUs are capable of processing the commands from the ECU directly. The output signals can be either a PWM signal or a CAN (controller area network) signal, further driving the high power circuitry involving motors for actuation. The CAN network is used to communicate with other peripheral ECUs.

With increasing safety and performance demands, the ECU itself has undergone a revolution. The ECUs should operate in real time, which means that the ECU should perform and respond in conjunction to physical processes

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

outside the ECU that demands high computational power to execute complex algorithms. Advancements in multilayer printed circuits, surface-mounted devices eliminating loose wires, and application-specific integrated circuits help make the ECU more compact, versatile, and robust. The ECU should withstand severe electromagnetic radiations to work under any severe conditions without failure. The construction technology of the ECU must meet the stringent demands of electromagnetic compatibility (EMC) tests and be able to withstand severe magnetic radiation, vibrations, and environmental conditions.

As outlined in Section 2.2, the ECU uses various closed- and open-loop logics along with complex algorithms to define the output actions. Thus, it is important to understand the concept of closed and open loops. Section 2.3 will discuss the basic idea behind closed- and open-loop control theories.

2.3 Concept of open- and closed-loop controls

Both open- and closed-loop controls are applied in EDC for engine management. The application is defined by the need and purpose of the control area. To describe these control systems, a simple example of room temperature control is considered for discussion.

As shown in Figure 22, the room temperature control can be managed by adjusting the valve that controls the airflow through a heat exchanger. In the open-loop control case, the position of the air control valve is reached using a characteristic curve, which is a function of set point. A characteristic curve is stored in the ECU, which determines the unique value for every desired room temperature set point. No adjustment to the control valve is made to compensate the variations in the room temperature with respect to the set point.

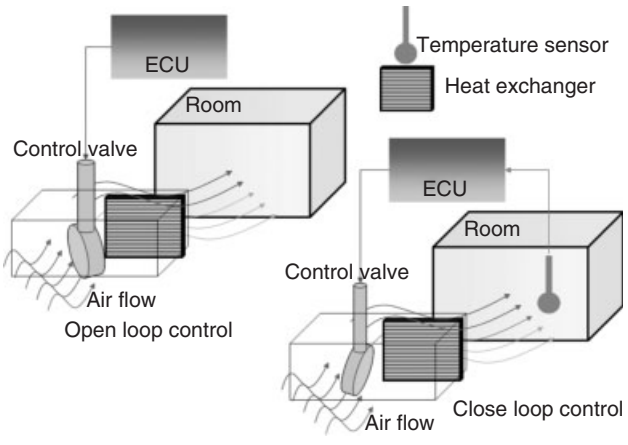


Figure 22. Open- and closed-loop examples for room temperature control. (Reproduced by permission of FEV. Inc.)

In the closed-loop control case shown in Figure 22, the ECU compares the set point with measured room temperature to determine the deviation between the desired and the measured temperatures.

This deviation is used to finely adjust the control valve to either open further or close to meet the set point request. The closed-loop control system allows for minimizing the error in maintaining the room temperature at the set point.

For more dynamic systems, a combination of predefined values from the open-loop control and deviation-based fine adjustment from the closed-loop system is used to provide faster response times.

The EDC system carries several such control system combinations to address air and fuel management during steady-state and transient engine operations.

2.4 Driver demand to fueling command

An engine is the main source of energy for propulsion in a vehicle; thus, the engine management is the center for safety, comfort, and performance. Figure 23 shows an example of a common rail system integrated into the vehicle, with the ECU to manage the fuel and air around the engine for efficient operation (Robert Bosch GmbH, 1998).

However, the rapid integration of electronics in every portion of the vehicle system such as transmission, braking, and body has opened up the doors for control system integration as shown in Figure 24. For vehicle management, multiple on-board ECUs are combined with an engine ECU that is used as the main controller for the power management. The engine ECU has to make significant interventions and satisfy demands received from various parts of the vehicle to ensure that engine power is managed. The demands of the driver need to be satisfied with consideration given to vehicle safety and drivability.

As shown in Figure 25, the accelerator pedal signal from driver demand passes through several signal interventions and conditioning before being converted to an equivalent fueling command. The interventions to the driver demand come from the other various control units on the vehicle such as the transmission controller. For example, the transmission controller is demanding torque reduction during gear change, whereas the traction control unit is demanding torque optimization to prevent wheel spin along with the electronic stability control asking for torque variation and split among the wheels to prevent the vehicle from going out of control. These interventions usually limit the driver-demanded torque. Driver demands are the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

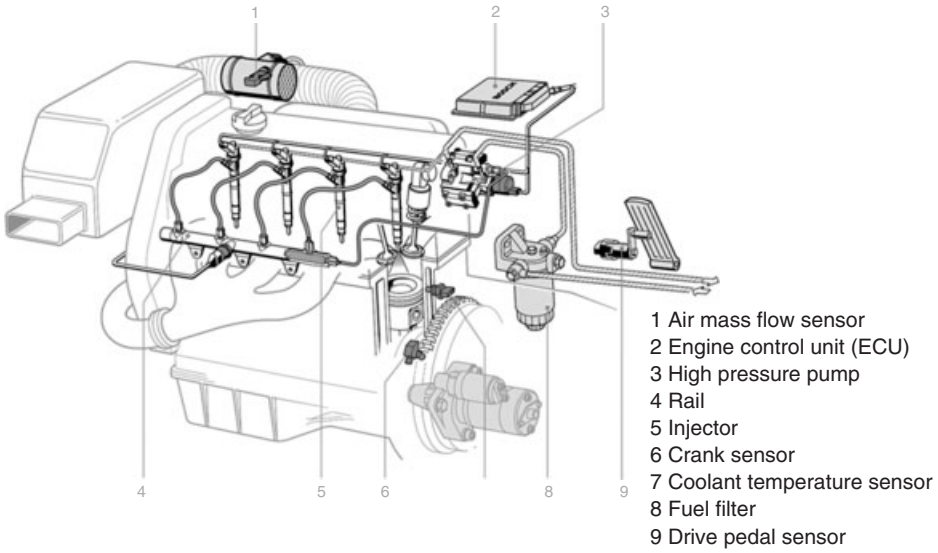


Figure 23. Common rail system installation on vehicle. (Reproduced by permission of Robert Bosch GmbH.)

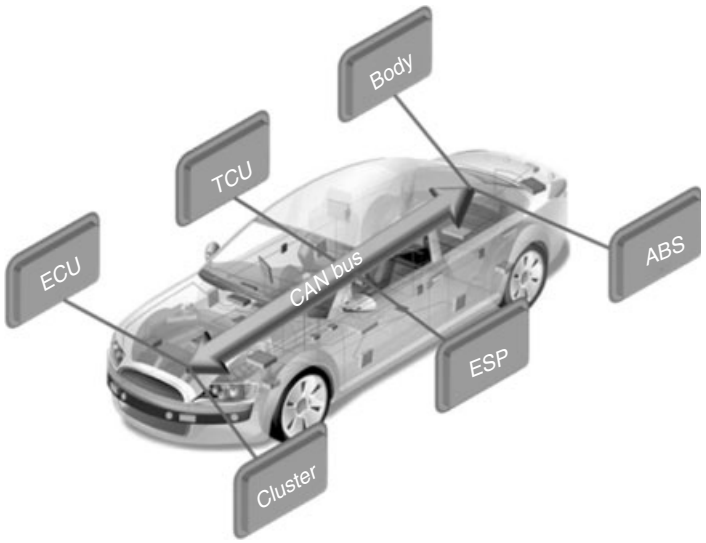


Figure 24. Vehicle management. (Reproduced by permission of FEV. Inc.)

highest of all torques; accessory load requests add on top of the driver demand to compensate for accessory power needs such as AC and alternator. However, the final torque request is limited against engine protection before being converted to fueling command. The engine protection algorithm inside the ECU limits the maximum fueling to protect the engine against maximum cylinder pressure. There are also engine performance-based torque limitations such as smoke limitations, which are dynamic and transient, based on the amount of air available at anytime in the combustion chamber.

Electronic control systems led to more efficient engine and vehicle management from energy efficiency to safety and performance, leading a way to integrate the state-of-the-art control systems.

Figure 26 shows vehicle operation control unit demands for a certain power output from the engine to overcome losses and meet driver demands. The driver is the closed-loop control element of engine torque; the driver decides how much power he or she needs at any given time. The driver demand results in speed and torque to operate the engine at desired power levels. The speed and torque

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

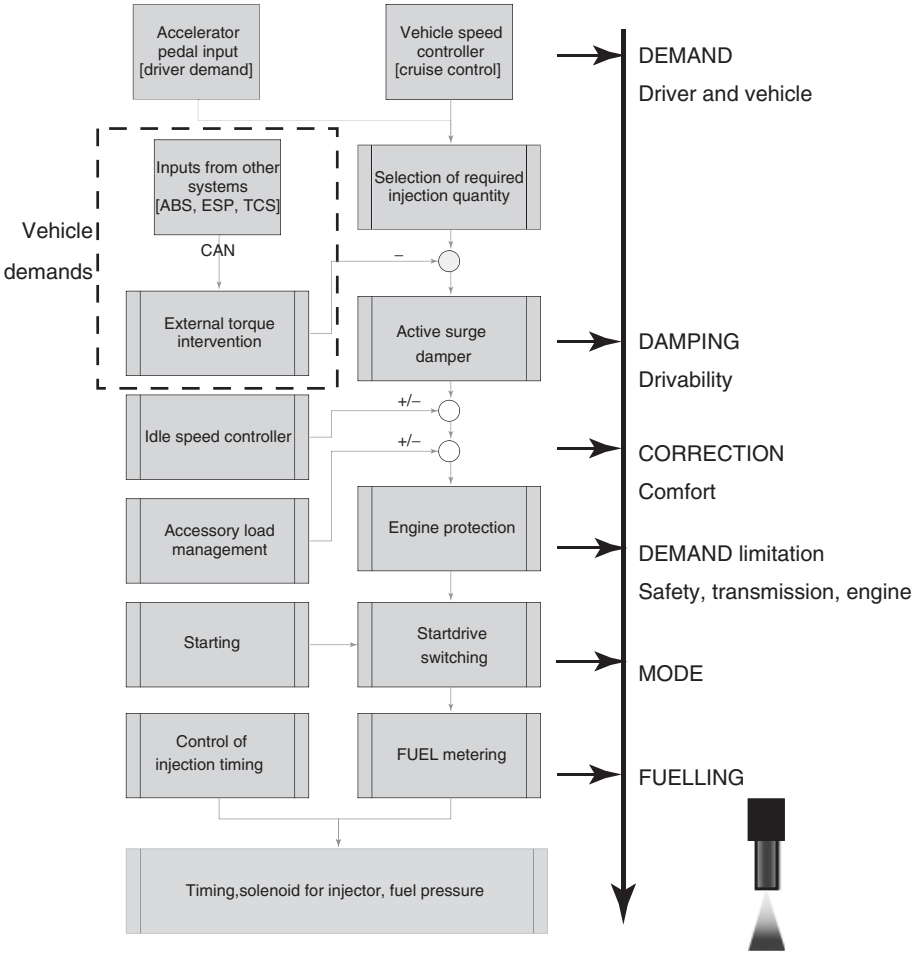


Figure 25. Fuel quantity calculation. (Reproduced by permission of Robert Bosch GmbH.)

defines the operating point of an engine, and the operating point further defines the desired air and fuel injection characteristics required to maintain engine operation at optimum performance.

In case of the diesel engine, the engine torque is defined by the fuel quantity; depending on the power requirement, the engine speed will be adapted to the defined fuel quantity. As shown in Figure 27, the diesel fuel injection is a very complex hydraulic process. Although the mass of fuel injected matters inside the combustion chamber for the energy to generate sufficient power, the fuel injection handling is done on volume basis. The fuel injection quantity is influenced by various parameters such as fuel pressure, fuel temperature, density, viscosity, nozzle hole profile or discharge coefficient, and the duration of nozzle opening. The nozzle opening time for fuel injection is defined in microseconds and is known as (*ET*). ET is a function of fuel pressure and desired fuel quantity defined in a map.

As outlined earlier, the operating point that is defined by engine speed and fuel quantity desired is used to define the rail pressure set point for further closed-loop control. This operating point also includes the injection timing, that is, when to open and close the injector to deliver fuel into the combustion chamber with respect to TDC. An example of driver demand to fuel quantity handling in a common rail fuel injection system is shown in Figure 28 (Robert Bosch GmbH, 1998).

Maps are three-dimentional matrices; the *X*-axis usually carries engine speed, whereas the *Y*-axis carries fuel quantity command required for a desired torque. The *Z*-axis carries the corresponding values for the set point determination, which defines the characteristic of a map. In the example shown in Figure 29, the maps define the set point for the start of injection in crank angle degrees with respect to TDC, which further sets the open-loop energizing system to operate the solenoid on the injector. However, rail pressure set point is first compared to the rail pressure sensors

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

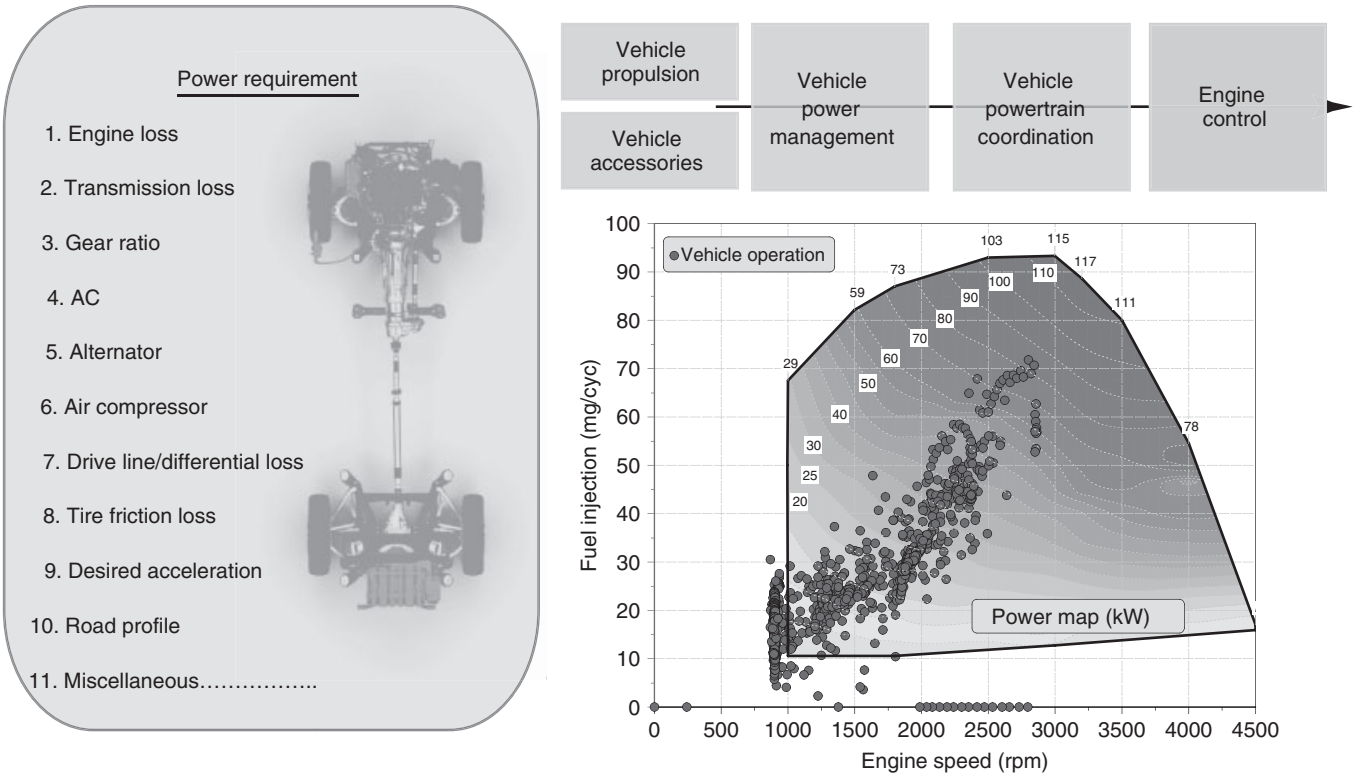


Figure 26. Power demands to engine control unit. (Reproduced by permission of FEV. Inc.)

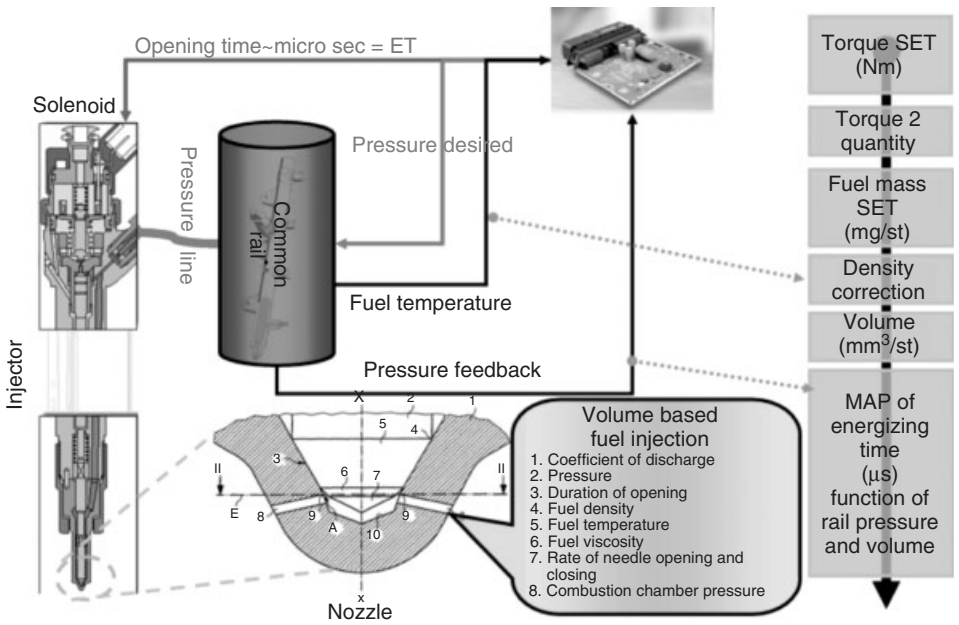


Figure 27. Fuel quantity delivery. (Reproduced by permission of Robert Bosch GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

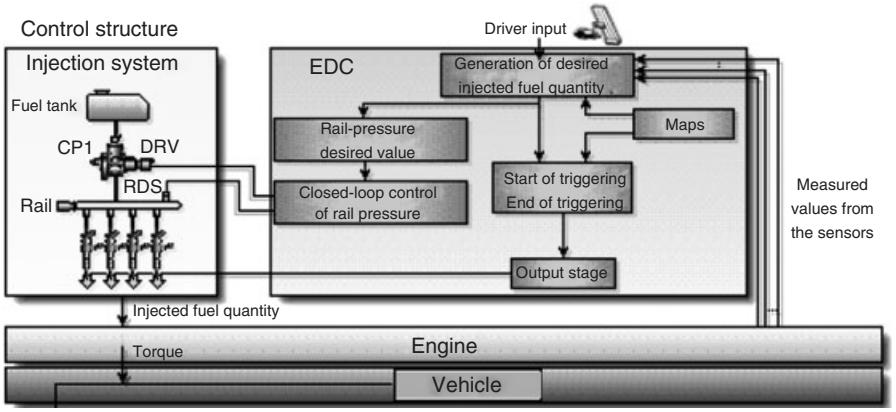


Figure 28. Driver demand to fuel injection control, an overview. (Reproduced by permission of Robert Bosch GmbH.)

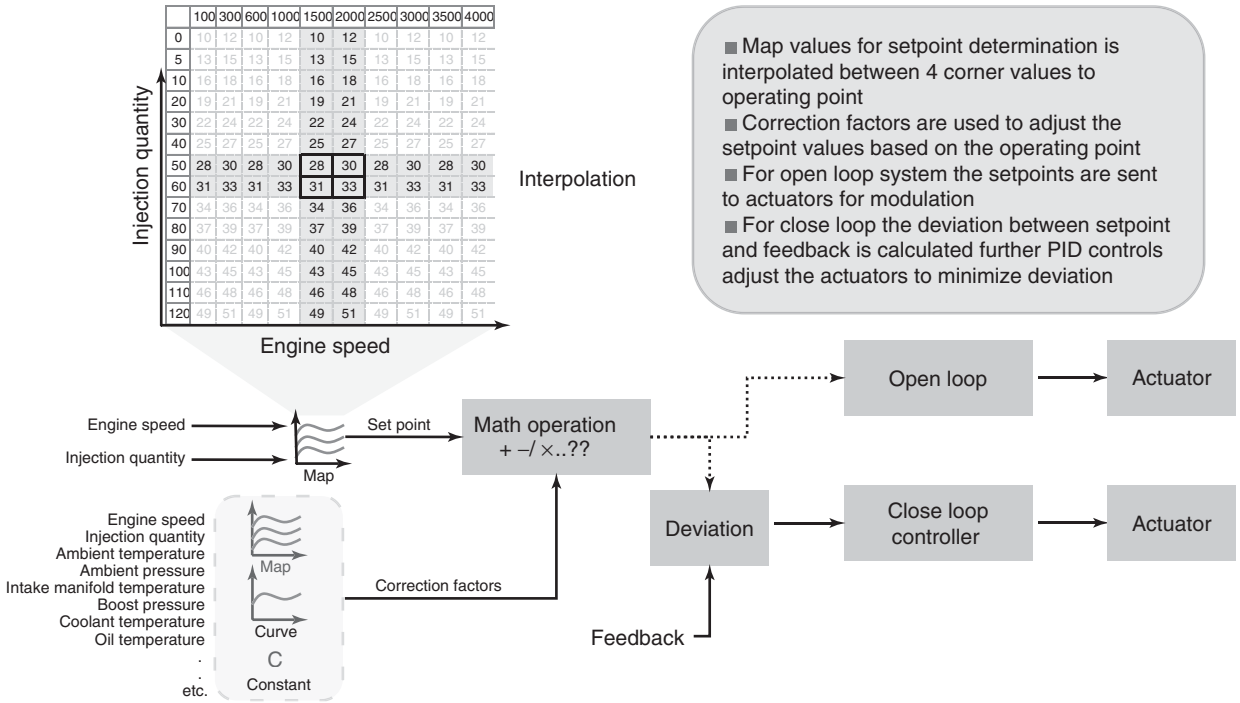


Figure 29. Set point determination and control. (Reproduced by permission of Robert Bosch GmbH.)

feedback and further closed-loop control of rail pressure is achieved. Figure 10 shows a brief overview of typical set point determination and flow of signals for the open- and closed-loop control architectures.

REFERENCES

Busch, R. (2004) Advanced Diesel Common Rail Injection System for Future Emission Legislation. *2004 Diesel Engine Emissions Reduction (DEER) Conference Presentations*, Coronado, CA.

Leonhard, R., Warga, J., Pauer, T., *et al*. (2008) *Bosch 2000 bar common rail system for passenger cars and light duty vehicles*. 29th International Vienna Motor Symposium.

Mahr, B. (2002) Future and Potential of Diesel Injection Systems. *Thiesel 2002, Conference on Thermo- and Fluid-Dynamic Processes in Diesel Engines*, Valencia, Spain.

Mattes, P., Boecking, F., and Kampmann, S. (2004) Piezo-Inline-Injektoren. Diesel- und Benzindirekteinspritzung, Tagung, Haus der Technik, Prof. Dr.-Ing. H. Tschöke, Berlin.

Potz, D., Christ, W., Dittus, B., *et al*. (2002) Dieseldüse–die entscheidende Schnittstelle zwischen Einspritzsystem und Motor. Dieselmotorentechnik 2002, Aktueller Stand und Darstellung

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

neuerster Entwicklungen und Entwicklungsziele, M. Bargende, U. Essers, Expert Verlag.

Robert Bosch GmbH (1998) *Bosch Common Rail Introduction*.

Robert Bosch GmbH (2000) *Bosch Automotive Handbook*, Bentley Publishers, Germany. ISBN: 9780837606149

Robert Bosch GmbH (2004) *Diesel-Engine Management*, 3rd edn, John Wiley and Sons, Ltd, Chichester.

FURTHER READING

Hoffmann, K.-H., Hummel, K., Maderstein, T., and Peters, A. (1997) Das Common-Rail-Einspritzsystem–ein neues Kapitel der Dieseleinspritztechnik. *MTZ*, **58** (10), 572–582.

Kampmann, S., Dohle, U., Hammer, J., *et al*. (2005) *Common rail systems to achieve oncoming EU emission standards*. 26th International Vienna Motor Symposium.

Leonhard, R. and Warga, J. (2008) 2000 bar diesel common rail by Bosch for passenger cars. *MTZ*, **69**(10), 834–840.

Robert Bosch GmbH (2001) *Electronic Diesel Control EDC*. The Bosch Yellow Jackets, Edition 2001.

Robert Bosch GmbH (2011) *Bosch Kraftfahrzeugtechnisches Taschenbuch*, 27th edn.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto150
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Drive Cycles

Mark E. Case and Dean Tomazic
FEV Inc., Auburn Hills, MI, USA

1 Introduction 1
2 Standard Drive Cycles and Regulatory Limits 2
3 Vehicle Testing Process and Measurements 16
4 Vehicle Testing Facilities 22
Useful Websites 25
Regional Websites 25
Abbreviations 26
Further Reading 26

1 INTRODUCTION

Starting in the late 1940s, research in the causes of air pollution accelerated, led by California in a world effort. Emissions of pollutants from light-duty vehicles were directly associated with poor ambient air quality in the Los Angeles basin. Government agencies began researching the causes for smog forming pollutants, and also vehicle test facilities, and simultaneously what might represent typical driving conditions of a vehicle in real-world conditions that could be simulated.

From this research, different approaches to representing the vehicle driving condition were made. In the United States, driving cycles were developed to represent the actual vehicle operation in traffic on public roads primarily in California. Traces of vehicle speed were recorded over time in city and highway driving environments. In Europe and Japan, typical operating modes were identified, and driving cycles were created to provide a representation of these typical modes. In either case, the goal is identical: represents the operating conditions of the powertrain in terms of vehicle speed along a time trace and provide a representative set of conditions under which the emissions and fuel economy of the vehicle could be measured against regulatory limits.

Development began on a test facility that could provide the capability to measure output of emissions and consumption of fuel from a vehicle in a controlled environment. To provide the capability to exercise a vehicle powertrain in a static facility in a way that is representative of actual use, it was desirable to make the ability to repeatedly operate the vehicle in an environment that simulates the operation of a vehicle on a road, but in a facility where instrumentation can be conveniently and permanently located. Standard measurements of gaseous emissions including carbon monoxide (CO), carbon dioxide (CO_2), unburned hydrocarbons (generically HCs), and oxides of nitrogen (NO_x) are made using an emissions bench that collects and samples the gases from the tailpipe of the vehicle. Because consumption of HC fuels produces CO_2 in proportion to the quantity of fuel consumed, the rate of fuel consumption, or inversely fuel efficiency, could also be measured.

Theoretically, the driving cycle and test facility provides also a repeatable standardized test that can be replicated and controlled.

The major driving cycles from the United States, Europe, and Japan were all standardized and written into each country's regulations for exhaust emissions and fuel economy. All three major regulatory regimes—the United States, Europe, and Japan—have modified or introduced supplementary driving cycles that are designed to create a more complete emissions relevant operating region under regulatory limits and monitoring. In the United States, these incremental cycles do not replace the first cycle but expand

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

the coverage of the operating map under legal control. The expanded coverage can be a wider vehicle operating range (e.g., aggressive driving) or ambient weather conditions (e.g., cold or hot climate).

The test facilities established for measurement of emissions and fuel economy have become essentially a world industry standard. In each major market, with individual drive cycles, the facility is usually capable of testing all cycles from all major regulatory regimes. In some cases, the regulated emissions constituents are different; for example, California regulates non-methane organic gases (NMOGs), whereas earlier US federal measurements were total hydrocarbons (THCs), and current limits are in non-methane hydrocarbons (NMHCs). The key components of an emissions/fuel economy test facility include the room with ambient condition control, chassis dynamometer that applies resistance representative of the load required to move the vehicle, and the emissions system that samples and quantitatively measures the gaseous emissions from the tailpipe, inferring fuel consumed by sampling CO_2 emissions. In this chapter, regulatory limits for the passenger car classification of light-duty vehicles are provided.

Because the driving cycle is intended to provide representative load conditions to the powertrain, there is a process to follow that establishes chassis dynamometer coefficients that adequately simulate the driving resistance in each specific vehicle. A reduction in the driving load generally reduces emissions of CO_2 and improves fuel economy.

Vehicle manufacturers are required to estimate the durability of the emissions control system of a vehicle. The degradation in emissions over vehicle usage is estimated and must be enacted and accounted for on typical emissions test results that are conducted on new or lightly used vehicles. It is implied that each vehicle must meet the regulated limit for each constituent for the entire useful life of the vehicle, and forms of demonstration of compliance from vehicles in consumer use are enacted in the major regulatory markets.

2 STANDARD DRIVE CYCLES AND REGULATORY LIMITS

2.1 The United States

In the United States, the original driving cycles developed for emissions and fuel economy were created by the Environmental Protection Agency (EPA) to represent driving in city and suburban driving conditions mainly in Los Angeles. Since the late 1970s, the standard emissions and fuel economy cycle applied toward emissions and fuel economy testing has been the Federal Test Procedure

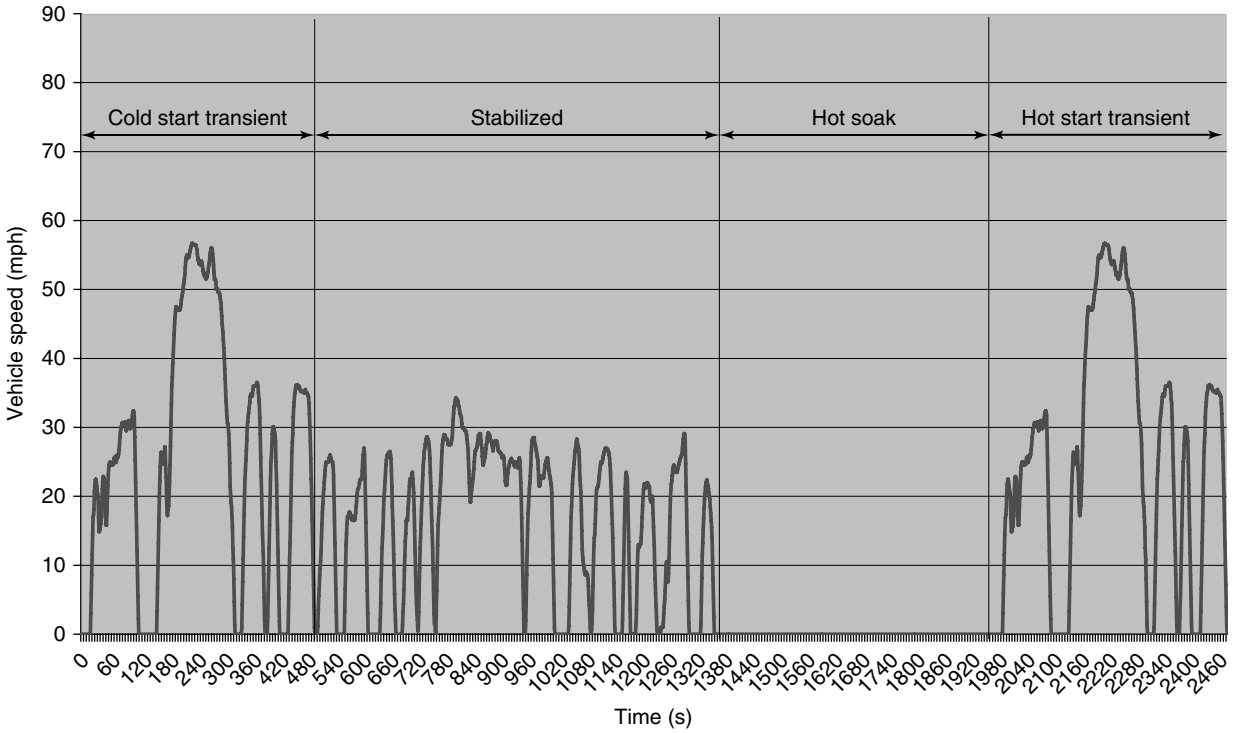


Figure 1. FTP-75 cycle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

FTP-75 cycle, also referred to as *EPA-III*. Shownn its entirety in Figure 1, this cycle was the exclusive cycle required to certify tail-pipe emissions from a passenger car or light truck for the US market until 1994 and still represents the primary emissions and part of the fuel economy regulatory cycle in 2013.

The FTP-75 cycle is intended to present a typical driving trip from cold start, followed by a shutdown and 10-min-long soak period, a hot restart, and additional short driving cycle. In total, the FTP-75 cycle represents a driving distance of 11.04 miles in two sections, a duration of 1874 s, at an average speed of 21.2 miles/h.

The FTP-75 cycle is made up of an initial cold start after a prescribed soak time and city driving cycle that is 505 s long. Without a stop, the cycle continues with a stop-and-go driving portion followed by engine shutdown. After a 10-min hot soak, the driving cycle is repeated with both first and second phases after a hot start. In actual testing, the second stop-and-go cycle is not repeated because the completion of the cycle in the first half is not expected to change, as both phases are conducted after the vehicle has reached operating temperature. Shown in Table 1, weighting factors representing this double counting of the stabilized stop-and-go phase, and accounting for the difference in distance traveled for each phase, have been developed and have been in use from the inception of the cycle.

Table 1. FTP-75 weighting factors.

Portion of Cycle	Name	Distance (miles)	Time (s)	Weighting Factor
Phase 1	Cold start transient	3.6	505	0.43
Phase 2	Stabilized stop-and-go	3.9	864	1.0
Phase 3	Hot start transient	3.6	505	0.57

Emissions regulations are generally written in an allowed emissions constituent in mass, which is measured or sampled over the entire driving cycle, and then divided by the distance traveled. In US regulations, the units are in grams of constituent per mile traveled. The regulated maximum integrated emissions for HC, NO_x, and CO emissions constituents are shown from the original regulation limits through 2001 in Table 2. There was an early version of the FTP-75 cycle called *FTP-72* for some early model years. The FTP-72 cycle only differs from the FTP-75 cycle in the lack of the hot start portion of the cycle. An estimate of emissions from a typical passenger car is made before 1968, and the limits before 1975 were set on the early FTP-72 driving cycle. In later years, more emissions constituents including particulate matter (PM) and formaldehyde have been added to regulation, but the test cycle remains unchanged. Originally, California and the United States required that emissions measured from the tailpipe of the vehicle should be guaranteed by the vehicle manufacturer for 50,000 miles, a value that was termed the useful life of the vehicle. Limits of emissions constituents were promulgated at 50,000 miles useful life and extended to 100,000 miles beginning in 1994. Beginning with this model year, emissions limits were published and required to be met at 50,000 miles intermediate life and 100,000 miles full useful life.

A companion to the FTP-75 cycle is the highway fuel economy (HWFE) cycle, which is representative of highway driving in the US market. This driving cycle is conducted immediately following the FTP-75 cycle without engine shutdown. If not conducted immediately, a warm-up cycle is required. The HWFE cycle, shown completely in Figure 2, represents 10.2 miles of driving with maximum vehicle speed of 57 miles per hour, a typical highway vehicle speed in the 1970s when the cycle was created, and

Table 2. US federal tail-pipe emissions limits, as measured on FTP-75 cycle.

	CO (g/mile)		HC (g/mile)		NO_x (g/mile)		
Useful Life (miles)	50 k	100 k	50 k	100 k	50 k	100 k	
Pre-1968[a]	87	—	8.8	—	3.5	—	—
1970–1971[b]	34	—	4.1	—	—	—	—
1972[b]	28	—	3	—	—	—	—
1973–1974	39	—	3.4	—	3	—	—
1975–1976	15	—	1.5	—	3.1	—	—
1977–1979	15	—	1.5	—	2	—	—
1980	7	—	0.41	—	2	—	—
1981	3.4	—	0.41	—	1	—	—
1987–1993	3.4	—	0.41	—	1	—	—
1994	3.4	4.2	0.25	0.31	0.4	0.6	Cold CO requirement
2001	3.4	4.2	0.075	0.09	0.2	0.3	Phase-in 1993 . . . 1995

[a] Typical value.
[b] Substantially different driving cycle makes direct comparison less clear.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

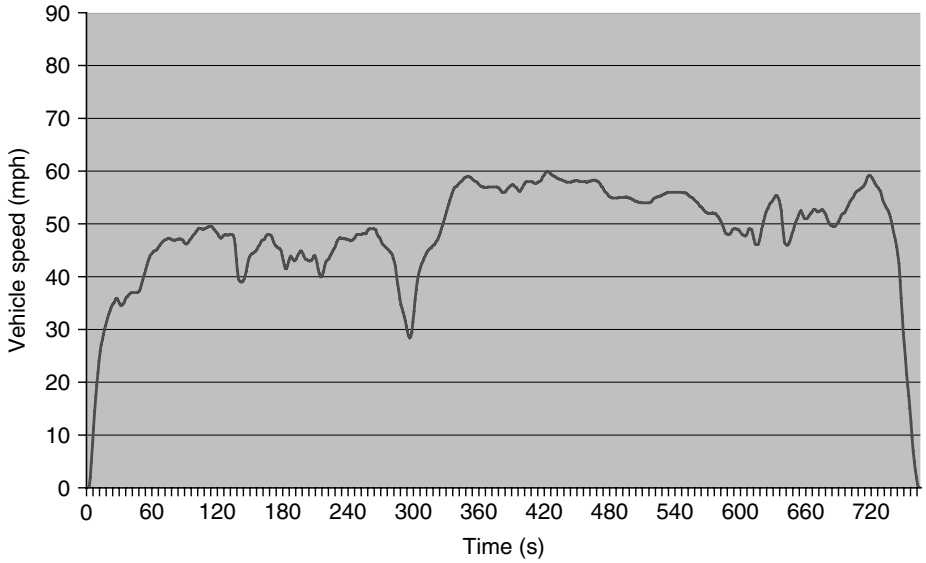


Figure 2. Highway Fuel Economy (HWFE) cycle.

the national speed limit in the United States was 55 miles per hour. In US federal regulations, the HWFE cycle was considered exclusively as part of a fuel economy regulation, not for tail-pipe emissions regulations.

Together, the FTP-75 and HWFE cycles are published for light-duty vehicles as city and highway ratings on the window sticker of new vehicles and combined for a single fuel economy cycle number. The combination of FTP-75 (city) and HWFE (highway) results in the combined fuel economy that is regulated by the US Department of Transportation under Corporate Average Fuel Economy (CAFE) limits. Equation 1 shows the calculation of combined fuel economy with 55% city cycle (FTP-75) and 45% highway cycle (HWFE), which is then applied as a single number to CAFE limits.

$$\text{Combined F.E. (MPG)} = \frac{1}{(0.55/\text{FTP-75 MPG}) + (0.45/\text{HWFE MPG})} \quad (1)$$

Using Equation 1, example fuel economy measured on the FTP-75 (city) and HWFE (highway) cycles results in a combined fuel economy level shown:

FTP city: 27.74 MPG
FTP highway: 40.14 MPG
FTP combined: 32.22 MPG

Originally developed for the 1975 model year, but subsequently delayed in implementation until 1978, the CAFE rules targeted an increase in the average fuel economy of a light-duty passenger car in the United States by 50% by 1985 from the typical level of 1973. These requirements were phased in as a response to the oil shortages of the early 1970s, which caused significant economic disruption in the United States and around the world. The target of 27.5 MPG fleet average for passenger cars was delayed but finally achieved in 1990 calendar year. There were no changes to this requirement for passenger cars and small increases for light trucks between 1990 and 2010, after which new increases were again mandated. Figure 3 shows the CAFE requirement history for passenger cars and light trucks sold in the US market. Fleet average requirements for passenger cars and light trucks are published from 1978 until 2016 model years, with proposed requirements from 2017 until 2025 model years also included. Beginning with 2011, these regulations established a new basis for fuel economy targets based on the vehicle footprint, which is defined as the vehicle wheelbase multiplied by vehicle track width.

The regulated CAFE limits for fuel economy are still based on the original formulation of combined fuel economy, which has led to the new terminology of *unadjusted* fuel economy (raw FTP-75 and HWFE values). Vehicle manufacturers are required to publish *adjusted* fuel economy values which are calculated using a different formula. Phased in beginning with the 2008 model year, the new five-cycle fuel economy values are calculated using Equations 2 and 3.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

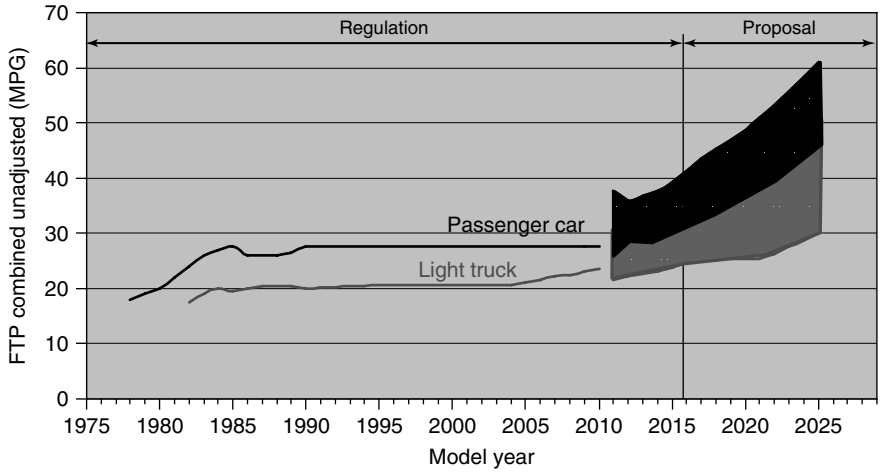


Figure 3. US Corporate Average Fuel Economy (CAFE) requirements.

$$\begin{aligned}
\text{City FE} &= 0.89\frac{1}{\text{Start FC} + \text{Running FC}} \\
\text{Start FC} &= 0.33\frac{(0.76 \times \text{Start Fuel}_{75} + 0.24 \times \text{Start Fuel}_{20})}{3.5} \\
\text{Running FC} &= 0.70\left(\frac{0.48}{\text{Bag2}_{75}\text{FE}} + \frac{0.41}{\text{Bag3}_{75}\text{FE}} + \frac{0.11}{\text{US06City FE}}\right) + 0.30\left(\frac{0.5}{\text{Bag2}_{20}\text{FE}} + \frac{0.5}{\text{Bag2}_{20}\text{FE}}\right) \\
&\quad + 0.133 \times 1.083 \times (\text{A/CFC}) \\
\text{MPG Based City FE Label Value} &= \frac{1}{\left(0.002549 + \frac{1.2259}{\text{FTP FE}}\right)}
\end{aligned} \tag{2}$$

$$\begin{aligned}
\text{Highway FE} &= 0.89\frac{1}{\text{Start FC} + \text{Running FC}} \\
\text{Start FC (gallons per mile)} &= 0.330\left[\frac{(0.76 \times \text{StartFuel}_{75} + 0.24 \times \text{StartFuel}_{20})}{60}\right] \\
\text{Running FC} &= (1.012)\left(\frac{0.79}{\text{US06 Highway FE}} + \frac{0.21}{\text{HFET FE}}\right) + 0.133 \times 0.377 \times \text{A/C FC} \\
\text{MPG Based Highway FE Label Value} &= \frac{1}{\left(0.000308 + \frac{1.4030}{\text{HFET FE}}\right)}
\end{aligned} \tag{3}$$

The new *adjusted* values are required for publication on the window sticker of new vehicles sold in the United States, and only these values can be used in advertising. An example window sticker is shown in Figure 4.

Beginning with a phase-in starting in the 2004 model year, the United States Environmental Protection Agency (US EPA) federal requirements were substantially updated to a new regulatory regime called *Tier 2*. First, reduced tailpipe emissions regulatory limits were imposed, still using the FTP-75 cycle, and the full useful life of a vehicle was extended to 120,000 miles. The new limits are published as bins allowing the manufacturer the discretion to develop vehicles meeting different emissions standards in the same model year. In addition, a fleet average NO_x standard was

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

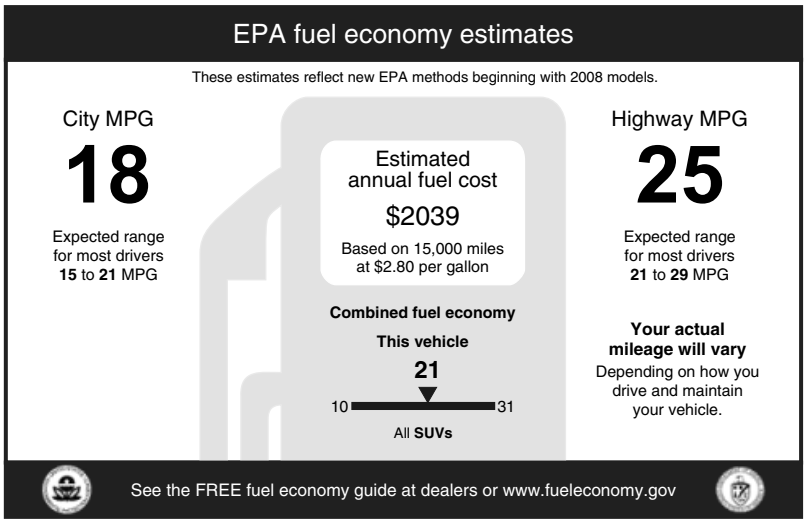


Figure 4. Required US fuel economy window sticker from the 2008 model year.

adopted that all vehicles in the manufacturer's sales for that model year must achieve, and there are caps to the NO_x emissions limit that effectively cancel the least stringent bins 9–11 by the 2007 model year. The more stringent bins no longer have 50,000 miles intermediate requirements. The individual bins and their limits are documented in Table 3.

As part of the Tier 2 emissions program, EPA enacted fleet average HC and NO_x emissions limits to be phased in over several years. By the 2007 model year for passenger cars and the 2009 model year for all light trucks, NO_x emissions must average less than 0.07 g/miles (bin 5). In addition, by the 2013 model year, fleet average HC emissions must average less than 0.3 g/miles; the test for fleet average HC emissions is the FTP-75 test conducted at ambient temperature of 20°F (−7°C).

Because the combination of FTP-75 emissions cycle and HWFE cycle has been the legal requirement for over 35 years for both emissions and fuel economy regulations in the United States, and because the EPA and California have been collecting data from all vehicles certified for their respective markets over that time period, the history of fuel economy and emissions measurements is both telling and valuable. US fleet average fuel economy for passenger cars and light trucks 1978 through 2011 is shown in Figure 5.

Because in early years of the regulatory process, there was no fleet average requirement and therefore no collection of the documented average emissions from a vehicle during this time period, a collected assessment of the average emissions from vehicles is not readily available. Therefore, as an illustration, the regulatory limits in the three major emissions constituents in grams/miles on the FTP-75 cycle from unregulated to the most recent Tier 2 (bin 5) level are plotted in Figure 6. The regulatory limits represent a

Table 3. US Tier 2 passenger car FTP-75 emissions limits.

Useful Life (miles)	CO (g/mile)		NMOG (g/mile)		NO_x (g/mile)		Formaldehyde (g/mile)		PM (g/mile)	
	50 k	120 k	50 k	120 k	50 k	120 k	50 k	120 k	50 k	120 k
Bin 11	3.4	4.2	0.195	0.28	0.6	0.9	0.022	0.032	N/A	0.12
Bin 10	3.4	4.2	0.125	0.156	0.4	0.6	0.015	0.018	N/A	0.08
Bin 9	3.4	4.2	0.075	0.09	0.2	0.3	0.015	0.018	N/A	0.06
Bin 8	3.4	4.2	0.1	0.125	0.14	0.2	0.015	0.018	N/A	0.02
Bin 7	3.4	4.2	0.075	0.09	0.11	0.15	0.015	0.018	N/A	0.02
Bin 6	3.4	4.2	0.075	0.09	0.08	0.1	0.015	0.018	N/A	0.01
Bin 5	3.4	4.2	0.075	0.09	0.05	0.07	0.015	0.018	N/A	0.01
Bin 4	N/A	2.1	N/A	0.07	N/A	0.04	N/A	0.011	N/A	0.01
Bin 3	N/A	2.1	N/A	0.055	N/A	0.03	N/A	0.011	N/A	0.01
Bin 2	N/A	2.1	N/A	0.01	N/A	0.02	N/A	0.004	N/A	0.01
Bin 1	N/A	0	N/A	0	N/A	0	N/A	0	N/A	0

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

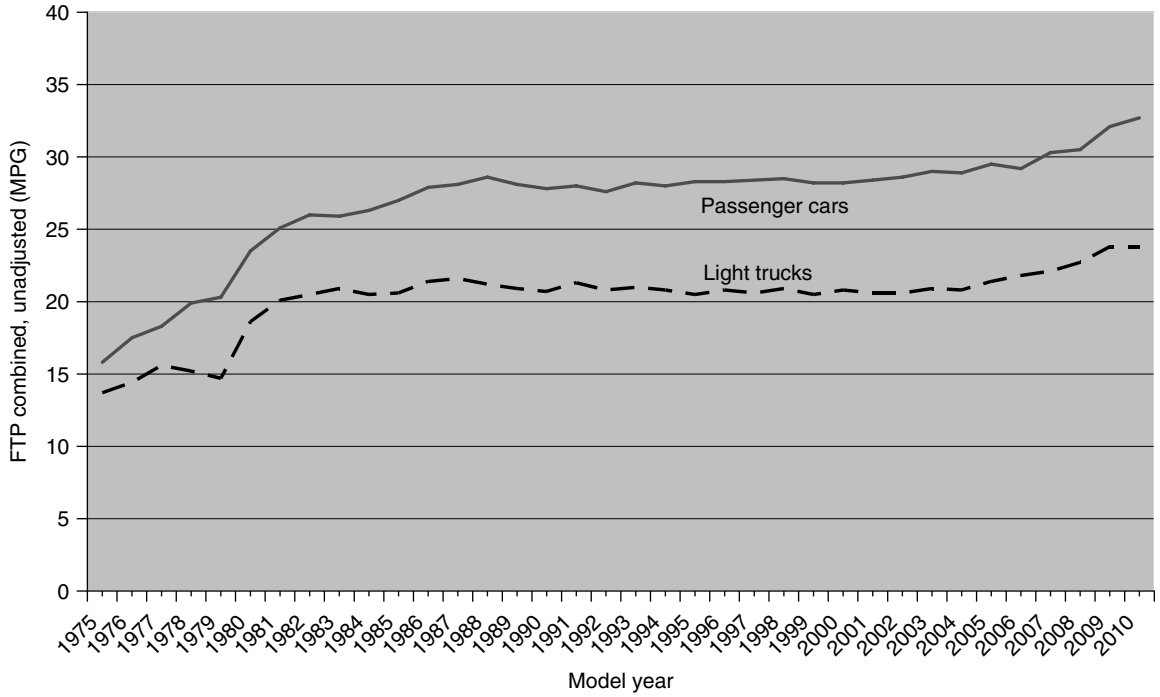


Figure 5. US light-duty vehicle fleet fuel economy trend.

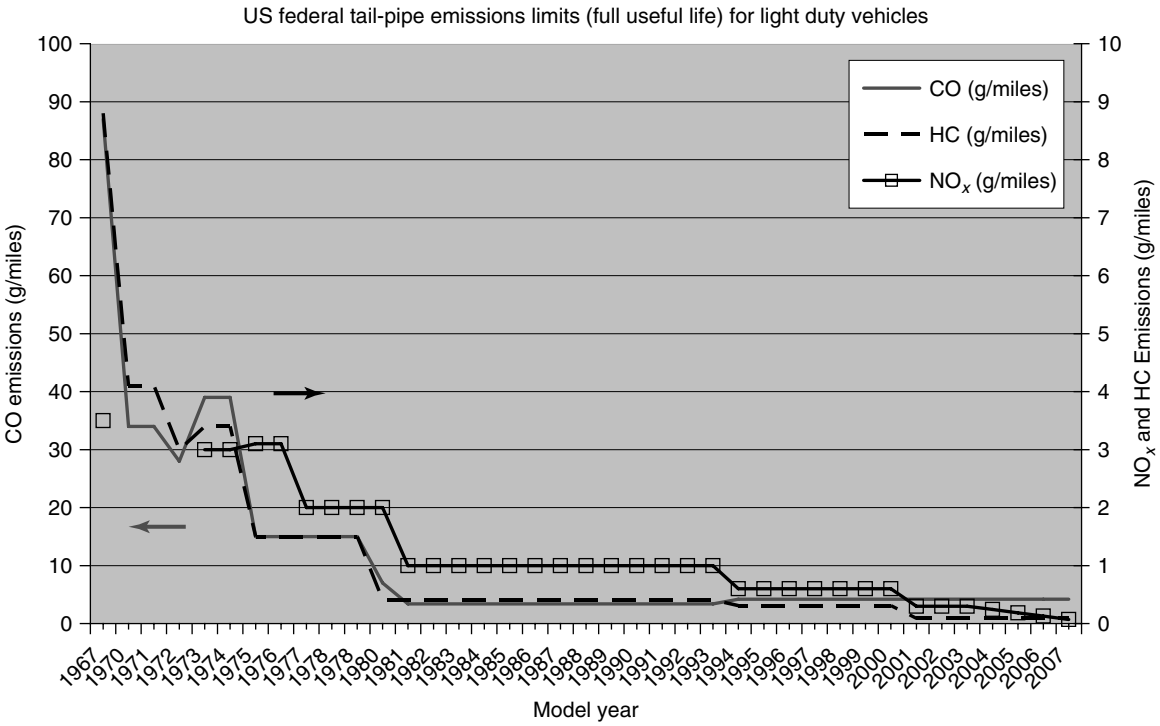


Figure 6. US regulatory limits for major emissions constituents.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

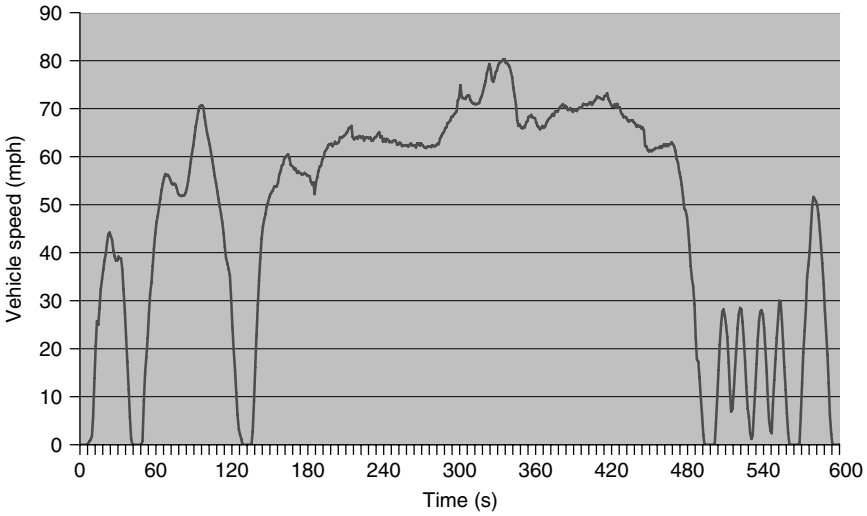


Figure 7. US06 cycle.

reduction in tail-pipe emissions from a typical unregulated light-duty vehicle of 95% (CO), 99% (HC), and 98% (NO_x).

With phase-in beginning with the 2001 model year, two additional driving cycles were added to the certification regime for the US market. These cycles, termed *Supplemental Federal Test Procedure* (*SFTP*), have the goal of expanding the operating region of driving to be more representative of faster and more aggressive driving conditions and operation with the air conditioning in the vehicle operating.

The first cycle was designed to cover the operating regime typical of more aggressive driving styles. Because the original FTP-75 cycle is a relatively lightly load cycle with moderate accelerations, the new US06 cycle was designed to monitor operation in areas that had previously been outside the regulated cycles. The US06 cycle, shown in Figure 7, represents 10 min of driving with vehicle speed up to 80 MPH.

A second supplemental cycle called *SC03* was designed to test vehicle operating in hot ambient environmental conditions with the passenger compartment air conditioning system operating at maximum. The air conditioning system can present a significant load on the engine and therefore significantly change the emissions and fuel economy performance levels. The new SC03 cycle, shown in Figure 8, also represents 10 min of driving, at ambient

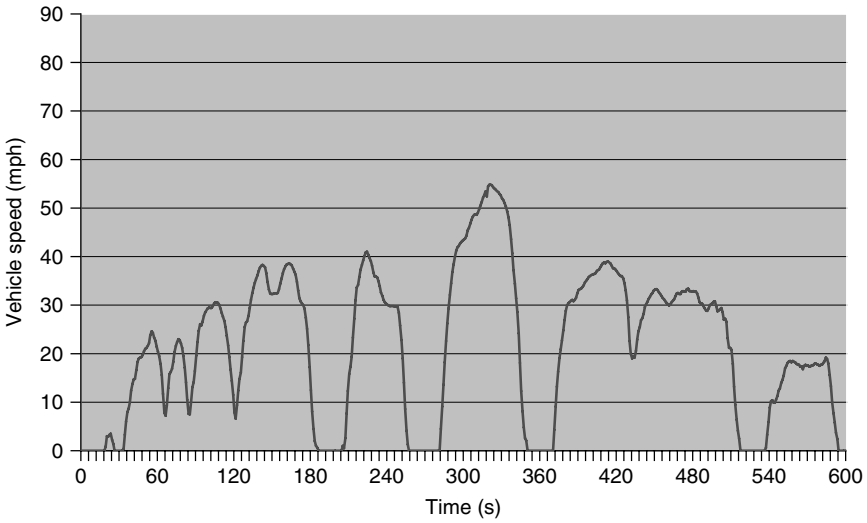


Figure 8. SC03 cycle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. US SFTP passenger car limits.

US06		SC03	
NMHC+NO_x (g/mile)	CO (g/mile)	NMHC+NO_x (g/mile)	CO (g/mile)
0.14	8	0.2	2.7

temperature of 95°F, with simulated solar load applied to the vehicle; the air conditioning system is set to maximum.

US federal limits for SFTP cycle performance for the Tier 2 4000 miles condition are documented in Table 4. Regulatory limits for California and later US federal requirements have been adjusted.

The US federal and California have created a complex set of regulations depending on the vehicle size and use. For US regulatory purposes generally, passenger cars are used primarily for the purpose of transporting people, whereas light- and medium-duty trucks are used primarily for transporting material or for use off paved roads. Emissions limits for passenger car legislated values are quoted throughout this chapter. However, there are multiple additional categories of vehicles including light-duty trucks and medium-duty passenger vehicles (i.e., buses) that have to meet somewhat less stringent requirements on an emissions mass per mile traveled basis. In the United States, there is a somewhat arbitrary line between light-duty vehicle (light truck and passenger car) and medium/heavy-duty truck with gross vehicle weight (GVW) of greater than 8500 pounds. The definition of GVW is the total weight of the vehicle including maximum payload stated by the manufacturer.

2.1.1 California

Because California faced severe local issues related to smog and ambient pollution related to geography with increased population and passenger car use, it was the first jurisdiction in the world to enact specific emissions control requirements. In 1961, the first regulations for positive crankcase ventilation (PCV) were required for new light-duty vehicles sold in California. Limits for HC and CO tail-pipe emissions were enacted in 1966. The first standard for NO_x from passenger cars was implemented by California in 1971. Although not entirely related to the regulation of automobile emissions, there is documented improvement in ambient air quality over the time period during which passenger car tail-pipe emissions were reduced, even though the population and number of light-duty vehicles increased.

California is the only state in the United States allowed to develop and promulgate emissions standards that are unique from the federal standards, although other states are allowed to adopt California standards.

Table 5 documents the early California emissions limits, although the 1960s limits conducted on a different driving cycle without comparable emissions sampling standards are excluded; before 1975, an early variant called the *FTP-72 cycle* was applied. The California definition of full useful life of a vehicle was extended to 100,000 miles for the 1980 model year, 14 years in advance of US federal regulations.

With phase-in beginning in the 1994 model year, California introduced a new emissions regime called the *Low Emissions Vehicle* (*LEV-1*) program. New categories of vehicles were created to provide automotive manufacturers the flexibility to certify vehicles in different categories.

Table 5. Early California passenger car emissions limits.

	CO (g/mile)		HC (g/mile)		NO_x (g/mile)	
Useful Life (miles)	50 k	100 k	50 k	100 k	50 k	100 k
1973	39	—	3.2	3		—
1974	39	—	3.2	—	2	—
1975	9	—	0.9	—	2	—
1976	9	—	0.9	—	2	—
1977	9	—	0.41	—	1.5	—
1978	9	—	0.41	—	1.5	—
1979	9	—	0.41	—	1.5	—
1980	9	9	0.39	0.39	1	1.5
1981	3.4	7	0.39	0.39	1	1.5
1982	7	7	0.39	0.39	0.4	1.5
1983	7	7	0.39	0.39	0.4	1.5
1984\td\87	7	7	0.39	0.39	0.4	1
1988	7	7	0.39	0.39	0.4	1
1989\td\94	7	7	0.39	0.39	0.4	1
1995	3.4	4.2	0.25	0.31	0.4	0.6

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

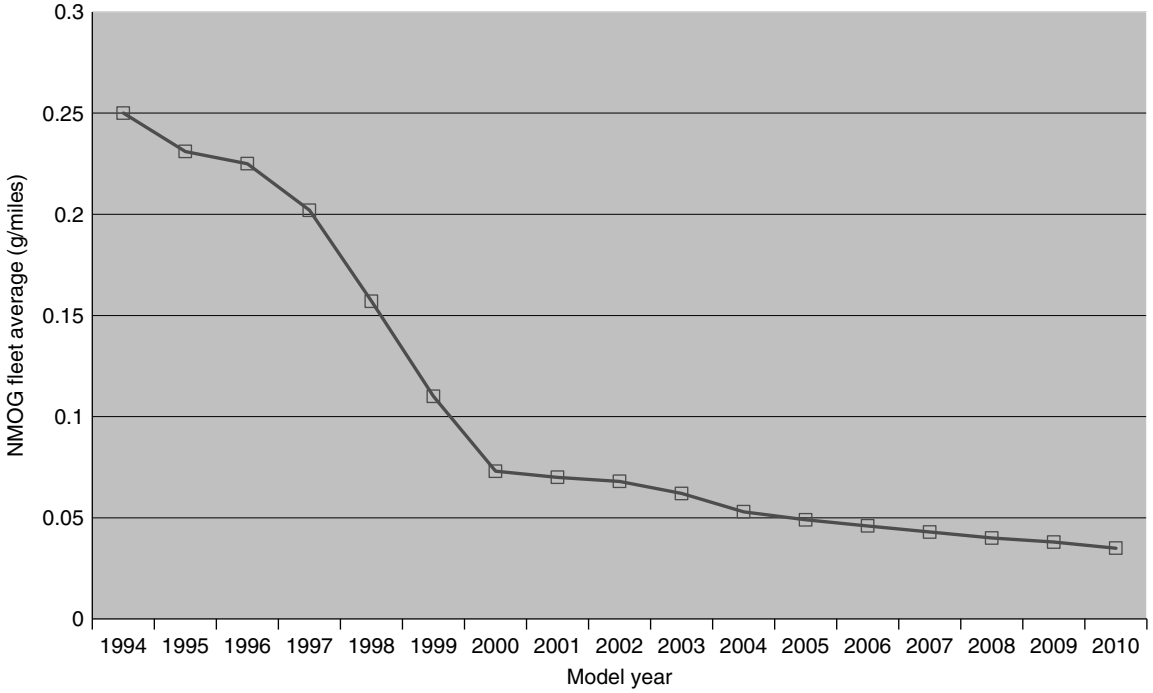


Figure 9. California fleet average hydrocarbon (NMOG) emissions limit.

A fleet-average HC emission rule was established, with targets for each calendar year progressively lower. This fleet-average HC emissions, or more correctly NMOGs, requirement is shown beginning with the 1994 model year in Figure 9. In addition, the new federally developed SFTP cycles were added to the regulation; supplemental emissions limits to be met on each cycle for each constituent were also mandated. New categories of vehicles called *transitional low emission vehicle* (*TLEV*), *LEV*, and *ultra low emission vehicle* (*ULEV*) were created. The full useful life requirement for vehicles was increased to 120,000 miles. The California LEV-1 limits for passenger cars are summarized in Table 6.

With phase-in beginning in the 2004 model year, California introduced a new emissions regime called *LEV-2* program. Another new category of vehicle was created to another even lower category level, termed the *super ultra low emissions vehicle* (*SULEV*). Fleet-average HC emissions levels were extended until the 2010 model year. The California fleet-average NMOG emissions level for both LEV-1 and LEV-2 regulatory periods from 1994 to 2010 is shown in Figure 8. Table 7 summarizes the passenger car LEV-2 emissions requirements.

With phase-in anticipated in the 2014 model year, California has *proposed* and implemented an extension of the LEV-1 and LEV-2 to the revised LEV-3 program. New categories of vehicles were added to the existing levels. Summed fleet-average HC and NO_x emissions levels were extended until the 2022 model year. Table 8 summarizes the passenger car LEV-3 emissions requirements.

2.1.2 *Inspection and maintenance*

To comply with federal clean air requirements, some states and municipalities require regular emissions measurements to confirm adequate emissions are maintained in use by a

Table 6. California passenger car LEV-1 emissions limits.

	CO (g/mile)		HC (g/mile)		NO_x (g/mile)	
Useful Life (miles)	50 k	100 k	50 k	100 k	50 k	100 k
TLEV	3.4	4.2	0.125	0.156	0.4	0.6
LEV	3.4	4.2	0.075	0.09	0.2	0.3
ULEV	1.7	2.1	0.04	0.055	0.2	0.3
SULEV	1.7	2.1	—	0.072	0.2	0.3

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 7. California passenger car LEV-2 emissions limits.

	CO (g/mile)		NMOG (g/mile)		NO_x (g/mile)		PM (g/mile)		Formaldehyde (mg/mile)	
Useful Life (miles)	50 k	120 k	50 k	120 k	50 k	120 k	50 k	120 k	50 k	120 k
LEV_2	3.4	4.2	0.075	0.09	0.05	0.1	N/A	0.01	15	18
$ULEV_2$	1.7	2.1	0.04	0.055	0.05	0.07	N/A	0.01	8	11
$SULEV_2$	N/A	1	N/A	0.01	N/A	0.02	N/A	0.01	N/A	4

Table 8. California passenger car LEV-3 emissions limits.

Useful Life (miles)	CO (g/mile) 150 k	$NMOG + NO_x$ (g/mile) 150 k	PM[a] (g/mile) 150 k	HCHO (Formaldehyde) (mg/mile) 150 k
LEV_3	4.2	0.160	TBD	0.018
$ULEV_3$	2.1[b]	0.125	TBD	0.011[b]
ULEV70	2.1[b]	0.070	TBD	0.011[b]
ULEV50	2.1[b]	0.050	TBD	0.011[b]
$SULEV_3$	1.0[c]	0.030	TBD	0.004[c]
SULEV20	1.0[c]	0.020	TBD	0.004[c]

[a] Proposed 0.01 g/mile LEV II PM standard reduced to 0.003 g/mile for gasoline engines.
[b] Assumed ULEV III, ULEV70, and ULEV50 have same CO and HCHO standards as ULEV II.
[c] Assumed SULEV III and SULEV20 have same CO and HCHO standards as LEV II.

vehicle owner. For this monitoring program, an inspection cycle called *IM-240* was developed. The IM-240 cycle is similar to the first 240 s of the FTP-75 cycle. Figure 10 shows the IM-240 cycle.

Tail-pipe emissions limits from vehicles tested on the IM-240 cycle are established independently by each jurisdiction depending on the ambient air quality of the region and plan for achieving national air quality standards and depending on the model year of the vehicle being tested. Such a test is performed as a part of vehicle registration typically once every 2 years. In 2011, IM-240 testing is performed in only a few cities whose ambient air quality does not meet standards established by the EPA. In 2011, US states, cities, and municipalities that require tail-pipe emissions inspection and maintenance sampling are listed in Table 9.

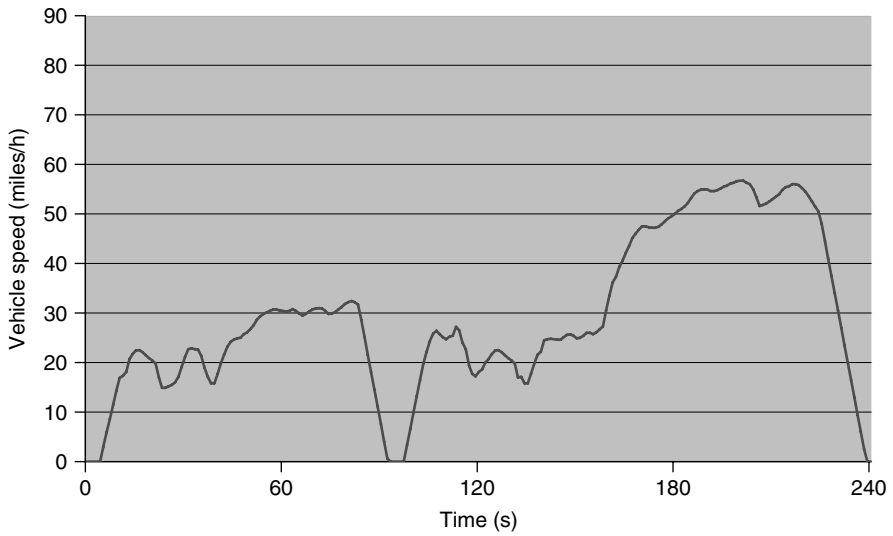


Figure 10. IM-240 cycle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 9. US municipalities with inspection maintenance requirements.

State	Emissions Inspection Requirements (at License Application)
Alaska	Anchorage municipality only; CO emissions only
Arizona	Phoenix and Tucson areas only
California	All vehicles from outside state, and all vehicles in Los Angeles, San Francisco, San Diego, and Sacramento metropolitan areas
Colorado	Denver and Boulder municipal areas and Northern range areas only
Connecticut	All vehicles
District of Columbia	All vehicles
Delaware	All vehicles
Georgia	Atlanta municipal area only
Idaho	Boise (Ada County) municipal area only
Illinois	Chicago and St. Louis municipal areas only
Indiana	Lake and Porter Counties only
Louisiana	Baton Rouge municipal area only; visual and OBD inspection only
Maine	Portland municipal area (Cumberland County) only
Maryland	All vehicles; idle test and OBD inspection only
Massachusetts	All vehicles
Missouri	St. Louis municipal area only
Nevada	Clark County (Las Vegas) and Washoe County (Reno) only; idle test and OBD inspection only
New Hampshire	All vehicles; OBD inspection only
New Jersey	All vehicles
New Mexico	Bernalillo County (Albuquerque municipal area) only
New York	All vehicles; OBD, gas cap, and visual inspection only
North Carolina	48 (of 100) counties; OBD inspection only
Ohio	Cleveland/Akron municipal area only
Oregon	Portland and Medford-Ashland municipal areas only
Pennsylvania	Pittsburgh, Philadelphia, and select counties only
Rhode Island	All vehicles
Tennessee	Memphis and Nashville municipal areas only
Texas	Dallas-Ft. Worth, Houston, Austin, and El Paso municipal areas only
Utah	Salt Lake City—Ogden municipal area only
Vermont	All vehicles; OBD test only
Virginia	District of Columbia municipal area
Washington	Seattle municipal area (five counties) only
Wisconsin	Milwaukee and suburban Chicago municipal areas only

The use of this type of program is typically applied to urban areas whose ambient air quality exceeds the national and state requirements set by the EPA and CARB. Such a program can support the implementation plan for a local community to improve ambient air quality in the impacted "nonattainment" areas.

Canadian provinces of British Columbia and southern Ontario also require vehicle emissions testing at regular intervals.

2.2 Europe

The European Union introduced vehicle emissions limits beginning with the 1992 calendar year for all member states. Before this year, some countries implemented early regulations beginning in the late 1960s, but implementation timing varied by country. For the original European-wide regulation, the MVEG-A driving cycle was defined as were the EURO-1 category of emissions limits. As in the United States and California, the emissions limits were defined as a mass of emissions produced over a distance traveled on the complete driving cycle. In the original regulation, measurement of gaseous emissions started 40 s after the engine start allowing a brief unregulated warm-up period. Beginning with EURO-3 limits for the 2002 model year, the cycle was modified to become the current MVEG-B cycle officially called the *New European Driving Cycle* (*NEDC*). Shown completely in Figure 11, the NEDC is made up two parts. The first part is made up of four replicates of a short urban driving cycle with vehicle speeds up to 50 km/h. The second part is a representation of highway driving made up of acceleration and higher vehicle speed up to 120 km/h. The complete NEDC is made up of both sections with total distance traveled of just over 11 km in 1180 s. For reporting fuel economy, the cycle is typically separated into urban (city) and extra-urban (highway) portions, but the emissions test and limits are conducted on the entire cycle.

From 1993 through today, the emissions limits for the European Union are shown in Table 10. The European regulations have recognized a difference between gasoline and diesel engines, and the limits are different for the engine type for all years up to EURO-6 level in the 2015 model year for new vehicles.

European regulators have required measurement and reporting of fuel consumed on the standard emissions cycle as the beginning of its inception. However, European-wide requirement to meet fuel economy standards, CO_2 emissions limits, was phased in beginning in 2012. Individual member states have implemented different forms of tax

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

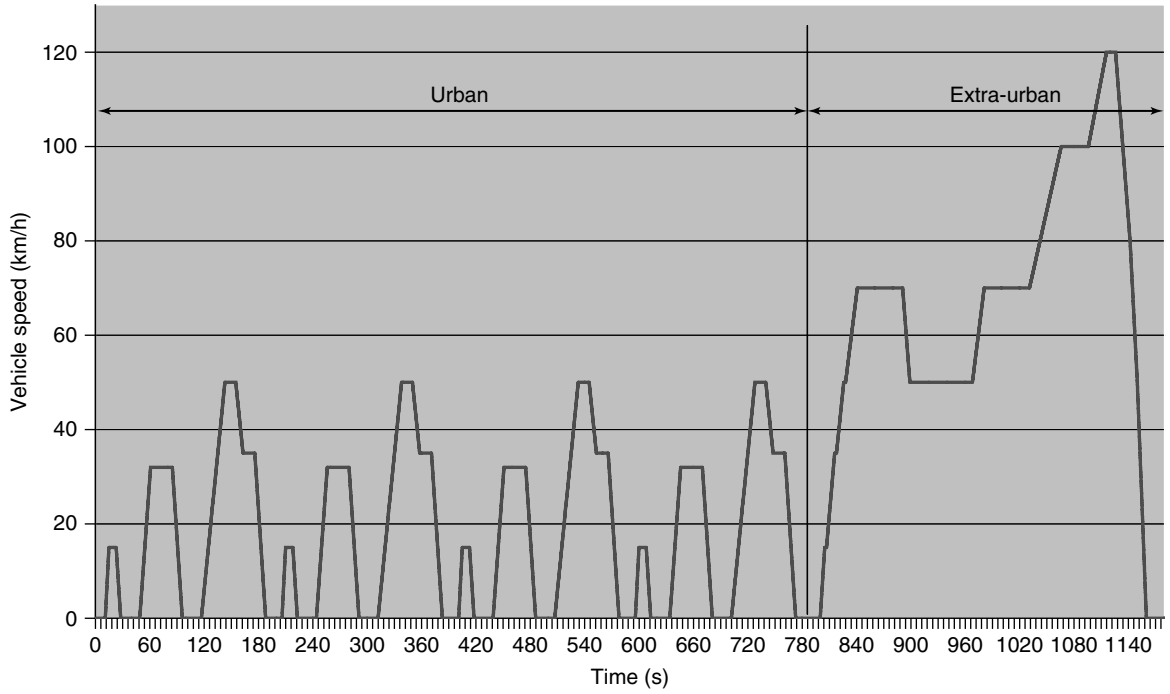


Figure 11. New European Drive Cycle (NEDC).

Table 10. European passenger car emissions limits.

		Gasoline			Diesel		
	Useful Life (km)	$HC+NO_x$ (g/km)	CO (g/km)	PM (g/km)	$HC+NO_x$ (g/km)	CO (g/km)	PM (g/km)
EURO-1 1993	80 k	0.95	2.72	0.14	0.95	2.72	0.14
EURO-2 1996	80 k	0.5	2.2	N/A	0.7	1	0.08
EURO-3 2000	80 k	0.35	2.3	N/A	0.56	0.64	0.05
EURO-4 2005	100 k	0.18	1	N/A	0.3	0.5	0.025
EURO-5 2010	160 k	0.16	1	0.005	0.23	0.5	0.005
EURO-6 2015	160 k	0.16	1	0.0045	0.17	0.5	0.0045

incentives or penalties. Fuel consumption and CO_2 emissions are measured using the same equipment and driving cycle.

There are in addition different limits for vehicles whose primary use is not for personal transportation. Light-duty commercial vehicles are regulated according to the same driving cycle and testing protocol, but the individual constituent limits are somewhat higher depending on the inertia class and test weight of the vehicle.

A new set of three driving cycles has been proposed by European regulators potentially for future emissions standards. The new Common ARTEMIS Driving Cycles (CADC) are developed under the European ARTEMIS project (ARTEMIS, Assessment and Reliability of Transport Emissions Models and Inventory Systems). The three cycles are defined for urban, rural road, and motorway, with maximum speed variants in the motorway segment of 130 or 150 km/h. Currently, the ARTEMIS Cycle, which is shown completely in Figure 12, is not used for emissions regulatory purposes.

2.2.1 Inspection and maintenance

The European Union requires vehicle inspections to be completed on light-duty vehicles, but emissions' testing is not included in the European wide inspection requirement. Some countries require emissions testing regularly as part of their vehicle inspection requirements. Germany, Spain, France, the Netherlands, and the United Kingdom have some form of emissions inspection on vehicles.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

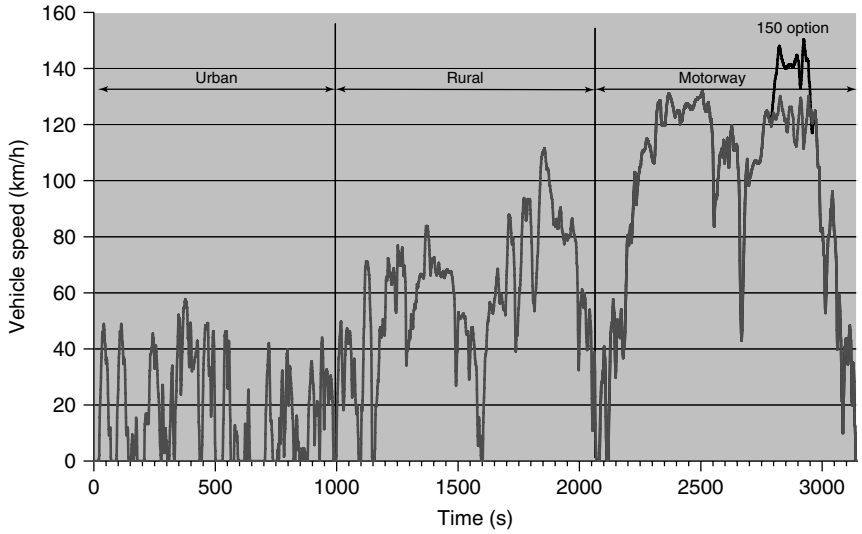


Figure 12. ARTEMIS Drive Cycle.

2.3 Japan

Beginning with the 1975 model year, Japan initiated a vehicle emissions program with a unique drive cycle more representative of driving in the Japanese market. Like the US and European regulations, the main emissions limits were established as a total mass of emissions constituent divided by the total distance traveled. The first Japanese driving cycle that regulated emissions from a warm vehicle was called the *10–15 mode cycle* is shown in Figure 13. A companion cycle called the *11 mode cycle* was also established for regulation of cold start emissions; on the 11 mode cycle, the regulated limits were in units of grams of pollutant per test, unlike most other cycles whose regulatory limits are stated on a distance traveled basis.

With phase-in beginning for the 2008 model year, the Japanese government implemented a new driving cycle to replace both existing cycles with a new test more representative of actual vehicle driving. Like the FTP-75 cycle, the new JC08 cycle shown in Figure 14 is representative of actual vehicle driving conditions on a public road.

Japanese emissions standards for passenger cars were first enforced for the 1975 model year for HC and CO

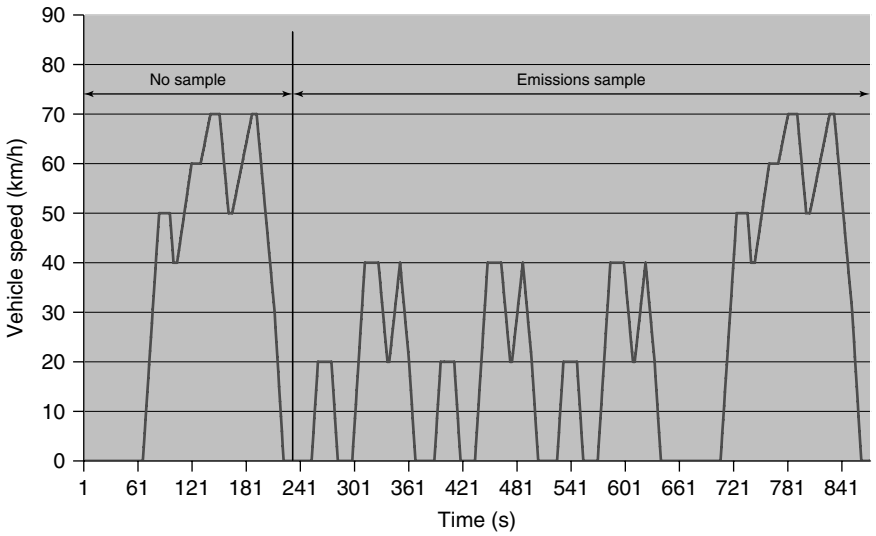


Figure 13. Japan 10–15 mode cycle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

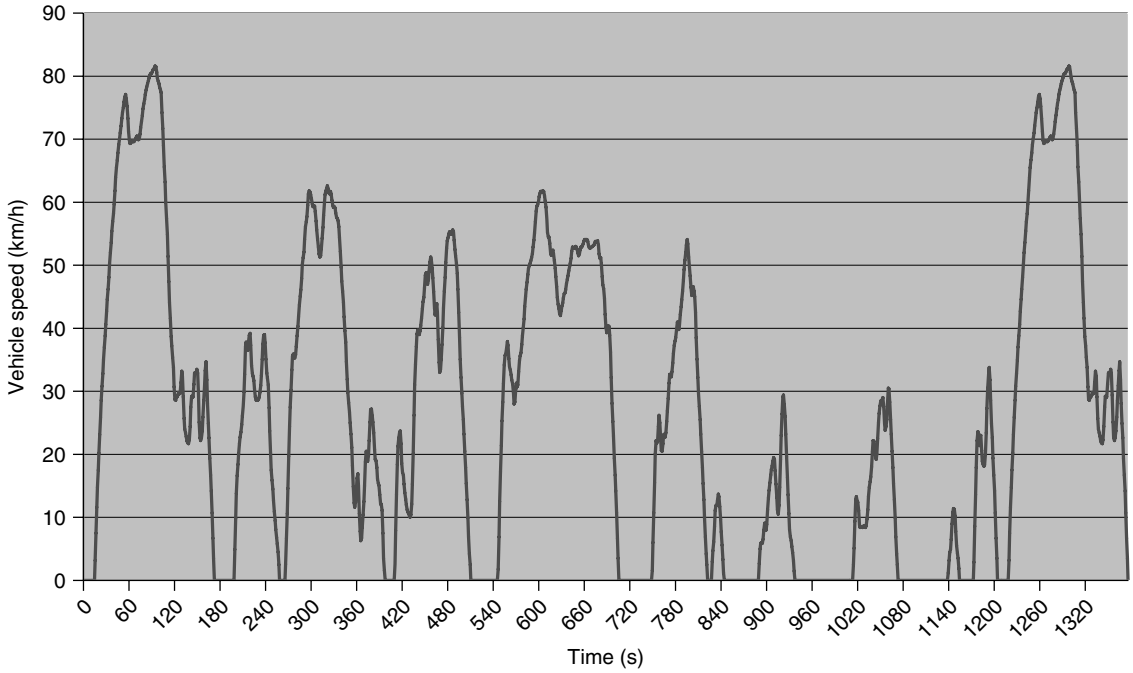


Figure 14. Japan JC08 driving cycle.

emissions, and 1978 for NO_x emissions. The standards were reduced for the 2001 model year. Emissions requirements with these early cycles are summarized in Table 11.

Together with the implementation of the new JC08 emissions cycle, much more stringent emissions limits (Table 12) were also phased in.

2.3.1 Inspection and maintenance

The Japanese government requires vehicle inspections to be completed on light-duty vehicles after 3 years of service and every 2 years thereafter, which includes emissions test requirements. The periodic inspection known as *Shaken* is

Table 11. Japan emissions limits from 1975 until 2005.

	CO		HC		NO_x		
Model Year	10–15 Mode (g/km)	11 Mode (g/test)	10–15 Mode (g/km)	11 Mode (g/test)	10–15 Mode (g/km)	11 Mode (g/test)	
1975	2.1/2.7	60/85	0.39/0.25	7.0/9.5	N/A	N/A	[a]
1978	2.1/2.7	60/85	0.39/0.25	7.0/9.5	0.25/0.48	4.4/6.0	[a]
2001	0.67/1.27	19/31.1	0.08/0.17	2.2/4.42	0.08/0.17	1.4/2.5	[a]

[a] The first value indicates the maximum permissible emissions; the second value indicates the average value for the vehicle type.

Table 12. Japan emissions limits from 2006.

	CO	HC	NO_x	PM	
Model Year	Combined Mode Test (g/km)	Combined Mode Test (g/km)	Combined Mode Test (g/km)	Combined Mode Test (g/km)	
2006	0.63/0.98	0.12/0.24	0.28/0.43	0.052/0.11	[a]
2009	0.63/0.84	0.024/0.032	0.14/0.19	0.013/0.017	[b]
2011	0.63/0.84	0.024/0.032	0.08/0.11	0.005/0.007	[c]

[a] 88% 10–15 mode hot start and 12% 10–15 mode cold start.
[b] 75% 10–15 mode hot start and 25% JC08 cold start.
[c] 75% JC08 hot start and 25% JC08 cold start.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

quite expensive. The emissions test is conducted at idle and only HC and CO emissions are measured.

2.4 Other

Most other countries with light-duty vehicle emissions standards base those requirements on a European, US, or Japanese standards, or even a combination. India has implemented a variation on the European emissions cycle called the *Modified India Driving Cycle*, which is essentially identical to the NEDC with the exception of maximum vehicle speed on extra-urban portion reduced from 120 to 90 km/h. Mexico currently (2011) allows either US standards from 1994 or EURO-4 levels from 2005.

3 VEHICLE TESTING PROCESS AND MEASUREMENTS

3.1 Representation of vehicle load

To represent correctly the powertrain operating regime, the chassis dynamometer should apply a resistance to motion that is consistent with the load required to move the vehicle forward on the road. There are industry standard processes for calculating the road load and applying it to the chassis dynamometer. Two categories of load are required to correctly represent this resistance. First, based on vehicle mass, a mass inertia is applied to resist vehicle acceleration. Second, as a function of the rolling resistance to motion, a quadratic resistance function that represents the increase in load as a function of vehicle speed is applied. Expressed as a quadratic equation, the coefficients are generally correlated with resistance to motion, rolling friction, and aerodynamic resistance.

To develop a set of dynamometer values that represent the actual vehicle on a road, a standard coast down test of a vehicle on a test track is performed. The vehicle test is conducted on a flat straight track in which the vehicle is coasted from high speed to low speed. The rate of deceleration is measured; effectively, the vehicle speed is measured over time and deceleration rate is calculated as a change in velocity over a time interval. Once the vehicle rate of deceleration without motive input from the powertrain is measured, the resistive force equation is calculated. When a vehicle is tested in a closed facility, the dynamometer resistance is calibrated to the actual road test, so that when a vehicle is coasted from high to low speed, the deceleration rate is comparable.

3.2 Deterioration factors

Because the exhaust emissions of a vehicle tend to increase with usage over the life of a vehicle, European and US regulatory regimes require a prediction of the deterioration is used to guarantee that the vehicle will produce emissions under the required limits over their entire useful life. The deterioration factor can be directly measured from a vehicle as it accumulates mileage. The US EPA has, for example, developed a standard cycle to exercise a vehicle over the full useful life. During the accumulation of mileage, emissions are measured on the regulated cycle. Linear regression is applied to calculate the exact deterioration of each emissions constituent from the new condition to the end of the required useful life. The US EPA Standard Road Cycle (US EPA SRC) used for demonstration of emissions durability is shown in Figure 15.

To conduct such a test to develop the emissions capability and deterioration for a life of a vehicle, this test cycle should be repeated until the limit mileage on the vehicle is obtained, typically 100,000–150,000 miles.

Because the catalytic converter in the exhaust is effectively responsible for the vast majority of the emissions control capability of a modern vehicle, the US EPA also allows a bench procedure for emissions deterioration. The catalyst assembly is aged using a bench on a cycle that is calculated to represent the deterioration experienced over the useful life of the vehicle. Aged exhaust components are prepared using bench methods. The test vehicle is tested with new and aged components for determination of the deterioration factor. According to US EPA regulations, the catalyst temperature profile and number of durability hours the exhaust system is exposed to is a function of the instrumented measurement of temperature the catalyst is exposed to over the Standard Road Cycle (SRC).

The US EPA allows manufacturers the flexibility to modify the process for determining the deterioration of emissions and durability depending on proprietary development processes and test cycles that are not part of the EPA standard process. To use a nonstandard cycle, the manufacturer must demonstrate the process and typically provide documentation of historical performance of vehicles in customer use.

In the European regulatory environment, standard assigned deterioration factors are permitted. The assigned deterioration factors are written into the regulation and available for all manufacturers. Assigned factors for each emissions constituent for the current series of regulations from EURO-1 to EURO-6 are tabulated in Table 13.

In both European and US markets, the manufacturer must warrant that vehicles in customer use must, when maintained according to the manufacturer's recommendations,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

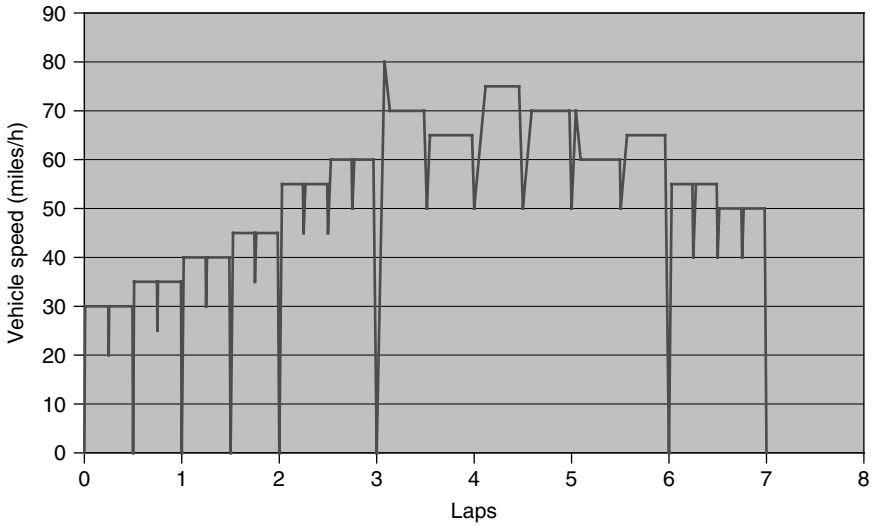


Figure 15. EPA Standard Road Cycle (SRC) for durability demonstration.

Table 13. European assigned deterioration factors.

		Deterioration Factors					
		Gasoline			Diesel		
	Useful Life (km)	HC+NO_x (g/km)	CO (g/km)	PM (g/km)	HC+NO_x (g/km)	CO (g/km)	PM (g/km)
EURO-1 1993	80 k	1.2	1.2	N/A	1	1.1	1.2
EURO-2 1996	80 k	1.2	1.2	N/A	1	1.1	1.2
EURO-3 2000	80 k	1.2	1.2	1.2	1	1.1	1.2
EURO-4 2005	100 k	1.2	1.2	1.2	1	1.1	1.2
EURO-5 2010	160 k	1.3	1.5	1	1.1	1.5	1
EURO-6 2015	160 k	1.3	1.5	1	1.1	1.5	1

perform within the regulated emissions limit to which they are certified to when new for the useful life of the vehicle. Theoretically and in practice, the regulating agencies have the authority to mandate recalls of vehicles when a model does not demonstrate emissions compliance and durability.

3.3 Emissions and fuel economy measurement system

Gaseous emissions from the tailpipe of a vehicle are collected in a constant volume sampling (CVS) emissions bench. A CVS system collects a sample of exhaust from a mixing tunnel in which the complete vehicle exhaust is mixed with air. Samples of diluted exhaust emissions are collected in bags and processed after completion of the test. The mass of the various constituents is calculated by direct measurement of the mass in the sample bag and calculation of the ratio between sample mass flow and tunnel flow. Several different types of measurement devices are employed for measurement of each constituent.

3.3.1 Emissions/fuel economy gaseous emissions

Gaseous emissions in the form of HCs, CO, CO_2, and NO_x are measured normally from the gaseous tail-pipe emissions from a vehicle. On an emissions chassis dynamometer facility, the standard method dilutes the exhaust sample in filtered air with a constant total volume flow rate at all operating conditions. Because the exhaust flow rate varies as a function of operating condition, the relative dilution of the exhaust varies also during testing. A small sample is continuously drawn from the diluted exhaust to a polytetrafluoroethylene (PTFE) (Teflon®) bag over the entire duration of the test phase. Similarly, a sample of the dilution air without exhaust is continuously drawn and sampled for comparison to the measured from the tailpipe of the vehicle. Because any constituents in the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ambient air are also collected in the emissions sample, the offset due to ambient is subtracted from the emissions samples.

By drawing a gaseous sample of emissions for a complete cycle and measuring and calculating the integrated total mass of each constituent for a cycle or portion of a cycle, this standard method provides no indication of the rate of emissions production as a function of the driving condition during the time domain of the cycle.

3.3.2 Hydrocarbon emissions

HC emissions are most often measured using a flame ionization detector, or FID. In principle, an FID measures the number of carbon–hydrogen bonds in a gas sample. A flame of hydrogen and oxygen is created. The HC sample is introduced through the flame, creating ions. An electric potential between the flame tip and the cathode is created, and the measured electric potential is proportional to the number of carbon–hydrogen bonds and therefore the HC concentration.

3.3.3 Oxides of nitrogen

NO_x, normally NO, but also NO_2 and N_2O, are measured using primarily chemiluminescence technology. First, the NO and NO_2 are typically converted into exclusively NO for measurement. NO is typically generated during high temperature combustion in an internal combustion engine, but when released into the low temperature atmosphere, NO is slowly converted to NO_2 at a relatively slow time constant. Therefore, for measurement, NO_2 is typically first converted to NO. In a chemiluminescence detector, NO and ozone (O_3) react to produce light, whose intensity is measured with a detector. The concentration of NO is proportional to the light intensity.

3.3.4 Carbon monoxide and carbon dioxide

CO and CO_2 are typically measured using an instrument that applies the principle of NDIR nondispersive infrared detector. Gas molecules of components such as CO and CO_2 interact with infrared radiation in specific wavelengths for each compound. An infrared radiation source is applied to a sample and the infrared absorption can be measured using a detector proportionally to the amount of the compound in the sample. Because the different compounds absorb infrared radiation sensitive to different wavelengths, a sampling system with multiple detection wavelengths can be used for both CO and CO_2 measurements.

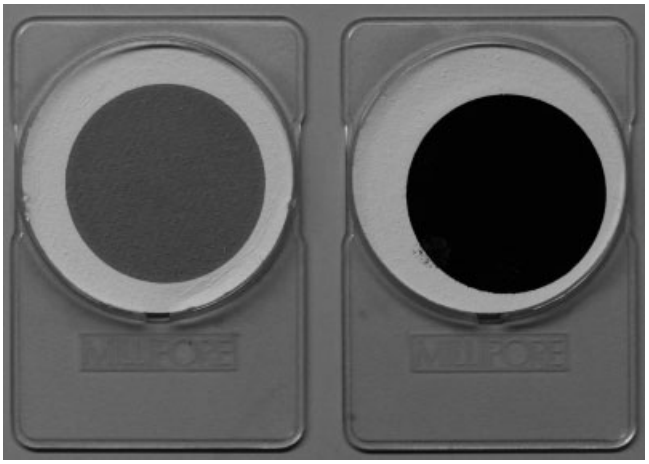


Figure 16. Two filters with deposited diesel engine exhaust particulate.

3.3.5 Particulate emissions

A special class of emissions measurements is required for vehicles powered especially by diesel engines. The particulate matter (PM) from the exhaust is sampled in a facility that dilutes and cools the exhaust. The mass of particulate is defined as the mass collected on a filter once the exhaust sample has been diluted and cooled in ambient air. The actual measurement properties of the filter are mandated in the regulations. The total mass of particulate that can be collected from the exhaust of an internal combustion engine is made up of several different sources including black carbonaceous compounds, heavy HCs that condense when cooled, sulfates, and water.

Two samples of particulate matter (PM) collected from the exhaust of a small industrial diesel engine are shown in Figure 16. The total mass of particulate collected on each filter sample is similar, but clearly the character of the particulate is significantly different. The lighter colored sample is measured at light-load conditions and is dominated by heavy HCs from fuel and lubricating oil and sulfates combined with water. The darker colored sample is measured at full-load conditions and is predominately made up of black carbon agglomerates.

By regulation, the mass of particle deposited on the filter, corrected for the ratio of sample compared to that of the total exhaust flow, is the definition of PM.

3.4 Example measurements and calculations

The following sequence of measurements is used as an example measurement of vehicle on the major components of the process for emissions and fuel economy. Some test

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

measurements are from a European process, and some are from the US process, but either illustrates a similar process.

3.4.1 Coast-down test

A test vehicle representative of the class of vehicle to be certified must perform a coast-down test on a test track. The results of the coast-down test are applied to the chassis dynamometer test facility to appropriately represent the driving load the vehicle experiences in the test cell. Depending on the manufacturer's preference, a single variant of a vehicle class may be tested; this vehicle should represent the vehicle with the maximum weight and highest rolling resistance of all vehicles represented by the class. A manufacturer may also choose to test each major powertrain or inertia class variant so that each version is tested with the lowest possible road load for that vehicle model.

The test vehicle is prepared for testing. The actual test is conducted on a straight and level track of suitable length. The vehicle is accelerated past the maximum speed of the driving cycles to which it will be evaluated—120 km/h on the NEDC. The test driver then places the transmission in neutral and the vehicle coasts from high to low vehicle speed. The time between each speed interval is measured and tabulated as shown in Table 14.

For such a test to be considered valid, typically several replicates of the test must be conducted with small variation between each sample, and the average interval times are used for further calculation. Test conditions and the facility definition are written. In most cases, it is to the advantage of the automaker to achieve the lowest resistive force and therefore reduce the relative motive power the powertrain must exert to move the vehicle in a simulated environment. Therefore, the vehicle preparation and ambient conditions are very often selected carefully. The test results must be corrected for the standard ambient conditions of the nominal chassis dynamometer test.

Table 14. Exemplary vehicle coast down test result.

Velocity (km/h)	Time Interval (s)	Force (N)
120	5.30	1069
110	6.00	944
100	6.87	825
90	7.82	725
80	8.99	630
70	10.41	544
60	11.88	477
50	13.68	414
40	15.59	364
30	17.33	327
20	19.73	287

The rate of change of vehicle speed is calculated across the intervals to determine an average acceleration rate over each 10 km/h increment.

With a documented vehicle mass, the resistive force applied to the vehicle by aerodynamic drag, rolling resistance, and rolling friction is calculated using the simple calculation of Equation 4:

$$\text{Force} = \text{Mass} \times \text{Acceleration} \tag{4}$$

Using the average acceleration from the coast-down intervals that represent a change in vehicle velocity over a measured time interval, the average force over this interval is calculated.

The resistive force applied by the chassis dynamometer should be representative of the force applied on a road. The force equation is simply represented by a quadratic equation of force as a function of vehicle speed, typically fit with a least squares regression calculation. Using the same coast-down measurements from Table 14, the quadratic coefficients can be calculated. In the United States the A, B, and C coefficients for this quadratic equation are also tabulated and available to the public. Figure 17 illustrates the resistance curve calculated from the measured coast-down times from Table 14, and the calculated A, B, and C coefficients for the quadratic equation associated with that curve. In practice, the quadratic terms are associated with functionality of a vehicle that these terms in some ways represent. Coefficient A, the square term, is associated with aerodynamics as this term is a squared function of velocity of the vehicle. Coefficient B, the linear term, is associated with rolling resistance. Coefficient C, the offset term, is associated with the constant force needed to move the vehicle from rest—mainly a force associated with tire rolling friction. The quadratic equation representing this resistive force curve is also shown with A, B, and C coefficients.

$$\text{A Coefficient: } 0.0549 \ (\text{N}/(\text{km/h})^2)$$

$$\text{B Coefficient: } 0.3443 \ (\text{N}/(\text{km/h}))$$

$$\text{C Coefficient: } 270.25 \ (\text{N})$$

3.4.2 Chassis dynamometer setup and vehicle preparation

When a vehicle is first to be tested on a chassis dynamometer facility, the initial setup must be performed to match the road load from the coast-down test. The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

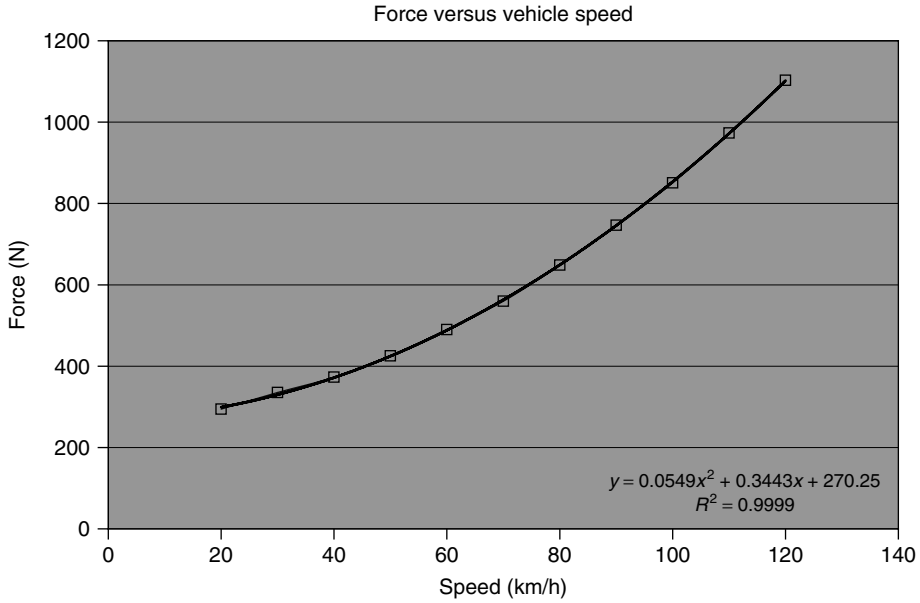


Figure 17. Resistive force and quadratic equation calculated from coast-down test.

quadratic coefficients are used initially. The measured coast-down time intervals when the test vehicle is decelerated from high speed to low speed are also considered. The chassis dynamometer facility does not ideally represent the vehicle on-road conditions, so in some ways the idealized environment of the facility is modified so that it represents as closely as possible the same resistance to motion a vehicle would experience on road driving. Once the chassis dynamometer matches the coast-down times, actual driving cycle testing can begin.

A test vehicle may be prepared with instrumentation for testing, or in some cases, the unmodified vehicle is provided for testing. A known fuel with measured properties including number of carbon atoms must be used for accurate results, because the measurement of CO_2 emissions is made directly, and fuel consumed during the test is inferred from these measurements. For either US FTP or European NEDC testing, a vehicle must be tested first on a known cycle and then soaked for a prescribed period of time at specified humidity and temperature. Before a driving cycle begins, the vehicle is rolled on the facility and fixed to the floor; the drive wheels are placed in contact with the dynamometer roller. The tail-pipe exhaust is captured by the emissions sampling system. An external cooling fan supplies cooled air to the vehicle cooling system to maintain standard conditions for the powertrain; in newer facilities, a cooling fan is controlled proportionally to the vehicle speed on the driving cycle. Ambient conditions must be maintained during the soak period before a test and also during the entire test phase.

Some incremental new driving requirements require testing the vehicle after soaking and/or while operating at different ambient conditions. In these cases, the room of the test facility must have the capability of controlling the ambient conditions. Effectively, this requires a test facility that must be able to soak vehicle for normal ambient temperatures of 20°C, for cold ambient temperatures of −7°C, and at elevated temperatures of 35°C for US SFTP requirements.

3.4.3 Test performance

The driver starts the engine and follows the instructions provided in a monitor outside the vehicle. The key instruction to follow is a vehicle speed trace in kilometer/hour that the vehicle should follow. The vehicle speed trace is prescribed in the NEDC as shown in Figure 10.

The driver is evaluated whether the prescribed driving trace is met with the actual driving cycle performed during the test using a regression analysis. If the actual driving is not close enough to the prescribed cycle, the results are invalid. Figures 18 and 19 show the actual vehicle speed measured during the completion of an NEDC, overlaid with the prescribed cycle on the complete cycle, and for a short segment. Generally, the vehicle speed must be within a tolerance of the target, and the average deviation over the complete cycle must also be within a specified tolerance.

For maximum repeatability, automated "robot" drivers have been developed and are in use in some facilities. Such

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

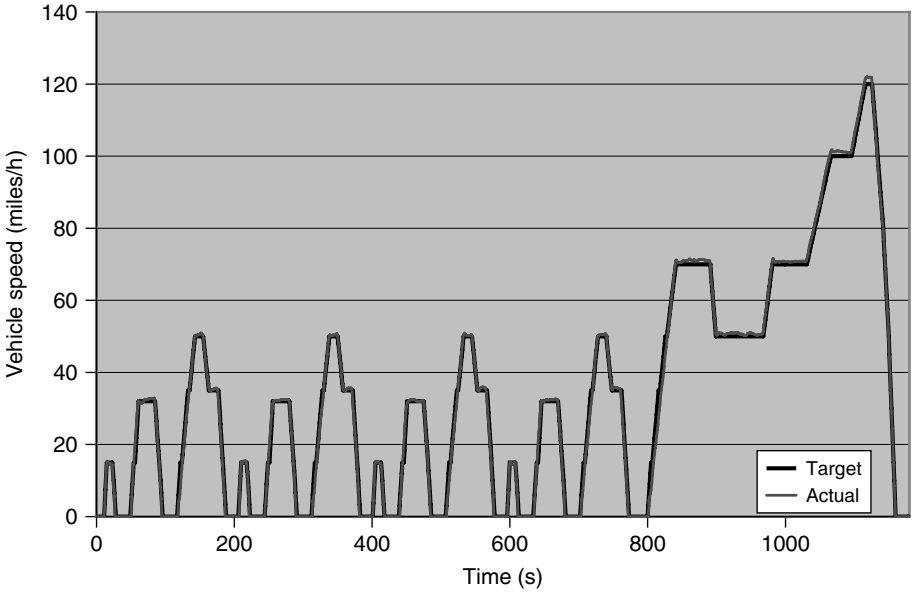


Figure 18. Target and actual vehicle speed on complete NEDC.

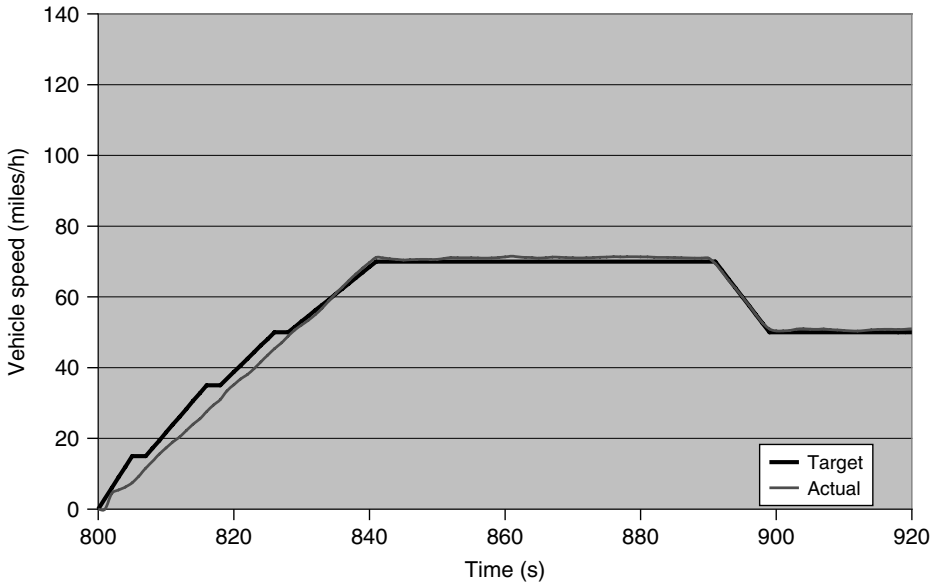


Figure 19. Target and actual vehicle speed on complete NEDC—detail.

automated drivers improve test to test repeatability and can remove some variation in results attributed to the driver.

3.4.4 Emissions postprocessing and cycle calculation

Once the vehicle completes the driving cycle, the engine is turned off and emissions sampling equipment stopped. Measurements of the regulated emissions are entirely post-processed from the emissions constituents that are collected in Teflon sampling bags. On the NEDC, two bags are required: one for the urban portion and other one for the extra-urban portion. The concentration of each emissions constituent captured during the driving cycle is measured by direct measurement of each constituent and the concentration. The total amount in grams of each constituent is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 15. Sample emissions test summary report for FTP-75 test.

EMISSIONS TEST SUMMARY REPORT		
Vehicle ID: FEV_025M993	Test ID: FEV-EPA75_2011.07.27	
Test Type: EPA75	Facility: Cell 1	Start Time: 07/27/2011 09:32:23
Shift Sched: Manual_6	Trace End: 07/27/2011 10:12:57	Start Odometer: 5450
Fuel Type: California Phase 2	Inertia Weight: 3500	
Road Load Coeff A: 26.51	Cell Temp Set Pt: 75	
Road Load Coeff B: −.4425	Altitude Set Pt(ft.): 930	
Road Load Coeff C: 0.02763	Hum. Set Pt (Grains): 50.00	

	HC	CO	NO_x	CO_2	NMHC	CH_4	NMOG	Volume MPG	EPA Correlation MPG
CVS Mass Results (g/mile)									
Phase: 1	0.06087	0.60386	0.00744	326.0	0.04815	0.01331	0.0000	26.3487	26.65
Phase: 2	0.00107	0.08579	0.00000	327.2	0.00023	0.00088	0.0000	26.3481	26.65
Phase: 3	0.00296	0.15062	0.00385	280.3	0.00028	0.00280	0.0000	30.7574	31.11
CVS Weighted Mass Results (g/mile)									
	0.01401	0.21123	0.00261	314.1	0.01020	0.00399	0.0106	27.4183	27.74

divided by the distance traveled for that phase or totaled for the complete driving cycle. Because the complete vehicle exhaust flow is too large and bulky to efficiently store and hold, only a small portion of the total exhaust flow is diverted to the emissions sampling system.

An exemplary table of emissions and fuel economy measurements from a vehicle test conducted on the FTP-75 cycle is given in Table 15.

In the regulated emissions constituents and levels, only the final bag results that are the integrated total emissions values are required. Such a test is the only accepted method used for the regulatory bodies. However, for engineering, it is invaluable to know at what event in a driving cycle the emissions rates are highest or lowest and, in addition, the aftertreatment system efficiency. For these product development requirements, emissions measured in a vehicle facility typically include also emissions sampled instantaneously and often both before and after the aftertreatment system to calculate the effectiveness. Table 16 shows emissions collected instantaneously during the completion of an FTP-75 test. In this case, the emissions before and after the catalytic converter are also measured that allows calculation of conversion efficiency. With this information, the attribution of performance to specific operating points in the cycle can be performed.

These modal emissions results are useful for product development, but are not performed according to the regulations and are not exactly identical to the prior integrated values. For example, the fuel economy estimate from this method is about 10% lower than that measured using the CVS method. It should be clear from the amount of information available that such a method greatly increases the use for a developer. For example, the conversion efficiency of the catalyst and allocation of emissions to the individual operating conditions are only available from modal measurements.

Unlike other emissions constituents, diesel particulate emissions are sampled separately from other gaseous emissions and continuously throughout the cycle. A sample of the exhaust is routed to a dilution tunnel in which it is mixed with ambient air. A portion of the diluted and cooled sample is drawn through a filter. The amount of mass deposited on the filter is defined as the mass of particulate. In most light-duty automotive facilities, only a portion of the complete exhaust is introduced into the dilution tunnel. Then only a portion of the diluted sample is routed through the filter for measurement. The ratio of the total vehicle exhaust flow rate to the sample flow rate and then the ratio of the tunnel flow to the probe flow rate must be continuously calculated or controlled.

4 VEHICLE TESTING FACILITIES

Several different types of facilities have been developed for automotive product development in a static location.

4.1 Emissions chassis dynamometer

For measurement of emissions and fuel economy according to the procedures required by the major regulatory regimes,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 16. Modal emissions summary for FTP-75 test.

		Tailpipe								Engine				Converter Efficiency %		
		HC	NMHC	CH4	CO	NO_x	CO_2	ExVol	MPG	HC	CO	NO_x	CO2	HC	CO	NO_x
Individual cycles: (g/mile)	Time-63	0.9250	0.7623	0.1815	9.0613	0.2540	750.4	25.4		3.7762	27.482	3.2925	680.7	75.5	67.0	92.3
	Cycle1	0.3244	0.2649	0.0664	3.0332	0.0901	486.0	47.1	17.5270	2.5430	14.624	2.7280	433.6	87.2	79.3	96.7
	Cycle2	0.0100	0.0057	0.0048	0.0997	0.0021	307.5	89.0	27.9687	1.5539	10.787	2.9311	283.7	99.4	99.1	99.9
	Cycle11	0.0013	0.0010	0.0004	0.0899	0.0054	278.2	57.7	30.9897	1.5478	9.2955	1.8535	253.5	99.9	99.0	99.7
	Cycle19	0.0100	0.0022	0.0087	0.1460	0.0652	332.9	34.0	25.8640	1.7547	10.694	2.4204	310.3	99.4	98.6	97.3
Phase 1 modal (g)	IDLE	0.1408	0.1191	0.0242	1.6235	0.0242	105.6	17.4	0.0794	0.6314	5.4414	0.1644	108.9	77.7	70.2	85.3
	ACCEL	0.0767	0.0546	0.0247	0.6252	0.0422	552.3	80.3	13.739	3.0751	20.5726	5.8598	510.2	97.5	97.0	99.3
	CRUISE	0.0164	0.0124	0.0044	0.1187	0.0045	524.6	74.7	29.811	2.3658	14.2974	4.5999	488.0	99.3	99.2	99.9
	DECEL	0.0091	0.0073	0.0021	0.0404	0.0018	142.6	24.3	54.174	0.9323	4.5563	0.5688	112.5	99.0	99.1	99.7
	CRANK	0.0000	0.0000	0.0000	0.0001	0.0000	0.1	0.3	0.0000	0.0004	0.0136	0.0000	1.1	94.9	99.1	78.4
	TOTAL	0.2430	0.1934	0.0555	2.4079	0.0726	1325.2	196.9		7.0049	44.8812	11.1929	1220.6	96.5	94.6	99.4
	Phase: 1 Equivalent Mass Results (g/mile)															
		0.0677	0.0538	0.0154	0.6702	0.0202	368.9	196.9	23.279	1.9498	12.493	3.1156	339.7	96.5	94.6	99.4
Phase 2 modal (g)	IDLE	0.0006	0.0005	0.0001	0.0082	0.0063	85.4	15.7	0.0605	0.671	3.8396	0.243	99.1	99.9	99.8	97.4
	ACCEL	0.005	0.0028	0.0027	0.2313	0.0155	747.9	113.1	14.752	4.104	25.155	7.1597	710.2	99.9	99.1	99.8
	CRUISE	0.0018	0.0017	0.0002	0.0819	0.0049	454.8	68.4	34.194	2.4304	13.621	2.6914	423.8	99.9	99.4	99.8
	DECEL	0.0008	0.0006	0.0002	0.0346	0.0015	145.9	25.6	44.642	1.2627	7.045	0.3146	106.2	99.9	99.5	99.5
	TOTAL	0.0081	0.0055	0.0032	0.356	0.0283	1434	222.8		8.468	49.66	10.409	1339.3	99.9	99.3	99.7
	Phase: 2 Equivalent Mass Results (g/mile)															
		0.0021	0.0014	0.0008	0.0926	0.0074	372.9	222.8	23.1	2.2023	12.916	2.7071	348.3	99.9	99.3	99.7
Phase 3 modal (g)	IDLE	0.0009	0.0003	0.0006	0.0054	0.0032	49.8	9.2	0.0346	0.3709	1.7934	0.1327	59.4	99.8	99.7	97.6
	ACCEL	0.0103	0.0024	0.0090	0.4060	0.0373	490	75.1	15.023	2.7262	20.9229	5.9922	464.0	99.6	98.1	99.4
	CRUISE	0.0027	0.0018	0.0012	0.1378	0.0156	440.9	66.2	35.453	1.9754	13.8007	4.1745	423.1	99.9	99.0	99.6
	DECEL	0.0008	0.0006	0.0004	0.0397	0.0018	109.9	19.5	71.198	0.7433	3.6897	0.4875	89.6	99.9	98.9	99.6
	CRANK	0.0000	0.0000	0.0000	0.0000	0.0000	0.5	0.1	0.0000	0.0005	0.0013	0.0000	0.7	97.6	97.2	98.8
	TOTAL	0.0148	0.0051	0.0111	0.5889	0.0579	1091.1	170.1		5.8162	40.2080	10.7870	1036.8	99.7	98.5	99.5
	Phase: 3 Equivalent Mass Results (g/mile)															
		0.0041	0.0014	0.0031	0.1643	0.0162	304.4	170.1	28.329	1.6225	11.217	3.0092	289.2	99.7	98.5	99.5
	Weighted Total Modal Equivalent Mass Results (g/mile)															
		0.0163	0.0123	0.0045	0.2323	0.0124	353.2	592.2	24.3897	1.9904	12.3605	2.8750	330.3	99.2	98.1	99.6

Encyclopedia of Automotive Engineering. Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

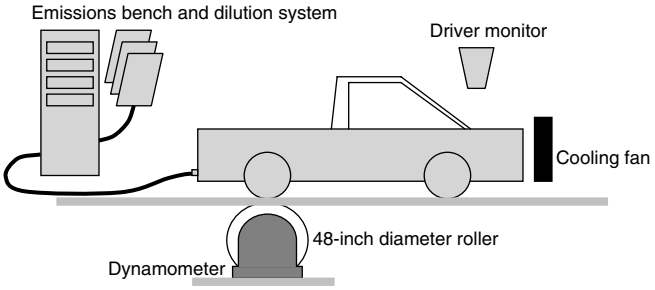


Figure 20. Emissions chassis dynamometer facility.

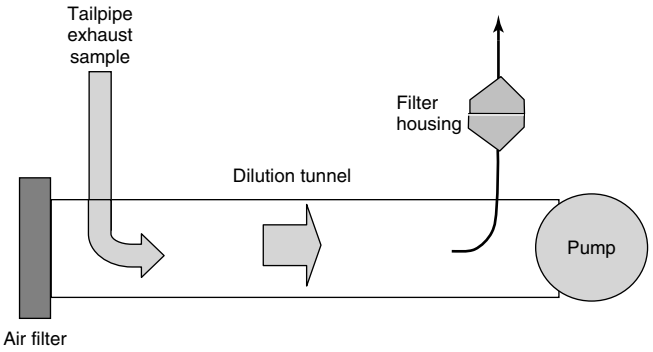


Figure 21. Particulate emissions dilution tunnel.

the emissions chassis dynamometer facility must consist of the major components shown in Figure 20. First, the facility must have rooms that control temperature and humidity in which the vehicle is soaked before testing. The capability to control temperature and humidity in the actual room in which testing is conducted must also be provided. Often, a facility with multiple test cells has attached a soak room in which multiple vehicles are soaked in preparation for testing.

The chassis dynamometer itself consists of a large drum on which the tires of the drive axle of the test vehicle sit. The drum is both restrained and in some cases driven by an electric dynamometer controlled by microprocessor-based system. The emissions from the vehicle tailpipe are collected by a duct attached to and sealed to the tip of the exhaust of the vehicle. The complete vehicle exhaust sample is routed to a dilution tunnel and mixed with sample air, which is pumped through at fixed flow rate. The sample air is controlled for temperature and pressure and humidity in regulated tolerance range. For facilities that are located where the ambient air contains significant levels of CO or HCs, the sample air can first be treated with filtration or catalyst technology to reduce the ambient levels introduced with the dilution air. A small sample of the diluted exhaust is drawn into a set of bags and stored for processing after the test is completed. If modal emissions are also measured, an additional sample probe is placed in the dilution tunnel for additional measurements, and a second set of emissions analyzers that can measure at higher sampling speed is required (Figure 21).

4.2 Environmental chassis dynamometer

An environmental facility is similar to the facility used for emissions measurement, with the addition of ambient environmental conditions. For emissions requirements, the ambient temperature range for testing must be extended to a temperature of -7°C (20°F) for both European and US requirements. Also for US requirements, an ambient temperature range must be extended to ambient temperatures of 95°F (35°C).

For evaluation of the performance of a powertrain at conditions that are outside the normal range, but still experienced by customers, an environmental facility may also be specified for continuous operation at cold or hot ambient temperature, high altitude, testing, and various humidity levels.

4.3 Evaporative emissions test facility

Emissions of volatile organic compounds (HCs) that are not from the tailpipe of running engine but from liquid fuel evaporating from the vehicle are also regulated in United States and California. The facility for measuring this category of emissions is called a *sealed housing for evaporative determination* (*SHED*). Such a facility is used to capture the HCs that are emitted by a vehicle from fuel evaporating from the fuel system components and fuel tank when the vehicle is running or not running. In the United States, the emissions of fuel vapor during the fuel tank refueling process are also monitored and limited. Typically, the mass emissions of HCs are limited during an extended test period under which the vehicle is tested, soaked at high temperature, or refueled. The low speed New York City Cycle, shown in Figure 22, which was developed in similar time frame to the FTP-72 cycle is employed for some parts of evaporative emissions testing for US requirements.

Recently, hybrid vehicles with potential for energy storage and energy recovery have required additional standards to represent the state of charge applied at the beginning, and how much that available energy can be depleted and included in the cycle calculations for emissions and fuel economy.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

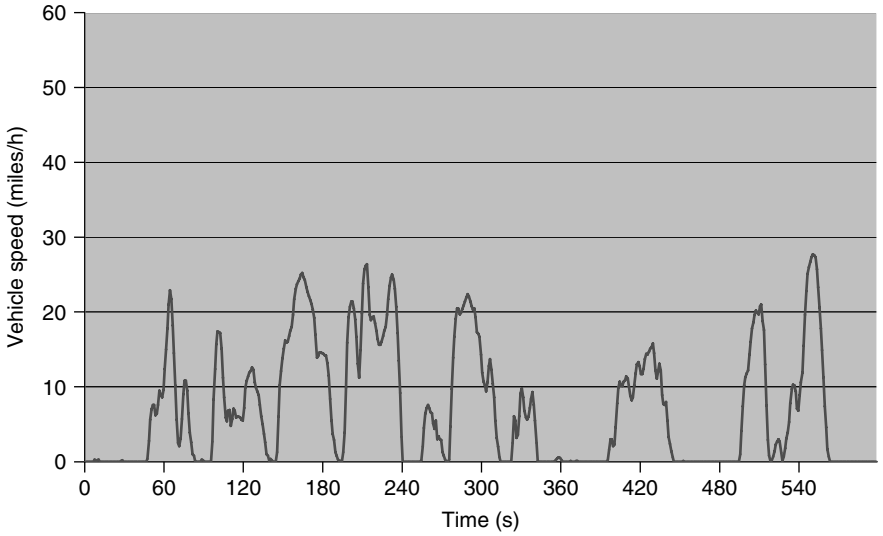


Figure 22. New York City Cycle.

USEFUL WEBSITES

http://www.epa.gov/ld-hwy.htm
http://www.arb.ca.gov/html/lawsregs.htm
http://www.arb.ca.gov/html/brochure/history.htm
http://ec.europa.eu/environment/air/transport/road.htm
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31970L0220:en:NOT
http://www.env.go.jp/en/air/aq/mv/st andards.html

REGIONAL WEBSITES

British Columbia: http://www.aircare.ca/
Ontario: http://www.ene.gov.on.ca/environment/en/category/drive_clean/index.htm
Japan: http://www.navi.go.jp/english/index.html
United Kingdom: http://www.direct.gov.uk/en/Motoring/OwningAVehicle/Mot/DG_4022109
Alaska: http://doa.alaska.gov/dmv/reg/imtest.htm
Arizona: http://www.myazcar.com/
California: http://www.dmv.ca.gov/vr/smogfaq.htm#BM2540
Colorado: http://www.aircarecolorado.com/
Connecticut: http://www.ctemissions.com/
District of Columbia: http://dmv.dc.gov/serv/inspections.shtm
Delaware: http://www.dmv.de.gov/services/vehicle_services/other/ve_other_general.shtml
Georgia: http://www.cleanairforce.com/
Idaho: http://www.emissiontest.org/
Illinois: http://www.epa.state.il.us/air/vim/index.html
Indiana: http://www.in.gov/bmv/2655.htm
Louisiana: http://www.lsp.org/weights.html
Maine: http://www.maine.gov/dps/msp/vehicles_inspections/motor_vehicle_inspections.html
Maryland: http://www.mva.maryland.gov/MVA-Programs/VEIP/default.htm
Massachusetts: *http://www.massvehiclecheck.com/*
Missouri: http://dor.mo.gov/motorv/help.php#emissions
Nevada: *http://dmvnv.com/emission.htm*
New Hampshire: *http://www.nhinspect.com/*
New Jersey: http://www.state.nj.us/mvc/Inspections/index.htm
New Mexico: http://www.cabq.gov/vehicle-pollution-management/emissions-testing
New York: http://www.nydmv.state.ny.us/vehsafe.htm
North Carolina: http://daq.state.nc.us/motor/inspect/
Ohio: http://epa.ohio.gov/Default.aspx?tabid=2425
Oregon: http://www.oregon.gov/ODOT/DMV/vehicle/emissions.shtml
Pennsylvania: http://www.drivecleanpa.state.pa.us/default.htm
Rhode Island: http://www.riinspection.org/
Tennessee: http://www.state.tn.us/environment/apc/vehicle/
Texas: http://www.txdps.state.tx.us/vi/inspection/item_insp.asp
Utah: http://dmv.utah.gov/registerinspections.html#safety
Vermont: http://dmv.vermont.gov/safety/laws/emissions
Virginia: http://www.deq.state.va.us/mobile/mobfaq.html
Washington: http://www.dol.wa.gov/vehicleregistration/emissions.html
Wisconsin: http://www.wivip.com/

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ABBREVIATIONS

CVS	constant volume sampler
NO_x	oxides of nitrogen (mainly NO and NO_2)
HC	hydrocarbons
THC	total hydrocarbons
CO	carbon monoxide
CO_2	carbon dioxide
NMOG	non-methane organic gases
NMHC	non-methane hydrocarbons
SHED	sealed housing for evaporative determination
CAFE	Corporate Average Fuel Economy

FURTHER READING

1. Degobert, P. (1992) *Automobiles and Pollution*, Editions Technip (ISBN 2-7108-0628-2), Paris.
2. Faiz, A., Weaver, C., and Walsh, M.P. (1996) *Air Pollution from Motor Vehicles*, the International Bank for Reconstruction and Development, Washington: D.C.
3. Committee on State Practices in Setting Mobile Source Emissions Standards, National Research Council (2006) *State and Federal Standards for Mobile-Source Emission*, The National Academies Press: Washington, DC. Available at: http://books.nap.edu/catalog.php?record_id=11586.
4. United States Environmental Protection Agency (2010) *Light-Duty Automotive Technology, Carbon Dioxide Emissions, and Fuel Economy Trends: 1975 Through 2010*, EPA420-R-10-023, U.S. Environmental Protection Agency: USA. Available at: http://www.epa.gov/oms/cert/mpg/fetrends/420r10023.pdf.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto151
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fuel Economy Optimization

Stephen F. Bowyer, Dean Tomazic, and Gary W. Rogers
FEV, Inc., Auburn Hills, MI, USA

1 Introduction 1
2 Overview of Future Fuel Economy Regulations 1
3 Overview of Internal Combustion Engine Efficiency 2
4 General Fuel Efficiency Enablers 4
5 SI Engine-Specific Enablers 8
6 CI Engine-Specific Enablers 12
Further Reading 14

1 INTRODUCTION

The efficiency of a thermomechanical system is defined as the ratio of the mechanical work done versus the amount of fuel energy input per unit time. As applied to the automotive engine, the chemical energy in the fuel is converted into heat, which is in turn converted into the mechanical work extracted from the crankshaft. The fuel efficiency (fuel quantity/distance traveled) and fuel economy (distance/fuel quantity) of the vehicle are related to the vehicles mass and other dynamic parameters associated with the vehicle being propelled such as aerodynamic and rolling resistances. To achieve the lowest possible fuel consumption, while maintaining high customer satisfaction, in a given vehicle powered by an internal combustion engine, parameters such as mechanical design, fluid dynamics (gas exchange), combustion cycle, exhaust emissions, torque, electronic controls, noise and vibration, and torque versus speed must be optimized. In some cases, higher vehicle performance must be sacrificed to achieve government-regulated fuel economy and exhaust emission standards. This chapter outlines the influences major components and subsystems have on thermodynamic energy conversion and the mechanical losses that limit total system efficiency and identifies advancements being developed to optimize fuel economy.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 OVERVIEW OF FUTURE FUEL ECONOMY REGULATIONS

The United States recently enacted legislation requiring automotive manufacturers meet tough new Corporate Average Fuel Economy (CAFE) standards. These new standards will score a vehicle's fuel economy with respect to the size of the vehicle's footprint. The new CAFE standard will require the manufacturers to sell vehicles in order to obtain a fleet average of approximately 54.5 mpg by 2025 or receive heavy fines. There are some variances to this target as the CAFE targets are adjusted by the types of vehicles the manufacturer sells and, therefore, typical footprint of the vehicles. By comparison, the former CAFE standard was 27.5 mpg for passenger cars.

The European Union (EU) has already converted largely to more fuel-efficient vehicles because of the fuel taxes levied within the region. Further, EU regulations under discussion stipulate that manufacturers must limit their European fleet average CO_2 emissions to 130 g/km by 2012. Ongoing negotiation may push future CO_2 levels below the 100 g/km level by 2020.

The regulations in place and in discussion in developed economies and the fuel costs in developing countries are putting new pressure on automotive manufacturers. As

such, many of the fuel-saving technologies outlined in this chapter are under research, being developed, and beginning to be implemented in future engine programs faster than ever before.

3 OVERVIEW OF INTERNAL COMBUSTION ENGINE EFFICIENCY

Internal combustion engines have been the primary source of mobile power for more than a century. Two combustion cycles, spark ignition (gasoline or alcohol) and compression ignition (CI) (diesel), have dominated the automotive industry since its inception for their excellent power density and load flexibility. However, there are significant inefficiencies inherent with both cycles and differences in efficiencies between the two resulting from the way load is controlled.

Most modern automotive gasoline engines operate at approximately 36% overall engine efficiency, whereas diesel engines are capable of just over 40%. The largest share of wasted energy for both cycles escapes the engine through the exhaust. The loss of 25–45% of the fuel's chemical energy through the exhaust is a result of the difficulty in extracting energy from high temperature but low pressure gas. The remaining losses include heat transfer to the cooling system, internal friction of the engine, and unburned fuel. An example of the energy split from a turbocharged, direct-injected gasoline engine is shown in Figure 1 illustrating the distribution of fuel energy for a high power density engine.

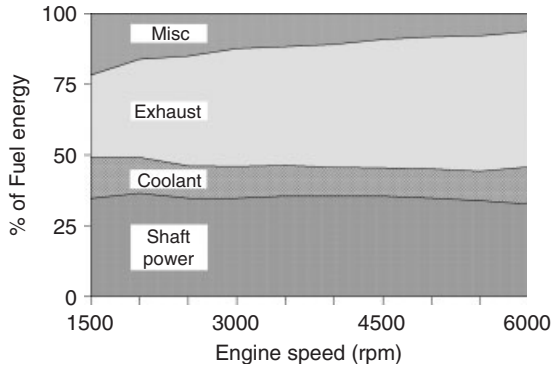


Figure 1. Energy split for a turbocharged, direct-injected gasoline engine at full load. (Reproduced by permission of FEV Inc.)

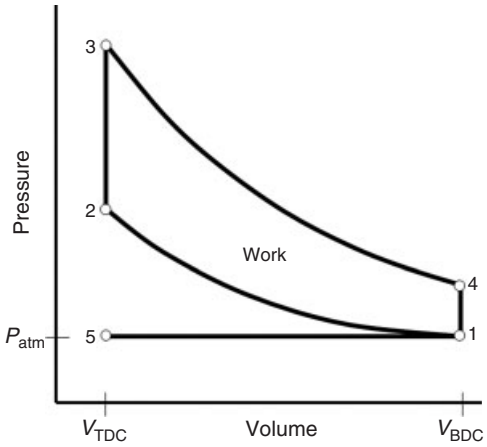


Figure 2. P–V diagram for an ideal Otto cycle. (Reproduced by permission of FEV Inc.)

3.1 Spark-ignited engines

Spark-ignited (SI) engines operate using the Otto cycle, a theoretical constant volume combustion process as shown in Figure 2. In this process, fresh charge is pulled into the cylinder as the piston expands the volume in the chamber from points 5 to 1. From there, the air/fuel mixture in the cylinder undergoes adiabatic compression from points 1 to 2. Heat is added to the cylinder through combustion at such a rapid rate that it occurs at a constant volume condition as shown from points 2 to 3. From point 3, the gas undergoes adiabatic expansion to point 4 before being exhausted to atmosphere where the conditions in the cylinder return to their starting point at point 1. The work output of the engine is defined by the area bounded by points 1–4.

The efficiency of this cycle is determined by the compression ratio. Equation 1 defines this relationship.

$$\eta_{th,v} = 1 - \frac{1}{r_v^{\gamma-1}} \quad (1)$$

where γ = ideal gas constant and r_v = compression ratio.

In practice, several aspects of the P–V diagram vary from that in Figure 2. The charge is compressed nearly adiabatically from points 1 to 2 but there is some heat transfer between the gas and the surrounding chamber. While combustion duration is very short for modern engines, typically over 30–50° crank angle, it is not a constant volume process. As such, the combustion must be initiated before top dead center (TDC) and continues for several crank degrees beyond TDC giving a rounded-off appearance between points 2 and 3. Expansion of the combustion gas is not purely adiabatic, as a small amount of heat is transferred to the components making up the combustion chamber. The exhaust valve opens before the piston reaches bottom dead center (BDC) in order to provide time for the combustion pressure to be relieved before the piston begins the exhaust stroke on its way to point 5. Finally, flow losses through the induction system and intake valve

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

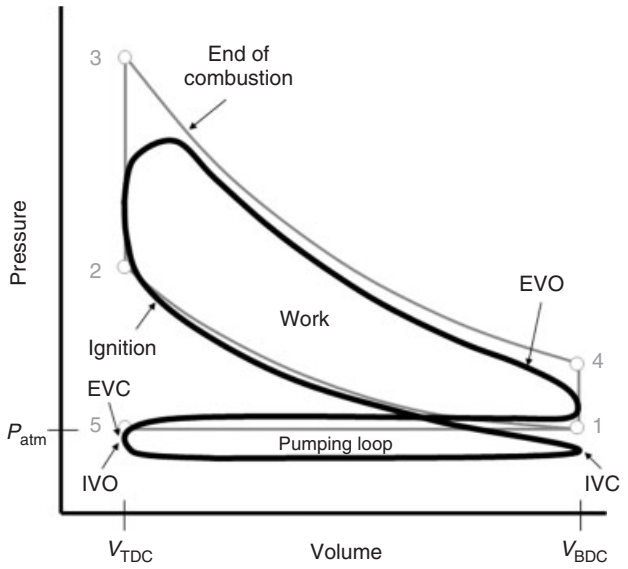


Figure 3. Ideal Otto cycle P–V diagram with actual gasoline cycle overlaid. (Reproduced by permission of FEV Inc.)

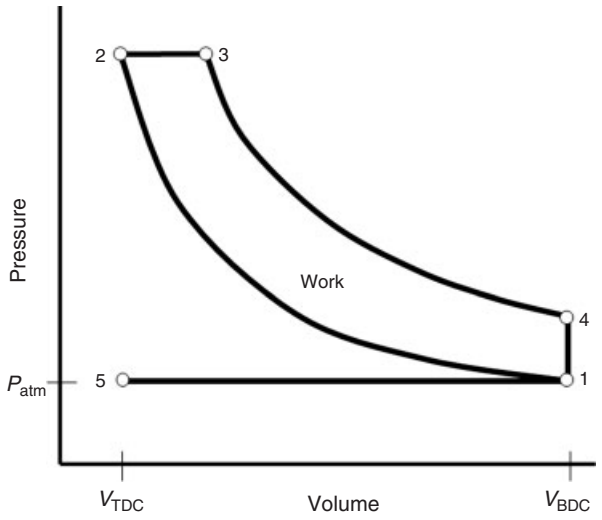


Figure 4. P–V diagram for an ideal diesel cycle. (Reproduced by permission of FEV Inc.)

cause a negative depression above the piston relative to the pressure in the crankcase. This pumping loop of negative work reduces the available output of the engine. An overlay of P–V diagrams for an ideal and actual cycle is shown in Figure 3.

For SI engines, load demand by the driver is varied through the use of a throttle blade, which meters the amount of air allowed into the engine. Throttling of the incoming air charge reduces engine output by reducing the amount of fresh charge allowed into the engine that is desired. In addition, though, the amount of work required for the pumping loop is increased, as manifold pressure is reduced further below atmospheric pressure. All of these modifications to the cycle cause the real efficiency of an SI engine to deviate from the efficiency calculated from Equation 1. The most significant of these losses are created by the negative work from the pumping loop. Reduction in the amount of work required for load metering is the major focus for improvement in gasoline engines.

3.2 Compression ignition engines

CI engines operate using the Diesel cycle, a theoretical constant pressure combustion process as shown in Figure 4. In this process, fresh charge is pulled into the cylinder, as the piston expands the volume in the chamber from points 5 to 1. From there, the air in the cylinder undergoes adiabatic compression from points 1 to 2. Heat is added to the cylinder through the combustion of fuel. The fuel is sprayed into the chamber at a rate such that the pressure in the cylinder remains constant from points 2 to 3. From point 3, the gas undergoes adiabatic expansion to point 4 before being exhausted to atmosphere, where the conditions in the cylinder return to their starting point at point 1. The work output of the engine is defined by the area bounded by points 1–4.

The efficiency of this cycle is determined by the compression ratio and an inverse of the cutoff ratio. The cutoff ratio is defined by the distance the piston travels while heat is being added to the chamber divided by the total stroke. Thus, for a given stroke, a short heating time and long post-expansion provide the best efficiency. Equation 2 defines this relationship.

$$\eta_{\text{Diesel}} = 1 - \frac{1}{(\gamma \times {r_v}^{\gamma-1}) \times \left[\dfrac{({r_c}^{\gamma} - 1)}{(r_c - 1)}\right]} \tag{2}$$

where γ = ideal gas constant, r_v = compression ratio, and r_c = cutoff ratio.

As with the SI engine cycle, real-world diesel P–V diagrams vary from ideal. As with the SI engine, both compression and expansion strokes are not quite isentropic, as heat is transferred between the combustion chamber and gas during these processes. Multiple fuel-injection events before TDC are frequently used to mitigate noise. Fuel delivery at this point does not match a pure constant pressure system, which also contributes to the deviation between points 2 and 3. Like the SI engine, the diesel engine requires that the exhaust valve opens early in order to allow for blowdown before the piston begins the exhaust

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

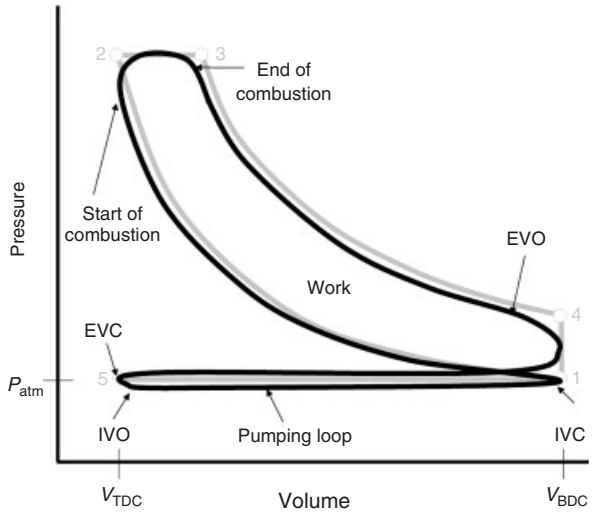


Figure 5. Ideal diesel cycle $P-V$ diagram with actual diesel cycle overlaid. (Reproduced by permission of FEV Inc.)

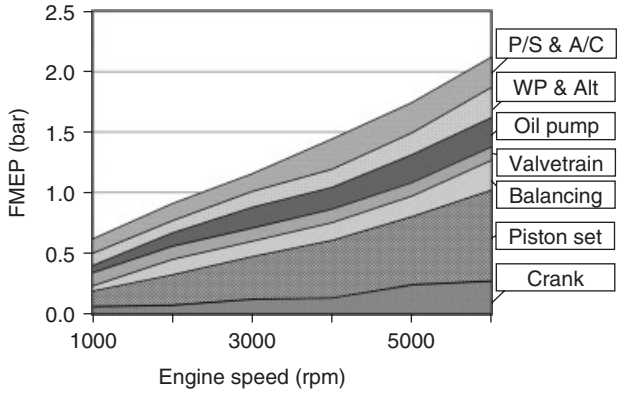


Figure 6. Motored engine friction by strip method. (Reproduced by permission of FEV Inc.)

stroke causing a deviation at point 4. Pushing the exhaust gas out of the cylinder and pulling fresh air into the cylinder creates a pumping loop caused by the flow losses through the valves as the piston moves to and from point 5. Deviations from the ideal cycle can be seen in Figure 5.

Load in a diesel engine is managed by the amount of fuel injected into the combustion chamber. This provides a positive advantage relative to the SI engine cycle at part load. By comparison, the diesel-pumping loop actually decreases at part load because of the reduced port velocities during low boost conditions. Diesel exhaust, however, cannot be cleaned with a simple three-way catalyst (TWC) as the SI engine is. The significant oxides of nitrogen (NO_x) and particulate emissions created by the diesel engine require significant aftertreatment systems, which add cost and reduce the efficiency of the diesel engine. The efficiency losses are caused by increasing the effort to pump exhaust gases through the exhaust system and occasional exhaust system regeneration process, which burns additional fuel to clean particulate filters and NO_x traps.

4 GENERAL FUEL EFFICIENCY ENABLERS

4.1 Reduction in parasitic losses

4.1.1 Friction

Reduction in engine friction helps improve fuel consumption by reducing the work that is consumed by the various sliding motions within the engine. An example of an engine's friction contribution by subsystem is shown in Figure 6.

4.1.1.1 Piston and ringpack. As can be seen in Figure 6, the piston and ringpack are the largest contributors to the engine's overall friction. Several enablers are used to reduce friction in this assembly but the most effective are reductions in pretension of the piston rings. In order to achieve the greatest reduction in piston ring tension, great effort must be put into minimizing cylinder bore distortion. Typically, the most challenging order of distortion is related to the number of cylinder head fasteners used per cylinder. In most automotive applications, fourth-order distortion is then most prominent and should be reduced below 5 μm in order for reduced ring pack tensions to be considered. If bore distortion can be managed, significant reductions in oil control ring tensions can be achieved.

Another option to reduce piston group friction is to offset the crankshaft by somewhere between 0 and 20 mm in the direction of major thrust as illustrated in Figure 7. This puts the connecting rod in a more vertical orientation during high cylinder pressures, thus reducing the side force of the piston on the cylinder bore. However, the friction is increased during the compression stroke. During high load conditions, the increased compression stroke friction is offset by the reduction in friction during the power stroke providing a net friction reduction. At low loads, however, the friction improvement during the power stroke friction may not fully compensate the increase in the compression stroke, potentially leading to a net increase in friction. As such, engines that will see high load for most of their operation, such as hybrid applications, range-extending engines, and extreme downsize applications, will see the most benefit for offset crankshafts. Thus, it is important to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

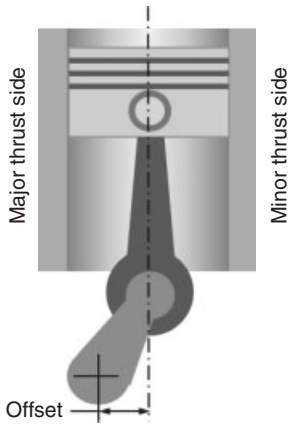


Figure 7. Illustration of crankshaft offset. (Reproduced by permission of FEV Inc.)

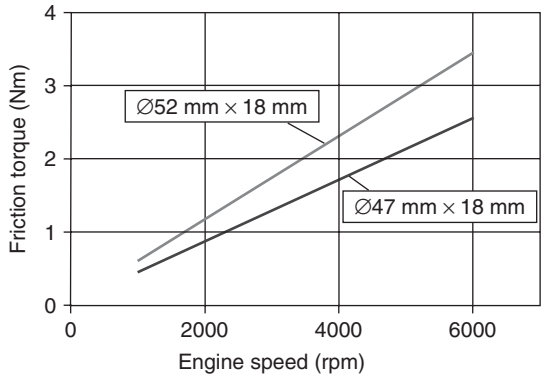


Figure 8. Friction comparison of different bearing sizes. (Reproduced by permission of FEV Inc.)

know the intended drive cycle of the engine to determine the magnitude of offset and benefit of doing so.

4.1.1.2 Dynamic seals. New dynamic seals on the market have the ability to reduce friction of rear main and front main seals. These seals, with more compliant lip designs and lower friction coefficient materials, have the ability to reduce seal friction by 30–70% compared to conventional polytetrafluoroethylene (PTFE) (Teflon®) seals. Reducing the speed of the shaft in contact with the seal contacts by reducing shaft diameters can also have measurable benefits. Great effort should be made during the design phase to minimize shaft diameters riding on dynamic seals and thereby reducing the frictional drag caused by the seals.

4.1.1.3 Bearings. Reductions in the sizes of various engine bearings throughout the engine can provide a measurable reduction in engine friction. Figure 8 illustrates a comparison of a main bearing that is 52 × 18 mm (diameter × width) to one that is reduced to 47 × 18 mm. Here, a reduction in driving torque of 0.3 Nm at 2000 rpm for a typical four-cylinder engine with five main bearings can be seen. At rated engine speeds, an additional half of a kilowatt is consumed by the friction of the larger bearing.

There have been many investigations into utilization of roller bearings in place of various plain journal bearings throughout the engine; however, these have not been implemented in production engines because of various durability issues, cost relative to plain bearings, and relatively small frictional improvement at most engine speeds. Owing to the efficiency at high speeds for plain bearings, it is anticipated that plain bearings will remain the dominant bearing for the foreseeable future. The valvetrain, however, which rotates at half of the engine's speed may see some roller bearings used, particularly in the highly loaded front journal. Camshaft tappets, finger followers, and rocker arms have already seen significant penetration of roller use in order to improve valvetrain friction and allow the use of more aggressive valve events.

4.1.1.4 Advanced coatings. Several coatings throughout the engine have been evaluated for improved friction behavior. The most common coating used in the engine for improved friction and scuff resistance is a molybdenum disulfide coating, which is screen printed onto the piston skirt. Another coating that is receiving a lot of attention for its improved friction characteristic is diamond-like carbon (DLC). DLC is a coating that is formed and leaves a very hard and smooth surface on the part providing a very low friction coefficient. Several engines now employ DLC coating on surfaces that slide with only splash oil lubrication such as gudgeon pins and valvetrain parts.

4.1.1.5 Advanced lubricants. Advanced lubricants can also help improve an engine's frictional losses. Synthetic engine lubricants can provide improved friction features while simultaneously improving temperature capability. Friction is not the only requirement for these advanced engine lubricants, as features such as engine sensor compatibility and soot loading are also important and must be considered.

4.1.2 Accessories

Engine accessories can create a significant drag on the engine as shown in the friction strip measurements in Figure 6. Many of these components impart a load on the engine even when they are not in use. The parasitic

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

losses associated with conventional accessory drives, in combination with advanced fuel economy enablers such as start/stop and full hybrid powertrains, have driven significant changes to accessory drives.

Power steering pumps are one of the most common components to be removed from the accessory drive in conventional powertrain configurations. Conventional pumps bypass the fluid when the driver does not demand power-assisted steering. Although it is not pumping the fluid to a high pressure during this standby mode, hydraulic work is being conducted with no benefit to the driver. Electric power-assisted steering (EPAS) and electro-hydraulic power steering (EHPS) are replacing accessory drive power steering units. The power-assist feature is provided by an electric motor or hydraulic pressure generated by an electric pump. In either system, the electric motor operates only when commanded by the driver, thus eliminating the standby losses of a pump in the accessory drive.

AC compressors are also starting to be removed from the accessory drive and converted to electrical operation. Conventional compressors have a clutched pulley that eliminates all but the bearing friction within the compressor. Although converting mechanical energy from the crank to electrical energy through the alternator and then back to mechanical energy of the compressor is inefficient, it is necessary to enable other fuel-saving technologies such as stop/start (engine shutdown during prolonged idle to reduce fuel consumption) and full hybrid powertrains.

Some manufacturers have started using electrical coolant pumps to better match cooling demand to cooling supply. Conventional pumps provide coolant flow based on engine speed because of the direct couple to the engines' crankshaft. This means that a cooling system sized to provide cooling power at high load is oversized during low load conditions. An electric pump's speed is decoupled from the engine and has the ability to match the engine's cooling requirements independently, thereby eliminating the excess hydraulic work performed by a mechanical pump.

4.1.3 Thermal management

Advanced cooling systems can be employed to reduce engine friction and improve fuel economy. These systems include computer-controlled thermostats or electric cooling pumps to meter the coolant flow through the engine. This enables engine calibrators to allow the engine to run hotter during low load, which keeps engine oil hotter and less viscous while also allowing the engine to receive more cooling during high demand to protect the durability of the engine's structural components.

There are several configurations that can be used in an advanced cooling system to speed up the heating of engine and transmission fluids. These strategies can be particularly effective if the engine is equipped with an exhaust gas recirculation (EGR) cooler. In such a configuration, coolant is first circulated through the EGR cooler to absorb heat from the exhaust gas and then routed through the oil cooler where heat is then transferred to the oil. The reduced viscosity of hot oil reduces the friction of the engine during cold start operation and improves fuel consumption during this cycle.

4.1.4 Idle

The fuel an engine consumes while idling at a stop light is wasted, as the intent is strictly to keep the engine in standby mode. Start/stop strategies are now being implemented to reduce the fuel consumed during idle. The engine automatically turns off when the vehicle is stopped and then automatically restarts when the driver presses the accelerator pedal. Start/stop systems require a more robust starter motor and may require one-way clutches in the system, such as at the flywheel and alternator, to manage various inertias in the engine system.

4.1.5 Lubrication system

Similar to the cooling pump, the lubrication pump is sized for one or two key load conditions but is tied to the engine's speed. As such, the lubrication pump provides an oversupply of oil through much of the engine speed range. The lubrication pump is also sized for operation at maximum operating temperatures and maximum design tolerances (e.g., bearing clearance), which also implies that at normal operating temperature with normal tolerances, the pump capacity is larger than required. Two-stage and variable displacement lubrication pumps are sometimes employed to limit these losses. Two-stage pumps have two sets of pump elements in the pump body with the high capacity pump being used at low speeds and the low capacity pump being used at high speeds.

Variable displacement pumps are vane-type pumps, which use oil pressure to vary the eccentricity of the pump. These pumps automatically adjust to the lubrication flow rate required and are tuned by sizing a spring that determines the maximum oil pressure achieved by the pump. While vane-type pumps are inherently inefficient because of the size of the leak paths within the pump, these pumps can offer a significant overall improvement to the engines fuel economy.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.2 Cycle efficiency

4.2.1 Atkinson cycle

A modified version of the cycle invented by James Atkinson has been developed in recent years. In this cycle, the intake valve is either held open longer or closed before BDC, such that fresh charge that is trapped in the cylinder is reduced by pushing some portion of the charge back into the inlet tract before the valve closes or reducing the charge pulled into the cylinder by cutting the intake event short thereby making the expansion stroke significantly longer than the effective compression stroke. This extended expansion stroke reclaims some of the lost work when the cylinder pressure is released to the exhaust system during the blowdown process. The effective compression ratio in this cycle is significantly lower than the geometric compression ratio because of the shortened compression stroke after the intake valve is closed. As such, the geometric compression ratio must be increased in order to mitigate the inefficiency caused by a lower compression ratio.

The reduced effective displacement of the engine needs to be mitigated to keep the power density at an acceptable level. Most applications using the Atkinson cycle are hybrids, which use the electric motor to augment the engine's power. It is possible, though, that application of variable valve timing in conjunction with a variable compression ratio system would be able to convert an Atkinson cycle engine with high geometric compression and late valve timing to a conventional Otto cycle combustion system for high load operation.

4.2.2 Miller cycle

Ralph Miller developed the Miller cycle to mitigate the power loss associated with the Atkinson cycle. In this cycle, a supercharger is used to boost the engine to power density levels of a conventional Otto cycle. The additional benefit that is realized with compressing the charge air outside of the cylinder is that the heat generated from the compression of the gas can be removed before it is ingested in the engine. As such, a higher cylinder pressure can be achieved at the conclusion of the compression stroke without the risk of knock in SI engines or increased NO_x production in CI engines.

4.3 Waste gas energy

The majority of lost energy from an internal combustion engine is the high temperature exhaust gas as shown in Figure 1. There are several strategies to recover some of the energy lost to the exhaust, the most common of which is turbocharging. The expansion of exhaust gas through a turbocharger spins a turbine wheel, which is attached to a compressor wheel that compresses the incoming intake charge. Turbochargers do increase the pumping work of the engine; therefore, they do not improve fuel economy by themselves but their ability to enable downsizing of the engine and higher thermal efficiencies because of higher in-cylinder peak pressures can provide a significant improvement in overall fuel efficiency.

Turbo-compounding operates under the same process as the turbocharger in the way it extracts energy from the exhaust. However, the shaft is not connected to a compressor wheel but is instead mechanically connected to the output shaft of the engine or, in some cases, can be connected to an electrical generator or hydraulic pump. For the systems realized to date, the friction of the system creates a drag on the engine at low loads but offers net positive work output at high loads. A beneficial turbo-compounding system has not been realized to date for automotive applications because of this effect but it has been found to be beneficial in heavy-duty trucking applications because of their duty cycle with extended high load operation.

4.4 Engine mass

Reduction in engine mass does not improve the specific fuel consumption of the engine but can play a role in the total fuel economy of the vehicle. Particular attention should be paid to the largest powertrain in the vehicle. Mass reductions for the heaviest powertrain in a particular vehicle will allow reductions to the vehicle structure, suspension, and braking systems, which compound the benefits that the engine can achieve on its own.

Downsizing the engine's displacement can have beneficial mass reductions as well depending on the level of downsizing. In particular, if the largest engine available in a vehicle can be reduced from a V8 to a V6 or a V6 to an I-4, the maximum powertrain mass that must be accounted for in a vehicle can be reduced and the vehicle mass reductions discussed earlier can be realized.

When considering mass reduction, materials are always the first consideration. Lightweight materials such as magnesium, composite plastics, and aluminum have seen significant penetration in modern automotive applications. Beyond this, several hybrid cylinder block and head construction concepts have been considered with some market penetration in recent years. These hybrid structures utilize a lightweight material for most of the component while using other materials for their strength or wear properties. Audi and BMW both have cylinder blocks in production that utilize an aluminum structure at the core of the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

component but with a lightweight magnesium casing that encloses the engine structure for extreme weight reduction.

Mass reductions of components within the engine can make secondary mass reductions possible. In particular, piston and connecting rod mass reductions can reduce crankshaft mass because of reduced balance mass requirements. In addition, reduced reciprocating and rotating mass forces allow for reduced cylinder block structure, especially in high speed gasoline applications.

4.5 Hybrid strategies

Engines that are used in hybrid applications or as range extenders for electric vehicles are not required to have the same range of flexibility as those used in a conventional powertrain. This enables the powertrain to be tuned for a very narrow speed and load band. Valve events, intake and exhaust tuning, coolant pump sizing, and lubrication pump sizing can be optimized for fuel economy when the engineer is not concerned with the extreme or obscure load conditions required from conventional powertrains.

5 SI ENGINE-SPECIFIC ENABLERS

5.1 Throttling losses

SI gasoline engines exhibit several losses resulting in lower engine efficiency. The majority of the losses are based on throttling of the intake charge because of the quantitative engine load control principle. Load control through throttling of the intake charge leads to high pumping losses and a nonoptimized in-cylinder combustion process.

5.1.1 Exhaust gas recirculation

EGR as a means to reduce engine-out NO_x emissions on diesel engines was already introduced decades ago. Increasing the inert gas (CO_2) content with high specific heat capacity inside the cylinder, it results in lower in-cylinder peak temperatures and thus reduced NO_x emissions. Although a common technology for all automotive diesel engines exists nowadays, its application on gasoline engines has been limited. However, EGR can be used in SI engines to displace fresh air/fuel charge and allow for reductions in throttling losses and consequently result in reductions in fuel consumption. It can be assumed that, in combination with further future tightening of emissions legislation such as EU6 and LEV III, it will be utilized more and more on upcoming applications. In addition to exhibiting beneficial behavior at part-load conditions, EGR also shows significant advantages during high load conditions by mitigating knocking while simultaneously reducing fuel consumption due to drastically reduced enrichment requirement.

EGR can be introduced back into the cylinder either internally or externally. Internal EGR is achieved by holding both intake and exhaust valves open simultaneously, known as *valve overlap*, during gas exchange near TDC. For fixed valve timing, the amount of EGR induced is mostly a function of engine speed and load. However, as most modern gasoline engines are equipped with individually controllable exhaust and intake cam phasers, the overlap can be freely adjusted as a function of operating conditions allowing for the desired amount of EGR. Greater reductions in throttling losses are seen, as more EGR is drawn back into the cylinder. However, the amount of EGR that the engine is capable of operating with is limited. Each engine will reach a level of EGR dilution at which misfiring becomes prevalent, a condition that needs to be avoided under all circumstances. Generally, high rates of charge motion can increase an engine's EGR tolerance and the amount of throttling losses that can be mitigated using this method.

External EGR can be achieved taking advantage of the positive pressure differential between the exhaust and the intake systems. Connecting the exhaust with the intake using a pipe and an EGR valve to meter the amount of EGR is common. Depending on the application, an additional EGR cooler can be installed to reduce the EGR-diluted intake charge temperature. This countermeasure reduces the thermal throttling caused by heating and expanding the intake gas and thus maximizes performance.

5.1.2 Downsizing

As SI engines exhibit the biggest efficiency losses during part-load operation and most engines are operated under low part-load conditions for a large portion of their entire operating regime, one approach is to reduce engine displacement to raise engine load and thus improve engine efficiency. However, for a naturally aspirated (NA) engine, that would result in lower engine power and torque, which is unacceptable. Therefore, supercharging or turbocharging to make up for that loss in performance is applied. This process is called *downsizing*, where the engine displacement is reduced, in some cases, by approximately 50%, compared to a NA engine while increasing or at least maintaining engine performance of the NA engine. Appropriate downsizing, also often referred to as *rightsizing*, can result in fuel consumption reductions approaching 15–20% depending on the applicable drive cycle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

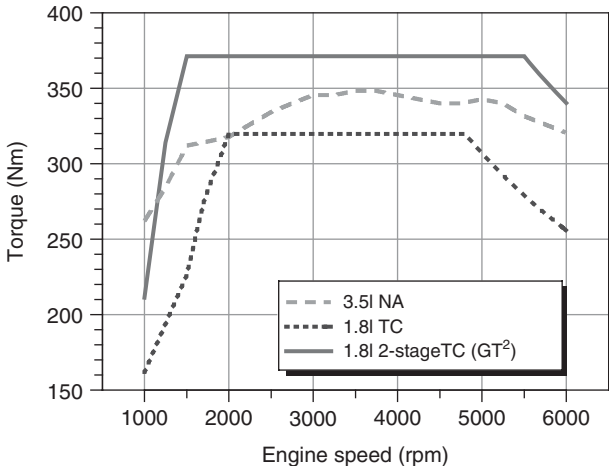


Figure 9. Performance comparison of a large naturally aspirated engine to that of two smaller turbocharged engines. (Reproduced by permission of FEV Inc.)

Meanwhile, turbocharged, direct-injected (DI) engines are able to achieve brake mean effective pressure (BMEP) levels above 20 bar currently with levels approaching 30 bar in the near future. This allows replacing 3-L class six-cylinder NA engines with 2-L class four-cylinder engines. To further optimize the entire full-load curve, two-stage turbocharging systems with intercooling are under development. These engines can provide excellent transient response behavior, high low-end torque and high rated power and torque over a wide range of engine speed for superior vehicle performance. In the near future, state-of-the-art two-stage turbocharged intercooled engines will be capable of specific power ratings of 100–120 kW/L and specific torque ratings of 180–200 Nm/L. Figure 9 depicts a downsizing example starting with a 3.5-L NA V6 engine compared to a 1.8-L engine with a single turbocharger and two-stage turbocharging.

Combinations of supercharger and turbocharger are possible and have already successfully been introduced into the market. This combination can provide superior transient response at low engine speeds because of the direct drive of the supercharger and may prove to be a good combination for large vehicles with downsized engines.

5.1.3 Lean burn concepts

5.1.3.1 Gasoline. In contrast to the conventional stoichiometric operating mode of the gasoline engine, the application of a lean-burn concept offers additional gains in engine efficiency because of de-throttling of the engine and increased peak pressures offering improved cycle efficiency. In general, distinction is made between lean-burn homogenous and stratified engines. While lean burn, in general, indicates an over-stoichiometric ($\lambda > 1$) engine operation resulting in excess oxygen in the exhaust, the stratified combustion system employs a local enrichment around the spark plug at ignition, whereas the homogenous lean-burn engine is using a globally constant in-cylinder $\lambda > 1$. The homogenous lean-burn engine can use either port fuel injection (PFI) or a direct injection (side or central DI) system. The stratified system, however, can only be realized using a centrally located injector close to the spark plug that develops a locally rich air–fuel ratio for ignition while the remaining charge throughout the remainder of the cylinder is significantly leaner.

Despite the significant efficiency advantage of a lean-burn concept, the change in emissions treatment is substantial, as the standard TWC commonly used to treat gasoline engine emissions cannot be applied. Although the engine-out hydrocarbon (HC) and carbon monoxide (CO) emissions are lower based on the excess oxygen in the exhaust, the amount of NO_x is increased. In addition, a typical three-way exhaust catalyst is relatively efficient at oxidizing HC and CO emissions with lean mixtures, whereas the NO_x reduction efficiency is nearly zero when operating lean. Therefore, for stratified engines, a separate NO_x reduction aftertreatment device is essential. Over the past years, two systems, the selective catalytic reduction (SCR) system using urea as a reductant and the lean NO_x trap (LNT), also often referred to as *NO_x absorber catalyst* (*NAC*), have been developed and successfully applied. Both systems have already been successfully used in diesel engines, whereas the LNT has been used in some lean-burn gasoline production engines.

Although the emissions can be reduced to the required legislated levels, the issues of increased cost and overall efficiency gain are subject to further refinements, as the LNT needs to be kept at operating temperature (above "light-off") to ensure proper reduction of the NO_x. This requires the engine management system to artificially raise exhaust gas temperature by inducing inefficiencies such as spark retardation or early exhaust cam phasing. This effect is more pronounced with lower base exhaust gas temperature levels. Therefore, catalyst heating is required more often for less-aggressive drive cycles, diminishing the advantage of the lean-burn concept. In addition, the LNT technology requires extremely low fuel sulfur content, as fuel sulfur in form of sulfates occupies the receptive catalyst sites in the LNT; therefore, rendering it useless over time. As a result, desulfation is required to remove the sulfur from the LNT during a rich and high exhaust gas temperature operating period. Again, this operating point utilizes fuel energy to clean the LNT further reducing the efficiency

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of the concept. Reduced sulfur fuel may provide a future solution for lean-burn concepts.

Much like current diesel engines, SCR systems might find their way into the market place for gasoline engines as well. However, it is not known if the requirement of the driver to periodically refill the reductant tank with separate fluid will achieve widespread market acceptance in all markets.

Although it is desired to operate the engine as lean as possible for efficiency reasons and the side effect that with increasing excess air, beyond λ of 1.2, the NO_x engine-out emissions start to decline drastically, misfiring begins to limit the capabilities of the combustion system as the lean threshold is approached. Increased charge motion and higher ignition energy can help push the threshold but engine operation above relative air–fuel ratios of 1.5 remains challenging.

Another combustion system modification required because of lean engine operation is related to the ignition system. Owing to the increased in-cylinder gas pressure levels (especially in case of turbo- and supercharged engines), the energy level at the spark plug is higher to ensure proper spark and thus combustion initiation. This necessitates a different, high energy system and special spark plug materials (e.g., iridium electrodes) to minimize electrode erosion because of the increased spark energy.

5.1.3.2 Compressed natural gas. Compressed natural gas (CNG) due to its global availability in large quantities and its excellent fuel properties for SI engines represents a beneficial alternative to gasoline. Especially, owing to its high knock resistance and the superior mixture formation, CNG is gaining importance. CNG is also a very suitable fuel for lean-burn concepts. As most CNG automotive applications use port injection systems, only homogenous lean-burn concepts can be found in the market place. However, it can be expected that future DI CNG injectors will also allow for stratified lean-burn configurations further pushing the lean operating limit and overall engine efficiency.

5.1.4 Cylinder deactivation

Reducing the number of active cylinders when an engine is at part load allows for the remaining cylinders to provide the same amount of power, but at a reduced level of throttle. Deactivation of the cylinders is accomplished in both overhead cam and pushrod engines by allowing lost motion in the valvetrain. Either collapsing finger followers for overhead camshaft engines (OHC) or collapsing tappets in overhead valve engines (OHVs) are activated and deactivated with pressurized engine oil to control the actuation of the valves. Controlling cylinder usage by means of valvetrain actuation allows for the deactivation of the cylinders without the pumping losses that would remain if deactivation were to be enabled simply by fuel cutoff.

Considerations for engine balance tend to lead to deactivation of half of the total engine cylinder count. As such, the load requested from the engine must be less than half of the total engine output and some additional coverage to allow for seamless transition from deactivated to fully operational cylinder counts in order for the engine to operate with deactivated cylinders. Engine vibration at idle with cylinders deactivated typically exceeds the comfort level of most customers in developed markets. Thus, full engine operation is typically required at idle, which further limits the utilization range of cylinder deactivation. As the operating window is narrowed by these considerations, the effectiveness of cylinder deactivation on engines with smaller cylinder count becomes smaller making cylinder deactivation a better option on V8 and V6 engines than on I-4 configurations.

5.1.5 Fully variable valve lift and duration

A significant amount of development effort has been invested in the development of fully variable inlet valve lift and duration known as *variable valve lift* (*VVL*). There are electromechanical, hydraulic, kinematic, and pneumatic systems being developed or on the market. There are several advantages to having real time control over intake valve motion. In a conventional, mechanical camshaft valvetrain, inlet valve lift and duration are fixed by the profile of the camshaft. This profile is a compromise between the desired events at high and low engine speeds and high and low engine loads. At high engine speeds and loads, it is desired to keep the intake valve open longer and at higher lift for more complete cylinder filling. This is contrary to the short duration and low lifts desired at idle.

With some fully VVL systems, the intake valve movement is not limited to a single event per combustion cycle. Instead, a second lift event can be included to increase in-cylinder turbulence and enhance fuel mixing and combustion speed or increase the amount of internal EGR.

Another use of VVL is to reduce or eliminate the intake air throttle and reduce the throttling losses of the engine. As discussed in Section 3.1, work is performed by the piston at low loads, as it pulls intake charge through the open inlet valve. The difference in pressure between the throttled intake manifold and positive crankcase pressure generates the pumping loop shown in Figure 3. The piston conducts the work throughout the intake stroke. With a VVL system, the valve can be opened for a short duration at the beginning of the intake stroke allowing the pumping loop to resemble

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the engine at full load. Once the appropriate amount of mass is induced, the intake valve is closed and the piston begins to perform negative work, as it expands the gas during the final travel to BDC. As such, the average cylinder pressure throughout the intake stroke is higher than during for a conventionally throttled engine.

5.2 Combustion efficiency

5.2.1 Compression ratio

As shown in Equation 1, the maximum potential efficiency of an SI engine is limited by the compression ratio. Therefore, it is important to maximize the effective compression ratio of the engine in order to achieve the highest possible engine efficiency. This section outlines some key parameters to consider when attempting to maximize an engine's compression ratio.

5.2.1.1 Combustion chamber design. Besides some three-valve and five-valve exceptions, the majority of current gasoline engines are four-valve engines with a central spark plug in a pent-roof type combustion chamber arrangement. With two intake valves on one side and two exhaust valves on the other side, a thermal gradient across the cylinder head is inevitable. Therefore, in order to maximize the compression ratio of the engine, cooling of the combustion chamber must be considered during the design phase of the engine to avoid hot spots in the combustion chamber and mitigate knocking.

The central spark plug location is also beneficial for maximizing the engine's compression ratio. The central location allows for basically identical flame propagation distance in each direction fostering a stable combustion. As the flame travels through the combustion chamber, the charge mixture that is farthest from the spark plug, known as *end gas*, absorbs heat from the combustion process and is exposed to higher pressure in the combustion chamber. As this gas absorbs heat and is compressed, it becomes more susceptible to self-ignition. Shortening the time it takes for the flame front to reach the end gas reduces this risk and allows the compression ratio to be increased relative to noncentralized spark plug engines. Regardless of plug location, it is of utmost importance to provide sufficient cooling to the spark plug to avoid overheating of the plug and providing a preignition source.

5.2.1.2 Exhaust system. Besides the combustion chamber geometry, the intake and exhaust port geometries play an important role. Focusing on scavenging of the combustion chamber, some newer integrated exhaust manifold designs combine the cylinder head and exhaust manifold and provide reduced resistance in the exhaust port. By doing so, reduced levels of residual gas are retained in the cylinder during full-load operation thereby reducing the overall charge temperatures and risk of engine knocking.

Another option for the exhaust system is to utilize long primary exhaust runners. Long runners allow for the cylinder exhaust process to occur without interference from pressure waves of the adjacent cylinder. The high pressure waves from an adjacent cylinder beginning the exhaust blowdown process can force hot exhaust gas back into a cylinder that is trapped by a closing exhaust valve. This process increases the charge temperature for the next combustion cycle requiring a lower compression ratio.

5.2.1.3 Intake system. On the intake side, the port geometry is crucial in terms of charge motion generation allowing for stable combustion over the entire engine operating map. Designing the port for optimized tumble motion and high in-cylinder flow velocities, resulting in fast combustion and high EGR tolerance while in combination with the piston top design preserving the charge motion until ignition occurs, is the prime focus of each combustion system development effort. The increased combustion speeds created by high charge motion reduce the amount of time available for end gas heating and allow for higher compression ratios.

5.2.1.4 Variable compression ratio systems. The engine's compression ratio is an important parameter defining the efficiency, performance, NVH behavior, and engine-out emissions. While under cold start and part-load operations, a high compression ratio is desired. However, high compression ratios limit the full-load performance, as it promotes knocking at higher loads requiring enrichment of the air–fuel ratio ($\lambda < 0.9$) and ultimately leading to severe efficiency losses. Therefore, the compression ratio for each application needs to be selected very carefully because of these compromises.

As the compression ratio always represents the best acceptable compromise between efficiency and performance, it is even worse considering flex-fuel applications such as E85 or CNG. In that case, despite the fact that a much higher compression ratio could be employed delivering much higher engine efficiencies, the engine will be operated at the same low compression ratio as designed for operation with gasoline only. Thus, a conventional fixed compression ratio engine cannot take advantage of the significantly lower knock sensitivity of E85 or CNG.

Considering upcoming CO_2 emission limits worldwide and fuel consumption regulations, one major technology considered for future applications are mechanisms capable of varying the engine's compression ratio. This technology,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

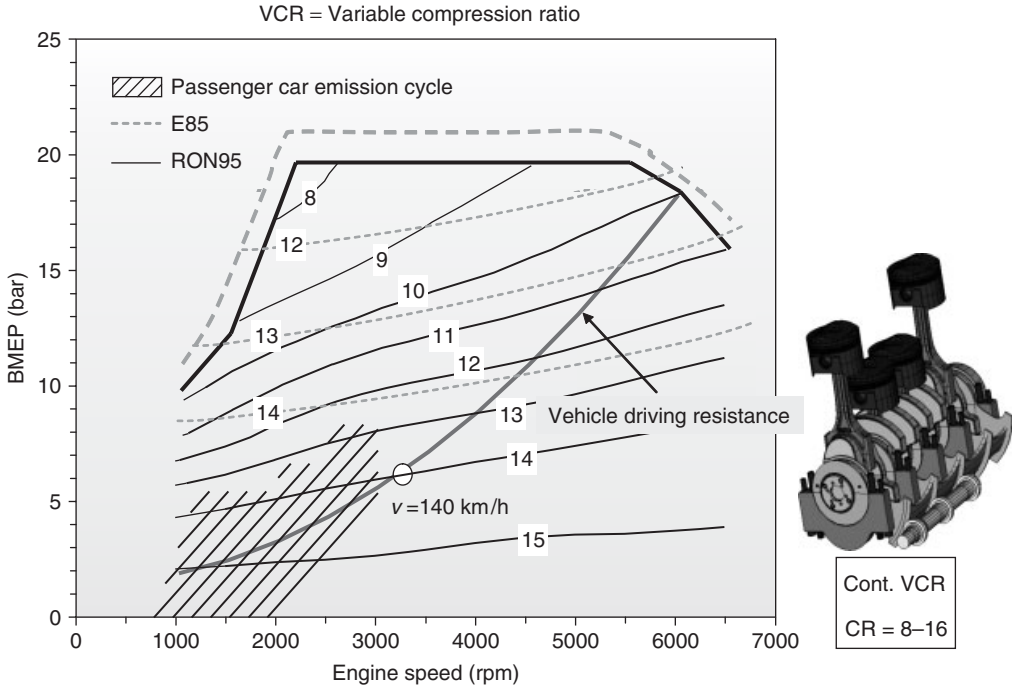


Figure 10. Ideal compression ratio based on engine load and speed illustrating the benefits of a variable compression ratio mechanism. (Reproduced by permission of FEV Inc.)

in combination with downsizing, allows for fuel consumption reductions up to 25% compared to a NA engine configuration of identical performance. In addition, it offers improvements to cold start behavior, cold start emissions, reduced or eliminated enrichment during high load, and more optimal operation on alternative fuels. An illustration of a VCR system and associated capability is shown in Figure 10.

In addition, by dynamically optimizing the compression ratio during engine operation, the EGR tolerance and thus emissions and fuel consumption can be further improved during part- and full-load conditions.

5.2.1.5 In-cylinder temperatures. Advanced cooling system technologies such as electrified thermostats and electric water pumps offer further potential for reduction of fuel consumption. By allowing the cylinder liner to run hotter during part-load operation, but increasing cooling capacity during high load conditions, it is possible to increase the compression ratio beyond that of a fixed, mechanical cooling system.

Another possibility to mitigate knocking and increase the engine's compression ratio is spray bore liner technology. Here, the cast iron liners used in most aluminum cylinder blocks are removed and replaced with a thin coating that provides better heat transfer between the in-cylinder gases and the cooling jacket. This ultimately reduces the heat-up of the fresh intake charge during the intake stroke and also improves heat removal during the compression stroke. Both effects assist in mitigating knock and allow for increased compression ratios.

5.2.1.6 Fuels. The fuel used in an SI engine is probably the easiest method for achieving higher compression ratios. Higher octane-rated fuels are less prone to autoignition and can tolerate higher temperatures associated with higher compression ratios. The issue, however, is market acceptance of purchasing higher cost fuels for the vehicle. Therefore, countermeasures must be made to protect the engine should an insufficient octane fuel be used by the end consumer. In case the octane rating of the fuel is insufficient, modern engines are equipped with a knock sensor as a part of the engine management system that will retard spark timing to mitigate knocking.

6 CI ENGINE-SPECIFIC ENABLERS

In comparison to the gasoline engine, the diesel engine due to its lean operating mode, higher compression ratio, and the unthrottled operation has major advantages making it the most efficient combustion engine available on the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

market. On the basis of these advantages, it can be found in applications ranging from small passenger cars up to large bore engines for ships, power generation, and other industrial applications. However, in contrast to gasoline engines, cleaning the exhaust of a diesel engine is significantly more challenging. Thus, the majority of the effort in diesel engines is the development of the combustion system to produce fewer engine-out emissions and the development of aftertreatment systems that minimize their impact on engine efficiency.

6.1 Efficiency

6.1.1 Combustion system

Each combustion system applied in a certain market must be engineered to comply with the corresponding regulatory agencies. Differences in fuel properties such as biodiesel and sulfur content and cetane number can have a huge impact on how a combustion system is configured and optimized. Therefore, implementation of an in-cylinder pressure sensor in combination with a glow plug can improve the adjustment to different fuel properties and thus help optimize the efficiency of the engine. Diesel fuel sulfur content is low in Western Europe and North America (<10 ppm in both cases), whereas other markets, especially in developing countries, exhibit significantly higher sulfur concentrations requiring hardware and calibration changes. Meeting future, more stringent, emission standards will require ultra-low sulfur diesel (ULSD) fuel worldwide in order to not jeopardize the sulfur-sensitive exhaust aftertreatment systems. In contrast to sulfur content, the cetane numbers between Europe and North America are very different. With the European norm EN590 requiring a cetane number of at least 51, the US market 2D fuel requirement is between 40 and 50 according to the EPA (Environmental Protection Agency). In addition, California requires a range between 47 and 55. In order to avoid excessive white smoke emissions (unburned HCs) during engine start and simultaneously improve cold start capability, the compression ratio as a part of the combustion system must be chosen very carefully. While a high compression ratio is desired for good starting behavior defined by low emissions, fast start-up, and low noise, the compression ratio is desired to be lower at higher loads and speeds to limit pressure and production of NO_x emissions.

6.1.1.1 Compression ratio. Although a higher compression ratio is usually beneficial with respect to engine efficiency, it results in high in-cylinder peak pressures and undesired high NO_x formation at higher engine speeds and loads, especially for high output downsized engines. Therefore, the right compromise, considering all aspects in terms of performance, emissions, fuel consumption, NVH, and fuel properties, is crucial. Current development efforts are also looking at variable compression ratio mechanisms, which allow for the optimal compression ratio to be selected as a function of engine speed and load. With such a mechanism in place, all the trade-offs described earlier could be eliminated.

6.1.1.2 Intake system. Minimizing the pressure loss in the intake system in order to maximize the air throughput, it is desired to provide the highest amount of air to fully oxidize (combust) the fuel injected on a per-cycle basis and minimize pumping work. However, the charge motion generated in the intake port is as important as the quantity of air provided. The design of the usually separated intake ports in case of a four-valve engine configuration is usually split into a swirl port and a filling port to achieve the best compromise. Charge motion is required to assure sufficient mixing inside the cylinder before and during combustion to optimize combustion efficiency and minimize emissions. The overall swirl level has been slightly reduced over the past years, as the higher injection pressures available in modern diesel fuel systems promote improved mixture preparation. Lower required swirl levels allow for higher flow ports that improve volumetric efficiency and power density and reduce pumping losses in comparison to more restrictive high swirl ports.

To further enhance the charge motion with minimal flow penalty is to incorporate an offset chamfer at the intake valve seat. This offset chamfer, often referred to as a *swirl chamfer*, promotes swirl at low valve lifts during the initial opening and especially closing of the intake valve. This has proven to be a very effective measure to improve mixture preparation, combustion, and thus fuel efficiency. In combination with a VVL system, this approach can result in not only significantly improved combustion efficiency but also lower emissions.

6.1.1.3 Fuel spray and piston crown. The piston crown geometry and injector spray pattern have a major influence on the combustion efficiency and emissions performance of the engine. The interaction between the fuel spray and the combustion bowl in the piston need to be developed simultaneously to achieve the best possible combination. Such systems are typically developed analytically during the early stages of engine development but test hardware must be used to confirm the final selections.

6.1.1.4 Exhaust system. To further improve the overall efficiency, the backpressure generated by the typical

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

exhaust components such as piping and muffler in combination with the required aftertreatment components such as diesel particulate filter (DPF) and LNT or SCR needs to be minimized to avoid unnecessary pumping work.

6.1.2 Aftertreatment system effects

Meeting regulated emission standards worldwide has become a challenging task for automotive diesel engines because of the lean operating mode and the particulate matter (PM) contained in the exhaust emitted. However, the applied aftertreatment system technology can require additional engine operation modes, increase the total backpressure of the exhaust system, and require additional fuel injection events. All can increase the total amount of fuel consumed.

With the initial introduction of diesel oxidation catalysts (DOCs) in combination with DPFs in the European and the US markets exhibiting PM filtration efficiencies of 95% and higher, the PM emissions can be contained within the regulated limits. Further reductions can be expected to be achieved based on continuous improvements in DPF technology related to parameters such as materials (ceramic vs metal based), porosity, and pore size. As the backpressure and subsequently also the fuel consumption increase due to PM trapped in the DPF, the ECU will enforce periodic filter regeneration. During that process, the engine management system demands a secondary injection inside the cylinder later in the expansion cycle. By generating an exotherm across the DOC to provide a sufficiently high exhaust gas temperature level to oxidize the PM trapped in the DPF, the backpressure will (in case of a successful regeneration) return back to the original value. Depending on the regeneration strategy, the entire exhaust system layout in terms of heat losses (insulation) and location, potential catalytic coating inside the DPF, the consistency of the trapped PM, and the potential use of a fuel-borne catalyst to reduce the soot ignition temperature define the loss in efficiency due to the regeneration. Current development efforts continuously improve the efficiency of the DPF system focusing on areas such as active and passive thermal managements as well as DPF materials and substrates. Further losses that need to be accounted for are based on potential oil dilution caused by the late (post) in-cylinder injection to provide sufficient HC molecules needed to generate the exotherm across the DPF. On the basis of the nozzle layout, the rail pressure applied, and the injection timing, it is of utmost importance to minimize oil dilution to the highest possible extent.

In addition to the PM emissions, the stringent NO_x emissions pose another challenge to vehicle manufacturers. In order to meet these limits, two different systems have been established in the marketplace, SCR catalyst and LNT. Requiring a reductant in the form of liquid urea, which is injected into the exhaust stream upstream of the SCR catalyst, this system can achieve NO_x conversion ratios well above 90%. However, a separate tank system aboard is required to provide the urea in its demanded quantities. By comparison, the LNT system does not rely on another liquid but utilizes the diesel fuel already aboard to regenerate the LNT by temporarily enriching the exhaust. With the LNT directly using diesel fuel to regenerate, the LNT system exhibits a loss in efficiency. In comparison, the SCR catalyst uses a separate fluid to allow meeting the NO_x standards and does not require additional fuel heating for regeneration. In both cases, the efficiency is reduced because of the increased backpressure caused by the flow-through catalyst substrate and the fuel/reductant necessary to reduce NO_x to required levels. Owing to the overall stringent NO_x emission targets, a further advanced beginning of injection (BOI) to gain efficiency can unfortunately not be realized as BOI retardation (resulting in deteriorated efficiency), application of EGR, and NOX aftertreatment means must be combined in order to achieve NO_x tailpipe emission levels that are safely below the mandated limits.

FURTHER READING

Bauer, H. (ed.) (2011 Print) *Bosch Automotive Handbook*, 8th edn, Robert Bosch GmbH, Stuttgart.

Department of Energy, Fuel Economy. n.d. Web: October 10, 2012. http://www.fueleconomy.gov/.

National Highway Transportation Safety Administration, CAFE—Fuel Economy. n.d. Web: October 10, 2012. http://www.nhtsa.gov/fuel-economy.

Heywood, J.B. (1988 Print) *Internal Combustion Engine Fundamentals*, McGraw-Hill, Inc, New York, NY.

Taylor, C.F. (1985 Print) *The Internal-combustion Engine in Theory and Practice*, MIT, Cambridge, MA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto152
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

EV Auxiliaries

Gang Jin and Yuefeng Liao

BAIC Motor Technology Center, Beijing, China

1 Introduction 1
2 DCDC Converter 3
3 HVAC System 8
4 Electric Vacuum Booster Systems 14
5 Electric Power Steering Systems 17
6 Summary 18
Acknowledgment 18
Endnotes 18
References 19

1 INTRODUCTION

In a conventional vehicle equipped with an internal combustion (IC) engine, power generated by the engine is transferred to the various auxiliary systems through different mechanisms. As shown in Figure 1, the high power consumption, continuous operation devices such as alternator, hydraulic power-assisted steering (HPAS) pump, air conditioning (AC) compressor, and engine coolant pump are directly driven by the auxiliary belt off the engine crank shaft. The alternator converts the kinetic energy into electricity that powers various electrical loads and charges the 12V onboard battery. The vacuum incurred by the engine intake system[1] provides the assisted brake force through the brake booster. Finally, a portion of the residue thermal energy from the combustion process is transferred to the HVAC heating core through the engine coolant circulation system and heats up the passenger compartment.

For electric vehicles (EVs), the energy source is changed from liquid fuel to rechargeable battery and the power generating device from the IC engine to the electric motor. As the energy density of the battery is significantly lower than that of the liquid fuel, major design changes must be made in order to maximize the energy utilization efficiency to achieve commercially viable driving range and operation cost. The design shall also observe the current state of the art that (i) the electric motor has much higher energy conversion efficiency (around 85–95%) than that of the IC engine (around 20–30%); (ii) the transfer of energy, including routing and buffering, can be more efficiently conducted via the electrical path than the mechanical links (belts, hydraulic, or pneumatic circuits); and (iii) the precision control of the actuation dynamics, for example, power on demand and fast transient that help to conserve energy, can be more easily achieved by dedicated electrical actuators than direct mechanical links.

A typical architecture of the EV auxiliary system is shown in Figure 2. The traction battery pack provides the main source of energy to the auxiliary system in the form of high voltage electricity (typically 200–400 V). The high voltage is a prerequisite for driving efficiency of the traction motor and other high power consumption devices (typically over 1000 W) such as the AC compressor and the PTC (positive temperature coefficient) heater of the HVAC system.[2] The direct current to direct current (DCDC) converter converts the high voltage electricity to 12 V that powers the conventional electric power net. This enables the carryover of vast electric components designed for the conventional vehicles, for example, lighting, infotainment, and power convenience that typically consume less than 100 W of power and can be safely powered by the 12 V net. For devices that consume a few hundred watts of electric power,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

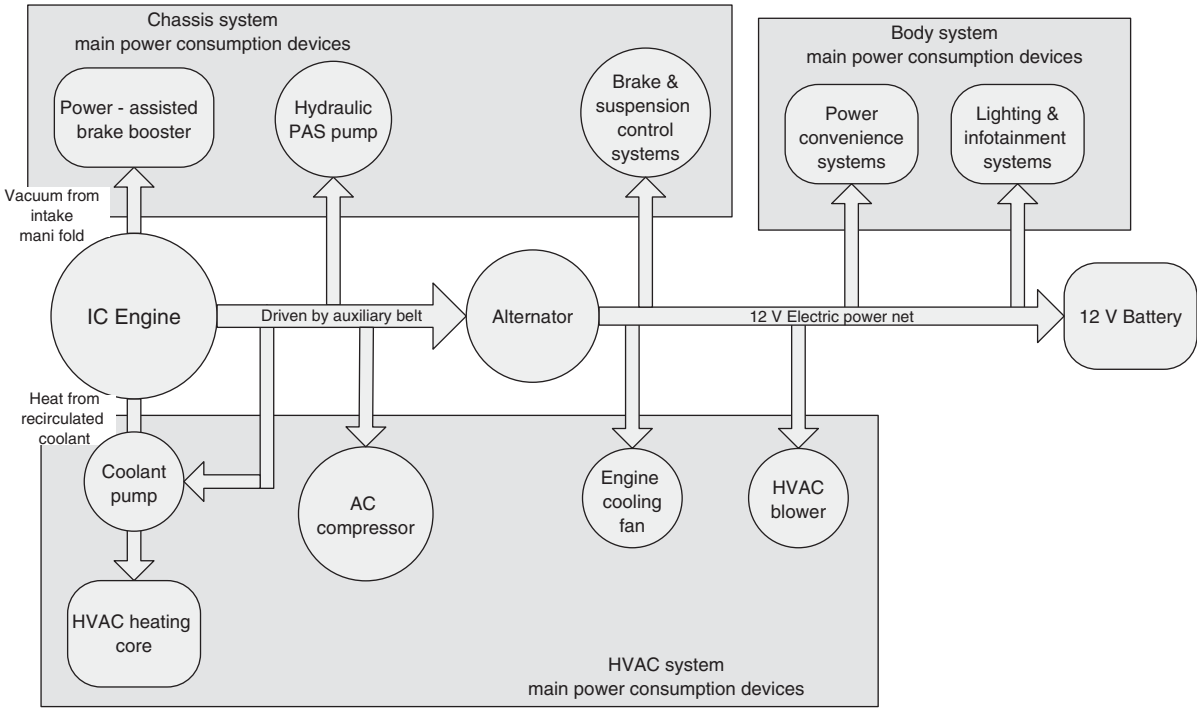


Figure 1. Conventional vehicle auxiliary systems.

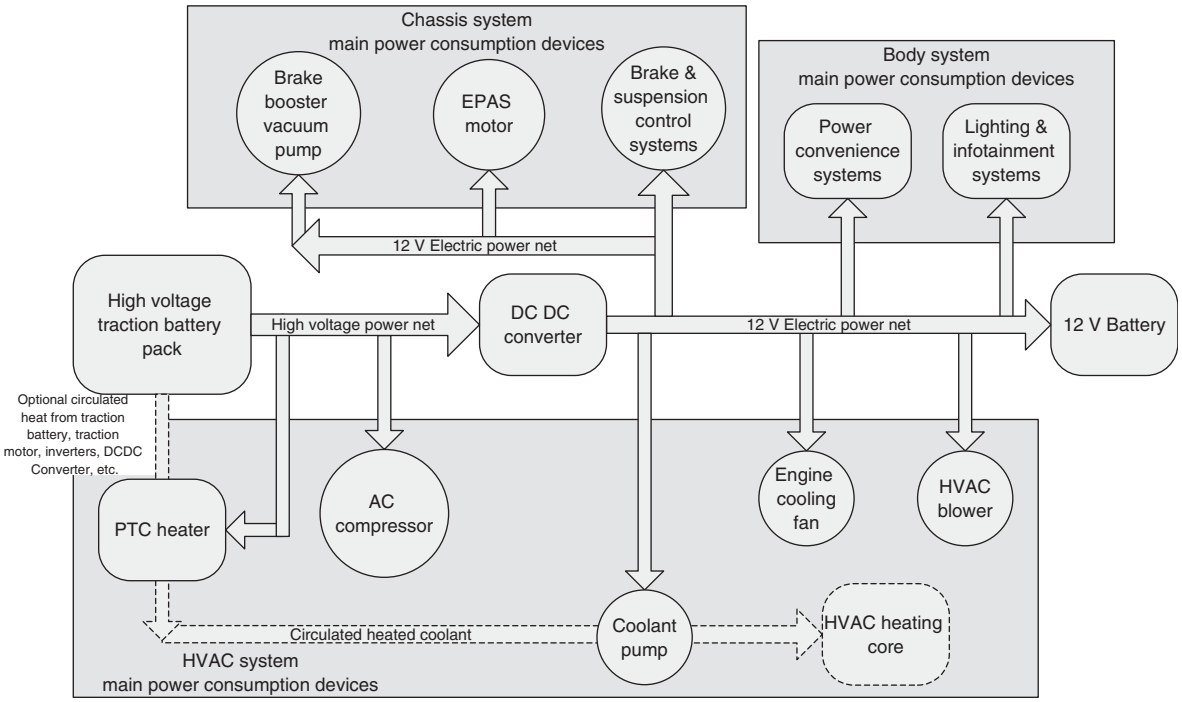


Figure 2. Electric vehicle auxiliary systems.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Estimated impact of auxiliary functions on EV driving range.

System	Function	Range Impact	Comments
HVAC	Heating	Up to 35%	Highly dependent on ambient conditions and vehicle settings
	Air conditioning	Up to 30%	
Chassis	Power steering	Up to 5%	Depending on use
	Power brake	A few percentage	
Body	Power convenience	A few percentage	Depending on use
	Lighting and infotainment	Up to 10%	

for example, engine cooling fan, HVAC blower, EPAS (electric power-assisted steering) motor, and brake booster vacuum pump, EV adaptation offers the opportunities to move these devices to the high voltage power net for better energy efficiency (Shiraki *et al*., 2012). This, however, is still largely under research investigation because of high voltage safety concerns and the currently achievable cost versus performance balance.

This chapter presents an overview of the EV auxiliaries with a focus on subsystems or components that are unique or have major changes over the conventional vehicles, namely the DCDC converter, the HVAC system, and the electric-driven vacuum booster and EPAS motors. As these devices consume (or conduct) a few hundred to a few thousands Watts of power, their impact to the EV driving range can be significant. An estimate is provided in Table 1. The theme of the system design shall then be centered on energy efficiency subjected to safety and cost constraints that align with the design of the total EV. This should be apparent from the subsequent discussions.

2 DCDC CONVERTER

The EV has a 12 V auxiliary battery to operate the lights, radio, and other equipments. The 12 V battery in a gasoline-powered vehicle is recharged with an alternator driven by the engine. In an EV, the auxiliary battery is recharged with the use of a DCDC converter with power from the high voltage battery pack. The conversion efficiency of today's state-of-the-art DCDC converter is above 90% with an output power rating typically around a few thousands Watts. This compares to about 70% for the alternator in a conventional vehicle with the same power rating. Considering the additional power loss to generate the kinetic energy to drive the alternator, the packaging constraint, and the additional weight of the alternator, it is apparent that the DCDC converter is a more efficient solution to power the 12 V net directly from the traction power net.

2.1 Typical DCDC converter topologies

Over the past few decades, many topologies have been developed for DCDC converters to suit the special needs of different applications (Gu, Lu and Qian, 2004). For EVs, the DCDC converter must satisfy a few baseline requirements. First, it must provide galvanic isolation from the high voltage source to the low voltage output for safety concerns. Second, it must have high conversion efficiency of at least 80% and preferably above 90% over a wide operation load range. Third, it must meet the stringent environmental, EMC, and reliability requirement for automotive applications.

The most commonly used DCDC converters for EV applications have the basic configuration as shown in Figure 3. After passing the high voltage filter to suppress high frequency noise, the DC current is converted to AC current by means of a switching circuits typically operating at tens to hundreds of kilohertz. This AC signal is down converted from high voltage to low voltage by a transformer with an $n:1$ ratio between the primary and the secondary windings. The down converted AC signal is then rectified and passed through the low voltage filter to feed the low voltage load. With proper design, the output voltage can be precisely controlled by the duty cycle of the switching circuits for a wide range of input voltage and output load. The transformer provides the required galvanic isolation between the high voltage input and the low voltage output. The power density and the conversion efficiency largely depend on the circuit design for the DCAC inverter and the ACDC rectifier and the selection of switching components and the transformer.

Depending largely on the design of the DCAC inverting circuits, different topologies for the DCDC converter have been proposed. Here, two types of topologies that have seen most applications in EV DCDC designs are briefly introduced. Figure 4 shows the configuration and key operation waveform of a *single end forward* DCDC converter.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

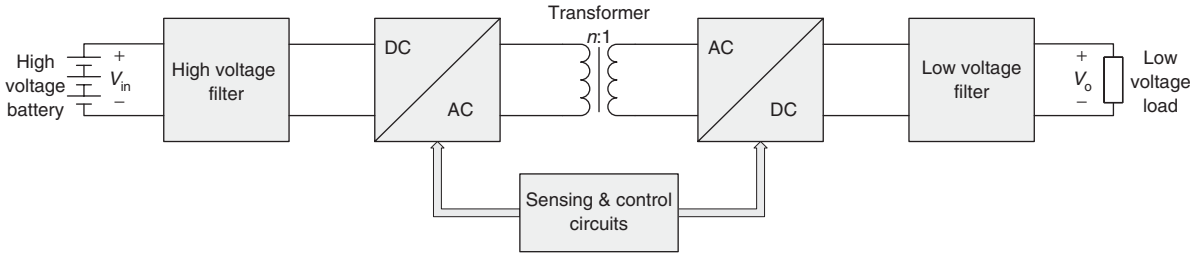


Figure 3. Basic DCDC converter configuration.

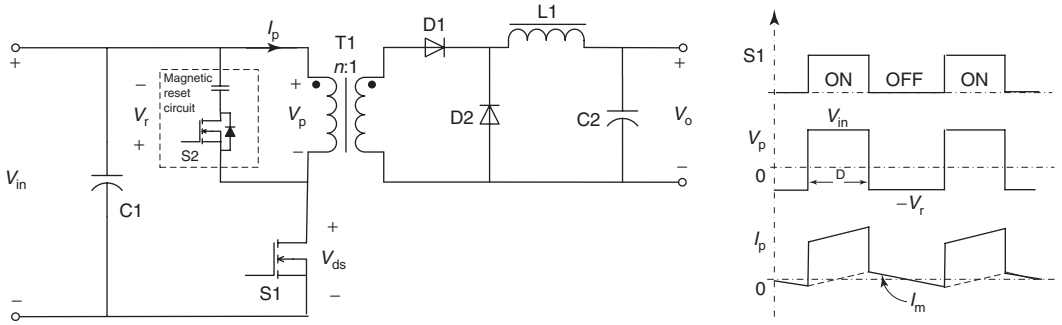


Figure 4. Single-ended active clamp forward DCDC converter.

In this type of converters, only one main switch S1 is employed on the primary side of the transformer. When the switch S1 is turned on, the transformer primary winding is connected to the input voltage V_{in} and energy is transferred from the input source to the output load through the transformer. When the switch S1 turns off, the transformer primary winding is connected to a magnetic reset circuit that applies a negative voltage $-V_r$ to the primary winding and resets the magnetizing current I_m. There are different ways to design the magnetizing reset circuits, for example, by an auxiliary winding or by an active clamping circuits as shown in the figure. In the latter design, the switch S2 turns on and off out of phase with S1 to provide a recirculation path for the magnetizing current. As the transformer primary winding is alternatively applied with V_{in} with duration of DT, and with $-V_r$ with duration of $(1-D)T$, where D represents the duty cycle and T the switching period, the following equation must be true in order to maintain a balanced current in the primary winding when the DCDC operates in stationary mode

$$V_{in}DT = V_r(1-D)T \tag{1}$$

This leads to

$$V_r = V_{in}\frac{D}{1-D} \tag{2}$$

It can be seen that when the duty cycle D approaches one, the clamping voltage V_r could be very large compared to V_{in}. For this reason, the main switch S1 must be designed with a sufficiently high blocking voltage ($>V_{in}+V_r$) for this type of topology or the control duty cycle D must be limited. As no power is transferred to the secondary side of the transformer when S1 is open and the current recirculates through S2, the power output of this type of converter is limited (typically up to a few kilowatts).

Now on the secondary winding side of the transformer, when the main switch S1 turns on, the diode D1 becomes positively biased and conducts the current through the low frequency passing filter circuit to the load. When S1 turns off, D1 becomes negatively biased and blocks the current in the secondary winding. As a result, D2 turns on to provide the recirculation path for the inductor L1 current. During stationary operation, the inductor L1 must maintain a balanced current flow that implies that

$$\left(\frac{V_{in}}{n} - V_o\right)DT = V_o(1-D)T \tag{3}$$

This gives to

$$V_o = \frac{V_{in}}{n}D \tag{4}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

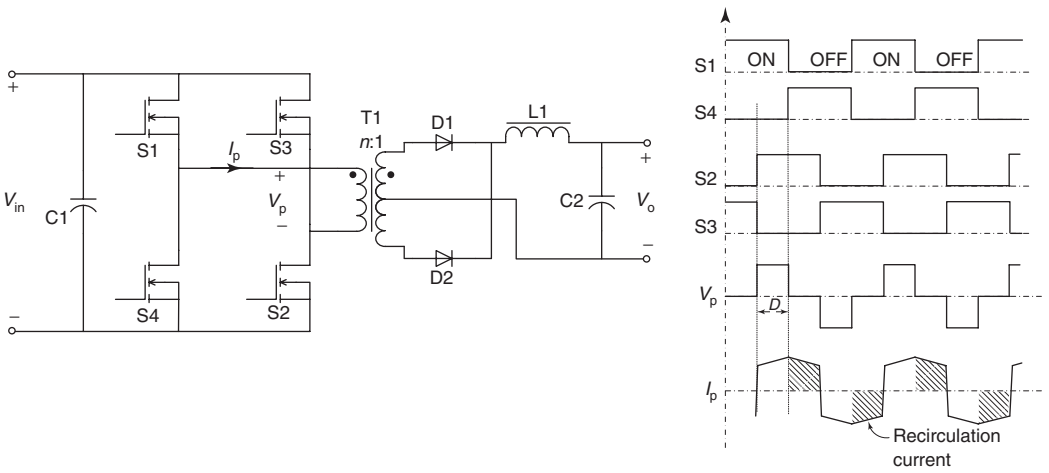


Figure 5. Phase-shifted full-bridge DCDC converter.

Thus, the output voltage can be changed by linearly adjusting duty cycle D based on the input voltage V_{in} and the primary over secondary winding ratio n.

For increased power output, the DCDC converter can be designed with a full bridge topology on the primary side inverting circuits.[3] The commonly used *phase shift full bridge* topology is shown in Figure 5. In this configuration, the primary winding is connected to the center point of the two half bridges formed by S1–S4 and S3–S2 respectively. When the transformer operates in the stationary state, there are four primary modes of operation: (1) when S1 and S2 are turned on, the primary winding is connected to the power input V_{in} and the primary current I_p is positively charged up; (2) when S1 turns off and S4 turns on, the primary winding is shorted (both sides connected to ground) and the primary current recirculates; (3) when S2 turns off and S3 turns on, the primary winding is *reversely* connected to power input V_{in} and the primary current I_p is negatively charged up; and (4) when S4 turns off and S1 turns on, the primary winding is shorted (both sides connected to power input) and the primary current recirculates. It should be noted that the minimum required blocking voltage for the four main switches is V_{in}, which is lower than that of the single-ended forward DCDC converter.

On the secondary winding side, the reflected current flows through diode D1 to the load during mode (1) and mode (2) and through diode D2 during mode (3) and mode (4). However, only during mode (1) and (3), power is transferred from the power input to the load, and thus the combined duration of mode (1) and (3) over the total switching period (all four modes) defines the effective duty cycle D. The input to output voltage map can be similarly derived as that of the single-ended forward DCDC converter; in fact, it takes the same form as that of Equation 4.

It is noteworthy that there are multiple ways to improve the efficiency of the DCDC converters by optimizing the circuit design on either side of the transformer. On the primary side, the most widely used technique is the so-called *zero voltage switching* (ZVS) where the main switches are turned on and off when the voltage across it passes through zero. This minimizes the switching loss and the electromagnetic emission, which is essential for high switching frequency operation. On the secondary side, low impedance active switches, for example, power MOSFET, can replace the passive diodes that turn on and off in synchronous with the main switches on the primary side to rectify the output current. This *active synchronous rectification* is especially effective when the regulated output voltage is relatively low compared to that of the diode voltage drop. Both techniques have seen wide applications in EV DCDC converter designs.

2.2 Key DCDC converter design parameters

To select a proper DCDC converter for an EV application, its key design parameters must be carefully examined. In this section, we review the parameter design based on a real-world application. The Beijing® E-150 is a pure electric compact sedan launched in 2012. The DCDC converter selected for this vehicle is a Dilong® DZEH2000S9 series product designed especially for EV applications. As shown in Figure 6, the DCDC converter is packaged in the engine compartment along with the traction motor, the inverter, the ACDC charger, and other auxiliary systems.

Refer to Table 2 for a list of the DCDC converter's key parameters and specifications. The first group of parameters

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

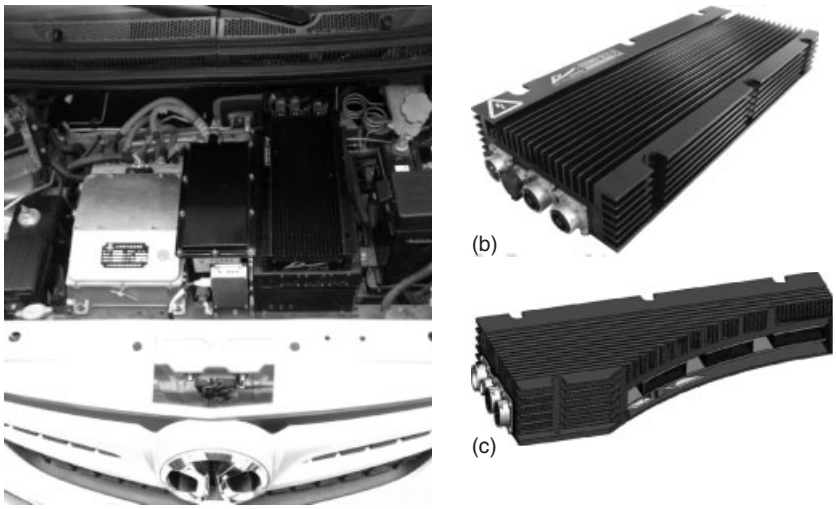


Figure 6. (a–c) DCDC converter for Beijing E-150 pure electric sedan.

Table 2. Dilong® DZEH2000S9 series DCDC converter key parameters and specifications.

Electrical Functional Performance			
Rated input voltage	Minimum: 220 V DC	Type: 360 V DC	Maximum: 410 V DC
Rated input current (fully loaded at rated output voltage)	9.8 A	6.1 A	5.37 A
Rated output voltage	14 V DC		
Rated output current	143 A		
Rated output power	2000 W		
Output current limit (in current regulation mode)	150–165 A		
Voltage regulation accuracy	≤±1% V	Input 220–410 V DC	Output 10–100% Load
Load regulation	≤±0.5% V	Input 360 V DC	Output 10–100% Load
Source regulation	≤±0.2% V	Input 220–410 V DC	Rated output load
Output ripple	≤±1% V	20 *M* Oscilloscope and twisted pair cable	
Efficiency	90%	Input 360 V DC	Output 14 V DC 140A
Switching frequency	200 ± 10% kHz		
Functional and Safety Protection			
Input under voltage protection	210 ± 10 V DC	Auto reset	
Input over voltage Protection	420 ± 10 V DC	Auto reset	
Output over voltage protection	17 ± 1 V DC	Manual reset	
Over temperature protection	95 ± 5 °C	Manual reset	
Input reverse polarity protection	Output inhibited		
Rated insulation voltage	Input to output	2000 V DC 1min	Leakage current ≤1 mA
	Input to shell	2000 V DC 1min	Leakage current ≤1 mA
	Output to shell	500 V DC 1min	Leakage current ≤0.5 mA
Insulation resistance	≥20 MΩ		
Environmental Requirement			
Ambient temperature	−40 to 70°C (operation)	−40 to 85°C (storage)	
Ambient humidity	5–85% RH (operation)	5–95% RH (storage)	
Cooling mechanism	Nature air convection		
Ingress protection rating	IP66 (IEC 529)		
Mechanical loading	Passenger vehicle body mounting (ISO 16750)		
Dimensions and Weight			
Volume (length × width × height)	390 × 164 × 52 mm		
Weight	≤4.5 kg		

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

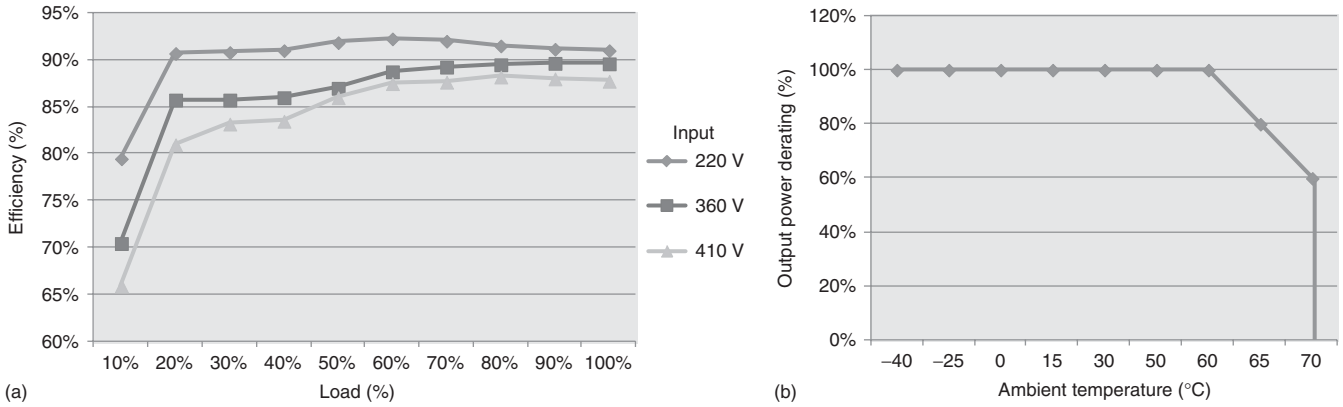


Figure 7. (a,b) DCDC converter efficiency and output power derating.

specifies its electrical performance. It supports a wide input voltage range from the traction battery (from 220 to 410 V) to ensure maximum achievable driving range. The output is precisely regulated at 14 V to feed the conventional low voltage power net. The rated output power is 2000 W with an efficiency rating around 90% under high load conditions. To achieve the aforementioned design performance (wide input voltage range, high output power, and high conversion efficiency), it employs the phase-shifted full-bridge topology as introduced in the previous section with ZVS technology. It should be noted that to ensure robust and reliable operation, the output power must be properly derated based on ambient temperature (Figure 7b). The ambient temperature directly aggravates the operation temperature of the heat-generating devices within the DCDC converter and hence under high ambient temperature condition, the output power shall be properly reduced to limit the internal rise of temperature. Furthermore, the conversion efficiency highly depends on input voltage and output power loading conditions (Figure 7a). With the increase of the input voltage and the decrease of the output power, the conversion efficiency can be significantly reduced.[4]

The DCDC converter must also be protected from inadvertent conditions such as abnormal input/output voltages and an internal temperature exceeding the range of safe operation. In addition, the input, output, and the DCDC converter enclosure must be electrically isolated for operation safety. This is measured by the leakage current when a high voltage potential is applied between input and output of the converter and by the insulation resistance.

Finally, the DCDC converter must be designed to operate properly in the environment where it is packaged inside the vehicle and the environment where the vehicle will be operated. This relates not only to the ambient temperature as mentioned earlier but also to the electromagnetic interferences,[5] mechanical loading (e.g., vibration and shock), foreign material ingression (e.g., water, dust, and chemicals), and other environmental conditions.[6] For under hood installation, a minimum ingression protection rating of IP66[7] is required. With a sealed enclosure, the DCDC converter is cooled by nature air convection. This limits its heat transfer efficiency and hence requires large area of heat sinks (Figure 6). Liquid cooling can be designed but with the added complexity of coolant circuit construction.

2.3 DCDC converter design trend

To meet the stringent requirements for EV applications, the design of DCDC converters continuously evolves. This includes the invention of new topologies to achieve multiple design objectives, for example, wide range of input voltage, minimum voltage stress on the switches, high effective switching duty cycle, and soft switching for the main switches to achieve maximum conversion efficiencies for a wide range of operation conditions (Zhang, Huang and Gu, 2002). Innovative products are engineered and commercialized that have unique internal architect to enable new features and functions, for example, multiple voltage outputs for multiple power nets, and bidirectional conversion for fuel cell EV (Su and Peng, 2005). In addition, to maximize the volume efficiency and simplify the high voltage wiring harness design, the DCDC converter and other power electronics, for example, the traction motor control inverter and the onboard household charger, are increasingly integrated together in an EV application. For example, the VW Golf EV boasts an integrated traction motor control inverter with DCDC converter installed in the engine compartment. The Mitsubishi i-MiEV has an integrated onboard charger with DCDC converter. The Beijing E-150 is also fitted with an integrated power module of onboard charger and the DCDC converter (Figure 8). The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

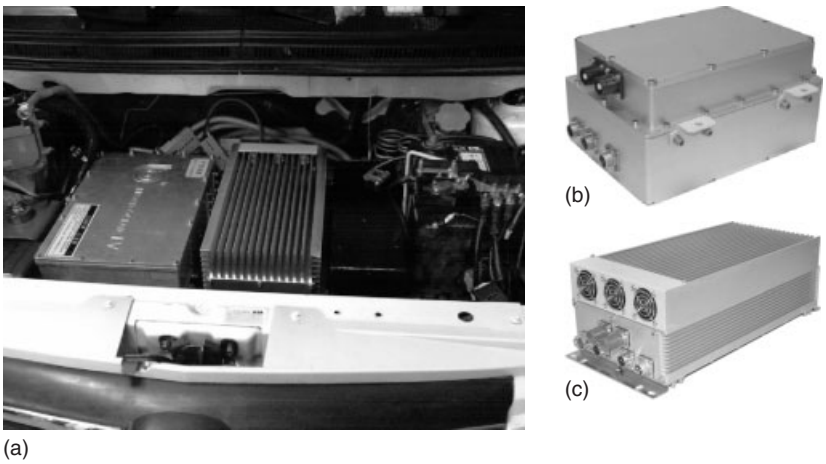


Figure 8. (a–c) DCDC converter with onboard charger for Beijing E-150 pure electric sedan.

product from Shinry® Technologies Corporation contains a 3.3-KW charger and a 1.5-KW DCDC converter in one housing (Figure 8a dimensions 430 × 220 × 148 mm) with nature air convection cooling. The series of products also support liquid cooling (Figure 8b, with dimensions 280 × 234 × 150 mm) and forced air cooling (Figure 8c, 485 × 200 × 146 mm) mechanisms for different EV applications. The liquid-cooled version requires significantly less packaging space compared to that of the nature or forced air convection versions. It also has a better protection for foreign object ingression with an IP rating of IP67. The topology for this DCDC converter is the single-ended active clamp forward type.

3 HVAC SYSTEM

The move toward battery-powered EVs poses several challenges when it comes to the heating, ventilation, and air conditioning (HVAC) system design. Safety regulations require all vehicles to have adequate heating and defrosting capacities. The heater/defroster system is easily operated in a conventional gasoline-powered vehicle using the waste heat from the engine combustion process. The EV is powered by an electric motor running on batteries that produces much less heat due to its high efficiency (roughly 90% for electric motors compared to 30% for IC engines). A dedicated heat source is often required to supplement the waste heat scavenged from the energy conversion devices (e.g., the traction motor, the traction battery, the DCAC inverter, and the DCDC converter) on the EV.

The AC system on a conventional vehicle consumes significant amount of power (around 15%) in very hot weather with the AC compressor being the largest power consumption device (around 50%) in the system. This will have a major impact to the driving range when the system is refitted in an EV. The first design change is to directly drive the AC compressor from the traction battery with a dedicated inverter often integrated within the motor assembly. This not only improves motor efficiency but also enables the compressor to produce externally controllable variable displacement that helps to improve system level performance.

Both the auxiliary heater and the electric AC compressor can cause significant reduction to the EV driving range (around 30%) in extreme weather and their design is crucial to the success of the initial conversion from conventional to EVs. Along with that, innovative concepts are being developed for EV applications to further minimize energy usage by HVAC system while maintaining a safe and comfortable driving environment. There are also unique requirements to properly condition the EV powertrain system for optimal performance and long-term reliability. All of the above forms an integral part of the HVAC system and its design is reviewed in this section.

3.1 The AC subsystem

The fundamental working principle of all AC systems is the same that constitutes the four basic processes in physics (compression, condensation, expansion, and evaporation) accomplished in four major mechanical components of the system (compressor, condenser, expansion valve, and evaporator) interconnected with tubes filled with the AC refrigerant (typically R134a). The compressor compresses cool, low pressure refrigerant gas, causing it to become hot, high pressure gas. This hot gas runs through the condenser where heat is extracted by (forced) convection of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

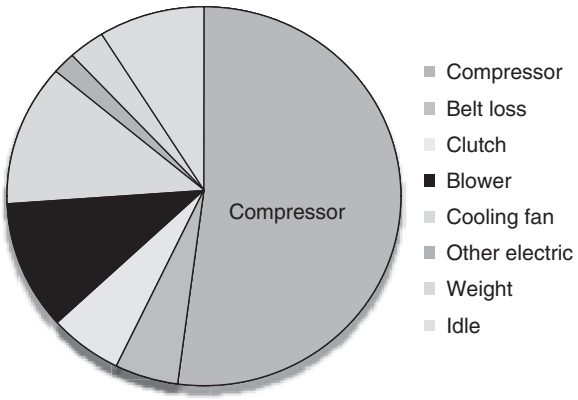


Figure 9. Energy usage in conventional AC system with fixed displacement compressor.

lower temperature ambient air and it condenses into liquid. The pressurized refrigerant flows through the expansion valve, which is basically a constricted orifice that reduces pressure of the liquid refrigerant so that it can expand and partially vaporize. The low pressure liquid/vapor mixture runs through the evaporator and completely vaporizes. During this process, heat is absorbed from the ambient air that is forced through the evaporator by the blower. The circulated low temperature air then cools down the passenger compartment. The cool, low pressure refrigerant gas is again compressed by the compressor and recirculates in the AC system.

The heart of the AC system is the AC compressor. It consumes the highest amount of power in the conventional AC system driven off the engine auxiliary belt (Figure 9). For this reason, significant effort has been devoted to the design and control of the compressor to maximize AC system efficiency. One of the most common improvements is to use an externally controlled variable displacement compressor.

In a conventional vehicle, the fixed displacement AC compressor is turned on and off based on several real-time parameters, among them are (i) the subjective request from the operator, (ii) the evaporator temperature, (iii) the compressed refrigerant pressure, (iv) the interior and exterior temperatures, and (v) the key engine operation status. As the turn-on transition of the compressor requires significantly more power than steady operation, this mode of control is less efficient compared to the variable displacement control (Figure 10). Another advantage of using variable displacement compressor is the ability to precisely control the surface temperature of the evaporator to keep it in the optimal region (typically between 2°C and 5°C). In Figure 11a, the evaporator temperature fluctuates without displacement control. This is compared to Figure 11b where the evaporator temperature remains constant with displacement control. Studies have shown that systems with externally controlled compressor can improve COP by 25% and fuel consumption by 30% in both city and highway drive modes.

The compressor suitable for EVs must have controllable rotary speed (variable cooling capacity), similar to the stepless variable frequency control of home AC. This corresponds to the externally controlled variable displacement compressor in a conventional vehicle but is driven by a dedicated inverter (that converts the high voltage DC to high voltage AC with variable frequency and amplitude) instead of the auxiliary belt. Owing to the difference in the driving mechanism, it can achieve higher efficiency, lower vibration and noise, and more accurate control (Hendricks, 2003). The electric compressor adjusts the cooling capacity in real time by changing the rotary speed and keeps the evaporator surface temperature at the desired level. Most commonly used control algorithm takes on the popular proportional, integrative, and derivative (PID) form with adjustment based on the large time delay of temperature change. Optimal control of the compressor to minimize energy usage is the core of the AC system design for EVs.

In the following, an industry-leading design, the Nanjing Aotecar® ATC-E26A electric compressor, is introduced to

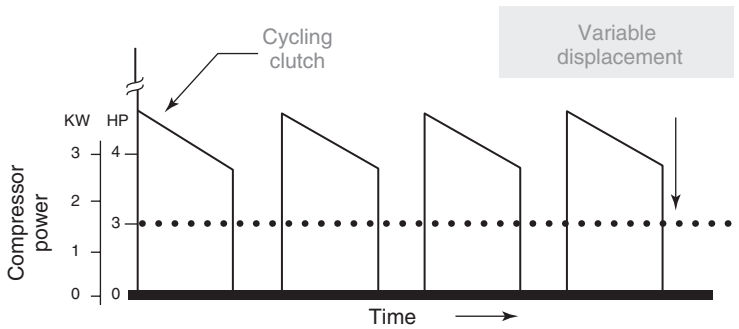


Figure 10. Comparison of power consumption between fixed and variable displacement compressor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

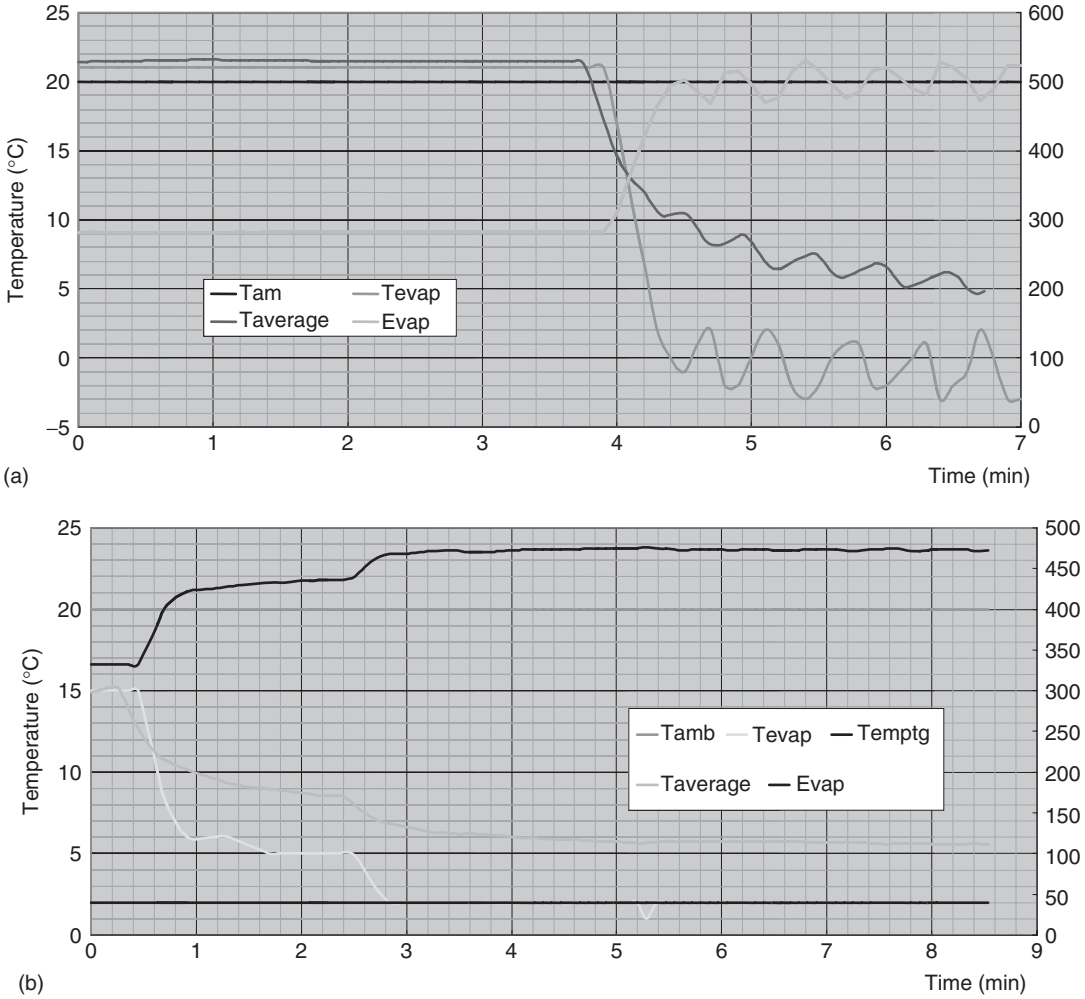


Figure 11. (a,b) Performance comparison between fixed and variable displacement compressor.

illustrate the key features and performance measures of an electric compressor. This device has been successfully integrated into the production Beijing E-150 pure EVs. Refer to Figure 12 for the internal structure of the assembly and Figure 13 for its external appearance and key performance curves. This device has the following main features and its key parameters are listed in Table 3.

(a) One piece design with integrated compressor, motor, and onboard inverter. The motor is directly cooled by the low temperature, low pressure-inhaled refrigerant. The inverter is cooled by its heat sink pressed against the cooler part of the compressor shell (close to the intake port).

(b) Under voltage protection. To protect the battery and the compressor motor, the inverter will automatically shut off the output when the power supply voltage is too low (less than 260 ± 5 V). The inverter will automatically reset if the supply voltage returns to normal (higher than 275 ± 5 V).

(c) Over current protection. When the input or the output current is too high, the inverter will automatically shut off the output to protect the inverter and the compressor motor.

(d) Variable speed control. Set the compressor rotary speed using a 400-Hz 12 V PWM signal by linearly adjusting the duty cycle (0–100% duty cycle corresponds to 1000–6500 RPM). The speed can also be controlled through a CAN interface.

3.2 The heating and defrosting subsystem

While IC engines generate a lot of waste heat, making it easy to warm the passenger compartment and clear the front windscreen frost in cold weather, EVs produce very

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

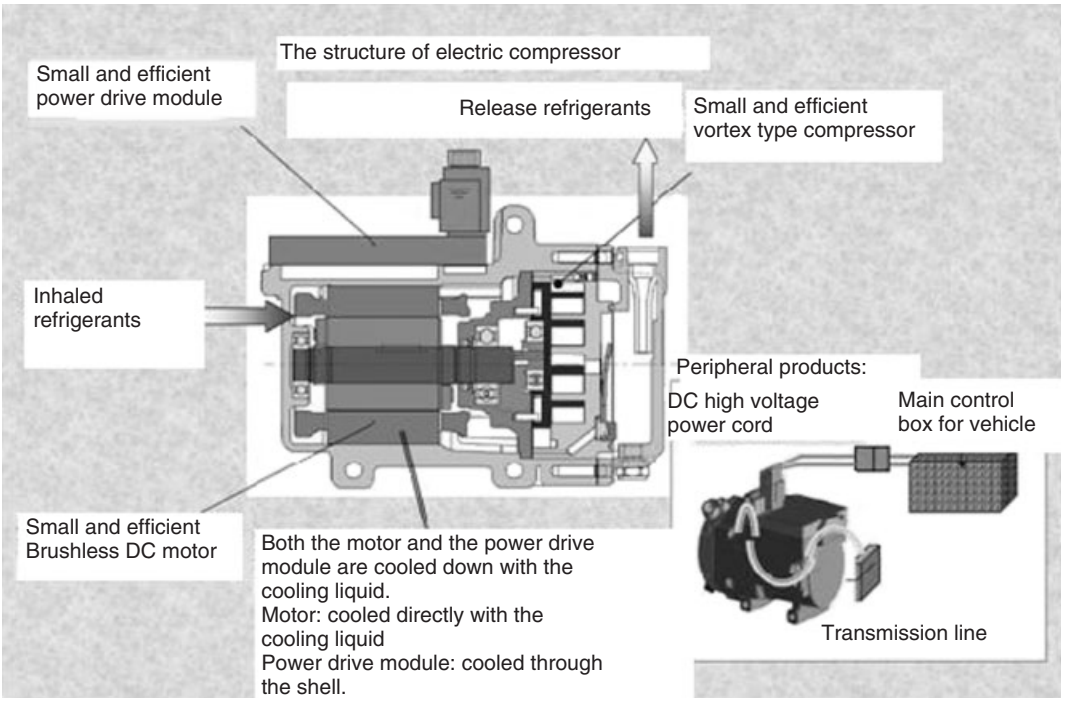


Figure 12. Electric AC compressor internal construction.

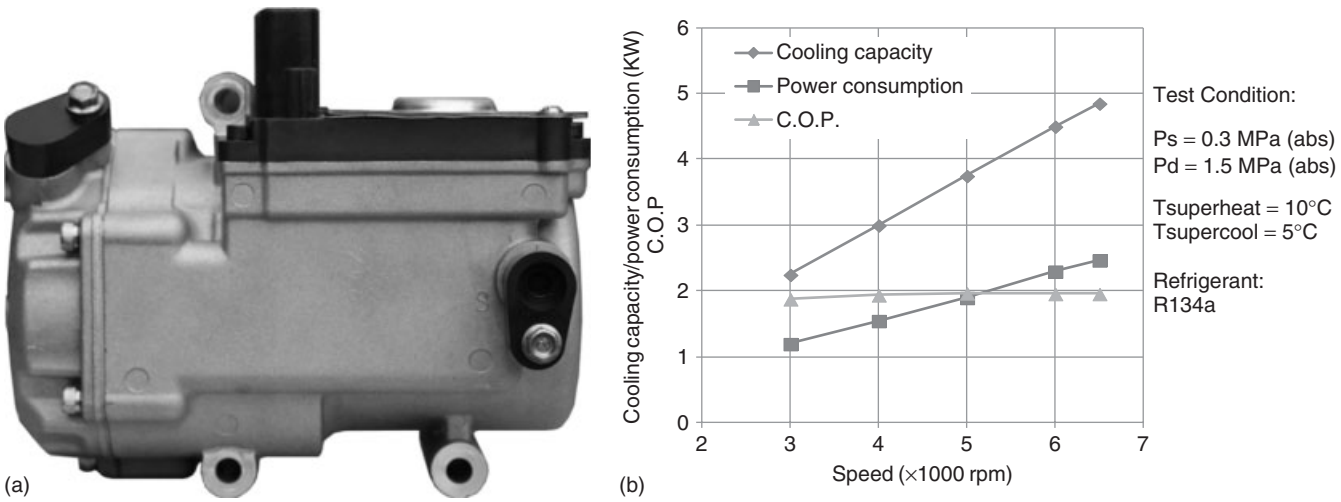


Figure 13. (a,b) Electric AC compressor with performance plot.

little excess heat. Furthermore, owing to material insulation constraints and semiconductor thermal characteristics, the electric motor and drive systems operates at a much lower temperature compared to the IC engine system. The IC engine operates optimally when the coolant temperature is around 90°C (normal operation range 85–100°C for gasoline engine and 80–90°C for diesel engine). For the traction motor, it is 60°C. As a result, it is often necessary to have additional (dedicated) heat sources in an EV.[8]

There are two main types of heater available on the market: the fuel burning heater and the PTC electric heater. Heaters using liquid fuel (e.g., gasoline, diesel, or ethanol) do not use electricity from the battery pack and will have no impact to the driving range of an EV. However, the emissions from the exhaust have a negative effect on the environment and open flame burning requires adequate protective measures for fire safety. The PTC electric heater uses a special type of ceramic that has positive thermal

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Aotecar® ATC-E26A electric compressor key parameters and specifications.

Inverter			
Rated input voltage	Minimum: 260 V DC	Type: 336 V DC	Maximum: 380 V DC
Rated input power	2437 W		
Control voltage range	9–15 V DC		
Output frequency range	100–375 Hz		
Maximum output current	25 A		
Overload tolerance	150% rated load		
Maximum control current	500 mA		
Insulation resistance (input to shell)	$\geq$50 MΩ		
Circuit protection	Under voltage, over voltage, over current, short circuit, loss of phase, and IGBT over temperature		
Operational ambient temperature	−30°C to 85°C		
Ingress protection rating	IP67 (IEC 529)		
Motor			
Type	Sensorless DC brushless motor (six Poles)		
Rated input voltage	336 V DC		
Rated power	2437 W		
Rated speed	6500 RPM		
Minimum speed	1000 RPM		
Speed error	<1%		
Operational ambient temperature	−30°C to 105°C		
Compressor			
Displacement	27 cc/rev		
Rated speed	6500 RPM		
Dimensions	208 ×121 × 176 mm		
Weight	4.5 kg		
Refrigerant	R134a		
Antifreeze lubricant	RL68H (120 mL)		
Cooling capacity	4875 W		

coefficient of resistance (i.e., resistance increases upon heating). This class of ceramics (often barium titanate and lead titanate composites) has a highly nonlinear thermal response such that it becomes extremely resistant above a composition-dependent threshold temperature. The rise in resistance is experienced within a fairly small temperature window of a few degrees centigrade. This attribute of the PTC results in a heating element that self-regulates to a preset temperature and automatically varies its wattage in order to maintain that preset temperature. The behavior causes the material to act as its own thermostat.

Compared to the fuel heater, the PTC heater has fast heating response times and plateau once the predefined reference temperature is reached. Because of these characteristics, it is possible to directly replace the conventional heater core by a PTC heater with a preset temperature threshold (i.e., the Curie temperature point) below the highest rated temperature of the surrounding plastic enclosures.[9] Currently, for this type of application, the rated power is limited between 2 and 3 KW for the PTC heater. This power level can meet the general requirements for heating and defrosting but is not ideal under low temperature (e.g., below −10°C) or for large vehicles. To increase the heating capacity, a PTC coolant heater with a higher power rating (greater than 4 KW) is often installed. Another added benefit of the coolant heating system is that the heated coolant can be used to warm up the traction battery pack to ensure optimal performance during start-up at low temperature. However, this type of system does require more packaging space for the added coolant circuits and will consume more electric energy.

The Changchun Chengde® air heating PTC that is installed in the production Beijing E-150 pure EVs is shown in Figure 14 with its temperature—resistance characteristic curve. The key product parameters are given in Table 4.

3.3 EV HVAC system design trend

In the initial effort to convert a conventional vehicle to an EV, the emphasis has been on minimizing the component change to achieve viable system cost and reliable operation. For the HVAC system, the basic approach is to replace the belt-driven compressor with an electric compressor and to add a supplemental heating source to satisfy regulation and comfort requirements. Owing to the large power consumption of the two devices, the control mechanisms will have

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

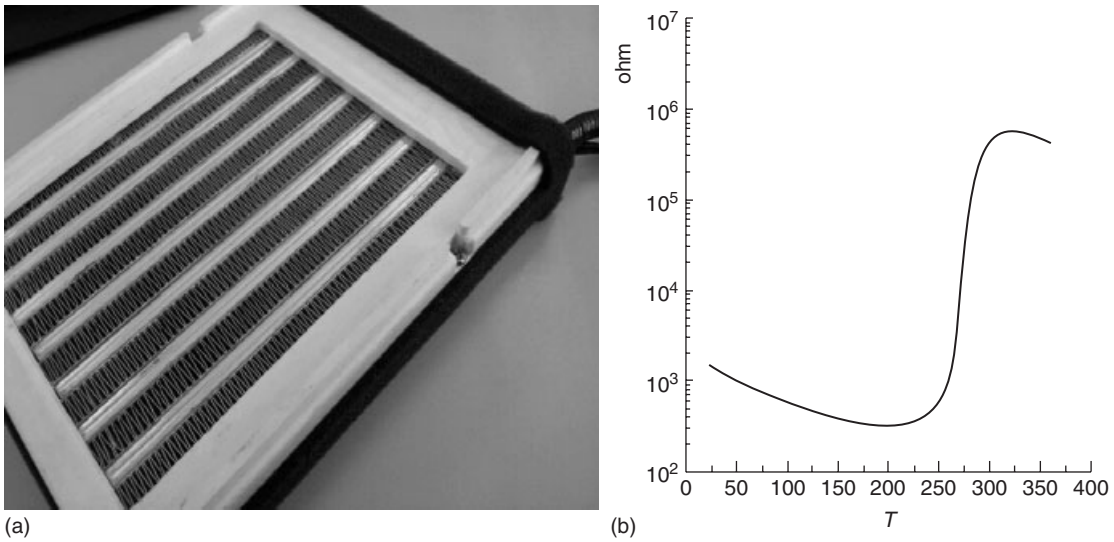


Figure 14. (a,b) Chengde® PTC air heater with temperature–resistance characteristic curve.

Table 4. Chengde® air heater PTC parameters.

Technical Parameters		Test Condition
Rated input voltage	384 V DC	384 V DC
Rated power	2000 W	Ambient temperature: $25 \pm 1^{\circ}$C
Rated power variations	−10%∼+10%	Applied voltage: 384 ± 1 V DC Wind speed: 4.5 m/s
Maximum start-up current	13 A	Ambient temperature: $25 \pm 1^{\circ}$C Applied voltage: 384 ± 1 V DC
Cold resistance	80–300 Ω	Soaked in $25 \pm 1^{\circ}$C ambient temperature for 30 min

to be adapted as well. This involves, for example, (i) adding direct control of the heating element (just like the AC activation switch in a conventional vehicle) for customer to consciously control the energy consumption (and hence driving range), (ii) dynamic regulation of the surface temperature of the heater core, and (iii) dynamic adjustment of the compressor cooling capacity for maximum level of system efficiency based on vehicle and environmental conditions.[10]

In addition, new concepts are being developed to further improve system level performance. One approach is to localize heating and cooling only to the driver and occupied passenger locations instead of the whole cabin (Weissler, 2012). It involves redesigning the HVAC air distribution system (e.g., the HVAC case, the temperature/mode control doors, and the air ducts) to direct airflow to the spots where the customer is most sensitive, for example, the ankle, the chest, and the face. By selectively enhancing airflow, and together with the use of climatically controlled seats, the objective is to make the driver and passenger feel more comfortable without actually increasing the heating or cooling output. A second approach is to precondition the passenger compartment before the trip while the vehicle is being charged from an external power supply. New thermal energy storage systems are being developed that have volumetric energy density greater than that of electrical batteries with charging and discharging achieved via solid-state converters (Green Car Congress, 2011). Thanks to the recent advances in the Telematics technology, the whole process can be conveniently automated or remotely controlled by the customer. Yet, the fundamental improvement of the heating performance comes with the incorporation of a heat pump system. The COP of the PTC-based system is below 1.0 (typically 0.95). The COP of a heat pump system can reach over 2.3. This is demonstrated by the Delphi Unitary HPAC (heat pump air conditioner) System (Kowsky *et al.*, 2012). In addition to improved heating efficiency, the system also has the potential to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

integrate powertrain components (e.g., the traction battery, inverter, and motor) cooling with the HVAC system for optimal vehicle level performance.

4 ELECTRIC VACUUM BOOSTER SYSTEMS

The vast majority of vehicles are equipped with vacuum servo brake system to provide assistance to the driver by reducing the braking effort. The vacuum source of the vacuum booster system for vehicles with normally aspired IC engines is from the engine intake manifold. The vacuum negative pressure could reach 0.05–0.07 MPa. For pure electric or fuel cell vehicles refitted from the traditional vehicle, the vacuum source is lost because of the removal of the engine assembly. The braking force generated solely by the human cannot meet the requirement of the servo brake system. A common solution is to restructure the booster system by adding an electric vacuum pump that can provide sufficient vacuum pressure to the brake booster system. The basic configuration of an electric vacuum booster brake system is shown in Figure 15. It consists of the electric vacuum pump, the vacuum tank, and the real-time controller. The mechanical structures of vacuum pumps used for brake servo systems include the rotary vane (with oil lubrication or dry type) and the diaphragm type. This will be discussed in the following sections.

4.1 Rotary vane pump with function of alarming

The production Beijing E-150 pure EVs are equipped with the rotary vane type vacuum pump as shown in Figure 16.[11] The structure diagram of electric vacuum booster brake system is displayed in Figure 17.

The rotary vane pump is mainly composed of the stator, rotor, rotary vane, upper cover plate, lower cover plate, springs and other parts. The rotor is eccentrically mounted inside of the stator cavity with spring loaded sliding rotary vanes. When the rotor is driven by the electric motor, the rotary vanes slide out of the rotor and are pressed against the inner lining of the stator due to the centrifugal force.

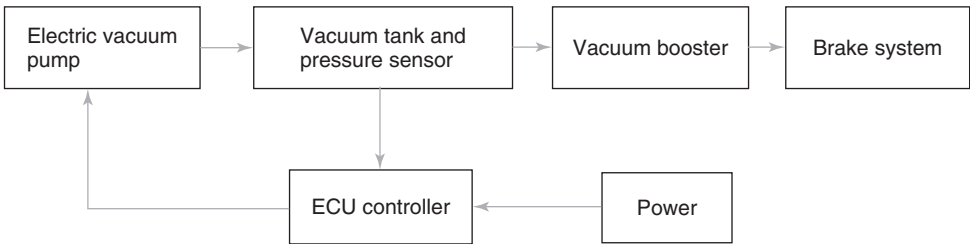


Figure 15. Schematic diagram of electric vacuum booster brake system.

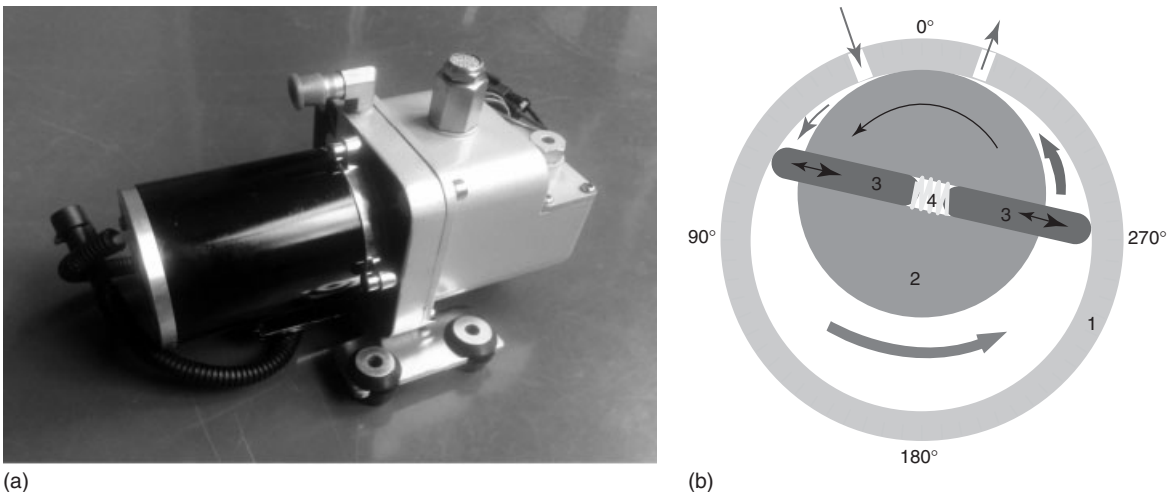


Figure 16. (a,b) Rotary vane pump with alarm function. (Illustration of internal structure. Reproduced from Wikipedia (http://en.wikipedia.org/wiki/Rotary_pump).)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

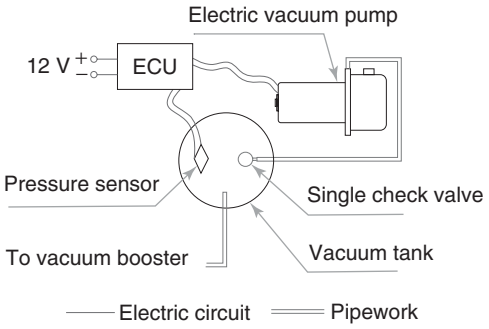


Figure 17. Structure diagram of electric vacuum booster brake system.

This creates the vane chambers that do the pumping work as follows. On the intake side of the pump, the vane chambers are increasing in volume. These increasing volume vane chambers are filled with air forced in by the inlet pressure. The inlet pressure is actually the pressure from the system being pumped which is the vacuum booster tank in this application. On the discharge side of the pump, the vane chambers are decreasing in volume, forcing air out of the pump into the atmosphere. With fixed inlet pressure, the pump drives out the same volume of air with each rotation.

A pressure sensor measures the pressure of the vacuum tank. A simple strategy of threshold comparison with hysteresis control is implemented with the following functions.

(a) *Prestart Self-Checking*. Implement the self-check function after vehicle start to detect system failure and activate the alarm.
(b) *Start*. The vacuum pump starts to work when the measured tank pressure reaches threshold P_a.
(c) *Pressurization*. When the pressure reaches threshold P_b, the vacuum pump continues to increase the degree of vacuum for an adjustable delay time of 5–20 s.
(d) *Stop*. When the pressure reaches threshold P_c, the vacuum pump stops working immediately.
(e) *Alarm*. When the pressure reaches threshold P_d, the alarm is activated.

The primary performance criteria of a vacuum pump include the vacuum level that can be produced, the rate of air removal, and the required input power. Other parameters include the temperature effect, the noise level, and the service life. These are listed in Table 5.

Table 5. Specifications of the rotary vane electric vacuum pump assembly.

Parameters	Specifications
Rated voltage	12 V
Antileak tightness	At -66.7 ± 5 KPa vacuum level, the pressure loss shall be less than 0.3 KPa during a period of 15 s
Maximum suction	≥ -85 KPa
Operating current	≤ 15 A
Noise	≤ 75 dB(A)
Operation temperature	-30°C to100°C
Corrosion resistance	No more than 2 diameter corrosion per 100 cm^2
Pumping capacity	With test volume of 2 L, time for pressure initiation is: 1. From atmosphere pressure to -50 kPa: less than 4 s 2. From atmosphere pressure to -70 kPa: less than 7 s 3. From -40 to -70 kPa: less than 4 s
Service life	Activation cycles: 1. Normal temperature: minimum 200 K cycles 2. Low temperature (-30°C): minimum 50 K cycles 3. High temperature (100°C): minimum 50 K cycles

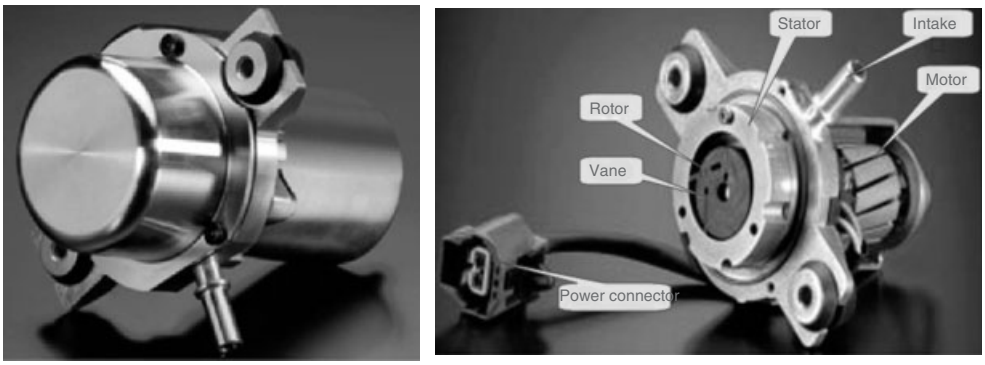
4.2 Auxiliary dry rotary vane pump

The vacuum pump introduced in the previous section requires oil lubrication. The oil-lubricated types have distinct advantages if proper maintenance is provided. They can usually provide higher vacuums because the lubricant acts as a sealant between moving parts. In addition, they usually last longer than oil-less units in normal service because of their cooler operation and less susceptibility to corrosion from condensed water vapor. Yet, oil-less pumps are preferred for the brake booster application on the basis of avoiding the cost and time of regularly maintaining the oil fillings. This is particularly important when the pump is to be mounted in an inaccessible location. The dry rotary vane pump presented in this section is one type of oil-less vacuum pump that uses specialized rotary vanes to eliminate the need of the lubricating oil. The diaphragm pumps introduced in the next section are also designed to be oil-less.

The auxiliary dry rotary vane pump with its internal structure displayed is shown in Figure 18. The internal structure, operation principle, and control strategies between the oil-less and the oil-lubricated pumps are quite similar, which will not be detailed again here. The major technical features of the dry pump are listed below:

- Graphite rotary vane structure;
- Good sealing performance;
- Maintenance-free, without oil refilling;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) (b)

Figure 18. (a,b) Auxiliary dry rotary pump outline.

Table 6. Technical parameters of auxiliary dry rotary vane pump.

Parameters	Specifications
Rated voltage	12 V
Operation voltage	6–16 V
Rated current	10 A
Starting current	≤60 A
Operation temperature	−40°C to 120°C
Noise	≤75 db(A)
Performance	With test volume of 3.2 L, time for pressure initiation is: 1. From atmosphere pressure to −50 kPa: less than 5.5 s 2. From atmosphere pressure to −70 kPa: less than 12 s
Service life	≥300 K Cycles

- High pumping speed with shorter vacuum restoration time
- Compact and light weight
- Wide operating temperature range: −40°C to 120°C.

The technical parameters are listed in Table 6.

Table 7. Main technical parameters of diaphragm pump.

Parameters	Specifications
Rated voltage	13 V
Average current consumption (13 V; room temperature; after run-in)	<15 A
Operation temperature	−40°C to 110°C 120°C for limited time
Maximum degree of vacuum (13 V; room temperature; after run-in)	Great than 86% of ambient pressure
Noise level (13 V; room temperature; after run-in)	<60 dB(A)
Motor speed	≈3300 rpm
Pumping capacity (13 V; room temperature; after run-in)	With 5I Booster, time for pressure initiation is: 1. From atmosphere pressure to −50 kPa: less than 3.5s 2. From atmosphere pressure to −70 kPa: less than 6.5 s
Maximum working load (>80 h, 90°C)	100%
Service life	>1200h >1.2 *M* Cycles
Product weight	<2.6 kg

4.3 Diaphragm vacuum pump

A production diaphragm pump for EV applications is shown in Figure 19.[12] Its internal structure is revealed in Figure 20. The diaphragm pump creates vacuum by flexing of a diaphragm inside a closed chamber. The reciprocating action of the diaphragm is driven by an electric motor. The check valves on either side of the chamber ensure single direction of airflow. Diaphragm pumps have good suction lift characteristics and handle well with impurities in the air. The flow rate primarily depends on the effective working diameter of the diaphragm, its stroke length, and the motor speed. Diaphragm pumps also have good dry running characteristics and can be very power efficient. The detailed technical parameters for the pictured product are listed in Table 7.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

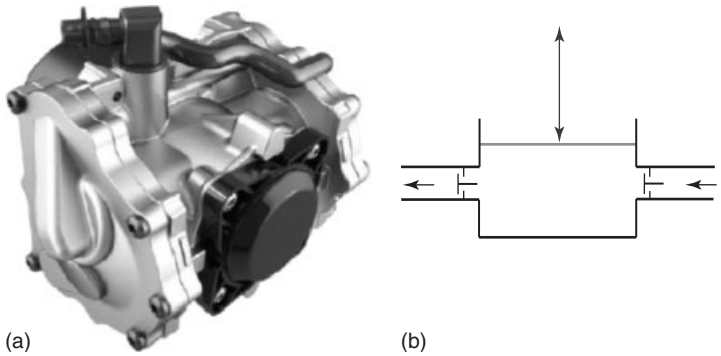


Figure 19. (a,b) Outline drawing of diaphragm pump. (Illustration of internal structure. Reproduced from Wikipedia (http://en.wikipedia.org/wiki/Diaphragm_pump).)

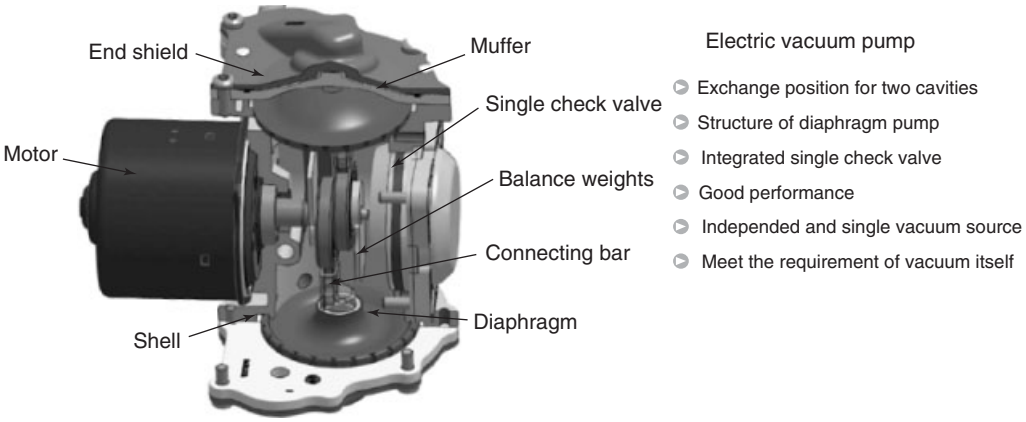


Figure 20. Internal structure of diaphragm pump.

5 ELECTRIC POWER STEERING SYSTEMS

Currently, almost all production EV models are equipped with electric power steering (EPS), also refereed to as electric power assisted steering (EPAS), systems of different types. As in conventional IC engine vehicles, EPS systems are categorized into the column-assist type (C-EPS), the pinion-assist type (P-EPS), and the rack-assist type (R-EPS), according to the location of the booster actuation. However, unlike in conventional vehicles, electro-hydraulic power steering (EHPS) systems are seldom used in EV's because of the need for a separate power source.

The EPS system is in general composed of a torque sensor, an electronic control unit, a power amplifier module, an electric motor, and a reduction gear mechanism. The electronic control unit calculates the required steering assist torque according to the torque sensor output and controls the rotation of the electric motor through the power amplifier module. The motor output drives the conventional steering rack and pinion mechanism after the reduction mechanism to produce the corresponding assist steering torque. The electronic control unit also communicates with other vehicle subsystems through onboard network to further enhance steering control accuracy and comfort and to optimize fault diagnosis. The EPS system must meet the reliability requirements and electromagnetic compatibility requirements of an automotive electronic system.

Note that the design of EPS systems in EVs follows the same basic principles of that in conventional IC engine vehicles. Therefore, interested readers are recommended to reference the related materials as provided in New Electrical Power Steering Systems for more details. For the remainder of this section, two production EPS systems, both from Zhuzhou Elite Tech Corporation, are briefly presented. The first is the column-assist type EPS as shown in Figure 21 with technical parameters in Table 8. The second is the pinion-assist type EPS as shown in Figure 22 with technical parameters in Table 9. These systems have been fitted to the Beijing E-150 pure EVs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 21. Production C-EPS system. (Reproduced with permission from Zhuzhou Elite Tech Corporation.)

Table 8. Technical parameters of the C-EPS system.

Devices	Parameters	Specifications
Controller	Supply voltage	10–16 V
	Maximum operating current	35 A
Motor	Rated voltage	12 V
	Rated power	200 W
	Maximum operating current	35 A
	Maximum output torque	2 Nm
Sensors	Supply voltage	3.3 V
	Main torque voltage	0.66–2.64 V
	Assist torque voltage	0.66–2.64 V
	Medium voltage between main and assist torques	1.65 V
Column assembly	Maximum steering wheel output torque—manual	7 Nm
	Reduction ratio	16.50
	Maximum steering wheel output torque—assisted	40 Nm



Figure 22. Production P-EPS system. (Reproduced with permission from Zhuzhou Elite Tech Corporation.)

Table 9. Technical parameters of the P-EPS system.

Parameters	Specifications
Angle ratio of steering gear	44.15 mm/rev
Angle of gear and rack shaft	20°
Center distance between pinion and rack	16.5 mm
Rack displacement	±71.5 mm
Rack width	24 mm
Number of pinion	8
Module of pinion	1.75
Pitch radius of pinion	7.724 mm
Rated current of EPS motor	52 A
Rated output torque of EPS motor	2.36 Nm
Rated power of EPS motor	360 W
Reduction ratio between worm and gear	1 : 18

6 SUMMARY

This chapter serves as an introduction of the auxiliary systems for EVs. These systems complement the main power house of the EV (namely the traction battery, the traction motor, and the vehicle central control unit) and help create a safe and comfortable driving environment. The systems covered in this discussion include the DCDC converter that powers the conventional 12 V low voltage net, the HVAC system that is essential for driving comfort and hence safety, the electric vacuum booster system for assisted power braking, and the EPS system that increasingly becomes an integral part of the total vehicle dynamic control systems. It shall be clear from the presentation that the theme of the design is energy efficiency, which coincides with that of the total EV. Yet, the drive for sustainable mobility shall also take advantages from technologies in other fields. The particular one worth mentioning is the technologies for connectivity that intimately link between the vehicle, the customer, a (typically cloud-based) back-end server, and from there, an integrated supporting infrastructure. Together, they provide the foundation for the success of EVs.

ACKNOWLEDGMENT

The authors would like to acknowledge the support from their colleagues including but not limited to Dr. W. Gao, Dr. L. Gu, Mr. H. Zhou, Dr. Y. Wei, Dr. R. Wu, Dr. G. Huang, and Dr. A. Cooprider. The authors appreciate the guidance that the associate editor Prof. C.C. Chan had kindly offered to the team.

ENDNOTES

1. For naturally aspired engines. For turbo charged engines, a dedicated vacuum pump may be required.
2. This can be seen from the equation $P_{\text{Loss}} = I_{\text{Load}}^2 \times R_{\text{Conductor}} = (P_{\text{Load}}/U_{\text{Load}})^2 \times R_{\text{Conductor}}$ that with given load power P_{Load} and conductor resistance $R_{\text{Conductor}}$,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the power loss in the conductor P_{Loss} is inversely proportional to the square of the voltage across the load U_{Load}.

3. Another approach is to have multiple windings on the primary side that turn on in sequence to (continuously) inject current to the secondary winding side. Thus, the effective duty cycle is multiplied without jeopardizing the maximum blocking voltage of the main switch. This is the so-called *multi phase* DCDC converter that has additional benefits such as (i) the heat generating components (the winding and the main switch) are spread out so that the power density of the device can be increased and (ii) if one primary winding circuit fails, the device can still operate with reduced power.
4. At higher operating temperature, the conversion efficiency can also be reduced because of higher conduction loss in the transformer (copper loss) and the main switching semiconductors.
5. Electromagnetic compatibility (EMC) is defined as the ability of an equipment or system to function satisfactorily in its electromagnetic environment without introducing intolerable electromagnetic disturbances to other devices in that environment. Much of the existing methods to address EMC of motor drives (primarily conducted and radiated emissions) are based on "legacy" low voltage components and systems. The high voltage (typically 200–400 V) and high current (hundreds of amperes peak demand) in EV propulsion systems will require new analysis approaches, test methods, and effective use of simulation/modeling. The standards are also under development by various governmental agencies and industrial institutions. The reader may refer to the publications from EC, ISO, IEC, SAE, and other issuing bodies for the latest status.
6. For example, when operating at high altitude with reduced air pressure, the cooling capacity has to be derated compared to that of sea level. A rule-of-thumb formula for the derating factor is given as $1\frac{h}{17{,}500}$, where h is the height over sea level in meters.
7. Under IEC529 designation, the first digit "6" implies "dust tight," or no ingress of dust (at a partial vacuum of 20 mbar inside the enclosure); the second digit "6" implies "powerful water jets," which means water projected in powerful jets against the enclosure from any direction shall have no harmful effects.
8. It should be noted that for vehicles equipped with highly efficient small diesel engines, it is also often necessary to have a dedicated heat source such as the fuel burning heater or the PTC electric heater that are discussed in this section for sufficient heating in very cold weather.
9. Owing to its relatively high surface temperature, care shall be taken when packaging the PTC heater to ensure safety of operation. This includes keeping plastics, sponge, and other temperature-sensitive components at safe clearance to the heating element and adequate ventilation.
10. It should be noted that the aforementioned technologies are not unique to electric vehicles but have seen applications in (high end) conventional vehicles for better fuel efficiency.
11. The internal structural diagram of the rotary vane vacuum pump as displayed in Figure 16 is based on Wikipedia, http://en.wikipedia.org, article on Rotary vane pump, March 2013.
12. The internal structural diagram of the diaphragm vacuum pump as displayed in Figure 19 is based on Wikipedia, http://en.wikipedia.org, article on Diaphragm vane pump, March 2013.

REFERENCES

Green Car Congress (2011) Sheetak and Delphi to develop efficient HVAC for EV/HEVs; $4.7M ARPA-E award, http://www.greencarcongress.com/2011/12/sheetak-20111207.html (accessed 7 December 2011).

Gu, Y., Lu, Z., and Qian, Z. (2004) DC/DC topology classification and selection criterion, *Journal of Zhejiang University (Engineering Science)*, **38**(10), 1375–1379.

Hendricks, T.J. (2003) Multi-Variable Optimization of Electrically-Driven Vehicle Air Conditioning Systems Using Transient Performance Analysis, *Proceedings of VTMS 6: Vehicle Thermal Management Systems*, Brighton, UK, pp. 135–148, 18–21 May 2003.

Kowsky, C., Wolfe, E., Leitzel, L., *et al*. (2012) Unitary HPAC system, *SAE International Journal of Passenger Cars—Mechanical Systems*, **5**(2). DOI: 10.4271/2012-01-1050

Shiraki S., Kudo H., Tago M., *et al*. (2012) A Novel Concept of High Voltage Auxiliaries and its Feasibility Study on Blower Motors, *Proceedings of FISITA 2012 World Automotive Congress*, Lecture Notes in Electrical Engineering 194, doi: 10.1007/978-3-642-33829-8_24, Springer-Verlag Berlin Heidelberg 2013.

Su G. J., Peng F. Z. (2005) A Low Cost, Triple-Voltage Bus DC-DC Converter for Automotive Applications. *the IEEE Applied Power Electronics Conference and Exposition (APEC)*, Austin, Texas, Vol. 2, pp. 1015–1021, March 6–10, 2005.

Weissler, P. (2012) Zone, Spot HVAC Systems Could Save Fuel or Extend EV Range, *Automotive Engineering online* http://www.sae.org/mags/aei/saewc/11013 (accessed 18 May 2012).

Zhang A., Huang G., Gu Y. (2002) Asymmetrical Full Bridge DC-to-DC Converter. United States Patent, Patent No. 6,466,458 B2, October 15.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto153
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Steel Processing: Formability of Steel Sheets and Tailor-Welded Blanks for Automotive Applications

Kaushik Bandyopadhyay and Sushanta K. Panda

IIT Kharagpur, Kharagpur, India

1 Introduction 1
2 Deep Drawing of Anisotropic Steel Sheet 4
3 Formability of Tailor-Welded Blank 19
4 Conclusions 24
Acknowledgments 24
Related Articles 24
Endnotes 25
References 25

1 INTRODUCTION

The intrinsic ability of metals and alloys to be worked into various shapes without removing materials forms the basis for manufacturing operations known as *metal forming* or *plastic deformation processes*. Metal forming processes can be broadly classified into two types as bulk forming processes and sheet metal-forming processes. In sheet metal forming, the *working material* often referred to as *blank* has a lower dimension in one direction compared to other two directions. The complete discussion on steel processing including iron making, primary and secondary steel making, hot rolling, cold rolling, extrusion, forging, joining, machining, coating, and heat treatment is very complex and time consuming, and hence, beyond the scope of this chapter. However, the various aspects of sheet steel processing are discussed. In sheet metal processing, plastic deformation is imposed in the blank by punch and die tooling assembly to achieve a required three-dimensional shape. Tensile stress plays a major role during deformation. Commonly used sheet metal-forming processes are bending, deep drawing, ironing, and stretch forming (Figure 1).

In sheet bending, the desired bent or curved shape is given to the sheet material by deforming about the bending axis with little or no change in surface area. Mostly, sheet bending operations are U-bending, air bending, arc bending, and V-bending process. In bending, inner side of the sheet metal across the thickness is compressed and outer side is stretched. Through successive bending, complex form can be achieved. Sheet metal is drawn into desired flat end hollow cylinder or box shape with the help of flat bottom punch and dies in deep drawing process. A blank holder is used to prevent wrinkling in the flange portion of the cup. In deep drawing, the thickness of the cup wall is not constant. It is minimum near the cup bottom and maximum at the top of the cup. Ironing is the process in which a cup height is increased at the expense of wall thickness, making the wall thickness more uniform. This can be achieved by forcing the cylindrical cup through an ironing die in which the punch–die clearance is smaller than the metal thickness. Ironing may be done along with deep drawing or after deep drawing as a separate step. In stretch forming, blank is clamped rigidly by jaws along the edges and stretched over a punch or die block. Generally, the punch has a spherical or convex shape. In stretching, mainly tensile stress plays the major role to deform the sheet. Complex autobody stamping requires a combination of stretching and drawing with which bending and unbending are associated (Mellor, 1981).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

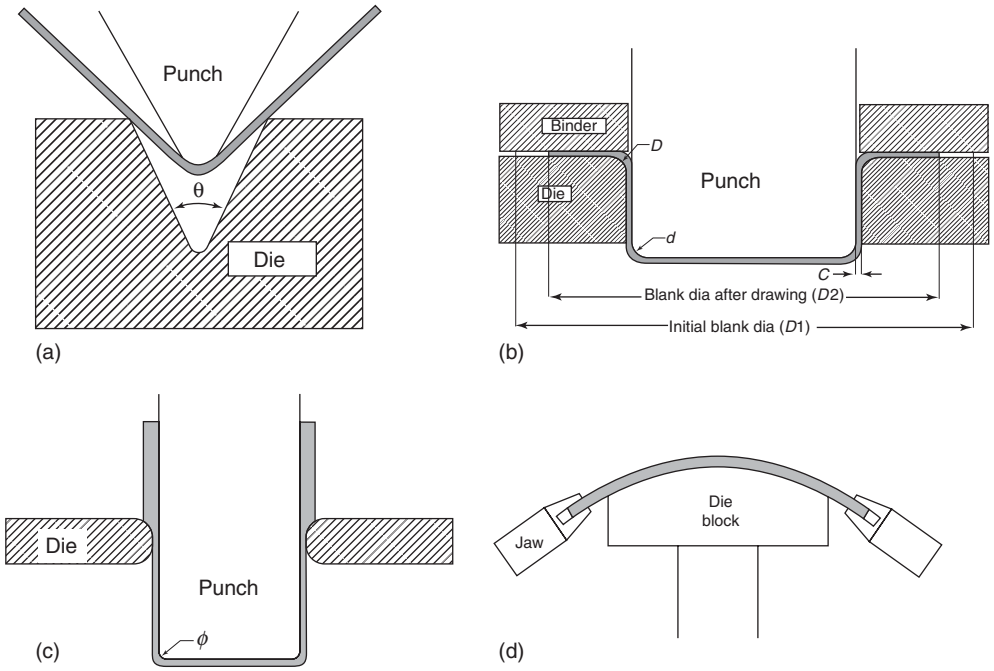


Figure 1. Schematic diagrams of some basic sheet forming processes: (a) bending, (b) deep drawing, (c) ironing, and (d) stretch forming.

The ability of a sheet metal to be worked into different shapes without flow localization under complex conditions of loading and deformation is called *formability*. Flow localization is an unstable flow where the deformation is confined to one zone of the sheet metal. This is also referred to as *onset of necking*, which leads to fracture on further deformation. When a stamping tears during forming, the tear is a visible indication that metal has been worked beyond its prevailing forming limit. Whether or not a particular sheet of metal can be formed without failure depends on many factors such as material properties, surface conditions, blank size and shape, lubrication, press speed, blank holder pressure, and punch and die design. Because of this, source of flow localization problem is difficult to pinpoint and quantify. Hence, formability is an attribute that has no precise, universal meaning. Because of multitude of materials, stamping designs, and press conditions, there are no standard valid rules for improving the formability of a stamping through changes in tool design or process parameters (Mellor, 1981; Rao and Sing, 2000).

1.1 Materials for automotive manufacturer

Typical automotive steel grades include low strength steels such as interstitial free (IF), drawing quality (DQ), extra deep drawing (EDD) steel, and conventional high strength steel such as high strength low alloy (HSLA). These low carbon steel sheets have long been the work horse material in automotive manufacturing and consumer industries because it can be stamped into inexpensive, complex stamping at a very high production rate (Panda, 2007). However, the driving force for the increased use of recently developed advanced high strength steel (AHSS) has been their superior strength and enhanced formability compared to HSLA. In addition, these steels have excellent crash energy absorbing behavior. The various types of commercial available AHSSs are dual phase (DP), transformation-induced plasticity (TRIP), complex phase (CP), twinning-induced plasticity (TWIP), and martensitic steel (MS). These AHSSs are primarily steels with a microstructure containing a phase other than ferrite, pearlite, for instance, martensite, bainite, austenite, and/or retained austenite in quantities sufficient to produce favorable mechanical properties. Application of AHSS in automotive body structures, for example, motor compartment rail, bumper, B-pillar, and door panel, can significantly increase the crashworthiness of the vehicle with reduction in weight (Advanced High Strength Steel Application Guidelines, 2009). It was reported by Worldsteel Association[1] that replacement of conventional steels by AHSS for the car body or the body in white (BIW) resulted in 17–25% weight savings. Weight saving has a remarkable influence on a vehicle's efficiency, with researches showing that 1% reduction in a vehicle weight can result in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Representative chemical composition of different low carbon steels with carbon equivalent.

Materials	C	Mn	Si	Al	Ti	Nb	CE[a]
IF	0.008	0.088	0.006	0.068	0.056	0.026	0.018
IFHS	0.0011	0.4221	0.0115	0.049	0.046	0.0012	0.039
HSLA	0.08	0.827	0.454	0.048	0.013	0.029	0.18
DP600	0.125	1.567	0.193	0.045	0.0219	0.0018	0.39
DP980	0.165	1.483	0.341	0.051	0.0021	0.0014	0.41

[a]

$$\mathrm{CE} = \mathrm{C} + \mathrm{A(C)}\left(\frac{\mathrm{Mn}}{6} + \frac{\mathrm{Si}}{24} + \frac{\mathrm{Cr}+\mathrm{Mo}+\mathrm{V}}{5} + \frac{\mathrm{Cu}}{15} + \frac{\mathrm{Ni}}{20} + \frac{\mathrm{Nb}}{5} + 5\mathrm{B}\right)$$

where $\mathrm{A(C)} = 0.75 + 0.25\tanh[20(\mathrm{C} - 0.12)]$.

Table 2. Mechanical properties of some automotive steel sheets.

Material (thickness)	Yield Strength (MPa)	Ultimate Strength (MPa)	Elongation (%)	*n*-Value	*K*-Value (MPa)	R_0	R_{45}	R_{90}	$\overline{R}$
IF(1.0)	132	278	50	0.28	550.1	1.75	1.50	3.07	1.96
IFHS(0.7)	191	364	39	0.26	645.3	1.22	1.18	1.39	1.24
HSLA(1.0)	359	431	26	0.13	760.9	1.02	1.23	1.16	1.16
DP600(1.2)	365	631	26	0.21	1097.6	0.80	0.96	1.03	0.94
DP980(1.2)	672	1058	12	0.10	1505.4	0.83	0.91	1.05	0.93

0.6–1% reduction in fuel consumption (Hrayashi, 1996). The chemical composition with carbon equivalent (CE) using Yurioka formula (Yurioka and Kasuya, 1995) and the mechanical properties of some automotive grade steel sheets are enlisted in Tables 1 and 2 for reference. The influence of these properties on forming behavior is discussed in various sections of this chapter where ever necessary.

1.2 Tailor-welded blank

Automakers are constantly searching for innovative means of reducing vehicle weight and manufacturing costs in order to meet ever-restricting fuel economy standards while remaining economically competitive. A promising opportunity to meet these seemingly conflicting requirements is through the use of tailor-welded blanks (TWBs). TWBs are blanks where multiple sheet metals of different shapes and thickness are welded together before forming into a three-dimensional component. Thus, the blanks can be tailored for a particular application, including not only sheets of different shape and thickness but also sheets of different quality, and with or without coatings on one or both surfaces. This trend of welding and forming of sheet metal pieces allows significant flexibility in product design, structural stiffness and crash behavior (crashworthiness), formability, and use of different materials in one component (Auto Steel Partnership, 1995; Wang *et al*., 1995). This method is becoming increasingly important, particularly in the automotive industry. Because each welded piece can have a different thickness (as guided by design considerations such as stiffness), grade of sheet metal, coating, or other characteristics, these blanks possess the needed characteristics in the desired locations of the formed part. As a result, productivity is increased, the need for subsequent spot welding of the product (e.g., a car body) is reduced or eliminated, and dimensional control is improved. TWBs offer several other notable benefits including decreased part weight, reduced manufacturing costs, increased environmental friendliness, and improved dimensional consistency (Lokka, 1997). In order to take advantage of these benefits, however, designers need to access the formability of TWBs early in the design process. Hence, researchers have worked on various aspects on formability of TWBs (Kinsey, Liu, and Cao, 2000; Chan and Chan, 2003; Kinsey, Viswanathan, and Cao, 2001; Waddell, Jacken, and Wallach, 1998). Currently, TWBs are used for manufacturing auto body parts such as front door inner, rear door inner, floor pan, A-pillar, center pillar (B-pillar), and body side frames (Pallet and Lark, 2001). The application of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

AHSS in TWB has the potential to significantly reduce part weight and improve product performance.

1.3 Hot stamping and coating

In the context of continuously increasing demands for reduced vehicle weight with improved crashworthiness qualities, a brief note about the hot press forming (HPF) or hot stamping technology is mentioned to produce AHSS parts for automobile application. Recently, different hot stamped parts find their applications in various chassis components: door beams, bumpers, pillars, roof rails, rocker rails, and tunnels. Hot stamping has two main variants: direct HPF and indirect HPF. In direct HPF, the blanks are heated to achieve full austenization (i.e., 900–950°C), and subsequently, they are press formed and quenched. However, the cold preformed parts are first austenized and then these parts are calibrated and quenched in the press during indirect HPF. After quenching, full martensitic transformation is obtained in the material resulted increase in the strength of the component (Fan, Kim, and De Cooman, 2009). This process can be exploited to tailoring properties of the blanks. It can be achieved by changing the cooling rate below the critical rate to avoid martensite formation in some parts or heating the blank below AC_3 temperature, which results incomplete austenization. Different techniques such as introducing segmented tools having different thermal conductivity at different zone, differentiating temperature of different region of die by heating, local heating of the blank, and so on (Karbasian and Tekkaya, 2010).

Coating using molten zinc by electroplating, spraying, and hot dip galvanizing is used for protection against corrosion in a wide variety of low carbon steel finished products. Under atmospheric conditions with high moisture content or condensation on the metal surface, zinc hydroxide forms, and this film reacts with carbon dioxide to form insoluble zinc carbonate that shields zinc from the outside environment. Thus, zinc carbonate is very protective and is responsible for the excellent corrosion resistance of zinc in the atmosphere (Porter, 1991). In addition, the zinc layer provides cathodic or sacrificial protection to the substrate steel under extreme corrosive environment as it is more electronegative than steel. Research shows that by annealing the hot dip galvanized coatings to produce galvannealed coatings, the corrosion rate can be substantially decreased (Zhang, 2000). This is due to the presence of various Fe–Zn intermetallic phases. Unfortunately, the great difference in linear thermal expansion coefficient between steel and Fe–Zn phases lead to development of large residual stresses at high heat input, and this may result detrimental cracks in coating. Mainly, coating of Al–Si layer is applied by researchers in direct HPF to prevent the formation of scale during hot forming. Recently, lots of research work is ongoing in the selection of various oxidation preventive oils, metallic, inorganic, and organic coatings with coolants for hot stamping applications (Karbasian and Tekkaya, 2010; Naderi *et al*., 2011).

2 DEEP DRAWING OF ANISOTROPIC STEEL SHEET

Deep drawing process is widely used in automotive industries for forming complicated sheet metal parts, for example, automotive body parts, fuel tank, and flash light at a very high production rate. This process is also popular for assessing formability of sheet metals in terms of limiting drawing ratio (LDR). In deep drawing operation, different regions of the half drawn cup is subjected to different states of stresses as shown in Figure 2. The flange region of the cup experiences a radial tensile stress and tangential (hoop) compressive stress resulting reduction in circumference of the blank while being pulled into the die cavity. The cup flange buckles when the tangential compressive stress reaches a critical value and wrinkling defects appear due to lateral deflection. However, the development of lateral deflection of flange is suppressed by application of compressive blank holding force. The cup wall is held tightly over the punch, whereas the drawing force is transmitted into the cup. Hence, the cup wall region is experiencing tensile longitudinal and hoop stress. The drawing punch force is limited to the maximum tensile load that can be carried by the cup wall and this in turn limits the depth of drawn cup with a tearing mark. The cup bottom experiences a minor biaxial tensile stress as this region is resting on the flat face of the punch from the beginning of deep drawing.

2.1 Anisotropy in sheet steel-Lankford parameter

The mechanical properties are not uniform in all direction of cold-rolled steel sheets because of preferred crystallographic orientation of grains (texture of polycrystalline). The anisotropic behavior of sheet metal is quantified in terms of Lankford parameter, often referred as *R*-value and it is defined as Equation 1.

$$R = \frac{\varepsilon_2}{\varepsilon_3} \tag{1}$$

where ε_2 and ε_3 are the true strains in width and thickness directions, respectively. As the thickness strain is difficult

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

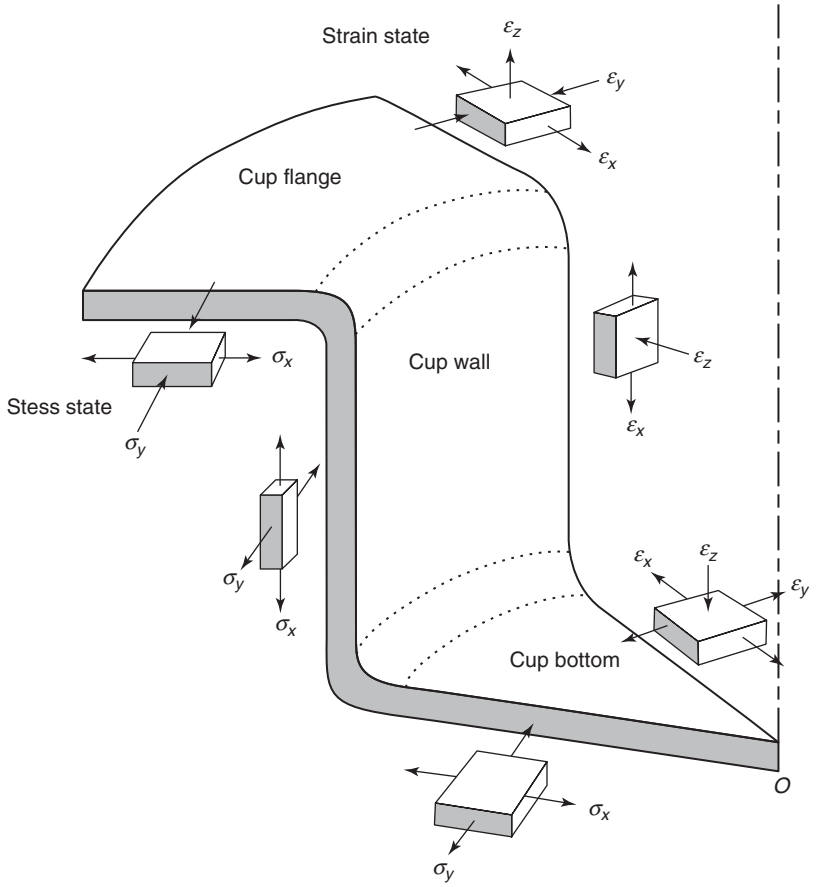


Figure 2. State of stress and strain at different regions of a deep drawn cup.

to measure, the volume constancy condition is invoked and the Lankford parameter can be calculated as Equation 2,

$$R = \frac{\varepsilon_2}{-(\varepsilon_1 + \varepsilon_2)} \tag{2}$$

where ε_1 is the true strain in length direction. Typically, *R*-value is measured from tensile specimens machined parallel, perpendicular, and at 45° to the rolling direction (RD). The *R*-value of the rolled steel sheet in these directions being represented by R_0, R_{90}, and R_{45} respectively. The average normal anisotropy coefficient is calculated as Equation 3.

$$\overline{R} = \frac{R_0 + 2R_{45} + R_{90}}{4} \tag{3}$$

Sheet metals with high $\overline{R}$-value have a low thickness strain compared to the strain in width direction. Hence, these sheets have a high resistance against thinning and are therefore said to have a good drawability. Another description of anisotropy is the variation of the *R*-value in the plane of the sheet, referred as *planar anisotropy* and it is defined as Equation 4.

$$\Delta R = \frac{R_0 - 2R_{45} + R_{90}}{2} \tag{4}$$

If $\Delta R = 0$, then the sheet metal is exhibiting planar isotropy, that is, same properties in all directions within the plane of the sheet. However, most of the sheet metals are planar anisotropic and exhibit different resistance while drawing into the die cavity in different direction. This leads to a wavy profile on the top edge of the completely drawn cup, which is referred as *earring*. These ears are usually removed through trimming operation after deep drawing and leads to lots of scrap materials. Hence, earring is a highly undesired phenomenon in deep drawing and referred as a *defect*.

2.2 Limiting drawing ratio (LDR)

Wrinkling, earring, and tearing are most common defects observed in deep drawn products and are shown in Figure 3.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

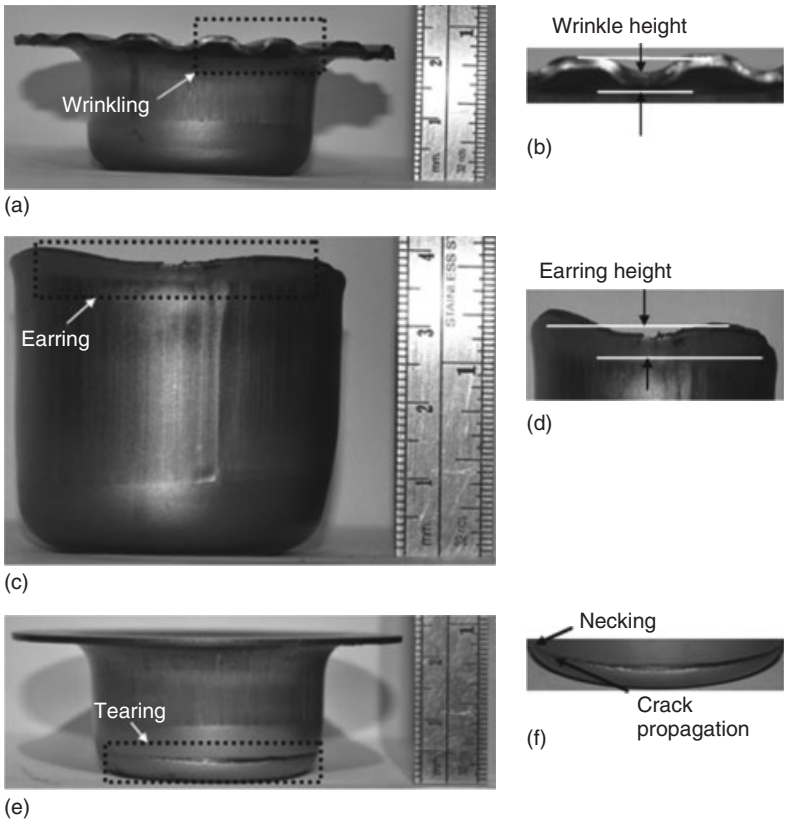


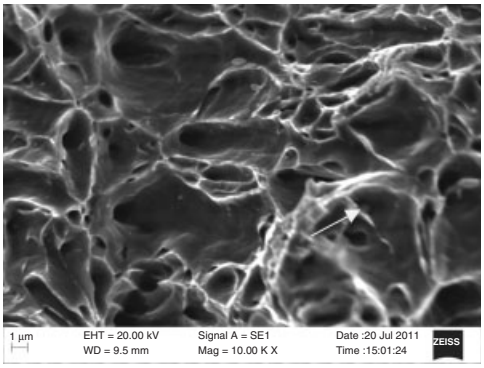
Figure 3. Deep drawing defects: (a) wrinkling, (b) wrinkling height, (c) earring, (d) earring height, (e) tearing, and (f) necking and crack propagation.

Two commonly found surface defects in deformed low carbon steel components are stretcher strain and orange peel. The stretcher strain is flame like pattern of depression on the surface, and it is associated with nonuniform deformation because of Lüder's band formation at yielding. This defect can be avoided by temper rolling. However, the stretcher strain will reappear because of strain aging during storage. Orange peel is a pronounced surface roughening in the deformed component made of steel metal of relatively larger grain size. This problem can be avoided using sheet metal with finer grain size so that deformation of individual grain is difficult to distinguish in bare eye (Hosford and Caddell, 2007). In deep drawing process, drawing ratio (DR) is defined as ratio of the initial blank diameter to the punch diameter. Formability in deep drawing is evaluated in terms of LDR, and it is the maximum DR that may be successfully drawn without wrinkling and tearing. The LDR depends on various tool design parameters, process parameters, and material parameters as listed in Table 3 in a very complex manner. Because of multitude of materials,

Table 3. Various factors influencing formability of sheet metal.

Design Parameters	Process Parameters	Material Parameters
Punch corner radius	Blank holding force	Sheet thickness
Die corner radius	Friction and lubrication	Grain size
Punch–die clearance	Mode of deformation	Material properties
	• In-plane (flat punch)	• Anisotropy (R-value)
	• Out-of-plane (domed punch)	• Strain hardening exponent (n-value)
		• Strain rate sensitive index (m-value)
Drawbead design and location	Punch speed	Inclusions and defects
	Strain path	

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Element	Weight %	Atomic %
O	1.03	3.48
S	**0.91**	**1.52**
Ti	1.51	1.70
Mn	**1.88**	**1.84**
Fe	94.68	91.46
Totals	100.00	

Figure 4. Elemental analysis of fractured surface of IFHS steel sheet.

stamping design, and press condition, there is no single analytical formula to predict LDR and defects.

Nonmetallic inclusions have significant influence on fatigue and impact strength apart from deep drawing behavior of steel sheets. Generally, some inclusions entrapped into the molten steel accidentally and others generate in the process of separation from steel when their solubility in the metal exceeds. A study of the fracture surface of IFHS steel specimen reveals the presence of sulfide inclusion (MnS) in the dimples. The crack initiates at the inclusion interface because of difference in plastic deformation, and the crack grows to generate microvoids. The accumulation of these voids leads to material separation as shown in Figure 4. Inclusions of sulfide such as MnS elongate during hot rolling, and these are also responsible for anisotropy in steel resulting different mechanical properties [YS (yield strength), UTS (ultimate tensile strength), *n*-value, and elongation] in different direction with respect to RD (Paul and Ray, 1997). Similarly, the presence of brittle Fe–Zn intermetallic phases at zinc–steel interface in the coated IF steel sheets appears to have a more dominant role in reducing the formability in stretch forming than the beneficial effect of reduced friction at the punch–sheet interface (Gupta and Ravi Kumar, 2006).

2.3 Numerical prediction of deep drawing behavior

Finite-element (FE)-based models of sheet forming processes are now capable of understanding the complex deformation mechanisms and giving outstanding improvements in economy of manufacture and product quality. Accuracy of this numerical modeling technique depends largely on the use of a constitutive material model that describes the deformation behavior, boundary conditions, and failure criteria.

2.3.1 Theoretical background on finite element method

In FEM (finite element method), the deformable sheet metal is discretized into number of shell or solid elements of volume, ve, and surface area, se, which are interconnected at nodes. The governing equation over the total deformation region consisting of ne-number of elements can be expressed in matrix form as Equation.

$$\sum_{i=1}^{\mathrm{ne}}\left[\int_{\mathrm{ve}}\left(B^{\mathrm{T}}E_{\mathrm{p}}B\,\widehat{u}+N^{\mathrm{T}}\rho N\ddot{u}\right)\mathrm{d}v - \int_{\mathrm{ve}}N^{\mathrm{T}}\mathbf{f}\,\mathrm{d}v-\oint_{\mathrm{se}}N^{\mathrm{T}}\mathbf{t}\bullet\mathrm{d}s\right]_i=0 \tag{5}$$

where B is the strain displacement matrix relating $\varepsilon_{ij} = B\,\widehat{u}$, E_{p} the constitutive matrix representing elastic–plastic modulus of the material, N the matrix of shape function for interpolating displacement inside element from the nodal displacement vector as $u = N\,\widehat{u}$, ρ the density of the material, $\ddot{u}_i$ acceleration of the body, and $\mathbf{f}$ and $\mathbf{t}$ the vectors representing body force and traction force, respectively. The Gauss quadrature technique is mostly applied for numerical integration for each element and the governing equation can be reduced to the following form (Equation 6) after assembling for all the elements.

$$[m]\{\ddot{u}\}+[c]\{\dot{u}\}+[k]\{u\}=\{F\} \tag{6}$$

where the global mass matrix $[m]=\sum_{i=1}^{\mathrm{ne}}\int_{\mathrm{ve}}N^{\mathrm{T}}\rho N\,\mathrm{d}v$, global stiffness matrix $[k]=\sum_{i=1}^{\mathrm{ne}}\int_{\mathrm{ve}}B^{\mathrm{T}}CB\,\mathrm{d}v$, the force vector $\{F\}=\sum_{i=1}^{\mathrm{ne}}\left(\int_{\mathrm{ve}}N^{\mathrm{T}}\mathbf{f}\,\mathrm{d}v+\oint_{\mathrm{se}}N^{\mathrm{T}}\mathbf{t}\bullet\mathrm{d}s\right)_i$, and the damping matrix is approximately selected in sheet metal forming as $[C]^t=\alpha[m]^t$ and is added to the system. The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

above-mentioned equilibrium equation is solved at a time step, t, during the sheet forming to obtain the nodal displacement vector at next incremental time, $t+\Delta t$ using central difference scheme as shown by the following two Equations 7 and 8. This procedure is referred as *explicit method*.

$$\{X\}^t = \{\mathbf{F}\}^t - \left[[k]^t - 2\frac{[m]^t}{\Delta t^2}\right]\{u\}^t - \left[\frac{1}{\Delta t^2} - \frac{\alpha}{2\Delta t}\right][m]^t\{u\}^{t-\Delta t} \quad (7)$$

$$\left[\frac{1}{\Delta t^2} + \frac{\alpha}{2\Delta t}\right][m]^t\{u\}^{t+\Delta t} = \{X\}^t \quad (8)$$

However, the explicit method is only conditionally stable with a time increment less than critical value governed by Courant criterion $\Delta t_{\text{critical}} = \left(L_{\min}/\sqrt{E/\rho}\right)$, where $L_{\min}$ is the smallest element size, E the Young's modulus, and ρ the density of the sheet metal (Reddy, 2005; Worswick, 2002).

2.3.2 *Material models for sheet forming*

Material models for sheet metal forming processes that are used in FE simulations are the concept of yield surface, the flow rule, and the hardening rule. The convex yield surface is defined as a function, $f(\sigma_{ij}) = 0$, which encloses the elastic region in the stress space. If the stress state in the deforming body lies on the yield surface, then the material will start to yield. For a small change in stress state directing outside the yield surface, the plastic strain increment is governed by normality rule as Equation 9.

$$d\varepsilon_{ij} = d\lambda\frac{\partial f}{\partial \sigma_{ij}} \quad (9)$$

where $d\lambda$ is an arbitrary constant.

This rule also shows the direction of the plastic flow as shown in Figure 5. During sheet metal forming, the flow strength increases on plastic strain in most of the metal. In the FE analysis, this work hardening or strain hardening description should be incorporated mathematically. There are various phenomenological models and Hollomon's power law hardening expression commonly used for steel sheets is shown in Equation 10.

$$\overline{\sigma} = K\varepsilon^{-n} \quad (10)$$

where K and n are the strength coefficient and strain hardening exponent, respectively.

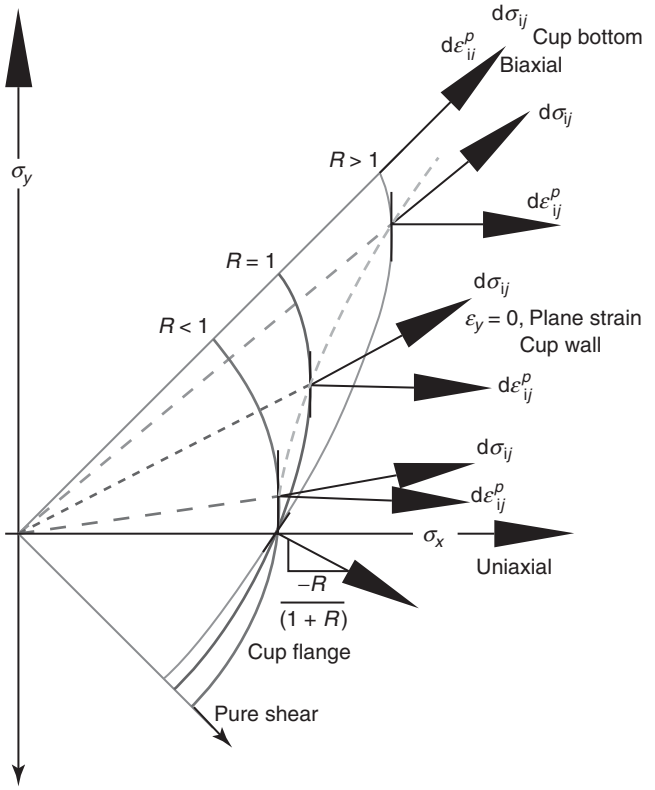


Figure 5. Schematic yield loci for Hill's and von-Mises's yield criterion with stress state of deep drawn cup.

The various yield criteria used extensively in numerical modeling of sheet metal forming are von-Mises, Hill's 48, Hosford, and Barlat89.

2.3.2.1 von-Mises yield function. The von-Mises yield criterion is one of the most widely accepted yield functions for isotropic materials. In sheet metal forming, plane stress condition along thickness direction (z-direction) is assumed while neglecting stresses in the plane of the sheet. The von-Mises yield functions for the plane stress condition is expressed as shown in Equation 11.

$$f = ({\sigma_x}^2 + {\sigma_y}^2 - \sigma_x\sigma_y) + 3{\sigma_{xy}}^2 - \sigma^{-2} = 0 \quad (11)$$

where σ_x and σ_y are the normal stresses, σ_{xy} shear stress, and $\overline{\sigma}$ is the effective stress of material. When $\overline{\sigma}$ reaches yield stress material starts to yield. The yield function is shown in graphical as in Figure 5.

2.3.2.2 Hill's 48 yield function. The most commonly used yield criterion for anisotropic sheet materials is the classical Hill's criterion proposed in 1948 (Hill, 1948), often referred as *Hill's 48*. It is expressed by a quadratic

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

function as Equation 12 for a plane stress condition,

$$2f = H(\sigma_x - \sigma_y)^2 + G{\sigma_x}^2 + F{\sigma_y}^2 + 2N{\sigma_{xy}}^2 - 1 = 0 \quad (12)$$

The relation between Hill's material anisotropy parameters H, G, F, and N can be expressed in terms of Lankford parameters R_0, R_{45}, R_{90} applying flow rule in the Hill's yield function as Equation 13.

$$\frac{H}{G} = R_0 \quad \frac{H}{F} = R_{90} \quad \frac{N}{F+G} - \frac{1}{2} = R_{45} \quad (13)$$

The Hill's 48 yield function can be further expressed as Equation 14.

$$2f = \frac{\begin{matrix} R_0R_{90}(\sigma_x-\sigma_y)^2+R_{90}{\sigma_x}^2+R_0{\sigma_y}^2 \\ +(R_0+R_{90})(2R_{45}+1){\sigma_{xy}}^2 \end{matrix}}{R_{90}(1+R_0)\sigma^{-2}} = 1 \quad (14)$$

The effect of normal anisotropy on the yield locus is shown in Figure 5.

2.3.2.3 Hosford yield function. Hosford proposed a nonquadratic yield function as shown in Equation 15 (Hosford, 1985).

$$f = \frac{1}{\{R_{90}(1+R_0)\}^{\frac{1}{a}}} \times [R_{90}|\sigma_x|^a + R_0|\sigma_y|^a + R_0R_{90}|\sigma_x - \sigma_y|^a]^{\frac{1}{a}} - \overline{\sigma} = 0 \quad (15)$$

It can be observed that the yield surface depends on the exponent a as shown in Figure 6. An important drawback of this criterion is caused by the lack of shear stress.

2.3.2.4 Barlat yield function. Barlat and Lian (1989) proposed a nonquadratic yield function, often referred as *Barlat89*, which would accurately predict both normal and planar anisotropic behavior of materials. Under conditions of plane stress, it has the nonquadratic form as shown in Equation 16.

$$a|K_1 + K_2|^M + a|K_1 - K_2|^M + c|2K_2|^M = 2\overline{\sigma}^M \quad (16)$$

where the parameters K_1 and K_2 are defined as Equation 17.

$$K_1 = \frac{\sigma_x + h\sigma_y}{2} K_2 = \sqrt{\left\{\frac{\sigma_x - h\sigma_y}{2}\right\} + p^2{\sigma^2}_{xy}} \quad (17)$$

where σ_x and $\sigma_y$are the normal components of stress in an arbitrary set of orthogonal directions and σ_{xy} represents the shear stress. On the basis of crystalline plasticity considerations for face-centered cubic (FCC) materials,

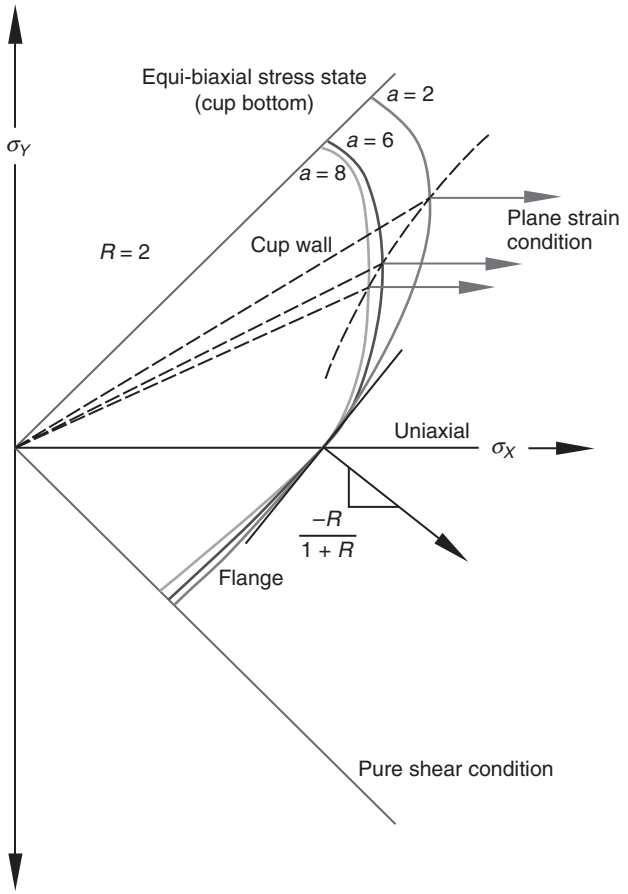


Figure 6. Schematic yield loci for Hosford's yield criterion with stress state of deep drawn cup.

Barlat determined that a value of $M = 8$ is appropriate, whereas for body-centered cubic (BCC) materials, a value of $M = 6$ is appropriate based on crystalline plasticity calculations. The material parameters, a, c, h, and p, can be determined using the Lankford coefficients, R_0, R_{45}, and R_{90}, with the help of Equations 18 and 19.

$$a = 2 - c = 2 - 2\sqrt{\left(\frac{R_0}{1+R_0}\right)\left(\frac{R_{90}}{1+R_{90}}\right)} \quad (18)$$

$$h = \sqrt{\left(\frac{R_0}{1+R_0}\right)\left(\frac{1+R_{90}}{R_{90}}\right)} \quad (19)$$

Figure 7 shows schematic drawing of Barlat's yield criterion along with the effect of M.

2.3.3 Failure criteria- forming limit diagram

Failures of sheet metal occur by strong local thinning/necking after diffuse necking. The deformation is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

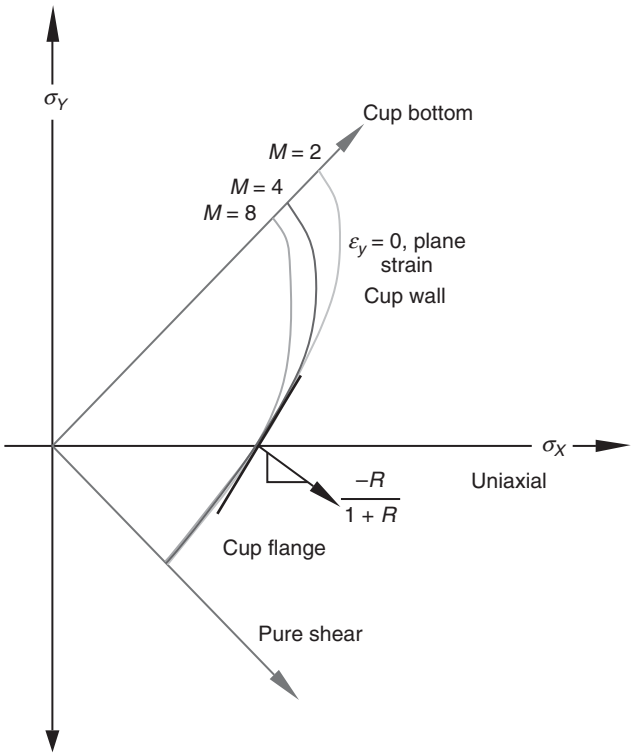


Figure 7. Schematic yield loci for Barlat's yield criterion with stress state of deep drawn cup.

confined to the localized necking region without further contraction of width of specimen and this region leads to tearing of sheet metal on further deformation. This localized necking is a visible indication that the sheet metal has been worked beyond its prevailing limit. Whether or not a particular sheet metal can be formed without failure depends on many factors such as material properties, surface conditions, blank size and shape, lubrication, press speed, blank holder pressure, and punch and die design. Because of this, source of flow localization problem is difficult to pinpoint and quantify. In the past a few decades, the technique of analyzing failures in sheet metal stampings through the use of circle grids and forming limit diagrams (FLDs) has gained considerable importance. The forming limits, determined by elaborate laboratory experiments, are typically represented in terms of limiting principal strains, known as the *FLD*. The diagram was first constructed by Keeler and Goodwin, following the concept of Keeler and Backofen (1963). The FLD represents the onset of localized (visible) necking over all possible combinations of strains in the plane of a sheet. Figure 8 shows the "Keeler–Goodwin" curve that characterizes the forming limits of a 1 mm thick low carbon steel sheet. This diagram describes the combinations of major and minor principal surface strains at which a highly localized zone of thinning or necking becomes visible on the surface of the sheet metals. The shaded area represents the scatter of results between strain combinations that are safe and those which cause failure. The lowest surface strain occurs at plane strain tension. The detailed procedure for determination of the FLD has been explained in several publications (Hecker, 1975; Hiam and Lee, 1978). After the initial work of Keeler and Goodwin, a large number of investigations (Ghosh and Hecker, 1975; Coubrough, Matlock, and Van Tyne, 1993; Talyan, Wagoner, and Lee, 1998) were carried out for experimental determination of FLDs of various materials such as deep DQ steel, EDD quality steel, HSLA steel, and stainless steel. Many theoretical models have been proposed to predict FLDs of sheet metals. Marciniak and Kuczynski (1967) developed a theoretical model for predicting the FLD based on the assumption of existence of an initial nonhomogeneity of the material in the form of a groove. This model was formulated for sheet metals subjected to biaxial tension, when the ratio of principal stresses lies in the range 0.5–1.0. This model was extended by Sowerby and Duncan (1971) to cover all cases of biaxial tension. Many more theoretical models were subsequently formulated based on the M–K model to predict the FLDs (Tadros and Mellor, 1975; Korhonen, 1978; Lu and Lee, 1987; Barlat and Lian 1989). The effects of strain hardening exponent (n), strain rate exponent (m), and strain ratio on FLD were analyzed by some researchers (Marciniak, Kuczynski, and Pokora, 1973; Graf and Hosford, 1990). Some approaches that are not based M–K hypothesis have also been proposed to predict limit strains (Jones and Gillis, 1984; Choi, Gillis, and Jones, 1989; Pishbin and Gillis, 1992). However, there has always been some deviation in the theoretically predicted FLDs from the experimental curves because of a variety of reasons.

To analyze sheet metal forming, Chow, Jie, and Hu (2003) determine the forming limit by applying Hosford nonquadratic yield criterion and the concepts of plasticity theory on Storen–Rice vortex theory (Storen and Rice, 1975). The details of the formulation of this theoretical FLD damage model are published elsewhere (Chow, Jie, and Hu, 2003). This damage model was also applied successfully in prediction of FLD of TWBs. In FLD, the limiting curve intersects the Y-axis (major strain axis or the plane strain line) by an intercept called FLD_0. Despite all the recent enhancements of the computational models, they are not able to give very accurate predictions of the limit strains in all the cases encountered in practical applications (different materials, thickness, forming rates, temperatures, strain paths, etc.). Owing to this fact, the commercial finite element codes still make use of experimental FLDs or

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

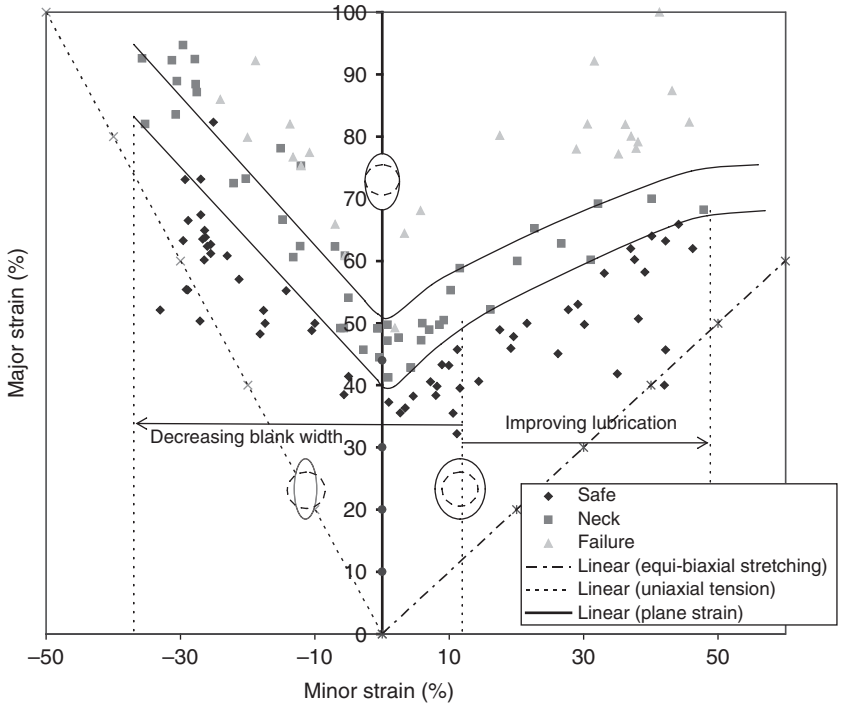


Figure 8. Forming limit diagram of a low carbon steel sheet showing different strain paths.

calculated FLDs with simple semiempirical models. Keeler and Brazier (1977) proposed an empirical relationship for calculating the limit strains corresponding to plane strain, FLD_0, using thickness and n-value of the sheet material. Equations 20, 21, and 22 are listed (all strain in the equation is true strain).

$$\mathrm{FLD}_0^{\mathrm{True}} = \ln\left[1 + (0.233 + 0.413t)\,\frac{n}{t}\right] \tag{20}$$

$$\varepsilon_1 = \mathrm{FLD}_0^{\mathrm{True}} - \varepsilon_{2,} \text{ while } \varepsilon_2 < 0 \tag{21}$$

$$\varepsilon_1 = \ln\left[0.6\left(\exp\left(\varepsilon_2\right) - 1\right) + \exp\left(\mathrm{FLC}_0^{\mathrm{True}}\right)\right] \tag{22}$$

where t is sheet thickness in inches and n strain hardening coefficient of the sheet.

Comparison of experimental and theoretical FLDs for HSLA, DP980, and DP600 steel sheets is shown in Figure 9. Experimental limit strains of HSLA sheet are close enough to Keeler–Brazier FLC; however, limit strains of DP980 and DP600 are matching well with the strains predicted by Storen–Rice vortex theory using Hosford yield criterion for $a = 6$ and 4, respectively. The near plane strain conditions show difference as much as 16%, but they are close to experimental strains in other regions. It can also be observed that yield criteria do not affect the forming limits in the tension–compression region, but it has remarkable influence in the tension–tension region of FLC. Moreover, as the order of Hosford yield criterion increases, the predicted limit strains decreases in the tension–tension region (right-hand side) of FLC. Suitable theoratical model of FLD has to be incorporated as a damage model in FEM. The calculated major and minor strain data points at each step of the deformation are superimposed on the FLD curve. The cup failure takes place at the cup depth when the strain state falls above the FLD while evaluating the formability.

2.3.4 *Prediction of limiting drawing ratio and earring*

In circular cup deep drawing, the method of determining LDR is well established. However, the LDR while drawing of noncircular parts can be calculated indirectly by replacing the punch and blank area by circles of equal size and calculating their respective equivalent diameter (Equation 23).

$$D_{\mathrm{p}} = 2\sqrt{\left(\frac{A_{\mathrm{P}}}{\pi}\right)}$$

$$D_{\mathrm{o\,max}} = 2\sqrt{\left(\frac{A\mathrm{o}}{\pi}\right)} \tag{23}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

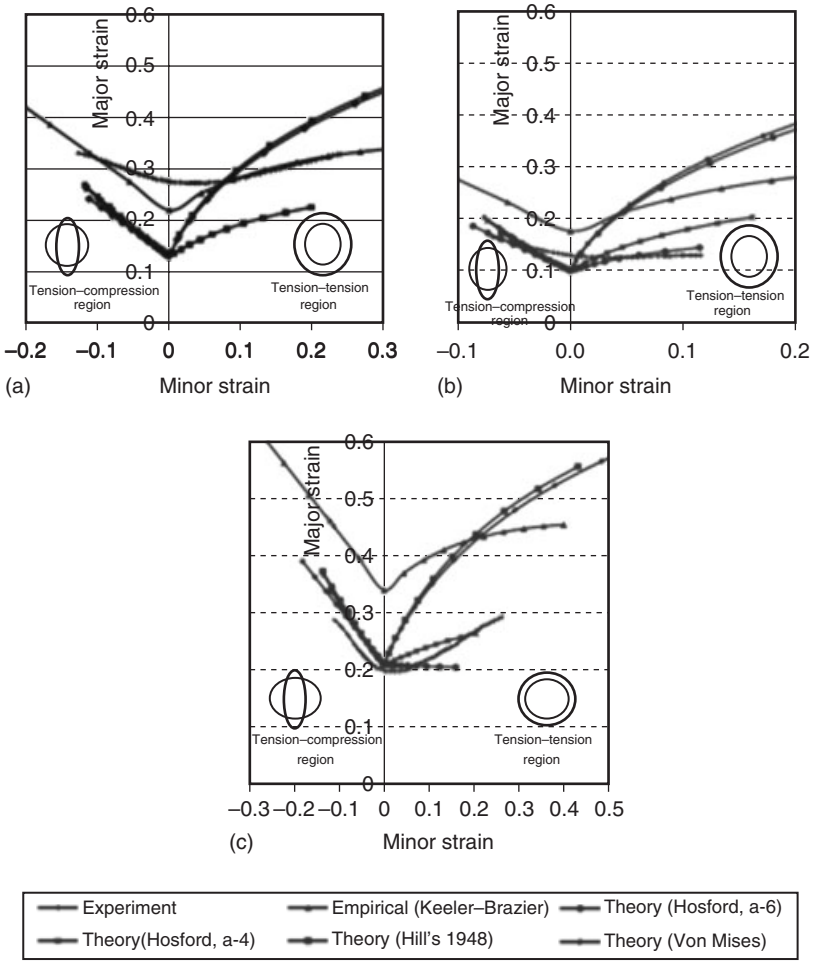


Figure 9. Comparison of experimental and theoretical forming limit curves: (a) HSLA, (b) DP980, and (c) DP600.

The DR can be calculated as for circular components by Equation 24.

$$\mathrm{DR} = \left(\frac{D_{\mathrm{o\,max}}}{D_{\mathrm{p}}}\right) \tag{24}$$

The LDR can be determined by two commonly used analytical relations namely Whiteley's (1960) and Leu's (1997) formula. Whiteley proposed the Equation 25 to predict LDR.

$$\mathrm{LDR} = \mathrm{e}^{\mathrm{f}\sqrt{(1+\overline{R})}/2} \tag{25}$$

According to this formula, the LDR increases with increasing $\overline{R}$ value. The formula proposed by Leu (Equation 26) considers the effect of both n and $\overline{R}$ value.

$$\mathrm{LDR} = \sqrt{\{\mathrm{e}^{2\mathrm{f}\,\mathrm{e}^{-n}\sqrt{(1+\overline{R})/2}}\} + \{\mathrm{e}^{2n}\sqrt{((1+\overline{R})/2)}\}} \tag{26}$$

The theoretical LDR from both the methods was calculated on the assumption that the drawing efficiency (f) is 0.9. The LDR obtained from FE simulations using Hill's, von-Mises's, and Barlat's yield criterion and analytical relationships was compared with the experimental results as shown in Figure 10. The LDR predicted by Barlat's yield criterion is very close to experimental LDR in both cylindrical and square box drawing cases (2.3% and 0.9% deviation). Whiteley's formula over predicted the LDR for both cases. However, Leu's formula over predicted the LDR in case of cylindrical cup and under predicted the LDR of square cup. These analytical relations do not consider design parameters and blank geometry and hence LDR predictions deviate from experimental results. The von-Mises's yield criterion, which is an isotropic yield criterion, does not consider the influence of normal anisotropy parameter while evaluating LDR. For both cases, Hill's yield criterion predicted higher LDR than experiment as the cup wall strength increased during cup drawing following plane strain deformation path. The increase in cup wall strength during plane strain deformation is predicted to be higher by

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

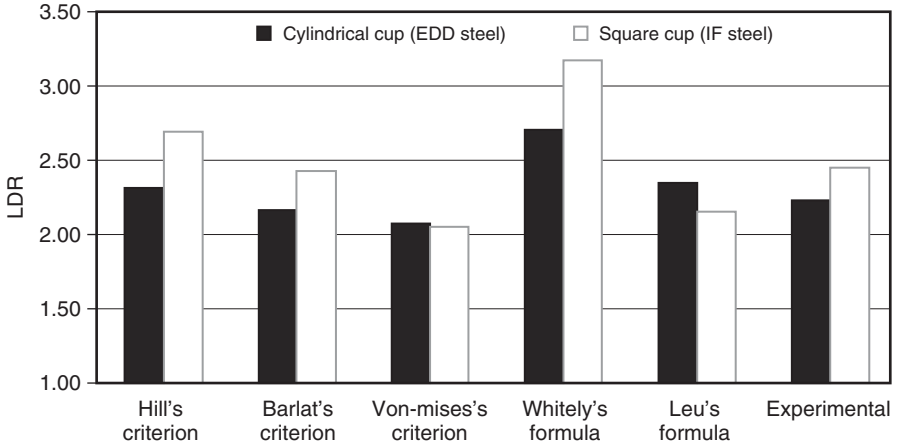


Figure 10. Comparison of LDR obtained through different methods for cylindrical cup drawing (EDD steel) and square cup (IF steel).

Hill's criterion and hence higher LDR is predicted than the Barlat's yield criterion (Barlat and Lian, 1989).

The EDD and IF steels both have high normal anisotropy $\overline{R}$ and planar anisotropy ΔR values, so significant earring was observed in the drawn cups. It can be seen that four ears have formed in case of cylindrical cup, at 0° and 90° to the RD because the ΔR-value of this material is positive. In FE-simulated cups, ears were formed in case of Barlat's and Hill's yield criteria because these criteria incorporated both normal and planar anisotropy. However, earring was not observed in case of von-Mises's yield criterion, which is an isotropic yield criterion (Figure 11). The cup is divided into angles with respect to RD to determine the cup height and the earring profile obtained for cylindrical cup from the FE simulations using Hill's and Barlat's yield criterion was compared with the experimental profile (Kishore and Ravi Kumar, 2002) as shown in Figure 12. The cup height of the experimentally drawn cup was nonuniform because of eccentricity in placing the blank holder on the blank and because of lack of exact centering of the blank. The cup height predicted by the FE simulations at points of ear formation is lower as compared to the actual cup height at ears. However, it is very difficult to point out the yield criterion that is predicting the ear profile in a better manner. Hence, the percentage ear height with respect to minimum cup height was determined and compared as shown in Figure 13. It was observed that the Hill's yield criterion predicted the ear profile more accurately. It was calculated that the maximum percentage ear height was 14.22% in the experiment and 14.27% in the FE-simulated cup using Hill's yield criterion, respectively. However, the maximum percentage ear height was 6.83% for the cup drawn using Barlat's yield criterion. Four ears have formed in case of square cup at the four corners representing 45° and 135° to the RD. The blank material available on die opening offers less resistance at side edge compared to diagonal direction as more material is available to deform in diagonal direction. The square cup obtained from the FE simulations using the three yield criteria has predicted earring in the cups. It is observed from the Figure 11 that square cup is having ears in spite of using von-Mises's yield criterion in FE simulations. Similar to the cylindrical cup drawing case percentage ear height with respect to minimum cup height has been determined for all the three cases and compared. It is observed that FE simulations are predicting similar ear height (as well as percentage ear height) with both Barlat's and Hill's yield criterion (Figure 14). Hence, it can be concluded that the geometry of tooling is mainly responsible for ear formation in square cups.

Numerical simulation was extensively studied incorporating different die radius, punch corner radius, and coefficient of friction between tooling for various steel sheet materials such as EDD, IF, HSLA, DP, DQ, and DDQ. It was observed that LDR increased with increase in punch corner radius and die corner radius, but the increase was more significant in case of die corner radius. The LDR decreased linearly with increase in friction. The LDR varied significantly with $\overline{R}$-value but strain hardening exponent had a negligible influence.

2.3.5 *Modification of blank shape to minimize earring*

Earring is highly undesirable as it adds on an additional processing step "trimming" and also metal representing ear will undergo deformation that demands extra load and work. One case study of modification of initial circular blank shape (blank diameter 82 mm) is studied in this

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

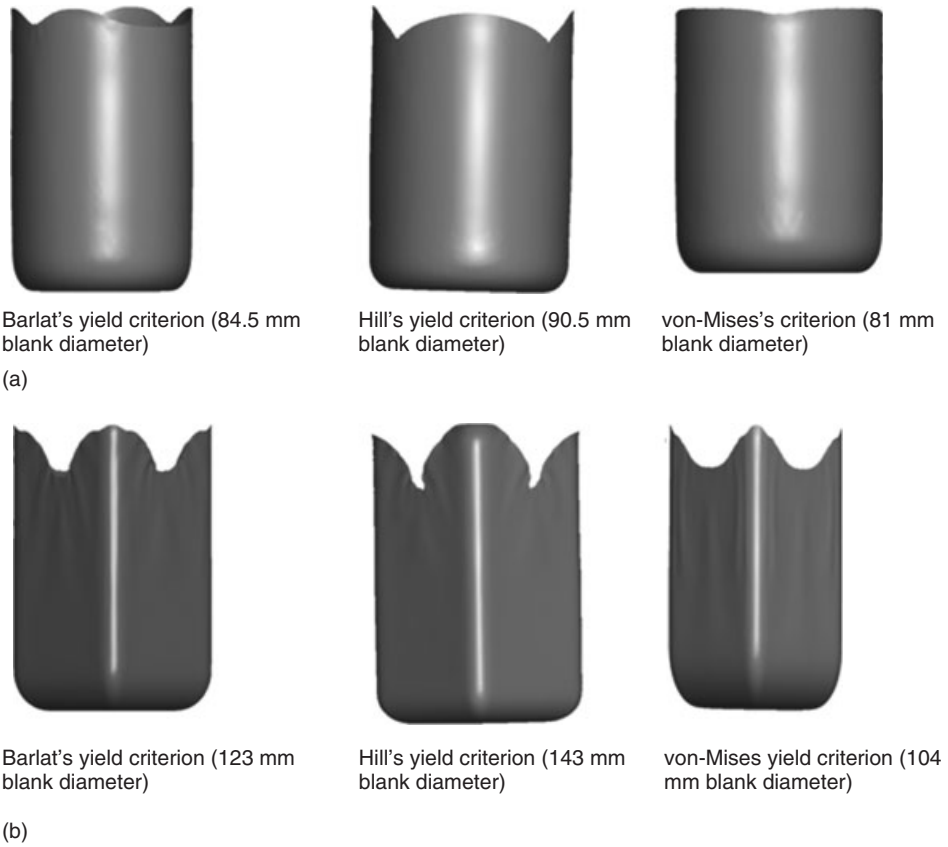


Figure 11. Earring defects predicted from FE simulations incorporating various yield criteria: (a) cylindrical cups from various initial circular blank diameter (EDD steel) and (b) square cups from various initial circular blank diameter (IF steel).

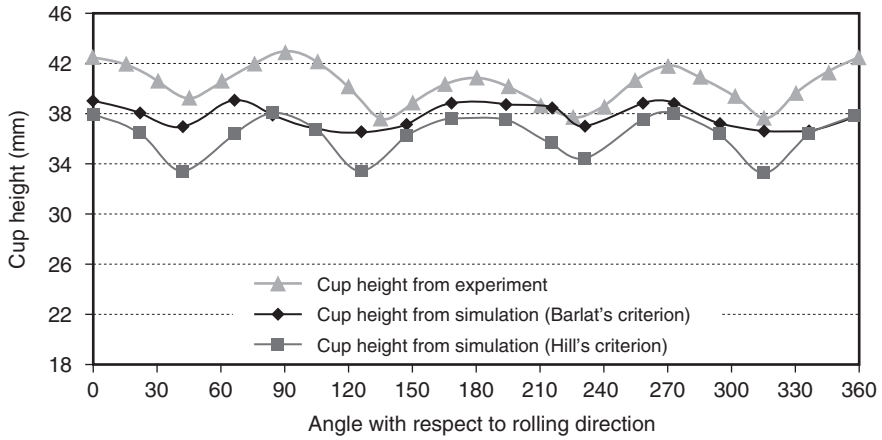


Figure 12. Comparison of cup height for a DR of 2.16 (84.5 mm blank diameter). (Experimental curve data taken from Kishore and Ravi Kumar, 2002.)

section. The modified X- and Y-coordinates were done as per the Equations 27 and 28.

$$\text{Modified X-coordinate} = R - \beta\Delta r_1 \qquad (27)$$

$$\text{Modified } Y\text{-coordinate} = R - \beta\Delta r_2 \qquad (28)$$

where R is the initial radius of the circular blank, $\Delta r_1 = R_0 - R_{45}$ and $\Delta r_2 = R_{90} - R_{45}$.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

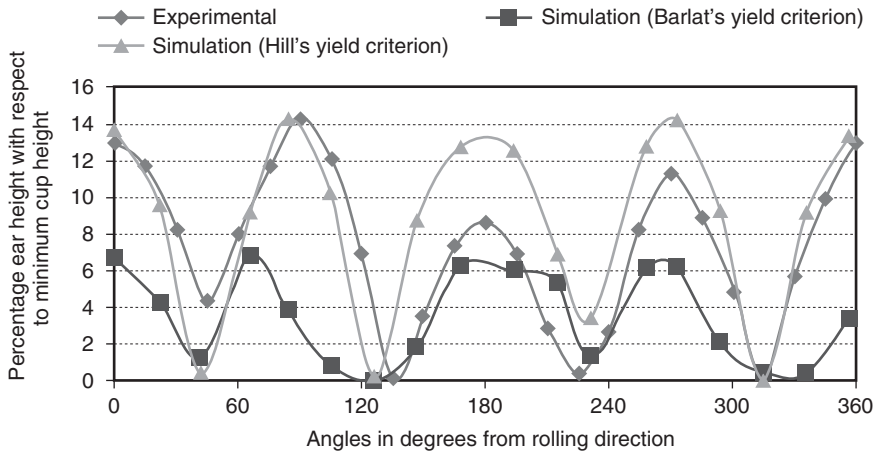


Figure 13. Comparison of percentage ear height above the minimum cup height based on experiment and yield criteria.

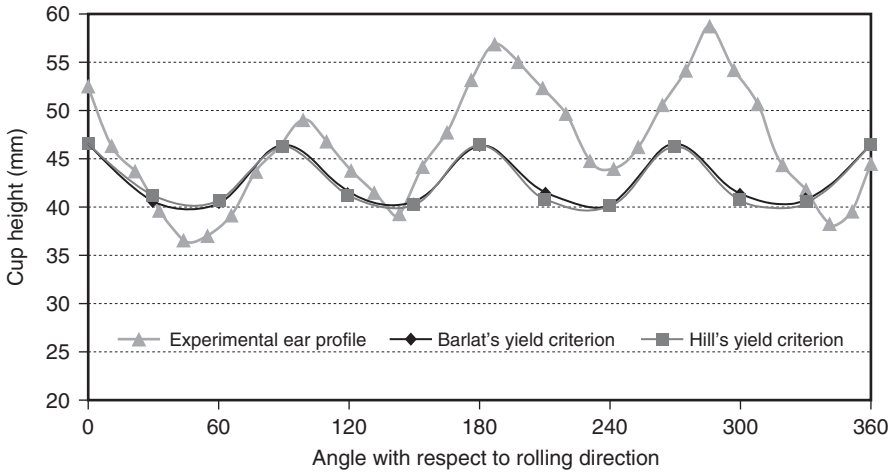


Figure 14. Comparison of cup height for a DR of 2.127 (108 mm blank diameter).

The detailed procedure was described elsewhere (Kishore and Ravi Kumar, 2002). Figure 15 shows the initial blank shape and modified shape of the initial blank using different $\beta = 3$, 6, and 8 values. Figure 16 shows the top view of the simulated cup at the stage when the ear touches the die corner radius. It was observed at this stage that the ear height in case of modified blanks was less than the initial circular blank shape. The completely drawn cups without and with modification are shown in Figure 17. It can be observed from figures that the ear height is decreasing in the case of modified blank after they have been completely drawn. The cup height reduced by 22.37% in the case of modified blank 1, 62.86% in modified blank 2, and 51.23% in the case of modified blank 3 as compared to initial circular blank. Comparison of cup height of conventional circular blank and modified blank shape is shown in Figure 18. The percentage ear height was determined in the case of initial circular blank and modified blank using approach 2 (Figure 19). It shows that use of modified blanks instead of circular blank while drawing cylindrical cups can save material during trimming operation. It was found from deep drawing experiments of circular and modified blank that the maximum load required to deform decrease by 7.0% because of the elimination of ears. It leads to less power consumption during deformation of modified blank that may increase cost effectiveness of the process. There are various theoretical techniques to estimate the load and power consumptions in deep drawing processes such as slab analysis (Hosford and Caddell, 2007), FEM (Worswick, 2002), upper bound technique (Agrawal, Reddy, and Dixit, 2008), and slip line field theory (Hosford and Caddell, 2007). However,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

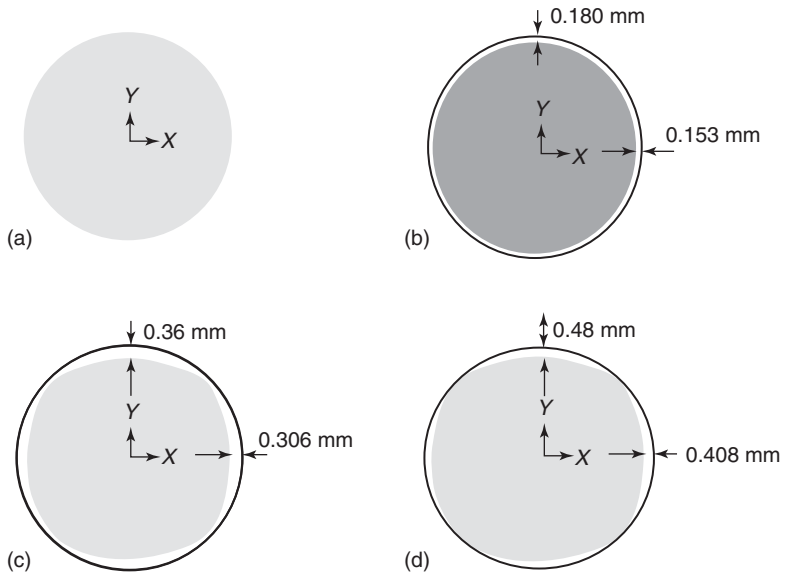


Figure 15. Different modified initial blank shapes corresponding to different values of β. (a) Initial blank, (b) modified blank 1, (c) modified blank 2, and (d) modified blank 3.

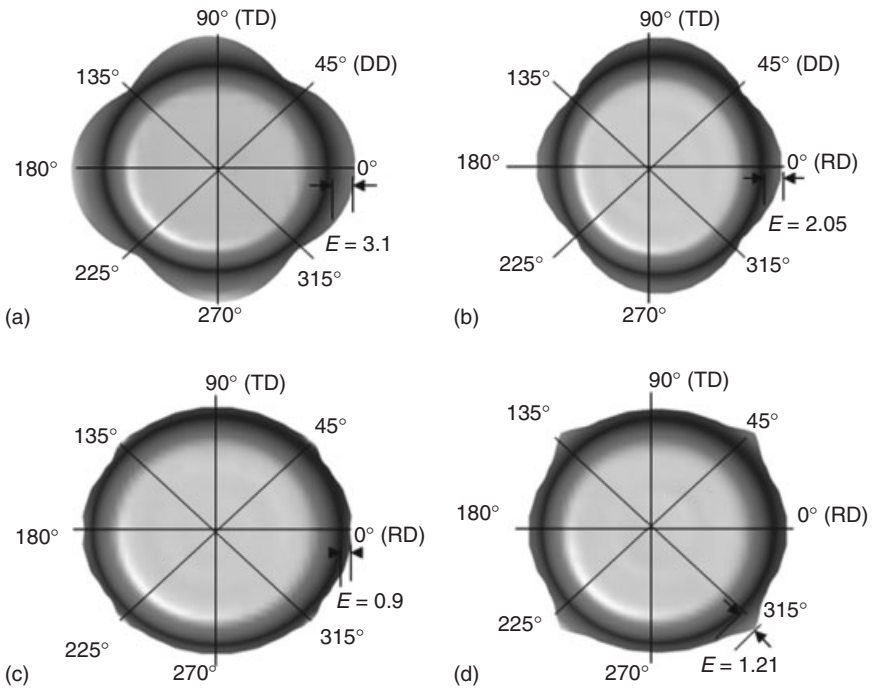


Figure 16. Drawn cups at the stage when the ears are just touching the die corner. (a) Initial blank, (b) modified blank 1, (c) modified blank 2, and (d) modified blank 3.

the details of these techniques are not within the scope of this chapter. To completely eliminate the waviness at the cup top need further optimization of the blank shape. This method of blank modification can be extended to other initial blank shapes for minimizing earring and wastage of scrap during automotive component stamping.

After finalizing the two-dimensional arbitrary shape to minimize the earring, the required shape has to be blanked

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

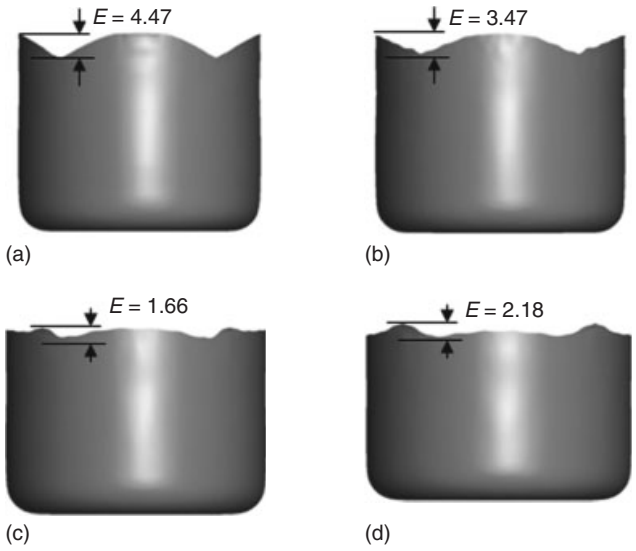


Figure 17. Completely drawn cups from circular and modified blanks. (a) Circular blank, (b) modified blank 1, (c) modified blank 2, and (d) modified blank 3.

out from the coil to reduce the amount of scrap. Owing to rapid development of CAD/CAM technologies, soft computing expert systems and optimization techniques, efficient methods of blank layout, or nesting from the strip or coil can be planned using progressive dies (Prasad, 1994; Ghatrehnaby and Arezoo, 2009). It was reported that approximately 30–50% of sheet materials procured is wasted as scrap. Proper utilization of the materials purchased is required to decrease the cost of the part produced, as only material cost shares 70% of the total cost (Ghatrehnaby and Arezoo, 2009). However, the factors to be considered in nesting to minimize the scrap are crystallographic orientation in the sheet (i.e., directionality of properties), direction of the burr, type of stock (strip or coil), type of press (mechanical or hydraulic), die cost, and size, shape, and thickness of the blank. TWB technology is also a best practice to reduce scrap production as the scraps can be utilized by welding and forming components for different applications.

2.3.6 *Deep drawing through tractrix die*

While deep drawing through conventional die using blank holder, the blank is bent through 90° as it passes over the die entry radius and then has to unbend as it straightened out to form cylindrical wall. This bending and unbending considerably exhaust ductility of the metal during deep drawing. Hence, contoured die was proposed by various researchers (Al-Makky and Woo, 1980; Narayanasamy and Sowerby, 1995) to increase the DR while eliminating the blank holder. The deep drawing can be possible in a single action press reducing the cost of the equipments and toolings. The die is referred as *tractrix die* if the die contour profile is a tractrix profile as shown in Figure 20a. The tractrix profile satisfying Equation 29 ensures only blank edge contact between blank and die during cup drawing reducing the frictional force. Bending occurs around the punch corner without any unbending action during deep drawing through tractrix die and hence a bigger cup can be drawn. It is found in practice that the exact tractrix die contour is difficult to fabricate, which will always give edge contact. Hence, modified tractrix die profile with a conical surface was proposed as shown in Figure 20b. It was observed that a sheet metal having very low normal anisotropy parameter, $R = 0.8$ had an LDR of 1.73 in

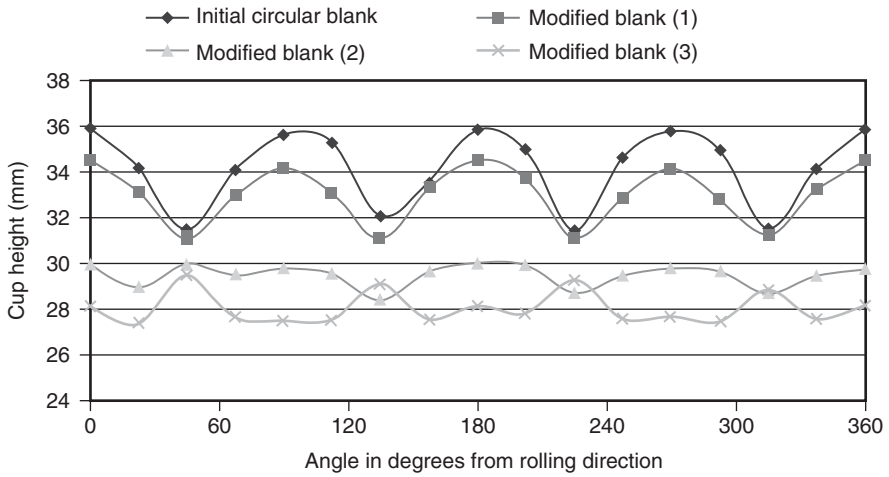


Figure 18. Comparison of cup height before and after modification of blank shape for DR of 2.1.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

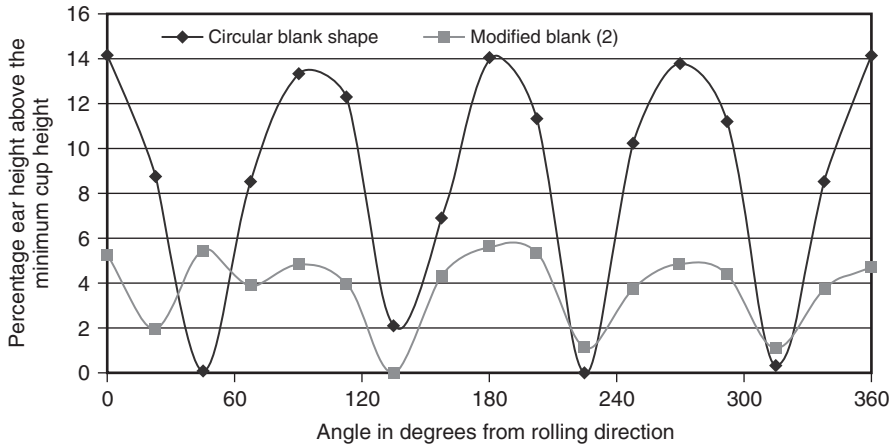
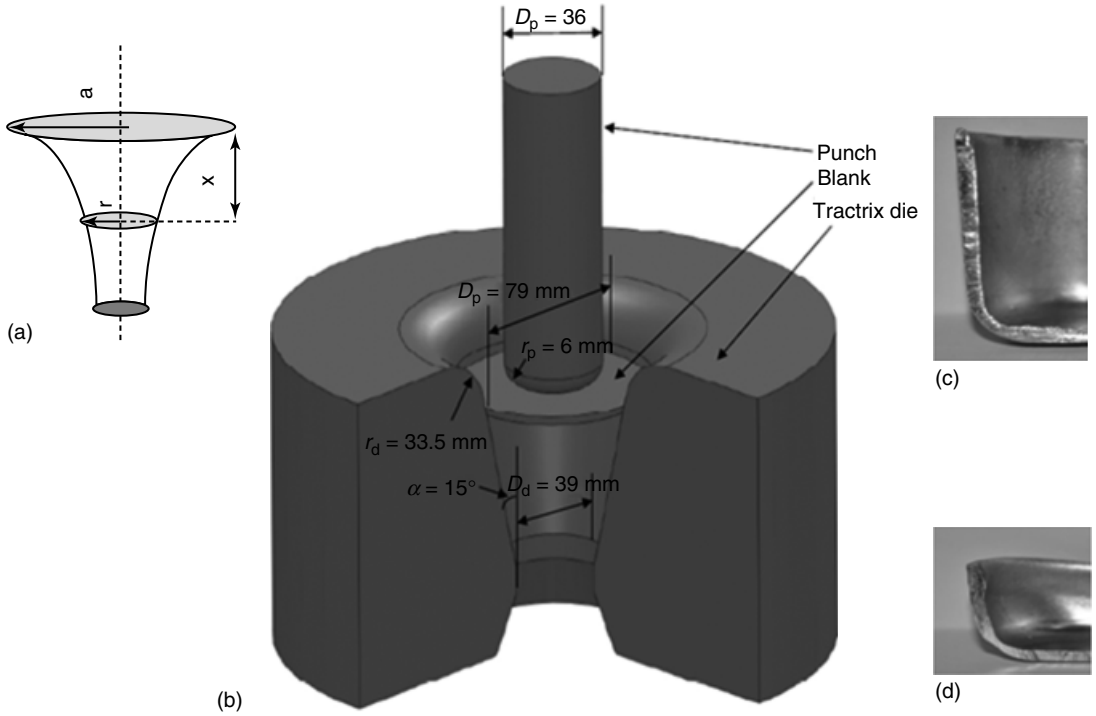


Figure 19. Comparison of percentage ear height above the minimum cup height based on modified and circular blank.



D_p = Punch diameter, D_d = dianmeter, D_b = Blank diameter, r_d = Die corner radius, r_p = Punch corner radius, α = Die angle.

Figure 20. (a–d) Modified tractrix die profile with a conical surface along with drawn cups.

conventional deep drawing, and the LDR was increased to 2.36 in tractrix die. The cross section of the maximum cups drawn by conventional die and tractrix die is shown in Figure 20d and c. The comparison of thickness profile obtained from experiment and numerical simulation is shown in Figure 21. However, it was observed that the sheet thickness and clearance plays a role for successful drawn cup. The drawn cup usually fails due to wrinkling and fracture. The safe working zone for a 75 mm circular cup drawing is shown in Figure 22. It can be very well

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

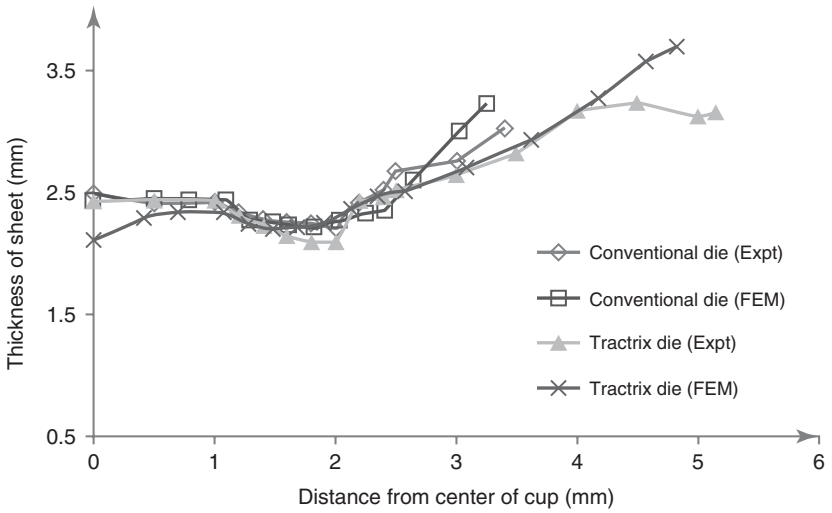


Figure 21. Comparison of thickness profile of deep drawn cup through conventional and tractrix die.

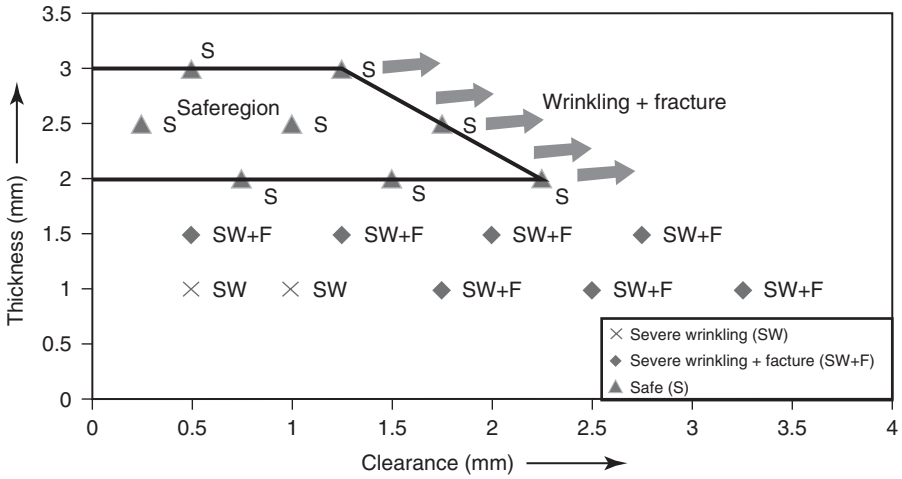


Figure 22. Safe working region in modified tractrix die.

concluded that the tractrix die is very much suitable for deeper cup drawing of thicker sheet metals.

$$x = a \times \ln\left(\frac{\sqrt{a^2 - r^2}}{r}\right) - \sqrt{a^2 - r^2} \qquad (29)$$

3 FORMABILITY OF TAILOR-WELDED BLANK

3.1 Tailor-welded blank fabrication-laser welding

The various fusion and solid-state welding techniques adopted for fabrication of TWBs are tungsten inert gas (TIG), metal inert gas (MIG), electron beam welding (EBW), laser beam welding (LBW), and friction stir welding (FSW). Among these processes, laser welding of steel sheets is frequently used by the industries because of narrow weld and HAZ (heat-affected zone) width with higher penetration. The different lasers adapted for welding processes are CO_2 laser, YAG (yttrium aluminum garnet) laser, diode laser, and fiber laser. Laser spot size in YAG laser and fiber laser is in micrometer level and hence is capable of producing very narrow weld width. Weld quality should be good and free of imperfection/defects as it may significantly affect postweld forming process. Some common welding defects are lack of fusion (due to misalignment of beam), lack of penetration (due to insufficient energy density of the laser source), concavity (due to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Weld width and maximum weld hardness in different laser-welded sheet metals.

Welded Blanks	Power (kW)	Scan Speed (mm min^{-1})	Weld Width (mm)	Maximum Weld Hardness (VHN)
IF–IF(both 1.0 mm)	1.8	6000	0.49	187
IFHS–IFHS (both 0.7 mm)	1.5	6000	0.62	227
HSLA–HSLA (both 1.0 mm)	1.8	6000	0.68	349
IF–IFHS (1.0 mm	1.6	6000	0.62	173
IF1.0–HSLA1.0	1.8	6000	0.72	352
DP980–DP980	1.8	1000	1.50	452
DP600–DP600	1.8	1000	1.25	398

lack of filling material in the weld zone), undercut (due to high surface tension of weld pool), porosity (due to evolution of absorbed gas from molten metal), and misalignment of the sheet metals to be welded. TWBs with above weld defects fail in the weld region during sheet-forming processes leading to drastic decrease in formability. Weld quality in TWBs is generally tested by visual inspection, observing weld macrostructure, microhardness test, tensile test of transverse weld specimen, and Erichsen cupping test. Successful fiber laser welding parameters for fabrication of TWBs are given in Table 4 for various automotive steel sheet grades. The weld width and hardness depends on laser power, scan speed, and the sheet material. The influence of sheet material chemistry in terms of CE on average weld hardness and relative formability is shown in Figure 23. The relative formability is calculated in terms of ratio between the Erichsen cup height of the laser-welded blank and corresponding parent metal. It can be observed that increase in CE from 0.002 to 0.41 in steel sheets decreases the Erichsen cup height of laser-welded blanks to 57% of that of parent metal. Hence, influence of weld zone in formability is significant at higher steel chemistry because of increase in weld hardness.

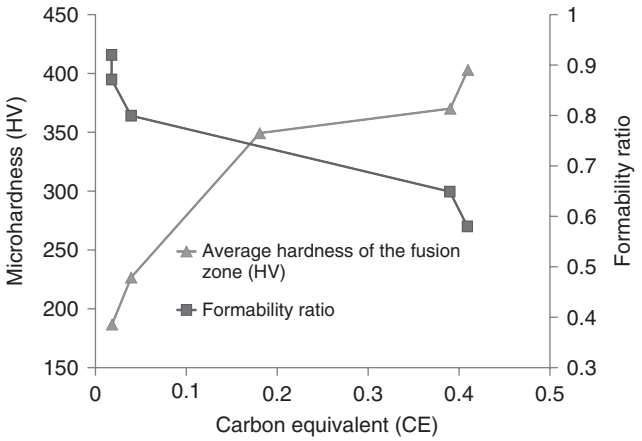


Figure 23. Dependence of hardness of the fusion zone and the relative formability of the LWBs on carbon equivalent.

The weld macrostructure and microhardness profile across the weld for DP980 and DP600 laser-welded specimen are shown in Figures 24 and 25. There are various observations those can be made for these weld. The first is that the weld fusion zone (FZ) is narrow and has excellent penetration into both the sides of sheet metal. The hardness (Figure 24) is observed to decrease in the outer HAZ region compared to parent DP980 metal and this is referred as *soft zone* (*SZ*). The peak temperature experienced by this region is below the AC_1 temperature, and hence, austenitization transformation is not possible in this region. However, tempering of preexisting martensite occurred during laser welding and 25% reduction of hardness was observed in this region. The SZ was not observed in the DP600 steel (Figure 25). However, hardness gradually increases toward the HAZ closer to the FZ. This inner HAZ region has experienced recrystallization and maximum hardness was recorded near the boundary of HAZ and FZ. FZ and inner HAZ are referred as *hard zone* (*HZ*). The presence of SZ is responsible for decrease in ductility and strength of DP980 laser-welded sheet because of strain localization and premature failure of the DP980 laser-welded specimen. This observation is clearly shown in Figures 26 and 27. The weld HZ and SZ significantly affect the formability. The influence of weld zone on formability can be characterized by standard Erichsen cup test. It was observed that DP600 laser-welded blanks showed lower cup height than that of parent metal because of reduced ductility of the hard weld zone and 39% reduction in cup height was observed for DP980 laser-welded blanks because of strain localization in the SZ during deformation. Hence, the presence of SZ is detrimental in forming of welded DP steels. However, TWB sample of IF–IFHS and IF–HSLA failure occurred in the weaker or thinner side justifying good quality weld (Figure 28). The fracture location depends on difference in properties and thickness, but the latter plays more critical role. The nonuniform deformation leads to decrease in ductility of the TWBs as shown in Figure 29.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

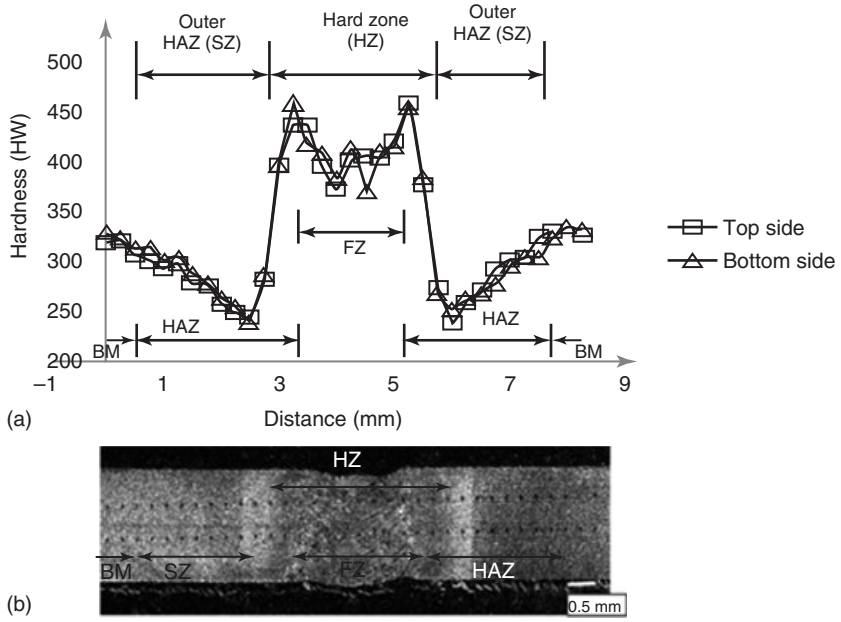


Figure 24. (a,b) Crossweld microhardness of laser-welded DP980 with weld macrograph.

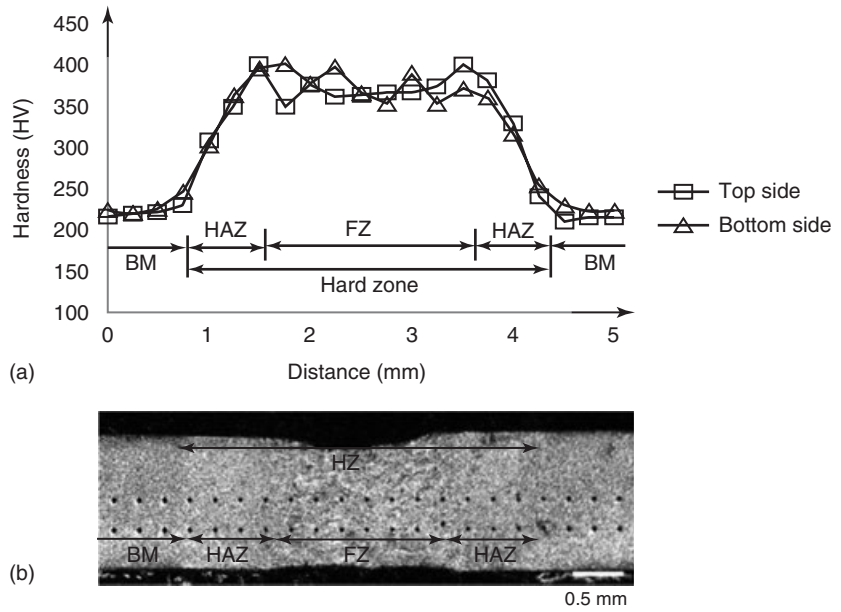


Figure 25. (a,b) Crossweld microhardness of laser-welded DP600 with weld macrograph.

3.2 Deep drawing behavior of tailor-welded blank

Cylindrical flange drawing operation was carried out by a 36 mm punch and the cup height just at the start of necking for 80 mm circular blank was considered as the measure of deep drawability. The drawn cups of the parent metals and welded blanks are shown in Figure 30. The IF steel has higher drawability with a cup height of 22.25 mm compared to 13.33 mm of HSLA sheet. This is due to higher resistance to thinning of IF steel as it has a higher anisotropy parameter (R-value) compared to that of HSLA. The reduction in cup depth in all the similar laser-welded blanks compared to the respective parent metals is about

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

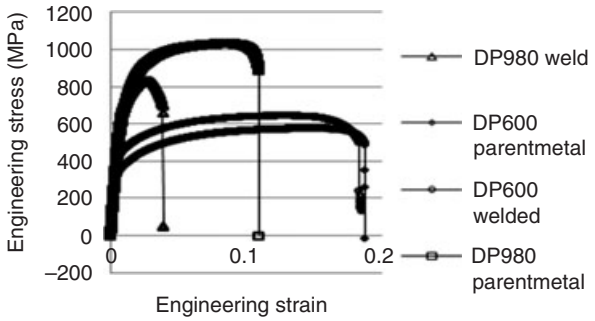


Figure 26. Engineering stress–strain diagram of base metals and AHSS laser-welded samples.

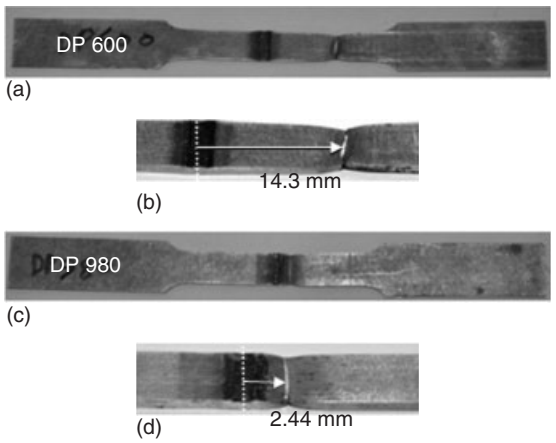


Figure 27. (a–d) Fractured tensile testing samples of AHSS laser-welded blanks.

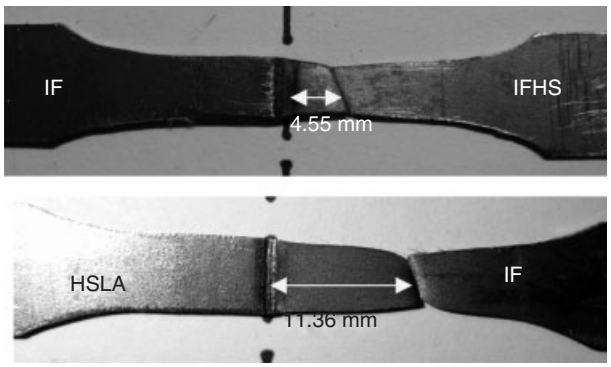


Figure 28. Transverse tensile tested tailor-welded blanks with fracture location from weld in millimeter.

5–7%. This is due to negligible influence of a very narrow weld zone on the drawability of the selected laser-welded sheets, and similar observation was also during uniaxial transverse tensile testing.

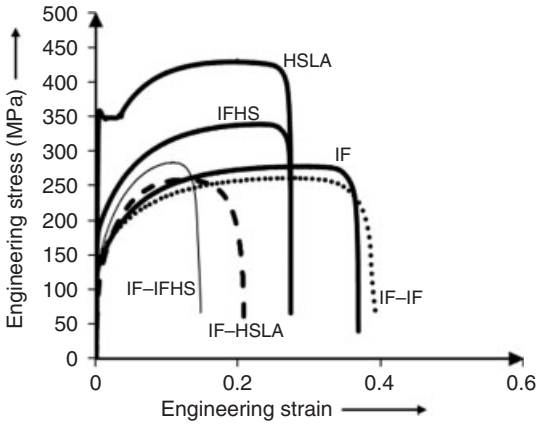


Figure 29. Engineering stress–strain curves of parent metals, IF–IF similar welded specimen and TWBs.

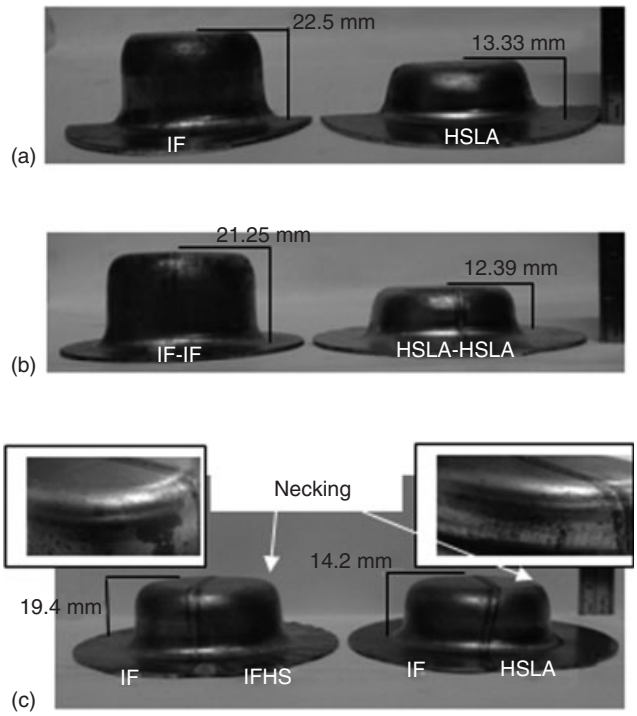


Figure 30. Deep drawn cups of parent metals and laser-welded blanks with cup height in millimeter: (a) cups of IF and IFHS parent metal, (b) cups of IF–IF and HSLA–HSLA similar welded blank, and (c) cups of TWBs with fracture locations.

3.2.1 Drawability and failure location of TWBs

Figure 30c shows comparison of failure locations in the drawn TWBs. In the IF–IFHS TWB with thickness ratio of 1.43, the cup depth was found to be 19.45 mm, which was more than the cup depth of the thinner parent metal

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

IFHS (18.25 mm) but lower than the thicker counterpart IF (22.5 mm). In the IF–HSLA TWB having strength ratio 1.56, the cup depth was found to be 14.20 mm, which is higher than the cup depth of HSLA parent metal (13.33 mm) but very much lower compared to IF parent metal. Hence, drawability depends on the material combination in TWBs. IF–IFHS TWB was found to have necking in the thinner IFHS side. However, in IF–HSLA TWB, necking was found to be initiating at the stronger HSLA side, which deviates from the fracture location in tensile testing (Figure 28). This is due to lower anisotropy parameter, *R*-value of HSLA, and hence it offers less resistance to thinning during deep drawing of IF–HSLA TWB leading to failure in HSLA side. Simulation results were found to agree very well with experimental cup height and failure location.

3.2.2 *Weld line movement and thickness distribution*

Weldline was found to move toward stronger or thicker side at the center of the drawn TWB cups. A maximum weldline movement of 2.57 mm was observed toward HSLA side at the cup center because of resistance from stronger side while drawing IF–HSLA TWB into the die cavity. In addition, the weldline moved by 4.02 mm in opposite direction, that is, toward the IF (weaker side) at the outer most cup flange. However, very negligible weld movement was observed in IF–IFHS TWB compared to IF–HSLA TWB. Figure 31 shows the comparison of weldline profile

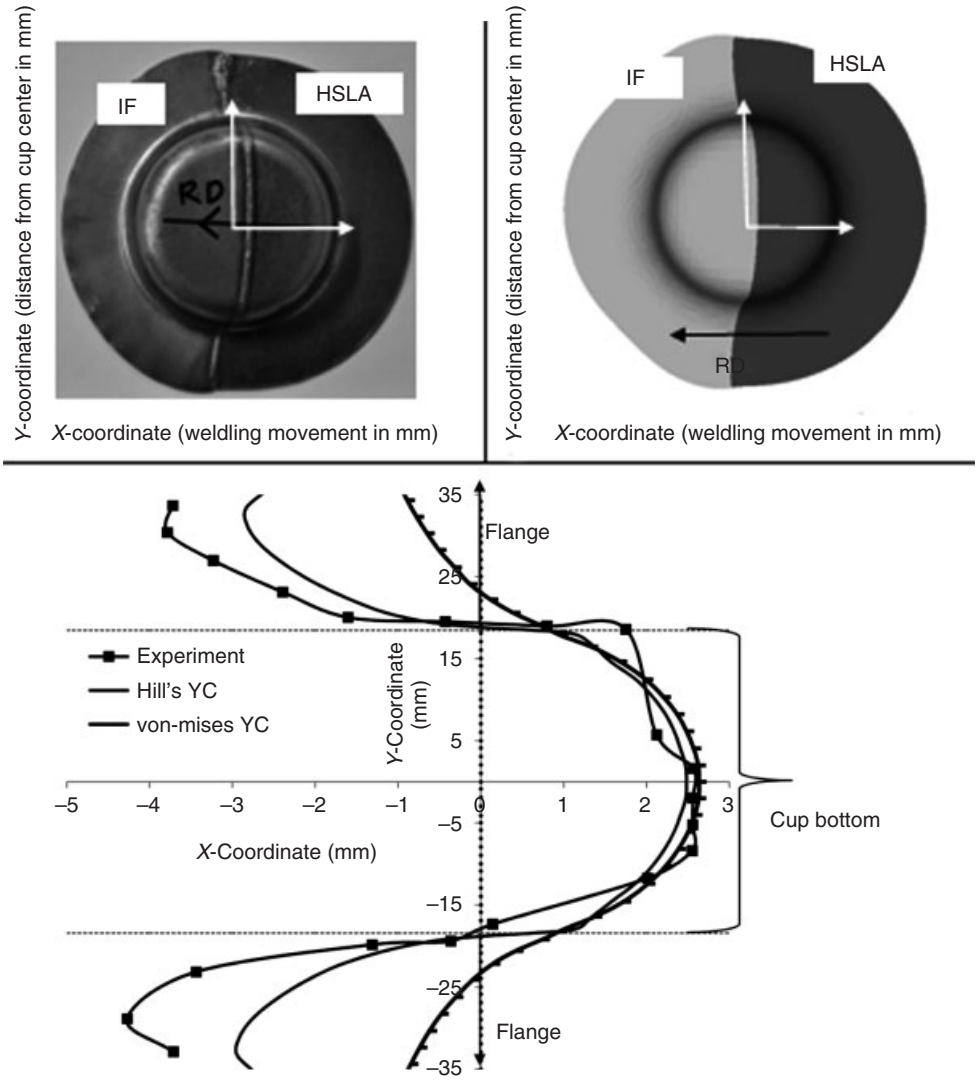


Figure 31. Comparison of experimental and numerical weldline movement in deep drawn IF–HSLA TWB cups.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

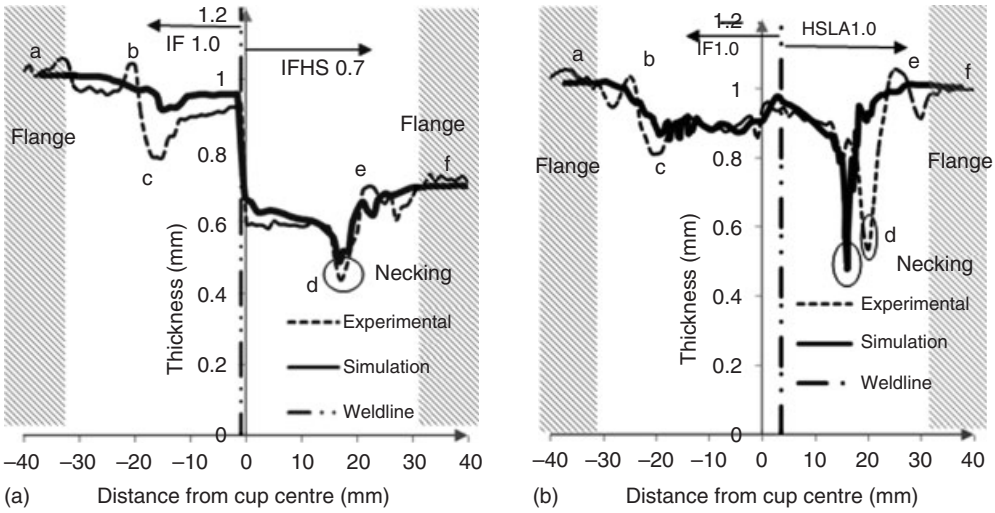


Figure 32. Thickness distribution in deep drawn cups across the weld in TWBs: (a) IF–IFHS TWB and (b) IF–HSLA TWB.

after deformation in the experiment and simulation for IF–HSLA TWB. The comparison of thickness distribution along RD (i.e., across the weldline) of both the TWBs is shown in Figure 32. It was observed that maximum thinning took place in the cup corner corresponding to the punch corner (at location-d). Minor thickening was also observed at the flange. In the IF–IFHS TWB, reduction in thickness was observed on the flat portion of the cup, although necking occurred in the thinner IFHS side corresponding at the cup corner, indicated by location-d. In IF–HSLA TWB, reduction in thickness on the flat portion of the cup was negligible and necking occurred at the cup corner region in HSLA side (location-d). These results revealed the influence of thickness ratio and *R*-value in thinning during deep drawing of TWBs.

4 CONCLUSIONS

Formability study is an important aspect of stamping lightweight automotive body components from various alternative sheet steels and TWBs. It was observed that forming behavior of the sheet metal is significantly influenced by various design, material, and process parameters. This chapter provides insights into various formability measures through laboratory scale experiments and FE-based CAE (computer-aided engineering) techniques. This chapter also highlighted prediction of earring, necking, and wrinkling during steel sheet deformation. Large reduction in formability of two AHSS laser-welded steel sheets, for example, DP980 and DP600, was observed because of hard weld zone and soft outer HAZ. The reduction in cup depth of all laser-welded IF, IFHS, and HSLA is negligible because of narrow weld zone without any soft outer HAZ. Hence, the presence of SZ was detrimental in forming of welded DP steels. The presence of SZ and various defects in weld zone of laser-welded blanks can be characterized by macro- and microstructures with microharness testing before forming. Fracture took place in thinner IFHS side of IF–IFHS thickness combination TWB and material with lower anisotropy HSLA of IF–HSLA material combination TWB. Weld line movement was observed toward HSLA in the cup bottom and toward the IF side in the flange leading to nonuniform deformation of TWBs indicated by thickness distribution. Hence, formability, weldline movement, and failure location in TWBs were influenced by additional parameters such as difference in thickness, difference in properties, and difference in weld zone properties.

ACKNOWLEDGMENTS

The authors are grateful to Dr R. Verma of TATA steel, India, and Dr A. Das of ESSAR steel, India, for providing the steel sheets for the present studies. The corresponding author is thankful to Department of Science and Technology, India, for sponsoring through Fast-track scheme. The authors are also thankful to Steel Technology Centre, Indian Institute of Technology Kharagpur, India, for providing the metallographic and computational facilities.

RELATED ARTICLES

Lightweighting Approach: A Historical Perspective

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ENDNOTES

1. Advanced High Strength Steel Application Guidelines, version 4, World Auto Steel, March 2009, www.worldautosteel.org.

REFERENCES

Advanced High Strength Steel Application Guidelines (2009) *World Auto Steel, version 4*, www.worldautosteel.org (accessed 12 November 2013).

Agrawal, A., Reddy, N.V., and Dixit, P.M. (2008) Optimal blank shape prediction considering sheet thickness variation: an upper bound approach. *Journal of Materials Processing Technology*, **196**, 249–258.

Al-Makky, M.M. and Woo, D.M. (1980) Deep-drawing through tractrix type dies. *International Journal of Mechanical Sciences*, **22**, 467–480.

Auto Steel Partnership (1995) Tailor Welded Blank Design and Manufacturing Manual. Technical Report.

Barlat, F. and Lian, J. (1989) Plastic behavior and stretchability of sheet metals. Part I: a yield function for orthotropic sheets under plane stress conditions. *International Journal of Plasticity*, **5**, 51–66.

Chan, S.M. and Chan, L.C. (2003) Tailor-welded blanks of different thickness ratios effects on forming limit diagrams. *Journal of Materials Processing Technology*, **132**, 95–101.

Choi, W., Gillis, P.P., and Jones, S.E. (1989) Calculation of the forming limit diagram. *Metallurgical Transactions A*, **20**, 1975–1987.

Chow, C.L., Jie, M., and Hu, S.J. (2003) Forming limit analysis of sheet metals based on a generalized deformation theory. *Transactions of the ASME*, **125**, 260–265.

Coubrough, G., Matlock, D., and Van Tyne, C. (1993) Formability of Type 304 Stainless Steel Sheet. Society of Automotive Engineers, Technical Paper 930814.

Fan, D.W., Kim, H.S., and De Cooman, B.C. (2009) A review of the physical metallurgy related to the hot press forming of advanced high strength steel. *Steel Research International*, **80** (3), 241–248.

Ghatrehnaby, M. and Arezoo, B. (2009) A fully automated nesting and piloting system for progressive dies. *Journal of Materials Processing Technology*, **209**, 525–535.

Ghosh, A.K. and Hecker, S.S. (1975) Failure in thin sheets stretched over rigid punches. *Metallurgical Transactions A*, **6**, 1065–1074.

Graf, A. and Hosford, W.F. (1990) Calculations of forming limit diagrams. *Metallurgical Transactions A*, **21**, 87.

Gupta, A.K. and Ravi Kumar, D. (2006) Formability of galvanized interstitial-free steel sheets. *Journal of Materials Processing Technology*, **172**, 225–237.

Hecker, S.S. (1975) Simple technique for determining forming limit curves. *Sheet Metal Industries*, **52**, 671–675.

Hiam, J. and Lee, A. (1978) Factors influencing the FLC of sheet steel. *Sheet Metal Industries*, **50**, 400–411.

Hill, R. (1948) A theory of the yielding and plastic flow of anisotropic metals, proceedings of the royal society of London. *Series A, Mathematical and Physical Sciences*, **193** (1033), 281–297.

Hosford, W.F. (1985) Comments on anisotropic yield criteria. *International Journal of Mechanical Sciences*, **27**, 423–427.

Hosford, W.F. and Caddell, R.M. (2007) *Metal Forming: Mechanics and Metallurgy*, Cambridge University Press, Cambridge, pp. 290–291.

Hrayashi, H. (1996) Forming Technology and Sheet Materials for Weight Reduction of Automobiles. *Proceedings of the 19th IDDRG Biennial Congress*, Eger, pp. 10–14.

Karbasian, H. and Tekkaya, A.E. (2010) A review on hot stamping. *Journal of Materials Processing Technology*, **210**, 2103–2118.

Keeler, S.P. and Backofen, W.A. (1963) Plastic instability and fracture in sheets stretched over rigid punches. *Transactions of the ASM*, **56**, 25–48.

Keeler, S.P. and Brazier, W.G. (1977) Relationship between laboratory material characterization and press shop formability. *Proceedings of Microalloying*, **7**, 517–530.

Kinsey, B., Liu, Z., and Cao, J. (2000) A novel forming technology for tailor welded blanks. *Journal of Materials Processing Technology*, **99**, 145–153.

Kinsey, B., Viswanathan, V., and Cao, J. (2001) Forming of aluminium tailor welded blanks. Society of Automotive Engineers, Technical Paper, 2001-01- 0822, pp. 673–679.

Kishore, N. and Ravi Kumar, D. (2002) Optimization of initial blank shape to minimize earring in deep drawing using FE method. *Journal of Materials Processing Technology*, **130–131**, 20–30.

Korhonen, A.S. (1978) On the theories of sheet metal necking and forming limits. *Journal of Engineering Materials and Technology*, **100**, 303–309.

Jones, S.E. and Gillis, P.P. (1984) A generalized quadratic flow law for sheet metals. *Metallurgical Transactions A*, **15**, 129–132.

Leu, D.K. (1997) Prediction of the limiting drawing ratio and the maximum drawing load in cup-drawing. *International Journal of Machine Tools and Manufacture*, **37**, 201–213.

Lokka, A. M. (1997) *An economic evaluation of tailor welded blanks in automotive application*. MS Thesis, MIT, USA.

Lu, Z.H. and Lee, D. (1987) Prediction of history-dependent forming limits by applying different hardening models. *International Journal of Mechanical Sciences*, **29** (2), 123–137.

Marciniak, Z. and Kuczynski, K. (1967) Limit strains in the processes of stretch-forming sheet metal. *International Journal of Mechanical Sciences*, **9**, 609–620.

Marciniak, Z., Kuczynski, K., and Pokora, T. (1973) Influence of the plastic properties of a material on the forming limit diagram for sheet metal in tension. *International Journal of Mechanical Sciences*, **15**, 789–805.

Mellor, P.B. (1981) Sheet-metal forming. *International Metals Reviews*, **26** (1), 1–20.

Naderi, M., Ketabchi, M., Abbasi, M., and Bleck, W. (2011) Analysis of microstructure and mechanical properties of different high strength carbon steels after hot stamping. *Journal of Materials Processing Technology*, **211**, 1117–1125.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Narayanasamy, R. and Sowerby, R. (1995) Wrinkling behaviour of cold-rolled sheet metals when drawing through a tractrix die. *Journal of Materials Processing Technology*, **49**, 199–211.

Pallet, R.J. and Lark, R.J. (2001) The use of tailored blanks in the manufacture of construction components. *Journal of Materials Processing Technology*, **117**, 249–254.

Panda, S.K. (2007) Formability of tailor welded blanks of low-carbon steels in stretch forming. PhD Thesis, IIT Delhi, India.

Paul, S.K. and Ray, A. (1997) Influence of inclusion characteristics on the formability and toughness properties of a hot-rolled deep-drawing quality steel. *Journal of Materials Engineering and Performance*, **6** (1), 27–34.

Pishbin, H. and Gillis, P.P. (1992) Forming limit diagrams calculated using Hill's non quadratic yield criterion. *Metallurgical Transactions A*, **23**, 2817–2831.

Porter, F. (1991) *Zinc Handbook: Properties, Processing and Use in Design*, Marcel Dekker, New York, pp. 100–102.

Prasad, Y.K.D.V. (1994) A set of heuristic algorithms for optimal nesting of two-dimensional irregularly shaped sheet-metal blanks. *Computers in Industry*, **24**, 55–70.

Rao, K.P. and Sing, W.M. (2000) On the prediction of the effect of process parameters upon forming limit strains in sheet metals. *International Journal of Mechanical Sciences*, **42**, 451–472.

Reddy, J.N. (2005) *An Introduction to the Finite Element Method*, 3rd edn, McGraw-Hill Education, New York.

Sowerby, R. and Duncan, D.L. (1971) Failure in sheet metal in biaxial tension. *International Journal of Mechanical Sciences*, **13**, 217–229.

Storen, S. and Rice, J.R. (1975) Localized necking in thin sheets. *Journal of the Mechanics and Physics of Solids*, **23**, 421–441.

Tadros, A.K. and Mellor, P.B. (1975) Some comments on the limit strains in sheet metal stretching. *International Journal of Mechanical Sciences*, **17** (3), 203–210.

Talyan, V., Wagoner, R.H., and Lee, J.K. (1998) Formability of stainless steel. *Metallurgical and Materials Transactions A*, **29**, 2161–2172.

Waddell, W., Jacken, S., and Wallach, E.R. (1998) The influence of the weld structure on the formability of laser welded tailored blanks. Society of Automotive Engineers, Technical Paper, 982396.

Wang, B.Y., Shi, M.F., Sadrina, H., and Lin, F. (1995) Structural performance of tailor welded sheet steels. Society of Automotive Engineers, Technical Paper, 950376.

Whiteley, R.L. (1960) The importance of directionality in drawing quality sheet steel. *Transaction ASM*, **52**, 154.

Worswick, M.J. (2002) Numerical simulation of sheet metal forming in *Metal Forming Science and Practice* (ed. J.G. Lenard), Elsevier, Oxford, pp. 135–181.

Yurioka, N. and Kasuya, T. (1995) A chart method to determine necessary preheat in steel welding. *Journal of the Japan Welding Society*, **35**, 327–334.

Zhang, X.G. (2000) Galvanic protection distance of zinc-coated steels under various environmental conditions. *Corrosion*, **56**, 39–143.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto155
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Processing of Polymers

Axel Kauffmann[1] **and Maik Ziegler**[2]

[1]*Baden-Wuerttemberg Cooperative State University Karlsruhe, Karlsruhe, Germany*
[2]*Daimler Trucks North America, Portland, OR, USA*

1	Processing of Polymers	1
2	Injection Molding	1
3	Extrusion	5
4	Thermoforming	8
5	Rotational Molding	10
6	Pressing	10
7	Foaming	11
8	Polyurethane Processes	12
9	Summary	13
Related Articles		13
References		13
Further Reading		13

1 PROCESSING OF POLYMERS

High volume automotive part production requires a range of polymer processing methods because of the diversity of parts required. Besides standard molding processes, several other special forming methods are available.

The following focuses on automotive-relevant molding and forming processes such as injection molding, which is used broadly for both nonstructural and structural parts; extrusion of semifinished products; thermoforming of larger parts; and foaming of lightweight parts.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

This chapter provides an overview of the most significant processes. Furthermore, some new and innovative process examples will be presented. This section focuses on thermoplastic material processes. Thermoset composite materials that are mainly processed by compression molding will be presented (see Processing of Polymeric Composites) (Figure 1).

2 INJECTION MOLDING

Injection molding is the dominate processing method used to produce plastic parts in high volume. It has fast cycle times and produces finished parts directly from raw materials. Molded parts are normally made to net shape and require little to no post finishing. Although thermoplastic materials are primarily used for injection molding, thermosets and elastomers can also be used.

In addition to simple mass-produced components, complex shapes can be automatically produced in a single operation. Injection molding is ideal for weight reduction, cost savings, and provides design flexibility to adjust form factor to fit space constraints.

2.1 Thermoplastic injection molding

Thermoplastic injection molding is the foundation for all other injection processes, and it is the most common plastic processing method. Piston type injection molding was introduced by the end of the 1950s. These machines melted polymer pellets in a heated cylinder and injected them using a piston.

At present, single screw injection molding is the most common method in which granular plastics are fed from



Figure 1. Environmentally friendly polymer materials in automobile manufacturing. (Reproduced with permission from Daimler © Daimler AG.)

a hopper into a reciprocating screw inside in a heated barrel in order to plasticize and shear the material into a relatively homogeneous melt. The melt collects at the front of the tip of the retreating screw. In the so-called injection phase, the screw is set back by hydraulic or mechanical force under pressure to axially feed (auger acts as a piston) the molten polymer into the cavity. The melt is injected at high pressure (usually 500–2000 bar) through a nonreturn valve on the injection nozzle and flows through a gate into the mold cavity. Following the high pressure injection, a lower packing pressure is applied until the material in the gate has solidified. Parts manufactured by injection molding have tolerances up to 1/100 mm or even better in special applications.

The two main sections of an injection molding machine are:

- The injection unit that processes and injects the molten polymer under high pressure into the mold.
- The clamping unit that holds the tool and provides an opening and closing function.

A three-zone screw is used for most thermoplastic processing. First, the feed zone is where the plastic pellets are gravity fed from a hopper into the barrel. Next, the compression zone is where the plastic is plasticized and compacted (degassed). Finally, there is the discharge zone, also known as *metering*, where the material is homogenized and finally injected through a nonreturn valve in front of the screw. Owing to increasing back pressure in the cylinder front, the screw moves axially backward. The nonreturn valve prevents the injection melt from flowing back into the screw. This arrangement allows the screw to function as an injection piston (Figure 2).

Typical process values for thermoplastic injection molding are

- mold temperature: 30–120°C, usually 50–80°C;
- melt temperature: 200–320°C;
- feed pressure: 0–50 bar;
- injection pressure: 200–1500 bar;
- packing pressure: 50–70% of injection pressure (Figure 3).

2.2 Elastomer injection molding

The primary difference between injection molding of elastomers and injection molding of thermoplastics is the

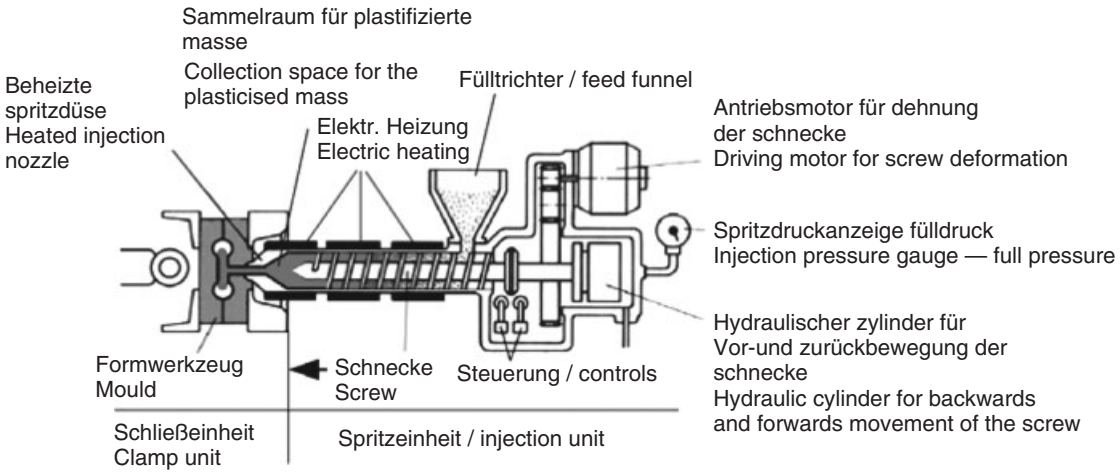


Figure 2. Shematic figure of an injection molding machine. (Reproduced from Domininghaus *et al*., 2012. With kind permission from Springer Science+Business Media.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 3. Injection molded storage compartments and dashboard, of a Freightliner Cascadia. (Reproduced with permission from Daimler © Daimler AG.)

temperature distribution inside the machine. In order to process elastomers, the cylinder around the screw is maintained uniformly at about 80°C using a water jacket to prevent premature vulcanization before the material being injected into the mold. In addition, the elastomer powders are fed by a special ribbon-shaped auger that generate little shear to the plasticized composition.

One additional consideration when molding elastomers is the material's low viscosity state, almost in a liquid form before injection. This process also requires more attention to the design of the tool surface; otherwise, the elastomer injection process is similar to thermoplastic injection molding.

Hoses, seals, and dampeners are typical elastomer molding applications. Elastomers are often used as a functional element when combined with other plastic or metal materials.

2.3 Thermoset injection molding

Injection molding thermosets is similar to the elastomer molding process. Thermoset materials harden at relatively low temperatures. Once cured, no additional processing is possible. The viscosity decreases with increasing temperature and increases during the chemical cross-linking phase. Therefore, the material has a narrow "process window" for optimal flow behavior. To avoid premature initiation of the chemical cross-linking, processing temperatures must be carefully controlled.

The injection unit mixes the chemical units (usually part A and part B) through the screw, which is shorter and has deeper cuts compared to that used for thermoplastic pellets. The compression in the thermoset screw is low because of the low viscosity of the uncross-linked materials. The molds require high surface quality and additional measures for reduced flash formation are necessary. The mold is heated to the curing temperature, which is commonly between 150 and 250°C for thermoset materials. Once 80% of the material has been cross-linked, the material has sufficient strength to be demolded. On demolding, the part will still be relatively hot and will continue to cross-link until a full cure is achieved. Injection molded thermoset components are often used in applications requiring higher temperature and chemical resistance than common thermoplastic materials. They also have high mechanical strength and dimensional stability, especially when reinforced with fibers. These applications are often found in automotive, rail, and aerospace industries.

Plastic processing associated with fiber reinforcement will be described in more detail in Processing of Polymeric Composites.

2.4 Special processes

In addition to traditional injection molding methods, there are a number of advanced injection molding processes (Table 1). The following table lists some processes, which are relevant for the automotive industry.

Table 1. Classification and types of special injection molding methods.

Classification	Variants
Multicomponent injection molding	Sandwich/coinjection Composite molding Multicolor molding Interval injection molding
Fluid and gas-assisted injection molding	Gas-assisted molding HELGA process Water injection technology (WIT) Structural foam molding (w/chemical expanding agent) MuCell process (w/physical expanding agent)
Low pressures injection molding	Cascade injection Compression injection Gas back-pressure technology
Insert molding techniques	Back injection (General) In-mold labeling In-mold decoration Insert moulding Back compression
Injection molding processes to join multiple components	Overmoulding technology Insert molding Hybrid technology In-mold assembly
Lost core processes	Lost core technology Multishell technology
Other special procedures	Injection molding with core pullers for undercuts (liquid silicone rubber) Injection Powder injection molding CD injection molding Cleanroom technology Combined compound/injection molding

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Owing to the large number of process varieties used for injection molding applications, the following describes a few relevant special procedures for the automotive industry as an example.

2.4.1 Multi-component injection molding

Multicomponent injection molding refers to the use of several polymers to produce high quality plastic parts, when different material properties or appearences are needed. Plastics with different colors and/or physical characteristics are combined within a single mold. The polymers can be injected at the same time or sequentially. They can be injected next to, opposed to, or one into the other. Multicomponent sandwich injection molding can be used for recycled materials by injecting them between virgin plastics to achieve a class A surface quality.

A typical example for multicolored injection molded parts are automotive tail lights made out of PMMA (polymethyl methacrylate) (Figure 4).

Multicomponent injection molding is used to produce cost-effective parts with a higher degree of functionality. The chemical compatibility of the materials must be ensured, as it has a strong influence on the adhesion between the injected components.

However, there is another advanced multi-injection molding process, which takes advantage of materials without chemical compatiblity. This process is sometimes called *inmold assembly* because the parts are injected next to, or into each other, and they can be easily disassembled because of low adhesion between the incompatible materials. Good examples are injection molded polymer screws or figurines with movable arms and legs, which are produced in a single tool. Multicomponent injection may also be used for products within daily use, such as packaging, electrical appliances, or for technical products such as rubber bearings, rollers, gaskets, dampers of various types, and housings within molded seals.



Figure 4. Injection molded tail light from Mercedes-Benz CLA. (Reproduced with permission from Daimler © Daimler AG.)

2.4.2 Gas-assisted molding

The principle of gas-assisted molding is to inject the polymer first, followed by injecting an inert gas (usually nitrogen) into the polymer core at approximately 100–200 bar. This process is especially used to produce lightweight parts by hollowing out thick sections. Lightweight parts with a well-defined shape and smooth outer skin can be produced economically with gas-assisted injection. The shrinkage of the part during the cooling process is reduced because of the internal gas pressure. In addition, this method allows for the molding of parts that would either be too difficult or impossible to mold using standard injection molding. The molding process is equivalent to injection molding, but costs are higher as there is additional equipment needed for gas injection.

Depending on the part design, the following are advantages of gas-assisted molding over standard injection molding:

- Design flexibility with respect to different wall thickness within one part
- High rigidity due to large, closed cross sections
- Low warpage and uniform shrinkage
- Reduced sink marks on large parts
- Weight savings of up to 50%
- Shorter cycle times compared to thick walled or solid parts.

2.4.3 MuCell process

MuCell is the name given to a special injection molding method developed by the company Trexel, which uses a microcellular foaming process. MuCell is a physical foaming process utilizing nitrogen in a supercritical state (supercritical fluid, SCF) as a blowing agent to generate micrometer- sized voids in thin-walled plastic parts. The SCF is introduced during the metering of the plastic melt. The foaming process begins when the melt is injected into the mold cavity. Compared to traditional injection molding, cycle times are reduced up to 40%. The process offers 50–75% improvement for key quality issues associated with injection molded parts such as flatness, warpage, and removal of sink marks. The process also contributes to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 5. MuCell valve cover. (Reproduced with permission from Trexel GmbH. © Trexel GmbH.)

weight savings on the order of 20%. For example, a valve cover produced by this method is lighter, has improved flatness, and requires 30% smaller machine size (350 tons instead of 500 tons for compact parts), which contributes to faster cycle times (Figure 5).

3 EXTRUSION

The key element of an extrusion line is the extruder (Figure 6). The most common machines are single-screw and twin-screw extruders. Single-screw extruders are characterized by simple constructions and are mainly used for plastification. The shaping of the melt to semifinished parts such as plates, slides, profiles, or coatings is realized by an extrusion die and the downstream equipment.

Twin-screw extruders with co- and counter-rotating screws are mainly used for polymer compounding, which includes coloring, degassing, and treating of the polymer melts. The co-rotating extrusion is suitable for economic compounding of plastics with very different characteristics and the counter-rotating twin-screw extruder is mainly used for the extrusion of PVC (polyvinyl chloride).

Beside the automotive industry where extrusion is mainly used for wiring harnesses and profiles, extruded products can be found in many industries such as building, packaging, furniture, and aerospace.

Furthermore, there are numerous special extrusion designs, such as the ring extruder (12 screws rotate around a fixed core), planetary extruder (satellite screws rotating around the main screw), Ram extruder (for the production of profiles of PTFE (polytetrafluoroethylene) and UHMW-PE (ultra high molecular weight polyethylene), Kneader- (includes a set of counter rotating kneading blades and a discharge screw), and pin extruder (with pins for additional shear flow).

In addition to the extruder, the complete extrusion line consists of a temperature-controlled heated die along with a cooling/shaping mold followed by a puller (caterpillar), and a cutting unit (saw). A collection table then collects the extruded and cut materials. All these elements are called the *extrusion line*.

Alternatively, an extruder can be used to custom compound plastic pellets. The starting polymers and additives are mixed in the extruder and then pelletized. The convention is to extrude strands, which are then cut and cooled in a water bath. The plastic strands are extruded through a perforated plate directly into a water tank. A rotating blade is used to cut the strands to a designated length as they exit the die. The compounded pellets produced by this process can be further processed in subsequent injection or compression processes.

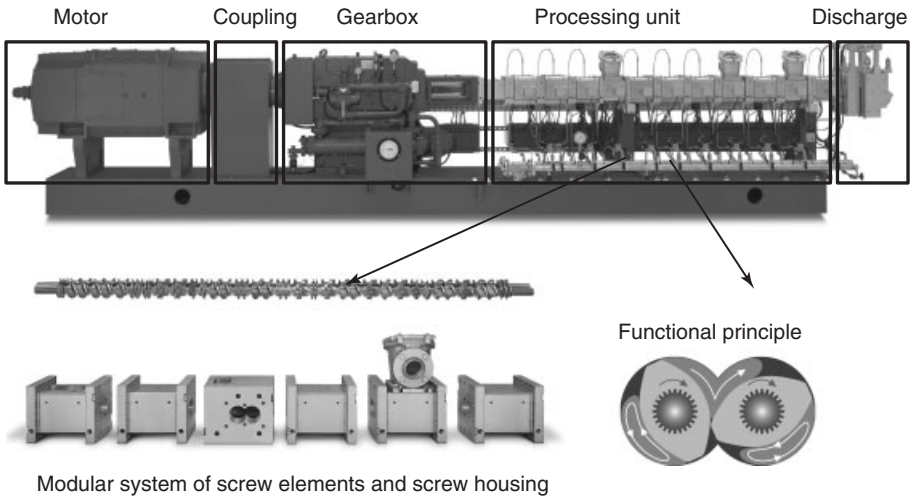


Figure 6. Extruder. (Reproduced with permission from Coperion GmbH, Stuttgart. © Coperion GmbH, Stuttgart.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 7. Exdruded elastomer door sealing of a Freightlinier-Cascadia. (Reproduced with permission from Daimler © Daimler AG.)

3.1 Extrusion of profiles

For tube and profile extrusions, the extruded material is maintained after leaving the shaping tool and is sized to specific dimensions as it is drawn through a cooled calibration tube. Often, it is necessary to apply vacuum between the material and the calibration tube during the cooling process. The material is then sequentially grabbed and pulled from the calibration tube using a crawler, which is moved by chains. The speed of the crawler largely determines the extrusion line speed (Figure 7).

Combining different plastics is often required to produce technical products. Multilayer structured products (up to 9 layers) can be produced by a coextrusion process. For this process, each molding material is plasticized in an independent extruder and introduced into the extrusion die together. With this method, different materials can be optimally homogenized and bonded.

Another important extrusion application in the automotive industries is sheet or film extrusion. A foil is extruded via a flat die and goes through calendering process or a tube is extruded, which is inflated, cut, and flattened (blown film extrusion). Coextrusion is also used for sheet production. For decorative parts, these films are later vacuum or thermo formed and connected to a injection molded part with soft polyurethane (PUR) foam (e.g., instrument panels).

3.2 Extrusion blow molding

In addition to pellet compounding, extrusion or injection molding is increasingly being used to create preforms for blow molded parts. A tubular preform can be created by extrusion with a volume from a few liters to greater than 10,000 L. By combining this method with a blow molding tool, it is possible to create lightweight, hollow-bodied components with three-dimensional surfaces.

The procedure for creating a blow molded part from a plastic extruded preform is as follows:

- Extrude a plastic tube directly into an open, two-part blow mold.
- Close blow mold containing the molten tube.
- Cut the extruded tube across the top surface of the blow mold.
- Insert blow mandrel into the blow mold and introduce compressed air into the cavity of the preform (molten plastic tube).
- Preform is inflated to final part dimensions by taking the contour of the mold surface.
- Cool, open, and eject finished plastic article from blow mold.

Material flash (excess plastic) that is commonly created by the mold pinch at the top and bottom of the part is often removed by a secondary operation automatically and can be recycled. Fasteners, or other functional components may be inserted into the blow mold before molding (Figure 8).

A typical automotive part produced using extrusion blow molding process is a lightweight, plastic fuel tank (PFT). This process offers possibilities to manufacture lightweight tanks with complex geometries and has been established as a successful alternative to tanks made from steel or aluminum. The PFT is generally produced by extrusion of a tubular preform of HDPE (high density polyethylene), which is pinched by the blow molding tool and inflated with air or nitrogen as described earlier. While the first automotive fuel tanks had comparatively simple geometries, they are now complex shapes such as a saddle tank (Figure 5). Since around 1994, the PFTs contain barrier layers to reduce diffusion of volatiles. (Karsch, 2001; Klee, Karsch, and Kempen 2000). PFTs are produced using both the coextrusion and the blow-molding operations. In 1994, Chrysler introduced the first PFTs utilizing a six-layer coextrusion material. At present, this is the most used method (Figure 9).

Increasingly, three-dimensional curved and bellowed tubes are produced using the extrusion/blow molding

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

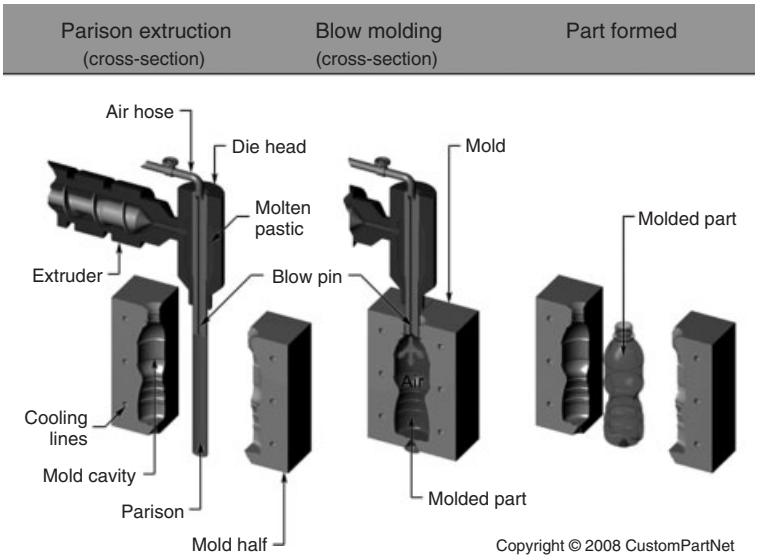


Figure 8. Extrusion/blow molding. (Reproduced with permission from CustomPartNet. © CustomPartNet.)



Figure 9. Extrusion/blow molded saddle tank. (Reproduced with permission from Kautex Maschinenbau GmbH. © Kautex Maschinenbau GmbH.)

process. In the example shown below, a robot is used to place the extruded tube along the mold contour. After insertion, the mold closes and the tube is inflated by compressed air (Figure 10). The advantage of this method lies in the produced part's material uniformity and net shape. No reworking of the article edges is required. The part is lightweight because of uniform wall thickness. Furthermore, this method considerably reduces amount of flash, which minimizes scrap material (Geiger).

Coextrusion can also be used for a range of applications requiring a combination of different plastic materials to address performance requirements. Air intake hoses, manifolds, axle boots, and shock absorbers are a few examples. A further development is sequential coextrusion. The individual components (hard or soft component) can be extruded into various storage heads, alternately released sequentially, and then blow molded.

Blow molding can also be combined with injection molding to reduce part weight and cost. Through a combination of these production processes, the cost of a charged air pipe can be reduced by about 25% over an aluminum component. Inside the blow mold, functional elements are simultaneously injection molded onto the blow molded part. For this example, fixing lugs will be attached to the blow molded tube by injection molding. Immediately after the extrusion is blown and before cooling, plastic tabs are injection molded on to the plastic tube creating a single functional part. The securely attached tabs can withstand a pressure up to 2.8 bar at 230°C. The part is 30% lighter than the aluminum tube.

High melt strength polyolefines (HMSPE) and high melt strength polypropylene (HMSPP) are the most common plastics for extrusion blow molding, but other thermoplastic materials used today include

- ABS (antilock brake system): panels, extenders, and exterior components;
- PVC, PET: transparent and round handle bar;
- PA: hydraulic reservoir and air intake in the engine compartment;
- PC: transparent transport container; and
- TPE (thermoplastic elastomer): substitution of elastomer components.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

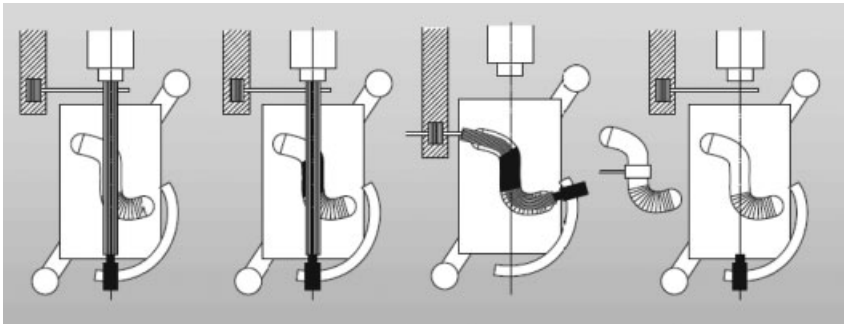


Figure 10. Three-dimensional curved tubes. (Reproduced with permission from Kautex Maschinenbau GmbH. © Kautex Maschinenbau GmbH.)

4 THERMOFORMING

Thermoforming is a forming process by which a plurality of different process steps allows the manufacture of a dimensionally stable plastic part. Essentially, a thermoplastic sheet is placed into a frame which is exposed to heat to soften the material. The material is then transferred over a single cavity mold and vacuum is applied. The vacuum draws the sheet to the mold surface forming the part. The material cools and the part is removed from the mold. Cooling can be accelerated by applying chilled air. The orientation of the molecular chains is frozen in their extended position during the cooling prosses.

Almost all amorphous and semicrystalline thermoplastics are suitable for thermoforming. This method can be used to produce large surface area technical parts and is widely used in packaging. A range of sheet thicknesses (0.05–16 mm) can be used. Foamed sheets can even be up to 60 mm thickness. Parts can be made from PC, PMMA, PA, and ABS semifinished sheets as well as fiber-reinforced composites and self-reinforced materials. In the automotive sector, thermoformed parts are often made from elastomers and thermoplastic polyolefins. The window for the thermoforming temperature is (Table 2):

- above the glass transition temperature for amorphous thermolasitc materials and
- below the crystalline melting temperature for semicrystalline thermoplastic materials.

The thermoforming process typically uses only one tool surface. This means that only one-contouring surface is possible. An advantage is that only one half mold needs to be designed, machined, and purchased. The process works well for parts with medium to large surfaces.

Interior and exterior hoods and panels, especially for commercial, construction, and agricultural vehicles

Table 2. Examples of thermoplastic thermoforming temperature ranges. (Reproduced from Domininghaus *et al*., 2012. With kind permission from Springer Science+Business Media.)

Plastic	Glass Transition Temperature (°C)	Melt Temperature (°C)	Forming Temperature (°C)
PC	∼145	—	150–180
PS	∼105	—	120–150
PP	∼0	∼165	150–165
HD-PE	∼−80	∼135	140–170
PET	∼75	∼245	100–120

Reproduced from Domininghaus *et al*., 2012. With kind permission from Springer Science+Business Media.



Figure 11. Charged air tube made by injection-bonding technology. (Reproduced with permission from Röchling Automotive AG & Co. KG. © Röchling Automotive AG & Co. KG.)

(Figure 11) are examples of thermoformed parts in the automotive business. Moreover, fenders and bumpers have been produced using this method including fiber-reinforced polymers and sandwich panel materials. Other vehicle components are manufactured by vacuum forming, including painted films, which are subsequently

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 12. Mercedes Benz Atego thermoformed rear wall interior trim. (Reproduced with permission from Daimler © Daimler AG.)

back-injected, or fuel tanks, which consist of two thermoformed parts welded together (Figure 12).

In addition to numerous packaging applications, thermoforming is increasingly being used as an alternative to injection molding, in particular, for small to medium sized production runs. For example, panels for machines and material handling devices are produced via thermoforming. Engineered components are usually made from positive–negative vacuum forming, mainly produced through automation (Figure 13).

In the first stage, the thermoplastic sheet is brought to the appropriate forming temperature. (i) By preblowing, the film is prestretched (ii) in order to achieve a uniform wall thickness distribution. Thereafter, the tool (iii) and an applied vacuum pull the film into the desired final shape (iv). After cooling, the plastic can be removed from the mold (without the risk of deforming).

The thermoforming process is usually in competition with injection molding. Thermoforming is advantageous over injection molding for the following cases:

- High component weight (up to 125 kg)
- Large surface area parts (up to 4 m)
- Low cost for low volume production (cost-effective tooling)
- Low cost for design change and color change
- Homogeneous multilayer composite parts.

The disadvantages compared to injection molding are as follows:

- Few design options (large undercuts are not possible)
- Precise temperature control required
- The use of semifinished goods does not allow for plastic formulation.

Low cost tooling enables quick part restyling and facelifts, which are desirable for the automotive industry. Furthermore, deep drawing methods offer high potential for lightweight construction.

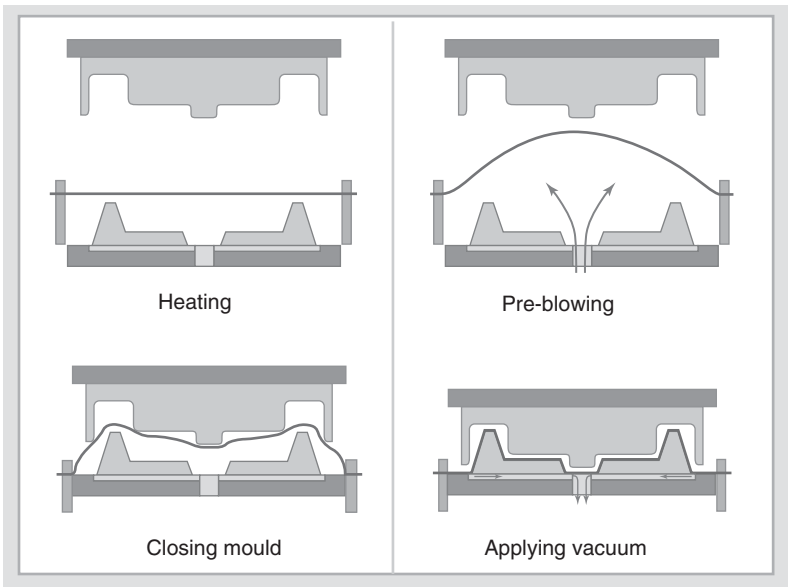


Figure 13. Steps in the positive–negative vacuum forming. (Reproduced from Domininghaus *et al.*, 2012. With kind permission from Springer Science+Business Media.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Twin-sheet thermoforming is a new variation that can be used to produce car air ducts. They are manufactured in sophisticated 3D geometries from PE (polyethylene) foam sheets and offer very lightweight and flexible properties as well as heat insulation for the surrounding components . For twin-sheet thermoforming, the process steps are heating, forming, and cutting at the same station. Two heated PE sheets are supported between the two half molds and are then formed and welded together simultaneously during the twin-sheet forming process. The final hollow body part is stamped directly in the tool. Compared to extrusion blow molding, polypropylene (PP)-air ducts made by the twin-sheet thermoforming method offers a 65% weight reduction.

5 ROTATIONAL MOLDING

Rotational molding is a well-known plastic processing method used for over 50 years. It is suitable to produce large volume, hollow-bodied parts in small quantities. Initially, only simple part geometries could be produced. Currently, there are numerous plastic resin materials tailored for the rotational molding process. Technically sophisticated and weight-optimized parts with complex shapes such as shaped fuel tanks, body parts, air intake ducts, and canoes can be made.

A rotational molding technique is also an effective process to produce decorative interior skins from PVC and TPE thermoplastic polymer powder.

In rotational molding, powder or liquid material is placed into the tool, which then rotates multiauxiliary in a heated chamber. The rotational motion provides a uniform distribution of the plastic material on the mold wall. The mold is then transferred to a cooling station, and the part is removed.

In the first step, the tool is filled with a measured quantity of powdery polymer or liquid raw material. Next, the tool is placed onto a two-axis rotational arm to insure complete distribution of the plastic materials along the wall surface. The tool is then moved into the heating chamber, whereas the tool is rotating allowing for the material to heat and melt gradually, or polymerize. Depending on the component, plastic, or process control, several heating stations may be required. Once uniform distribution of the material is completed, the tool moves to a cooling station. During cooling, rotation is maintained to prevent draining of the material or sagging of the part walls. The tool can be cooled by cold air, water spray, or direct immersion in a water bath. After sufficient cooling, the mold goes to the loading and unloading station. Here, the part is removed. Finally, the tool is refilled and the production cycle starts again.

Common powdered plastics used in this process include PE, PP, PC, PA, fluorpoylymere (PVDF and PFA), and EVA. Tanks, housings for machinery, transport and security containers, leisure, and water sports equipments such as kayaks, furniture, and toys are typical examples for rotational molded parts.

Advantages of rotation molding:

- "Pressure-free" manufacturing with uniform cavity pressure
- Low cost tooling
- Hollow body parts with a wide size range (from few ml up to several cubic meters)
- Economical for small batches
- Very low shear forces—parts with low internal stresses
- Thin-walled structures based on maximum part size possible
- Load inserts or labels possible.

Disadvantages of rotation molding:

- Long cycle time (about 10–30 min) compared to blow molding, thermoforming, and injection molding
- Raw materials must be preground
- Long residence time of the materials at high temperatures (good stabilization is required)
- Heating and cooling in one unit per tool
- High energy consumption
- Intrinsic heat-up and cool-down time.

A process variant is slush molding, which is used for sintering polymer powder into molded skins. The method can produce decorative skins with a constant surface quality, for example, for foils for dashboards. Especially in areas with tight radius, the surface grain will be of better quality compared with thermoformed skins. Furthermore, the production of continuous multicolored skins is possible.

6 PRESSING

Pressing methods are mainly used for the production of components made of fiber-reinforced plastics. For manufacturing these components, mechanical or hydraulic presses and two-or multipart tools are used. In order to shorten the cycle time, the materials are often preheated outside of the tool using radio frequency, microwave, infrared, or convection heating. Alternatively, pressing can be a means to mold parts from extrusion, which is connected upstream of the pressing process. For example, the extrusion machine is utilized to meter a predefined melt that

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

can be compressed to the final shape. Recent developments also use one or more twin-screw extruders in which the extrudate can be automatically handled and inserted into the press tool (see Processing of Polymeric Composites). In addition, high bending strength panels can be produced with this process, by inserting comingled continuous fiber textiles (glass/themoplast fibers) into the press, functioning as outer layers around the extruded core.

7 FOAMING

Polymer foams can achieve weight reductions of up to 98%, compared to their solid counterparts. The foaming process uses chemical or physical blowing agents to expand the material. In principle, all plastics, thermoplastics, thermosets, and elastomers polymers can be expanded. Plastic foams share a common manufacturing process such that the material at the beginning of the foaming process is in a malleable, fluid state, and then foamed and solidified into a foam structure. The main plastics converted to foams are PURs and the thermoplastic polymers such as PS (polystyrene), PE, PP, and PVC.

7.1 Extrusion foaming

Extrusion foaming is the simplest and most economical foaming method for thermoplastic materials. A conventional extruder is utilized to plasticize the polymer and uniformly mix and disperse a blowing agent into the melt. The mixture is dispensed through a cooling die to shape the foam to desired geometry and foam density. Typically, semifinished products such as profiles, sheets, and films are made with a foam core. These parts are often used directly for acoustic, mechanical, and thermal insulation or damping (pipe insulation and packaging).

There are a diverse range of automotive applications that use foam-extruded parts (sun visors, soft skin, door panels, and engine enclosures).

7.2 Particle foams

7.2.1 Extruded foam particles (EPS and EPP)

Particle foam can be produced in a special foam extrusion process. A physical blowing agent is injected into the melt and uniformly mixed and dispersed. Temperarure and pressure are controlled to allow nucleation to occur within the melt but not actual expansion. The material is pushed through channels forming long strands and expands during a sudden pressure drop as the material exits the die. The expanding material is cut by rotating knives into approximate spherical particals, which are cooled in a water bath.

7.2.2 Production of EPP foam particles in the autoclave process

The autoclave process is currently the most widely used method to produce foam particles, especially EPP (expanded polypropylene). In an autoclave, PP microgranules are suspended in a solution containing an inert gas and an expanding agent such as propane or butane. The mixture is stirred while heated and the propellant accumulates in the PP particles through diffusion. When the desired blowing agent content is reached, a valve is opened to release the particles into a lower pressure chamber. In the lower pressure environment, the impregnated propellant expands rapidly creating foamed beads. The cells do not collapse because the melted polymer freezes instaneously during the pressure drop.

7.2.3 Expansion of polystyrene (EPS)

EPS beads (expanded polystyrene) are also initially produced from microgranules. Styrene is polymerized in a suspension polymerization process in the presence of pentane, whereby the propellant gas becomes dissolved in the reacted product. Foaming the EPS microgranules occurs in a pressure vessel with steam. Pentane with a boiling point of just 35°C is evaporated and the granules inflate into foam beads. Approximately 50% of the blowing agent still remains in the foam particles and is used in a later process for further expansion. The produced foam beads must be processed into parts within approximately 1 week after the pre-expansion process.

7.2.4 Molding components from particle foam materials

For EPS molding, steam is used as an energy source, which heats and melts the surface of foamed particles in a shaping tool. Steam is evenly injected through a porous aluminum tool that is placed inside a steam chamber (Figure 14).

The EPS and EPP methods to mold foamed articles are quite similar, although the mechanism is different. During the EPS molding process, the EPS beads still contain a blowing agent. In the case of EPP, they do not. For EPS, the particles expand through the evaporation of the blowing agent. As the foam particles expand, they create internal pressure in the mold cavity, allowing the particles to fuse together to form a part. EPP, on the other hand, is injected with higher backpressure into the mold. The EPP

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

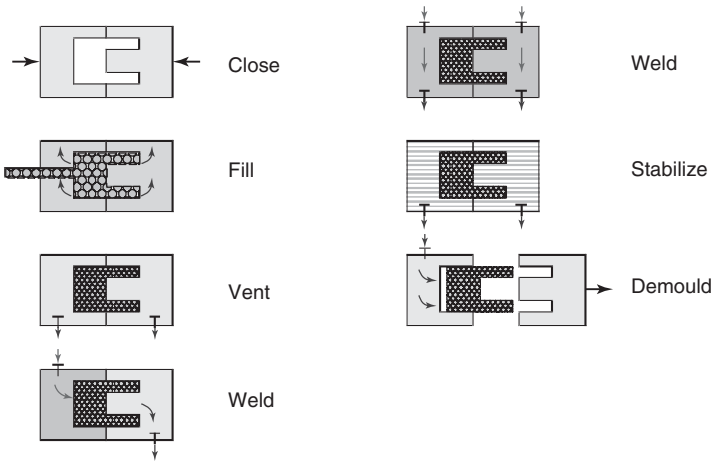


Figure 14. Schematic cycle sequence of moldings made from EPP and EPS.



Figure 15. Bumper with EPP inserts. (Reproduced with permission from Ruch Novaplast © Ruch Novaplast.)

particles expand as the mold is depressurized and then steam is injected to fuse the particles together. Subsequent to the molding process, the parts must be annealed to reach the final geometry and smooth the part surface. This is necessary for the removal of residual moisture contained in the cells, resulting from the steaming process. Six to eight hours are common anealing cycles at approximately 80°C.

EPS is mainly used for packaging and thermal insulation in the construction industry. EPP is currently applicable in the field of reusable transport packaging and increasingly in the automotive sector. For example, tool boxes, side impact protection, and head and knee cushions are made out of this material. Moreover, the majority of the car bumpers are equipped with EPP inserts. The market for EPP is constantly growing because of the lightweight properties



Figure 16. Polypropylene monomaterial composite with TPO skin. (Reproduced with permission from Taracell.)

and the good energy absorption. Compared with EPS, EPP offers higher mechanical flexibility and higher temperature resistance (Figure 15).

Typical applications for EPP in combination with other materials are EPP sun visors with injection molded frames and headrests using an EPP reinforcement structure in conjunction with a surface out of PUR foam.

A simple and cost-effective method for producing decorative and lightweight shaped parts can be achieved by combining a TPO skin (thermoplastic olefin based on PP) with EPP. This is a new, economical way of producing fully recyclable, decorative parts such as sun visors, pillars, and door trims (Figure 16).

8 POLYURETHANE PROCESSES

PUR is one of the most divese polymers existing in a thermoplastic or thermoset form and widely used in the automotive area. Accordingly, there are multiple production processes for PUR.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 17. Instrument panel. (Reproduced with permission from Daimler © Daimler AG.)

High strength PUR parts are made in low pressure or high pressure machines, in which the liquid components of polyisocyanate and polyol that may include additives and fillers, are mixed, and then form a reaction mixture. This process is known as the *reaction injection molding* (*RIM*) method. In principle, this method is used to make structural parts (porous core with virtually cell-free surface layers) and parts with predetermined, homogeneous density. The reactive mixture can be combined with fibers to increase the molded part strength. This is called *reinforced reaction injection molding* (*RRIM*). With additional continuous fiber reinforcement layers in the tool, the process is called *structural reaction injection molding* (*S-RIM*) or *structural reinforced reaction injection molding* (*SRRIM*).

PURs are also used as elastomers [mainly themoplastic polyurethane (TPU)] or as lightweight, flexible foams. The most important applications of PURs are in the furniture and upholstery industry, construction, and the automotive industry. Lightweight components are relevant for automotive applications, such as steering wheels, dashboards, seats, interior trim parts, and damping elements. Similar components are found in rail vehicles and aviation (Figure 17).

9 SUMMARY

Polymers are becoming more important in the auotmotive engineering society, as they are highly versatile, lightweight, and offer greater flexiblity for part design and component integration, compared with metals. This chapter only highlights some major polymer processes and applications that are being used by the industry. There are many more applications and developments in the polymer field, such as decorative exterior polymer films and polymer adhesives, which are increasingly playing an important role to simplify the assembly process. The field of polymer science continues to evolve rapidly and will offer new applications in the near future.

RELATED ARTICLES

Structure and Properties of Polymeric Composites
Processing of Polymeric Composites
Hybrid Structures
Recycling of Polymers & Composites
Biopolymers & Biocomposites

REFERENCES

Domininghaus, H., Elsner, P., Eyerer, P., and Hirth, T. (2012) *Kunststoffe. Eigenschaften und Anwendungen [Plastics. Properties and Applications]*, 8th edn, Springer, Germany.

Geiger http://www.geigerautomotive.com/ (accessed 24 August 2013).

Karsch, U.A. (2001) Aufbau von aktuellen Tanksystemen für Fahrzeuge des europäischen und amerikanischen Marktes in *Emissionen aus Kraftstoffsystemen von Pkw. 21./22. Februar 2001, Essen* Tagungsband. Essen, (ed. Haus der Technik)(Hrsg.).

Klee, W., Karsch, U.A., and Kempen, T. (2000) Barrieretechnologien. Ein Beitrag zur Emissionsminderung von Kraftstoffanlagen. Verein Deutscher Ingenieure, VDI-K (Hrsg.) Kunststoffe im Automobilbau. Tagung, Mannheim, 5. und 6. April 2000. Düsseldorf, 2000.

FURTHER READING

Bayer (2002–2006) Spritzgießen von Qualitätsformteilen – Verfahrenstechnische Alternativen und Verfahrensauswahl, ATI 1147 d, Bayer AG, Ausgabe 2002–06.

ContiTech AG (2009) *Pressemeldung*, http://www.contitech.de (accessed 24 August 2013).

Eyerer, P. (2005) Kunststoffkunde. Vorlesungsmanuskript mit CD. 13. Auflage, Fraunhofer IRB Verlag, Universität Stuttgart.

Eyerer, P., Hirth, T., and Elsner, P. (2008) *Polymer Engineering - Technologien und Praxis*, Springer Verlag, Berlin.

Kauffmann, A. (2011) *Verarbeitung von Kunststoffen. Handbuch Leichtbau: Methoden, Werkstoffe, Fertigung*, Carl Hanser Verlag, München.

Klempner, D. and Frisch, K.C. (1991) *Handbook of Polymeric. Foams and Foam Technology*, Carl Hanser Verlag, München.

Michaeli, W. (1999) *Einführung in die Kunststoffverarbeitung*, 4. Auflage, Hanser Verlag, München.

Ziegler, M. (2003) *Polypropylene Foam Parts for Automotive Applications* AUTO 2003, , Sovereign Publications Limited, London.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto157
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Operating Principles

Michael Bargende[1], Dietmar Schmidt[1], Hans-Jürgen Berner[2], and Michael Grill[2]

[1] *Universität Stuttgart, Stuttgart, Germany*
[2] *Research Institute of Automotive Engineering and Vehicle Engines Stuttgart, Stuttgart, Germany*

1 Classification of Internal Combustion Engines 1
2 Engine Operating Cycles 2
3 Characteristic Properties 9
4 Specific Metrics 13
References 16

1 CLASSIFICATION OF INTERNAL COMBUSTION ENGINES

Internal combustion engines (ICEs) are mechanical devices that thermodynamically convert the chemical internal energy of a fuel into mechanical work. This is accomplished by allowing the fuel to chemically react in the confinement of the cylinder, during which the chemical internal energy is converted into sensible internal energy. The transformation of the chemical internal energy into sensible internal energy is manifested by an increased temperature and pressure in the cylinder, which can then be expanded to extract mechanical work. Therefore, the chemical energy, which is contained in the fuel, is transformed into "thermal" energy, which is subsequently used to produce mechanical work.

Principally, the conversion process that produces mechanical work from combustion should be distinguished between the case of continuous-flow machines, for example, gas turbines or jet engines, and reciprocating engines. In continuous-flow machines, work is achieved using the kinetic energy of the working fluid. In reciprocating engines, for example, reciprocating ICEs, rotary engines, and steam and Stirling engines, the working chamber has rigid walls, with at least one movable wall (usually this is the piston). Owing to the volume change, combined with a varying gas pressure within the working chamber, power output is obtained.

Further differentiation of reciprocating engines could be made by considering external (steam and Stirling engines) and ICEs (reciprocating piston and rotary engines). In external combustion engines, the combustion process occurs outside the work producing chamber, whereas in ICEs, combustion takes place intermittently inside the work producing chamber, that is, the cylinder. In external combustion engines, energy transfer takes place by heat transfer processes from the combustion chamber to the working fluid (which is not the products of combustion in the external combustion chamber), which limits the viability of external combustion engines for mobile transportation systems because the timescales of the energy transfer are far too long. One advantage of internal combustion is that response times to load changes are extremely short. This makes the ICEs the preferred design for mobile transportation, which is characterized by frequent transient and dynamic change of torque requests. Maximum pressures and temperatures arise only for short times, and thus, thermal and mechanical stresses on the power unit are limited. Moreover, power density is high. However, one problem for ICEs is the short time available for the fuel–air mixture generation, needed to completely oxidize

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

the fuel–air mixture. This can lead to decreased efficiencies because of imperfect and incomplete combustion. In addition, combustion is always accompanied by formation of unwanted pollutants in (locally) high temperature and inhomogeneous (rich or lean) mixtures.

2 ENGINE OPERATING CYCLES

2.1 Design of reciprocating engines

Nearly all ICEs are of the reciprocating piston type, where the piston moves back and forth in a cylinder. Power is transferred from the piston by a connecting rod into the crankshaft, as shown in Figure 1. Owing to the rotation of the crankshaft, a cyclic movement of the piston takes place. At the top-dead-center (TDC) and bottom-dead-center (BDC) positions, the piston comes to rest and the corresponding volumes of the working chamber have minimum and maximum values, respectively. The minimum volume is called the *clearance volume* V_c, whereas the swept volume by the piston (the maximum minus the minimum volume) is called the *displacement volume* V_d.

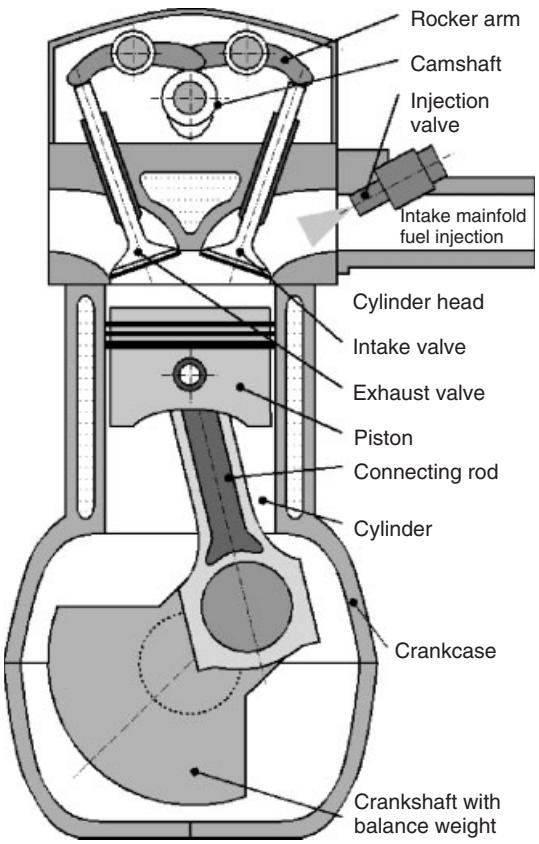


Figure 1. Cross-sectional drawing of a reciprocating engine.

2.2 Definitions and classifications

Figure 2 summarizes the classification of ICEs. It can be seen that the whole engine consists of many aspects involving interdisciplinary fields of activity.

2.2.1 Classification for process management and mixture generation

2.2.1.1 Spark-ignited engine. The spark-ignited (SI) engine or the "Otto engine" is one of the dominant engines on the road today. The notation "Otto engine" refers to the inventor of the so-called four-stroke cycle process management (see later), Nikolaus Otto ((Sass, 1962); (Goldbeck, 1976)). It is an ICE with an open circuit of the working fluid for both reciprocating piston and rotary engines. Mixture preparation of fuel and air takes place as early as possible in order to get a well-mixed homogeneous mixture, which is compressed in the working chamber. It should be noted that the compression must be limited, to ensure that no autoignition due to the increased temperature occurs. At the time of maximum compression, which is near TDC, which is at the lowest volume, ignition is initiated by a spark plug (externally supplied ignition) and the air–fuel charge can be burned. (SI engine combustion system details are discussed in Spark Ignition Combustion.)

SI engines can further be differentiated into Otto engines operated with liquid or gaseous fuels. The fuel can be either delivered to the intake manifold (the fresh air inlet system) using a carburetor or an injection system or it can be injected directly into the cylinder. The mixture can be rich (excess fuel), stoichiometric, or lean (excess air). In case of a globally lean mixture, the charge can be stratified by a late injection in the compression portion of the cycle.

2.2.1.1.1 Gasoline-fueled engines. More than 95% of all SI engines worldwide are fueled with liquid gasoline of different qualities. With a mixture preparation in the intake system, the fuel typically evaporates completely before it enters the working chamber. This is true for part load condition where the manifold pressure is lower than the atmospheric pressure. For full load and especially turbo- or super-charged conditions, a significant part of the fuel enters the working chamber in liquid form, and the evaporation as well as the mixing process with the fresh air for creating a homogeneous fuel–air mixture takes places during the induction and compression stroke.

2.2.1.1.2 Gaseous-fueled engines. In this case, the fuel is gaseous under atmospheric conditions. The mixture generation takes place upstream of or within the working chamber.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

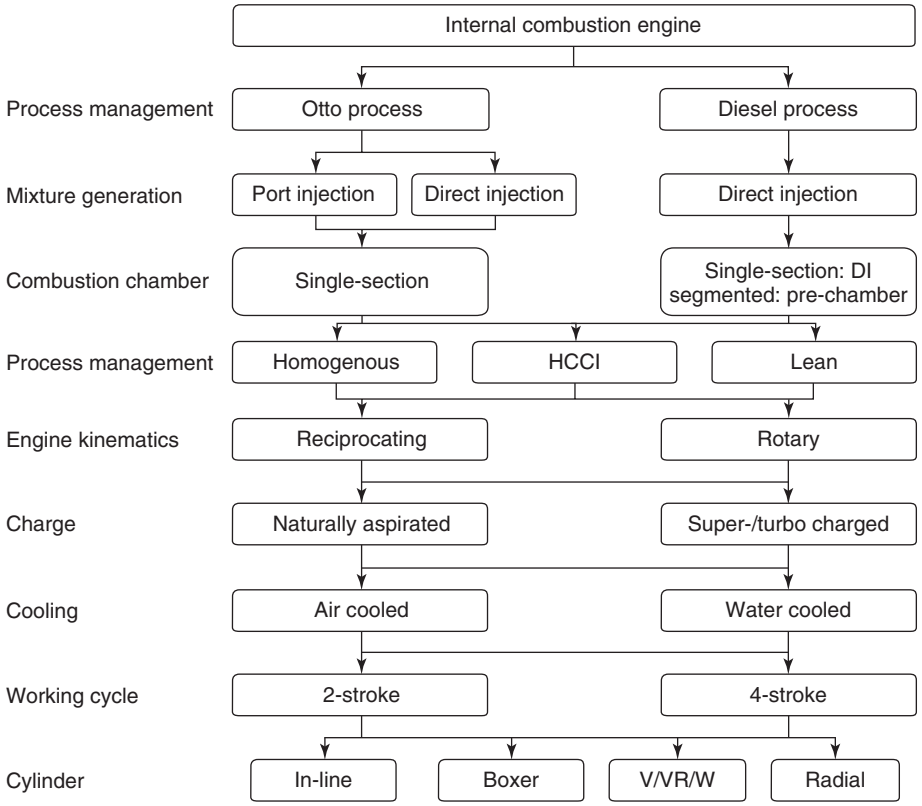


Figure 2. Main classification definitions in internal combustion engines.

Most gaseous-fuel engines are modified gasoline-fueled engines, thanks to the similar behavior of the different fuel types from a combustion perspective. Two options are more commonly used: liquid petroleum gas (mainly propane and butane, LPG) and methane-containing gas (e.g., natural gas and biogas, CNG and LNG), which is compressed and stored (i.e., to be transportable in acceptable volumes) in liquid or gaseous form (under high pressure, 200 bar, in the latter case) in special tanks. Engines using hydrogen as a fuel have been the subject of scientific studies for many years but do not have any significance in today's market for lack of infrastructure.

2.2.1.1.3 Otto engines having a carburetor. Liquid fuel is mixed with air in a carburetor before the mixture enters the working chamber. This is an old technology and will not be discussed in detail.

2.2.1.1.4 Injection Otto engines. Liquid fuel is injected directly into the air inlet system before the mixture enters the working chamber or directly into the cylinder, as can be seen in Figures 3 and 4.

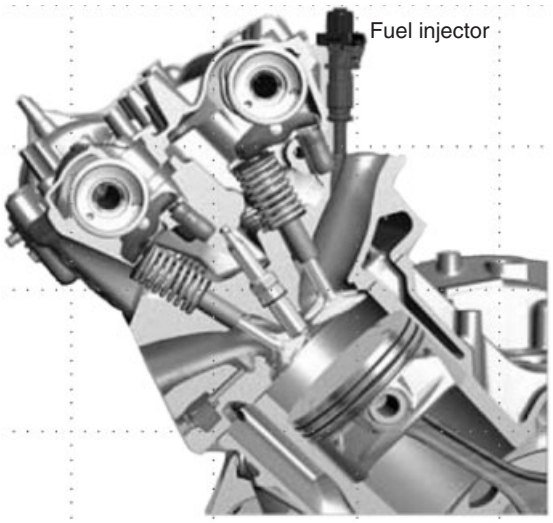


Figure 3. Otto engine with port injection. (Reproduced by permission of Daimler AG.)

2.2.1.1.5 Lean-burn engines. In lean-burn engines, globally very lean air/fuel ratios (AFRs) are used ($1.3 < \lambda < 1.7$, where λ is a characteristic property describing the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

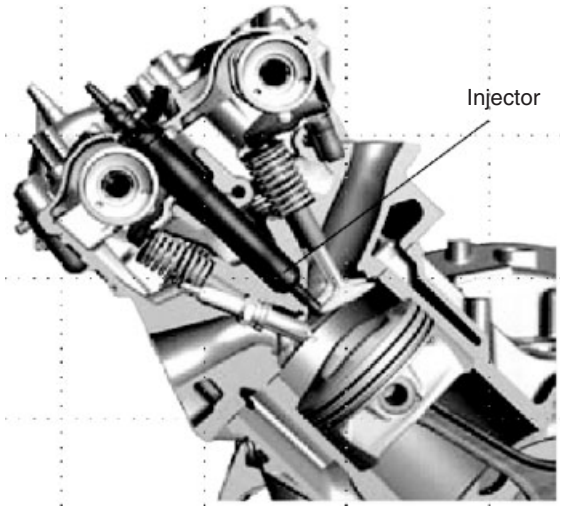


Figure 4. Direct injection Otto engine. (Reproduced by permission of Daimler AG.)

fuel to air ratio. For the definition of λ, refer to Section 3.2). Special attention is needed in order to ensure ignition of the lean mixture. Therefore, special design of mixture preparation, air-inlet and inlet manifold, combustion chamber, and ignition system is needed. One problem is that the very efficient exhaust gas after-treatment system known as the three-way catalyst, and used in stoichiometric mixture conventional Otto engines, cannot be used because of the lean AFR, which results in the presence of a significant quantity of oxygen in the exhaust stream. Additional and more complex exhaust gas after-treatment systems are necessary for the chemical reduction of nitrogen oxides in oxygen-rich exhaust gases from lean-burn engines.

2.2.1.1.6 Stratified-charge engines. In this case, an inhomogeneous mixture of fuel and air is desired. The mixture is ignited by the spark plug within a fuel rich zone (rich = excess of fuel). The hot gases of the rich combustion process get mixed with the mixture of lean areas (lean = excess of air), and thus, lean combustion takes place. Usually, direct injection (DI) of the fuel into the combustion chamber is used to achieve charge stratification. The globally lean fuel–air mixture leads to higher efficiencies and to lower fuel consumption. However, although engine-out-emissions can be reduced, additional exhaust after-treatment systems are necessary for the chemical reduction of nitrogen oxides in oxygen-rich exhaust gases, as mentioned earlier for lean-burn engines.

2.2.1.2 Compression ignition engine. Compression ignition (CI) engines include the most prevalent CI engine on the road today, the Diesel engine, as well as new CI engine technologies being researched, such as homogeneous charge compression ignition (HCCI) engines. The notation "Diesel engine" refers to Rudolf Diesel, the inventor of a combustion method initiated by autoignition in reciprocating engines ((Sass, 1962); (Diesel, 1892); (Diesel, 1913); (Diesel, 1893)). The Diesel engine is also an ICE having an open working fluid circuit. In contrast to Otto engines, the combustion chamber is filled only with air during compression. Therefore, the compression ratio can be much higher. During compression, the air temperature increases. Near the end of compression, which is near the TDC position of the piston and minimum volume in the cylinder, the fuel is injected. The fuel evaporates in the preheated air, mixes with the air, gets ignited by autoignition, and finally gets oxidized. (Diesel engine combustion system details along with advanced, low temperature Diesel combustion strategies currently being research are discussed in Diesel and Diesel LTC Combustion.)

Diesel engines further can be classified as Diesel engines using air injection (not in use anymore and not discussed further), with prechambers (divided chamber), and using DI.

2.2.1.2.1 Diesel engines with prechamber. The liquid fuel is injected into a small-volume prechamber, which is connected via a small passage to the main combustion chamber. During compression, airflow is introduced in the prechamber from the main chamber. When fuel is injected near TDC into the prechamber, mixing and autoignition (or glow-plug-assisted ignition) occur rapidly because of fluid motion, followed by a rich combustion. After the combustion starts, the reacting mixture blows down the passage into the main combustion chamber where an overall lean combustion takes place to complete the combustion process. Figure 5 depicts a Diesel engine having a swirl prechamber. Because of the generally lower efficiency of prechamber engines, owing to heat losses to the passage walls, these engines are not widely used anymore.

2.2.1.2.2 Direct injection Diesel engines. In a DI engine, liquid fuel is injected directly into the combustion chamber near TDC. High pressure injection systems are used to support fast mixture generation. The commonly used common-rail injector technology allows more than one injection event per operating cycle in order to optimize the combustion processes to reduce noise and pollutant emissions (e.g., pre-, main-, and postinjection). Nearly all Diesel engines are equipped with a DI system today (Figure 6).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

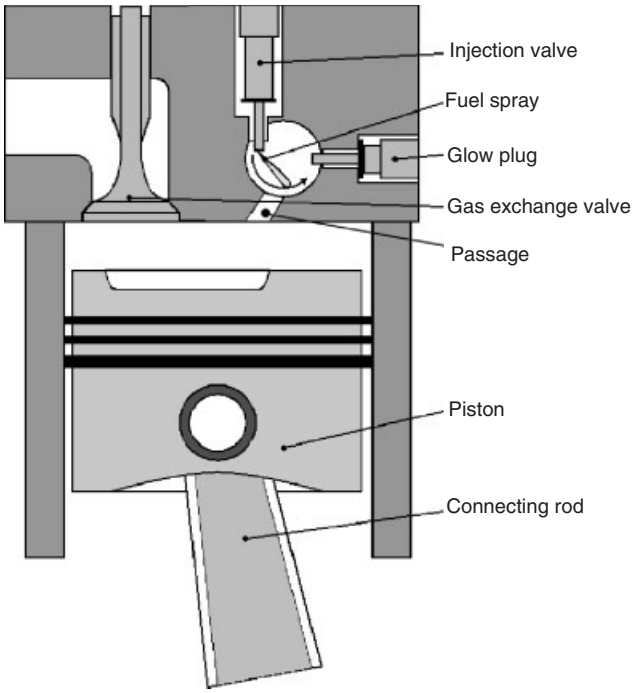


Figure 5. Cross-sectional drawing of a Diesel engine having a swirl prechamber.

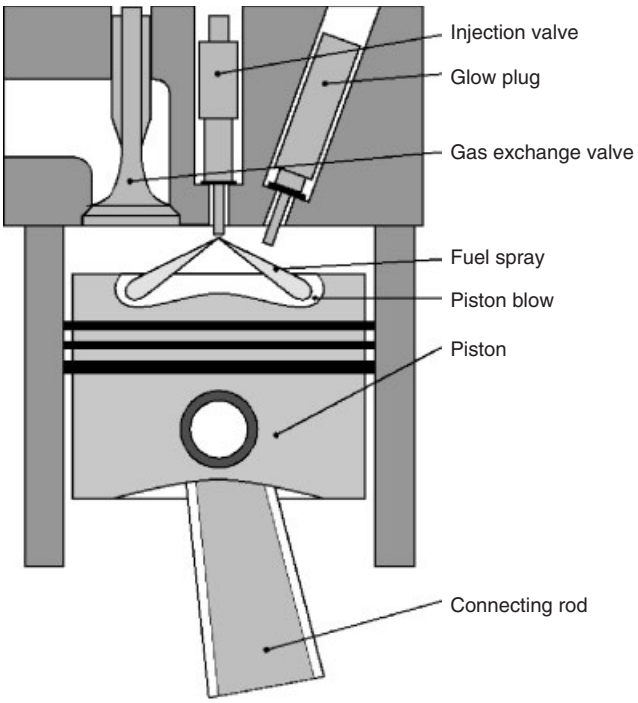


Figure 6. Cross-sectional drawing of a direct injection Diesel engine.

2.2.1.2.3 Homogeneous charge compression ignition engines. Creating a homogeneous mixture of fuel and air is usually not a big problem, especially with gasoline, because of its relative low evaporation temperatures. On the other hand, a controlled self-ignition of the homogeneous mixture close to the TDC position of the piston marking the end of the compression stroke is not easily realizable, because the ignition process mainly depends on very complex kinetic issues. Owing to high knock resistance (high RON numbers), gasoline needs very high temperatures—above 1100 K—for self-ignition at realistic engine speeds. This can be realized, for instance, by retention of hot burned gas in the cylinder from the previous combustion event. This retention process can be controlled, for example, by a variable valve actuation system. The benefits of the HCCI strategy are very low NO_x emissions and a better indicated efficiency due to very lean or dilute operating conditions leading to low combustion temperatures, very short combustion duration, and lower throttling losses. The drawbacks are the need for complex control, and higher combustion rates that lead to significant combustion generated noise. The high combustion rates and noise limit this combustion mode, in a highly premixed form, to part-load conditions and moderate engine speeds. However, test bench results suggest that compression ignition of stratified mixtures, coupled with EGR, can allow control of combustion rates and noise, especially at low and high load conditions. Practical application of this technology continues to be a subject of current research. (Advanced CI combustion strategies such as HCCI are discussed in detail in Advanced Compression-Ignition Combustion for Ultra-Low NO_x and Soot.)

2.2.2 Classification of intake condition

2.2.2.1 Naturally aspirated engine. Naturally aspirated engines induct air from the surrounding and discharge the exhaust gases into the surrounding.

2.2.2.2 Super- and turbocharged engines. In super- or turbocharged engines, the fresh air in the air intake system is compressed by external devices, for example, by a turbo- or supercharger. Thus, the density of the intake gas is increased, resulting in more fresh gas entering the cylinder and increased power and torque from the engine. There are two well-established methods for compressing the intake air, exhaust gas turbochargers and mechanical superchargers, with turbochargers being used more commonly in series production. In addition, two- or more-staged charging combinations of series of turbochargers and compressors can be used.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

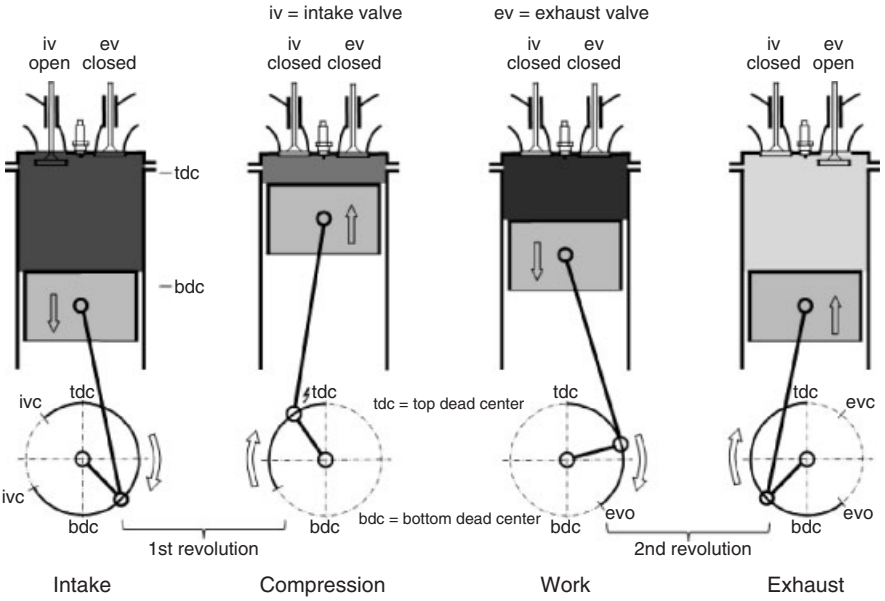


Figure 7. Four-stroke operating cycle (schematically).

2.2.3 *Classification of working cycle*

2.2.3.1 Four-stroke engine. Most of the engines (reciprocating and rotary) operate with the so-called four-stroke cycle, that is, for one complete working cycle, four strokes (piston movements), corresponding to two revolutions of the crankshaft, are necessary. Both Otto- and Diesel engines can use this cycle, which is shown in Figure 7. The four different strokes are as follows.

2.2.3.1.1 Intake stroke. Piston movement from the TDC to the BDC position with intake valves open. This draws fresh gas into the cylinder (fuel–air mixture for Otto and air for Diesel engines).

2.2.3.1.2 Compression stroke. Piston movement from the BDC to the TDC position with all valves closed. The gas inside the cylinder is compressed to a small volume compared to the volume at BDC position. During compression, the gas pressure and the temperature increase. Near the end of the compression stroke, combustion is initiated. In Otto engines, ignition takes place at the spark plug, and in Diesel engines, injection of the fuel and subsequently evaporation and autoignition occurs. When combustion starts, cylinder pressure increases rapidly.

2.2.3.1.3 Power stroke. Piston movement from the TDC to the BDC position with all valves closed. The movement starts at high pressure and temperature gas, which pushes the piston down because of the gas force onto the piston, rotating the crankshaft. The gas inside the cylinder rapidly cools down when volume is increased.

2.2.3.1.4 Exhaust stroke. Piston movement from the BDC to the TDC position with exhaust valves open (EVO). This causes the burnt gases to exit the cylinder.

In real engines, the valve opening and closing times are not precisely at the TDC and BDC positions but short times earlier or later. The four strokes are also visible in pV diagrams, where p is the pressure in the cylinder and V is the cylinder volume at any time. In Figure 8, an example for an Otto engine pV diagram at full load is shown.

2.2.3.2 Rotary engine (Wankel engine). The notation Wankel engine refers to its inventor Felix Wankel (Huf, 1961). Wankel engines work with the four-stroke principle, that is, the working chamber is opened and closed two times, whereas in two-stroke engines (see later) only once.

In contrast to reciprocating piston engines, no valves for the gas exchange are needed. The gas exchange process, similar to two-stroke engines, is realized by simple ports. For a schematic illustration of a Wankel engine, see Figure 9.

Advantages of rotary engines are their compact design, the smooth and low vibration performance, and the high rotational speeds. Problematic are the sealing of the combustion chamber, the nonoptimal design of the combustion chamber (from a combustion perspective), and the high one-sided thermal stresses (e.g., no cooling of the spark plug by inflowing fresh gas).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

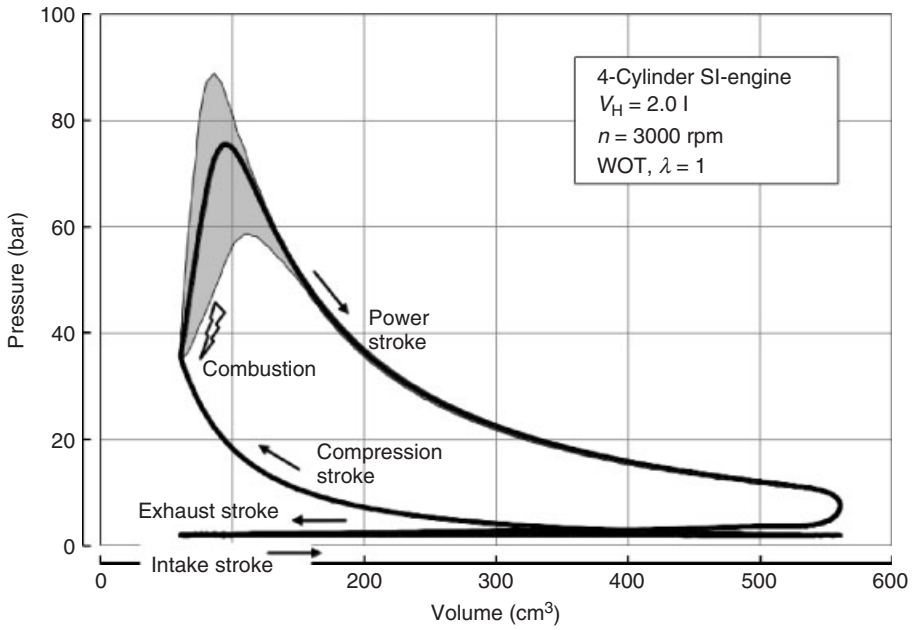


Figure 8. *pV* diagram of a four-stroke Otto engine at full load conditions. The gray area marks the region of cycle-to-cycle fluctuations.

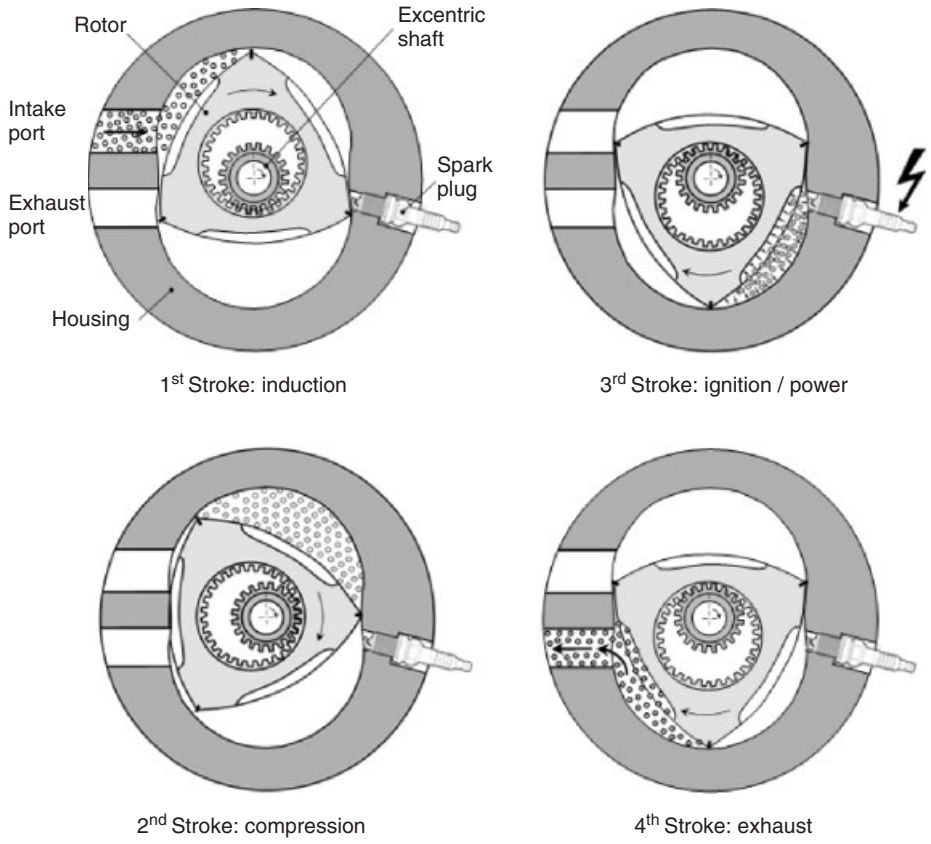


Figure 9. Cross-sectional drawing of a Wankel engine.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

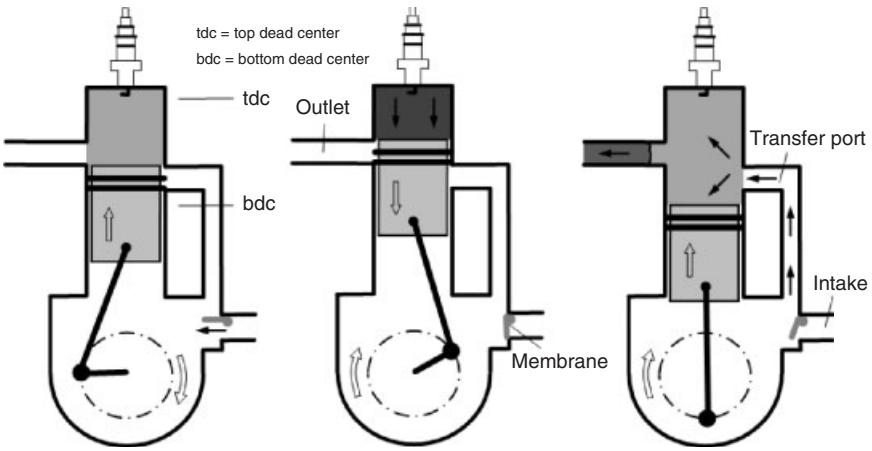


Figure 10. Two-stroke operating cycle (schematically).

2.2.3.3 Two-stroke engine. Only one rotation of the crankshaft is necessary for one power cycle in a two-stroke engine. The gas exchange mostly is managed by simple ports in the liner of the cylinder, which are opened and closed by the reciprocating motion of the piston. In Figure 10, a simple configuration for a two-stroke engine is shown. The two strokes are discussed in the following sections.

2.2.3.3.1 Compression stroke. Piston movement from around the BDC to the TDC position with inlet and exhaust ports closing at the beginning of the stroke. The gas mixture above the piston is compressed. Below the piston, the crankcase volume pressure drops below atmospheric pressure, until at a certain position, the piston opens the crankcase membrane and fresh gas can enter the volume below the piston.

2.2.3.3.2 Power stroke. Piston movement from the TDC to the BDC position with inlet and exhaust ports closed at the beginning. Ignition by spark plug occurs above the piston around the TDC position and rapid combustion takes place. The piston is pushed downward causing a rotation of the crankshaft. In the volume below the piston, the gas mixture is compressed and near the end of the stroke, the *transition from power to compression stroke* starts.

2.2.3.3.3 Transition from power to compression stroke. Exhaust ports open first, initiating the exhaust of burnt gases, followed by opening of the intake ports initiating induction of fresh compressed gas from the crankcase (or from an external supercharger), a process called *scavenging*.

In Figure 11, a corresponding *pV* diagram for a two-stroke cycle is shown. Because power is produced every stroke, these engines can have roughly twice the power

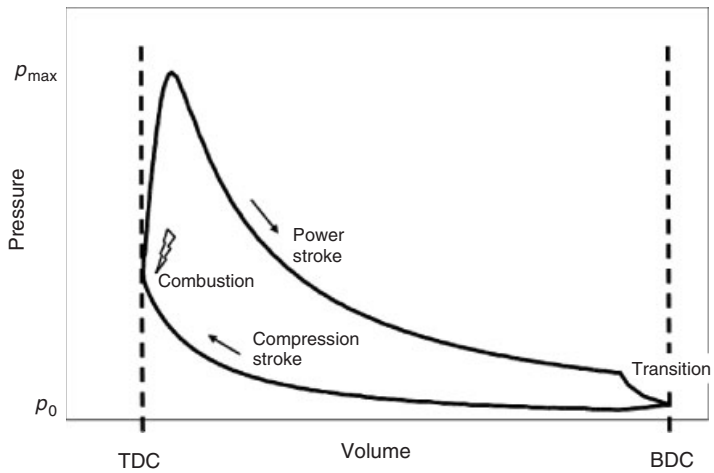


Figure 11. *pV* diagram of an Otto two-stroke cycle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

density of comparable four-stroke engines but have significant challenges with exhaust emission control. It should be stated that the remarks given are simplified and illustrative. It is also possible to realize the two-stroke cycle with a valved engine and mechanically driven auxiliary boosting system (Blair, 1990).

3 CHARACTERISTIC PROPERTIES

In this chapter, important properties are given, which are helpful to characterize engines. It is important to note differences in the subscripts, which are used to relate the meaning of the reference parameter to different variables (e.g., distinguish between the displacement of a single cylinder, V_d, and the displacement of the whole engine-all cylinders, V_D). For further reading, refer also to basic textbooks such as Refs ((Heywood, 1988); (Taylor, 1985a), (Taylor, 1985b); (Obert, 1973); (Taylor & Taylor, 1961)).

Nearly all important properties can be introduced by investigation of a single relation for engine power. This will be the guide throughout this chapter.

3.1 Compression ratio

The ratio between maximum and minimum volume per cylinder is denoted as compression ratio r_c. The maximum volume is the sum of the displacement volume V_d per cylinder and clearance volume V_c per cylinder at TDC. The minimum volume corresponds to the clearance volume. The compression ratio r_c then can be expressed as

$$r_c = \frac{\max V}{\min V} = \frac{V_d + V_c}{V_c} \tag{1}$$

Diesel engines usually are operated at high compression ratios to ensure autoignition of the mixture. Conversely, Otto engines have lower compression ratio, because unwanted uncontrolled autoignition, leading to shock waves inside the cylinder (engine knock), has to be avoided. Knock is favored at high pressures and temperatures. For typical numbers for r_c, refer to Table 1.

3.2 Power, work, and torque

The main important properties of ICEs are the actual power measured at the exit of the crankshaft, the brake power P_b [in units (W = Nm/s = J/s)], which is the usable power delivered. The brake power is related to torque T [in units (Nm)] and the angular or crankshaft rotational speed N [in units (1/s)] by

$$P_b = 2\pi \cdot N \cdot T \tag{2}$$

The torque, the work per unit of revolution, is given by the following expression

$$T = \frac{W_b}{n_R} \tag{3}$$

where $n_R = 2$ for four-stroke and $n_R = 1$ for two-stroke engines and W_b [in units (J)] is the brake work, respectively. The brake power relationship to brake work is therefore given by

$$P_b = 2\pi \frac{N}{n_R} \cdot W_b \tag{4}$$

The brake work can be expressed by the product of the input energy of the fuel mass ($m_f \cdot Q_{LHV}$) and the brake fuel energy conversion efficiency η_b

$$W_b = \eta_b \cdot m_f \cdot Q_{LHV} \tag{5}$$

with m_f being the input fuel mass [all masses are in units (kg)] and Q_{LHV} the fuel lower heating value [corresponding to the chemical energy stored in the fuel, in units (J/kg)].

Table 1. Typical values for compression ratio r_c, peak brake mean effective pressure *bmep*, peak mean piston speed S_p, peak-specific brake power related to either piston area $P_b/n_c A_p$ or displacement volume P_b/V_D, minimum specific fuel consumption sfc_b, and peak brake efficiency η_b for different engine types.

	r_c (−)	*bmep* (bar)	S_p (m/s)	$P_b/n_c \cdot A_p$ (kW/cm^2)	P_b/V_D (kW/cm^3)	sfc_b (g/kW h)	η_b (−)
Small two-stroke	8–15	8–12	18–22	0.65	220	260	35
Passenger car, Otto engine	10–11.5	10–15	14.5–24	0.45	45–55	230	35
Passenger car, Otto engine, turbocharged	8.5–12.5	13–23	15–19	0.53	80	240	35
Passenger car, Diesel engine, turbocharged	14–18	20–33	12.5–16.5	0.45	40–58	200	42
Heavy duty, Diesel engine, turbocharged	15–17	15–26	7–12	0.45	25–30	190	45
Large scale, slow two-stroke Diesel engine	12–17	20–28	6–8	0.7	10	150	54
Racing car engine, four-stroke Otto	10–16	25–35	20–28	2.3	300	250	
Racing car engine, four-stroke Diesel	13.5–15	25–45	15–16.5	2–4.2	80–110	220	

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

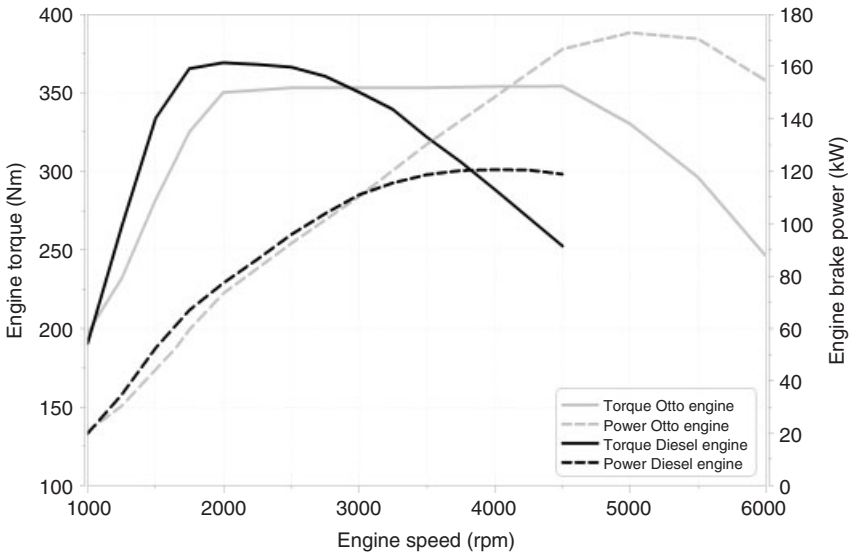


Figure 12. Typical curves for torque and power as a function of engine speed for both, an Otto engine (light gray curves) and a Diesel engine (black curves) of similar power and displacement volume, $V_D = 2\,\mathrm{L}$.

In Figure 12, an example for an Otto engine and a Diesel engine for brake power and torque versus angular speed is given. One recognizes the maximum values for both properties, occurring at different speeds. This is due to concurring processes depending on engine speed. At high speed, friction losses are high and the induction of a full charge of air into the cylinder becomes problematic, whereas at low speeds, heat losses are significant. In Table 1, typical values for the specific brake power for different engine types are given.

3.2.1 Air/fuel ratio

Introducing the AFR with the input air mass given by m_a

$$\mathrm{AFR} = \frac{m_a}{m_f} \tag{6}$$

the ratio of actual AFR to the stoichiometric $\mathrm{AFR_{st}}$ for a given mixture can be defined as

$$\lambda = \frac{\mathrm{AFR}}{\mathrm{AFR_{st}}} = \frac{m_a}{m_f \cdot \mathrm{AFR_{st}}} \tag{7}$$

and the equivalence ratio ϕ as

$$\phi = \frac{1}{\lambda} \tag{8}$$

These expressions relate the actual fuel mass m_f and air mass m_a to stoichiometric requirements. Stoichiometric conditions are when, theoretically, all the carbon and hydrogen bounds in the fuel molecules react completely with the O_2 in the air to convert all the fuel and O_2 into CO_2 and H_2O molecules. Typical numbers for λ for the different engine types are given in Table 2.

Table 2. Typical values for the air/fuel ratio as expressed by λ for different engine types.

Diesel engines	1.1	$< \lambda <$	6 . . . 10
Otto engines	0.7	$< \lambda <$	1.5
Otto engines ($\lambda = 1$ controlled)	0.99	$< \lambda <$	1.01
Stratified-charge Otto engines	1.2	$< \lambda <$	5

Then by substitution of Equation 7 into Equation 5 for the brake work, it follows that

$$W_b = \eta_b \cdot \frac{m_a}{\lambda \cdot \mathrm{AFR_{st}}} \cdot Q_{LHV} \tag{9}$$

3.2.2 Trapping efficiency

The trapping efficiency η_t denotes a relation between the intake air or the fresh gas mixture and the displacement of the engine. The notion trapping efficiency originally was used as a two-stroke engine metric. However, as in super- and turbocharged DI spark-ignition engines scavenging is also present, this efficiency is also commonly employed in four-stroke engines (Schmid *et al.*, 2011). The basic definition of the trapping efficiency η_t is given by

$$\eta_t = \frac{m}{m_A} = \frac{m}{n_c \cdot V_d \cdot \rho_A} \tag{10}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

with m being the charge mass in the cylinder (fresh gas per working cycle), m_A and ρ_A the ambient load mass and load density, respectively, n_c the number of cylinders, and V_d the displacement per cylinder.

For air inducting engines, such as Diesel or DI Otto engines, the load mass corresponds to the air mass in the cylinder

$$m = m_a \qquad \rho_A = \rho_a \tag{11}$$

whereas for engines inducing fuel/air mixtures, such as typical Otto engines, m and ρ_A are

$$m = m_a + m_f = m_a \left(1 + \frac{1}{\lambda \cdot \mathrm{AFR_{st}}}\right)$$
$$\rho_A = \rho_a \left(1 + \frac{1}{\lambda \cdot \mathrm{AFR_{st}}}\right) \tag{12}$$

Then it follows from Equations 9–12 that the brake work in both cases is

$$W_b = \eta_b \cdot \frac{\eta_t \cdot \rho_a}{\lambda \cdot \mathrm{AFR_{st}}} \cdot n_c \cdot V_d \cdot Q_{LHV} \tag{13}$$

For naturally aspirated engines, the air density ρ_a is approximated by atmospheric values. For turbocharged engines, values for pressures and temperatures are assumed to be equal to the corresponding values after the compressor or after the charge air cooler, as appropriate.

3.2.3 Charging efficiency

The charging efficiency η_{ch} denotes a measure of the cylinder fresh charge trapped mass m_c at the position when intake valve closes (IVCs). Similar to for m in Equations 11 and 12, it makes a difference between engines inducing air (Diesel engines) or fuel–air mixtures (Otto engines). Its definition reads

$$\eta_{ch} = \frac{m_c}{m_A} = \frac{m_c}{n_c \cdot V_d \cdot \rho_A} \tag{14}$$

3.2.4 Scavenging ratio

The scavenging ratio is a nondimensional characteristic property, which accounts for excess flow losses into the exhaust system during gas exchange. It is defined as

$$\eta_s = \frac{\eta_{ch}}{\eta_t} = \frac{m_c}{m} \leq 1 \tag{15}$$

The scavenging ratio is smaller than one when during exhaust valve opening times inducted fresh gas can exit the cylinder into the exhaust system. This is typical for piston-ported two-stroke engines.

3.3 Efficiency

3.3.1 Brake efficiency

Efficiency is defined as the ratio between work delivered and intake energy. In ICEs, the brake efficiency η_b describes the ratio of brake power output, for example, at the crank shaft, over the chemically energy contained in the fuel used during the combustion process

$$\eta_b = \frac{W_b}{Q_{in}} = \frac{W_b}{m_f \cdot Q_{LHV}} \tag{16}$$

3.3.2 Indicated efficiency

The indicated efficiency η_i of an ICE is defined as

$$\eta_i = \frac{W_i}{Q_{in}} = \frac{W_i}{m_f \cdot Q_{LHV}} \tag{17}$$

where the indicated work W_i corresponds to the integration over the full work cycle in the pV diagram (Figures 8 and 11) and represents the work that has been transferred to the piston face.

$$W_i = \oint_{\mathrm{cycle}} p\,\mathrm{d}V \tag{18}$$

For four-stroke engines, the integration is necessary for two crankshaft rotations, whereas for two-stroke engines, it only takes one revolution.

3.3.3 Gross indicated efficiency

For two- or four-stroke engines, if the pV integration is performed from BDC of the compression stroke to BDC of the expansion stroke (Figures 8 and 11), the result is the gross indicated work, W_{ig}, with the corresponding indicated gross efficiency η_{ig} being defined as

$$\eta_{ig} = \frac{W_{ig}}{Q_{in}} = \frac{W_{ig}}{m_f \cdot Q_{LHV}} \tag{19}$$

For the determination of W_{ig}, compression is assumed to start at BDC and expansion at TDC, regardless of differing opening times for the EVO and closing times for the IVC relative to BDC. This integral over the power cycle includes unused (lost) expansion work ΔW_E (calculated as the area of an extrapolated expansion pressure curve from EVO to BDC and the real pressure curve from EVO to BDC) and unused (lost) compression work ΔW_C (calculated as the area between the real pressure curve between BDC and IVC and an extrapolated pressure curve from IVC to BDC), illustrated in Figures 13 and 14 and discussed

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

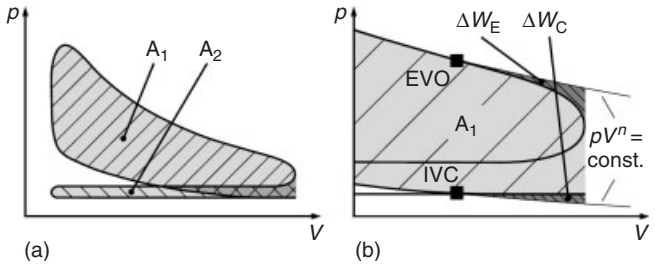


Figure 13. (a) pV diagram of a full four-stroke cycle. (b) Details explaining losses in work due to real start of compression and exhaust exit due to variable valve opening and closing times.

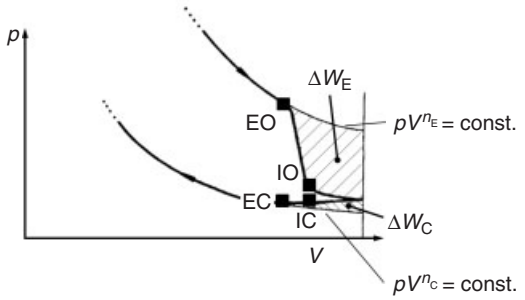


Figure 14. Details explaining losses in work due to real start of compression and exhaust exit due to variable valve opening and closing times in two-stroke engines.

additionally in Section 4.1. This definition is used in order to avoid dependencies of efficiencies on variable valve timing, which is a very common feature in today's ICEs. This dependency is considered with the gas exchange or pumping efficiency, which is discussed in Section 3.3.4.

3.3.4 Pumping efficiency

The pumping efficiency η_p is related to the work needed to do the gas exchange, called *pumping work* W_p. It covers the work needed to do the exhaust and intake strokes in four-stroke engines (Figure 8) and is determined by conducting a pV integration from BDC to BDC over the exhaust and intake strokes that include ΔW_E and ΔW_C explained in the last chapter. Thanks to this definition, which is valid also for any kind of variable valve timing, the integration is carried out from BDC to BDC. In two-stroke engines, W_p is solely the losses during compression, ΔW_C, and expansion, ΔW_E (Figure 14).

Pumping work is an energy loss mechanism to be minimized. The relationships between the pumping work and W_i and W_{ig} and the pumping efficiency and η_i and η_{ig} are given by

$$W_p = W_i - W_{ig} \qquad \eta_p = \frac{\eta_i}{\eta_{ig}} = \frac{W_i}{W_{ig}} = 1 + \frac{W_p}{W_{ig}} \tag{20}$$

3.3.5 Fuel conversion and combustion efficiency

Combustion efficiency describes the effects of incomplete and/or imperfect combustion. Incomplete combustion means that there are combustible components left in the exhaust gas, such as CO or H_2 present, for example, when chemical equilibrium is reached or when unburnt fuel remains in the exhaust for fuel-rich conditions. Imperfect combustion on the other hand describes nonequilibrium effects, such as local or global extinction of a flame, which leads to HC emissions and additional CO emissions. Usually, complete (or perfect) combustion never takes place.

Imperfect combustion is described by the fuel conversion efficiency η_f

$$\eta_f = \frac{Q_{in} - (Q_{HC} + Q_{CO})}{Q_{in}} = \frac{Q_f}{Q_{in}} \tag{21}$$

with Q_{HC} and Q_{CO} being the unreleased energy because of the presence of "additional" HC and CO emissions from imperfect combustion processes.

The HC and CO emissions due to incomplete combustion, for example, from rich combustion, are accounted for by the combustion efficiency term η_c. The fuel energy released during incomplete combustion is denoted by Q_c, and η_c is defined as the ratio of Q_c to the maximum possible energy release less that lost due to imperfect combustion, $Q_f = Q_{in} - (Q_{HC} + Q_{CO})$.

$$\eta_c = \frac{Q_c}{Q_f} \tag{22}$$

The losses due to incomplete and imperfect combustion could easily be measured (and/or calculated) and thus could easily be accounted for in the calculation of efficiency (Weberbauer *et al.*, 2005).

Using the assumptions, given in Vogt (Vogt, 1975), η_c can easily be determined from the AFR λ

$$\eta_c = 1.3373 \cdot \min(1, \lambda) - 0.3373 \tag{23}$$

or in a little more detailed form as given in Grill *et al.* (2007) by

$$\eta_c = 1 - \frac{Q_0}{Q_{LHV}} \cdot (0.415 \cdot L_{st} + 1) \cdot (1 - \min(1, \lambda))$$
$$Q_0 = 8.11 \text{ [MJ/kg]} \tag{24}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where $L_{st} = m_{a,st}/m_f$ describing the air mass needed for stoichiometric combustion.

3.3.6 Thermal efficiency and quality grade

Thermal efficiency, η_{th} (often also η_v), denotes the maximum possible gross indicated efficiency of the related ideal constant volume or constant pressure cycle for a given compression ratio. Therefore, it denotes the efficiency maximum, which for real cycles is never reached

$$\eta_{th} = \frac{W_c}{Q_c} = 1 - \frac{1}{r_c^{\gamma-1}} \tag{25}$$

with W_c being the maximum possible high pressure work of the related ideal constant pressure or volume cycle and $\gamma = c_p/c_v$ the adiabatic exponent given as the ratio of specific heat capacities at constant pressure and constant volume (which is assumed to be constant in the ideal process analysis). The thermal efficiency relates to the maximum releasable fuel energy, Q_c.

Thermal efficiency, fuel conversion efficiency, combustion efficiency, and gross indicated efficiency are related together by the definition of a combustion quality grade, η_g

$$\frac{\eta_{ig}}{\eta_{th}} = \eta_f \cdot \eta_c \cdot \frac{W_{ig}}{W_c} = \eta_f \cdot \eta_c \cdot \eta_g \tag{26}$$

The term η_g describes how much ignition timing, burn duration, the shape of the heat release rate, wall heat losses, leakage, and nonconstant gas properties influence efficiency.

3.3.7 Mechanical efficiency

Brake and indicated efficiency are related together by the mechanical efficiency η_m. η_m includes all losses in work between W_i and W_b, defined as W_f

$$W_f = W_i - W_b \tag{27}$$

Besides friction losses from the engine, W_f includes work that is needed to drive accessories, such as pumps, fans, and superchargers, as well as churning losses in the oil circuit or ventilation losses in the crank case. The mechanical efficiency reads

$$\eta_m = \frac{\eta_b}{\eta_i} = \frac{W_b}{W_i} = 1 - \frac{W_f}{W_i} \tag{28}$$

The overall system effectiveness could then be expressed as

$$\eta_b = \eta_m \cdot \eta_i = \eta_m \cdot \eta_p \cdot \eta_{ig} = \eta_m \cdot \eta_p \cdot \eta_f \cdot \eta_c \cdot \eta_g \cdot \eta_{th} \tag{29}$$

and brake power can be finally computed to be

$$P_b = 2\pi \cdot \eta_m \cdot \eta_p \cdot \eta_f \cdot \eta_c \cdot \eta_g \cdot \eta_{th} \cdot \eta_t \cdot \frac{\rho_a}{\lambda \cdot AFR_{st}} \cdot n_c \cdot V_d \cdot Q_{LHV} \cdot \frac{N}{n_R} \tag{30}$$

which is a general description of power output from ICEs.

Nearly all possibilities for increasing power can be derived from Equation 30. To increase power output of an ICE, the most effective measures are as follows:

- Increase in number of cylinders n_c or increase the piston displacement V_d
- Increase of intake air density ρ_a, for example, by turbocharging
- Increase of rotational speed N
- Increase of trapping efficiency η_t, for example, by an increase of charge by a vibration tube
- Increase of thermal efficiency η_{th} by increasing the compression ratio r_c
- Increase of mechanical efficiency η_m by reducing friction losses
- Reducing gas exchange work, increasing η_p
- Optimization of quality grade η_g.

4 SPECIFIC METRICS

One primary measure of ICEs is the piston displacement of a single cylinder V_d or of the whole multicylinder engine V_D. The piston displacement often is used to define specific metrics, which are commonly used to compare different engines.

4.1 Mean effective pressure

Definition of the brake, indicated, and friction mean effective pressures are

$$\text{imep} = \frac{W_i}{V_d} = \frac{1}{V_d}\oint_{\text{cycle}} p \cdot dV \qquad \text{bmep} = \frac{W_b}{V_d}$$
$$\text{fmep} = \text{imep} - \text{bmep} \tag{31}$$

Within the friction mean pressure fmep, not only losses due to friction inside the engines are included but also power needed for additional ancillary units. Measuring fmep is a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

complicated task, especially in turbocharged engines when mean pressures are high (<20 bar). Therefore, usually it is determined directly from the difference between imep and bmep. Typical values of bmep for different engines are given in Table 1.

In four-stroke engines, splitting the imep into a high pressure (gross indicated mean effective pressure) imep_g and a gas exchange part (pumping mean effective pressure) pmep is useful.

$$\text{imep}_g = \frac{W_{ig}}{V_d} = \frac{1}{V_d} \cdot \left[\left(\int_{BC}^{BC} p \cdot dV \right) + \Delta W_C + \Delta W_E \right]$$

$$\text{pmep} = \text{imep} - \text{imep}_g \tag{32}$$

ΔW_C denotes the work difference between real start of compression, when inlet valve closes, and the compression of an ideal cycle process starting at BDC. ΔW_E denotes the difference between a real exhaust exit, when the EVOs, and the ideal process followed by expansion at BDC (for details, see Figure 13).

For two-stroke engines, the mean pressure for the gas exchange process is directly determined from the losses during compression, $\Delta W_{C,}$ and expansion, ΔW_E (for details see Figure 14)

$$\begin{aligned} \text{pmep}_{2\text{-stroke}} &= \text{imep} - \text{imep}_g \\ &= \left[-\frac{1}{V_d} \left(\Delta W_C + \Delta W_E \right) \right] \Big|_{2\text{-stroke}} \end{aligned} \tag{33}$$

4.2 Specific fuel consumption

Instead of efficiency, often the specific fuel consumption *sfc* is used. The brake and indicated specific fuel consumptions sfc_b and sfc_i in units of (g [kWh]$^{-1}$) are

$$\begin{aligned} \text{sfc}_b &= \frac{m_f \cdot N}{n_R \cdot P_b} = \frac{1}{\eta_b \cdot Q_{LHV}} \\ \text{sfc}_i &= \frac{m_f \cdot N}{n_R \cdot P_i} = \frac{1}{\eta_i \cdot Q_{LHV}} \end{aligned} \tag{34}$$

using the following relations between power and efficiency

$$\eta_b = \frac{P_b}{\dot{m}_F \cdot Q_{LHV}} \qquad \eta_i = \frac{P_i}{\dot{m}_F \cdot Q_{LHV}}$$

$$\text{and} \qquad \dot{m}_F = \frac{m_f \cdot N}{n_R} \tag{35}$$

The indicated specific fuel consumption especially is used in single-cylinder engine research in order to estimate differences in combustion or gas exchange, in particular at low loads when unwanted changes of mechanical efficiency show big effects. Typical values for the specific fuel consumption for different engine types are given in Table 1.

4.3 Mean piston speed

The averaged piston speed is among others of importance for estimating the life of an engine. Sports car engines operate at mean piston speeds $>20\,\text{ms}^{-1}$, whereas large

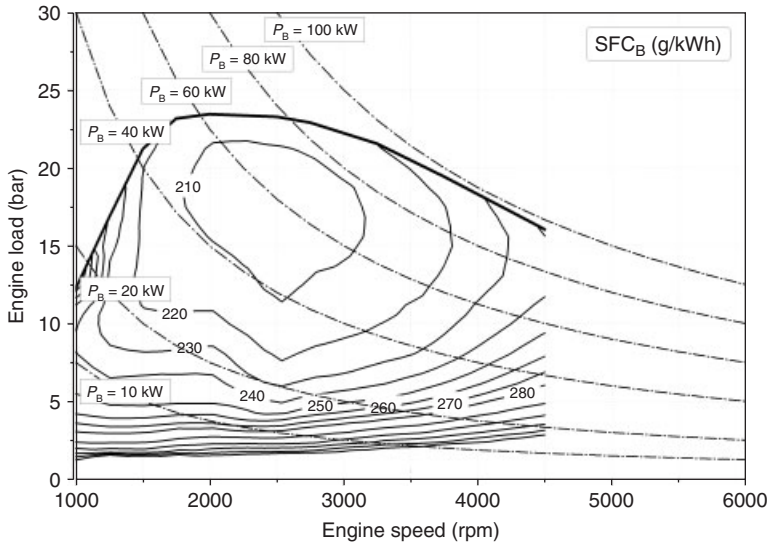


Figure 15. Typical characteristic map of specific fuel consumption and brake power for a Diesel engine ($V_D = 2\,\text{L}$). The isocurves denote either constant brake-specific fuel consumption (solid curves) or constant brake power (dashed-dotted curves).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

engines with long life times are operated below $10\,\mathrm{ms}^{-1}$. The piston speed also exerts strong influence on the in-cylinder fluid mechanics, which affects in-cylinder mixing processes and combustion rate. The averaged piston speed is defined as

$$\overline{S}_\mathrm{p} = 2 \cdot L \cdot N \tag{36}$$

where L being the stroke and N the rotational speed of the engine. Typical values for the mean piston velocity for different engine types are given in Table 1.

4.4 Engine characteristic maps

Engine characteristic maps describe a three-dimensional behavior of ICEs. Mainly engine-specific properties are plotted over rotational speed and engine load, with load being characterized by the torque T or brake mean pressure *bmep*. This allows a general description of the engine over the whole operating range. In Figures 15–17, characteristic maps of fuel consumption for a Diesel engine and two Otto engines (naturally aspirated and turbo charged) are shown.

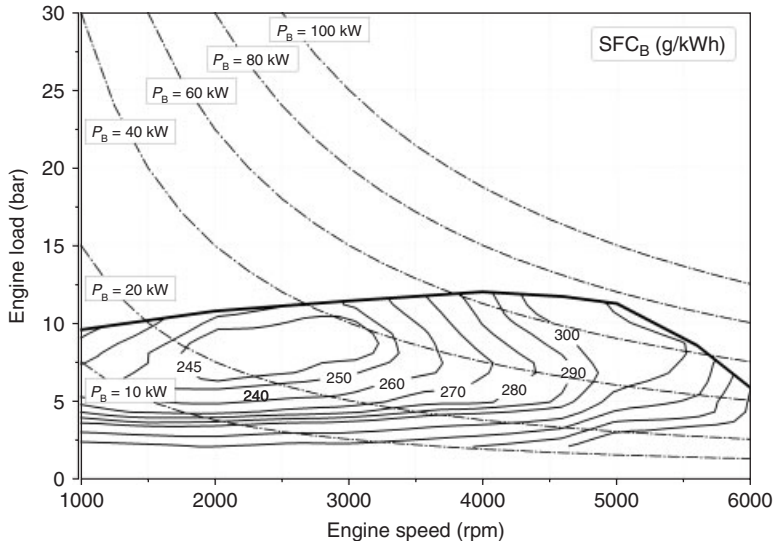


Figure 16. Typical characteristic map of specific fuel consumption and brake power for a naturally aspired Otto engine ($V_\mathrm{D} = 2\,\mathrm{L}$). The isocurves denote either constant brake-specific fuel consumption (solid curves) or constant brake power (dashed-dotted curves).

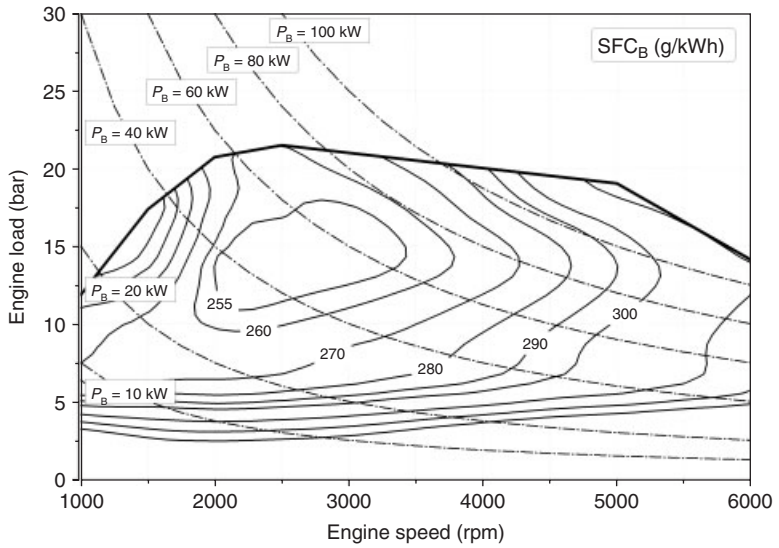


Figure 17. Typical characteristic map of specific fuel consumption and brake power for a turbocharged Otto engine ($V_\mathrm{D} = 2\,\mathrm{L}$). The isocurves denote either constant brake-specific fuel consumption (solid curves) or constant brake power (dashed-dotted curves).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

All engines have similar power and a displacement volume of 2 l. In the graphs, isocurves are plotted for constant specific fuel consumption and constant brake power. From these characteristic maps, the minimum fuel consumption can be obtained for a requested power.

REFERENCES

Blair, G. (1990) *The Basic Design of Two-Stroke Engines*, Society of Automotive Engineers, Warrendale, PA.

Diesel, R. (1892) Arbeitsverfahren und Ausführungsart für Verbrennungskraftmaschinen. An: ab 28. Febr. DRP Nr. 67207.

Diesel, R. (1893) *Theorie und Konstruktion eines Rationellen Wärmemotors zum Ersatz der Dampfmaschinen und der Heute Bekannten Verbrennungsmotoren*, Springer, Berlin, Heidelberg, New York. Reprint: Düsseldorf: VDI-Verlag 1986.

Diesel, R. (1913) *Die Entstehung des Dieselmotors*, Springer, Berlin, Heidelberg, New York.

Goldbeck, G. (1976) Die begründung der motorentechnik - die versuche von N. A. Otto 1860 bis 1876 *MTZ*, **37**, 109–114.

Grill, M., Schmid, A., Chiodi, M., Berner, H.-J. and Bargende, M. (2007) Calculating the Properties of User-Defined Working Fluids for Real Working-Process Simulations. SAE 2007-01-0936.

Heywood, J. (1988) *Internal Combustion Engine Fundamentals*, McGraw-Hill, New York.

Huf, F. (1961) Zur Geschichte der Rotationskolbenmaschinen. Automobil Revue.

Obert, F. (1973) *Internal Combustion Engines and Air Pollution*, Intext Educational Publishers, New York.

Sass, F. (1962) *Geschichte des deutschen Verbrennungsmotorenbaus von 1860–1918*, Springer, Berlin, Göttingen, Heidelberg.

Schmid, A., Grill, M., Berner, H.-J., and Bargende, M. (2011) Transient simulation with scavenging in the turbo spark-ignition engine *MTZ*, **71**.

Taylor, C.F. (1985a) *Internal Combustion Engine in Theory and Practice, Combustion, Fuels, Materials, Design*, vol. 1, 2nd edn Revised edn, The MIT Press, Cambridge, Massachusetts.

Taylor, C.F. (1985b) *Internal Combustion Engine in Theory and Practice. Combustion, Fuels, Materials, Design*, vol. 2, 2nd edn Revised edn, The MIT Press, Cambridge, Massachusetts.

Taylor, C.F. and Taylor, E.S. (1961) *The Internal Combustion Engine*, International Textbook Co, Cambridge, Massachusetts.

Vogt, R. (1975) Beitrag zur rechnerischen Erfassung der Stickoxidbildung im Dieselmotor. Dissertation. Technische Hochschule Stuttgart, Stuttgart.

Weberbauer, F., Rauscher, M., Kulzer A., *et al.* (2005) Allgemein gültige Verlustteilung für neue Brennverfahren *MTZ*, **66**.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto159
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Elastomeric Components for Noise and Vibration Isolation and Control in the Automotive Industry

Xiao-Ang Liu and Wen-Bin Shangguan

South China University of Technology, Guangzhou, China

1	Introduction	1
2	Stiffness and Damping Measurements	1
3	Rubber Technology	2
4	Engine Mounts	4
5	Suspension Bushings	9
6	Shock & Strut Mounts	10
7	Jounce Bumpers	12
8	Body and Subframe Mounts	13
9	Intermediate Driveshaft Mounts	15
10	Flexible Couplings	15
11	Mass Dampers	16
References		17

1 INTRODUCTION

Elastomeric isolators are widely used in industries to reduce noise and vibration. To use isolators effectively, the development engineer must design the isolators to satisfy multiple objectives, which typically include packaging restrictions, environmental criteria, motion control, load requirements, and minimum fatigue life, in addition to vibration isolation performance. An understanding of elastomeric material properties and the methods used to characterize elastomeric component behavior is necessary to achieve desired performance. In addition, the processing and preparation of rubber components and molding methods for elastomeric components are essential knowledge. Typical isolator used in automotive applications, including powertrain mounts, suspension bushings, shock-absorber bushings, flexible couplings, body and subframe mounts, and mass dampers are discussed in this chapter.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 STIFFNESS AND DAMPING MEASUREMENTS

A rubber isolator is usually simplified as three sets of elastic and damping elements in three perpendicular axes. The performances of a mount are evaluated by static and dynamic performances. The static property of a mount is the force versus displacement relations in the three perpendicular directions. The dynamic performances of a mount are usually evaluated with dynamic stiffness and loss angle (DSLA). A static force versus displacement relation is measured under quasi-static loading conditions, and the typical rates are between 10 and 25 mm per minute. In measuring DSLA of a mount, a preload corresponding to a part of the weight the isolator supported is applied firstly to one end of the mount, then a sinusoidal displacement excitation, $x(t) = X_0 \sin(\omega_0 t)$, is applied through the hydraulic actuator. The reaction force at another end of the mount, $F_T(t) = F_0 \sin(\omega_0 t + \phi)$, and the excitation displacement are acquisized at one time. The complex stiffness of the mount at $\omega = \omega_0$ is defined as, see Shangguan and Lu (2004a, b)

$$K^*(j\omega_0) = F[F_{\mathrm{T}}(t)]/F[x(t)]|_{\omega=\omega_0} = K_{\mathrm{s}} + jK_{\mathrm{l}} = K_{\mathrm{s}} + j\omega D \quad (1)$$

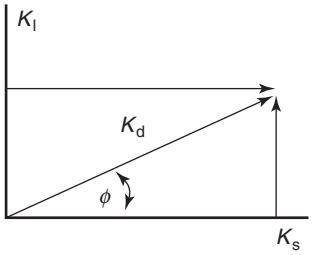


Figure 1. Phasor diagram for complex stiffness.

where F represents the Fourier transformations, K_s is the storage stiffness, K_l the loss stiffness, and D the damping coefficient. The dynamic stiffness, k_d, loss angle, ϕ, and damping coefficient, D, of a mount are calculated from

$$K_d = \sqrt{K_s^2 + K_l^2}, \quad \phi = \text{arctg}\left(\frac{K_l}{K_s}\right),$$
$$D = \frac{K_d \sin\phi}{\omega} \tag{2}$$

The definitions of the various types of stiffness identified respectively in the phasor diagram are shown in Figure 1 and are explained below:

Complex stiffness (K^*): It is the measured peak force divided by the measured peak displacement. The complex stiffness is the vector sum of the stiffness component in phase with the displacement and the stiffness component in phase with the velocity.

Storage stiffness (K_s): It is the proportionality factor between the displacement and the component of the force phasor that is in phase with the displacement.

Loss stiffness (K_l): It is the proportionality factor between the displacement and the component of the force phasor that is in phase with the velocity.

Loss angle (ϕ): It is the phase angle between the force and the displacement signals.

3 RUBBER TECHNOLOGY

Most vibration isolators in automotive industry are made of elastomeric compounds. Rubber is the general name for the existing variety of elastomeric materials, and the term *rubber* can refer to both natural and synthetic compounds.

Elastomers have the following three distinguishing characteristics, see Lewitzke and Lee (2001):

1. Resilience and energy storage—elastomers exhibit a high degree of recovery from large or small deformations through repeated cycling, returning approximately to the original dimension without suffering permanent damage, that is, resilience. Remarkable energy storage capacity and the ability to withstand repeated flexing attribute to rubber's utility as a spring material for mountings or suspensions for vehicles, machines, engines, and instruments.
2. Large deformability—elastomers exhibit the capability to withstand relatively high amounts of extension within the working range of the material, which, in extreme cases, can reach 300%, and the ultimate elongation may reach 1000% of the original length. Comparatively, steel may not be extended more than 0.5% of the original length without exceeding the elastic limit, and the ultimate elongation seldom exceeds 25%.
3. Low modulus—when exposed to large deformation, elastomers exhibit low stress levels. Rubber is seldom subjected to stresses >7 Mpa.

Natural rubber is produced by coagulating the latex, or sap, of the *Hevea brasiliensis* tree. Additional ingredients are also used in the formulations of both natural and synthetic rubber elastomers to provide specific performance characteristics, and the function of these ingredients is outlined in Table 1, see Lewitzke and Lee (2001).

Table 2 defines the major properties of the elastomers that are most commonly used in automotive vibration isolators.

Table 1. Additional ingredients used in the formulations of natural and synthetic rubber elastomers.

Ingredient	Comment
Base polymer	Natural rubber, poisoprene, butyl, and so on.
Carbon black	Reinforcement, hardness, dynamics, processability, and tensile strength
Sulfur	Cross-linking
Oils	Processability, dynamics, and hardness
Antioxidants and antiozonants	Prevent deterioration because of ozone and oxygen
Zinc oxide and stearic acid	Activate and stabilize the heat aging of the final mix
Accelerators	Speed up the curing process

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Properties of the elastomers used in automotive vibration isolators.

Elastomer	Properties	Applications
Natural rubber or polyisoprene (NR)	Available properties satisfy a broader range of engineering application than any other elastomer family. Excellent tensile strength and tear resistance	Powertrain mounts, suspension bushings, exhaust hangers, shock, strut mounts, front axle bushings, and rear differential mounts
Styrene-butadiene (SBR)	Reinforced or stiffer compounds offer properties only slightly lower than those of NR and IR, but more economical	Powertrain mounts and jounce bumpers
Synthetic isoprene (IR)	Similar to natural rubber. Slightly lower tensile strength and tear resistance	Powertrain mounts and suspension bushings
Poly-butadiene (BR)	Properties range a little below NR and IR. Resilience and low temperature flexibility better than NR and IR	Same as natural rubber
Polyurethane	Outstanding oil and solvent resistance; good impermeability; excellent aging; resistance to oils and gasolines; and ozone resistant	Body mounts, jounce bumpers, and suspension bushings
Polychloroprene	Moderate solvent resistance; excellent aging characteristics; and flame resistant. Approaches the broad engineering properties of NR and IR	Powertrain mounts and strut mounts
Butyl or polyisobutylene (IIR, CIIR)	Outstanding impermeability, chemically inert, excellent weathering resistance, high gum strength, high damping at moderate temperatures	Cradle and body mounts, jounce bumpers, and vibration dampers
Silicon (VMQ)	The highest and the lowest useful temperature ranges of all elastomeric compounds; superlative aging properties; radiation resistant; and reasonable oil resistance	Powertrain isolators and exhaust hangers

3.1 Processing and preparation of rubber components

A machine is used to mix the ingredients used in the formulation of elastomeric compounds. The ingredients are measured to a given weight and tolerance and then transferred to the Banbury mixer for mixing. After mixing, the rubber is dropped onto a twin-roll mill, where it is "kneaded" to induce a cross blending of the various elements that produces a homogeneous mixture. The rubber is then stripped off the mill and placed in a batch to cool. After cooling, the rubber can be sent either to an injection press or to an extruder for preparing slugs to a given weight for compression or transfer molding.

Gum rubber is a very viscous, partly elastic liquid, which has no practical engineering applications because the material has extremely poor tensile strength, elongation, and dynamic modulus and fatigue life. A vulcanization process is required to transform the rubber into a useful material.

The molecules of gum rubber are long and flexible. Vulcanization is a molecular cross-linking process that produces a molecular network by chemically linking together the independent chains of long molecules. The vulcanization process begins as soon as sulfur is added to the formulation, but time, temperature, and pressure are required to complete the process. The final formation of the cross-links occurs during molding, which is not only rapid but also controllable. Vulcanizates of natural rubber can be formulated to give good tensile strength and elongation over the complete range of hardness.

3.2 Measurement of the cure cycle

To verify that the desired performance attributes will be achieved during vulcanization, parameters that characterize the cure cycle during the molding process must be measured. The rheometer test instrument is used to measure those parameters.

The rheometer test instrument consists of a rotating grooved disk inside a closed cylindrical cavity. A test sample of uncured rubber is placed in the cavity to cure, while a disk embedded in the sample rotates in an oscillatory manner through a small arc. Torque is measured and recorded as a function of time on a rheometer chart. This chart will show initial viscosity, minimum viscosity, cure rate, and the final torque.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

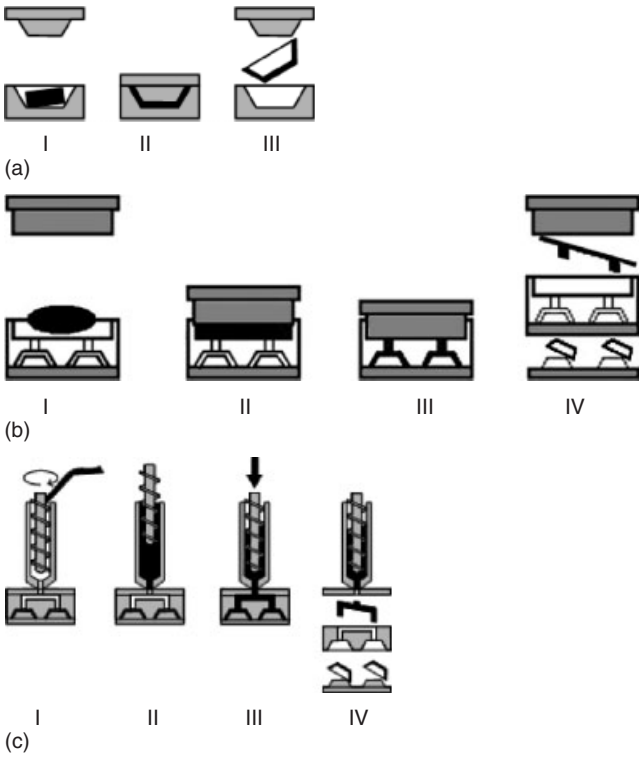


Figure 2. (a) Compression mold (semi-positive type), (b) transfer mold, and (c) injection molding.

3.3 Molding

There are three types of molding process used to manufacture rubber parts, which are described below, and the molding processes are shown in Figure 2.

1. Compression molding—a rubber slug is placed in the mold cavity. Pressure is applied as the lid is closed, which is maintained at a given temperature for a precise period of time.
2. Transfer molding—a rubber slug is placed in a pot located in the top part of the mold. As the press is closed, a ram forces the rubber through a sprue located in the bottom of the pot, and into the mold cavity.
3. Injection molding—a strip of rubber is fed into a holding pot. The holding pot contains a set of baffle gears that push a metered amount of rubber through a nozzle into the mold cavity.

Injection molding requires the shortest cure time and produces least amount of waste of the three molding processes. Table 3 provides comparative values for the time required for molding to full cure for a given part size and shape.

Table 3. Values for the time required for molding to full cure.

Modeling Type	Time
Compression molding	About 40 min
Transfer molding	20–30 min
Injection molding	4–8 min

Before molding, the metal inserts need to go through cleaning and adhesive coating operations to achieve good bonding of the rubber to metal. Cleaning operations consist of degreasing and wheel abrading. Primer and adhesive are applied to the metal parts in a two-coat operation, and then oven dried to prevent wipe-off of the primer/adhesive during molding.

After molding, pull tests are conducted, which measure the quality of the rubber-to-metal bond. Two criteria are typically used to qualify pull test results:

1. Ultimate pull force achieved before break, measured in newtons.
2. Visual Inspection—pull test results are qualified as a rubber failure, rubber/adhesive separation, and adhesive/metal separation. The acceptable pass condition is rubber failure only.

4 ENGINE MOUNTS

The functions of mounts are (i) to support the weight of the powertrain (weight support)—the powertrain weight should be well distributed among two to three load-carrying mounts; (ii) to control displacement of the powertrain center of gravity (CG) (motion control) under different quasi-static loads, such as maximum forward/reverse engine torque and shipping loads when the vehicle curies at different accelerations; (iii) to control powertrain vibration (vibration control) under tip in/back out (large torque change) and road impacts; (iv) to isolate chassis vibrations (vibration isolation) owing to the engine excitations at different operation conditions, such as idle, acceleration/deceleration, cruise, and switch on/off, see Yu, Naganathan, and Dukkipati (2001).

To satisfy with the design requirements of a powertrain mounting system (PMS, powertrain plus mounts), the mounts in a PMS are required to be soft (lower dynamic stiffness and lower damping) in order to isolate the engine excitation, and required to be hard (high stiffness and damping) in order to control the powertrain motion and vibration (see NVH Considerations in Engine Development). Therefore, the static force versus displacement (F–D) relation and DSLA of the mounts in a PMS should be designed carefully to meet the conflicting

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

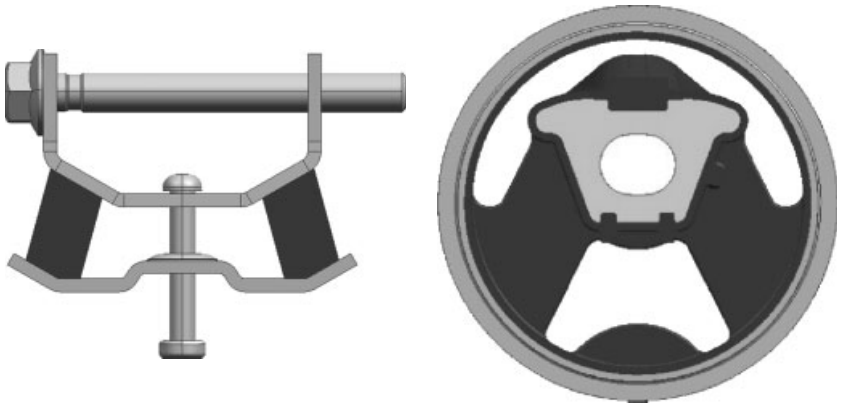


Figure 3. (a, b) Contours of complex mounts (wedge mount).

requirements for the mounts. Different kinds of mounts, from elastomeric mount (rubber mount) to hydraulic mount, and from passive mounts, semi-active mounts to active mounts, have been developed to improve the noise, vibration, and harshness (NVH) performance of a PMS, see Shangguan (2009).

4.1 Elastomeric mounts

Elastomeric mounts (rubber mounts) have been successfully used for powertrain mounts for many years. The contours of the rubber mounts are changed from standard parts (such as block or circular shapes) to complex mounts shown in Figure 3. The stiffness of the wedge mount in three axes can be adjusted by changing dimension and angle orientation of the rubber elements.

Bush-type rubber mounts used in a PMS are mainly used for controlling motion and vibration of a powertrain in the direction of crankshaft axis. Figure 4 shows a typical structure of a bush-type mounts.

The DSLA of a rubber mount under different excitation frequencies and excitation amplitudes can be found, see Shangguan and Lu (2004a, b). It is shown that the dynamic stiffness of a rubber mount under high frequency and small-amplitude excitations is larger than that under lower frequency and large-amplitude excitations. The damping of a rubber mount is relatively low, which could not meet the vibration control requirements for a powertrain under the ground excitations or output torque changes. The static F–D relation of a mount in one direction is usually nonlinear, and can be seen in Shangguan and Hou (2006).



Figure 4. Bush-type rubber mounts.

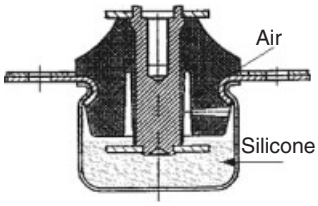
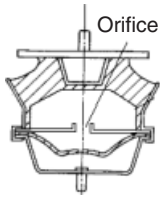
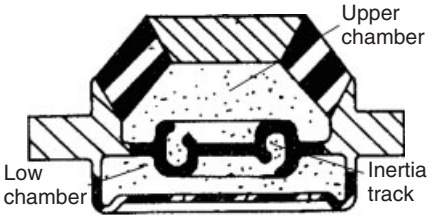
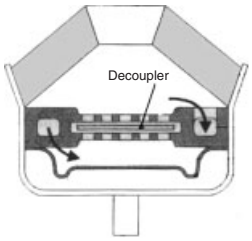
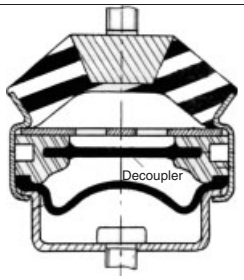
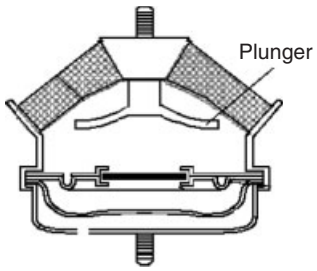
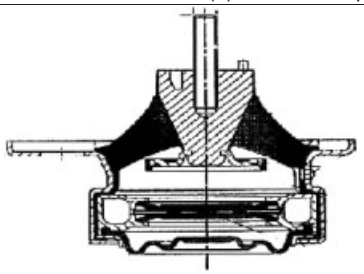
4.2 Hydraulic engine mounts

In 1962, Rasmassen patented a hydraulic engine mount (HEM) for increasing damping of a rubber mount, see Richard (1984). In 1985, all cars manufactured in GM are equipped with HEM, see Givers (1985). Today, HEMs are installed in almost all passenger cars. Moreover, it is validated that an HEM can improve the comfort and sound quality of a car, see Muller *et al*. (1994, 1996).

Different types of HEMs have been reported as shown in Table 4. The basic idea of an HEM is to use highly elastic rubber for vibration isolation and to use a hydraulic device

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Developments of hydraulic mounts.

Generation	Configurations		
First generation			
	(a) Viscous damped	(b) Orifice damped	(c) Inertia track damped
Second generation			
	(a) Movable decoupler	(b) Fixed decoupler	
Third generation			

(inertia track and decoupler) to generate the large damping at a prescribed frequency for vibration control. All types of hydraulic mounts reported in the literature are conceptually similar but different in detailed structure design. Table 4 gives the developments of hydraulic mounts.

4.2.1 *First-generation mounts: viscous-, orifice-, or inertia track-damped HEM*

A viscous-damped mount consists of a rubber spring and one chamber filled with a mixture of air and highly viscous fluids (e.g., silicone) or paste. The damping is created by forcing the silicon into the air gap, which makes the air is compressed and leads to bulge the rubber spring. This type of mount provides damping at broadband frequencies with a simple and cost-effective design. However, it is poor for isolating the engine small-amplitude excitations.

Orifice- and/or inertia track-damped HEMs: these two types of HEMs consist of two chambers, a rubber spring and a plate consisting of a fluid channels that forms an orifice or/and an inertia track. The chambers are completely filled with fluid. The upper chamber is a working chamber, and the pressure in this chamber changes greatly. The thickness of the rubber bellow in the bottom of an HEM is very thin that usually <3 mm. Therefore, the pressure change in the lower chamber approaches zero. The lower chamber is usually designed to absorb the transferred fluid. The large damping of an HEM is created by the fluid flowing from one chamber to another. The inertia effects of the fluid within the orifice or inertia track have great influences on the DSLA of an HEM at high frequency excitations; see Ushijima, Takano, and Kojima (1988).

4.2.2 *Second-generation mounts: hydraulic damped mounts with a decoupler and an inertia track*

The fluid in the inertia track cannot follow the excitation under the high frequency and small-amplitude excitations for the first-generation mounts, so the orifice or inertia track may be closed and the rising pressure in the upper chamber causes the increase in the DSLA of an HEM. To meet the conflicting requirements for vibration control and isolation of a powertrain, a decoupler is integrated with the inertia track to provide low stiffness for small-amplitude

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

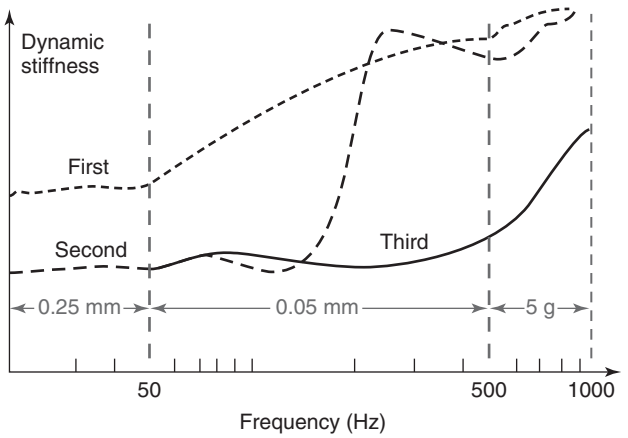


Figure 5. Dynamic stiffness of the three-generation HEMs under small-amplitude and high frequency excitations.

excitations, and to be blocked mechanically to generate large damping at large-amplitude excitations. The dynamic stiffness of an HEM without or with decoupler (the first- and second-generation HEMs) under high frequency and small-amplitude excitations can be found in Figure 5, see Tokushige, Funahashi, and Katayama (1999). It is seen that for the first generation of the HEMs, the dynamic stiffness increases if the excitation frequency is above 50 Hz, and the dynamic stiffness below 50 Hz is larger than that of the second-generation HEMs.

4.2.3 Third-generation mounts: inertia track-damped mounts with a decoupler and a plunger

If a component named "plunger" is incorporated in a second-generation mounts, a third-generation HEM is founded. The plunger subdivides the upper chamber and forms a ring-shaped channel against the chamber wall. The channel creates a high frequency inertia effect, which causes a decrease in dynamic stiffness. The frequency of the dynamic stiffness harden is determined by the radial clearance and the distance between the plunger and the rubber spring. In addition, the plunger is used as a restrictor. The dynamic stiffness of a third HEM under small-amplitude excitations can be found in Figure 5. It is shown that the frequency of dynamic stiffness harden is above 500 Hz, which is well above the frequency of dynamic stiffness harden for a second-generation HEM.

4.3 Torque struts

Torque struts are typically used in front-wheel-driven and transverse mounted powertrains (see Engine Configurations). A torque strut consists of a pair of elastomeric bushings or one elastomeric bush and one hydraulic bushing separated by a rod. A typical torque strut with one rubber bushing and one hydraulic bushing is shown in Figure 6. The function of a torque strut is to provide a torque against the powertrain pitch movement (a rotation around the engine crankshaft). In the design of a torque strut, the following factors need to be considered:

1. The resonance frequency of the mass-spring system consisting of the torque strut and bushing should be well above the fundamental orders of the engine firing frequency. This is accomplished using lightweight materials for torque strut rod and bushing housing, and by combining a soft bushing on one end of the torque strut and a stiff bushing on another end.
2. The torque strut applies a counteracting force to the load-carrying mounts in the longitudinal direction (front/after). Usually, the vertical stiffness of a mount increases rapidly with the increased longitudinal force, see Hofmann (2002). Accordingly, the stiffness and the positions of a torque strut should be designed with no consequential increase in the vertical stiffness of load-carrying mounts. An idea solution is to implement a longitudinal link with the load-carrying mounts. The link can support the longitudinal forces acting as torque strut but without increase the vertical stiffness of the mount. Such construction can be found in Figure 7.
3. In the assembly of a torque strut to a powertrain, care must be taken with the orientation of the torque strut. It will become unstable when it reacts a compression load that is along the torque strut rod direction. Even though the torque strut may be located such that it reacts forward drive loads in tension; under reverse torque conditions, torque strut will be loaded in compression, which may cause the torque to toggle over center.

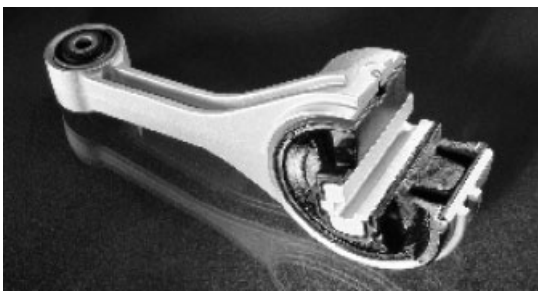


Figure 6. Torque strut.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

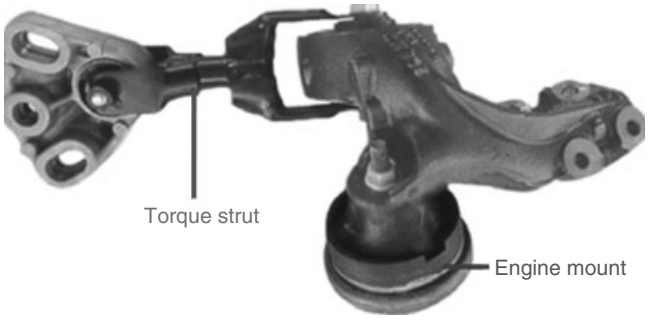


Figure 7. Integration of a torque strut and engine mount.

4.4 Semi-active mounts

A semi-active mount is an HEM with a mechanism to change the DSLA of the mount, and it is also called *switchable (on/off) mount*. A semi-active mount is used mainly for improving the car idle speed behavior. There are several types of semi-active mounts that are described below.

4.4.1 Vacuum-switchable type semi-active mount

The configuration of a vacuum-switchable type semi-active mount is shown in Figure 8. When a low stiffness is required in idling, the engine control module commands a signal to open a control valve, and a vacuum chamber is formed between the membrane and the rubber spring. The vacuum chamber decouples the fluid flow between the upper and the lower chambers through the inertia track. In driving conditions, the air in the vacuum chamber is sucked out, so the membrane is directly linked to the rubber spring and the mount behaves like a passive HEM with inertia track.

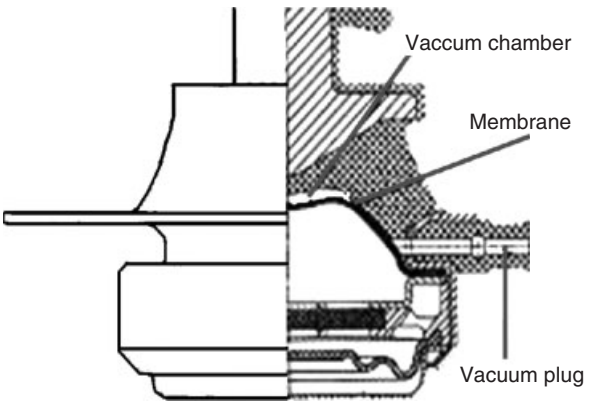


Figure 8. Vacuum-switchable type semi-active mount.

4.4.2 Orifice-controllable type semi-active mount

This type of semi-active mount consists of two chambers, a plate that forms an inertia track, and an orifice with a large diameter. The orifice can be opened or closed by a control valve. When a low stiffness is required, the orifice is opened, which allows the fluid to flow unrestricted between the two fluid chambers, resulting in a mount with low stiffness. When a high damping is required, the engine control module closes the orifice, causing fluid through the inertia track between two chambers, which produces high damping and stiffness in the mount.

4.4.3 Inertia track-controllable type semi-active mount

In some cases, especially with diesel engine, the idle vibration of a car is so significantly affected by high vibration levels transferred from the engine to the chassis through the engine mount, which means that a further reduction of the dynamic stiffness of a mount in the vertical direction is required.

The dynamic stiffness versus frequency cure of an HEM shows a notch before the stiffness increases because of resonance as Figure 9 indicates. This effect can be used to improve isolation of an engine in certain frequency ranges. The dynamic stiffness of an HEM near the notch frequency is far lower than the static stiffness of an HEM that is determined by its rubber spring. Experiments and simulations indicate that if the numbers of inertia tracks in an HEM increase, the notch frequency increases, see Zhang and Shangguan (2006). This concept is implemented in the design of a semi-active mount with controllable inertia track.

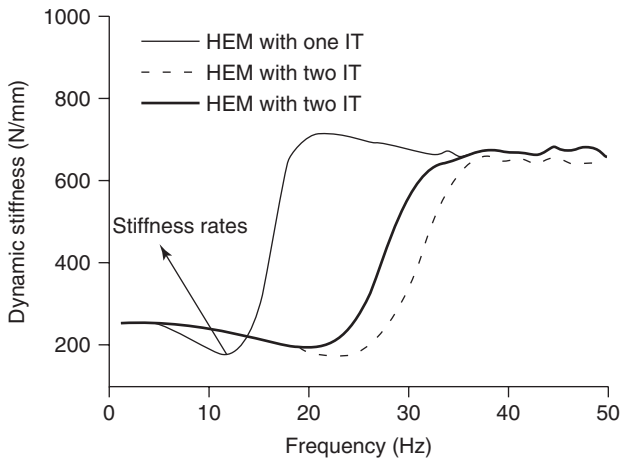


Figure 9. Dynamic stiffness of HEMs with different number of inertia tracks (IT: inertia track).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In this kind of semi-active mount, the HEM has two inertia tracks, one of which can be closed by an actuator. In idle, the control valve is open and makes the two inertia tracks work in parallel. Under this operation condition, the notch frequency of the mount is at a relatively high frequency, which results in minimum stiffness in the frequency at engine idle excitation (usually about 25 Hz), see Figure 9. By changing the inertia track sizes, the notch frequency can be adjusted. Under on-road condition, the control valve is closed and only one inertia track works. These result in large damping and stiffness at shake frequency as a conventional HEM.

4.5 Active mounts

An active mount consists of a passive mount (elastomeric or hydraulic mount), a vibration sensor, an electronic controller, and a force generating actuator. Active means that an active mount can suppress the structure vibration in actual driving conditions in very short time. Therefore, the controller for an active mount is typically implemented with the close-loop controller utilizing a linear sensor measurement. A sensor is mounted on the chassis side of the mount to measure the undesired vibration. A closed-loop control system is used to generate a secondary excitation force that is out of phase and equal in amplitude to the measured unwanted vibration, which is then applied by the linear actuator. Ideally, the secondary force cancels out the unwanted primary vibration, resulting in a zero level of transmissibility. An active mount requires significant external power and increases considerable cost and weight.

A simply type of active mount is to use electrorheological fluid (ERF) to replace ethylene glycol in a conventional HEM. The wall of the inertia track for a standard HEM is designed as electrode as shown in Figure 10. By applying high voltages, the viscosity of the ERF within the inertia track can be increased, which makes the same effect as reducing the size of the inertia track. If the voltages are high, the inertia track can be blocked. The dynamic performance of this mount can change from highly elastic nondamping mounts to highly damped mounts within one milli-second.

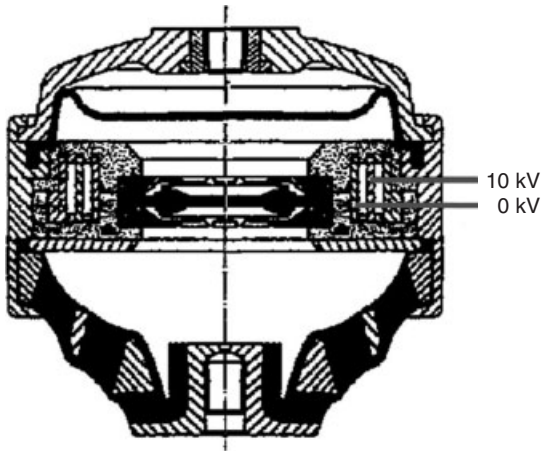


Figure 10. One type of active mount.

5 SUSPENSION BUSHINGS

Elastomeric bushings are used extensively in automobile suspension systems to withstand static, dynamic, and fatigue loads that occur under different road and driving conditions and to fulfill NVH and ride and handling targets of a vehicle. The bushings are functional parts of the control arms and are designed to attach the control arm to the front cradle as shown in Figure 11. The advantages of using elastomeric isolators in the suspension system are to eliminate the need to lubricate pivot joints. Table 5 shows suspension bushing usage.

Cylindrical bushing designs, as shown in Figure 12, are widely used in suspension systems to provide flexibility in torsion and tilt, as well as control for axial and radial displacements. The soft axial spring rate produces excellent isolation, and the high radial stiffness assures high stability. Figure 13 shows a cylindrical bushing in a trailing arm application.

A hydraulic bush in the rear control arm position shall be considered depending on vehicle requirement. It can provide large damping in axial or radial direction as shown in Figure 14. Figure 15 shows a suspension bushing with radial damping direction and vertical installation.

The dynamic characteristics of hydro bushing depend on its applications; the following gives some examples for application of hydraulic bush.

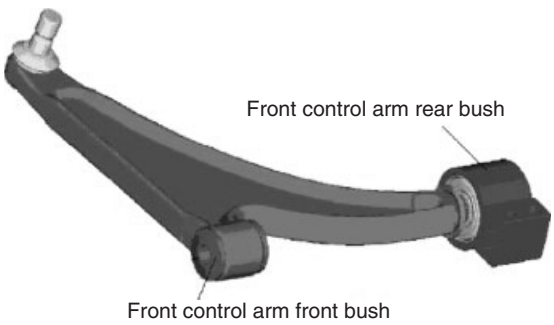


Figure 11. Bushings in a control arm.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 5. Bushing usage.

Type of Suspension	Bushing Locations
Solid axle, leaf—spring	Leaf-spring shackle
Twin I-beam—independent front suspension	Inner end of the axle; radius arm; and jounce bumper
Double wishbone—independent front suspension	Strut/shock mount; upper and lower; control arm; stabilizer bar; and strut bar
MacPherson strut—independent front suspension	Strut mount; jounce bumper; and anti-roll bar
Type 1 coil spring—independent front suspension	Lower control arm and upper control arm
Type 2 coil spring—independent front suspension	Lower control arm and upper control arm
Torsion bar—independent front suspension	Lower control arm; upper control arm; and eyehole pivot

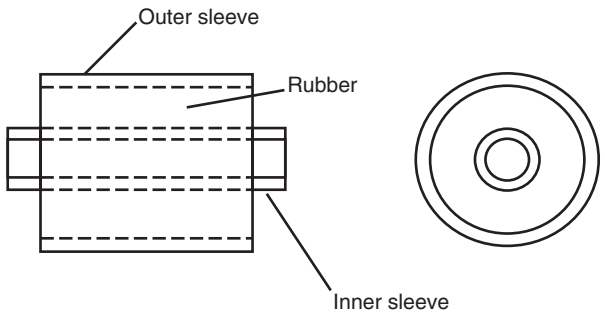


Figure 12. Cylindrical bushing design.

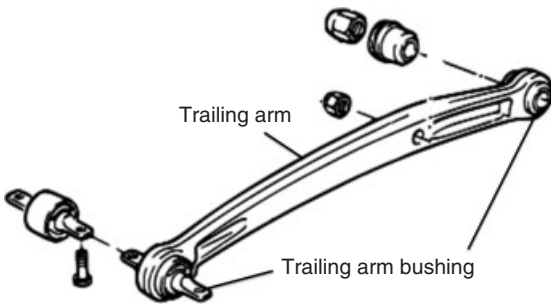


Figure 13. Trailing arm bushing.

One of the examples of the applications of hydro bushing is front lower control arm hydro bushings, which should suppress vibrations of the wheel suspension (wheel hop). Such disturbing vibrations lead to brake judder or steering wheel shimmy. Another application is twist beam hydro bushing at rear axle, which should reduce vibrations in fore/aft direction after driving over obstacles on the road. A third application of hydro damping is found at subframe bushings at front and rear suspensions. These hydro bushings reduce pitching vibrations of the subframes. All hydro bushings improve steering precision of the front and rear suspensions as well.

If a hydro bushing is replaced by a conventional bushing, one can gather the deterioration from the spider web diagram shown in Figure 16.

6 SHOCK & STRUT MOUNTS

The shock & strut mount is a component intended to provide isolation between the wheel suspension and shock absorber and the vehicle body structure, and provides position, motion control, damping, and load support of the shock absorber for the life of the vehicle. A typical



Figure 14. Hydraulic bushings.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

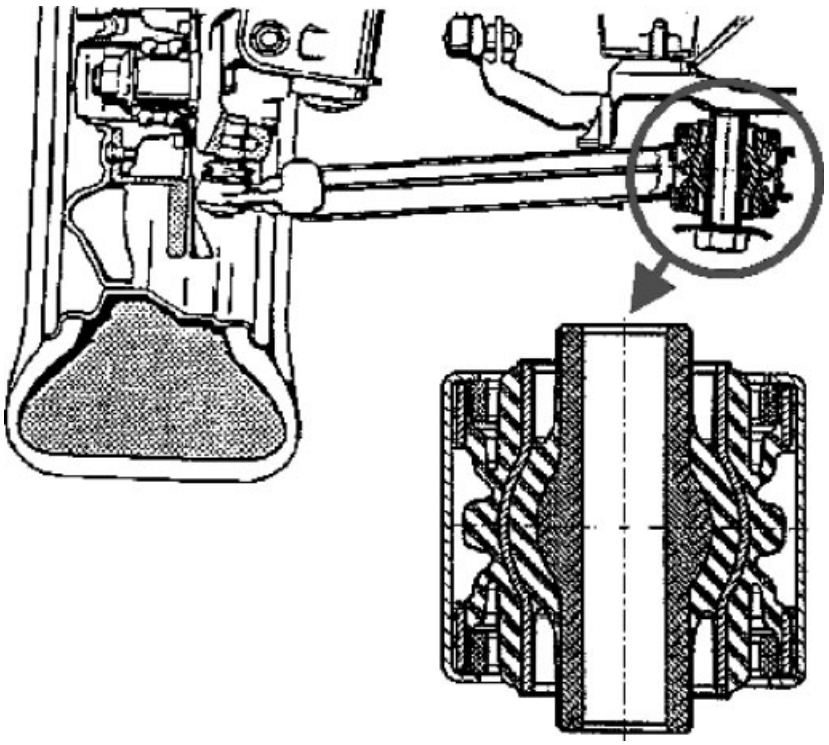


Figure 15. Suspension bushing with radial damping direction and vertical installation.

construction of a shock & strut mount is shown in Figure 17. The characteristics of the shock & strut mount is specified to obtain the correct behavior of the shock absorber under different load and driving conditions and to reduce the transmission of vibration forces from the shock absorber to the body/passenger compartment.

The flexibility of the shock & strut mount allows the strut angle to change to follow the travel of the lower ball joint, and the rubber reduces vibration and transmitted road noise. A bearing unit built into some mounts functions as the upper pivot point and forms the steering axis. When the front wheels are turned, the entire strut will pivot from the lower ball joint to the upper strut mount.

There are four types of strut & shock mount designs:

1. Single path shock & strut mount, which is shown in Figure 18: both the rod and the spring loads are isolated by the same rubber section. The rubber anchorage improves noise insulation. Initially, the deflection curve remains linear and then becomes highly progressive in the main work area, which is between 3 and 4 kN.
2. Dual path shock & strut mount, which is shown in Figure 19: the rod and the spring loads are isolated by different rubber sections. It provides good isolation, but is more expensive than the single path mount.

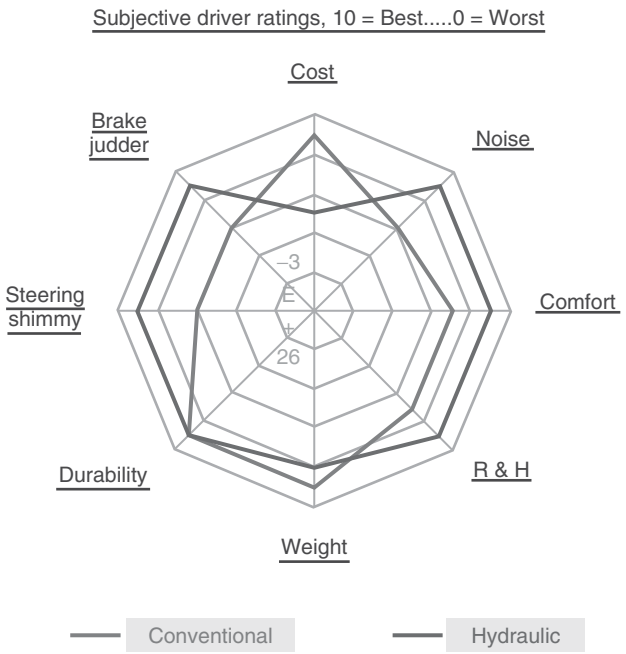


Figure 16. Comparison of conventional bushing and hydraulic bushing.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

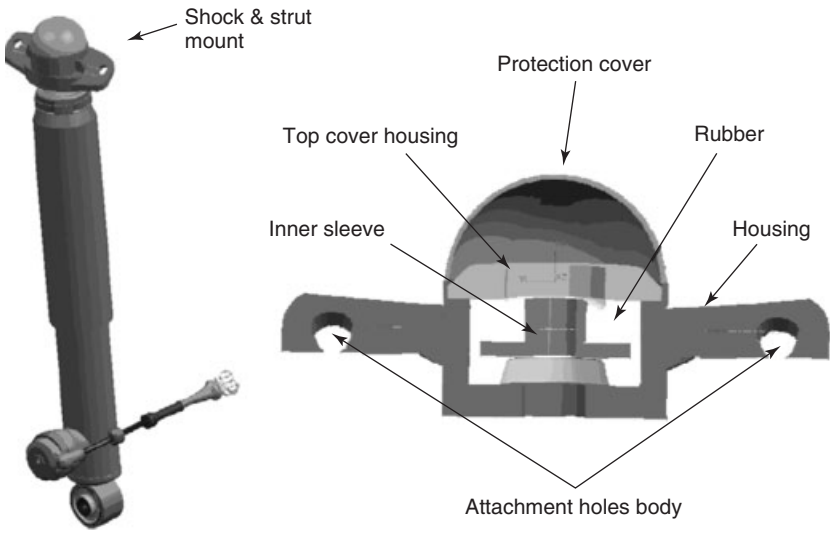


Figure 17. Construction of shock & strut mount.

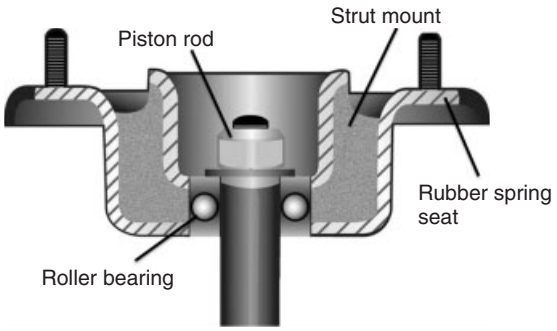


Figure 18. Single path shock & strut mount.

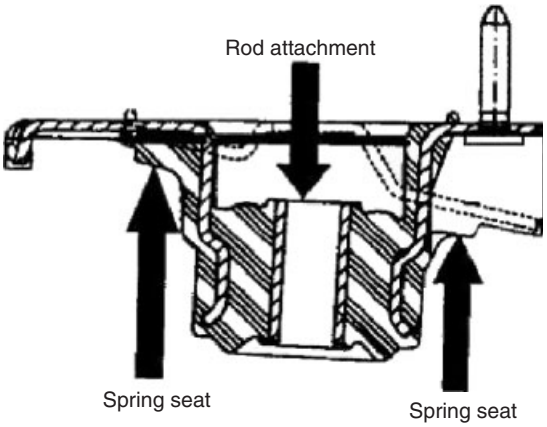


Figure 19. Dual path hock & strut mount.

1. Dual path with double isolation shock & strut mount, which is shown in Figure 20: the rod and the spring loads are isolated by different rubber sections as shown in Figure 20. The body spring and shock absorber forces are introduced into the body along two paths with variable rigidity.
   Inner path (1): The parts in inner path achieve good insulation from vibration and noise and improve the roll behavior of the body. High level of rigidity in a transverse direction is required for motion control.
   Outer path (6): The forces of the body springs are directed along the outer path, which has a considerably higher level of rigidity.

The force versus displacement for inner and outer paths is shown in Figure 21. It is seen that the stiffness in the inner path is much lower than that of the outer path.

Hydraulic top strut mounts shown in Figure 22 are also available on the market, not for damping but for isolation purposes. The fluid inertia effects within the mount are used to isolate tire-induced roll noise. By means of a dynamic decrease of stiffness, it is possible to reduce structure-borne noise through the strut mount.

7 JOUNCE BUMPERS

Jounce bumpers are typically installed on the chassis, strut assembly, control arm, or trailing arm assembly to increase the stiffness of the suspension system under high load conditions. They reduce spring stresses by transferring loads directly to the body or chassis under impact conditions. The material and shape of the jounce bumper establish the energy absorption capability of the design. These parameters also affect the maximum load transmitted

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

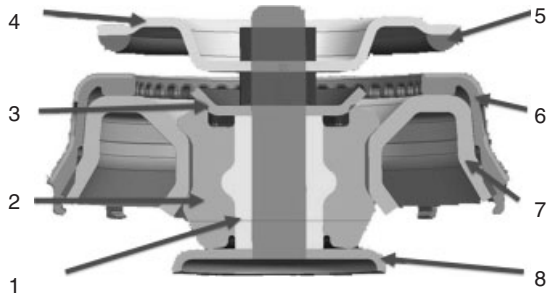


Figure 20. Dual path with double isolation shock & strut mount. 1 Inner Metal, 2 Primary/Shear Isolator, 3 Upper Rate Washer, 4 Reaction Washer, 5 Reaction Isolator, 6 Outer/Compression Isolator, 7 Main Stamping, 8 Lower Rate Washer.

to the body or frame during road impact events and the control of these loads will enable higher axle loading or conversely a bigger safety factor against overload.

The height of a given jounce bumper design is established by the difference between the maximum metal-to-metal clearance and the desired free travel. The amount of free travel is an important parameter that affects the ride characteristics of the vehicle. The maximum deflection of the bumper is the difference between the minimum clearance and the desired free travel. Figure 23 shows the jounce bumper design height and travel limitations.

Jounce bumper performance is evaluated by the following desired objectives:

1. High amount of energy absorption; magnitude of the peak force achieved during impact is minimized.
2. Gradual rate increase at initiation of contact (soft entry).
3. Low impact deflection; low rebound characteristics.

Butyl rubber and polyurethane are common jounce bumper materials, as these materials provide a high amount of energy absorption and the lowest rebound height. The bumper height, variations in taper, base area, bumper shape, and compound hardness are the other major parameters that affect the performance of a jounce bumper. The basic bumper shapes are shown in Figure 24.

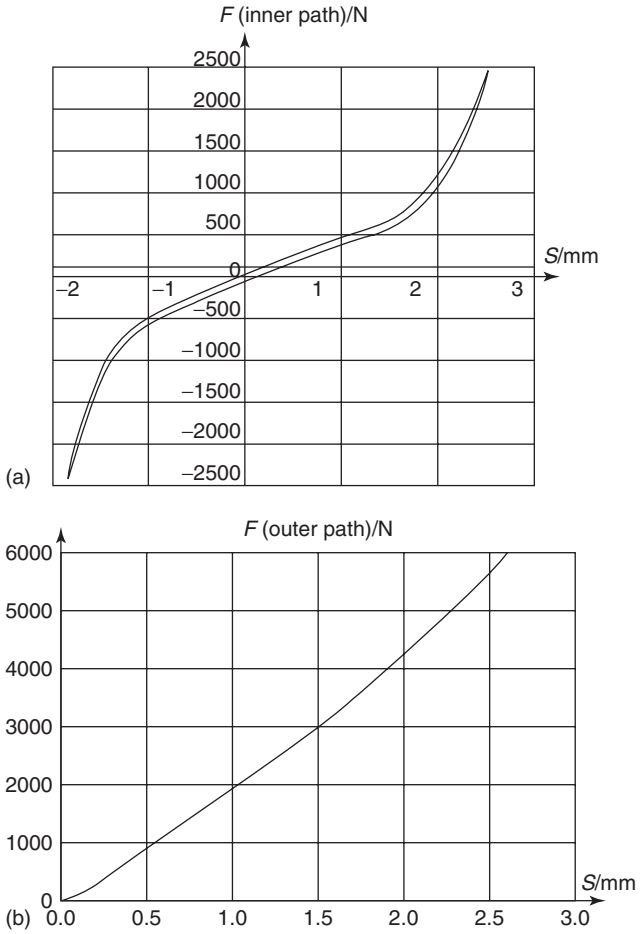


Figure 21. (a, b) Force versus displacement for inner and outer paths.

8 BODY AND SUBFRAME MOUNTS

8.1 Body mounts

Figure 25 shows the construction of a typical body/cab mount installation that consists of the load cushion, rebound cushion, the rebound cushion washer, and spacer sleeve. Body/cradle mounts are installed between the frame and the body/cab to isolate the passenger compartment from road noise, harshness, shake, and other vibrations in the chassis. The center bolt is tightened until the rebound cushion washer contacts the spacer sleeve, thus applying a compression force across both rubber parts. The load cushion supports the weight of the body. The external compression load increases the force on the load cushion and partially unloads the rebound cushion. The system behaves like two springs in parallel.

Body mounts must have the proper static rate to support the body at the proper height above the frame. Isolator dynamic rate must separate the natural frequencies of the body and frame from principal exciting frequencies, and the dynamic rate should be sufficiently low to isolate the vibrations. Isolator damping should be sufficiently high to control road-induced vibration. The polymer of the mount should have temperature stability and age resistance. Butyl rubber is used for most body/cradle mount applications, because butyl can provide more damping

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

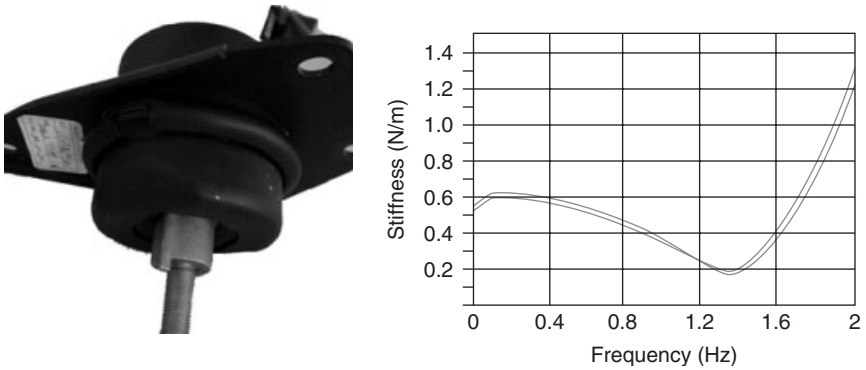


Figure 22. Hydraulic top strut mounts.

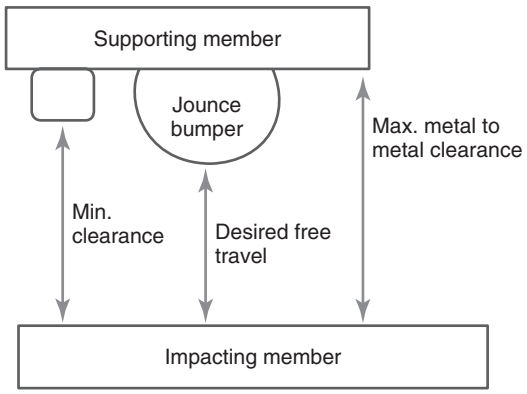


Figure 23. Jounce bumper traveling limitations.

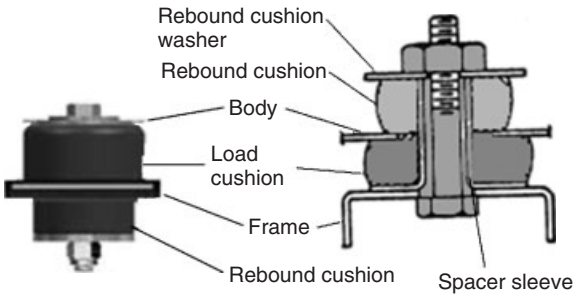


Figure 25. Body mount or cradle mount.

compared to natural rubber. Under body weight, the body/cab mount should not deflect more than 25% for durability considerations and set resistance. Under maximum operating conditions, the preload on the isolators should not allow either the load cushion or the rebound cushion to fully unload in order to maximize isolator durability of the mount and to maintain vehicle ride quality and stability.

8.2 Subframe mounts

The application of subframe designs enables the double isolation of structure-borne noise transmitted from engine to body, and thus absorption of noise emissions and provision of vibration comfort. The layout has to consider that the additional six degrees of freedom do not create any vibration problems.

The subframe normally contains four elastic components. In most cases, the mounts are pressed into the subframe and afterward assembled to the body. The elastic elements should provide a good isolation in vertical direction, but not decrease the handling performance. The corresponding designs are normally bush-type components that are soft in axial direction and that should be designed stiffer in longitudinal and lateral directions. Figure 26 shows a conventional design of a subframe mount.

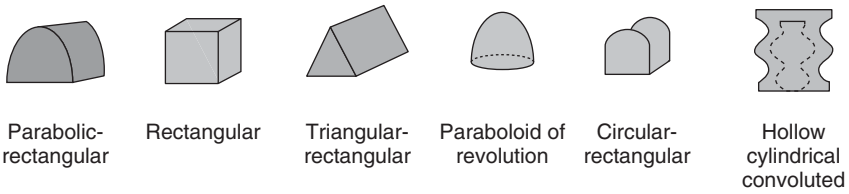


Figure 24. Basic jounce bumper shapes.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 26. Conventional subframe mount.

To avoid disturbing resonances of the elastically supported subframe, radial damper components are sometimes used to prevent lateral vibration amplification. Such a design is shown in Figure 27.

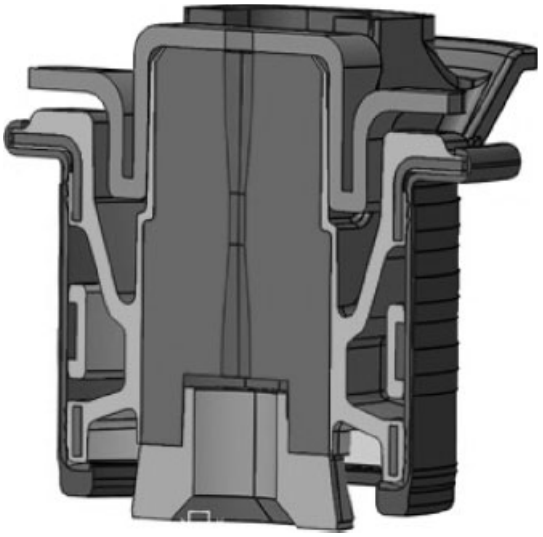


Figure 27. Hydraulic subframe mount.

9 INTERMEDIATE DRIVESHAFT MOUNTS

Rear-wheel-drive (RWD) driveshafts (propshafts) are frequently split into multiple components in order to keep the critical speed of the driveshaft below a desired limit. Two- and three-piece driveshafts require an intermediate driveshaft bearing-and-mount assembly to provide support. The flexible mounted bearing supports are shown in Figure 28. The rubber mounting provides the following two functions:

1. To accommodate a slight tilt of the shaft.
2. To act as a vibration damper and isolate any propshaft vibrations from the body.

The rubber-block bearing mount incorporates a rubber element that is mold-bonded to an external metal mounting flange and to the casing for the outer bearing race. The flexible bearing mount has a slot on each side of the rubber molding to improve flexibility. The V-section mount can fold and move about its mean position more readily as conditions demand. This also improves the vibration-damping properties of the rubber assembly. The double-fold rubber bearing mount also provides excellent damping properties, and provides greater rigidity for the bearing with increased articulation.

10 FLEXIBLE COUPLINGS

Flexible couplings are used to accommodate large angular and axial displacements with minimum resistance, reduce

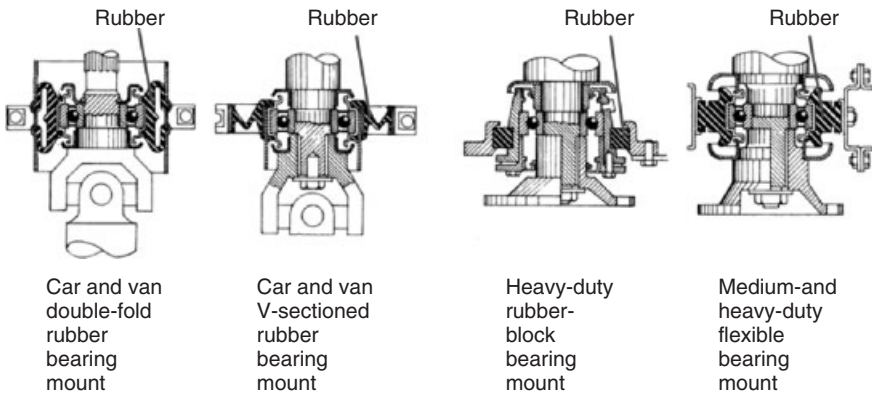


Figure 28. Intermediate driveshaft mounts.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

torsional vibration and noise, and absorb torque fluctuations. They can replace mechanical joints, eliminating the need for lubrication and averting metal-to-metal wear conditions. Often flexible couplings are used in driveshaft systems to provide torsional and axial flexibilities. When installed in an RWD driveshaft, a centering support device is also necessary. The rubber mounting in the coupling provides a flexible support for the bearing so that a slight tilt of the shaft can be accommodated. In addition, the flexible rubber element provides isolation for the body structure from propshaft vibrations.

Flexible coupling are used widely in the rear driveshaft for an all-wheel-drive (AWD) driveline, and in a front-wheel-drive (FWD) drive axle that accommodates axial and universal movements. It reduces loads on the gearbox casing and bearings and eliminates the need for sliding splines. The flexibility provided by the elastomeric coupling reduces driveline-induced noise and vibration (roughness) in the vehicle. Flexible couplings are also used on steering columns.

11 MASS DAMPERS

Often, in the situation of a noise or vibration problem that is caused by a resonant condition, it is not practical to reduce noise or vibration levels through isolation techniques alone. A concept that is frequently applied is the use of a mass damper, which utilizes the characteristics of a second-order system to obtain a noise or vibration reduction.

The model of a dynamic absorber is shown in Figure 29.

Commonly used vibration dampers include the control of bending modes in the transmission shaft by attaching the damper to the transmission shaft and the control of flexural bending modes of the axis by locating the damper in the axis as shown in Figure 30.

Another typical application of mass damper is the chassis absorber as shown in Figure 31. As attachment points for absorbers, areas are suitable where the strongest vibrations occur. Installed components such as batteries or pumps can also be used as absorber masses wherever possible.

Absorbers for rotational excitations are used in the drive line and on the end of the crankshaft and are known as torsional vibration dampers as shown in Figure 32. They considerably reduce the torsional resonance of the crankshaft, thus ensuring a long life.

Figure 33 shows variants of crankshaft torsional vibration dampers for passenger vehicle engines. Applications cover bonded and nonbonded solutions. In nonbonded versions, a pre-vulcanized rubber ring is inserted with a lubricant into

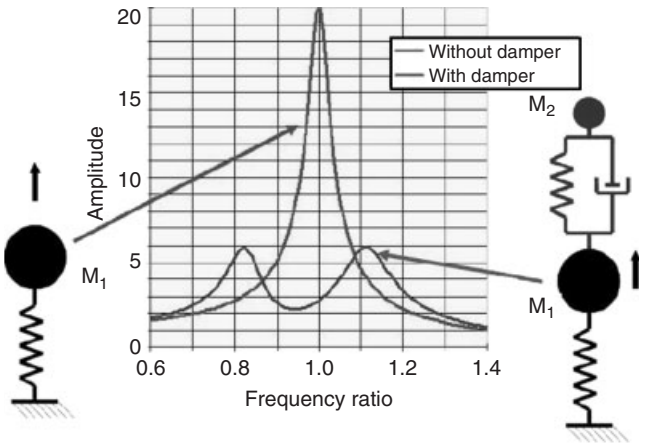


Figure 29. Vibration damper model.



Figure 30. Mass damper in transmission shaft.

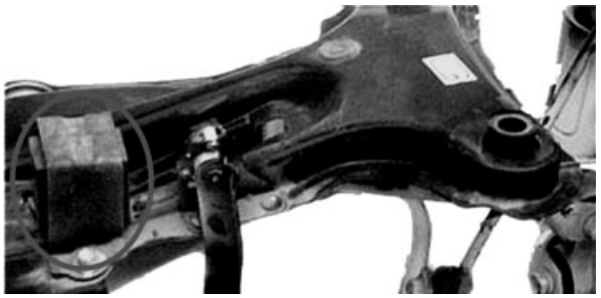


Figure 31. Mass damper in chassis.



Figure 32. Torsional vibration damper in crankshaft.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

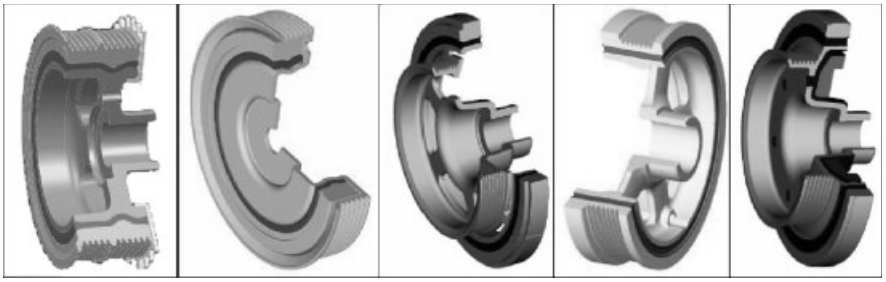


Figure 33. Several examples of torsional vibration damperst.

the gap between inertia and hub. The resultant precompression of the rubber ring enables the torsional vibration damper to transmit high forces without slipping.

REFERENCES

Givers, L. (1985) Technical highlights of the 1985 automobiles, *Automotive Engineering*, **92**, 39–51.

Hofmann, M. (2002) *Antivibration Systems: Fundaments, Designs and Applications*, Trelleborg Automotive Co. Ltd, http://www.mi-verlag.de.

Lewitzke, C. and Lee, P. (2001). Application of elastomeric components for noise and vibration isolation in the automotive industry. SAE Technical Paper Series 2001-01-1447.

Muller, M., Eckel, H.G., Leibach, M. *et al*. (1996). Reduction of noise and vibration in vehicle by an appropriate engine mount system and active absorbers. SAE Technical Paper Series 960185.

Muller, M., Weltin, U., Law, D. *et al*. (1994). The effect of engine mount on the noise and vibration of vehicles. SAE Technical Paper Series 940607.

Richard, A.M. (1984). Hydraulic mounts-improved engine isolation. SAE Technical Paper Series 840410.

Shangguan, W.-B. (2009) Engine mounts and powertrain mounting systems: a review, *International Journal of Vehicle Design*, **49**, 237–258.

Shangguan, W.-B. and Hou, Z. (2006) Strategies and calculation methods for automotive powertrain motion control under quasi-static loads, *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, **220**, 1131–1138.

Shangguan, W.-B. and Lu, Z.-H. (2004a) Experimental study and simulation of a hydraulic engine mount with fully coupled fluid structure interaction finite element analysis model, *Computers & Structures*, **82**, 1751–1771.

Shangguan, W.-B. and Lu, Z.-H. (2004b) Modeling of a hydraulic engine mount with fluid structure interaction finite element analysis, *Journal of Sound and Vibration*, **275**, 193–221.

Tokushige, M., Funahashi, Y., and Katayama, M. (1999) New technology of antivibration rubber products for automobiles (In Japanese), *Journal of the Society of Automotive Engineers of Japan*, **44**, 24–30.

Ushijima, T., Takano, K. and Kojima, H. (1988) High performance hydraulic mount for improving vehicle noise and vibration. SAE Technical Paper Series 880073.

Yu, Y.H., Naganathan, N.G., and Dukkipati, R.V. (2001) A literature review of automotive engine mount systems, *Mechanism and Machine Theory*, **36**, 123–142.

Zhang, Y.-Q. and Shangguan, W.-B. (2006) A novel approach for low-frequency performance design of hydraulic engine mounts, *Computers & Structures*, **84**, 572–584.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto160
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Gas Exchange—Breathing and Air Management

Francisco Payri, José M. Desantes, José M. Luján, José Galindo, and José R. Serrano

Universitat Politècnica de València, Valencia, Spain

1	Introduction	1
2	Quantitative Parameters	2
3	Flow through Orifices, Ports, or Poppet Valves	2
4	Valve Timing	5
5	Flow in Manifolds	7
6	Residual Gases and Exhaust Gas Recirculation	9
7	Volumetric Efficiency Curves	11
8	Pumping Friction Losses	13
Related Articles		15
References		15

1 INTRODUCTION

1.1 Objectives of the process

The objective of the gas exchange process is to expel the burnt gases from the engine cylinders and to replace them with fresh charge from the intake, thus permitting the subsequent cycle. These processes are called *exhaust* and *intake processes*.

From a wider point of view, the air management or handling includes not only the exhaust and intake processes but also all the processes taking place in the exhaust and intake lines. This involves the supercharging or turbocharging systems (see Turbocharging, Supercharging and Intake Boosting); the exhaust gas recirculation (EGR) and the exhaust gas energy management systems (see Exhaust Gas Energy Recovery). Moreover, the effects of aftertreatment (see Gas Aftertreatment Systems) and silencing systems (see NVH Considerations in Engine Development) in the exhaust, as well as those of charge air cooler and filter in the intake, have to be considered.

The gas exchange process is one of the key processes in the engine cycle for several reasons. First, the amount of fresh charge introduced into the cylinders limits the maximum amount of fuel that can be burnt and thus the mechanical power (see Operating Principles). Second, this process has an energy cost, the so-called pumping work (see Operating Principles), mainly because of pressure losses in the intake and exhaust paths. Third, the quality of the combustion chamber scavenging will influence the residual gas content and, in turn, the combustion process and emissions of the subsequent cycle. Finally, because of the sequential nature of intake and exhaust processes, pressure pulsations are generated at engine valves and travel through the intake and exhaust lines to the atmosphere, resulting in noise emissions (see NVH Considerations in Engine Development).

1.2 Four- and two-stroke engines

The main difference in the gas exchange process between four-stroke (4S) and two-stroke (2S) engines is that in the former the exhaust and intake processes are carried out in one complete engine revolution, whereas in the latter, they take only a fraction of an engine revolution. This makes the gas exchange process more complicated in 2S engines

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

than in 4S engines. First, the process has to be carried out over a smaller cycle period. Second, the exhaust and intake processes are simultaneous and it is difficult to control the charge composition.

1.3 Spark- and compression-ignited engines

The main difference between spark- and compression-ignited engines, as far as gas exchange is concerned, is the load control. In spark-ignited engines, as they may operate in a narrow range of equivalence ratio, the load is controlled by the amount of fresh charge (see Petrol Engines). Compression-ignited engines, however, may run within a wide range of equivalence ratios; therefore, the engine load is controlled by the amount of fuel regardless of the air quantity (see Automotive Diesel Engine Development Trends).

2 QUANTITATIVE PARAMETERS

Figure 1 shows the fresh charge balance during an engine cycle. The intake flow is split into two parts, the one that will be trapped in the cylinders and the other that will flow directly to the exhaust system (short-circuit). In addition, a fraction of the burnt gases from the previous cycle will remain within the cylinders as residuals for the subsequent cycle. Some parameters are commonly used to quantify the different effects.

Volumetric efficiency is calculated as the ratio of the intake mass flow to a reference flow needed to fill the swept volume at a given reference condition density, as Equation 1 shows.

$$\eta_v = \frac{m_{\text{delivered}}}{m_{\text{reference}}} = \frac{\dot{m}_a}{n i V_D \rho_a} \tag{1}$$

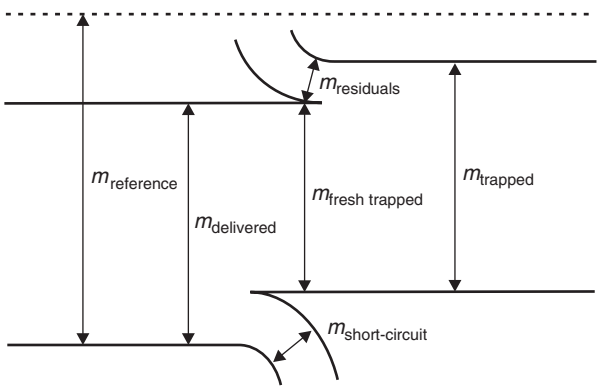


Figure 1. Charge balance during an engine cycle.

where $\dot{m}_a$ is the air mass flow, n is the engine speed, i is the number of cycles per revolution and takes the value of 1 in 2S engines and 0.5 in 4S engines, V_D is the total displaced volume, and ρ_a is the air density at a given reference conditions. This reference is the ambient in naturally aspirated engines and the boosting outlet conditions in supercharged engines (see Intake Boosting). The volumetric efficiency is commonly used in 4S engines to quantify the filling of the cylinders, because the amount of short-circuited mass is small in these engines. In 2S engines, the volumetric efficiency is usually called *delivery ratio*. Scavenging efficiency is the ratio between the fresh mass trapped in the cylinders and the total mass trapped, including residuals. Scavenging efficiency quantifies the proportion of residuals in the trapped mass.

$$\eta_s = \frac{m_{\text{fresh trapped}}}{m_{\text{trapped}}} = \frac{m_{\text{trapped}} - m_{\text{residuals}}}{m_{\text{trapped}}} = 1 - \frac{m_{\text{residuals}}}{m_{\text{trapped}}} \tag{2}$$

Trapping efficiency is the ratio between the fresh trapped mass and the delivered mass. Trapping efficiency quantifies the proportion of short-circuited mass in the delivered mass.

$$\eta_T = \frac{m_{\text{fresh trapped}}}{m_{\text{delivered}}} = \frac{m_{\text{delivered}} - m_{\text{short-circuit}}}{m_{\text{delivered}}} = 1 - \frac{m_{\text{short-circuit}}}{m_{\text{delivered}}} \tag{3}$$

3 FLOW THROUGH ORIFICES, PORTS, OR POPPET VALVES

The gas flow through singularities such as orifices, ports, or poppet valves, and the discharge to a cylinder or duct is a complex three-dimensional problem. Characterization of the permeability of these elements can be done using computational fluid dynamics tools described in Multidimensional Simulation or experimentally using a steady flow test rig schematically shown in Figure 2. A discharge coefficient is defined as the ratio between the actual (measured) mass flow rate ($\dot{m}$) and a reference mass flow rate ($\dot{m}_s$), as Equation 4 shows. Obviously, the choice of the reference mass flow rate will have a strong link to the value of C_D.

$$C_D = \frac{\dot{m}}{\dot{m}_s} \tag{4}$$

The reference mass flow rate, calculated for an isentropic compressible flow, will be obtained from Equation 5, using

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

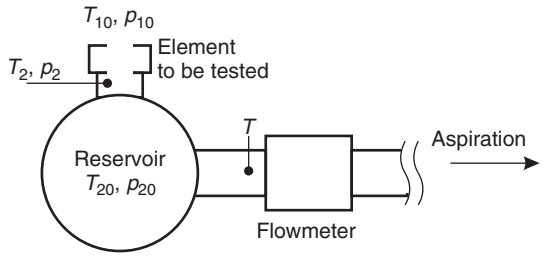


Figure 2. Schematic drawing of a steady flow test rig.

the nomenclature of Figure 2 where h is specific enthalpy; γ is the adiabatic index; R is the perfect gas constant; p is the pressure; T is the temperature; A_r is a reference or geometric flow area; and A_r is a reference flow area.

$$\begin{aligned}\dot{m}_s &= \rho A_r \sqrt{2(h_{10} - h_2)} \\ &= A_r \frac{\gamma p_{10}}{\sqrt{\gamma R T_{10}}} \left[\frac{p_2}{p_{10}}\right]^{1/\gamma} \\ &\quad \times \sqrt{\frac{2}{\gamma - 1}\left[1 - \left(\frac{p_2}{p_{10}}\right)^{(\gamma-1)/\gamma}\right]}\end{aligned} \tag{5}$$

The flow through the singularity for real engine conditions is adiabatic and unsteady, but steady flow test rigs under cold flow operation are considered, in spite of their simplicity, quite representative of the actual discharge coefficient at a given valve lift (Desantes, Benajes, and Urchueguía, 1995). Two reasons could be argued for this: the heat exchanged between the gas and the boundaries during the flow through the singularity is small compared with the total available energy of the flow (Ward-Smith, 1980); and flow velocity is usually much larger than valve (or piston) velocity so that quasi-steady conditions can be assumed (Benson, 1982).

When the flow is choked, that is, Equation 6, the flow velocity equals the local speed of sound and the mass flow rate becomes dependent only on upstream conditions, as shown in Equation 7. Under these conditions, Mach number equals one close to the narrowest cross section of the element under study.

$$\frac{p_2}{p_{10}} \leq \left[\frac{2}{\gamma + 1}\right]^{\frac{\gamma}{\gamma-1}} \tag{6}$$

$$\dot{m}_s = A_r \frac{\gamma p_{10}}{\sqrt{\gamma R T_{10}}} \left(\frac{2}{\gamma + 1}\right)^{\frac{\gamma+1}{2(\gamma-1)}} \tag{7}$$

The reference area, A_r, is easy to identify as the open area when considering a port, where it is also usually easy to calculate. However, when referring to the open area of a poppet valve, the choice is not as straightforward and, in fact, depending on the valve lift L_v (Figure 3), the minimum area exposed to the flow will change from the valve–seat interface to the port section.

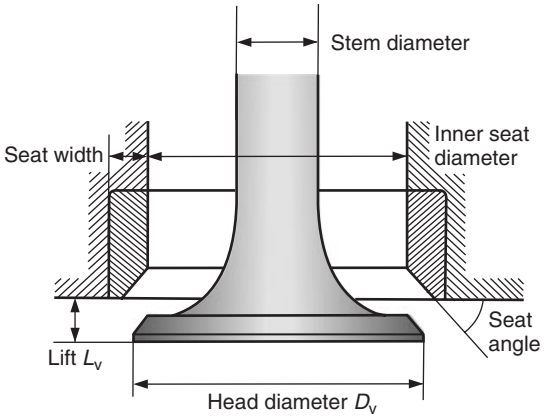


Figure 3. Parameters defining poppet valve geometry.

Two simple reference areas commonly used are the so-called *curtain area* $= \pi L_v D_v$ and *seat area* $= \pi D_v^2/4$. Figure 4 shows the experimental discharge coefficient versus the ratio L_v/D_v when using the two reference areas defined earlier. The discharge coefficient referred to the curtain area increases initially at low valve lifts and decreases for values of normalized valve lifts greater than 0.1. This nonlinear trend is due to the definition of the reference area, which increases linearly with L_v, and to the flow patterns through the valve–seat passage (Annand and Roe, 1974). For low valve lifts, the flow is attached to the walls, whereas beyond normalized valve lifts of 0.1, the flow pattern changes to a free discharge with detached flow. The discharge coefficient referred to the port area explains why actual engines are designed with maximum L_v/D_v

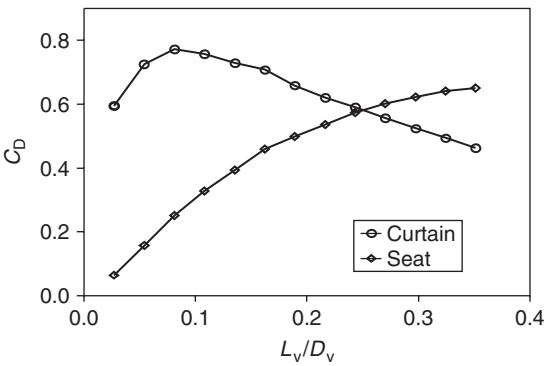


Figure 4. Discharge coefficient versus normalized valve lift.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ratios in the range 0.25–0.3. Increasing valve lift beyond this value has little influence on the discharge capacity of the valve, whereas mechanical loads on the valve train system do increase.

Inlet and exhaust valves and ports are usually the smallest cross sections in the intake/exhaust system, especially when they are not completely open. Consequently, highest flow velocities and flow compressibility effects are most noticeable at these elements, increasingly so as engine rotational speed increases.

The nondimensional parameter used to quantify the importance of the compressibility effects is the Mach number, that is, the ratio between flow velocity and the local speed of sound. During the intake (or exhaust) stroke, flow velocity in the port or valve is continuously changing because of variable thermodynamic conditions up- and downstream and a variable opening section, as Equation 5 reflects. Wave action models (Benson, Horlock, and Winterbone, 1986) are able to link the thermodynamic port conditions to the in-cylinder conditions with crank-angle resolution during the entire engine cycle making use of the discharge coefficient defined in Equation 4.

Intake flow, neglecting dynamic effects in the manifold, is controlled by the pressure difference across the intake valve generated by the piston velocity during its descending stroke (Benson, 1982). To simplify the problem and to have an overall estimator, different cycle-averaged Mach numbers have been defined. Taylor (1985) proposed the following index for the intake process

$$M_T = \frac{A_p \cdot S}{\overline{C}_D \cdot A_r \cdot a} \tag{8}$$

where A_p being the piston area, S the averaged piston speed, a the local speed of sound, and $\overline{C}_D \cdot A_r$ the average effective flow area of the intake valve during the intake stroke.

Figure 5a shows the volumetric efficiency of different engines versus the Taylor's Mach index. All the curves collapse on a single curve, which means that M_T is a good universal estimator. Figure 5 shows that for $0.3 < M_T < 0.6$ the volumetric efficiency is kept almost constant, suggesting a small effect of compressibility, whereas for values higher than 0.6, a clear deterioration of the volumetric efficiency is observed.

Fukutani and Watanabe (1982) proposed an alternative inlet Mach index, given by Equation 9, corrected with the volumetric efficiency (η_v) and the intake opening duration in crank angle degrees (IVC–IVO), where IVC and IVO are the inlet valve closing and inlet valve-opening crank angles, respectively

$$M_{FW} = M_T \frac{180}{IVC - IVO} \eta_v \tag{9}$$

The plot of the volumetric efficiency versus this corrected index appears in Figure 5b. A sudden reduction of volumetric efficiency is observed for values of $M_{FW} > 0.4$. The reduction of volumetric efficiency with increasing intake Mach index is due to the appearance of choked conditions in the valve seat passage during the intake stroke. Both indexes defined earlier are sensitive to this effect but probably the second index is able to establish a more precise limit for design purposes.

The three ways to keep a high volumetric efficiency at high engine speed are refining the design of the intake pipe-valve-seat arrangement, which in turns means a higher discharge coefficient; properly selecting the IVC and IVO; and, of course, increasing the reference area, usually by increasing the number of intake valves.

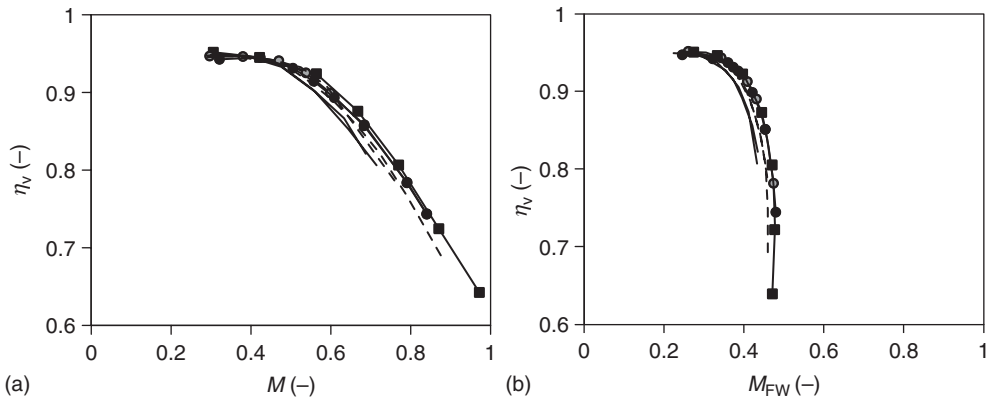


Figure 5. Volumetric efficiency versus intake Mach number for several engines: (a) Taylor definition and (b) Fukutani–Watanabe definition.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

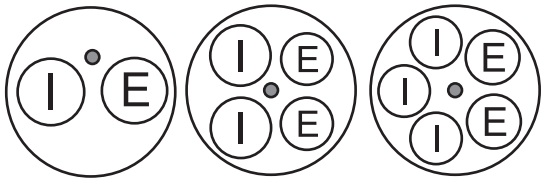


Figure 6. Schematic drawing of different cylinder head arrangements.

Exhaust flow, neglecting dynamic effects in the exhaust manifold, usually has two different periods (Figure 15). During the blowdown period, extending from exhaust valve opening (EVO) to bottom dead center (BDC) approximately, the flow is choked and controlled by the thermodynamic conditions in the cylinder. Around BDC, in-cylinder pressure has dropped to values similar to the exhaust back pressure, and from BDC to exhaust valve closing (EVC) near top dead center (TDC), exhaust flow through the valve is controlled by the piston movement during its ascending stroke, usually with Mach numbers lower than one (subsonic flow) during all this period.

Figure 6 shows schematic drawings of two-, four-, and five-valve cylinder head arrangements with values for the available reference intake (A_{IN}) and exhaust (A_{EX}) areas normalized to the piston area. Four-valve cylinder heads allow an increase of more than 40% in the available intake area and more than 20% in the exhaust area, whereas five-valve arrangements have little influence on the intake area but allow almost 20% more exhaust area.

Figure 7 shows the influence of the available intake and exhaust areas on volumetric efficiency and pumping losses versus engine speed. Figure 7 shows three curves in each chart, the first curve is for a reference effective flow area of an engine with one intake valve and one exhaust valve; the second curve corresponds also to an engine with two valves in the cylinder head but with an effective flow area 20% higher in each valve; and the third curve is for an engine with four valves in the cylinder head that corresponds to an effective flow area about 40% higher than the first configuration. From Figure 7, it is apparent that the intake flow area, and accordingly the intake system arrangement, has a large influence on the volumetric efficiency but less influence on the pumping losses of the engine. The exhaust process and the flow area through the exhaust valve on the other hand are quite irrelevant from the point of view of the volumetric efficiency of the engine, with only minor influence at very high engine speed due to the increase in residual gases. However, exhaust flow area is key in evaluating the pumping losses as explained in Section 8.

4 VALVE TIMING

The cam profile design directly affects the air management process and therefore the volumetric efficiency of the engine. The ideal valve operation would be the one

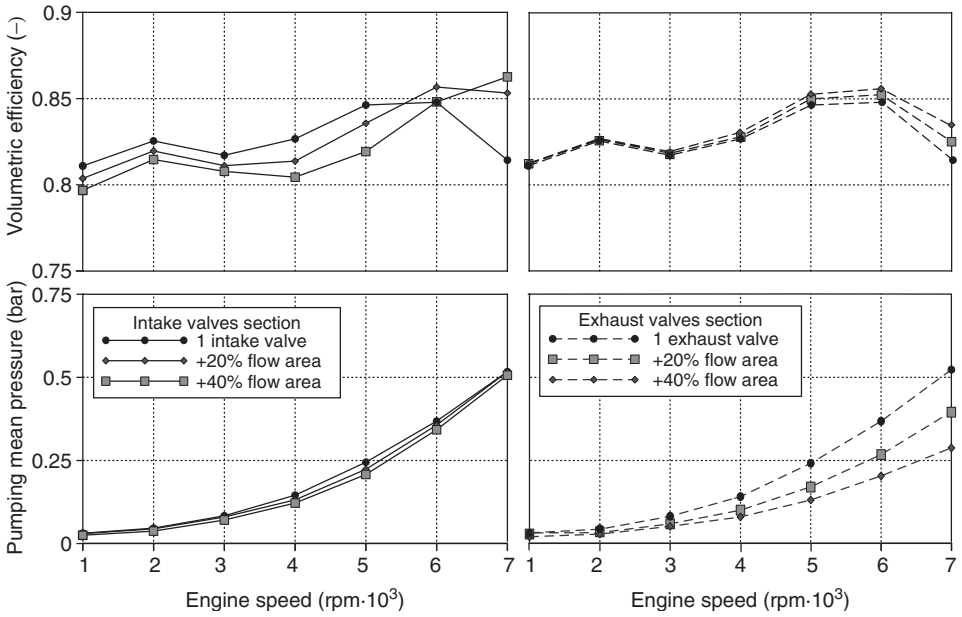


Figure 7. Influence of the intake and exhaust valve effective flow area on volumetric efficiency and pumping losses.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

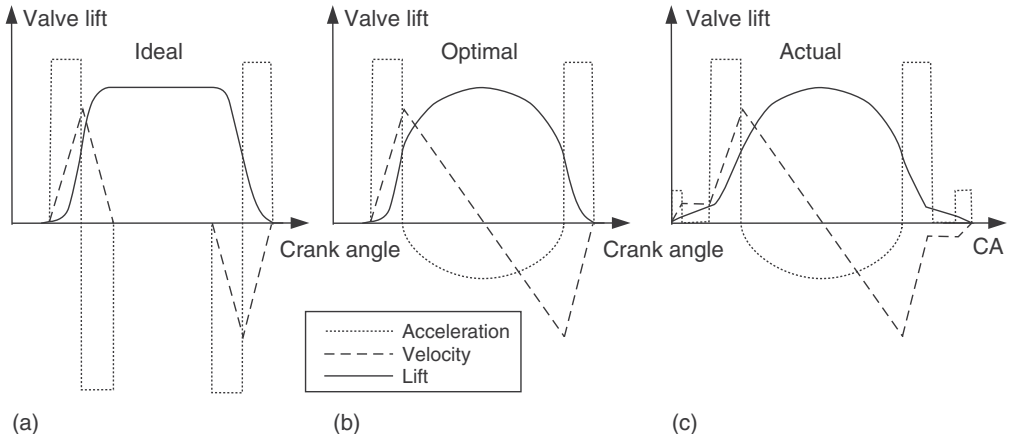


Figure 8. (a–c) Lift, velocity and acceleration of various cam profiles.

that allows maximum area for the gas flow and achieves instantaneous opening and closing with the desired timing. However, owing to stress limitations and dynamics, actual valves driven by a camshaft cannot open or close instantly but must follow a certain law. Figure 8a shows the ideal cam profile along with the optimal one (Figure 8b) that considers only valve or tappet clearance effects and the actual one (Figure 8c), which considers limitations because of both valve clearance and cam take off from the tappet.

In addition to the design of the cam profiles, it is necessary to determine the timing that achieves the optimal valve-lift law. Figure 9 represents a conventional valve-lift law.

The main objective of the valve-opening and valve-closing laws is to achieve optimal gas exchange with the minimum work (pumping losses). The selection of the values for the opening and closing angles must meet the following criteria:

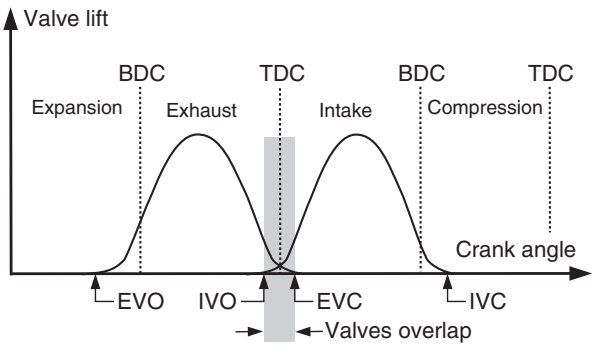


Figure 9. Exhaust and intake valve lift during a four-stroke engine cycle.

EVO: Allows the combustion gases to come out of the cylinder. A trade-off between reduction of pumping losses and reduction of piston work must be achieved as will be further discussed in Section 8. In turbocharged engines, the work produced by the turbine must also be taken into account. Typical values are between 30° and 70° before BDC.

EVC: Allows the completion of the combustion gas scavenging around TDC taking advantage of exhaust gas flow momentum. Typical values are around 2° and 15° after TDC.

IVO: Allows the entrance of fresh charge, driven by exhaust gas ejection during valve overlap period, which helps to sweep the remaining combustion gases out of the chamber around TDC. Typical values are between 2° and 15° before TDC.

IVC: Gives an extra time to fully fill the cylinder, taking advantage of the intake flow momentum achieved during the intake stroke. The optimum closing angle is when inlet flow velocity reaches zero and begins to invert its flow direction; typically between 20° and 40° after BDC.

4.1 Variable valve timing (VVT) systems

In order to determine the different parameters related to the valve motion laws, various specific aspects of reciprocating engines must be considered. Such aspects will not only be related to performance but also to other aspects such as reliability and durability of the different engine parts, as well as noise.

However, selecting the optimum parameters for valve motion laws in certain operating points does not ensure that those parameters will also be optimum for other operating

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

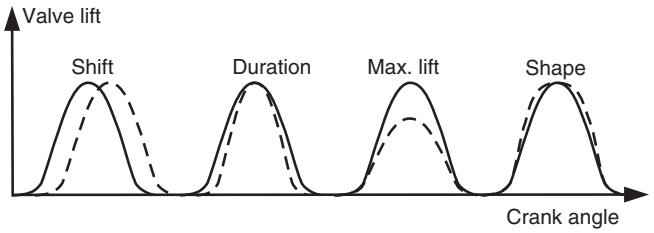


Figure 10. Valve-lift law, possible variations in a VVT system.

points. That is why VVT systems have been developed, which allow finding the optimum solution for different operating points, therefore providing a wider range of optimum engine operation.

Different configurations of VVT can be found. The valve-lift law parameters that can be adjusted with these VVT systems are valve-lift phasing at constant profile; valve-lift duration at constant maximum lift; valve-lift maximum lift at constant duration and phasing and valve-lift profile at constant duration, phasing, and maximum lift. Figure 10 shows the scheme of the described possibilities. There are also some VVT systems that may combine some of these parameters.

The valve-opening laws may be modified in a continuous way (infinite possible configurations of valve-lift laws ranging between two extreme values) or in a discrete way (allowing for two or three possible configurations) for further information, see Valvetrain development and Engine Management Systems. Figure 11 shows a comparison between a conventional camshaft and a real valve-lift law configuration for several operating conditions of an SI engine with 3.6 L of displacement volume equipped with a VVT system.

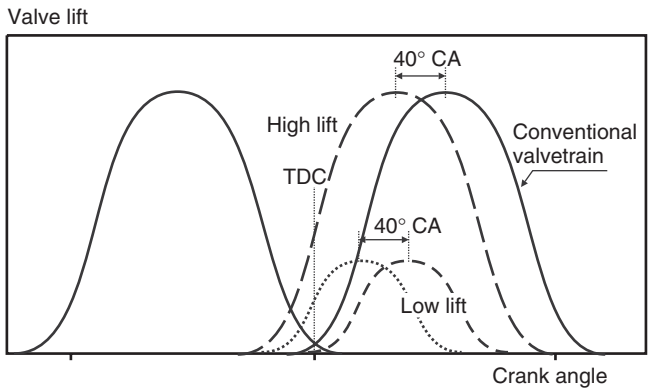


Figure 11. Valve-lift laws in a 3 l. SI engine for a sport car.

5 FLOW IN MANIFOLDS

Intake and exhaust processes are essentially unsteady because of their sequential nature. This produces unsteady pressure conditions at the valves that are transmitted throughout the entire intake and exhaust lines. Flow in manifolds can be usually considered as one dimensional (Winterbone and Pearson, 2000), and if heat transfer and friction are negligible, the flow equations become hyperbolic (Benson, 1982). The solution for these kinds of equations can be depicted as pressure and velocity perturbations that travel forward (p^+, u^+) and backward (p^-, u^-) interacting with discontinuities. Pressure p and velocity u perturbations can be related by means of Equation 10 introduced by Earnshaw (1860)

$$u = \frac{2a_0}{\gamma - 1}\left[\left(\frac{p}{p_0}\right)^{\frac{\gamma-1}{2\gamma}} - 1\right] \tag{10}$$

where a_0 and p_0 stand for the speed of sound and pressure at unperturbed flow conditions, respectively. As it can be extracted from the equation, overpressure waves can be associated to gas movement in the same direction of the wave, whereas rarefaction waves move in the opposite direction. In the first case, there is a "blowing" effect and in the second a "suction" effect in the direction of the wave.

Forward and backward waves travel within the ducts at $a+u$ and $a-u$ velocities, respectively. In a nonhomentropic flow, the shape of the waves is modified because of nonlinear effects, heat transfer, friction, or area changes (Benson, 1982). Pressure and associated velocity waves are also modified when they go through a discontinuity, such as open or closed ends, sudden area changes, branches, and temperature or composition discontinuities.

In closed ends, the zero flow velocity condition leads to a reflected component with the same amplitude as the incident wave. In open ends, pressure remains constant implying that pressure waves are reflected with the same amplitude; but overpressure waves are reflected as rarefaction waves and vice versa. The behavior at a sudden area change is shown in Figure 12. In a sudden expansion, an overpressure wave is reflected as a rarefaction wave with smaller amplitude and transmitted as a smaller amplitude overpressure wave. The sudden contraction behaves as a partially closed end. In both cases, the reflected and transmitted amplitudes depend on the area change ratio.

A pipe junction also behaves as a sudden expansion. When a pressure wave arrives from one of the pipes, it is reflected as a function of the ratio between the pipe area and the addition of the areas of the rest of the pipes (Winterbone and Pearson, 2000).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

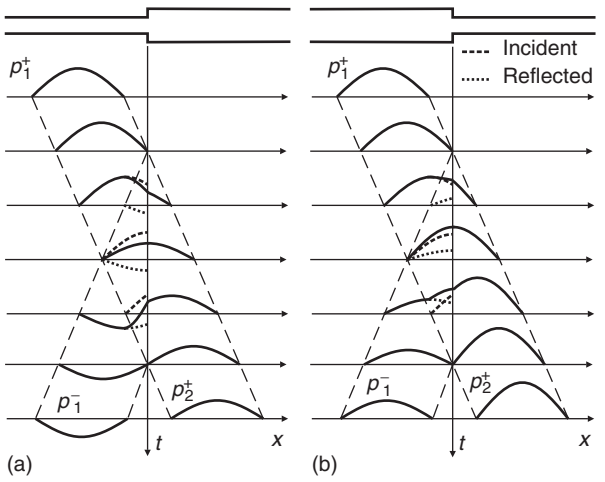


Figure 12. (a,b) Reflection and transmission of pressure waves through sudden area changes.

Flow in manifolds has been modeled by solving the flow equations for one-dimensional nonviscous flow. The first solution technique was based on the Method of Characteristics (MoC) in which the partial derivative equations are converted into total derivative equations over the so-called characteristic lines (Benson, 1982). This method has a graphical solution used before the arrival of digital computers. Later on, the MoC was adapted to numerical solution techniques that can be used in concurrence with the volume or the finite difference techniques (Blair, 1996). Solution schemes are classified according to the solution accuracy as first-order (MoC, Lax-Friedrich, one-step Lax-Wendroff), second-order (two-step Lax-Wendroff, McCormack), and higher order schemes (total variation diminishing, TVD; flux-corrected transport, FCT; and conservation element and solution element, CESE) (Payri *et al*., 2004).

Pressure waves produced during the intake and exhaust processes are a source of noise at the intake and exhaust mouths (see NVH Considerations in Engine Development). However, some advantages can be obtained using them to improve the engine filling or reduce the pumping work, both in the intake side and in the exhaust side (Winterbone and Pearson, 2000).

5.1 Dynamic effects at the intake

The aspiration of the cylinders during the intake process generates a pressure decrease at the proximity of the intake valves that travels as a rarefaction wave through the intake runners. The shape of the rarefaction wave is nearly sinusoidal. The rarefaction wave can be reflected at different discontinuities through the intake line. When the wave encounters a sudden expansion, a pipe junction, or a plenum, it is partially reflected as an overpressure wave coming back to the cylinders. This usually happens at the manifold, heat exchangers, air filters, compressors, and open ends. The reflected overpressure wave increases the pressure at the intake valve leading to an increase of the intake flow. This phenomenon is represented in Figure 13.

It is clear that the reflected wave must arrive before the intake valve closes. It has been proved that the maximum benefit is obtained if the wave arrives at the middle of the intake process. The time delay is a function of the length from the valve to the discontinuity and the speed of sound. The crank angle delay (θ) depends also on the engine speed. Therefore, for a given length and temperature, the best tuning is only achieved at a certain engine speed. This effect can be accounted for with the Strouhal number (Str), defined in Equation 11, as the ratio of the natural frequency of the system to the frequency of the pressure pulsations produced by the cylinder. For a closed/open end, the natural frequency is the speed of sound a divided by four times the system length L, and for a 4S engine, the frequency of the pulses is half of the engine speed n. Attending to Equation 11, this can be rearranged as a ratio between 360 crank angle degrees and the shift between generated and reflected pulse (θ), defined as shown in Figures 13 and 14 (Benajes *et al*., 1997)

$$\mathrm{Str} = \frac{f_{\mathrm{intake}}}{f_{\mathrm{engine}}} = \frac{\frac{a}{4L}}{\frac{n}{2}} = \frac{a}{2nL} = \frac{360}{\theta} \qquad (11)$$

Finally, the amplitude of the rarefaction wave is related to the ratio between the displacement volume and the volume available in the intake system.

The improvement of volumetric efficiency by means of the use of dynamic effects in the intake side can be considerable (Benajes *et al*., 1997). For this reason,

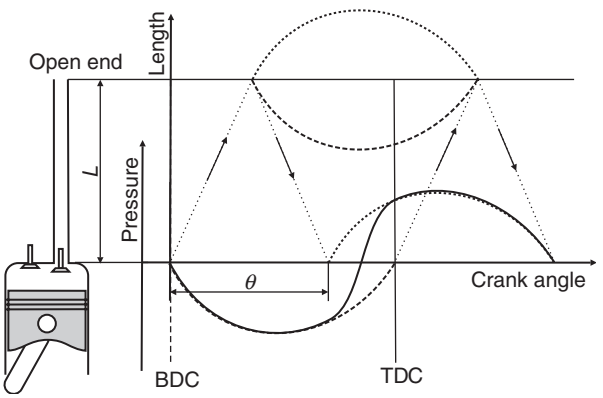


Figure 13. Scheme of pulse reflection at the intake valve.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

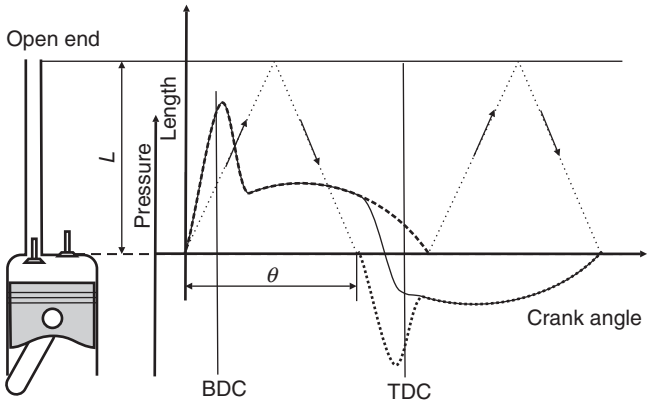


Figure 14. Scheme of pulse reflection at the exhaust valve.

this technique has been frequently used, especially in naturally aspirated engines. An issue to be considered for multicylinder engines is that the pressure waves produced at a cylinder can disturb the intake process of other cylinders. In order to reduce this negative effect, beyond a given number of cylinders (five in the case of 4S engines), the intake runners are separated into different manifolds (Winterbone and Pearson, 2000).

5.2 Dynamic effects at the exhaust

Dynamic effects at the exhaust side are similar to those explained earlier. However, some differences can be pointed out. First, the exhaust pulsations are overpressure waves generated because of the discharge of the cylinders and their shape has two separate peaks. In addition, the amplitude of the exhaust pulsation is larger than the amplitude of the intake one. This is depicted in Figure 14.

The overpressure wave travels through the exhaust ducts and can be reflected at discontinuities. When the wave encounters an increase of the duct cross section, it is partially reflected as a rarefaction wave that travels back to the cylinders and reduces the pressure at the exhaust valve. This reduces the pumping work and can also improve cylinder filling and scavenging if the reflected wave arrives during the valve overlap in a 4S engine or during the scavenging phase in a 2S engine (Blair, 1996). The reflection of the exhaust pulse can be produced at the exhaust manifold junctions, tapered pipes, catalysts, mufflers, or open ends.

Other elements such as convergent tapered pipes or turbines can reflect the exhaust wave as an overpressure wave. This can be used to reintroduce fresh charge that would be otherwise short circuited to the exhaust system. This technique is often used in naturally aspirated 2S engines as a way to improve scavenging and trapping efficiencies (Blair, 1996).

The use of tuning effects for 4S engines in exhaust systems is not as common as in intake systems for one main reason: because of the higher temperature at the exhaust side, the speed of sound is also higher; this leads to the tuned exhaust having greater length for a similar engine speed. In addition, the benefits in terms of volumetric efficiency are reduced compared to those obtained by intake tuning. Finally, turbocharged engines are not usually tuned as it is preferred to place the turbine as close as possible to the cylinders in order to reduce thermal losses.

The interferences between the different exhaust processes of a multicylinder engine are an issue as well. Furthermore, as the exhaust effective duration is longer than the intake duration, the problem is even larger. A way to reduce interference between cylinder exhausts in four-cylinder 4S engines is by means of separated manifolds (Galindo *et al*., 2004). Another way is the use of a proper design of the pipe junctions at the exhaust manifolds. The so-called pulse converters are designed to produce an ejector effect that reduces this negative effect. Two-scroll turbocharger turbines have a similar effect on reducing interferences (Watson and Janota, 1982).

5.3 Variable geometry manifolds

Dynamic effects can be exploited only at given engine speeds in which the reflected waves arrive at the cylinder valve at the appropriate moment. In some cases, two reflections at different discontinuities can be utilized at two engine speeds. A solution to overcome this limitation is to change the length that the pressure waves have to travel between the valve and the discontinuity. In these variable geometry manifolds, the length of the runners is modified to tune the reflected wave to the engine speed (Winterbone and Pearson, 2000).

6 RESIDUAL GASES AND EXHAUST GAS RECIRCULATION

6.1 Definition of residual gas and its effects on engine operation

Usually, the gas exchange process in reciprocating engines does not achieve complete removal of combustion gases in every cycle. It is therefore usual that at IVC a fraction of the total charge is still present in the cylinder coming from earlier combustion processes. This is known as *residual gas fraction*.

The residual gases will have a number of effects on engine operation. On the one hand, regarding engine performance, residual gases will reduce the work output per cycle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

This is the result of the residual gases taking up a certain volume in the cylinder, which is no longer available to the fresh charge, and thus reducing the oxygen available for combustion. This effect is worsened by the high specific volume of the hot residual gases. In homogeneous-charge, spark ignition (SI) engines, the residual charge directly limits the amount of fuel available for each cycle, because it limits the maximum amount of air–fuel mixture that can be drawn into the combustion chamber. In compression ignition engines, residual gases take up volume that will not be available to the fresh charge (which, in this case, is composed solely of air) and thus reduce the amount of oxygen available for achieving a convenient fuel oxidation. Moreover, the residual gases slow down combustion in a twofold way:

1. By reducing the probability of oxygen reacting with fuel due to the presence of an inert gas in the combustion chamber.
2. By reducing local combustion temperatures due to the higher specific heat of the residual charge with respect to that of the fresh charge.

On the other hand, residual gases will have an impact on pollutant emissions as residual gases modify combustion parameters. The most obvious effect (mainly on diesel engines) is the reduction of nitrogen oxides (NO*x*) generation because of the already discussed reduction of local combustion temperatures (see NO_x Formation and Models). This will limit the chemical reactions that result in NO*x* formation. This aspect became paramount in the last decade of twentieth century (Uchida, Daisho, and Saito, 1993) because of the focus of the different pollution regulations on reducing NO*x* emissions (e.g., Euro III and following regulations).

6.2 Residual gas and EGR: characterization and control

Inert gases can appear in the combustion chamber by two different ways.

6.2.1 Internal generation of residual gases (internal EGR)

Gases in the chamber are the sum of those introduced by the backflow that appears in the intake and exhaust valves, and the gases that are not expelled after combustion (remaining gases in the combustion chamber). The mentioned backflow originates because of the pressure difference between the cylinder and the valves, generally during the valve overlap period. Figure 15 shows the instantaneous mass flow rate through the valves. Negative values represent backflows.

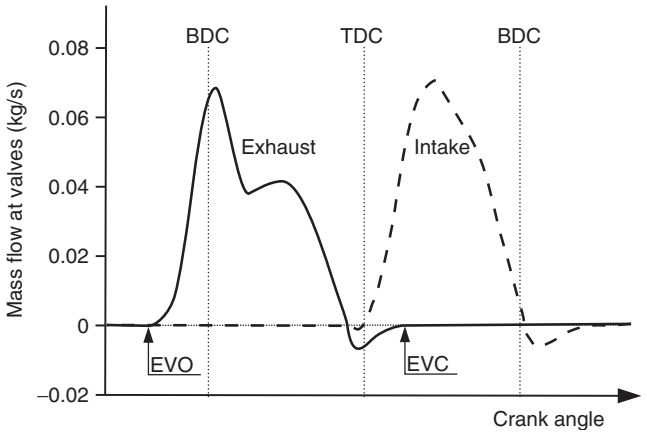


Figure 15. Instantaneous mass flow rate through intake and exhaust valves.

For each engine operating condition, valve-lift laws and geometry of the combustion chamber will determine the amount of residual gases that can be produced. Engines with conventional distribution systems generate a small amount of internal EGR (around 5% of the admitted gases). However, a number of strategies have achieved high EGR rates (higher than 40%) by modifying the values of the IVC and EVO crank angles. These strategies are the so-called intake valve prelift, exhaust valve postlift, and negative valve overlap (Benajes, Reyes, and Luján, 1996).

The internal EGR technique faces two fundamental problems that have hindered widespread application. On the one hand, it is not possible to control the amount of EGR produced, and, on the other hand, it is not possible to cool down the EGR gases, which would be very convenient for improving engine volumetric efficiency and reducing NO*x* emissions.

6.2.2 Recirculation of exhaust gases (external EGR)

This technique is aimed largely at reducing emissions of NO*x* in compression ignition engines (see Automotive Diesel Engine Development Trends). For this purpose, engines must be equipped with piping connecting the exhaust and intake systems. In SI engines, this technique finds little application, because NO*x* can be effectively removed by exhaust gas aftertreatment systems (see Gas Aftertreatment Systems and Stoichiometric Exhaust Emission Control).

Figure 16 shows a scheme for a turbocharged diesel engine with the typical external high pressure exhaust gas recirculation (HP EGR) and low pressure exhaust

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

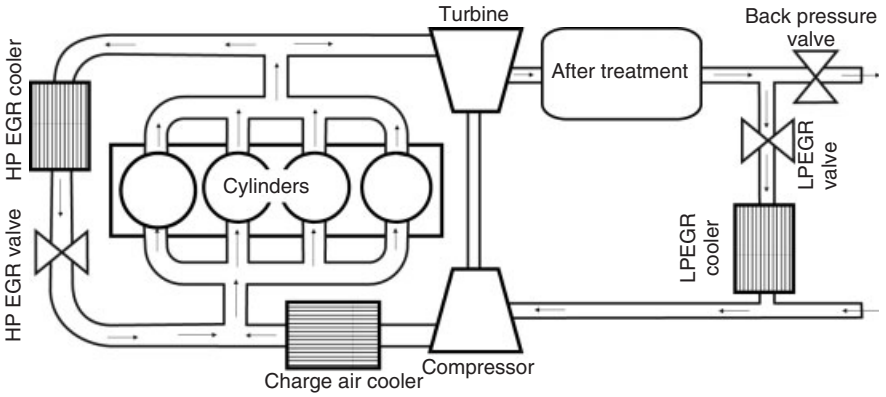


Figure 16. High pressure and low pressure EGR configurations.

gas recirculation (LP EGR) configurations, allowing for external recirculation of combustion gases.

High temperature (300–600°C) recirculated exhaust gases are usually cooled down by means of gas/liquid heat exchangers (HP EGR and LP EGR coolers). The gas temperature reduction has obvious impacts

1. It reduces the temperature of the mixture at the intake, which, in turn, reduces NO*x* generation.
2. It increases the density of the intake gas, thus increasing the amount of admitted gases (air + EGR).

EGR flow can be indirectly calculated by measuring CO_2 concentration at the intake (mixture of air and EGR) and exhaust ducts. EGR rate can be defined as the ratio of EGR mass flow rate to the total mass flow rate. This ratio is defined by Equation 12, where $[CO_2]$ refers to volumetric concentration

$$\mathrm{EGR} = 100\frac{m_{\mathrm{EGR}}}{m_{\mathrm{EGR}} + m_{\mathrm{FRESH_AIR}}} \approx \frac{[\mathrm{CO_2}]_{\mathrm{INTAKE}} - [\mathrm{CO_2}]_{\mathrm{AMBIENT}}}{[\mathrm{CO_2}]_{\mathrm{EXHAUST}} - [\mathrm{CO_2}]_{\mathrm{AMBIENT}}} \qquad (12)$$

Owing to the high cost of CO_2 analyzers, this calculation is still not implemented on mass-produced engines, even though it is commonly used in research facilities, where pollutant analyzers are essential.

The amount of inert gas introduced into the chamber (mainly CO_2 and H_2O) depends on the amount of recirculated gas and its composition. To this regard, the burned gas ratio (BGR) is defined as follows:

$$\mathrm{BGR} = \mathrm{EGR} \cdot \phi \qquad (13)$$

where ϕ is the equivalence ratio.

Oxygen concentration in the combustion chamber can be calculated as follows:

$$[\mathrm{O_2}]_{\mathrm{COMBUSTION}} = [\mathrm{O_2}]_{\mathrm{AMBIENT}}[1 - \mathrm{BGR}] \qquad (14)$$

The amount of recirculated gas is controlled by means of a valve placed in the EGR duct. The electronic control unit (ECU) regulates this valve based on the operating conditions of the engine (see Engine Management Systems).

Without using CO_2 analyzers (which are cost prohibitive), it is difficult to measure the amount of recirculated gas because of its high temperature and the high amount of particulates in the gas. Control of the EGR mass flow rate is achieved using an estimation of the EGR mass flow rate based on measurement of the mass flow rate of the fresh charge that enters the engine (which is easy to measure with hot wire devices and similar equipment), along with intake gas pressure and temperature measurements. Using Equation 1 and assuming constant volumetric efficiency, fresh charge displacement caused by the EGR can then be used as a means to calculate indirectly the produced EGR. Thus, the EGR amount is controlled in closed loop for each condition by setting the amount of fresh air, determined during engine calibration, and regulating this amount by controlling the EGR valve (see Engine Management Systems).

7 VOLUMETRIC EFFICIENCY CURVES

As a synthesis of the previous sections, it is worth considering the evolution of volumetric efficiency versus the two main engine operating parameters, engine speed and engine load. Figure 17 sketches the evolution of volumetric efficiency versus engine speed for different load levels for both SI and CI engine concepts.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

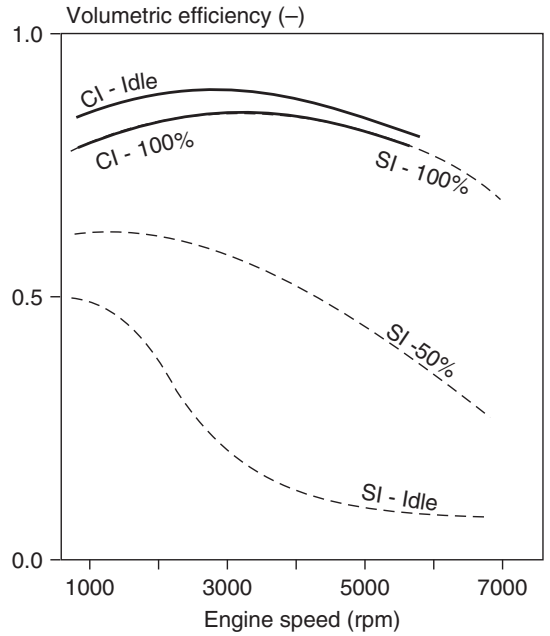


Figure 17. Volumetric efficiency versus engine speed (rpm) and load level (%) for SI and CI engines.

Focusing first on the evolution of volumetric efficiency versus engine speed at full load conditions, Figure 17 shows, with respect to the maximum, a decreasing volumetric efficiency trend when engine speed is reduced. This is true for either engine type, CI or SI. One main reason for the trend is an increase in air charge heating with the reduction of engine speed due to the lower speed and the resulting higher residence time of the air charge in the cylinders, and therefore, the higher the heat transfer from the hot cylinder walls to the intake air charge. Such heat transfer will reduce the average inlet charge density during the intake process and, consequently, the volumetric efficiency. Another equally important reason for this reduction is the increase of backflow from the cylinder to the intake ducts as the engine speed is reduced. The main reason for the backflow is that the engine valve timing is optimized for high engine speeds, that is, in order to use flow inertia and induction ram effects to improve cylinders filling and scavenging processes at rated power. Consequently, the high valve overlap periods and the late IVC, which are optimum for higher engine speeds, generate backflow that is detrimental to volumetric efficiency at reduced engine speeds.

Regarding the evolution of volumetric efficiency at full load conditions, starting from the maximum value and proceeding in the direction of increasing engine speed, Figure 17 shows an even steeper decreasing trend. The causes are twofold. The first is due to the increase of volumetric flow that increases flow friction and pressure losses in every element of the intake line, reducing the intake pressure and density. The second and main cause is due to the compressible fluid effects associated with the high Mach numbers at the intake valves, which choke flow and prevent significant mass flow increases when further increasing engine speed (Taylor, 1985). The previously commented induction ram effects can compensate, at higher engine speeds, the reduction of volumetric efficiency caused by compressible flow effects, in spite of the already mentioned backflows at lower engine speeds.

The balance between the described effects at low engine speed and high engine speed places the optimum volumetric efficiency in the medium engine speed range. Finally, tuning of cylinders, inlet and outlet ducts can generate a peak of volumetric efficiency taking advantage of the pulsating flow effects (Benajes *et al.*, 1997). Geometry of ducts may be designed for a specific range of engine speeds at which the tuning effects are maximized, considering engine service requirements and/or desired operating characteristics.

The described effects on volumetric efficiency affect in a similar way CI and SI engines at full load conditions. However, SI engines with port injection usually show lower peak values of volumetric efficiency because of flow losses from throttling (see Petrol Engines), higher residual gas fraction, intake manifold heating, and the presence of fuel vapor. This SI trend is changing although with the introduction of direct injection (DI) SI engines as full load injection is performed during the intake stroke. Introducing fuel directly inside the cylinders reduces charge temperature because of fuel vaporization and allows more trapped mass and greater volumetric efficiency.

The evolution of volumetric efficiency with engine load shows an opposite trend when comparing CI with SI engines. In the case of CI engines, and owing to the reduction of equivalence ratio with engine load, the temperatures of cylinder head walls are lower at partial loads than at full load. Therefore, the density reduction of air charge during the intake stroke due to heat transfer from the walls is less important at partial loads. Consequently, the volumetric efficiency at partial loads in CI engines is somewhat higher than at full load conditions, as it is indicated in Figure 17. This trend is exactly opposite the trend for SI engines. As SI load regulation is performed by throttling the intake line (see Petrol Engines) as already mentioned, the reduction of mass flow to maintain a constant equivalence ratio is detrimental to the SI engine volumetric efficiency. Figure 17 shows the progressive and important decrease of volumetric efficiency at any engine speed as a function of SI engine load.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

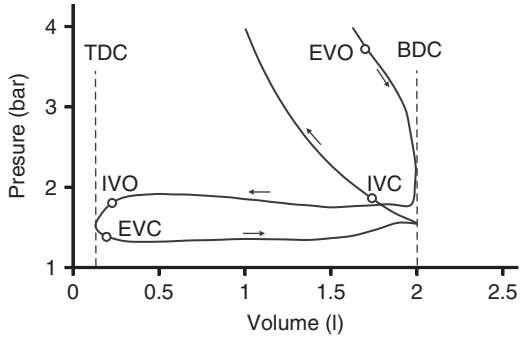


Figure 18. Counterclockwise low pressure loop corresponding to a four-stroke HD diesel engine at 1800 rpm and 25% load conditions.

8 PUMPING FRICTION LOSSES

Conceptually, pumping losses represent the necessary work to pump the combustion gases out of the cylinders during the exhaust stroke and to pull in the fresh charge for the next combustion process during the intake stroke. The pumping work per cycle can be calculated by integrating the low pressure evolution of the indicated $P-V$ diagram between the exhaust BDC and the intake BDC, as shown in Equation 15. A new effective pressure type variable (see Operating Principles) can be defined, the pumping mean effective pressure (pmep).

$$W_{\text{pumping}} = \int_{\text{BDC}_{\text{exhaust}}}^{\text{BDC}_{\text{intake}}} p(a)\mathrm{d}V(a) = \text{pmep} \cdot V_{\text{Displaced}} \tag{15}$$

Figure 18 depicts that the integral of Equation 15 includes an extra area outside of the low pressure counterclockwise loop. As this area is also considered when computing the indicated work, both magnitudes cancel when the difference between indicated and pumping work is considered. This difference is known as *net work* (W_{net}) and it can be calculated as shown in Equation 16, which defines the net mean effective pressure (nmep).

$$W_{\text{net}} = W_{\text{indicated}} - W_{\text{pumping}} = \text{nmep} \cdot V_{\text{Displaced}} \tag{16}$$

Figure 18 shows that any engine component contributing to increase in-cylinder pressure during the exhaust stroke or to decreased pressure during the intake stroke will increase pumping friction losses. Regarding pumping losses, these engine components can be separated into two groups: elements that are inside the cylinder head and those that are outside of it. The most important elements inside the cylinder head are the inlet and exhaust valves and, to a lesser extent, the inlet and exhaust ports. The pressure losses in these elements are called *valve flow work* by Heywood (1998), and they are responsible for the most important contribution to pumping work in diesel engines. The most significant pressure losses take place in the exhaust valve. The reason is that in order to prioritize the cylinder filling process with fresh air charge versus the cylinder exhaust process, the cross section of the exhaust valve is smaller than the corresponding cross section of the intake valve (Figure 6), even though similar volumetric flow rate has to flow through both of them.

Valve timing is the next parameter in importance that affects valve flow work. More specifically, advancing EVO (Figure 9) will help to maximize the balance between pmep and imep, that is, to maximize nmep. Figure 19a shows the effect of advancing EVO with respect to BDC (dashed line). In the case of the advanced EVO, the area reduction in the clockwise high pressure loop (vertical hatchings) is compensated by the spontaneous blowdown of the exhaust gases from the cylinder, thus achieving a reduction in pumping loop area (horizontal hatchings) because of a reduced in-cylinder pressure during the exhaust stroke.

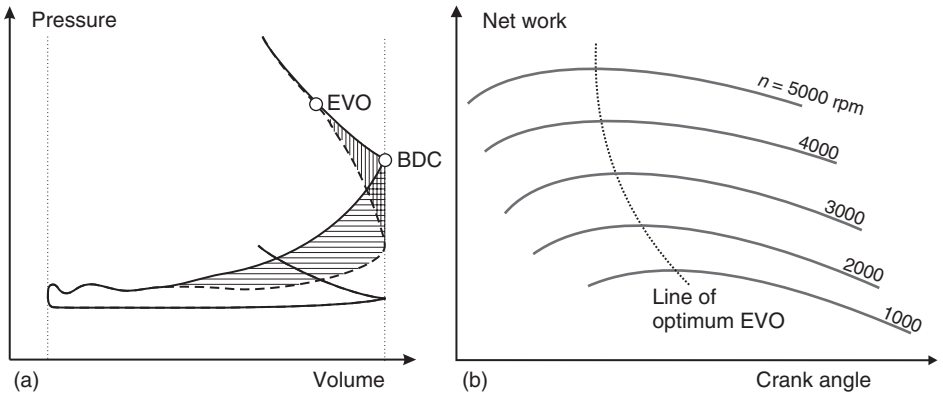


Figure 19. (a) Influence of EVO on network and (b) variation of optimum EVO as a function of engine speed.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The optimum value for EVO advance varies with engine speed, because the optimum advance should increase as the duration of the exhaust process is reduced, or, what is equivalent, as the engine speed increases. Figure 19b shows this advancement with respect to BDC in order to optimize nmep at different engine speeds. The dotted line that fits the maximum of every curve is the EVO that should be provided by a VVT system (see Engine Management Systems). In classical systems with fixed geometry camshafts, the optimum EVO is chosen at an intermediate engine speed.

In SI engines and at partial loads, the pressure losses due to throttling the intake charge overcome the importance of the exhaust valves with regard to their relative contribution to pumping friction losses; this is the case also in DI–SI engines, because the engine load remains controlled by the intake throttle.

Other elements outside the cylinder head whose design influences pumping work, but less than *throttling work* (Heywood, 1998), are the air filter, carburetor, intake manifold (on the inlet side) and the exhaust manifold, exhaust gas aftertreatment (catalytic converter, SCR, see Gas Aftertreatment Systems, and DPF, see Solid/Condensed Phase Aftertreatment Systems), muffler, and tail pipe (on the exhaust side, see NVH Considerations in Engine Development). The influence of exhaust line elements on pumping losses is threefold: the most obvious are pressure losses; following by ram effects that would help scavenging, especially during valve overlap period; and, last but not least, unsteady flow effects, already discussed in Section 5.2. Figure 20 shows an unsteady effect in the low pressure loop for two different θ values (Figure 14). If the rarefaction wave arrives around the middle of the exhaust stroke ($\theta = 90^{\circ}$), it can help to expel exhaust gases from the cylinders and to reduce pmep. However, if the rarefaction wave arrives around the valve overlap ($\theta = 180^{\circ}$), the unsteady effect does not help to reduce pumping losses,

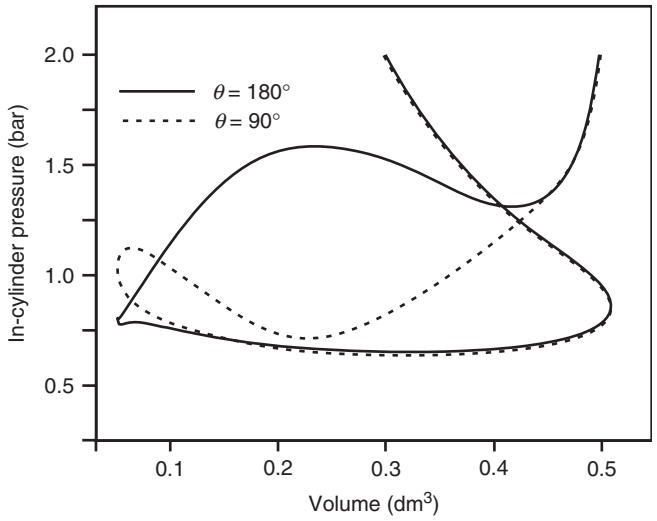


Figure 20. Influence of exhaust ducts tuning on counterclockwise low pressure loop. Results are from a naturally aspirated 2.0 L petrol engine.

although there is an aspiration effect that could increase volumetric efficiency and improve chamber scavenging.

It is also worth discussing separately, due to its relevance, the contribution to pumping losses of another element outside of the cylinder head: the turbocharger (see Intake Boosting and Turbocharging). For certain engine operating conditions, an efficient design (of both the turbocharger and its coupling with the ICE engine) can generate a clockwise low pressure loop that increases nmep over imep; despite the higher exhaust back pressure generated by the turbine (Watson and Janota, 1982). Figure 21a shows this effect in the low pressure loop of a heavy-duty (HD) diesel engine at medium speed (1200 rpm) and full load. It can be compared with typical counterclockwise low pressure loop at 1500 rpm and 50% load shown in Figure 21b for the same engine.

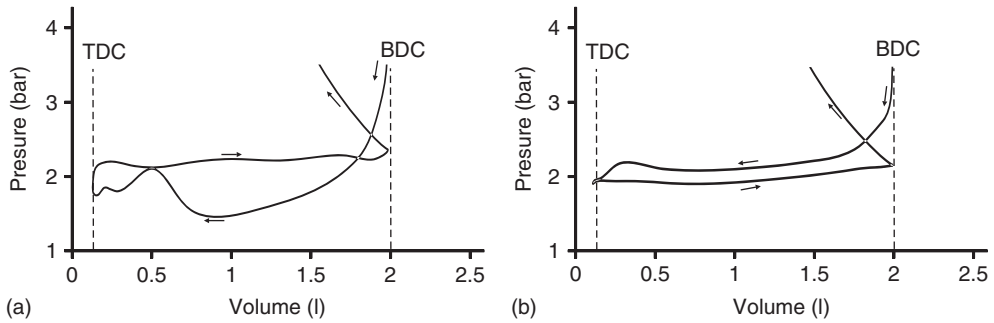


Figure 21. Low pressure loop in the $P-V$ diagrams of a turbocharged HD diesel engine. (a) 1200 rpm, full load and (b) 1500 rpm, 50% load.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In summary, with respect to how engine speed influences pumping losses, a twofold analysis has been made that covered the influence of engine speed on the optimum values of EVO and its influence on the design of the exhaust line, in order to take advantage of ram and unsteady flow effects. It can be added that generally the higher the engine speed, the higher the volumetric flow, and therefore the flow friction losses. With respect to the influence of engine load, it will be very significant in SI engines and pmep will increase as engine load is reduced because of throttling work. The influence of engine load on pmep for naturally aspirated CI engines is less important and it mainly depends on how the density of exhaust gases change and whether the volumetric flow is increased or not. In supercharged CI engines (see Turbocharging and Supercharging), the variation of mass flow with load will be much more significant than in naturally aspirated CI engines, and so will be the variation of pumping losses (Watson and Janota, 1982).

RELATED ARTICLES

Operating Principles
Thermodynamic Analysis
Fundamental Chemical Kinetics
Fundamental Combustion Modes
NO_x Formation and Models
AT Control—Actuation Methods & System Integration, Gear Choice, Gear Shift Strategy & Process, Adaptive Features
Particulate Formation and Models
In-Cylinder Flow
Zero- and One-Dimensional Methodologies and Tools
Multidimensional Simulation
Intake Boosting
Exhaust Gas Energy Recovery
Engine Thermal Management
Gas Aftertreatment Systems
Solid/Condensed Phase Aftertreatment Systems
Exhaust Emissions
Petrol Engines
Automotive Diesel Engine Development Trends
Valvetrain development
NVH Considerations in Engine Development
Turbocharging
Supercharging
Stoichiometric Exhaust Emission Control
Exhaust Emission Control Considerations for Diesel Engines
Engine Management Systems

REFERENCES

Annand, W.J.D. and Roe, G.E. (1974) *Gas Flow in the Internal Combustion Engine: Power, Performance, Emission Control, and Silencing*, G.T. Foulis, Sparkford.

Benajes, J., Reyes, E., and Luján, J.M. (1996) Modeling study of the scavenging process in a turbo-charged diesel engine with modified valve operation. *Proceedings of the IMechE Part C: Journal of Mechanical Engineering Science*, **210**, 383–393.

Benajes, J., Reyes, E., Galindo, J., and Peidro, J.L. (1997) Predesign model for intake manifolds in internal combustion engines. SAE Paper 970055.

Benson, R.S. (1982) *The Thermodynamics and Gas Dynamics of Internal Combustion Engines*, vol. 1, Oxford University Press, New York.

Benson, R.S., Horlock, J.H., and Winterbone, D.E. (1986) *The Thermodynamics and Gas Dynamics of Internal-Combustion Engines*, vol. 2, Oxford University Press, New York.

Blair, G.P. (1996) *Design and Simulation of Two-Stroke Engines*, SAE International, Warrendale, PA.

Desantes, J.M., Benajes, J., and Urchueguía, J. (1995) Evaluation of the non-steady flow produced by intake ports of direct injection diesel engines. *Experiments in Fluids*, **19** (1), 51–60.

Earnshaw, S. (1860) On the mathematical theory of sound. *Philosophical Transactions of the Royal Society of London*, **150**, 133–148.

Fukutani, I. and Watanabe, E. (1982) Air flow through poppet inlet valves—analysis of static and dynamic flow coefficients. SAE Paper 820154.

Galindo, J., Luján, J.M., Serrano, J.R., *et al.* (2004) Design of an exhaust manifold to improve transient performance of a high-speed turbocharged diesel engine. *Experimental Thermal and Fluid Science*, **28** (8), 863–875.

Heywood, J.B. (1998) *Internal Combustion Engines Fundamentals*, McGraw-Hill book Co, Singapore.

Payri, F., Galindo, J., Serrano, J.R., and Arnau, F.J. (2004) Analysis of numerical methods to solve one-dimensional fluid-dynamic governing equations under impulsive flow in tapered ducts. *International Journal of Mechanical Sciences*, **46** (7), 981–1004.

Taylor, C.F. (1985) *The Internal Combustion Engine in Theory and Practice*, MIT Press Cambridge, MA.

Uchida, N., Daisho, Y., and Saito, T. (1993) Combined effects of EGR and supercharging on diesel combustion and emissions. SAE Paper 930601.

Ward-Smith, A.J. (1980) *Internal Fluid Flow: The Fluid Dynamics of Flow in Pipes and Ducts*, Oxford University Press, New York.

Watson, N. and Janota, M.S. (1982) *Turbocharging the Internal Combustion Engine*, MacMillan Publishers Ltd, Southampton.

Winterbone, D.E. and Pearson, R.J. (2000) *Theory of Engine Manifold Design*, Professional Engineering Publishing Limited, London and Bury St. Edmunds.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto162
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Sandwich Materials

Heinz Palkowski[1], Olga A. Sokolova[1,2], and Adele Carradò[2]

[1] *Clausthal University of Technology (TUC), Clausthal-Zellerfeld, Germany*
[2] *Institut de Physique et Chimie des Matériaux de Strasbourg, Strasbourg, France*

1 Introduction 1
2 Theoretical Background: Classifications of Composite Structures 2
3 Joining Hybrid Materials 4
4 Industrial Application of Layered Composites 4
5 Metal–Polymer Adhesion 6
6 Manufacturing of Metal/Polymer/Metal Sandwich Sheets 8
7 Mechanics and Formability of the Hybrid Systems 9
8 Outlook 13
Acknowledgment 14
References 14

1 INTRODUCTION

The strong demand for novel materials with new functions generates considerable interest in the automotive community. Metals, ceramics, or polymers as mono materials cannot fulfill all technological needs for a variety of original applications. Researchers in chemistry and physics as well as engineers understand that to obtain materials with superior properties, they have to combine mono materials to hybrids. The definition of "*hybrid material*" given by Kickelbick is "*hybrid material* is used for many different systems across a wide area of different materials such as crystalline highly ordered coordination polymers and amorphous sol–gel compound materials with and without interactions between the inorganic and organic unit." A complete review of *hybrid materials* is given in the book by Kickelbick (2007).

For industrial applications, the development of a special *hybrid material*, the *matrix type*, is one of the most successful examples for a group of composites. A matrix is a system, in which a second type of material—in the form of fibers, particles, whiskers, lamellae, or a mesh—is integrated into a basic material. A typical example is a polymer reinforced by inorganic fibers, which possess improved mechanical properties. At present, these new systems are regularly used as lightweight materials with advanced mechanical properties for automotive or aviation applications.

In this chapter, *hybrid materials* are defined as three-layered metal–polymer–metal sheets, called *sandwich material*. They are laminates with two or more layers of minimum two different materials—called *laminae* (lamina: single ply or layer)—bonded together. With these combinations in a sandwich material, it is possible to "design" the properties of particular components with the right choice of mono materials, thus providing the functionality to fulfill the high demands on modern materials and structures (Harris, 1991). They can combine the advantages of various single materials together (e.g., low density, high bending resistance, energy absorption, and high load capacity with low weight) (Librescu and Hause, 2000; van Tooren, 2004). It should be pointed out that for sandwich materials, there are multiple possibilities of combining materials for core and skin layers.

To determine the mechanical properties of *sandwich materials*, common static, dynamic, and cyclic testing must

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

be performed as well as to define the conditions of failure, for example, by delamination of the layers.

According to the industrial requirements, various models and simulations of sandwich plates for deep-drawing, bending processes, or the crash behavior of sandwich material components in a vehicle body form the focal point of international research of steel–polymer sandwiches. Oh, Cho, and Kim (2005) numerically computed the delamination process of sandwich plates. Moreover, the spring-back effect was investigated using a numerical model (e.g., Liu and Wang, 2004). The study of damping, vibration, and acoustic properties of sandwich materials, which are fabricated using different polymer cores and steel or aluminum skins, is based on the theory of sandwich plates and a finite element (FEM) analysis (Chen, Hsu, and Chen, 1991; Engel and Buhl, 2011).

Using different modeling types (Roque and Thomson, 2005; Chen, Hsu, and Chen, 1991), the characterization of stress–strain conditions, structural changes, or other important effects can be determined and analyzed for both simple sandwich plates and the behavior of sandwich parts in a whole construction.

Special interest in the research of sandwich materials is given in the characterization of formability. The forming behavior of sandwiches under different loading conditions, especially bending and deep drawing has been studied intensively (Palkowski and Lange, 2007; Carradò *et al.*, 2010; Palkowski and Lange, 2008; Sokolova, Carradò, and Palkowski, 2011a; Carradò *et al.*, 2011a).

To industrially apply polymer-based sandwich materials, for example, as automotive parts, their behavior under thermal and mechanical joining (Section 3) has to be investigated, as these hybrids show both some weakness due to the core's thermal instability during the thermal joining and their elastic and/or plastic deformation by mechanical joining, thereby probably losing their prestressing force under dynamic load. In the case of thermal load, the polymer can be vaporized and cause metal corrosion (Palkowski *et al.*, 2006). A solution for this case can be provided by the use of local plate metal reinforcements (RE—reinforcing element), replacing the sandwich core at the place of future thermal or mechanical joining. Owing to the use of different materials for the REs (solid or mesh steel, see Section 3), their geometry, size, and the position of the inlays relative to the forming, these sandwiches can exhibit significantly different formability and the flow behavior of both the sandwich layers and the internal local inlays varies too (Sokolova, Carradò, and Palkowski, 2010, 2011a).

This chapter gives a short theoretical background to the use of composites in the automotive industry, followed by a study of sandwich materials that are often manufactured using a roll bonding process with a preliminary surface treatment. They consist of a polypropylene copolymer (PP–PE) (PP, polypropylene; PE, polyethylene) core and the covering AISI (American Iron and Steel Institute) SS316L austenitic stainless steel (316L) sheets. The mechanical behavior of 316L/PP–PE/316L sandwich sheets is presented together with their formability with and without metal reinforcements. This chapter concludes with a study of these systems' adhesion using a T-peel and shear test.

2 THEORETICAL BACKGROUND: CLASSIFICATIONS OF COMPOSITE STRUCTURES

Owing to the orientation, shape, and form of the components, composites can be applied in various fields of industry (Campbell, 2010; Jones, 1999; Lange, 1993). Some representatives of industrially applied sandwich materials can be classified as:

- reinforced composites with a thermoset or thermoplastic polymer matrix;
- particulate-filled composites;
- diffusion-bonded composites;
- sandwich composites.

The reinforced composites can have a polymer matrix as thermoset, thermoplastic, or elastomer. The reinforcements are used to modify specific properties, such as mechanical or electromagnetic ones, or to decrease the thermal expansion coefficient of the polymers, for example, embedding of electrical components in epoxy resins using the molding process (Rajput, 2007; Wijskamp, 2005).

Fiber-reinforced composites are lightweight composites consisting of a polymeric matrix (thermosetting resins) augmented with reinforcing fibers, some fillers, and pigments. They are classically applied for consumer products such as chairs, trays, helmets, water cooler bodies, tanks, and paneling (Vermeren *et al.*, 2003).

Some common matrix materials include epoxy or phenolic resins, polyester, polyurethane, and vinyl ester. Among these materials, resin and polyester are the most widely used ones. Epoxy, which has higher adhesion and less shrinkage than polyesters, is ranked second because of its high costs. Glass fibers are generally used as a reinforcing material in the resins and produce a combination of high strength and stiffness in the matrix. The direction of the fibers can be specially oriented or random (Kawai and Hachinohe, 2002). Fillers are commonly added to smooth

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the manufacturing process and to yield special properties. The most commonly used ones are carbon/graphite fibers, beryllium carbide, beryllium oxide, aluminum oxide, glass fibers, polyamide, and natural fibers.

Owing to many attractive advantages in comparison to the more widely used thermoset composites, *ultra-lightweight carbon fiber/thermoplastic composites* are of growing interest. As reported by El-Dessouky and Lawrence (2013), their advantages are principally based on the intrinsic properties of thermoplastic polymers (e.g., polyphenylene sulfide) used and the ability to be reshaped and reused or recycled.

A new and major application in their use is presented in the low carbon vehicle technology project (LCVTP) aiming at reducing the in-use CO_2 emissions for future vehicles (LCVTP, 2010). In particular, the LCVTP is a collaboration between leading automotive companies and research partners with the goal to revolutionize the way to design and produce vehicles, including full battery vehicles and hybrid vehicles, in order to significantly reduce carbon emissions, so, in the first step, reduce weight. Actually, in the LCVTP, an innovative seat back structure using rapid-stamp formed thermoplastic composites has been developed.

Reinforced composites with a thermoset or thermoplastic polymer matrix can be manufactured using different process routes such as liquid composite molding (Rudd *et al*., 1998), injection molding (Rosato, Rosato, and Rosato, 2000), or hot and cold press forming (Rajput, 2007).

These technologies can be used for metals and polymers as well as combinations of them. *In the liquid composite molding* process, the reinforcement fibers are already impregnated with a resin (called *prepreg*). *Injection molding* is a widespread process for the processing of thermoplastic materials. The addition of reinforcement materials to the injection molding compound allows increasing the mechanical strength of the final part. *Back injection molding* is used for high strength bond of metal sheets in combination with thermoplastics in series production for many applications. This technology allows to join a metal sheet with high quality surface with the back-side injected polymer, with or without fibers and bonding agents. Moreover, this process opens the way to reproduce symbols and structures to the metal sheet in one shot.

Press forming in this context refers to sheet molding compound and bulk molding compound, respectively; compounds consisting of a thermosetting matrix and long reinforcement fibers delivered as flat or volumetric batches. The compound is placed into a heated mold and then formed to the shape of the final part by closing the mold with sufficient pressure.

Particulate-filled composites are commonly used as high strength, lightweight plates with simple geometries because of the low formability of the polymer or the ceramic matrix (Rothon, 2003). The particles used for reinforcement include ceramics and glasses, small mineral particles or metal particles, such as aluminum, and amorphous materials, including polymers and carbon black. An example of the application of particulate-reinforced composites is the car tire, which has carbon black particles in an elastomeric polymer matrix. These particles are used to enhance the strength and to decrease the ductility of the matrix. Some of the useful properties of ceramics and glasses include high melting temperature, low high strength density, high stiffness, wear resistance, and corrosion resistance. Many ceramics are used because of their proper electrical and thermal insulations or their magnetic and piezoelectric properties. Some can even be used as superconductors. One major drawback of ceramics and glasses is their brittleness.

Diffusion-bonded composites are composites with a strong molecular bonding. They can be manufactured using different interfacial bonding mechanisms: modifying the wetting and adsorption of surface elements, generating diffusion on the surface of internal phases, creating chemical and mechanical adhesions of materials or magnetic diffusive joining of metallic compounds (Shirzadi and Wallach, 1997; Guo and Derby, 1993). These composites show a good combination of the properties required, but the reproducibility and precision of the manufacturing of such systems are not currently assured. Nevertheless, these kinds of composites show great potential for future applications.

Sandwich materials—understood to be *composites*—consist of at least two or more mono-material layers with different physical and mechanical properties. They can be produced with numerous core and skin materials (e.g., steel, aluminum alloy sheets, polymer foils, or fiber-reinforced plastics, with a polymer layer being homogenous or structured). They exhibit a great potential because they are simple to manufacture using lamination, press joining, or roll-bonding processes (Jones, 1999). Such a production process is described by Palkowski and Lange (2007) and Carradò *et al*. (2011a).

These materials can be highly deformed if the laminate combines highly deformable metallic materials or when the combination of materials demonstrates high plasticity. Compared to other types of composites, sandwich materials are often used as lightweight parts in the automotive and marine industries because of their easy production and the homogeneity of the layers. In the examples given, the objects will be thin metal–polymer–metal sandwich sheets.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3 JOINING HYBRID MATERIALS

Spot welding is a commonly used, wide spread process in the automotive industry. Nevertheless, the localized heating of the material may change the mechanical properties of the materials. For hybrid joints, a combination of joining techniques is often advised (Timings, 2008).

Adhesive bonding is mostly used in combination with spot welding. High quality glues offer the possibility to join dissimilar materials at low cost. Furthermore, using adhesive bonding allows avoiding the deformation of the components and the damping of noise and vibrations. At the same time, the adhesive joint can enhance the fatigue resistance because of the reduction of stress concentrations (Barnes and Pashby, 2000; Huck and Bosshard, 2007).

Hybrid joints consisting of spot-joints (welding) with adhesive bonding, which can generally be presented in three different methods: *fixing, injection, and sequential methods*. Using the *fixing method*, the adhesive is applied to the parts being joined; then, a spot-joint is made to complete the process, followed by the hardening of the adhesive layer. In the *injection method*, the parts are joined with a spot-joint before injecting an adhesive into the gap between the components. Capillary action causes the adhesive to spread through the joint. The adhesive layer then hardens. In the *sequential method*, the parts are joined after the adhesive has been applied. Subsequently, when the adhesive has hardened, the parts are spot-joined.

Laser welding (e.g., CO_2, excimer, and Nd : YAG lasers) presents some advantages compared with conventional welding techniques (Mackwood and Crafer, 2005). They can be summarized as a modest distortion of the components, a deep penetration welding, broader spectrum of "weldable materials," and an aptitude to process different kind of materials (Mackwood and Crafer, 2005). Most plastic products still need to be joined up. Formerly, these parts were often glued or screwed together with the drawback to be time consuming. Recently, as presented by Sieben and Brunnecker (2009), laser welding of plastics is used for joining automotive parts and medical equipment as well as consumer products.

Remote laser welding is a suitable process for applications where a good positioning and high track speeds are required. Owing to the high power density at the focal point, the material to be treated heats up until reaching its melting point, thus fusing and welding (Tsoukantas *et al*., 2005; Tsoukantas and Chryssolouris, 2006).

The riveting joint is one of the important joint methods to permanently fasten two thin-walled sheet-metal parts. The riveting technology is also enhanced by the introduction of *self-piercing rivets* (Hill, 1994; Barnes and Pashby, 2000). Self-pierce riveting is a high speed mechanical fastening technique for point joining of sheet materials. It is used strongly in the automotive sector. Even investigations on combining sandwich sheets to metallic mono materials are well known (Pickin, Young, and Tuersley, 2007). The advantages of this process are their ability to join dissimilar materials without thermal damage (Grote and Antonsson, 2009). Owing to the fact that in combination with adhesives, the self-piercing technology can provide leak proof joints of high fatigue strength and a high degree of reliability (Mortimer, 2005); it offers a viable alternative to spot welding in automotive assembly.

Joining by clinching is strongly used in the automotive industry for assembling sheet plate elements (Varis, 2003). The clinching joint is produced by local stamping of joined sheets without piercing and any heat effect on the material structure. For *hybrid systems*, an adhesive often is applied additionally to one of the components being joined. Once linked, the components are immediately clinched causing a releasing of the adhesive out of the joint. Finally, the joint is left to harden.

The *roll-bonding process* is a possibility for the continuous production of sandwich metal–polymer sheet material under defined temperature–pressure–time conditions (Palkowski and Lange, 2007; Sokolova, Carradò, and Palkowski, 2011a; Carradò *et al*., 2011a). More details are given in Section 6. Table 1 provides a brief comparison of the aforesaid joining processes. The imperfection of the sandwich construction cannot grant such joining properties because of the elastic polymer core at the place of joints and probably the earlier failure of sandwich constructions by long-time or dynamic loading. In all cases, a simple method of properly joining sandwich parts can be the use of local plate inserts (RE, reinforcement) in the core (Bozhevolnaya and Lyckegaard, 2005). The method of using such reinforcements is based on placing the inlay instead of the core material before joining the layers to a sandwich sheet. Applying the local RE in a sandwich core also reduces the use of complex adaptors, which can lead to an earlier failure of sandwich constructions. Moreover, adapting the REs to the specific needs (e.g., load) allows to create a graded property change. Figure 1 shows two examples for placing REs in a sandwich sheet.

4 INDUSTRIAL APPLICATION OF LAYERED COMPOSITES

Owing to the large differences in the properties of the partners, the manufacturing process for three-dimensional components by joining and/or forming differs from the shaping of mono materials. A sandwich (Zenkert, 1997) is normally a combination of a lightweight and a material of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Selection of joint technologies for metal and composite.

Assembly process	Access (1 or 2 sides)	Weldable materials	Heat distortions	References
Spot welding	2, good	Most metals	Yes	Michalos *et al*. (2010)
Adhesive bonding	1, very good	Most materials	No	Barnes and Pashby (2000), Huck and Bosshard (2007)
Laser beam welding	1, very good	Most materials	Yes	Mackwood and Crafer (2005), Sieben and Brunnecker (2009)
Remote laser welding	1, very good	Most materials	Yes	Tsoukantas *et al*. (2005), Tsoukantas and Chryssolouris (2006)
Friction stir welding	1, good	Most materials	No	Olsen (2007)
Riveting	1, good	Metals and plastics	No	Grote and Antonsson (2009), Hill (1994), Barnes and Pashby (2000)
Clinching	2, good	Metals and plastics	No	Varis (2003)
Roll bonding	1, very good	Metals and plastics	No	Palkowski and Lange (2007), Sokolova, Carradò, and Palkowski (2011a), Carradò *et al*. (2011a)

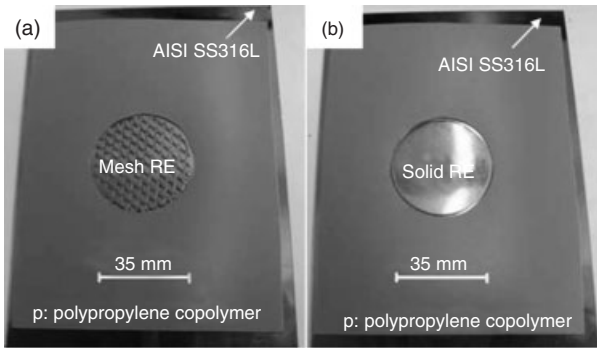


Figure 1. Reinforcements in a half sandwich with (a) mesh and (b) solid inlays.

higher strength. In general, the lighter material is used as the core layer, covered by stronger, stiffer skins to sustain the properties of the core. The soft core supports the metal sandwich skins under various loading, defines the distance between the cover layers, dissipates the shearing forces, and serves as a high damping material. Lightweight core materials can include polymer foils, balsa wood, polymer foams, and metallic, paper, or polymer honeycombs (Hull and Clyne, 1996). These materials have been used in various combinations with skins of carbon, glass, and/or aramid fiber-reinforced polymers, as well as aluminum, titanium, or steel. Single layers are usually connected by bonding (glue bonding, metal coating, extrusion, or molding), which guarantees high tensile and shear strengths. As a result, it was decided to employ sandwich materials in lightweight application (e.g., a sandwich having a honeycomb core was used for a car body (ultra light steel auto body for chassis components) (Milton and Grove, 1997). In Table 2, a classification for sandwich composites are listed, which are already employed in different fields of industry and are currently globally available.

Used as core materials, the aluminum-skinned honeycombs or PP foils between steel sheets demonstrate a weight reduction of approximately 50% with the same stiffness of a comparable steel sheet. In addition to this, using austenitic high grade steel (Euroinox, 2012) or special deep-drawing

Table 2. Classification of some representatives of industrially employed sandwich composites.

	Materials	References
GLARE©	Al/glass fiber in polymer/Al	Ibarra-Castanedo *et al*. (2011), van Tooren (2004), Burchitz *et al*. (2005), Botelho *et al*. (2006), Vermeren *et al*. (2003), Kawai and Hachinohe (2002)
Carall©	Al/carbon epoxy/Al	Almeida *et al*. (2008)
ARALL	Aramid fiber-reinforced aluminum laminates	Johnson (1986), Vogelsang (1983)
FML	Fiber metal laminate	Abdullah and Fahruddin (2009)
Hylite©	Al/polypropylene/Al (Al/PP/Al)	Burchitz *et al*. (2005)
Alukobond©	Al/PVC or mineral/Al	Alcan composite GmbH (2009)
Dibond©	Al/polyethylene/Al	Alcan composite GmbH (2010)
Bondal©	Steel/polypropylene/steel	ThyssenKrupp Stahl (2003)
TiGr©	β-Ti/carbon polyamide(PMC)/β-Ti	Burianek and Spearing (2001)
Hybrix©	Stainless steel/polypropylene/stainless steel	Jackson, Allwood, and Landert (2007)
Sollight©	Mild steel/polypropylene/mild steel	Jackson, Allwood, and Landert (2007), ArcelorMittal (2011)
Alulight©	Al/PVC or mineral/PVC	Jackson, Allwood, and Landert (2007), ArcelorMittal (2011)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

steel (ThyssenKrupp, 2008), an improved corrosion resistance and improved fatigue behavior can be achieved. Compared to other metals, such as aluminum or magnesium alloys, mild up to even high and ultra high grade steels exhibit a high elongation to fracture—and thereby, a high capacity for complex forming—and, at the same time, a high tensile strength (TS). One example for an automotive application of such a sandwich used for high damping and lightweight is the aluminum/PP/aluminum sandwich Kim, Kim, and Choi (2008). Hylite (Corus Group) actually followed by an improved sandwich sheet in the three-layered version steel/polyamide/steel by ThyssenKrupp Steel. The former was designed as a lightweight element and has already been employed as a roof, vehicle floor, and hood (Burchitz *et al*., 2005).

Bondal (ThyssenKrupp AG) sandwiches, composed of steel cover sheets and a polyolefin core in the range of some 10 microns, are used for sound damping of oil pans, valve, and transmission covers as well as garage doors or glass containers.

To improve their stiffness and strength, the use of glass fibers is foreseen because of their large availability and low price. The main handicap is their brittleness; indeed, subsequent to cooling down, the material cannot be deformed. Moreover, delamination or early failure under cyclic loading could be observed (Botelho *et al*., 2006). The capacity for high deformation is exhibited by the sandwich materials possessing a thermoplastic core.

5 METAL–POLYMER ADHESION

Adhesion is the physical condition of an interface layer, formed between two contacting phases by mechanical, physical, or chemical bonding (Ruokolainen and Sigler, 2008). The bondability of materials is the ability of various materials to interact between each other (Bischof *et al*., 1989).

Current research in adhesion is well presented by Possart (2005). The bonding agent or adhesive (e.g., epoxy) can permit the bonding between metal and polymer, when self-adhesion is not possible. The adhesive is commonly a viscous or temperature hardening agent. Figure 2 presents various types of contributions to adhesion in a sandwich plate consisting of "component 1" (e.g., metal skin), "component 2" (e.g., polymer core), and the "adhesive agent." The adhesion can be classified by the bonding mechanisms. Concerning metal/polymer/metal structures, the following types can be distinguished:

- *mechanical adhesion* strongly depends on the surface roughness and surface morphology of the partners;
- *chemical adhesion* is based on chemical reactions between the adhesive agent and adherends (bonded layers);
- *physical adhesion* depends on the physical processes, namely physical interpolar bonds based on the interaction forces (attraction and repulsive forces) going on

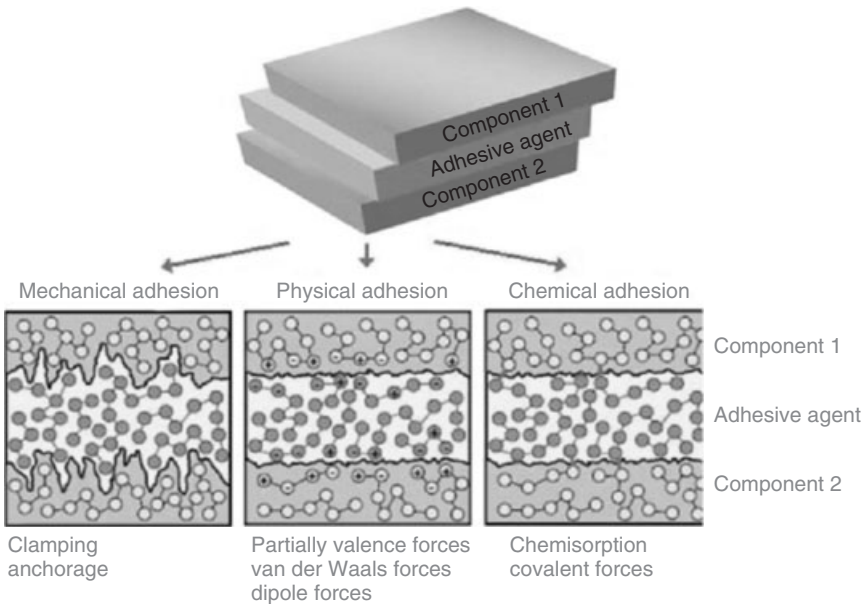


Figure 2. Introduction to bonding of different materials with an adhesive. (Reproduced from Schindel-Bidinelli and Gutherz, 1998 . © Wiley-VCH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

by the bonding of components (Schindel-Bidinelli and Gutherz, 1998).

Adhesion can influence the processing and forming properties and, of course, significantly affect the properties of the sandwich components. The effect of the adhesion on the mechanical and formability properties of steel/polymer/steel sheets was investigated (Ruokolainen and Sigler, 2008; Carradò *et al*., 2011b).

Different theories were presented in an attempt to understand the adhesive bonding of the metal–polymer. The mechanical adhesion between the two materials (by micro-mechanical anchorage) and the auto-adhesion (inter diffusion of motile macromolecules) are based on simple mechanisms, whereas the specific adhesion is linked to physical–chemical and thermodynamic phenomena.

In "*theory of polarization*," de Bruyne describes—based on the electrical polarity of the atoms or the molecular groups—the interactions between functional groups of adherents and adhesives (de Bruyne, 1939). Experiments have shown that the adhesive forces increase with an increase in the dipole moment of the atoms or molecules. To improve the adhesion of both the adherents and the adhesives, the polymer must contain some polar groups and provide the dipoles. Some examples of such polar functional groups include oxygen- or nitrogen-containing groups such as hydroxyl, carbonyl, carboxyl, ester, amino groups, epoxy, isocyanate, or nitrile.

In comparison to other theories based only on physical bonds between adherend and components, the "*chemical adhesion model*" assumes that not only intermolecular forces but also chemical reaction can occur (Reinhart, 1954). These strong chemical bonds can be formed, such as the covalent bonds in the case of polymers. Both an incomplete absorbability and high bonding energies of the molecules have been experimentally observed during the bonding. It was found that only valence bonds can lead to this irreversible process.

A surface treatment is necessary to change and/or to improve some of the polymeric and metallic surface properties without modifying the bulk properties (Chan, 1994). The low surface energy of the polymeric surface (de Bruyne, 1939) can be raised using surface treatments such as corona ozone activation or atmospheric plasma methods. Therewith, the bonding of the metal–polymer interface can be improved, too.

Corona (Baldan, 2004; Lahti and Savolainen, 2004; Sun, Zhang, and Wadsworth, 1999) and plasma treatments (Strobel, Lyons, and Mittal, 1994; Liston, 1991; Wertheimer *et al*., 2002; Minzari *et al*., 2008) are commonly used to introduce reactive groups onto a nonpolar polymer surface such as those present in a polyolefin film (polypropylene copolymer based on a mixture of PP and PE) and to improve the wettability of the surfaces. Wettability is a measure for the hydrophobicity of surfaces using the surface tension defined by the contact angle between, for example, a liquid on a solid. Sellin *et al*. (2003) have studied the effect of corona treatment on the polymer surface using this water drop contact angle measurement. They have shown that the presence of created polar groups by corona activation depends on the aging of the polyolefin surface after the treatment. In the review of Baldan (2004), the effect of surface preparation and influence of bonding method (adhesive agent) on the adhesion between the metal and polymer surfaces were analyzed. It was observed that surface treatments, for example, corona and plasma as well as cleaning of surfaces using chemical agents before the treatment can clearly improve the polarity of polymer surfaces.

The scanning electron microscopy (SEM) images of the PP–PE films show significant changes in the surface morphology induced by the corona treatment (Figure 3a, b). Holes (pores) and an irregular porosity distribution can be detected, which are important for the adhesion properties.

Following the corona treatment, it has been verified (Carradò *et al*., 2011b) that the adhesion—studied by T-peel test—between the metal and the polymers is improved as can be seen in Figure 4. In this test, the peel resistance—or initial crack strength (P_A)—is defined as $P_A = F_A/b$ (N/mm), where F_A is the peeling load (or crack force) and b is the width of the test piece (30 mm). The mean peel resistance is $P_S = \overline{F}/b$ (N/mm), where $\overline{F}$ is the adhesion strength and b the width of the sample (30 mm). The bar chart in Figure 3 depicts a strong increase in crack-peel resistance (28%) and a mean peel resistance (43%) with increasing corona exposure time from 20 to 120 s.

A linear tendency of improvement is observed. These results can be explained by the creation of polar chemical functional groups because of the corona surface treatment.

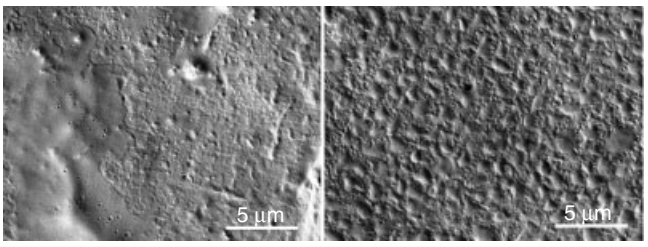


Figure 3. SEM image of polypropylene copolymer foil (PP–PE) before (a) and after (b) a corona discharge treatment for 120 s. (Reproduced from Carradò *et al*. 2011b. © John Wiley & Sons, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

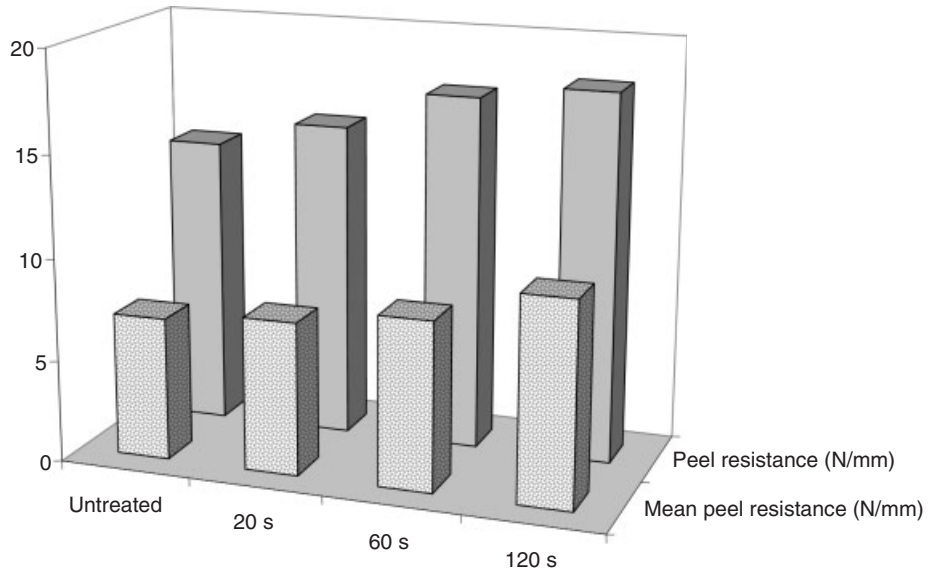


Figure 4. T-peel indexes (peel resistance and mean peel resistance) of polypropylene copolymer foil without and after corona treatment for different treatment periods (20, 60, and 120 s). (Mean ± SD, SD: standard deviation for $n = 6$). (Reproduced from Carradò *et al*. 2011b. © John Wiley & Sons, Inc.)

These groups also enhance the surface adhesion and wettability as stated by Zhang, Sun, and Wadsworth (1998).

The effect of corona treatment on the polymer under standard discharge in air is explained as an anchoring of atoms or molecules deposited onto the polymer surface (Bishopp, 2005). Moreover, the influence of the corona aging effect on the polymer surface's wettability was investigated by Pascual *et al*. (2008). The stable and high wetting capacity following a corona treatment permits corona-activated polymer sheets to be processed under standard discharge in air one day before bonding latest.

Finally, many parameters can improve the adhesion between polymers and metals to overcome the poor polarity of the polymers. In the hybrid steel/polymer/steel structure presented here, the surface is prepared by corona and/or plasma treatments before bonding. In order to chemically improve the adhesion, adhesive agents such as epoxy resin were used.

6 MANUFACTURING OF METAL/POLYMER/METAL SANDWICH SHEETS

To obtain and guarantee the desired properties of the sandwich composites, the production processes have to be adapted. The bonding of metal and polymer sheets can be performed in vacuum or standard atmospheres at room or elevated temperatures using a roll-bonding process. The advantage of the roll-bonding process is the ability to produce and to fabricate endless sheets under constant and reproducible conditions. The production process of metal–polymer sandwiches is modified in order to be carried out without adhesive agents using liquid or semi-liquid polymer cores in a press-rolling process (Kurt *et al*., 2004). A good overview of their applications and manufacturing techniques—with and without using adhesives—is given by Vinson (2005).

The AISI SS316L/polypropylene copolymer/AISI SS316L sandwich materials (SMS), manufactured in our laboratory, are produced using a roll-bonding process by means of a 2-high 10" rolling mill (Palkowski and Lange, 2008; Lange, Carradò, and Palkowski, 2009; Palkowski and Lange, 2007; Carradò, and Palkowski, 2009; Carradò *et al*., 2011a). Stainless steel AISI SS316L is a common alloy possessing excellent corrosion resistance and thermal stability, which facilitate its use in a wide range of metal–polymer–metal sandwich structures. Such steel sheets with a thickness of 0.5 mm are used for the skin material and a 0.6 mm thick polyolefin foil for the polymer core. The polyolefin, PP–PE, including PE, PP, poly(4-methyl-1-pentene), ethylene–propylene elastomer, and ethylene–propylene–diene rubber, is a widely used commercial polymer family (Chung, 2002). The PP–PE used was mixed with talc [$Mg_3Si_4O_{10}(OH)_2$], rutile (TiO_2), and barite ($BaSO_4$).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

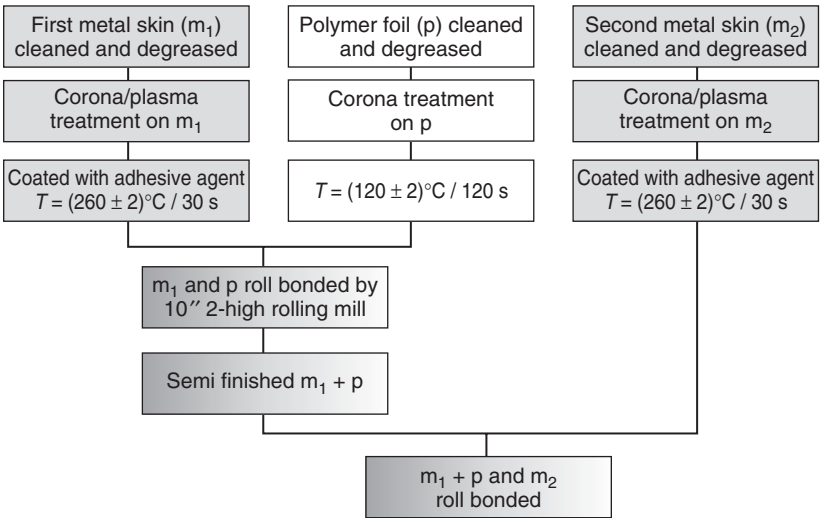


Figure 5. Laboratory roll-bonding process for the production of 316L/PP–PE/316L sandwich materials.

The roll-bonding manufacturing can be carried out as a continuous process. Under laboratory conditions, it was divided into two steps as sketched in Figure 5.

In the first step, the metal sheets (m_1) were cleaned and degreased with acetone and then coated with epoxy resin before being placed in a stationary convection furnace at (254 ± 2)°C for 30 s. Simultaneously, the PP–PE foil (p) was placed in a conventional furnace at a temperature of (120 ± 2)°C for 30 s to fit the bonding conditions with the epoxy resin. After activation of the adhesive, the upper sheet metal (m_2) was joined to the polyolefin foil using a 12" rolling mill. Before the roll bonding, a preliminary corona surface treatment was performed on the polymer core and the steel sheet was plasma treated. As discussed earlier, the corona and plasma treatments increase the surface energy and thereby improve the adhesion by means of activating the polymer (creating polar groups) and cleaning the metal before joining. The experimental conditions are given in Table 3.

As described in Section 3, a local strengthening of the hybrid structure can be achieved by a partial substitution of the core layer with a reinforcement (RE). A circular geometry was selected. Mesh (Figure 1a) and solid (Figure 1b) steel sheets were used for these REs. The diameters of the inlays were varied, the position of the local inlays was central to the deep-drawing shape or at a definite position offset from the center. The mesh inlays possessing a mesh size of $3 \times 3\,mm^2$ and a thickness of 0.5 mm were chosen to investigate the possibility and effect of additional weight saving for these sandwich structures.

Table 3. Parameters of the corona and plasma surface treatments.

Treatment	Parameters
Plasma on AISI SS316L sheets	8 kV, 500 W, time: min 20 s–max 600 s
Corona on PP–PE polypropylene copolymer foil	13 kV, 180 W, time: min 20 s–max 600 s

7 MECHANICS AND FORMABILITY OF THE HYBRID SYSTEMS

7.1 Fundamental laws of behavior

To describe how a sandwich material behaves is subjected to mechanical loading; some relations have been investigated to explain their performances.

The *rule of mixtures* refers to the expression (Equation 1) for the mechanical properties of a phase mixture in terms of the bulk or mechanical properties and the relative amounts of its constituent phases (Fan, Tsakiropoulos, and Miodownik, 1994).

$$\sigma_{US} = \sigma_{UA} V_A + \sigma_{UB} V_B \quad (1)$$

where σ_U and V indicate the uniaxial flow stress and volume fraction, and subscripts S, A, and B stand for the sandwich sheet and its A and B layers of components, respectively.

This law was tested and stated for describing the properties of fiber-reinforced composites with an acceptable degree of accuracy. However, the mixture rule for the sandwich materials has referred to the presence of negligibly small transverse stresses compared to axial or longitudinal stresses. The same law of mixtures can be modified

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to describe the Young's modulus for the composite by assuming a uniform axial strain for the whole sandwich composite.

$$\sigma_{US} = \sigma_{UA} V_A + \sigma_{UB} V_B$$
$$\varepsilon E_{US} = \varepsilon E_{UA} V_A + \varepsilon E_{UB} V_B$$
$$E_{US} = E_{UA} V_A + E_{UB} V_B \quad (2)$$

With the same indications as before and E_U: Young's modulus in the longitudinal direction.

Qualitative properties of some composite materials can be described using the *complementation theory*, which states that each constituent complements the other by separately contributing distinct properties; for example, an aluminum–steel composite, where aluminum provides the corrosion resistance and steel provides the strength (Milton, 2002).

The resulting properties of a composite material can also be described by the *concept of interaction*. It states that the property of one constituent is dependent on the property or the action of the others. The resulting composite's properties are usually intermediate to those of the constituents or can even exceed those of both; for example, ceramic matrix composites. The SiC/Si_3N_4, ceramic matrix composite, exhibits higher fracture toughness than that of its brittle constituents (Milton, 2002).

As an example for the mixture rule, the tensile test was used according to the standards (DIN 50114, 1981) to compare the SMS with the monolithic materials 316L and PP–PE (Figure 6). Yield strength (YS), TS, and elongation to rupture (ER) were measured using a universal testing machine. Because of anisotropy effects, specimens were taken at $\alpha = 0^\circ$, 45°, and 90° to the rolling direction of the sheets. Under uniaxial tensile load, the steel material and SMS show macroscopically elastic deformation until the YS is reached at about 620 and 400 MPa, respectively. Rupture occurs at a strain of approximately 53% and 50%, respectively. Additionally, the tensile test summarizes the mechanical properties of the 316L skin, the PP–PE core, and the SMS and shows the reduction in yield stress following the rule of mixture for the flow stresses of the sandwich sheets, given by Equation 3:

$$\sigma_{SMS} = \frac{\sigma_{316L} \cdot V_{316L} + \sigma_{PP-PE} \cdot V_{PP-PE}}{V_{316L} + V_{PP-PE}} \quad (3)$$

where σ and V indicate the uniaxial flow stress and volume fraction. The subscripts represent "SMS" the sandwich sheet, "316L" the metal, and "PP–PE" the polymer.

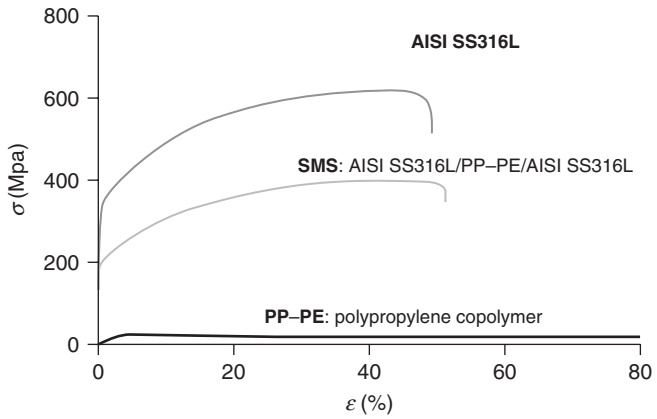


Figure 6. Stress–strain for 316L, SMS, and PP–PE at room temperature. Samples taken in rolling direction.

7.2 Formability, stretching, and deep drawing

Formability is the capacity of the sheet metal to be formed without failure (Narasimhan, Miles, and Wagoner, 1995). Thus, formability is not easily quantified, as it depends on numerous interacting factors such as material flow properties, ductility, tool geometry, tool material, lubrication, and deformation speed. Note that no single test can quantify the formability for all types of stamping applications (Bayraktar, Isac, and Arnold, 2005).

Numerical simulations using FEM (Ayari, Lazghab, and Bayraktar, 2009; Xiao *et al*., 2011) and the deep-drawing processes of the square cup test on metal or sandwich structures under load have been reported by a number of authors (Boesemann, Godding, and Huette, 2000; Weiss *et al*., 2007).

The size of the finished sample and the distribution of the deformations depend on the geometrical conditions and the degree of forming. Figure 7 shows characteristic areas of interest for a deep-drawn square cup as well as the typical stress states in a cup during deep drawing.

To investigate the formability of the SMS for deep drawing and stretching, tests according to Erichsen and Olsen (Olsen, 1920; Kaftanoglu and Alexander, 1961) as well as deep-drawing tests were performed. The stretch-forming behavior was characterized by the Erichsen test according to DIN 50101 (1979). Figure 8a illustrates a typical experimental setup. A cone-shaped spherical plunger deforms the sample, which is restrained around its periphery, until fracture occurs (Figure 8b). The height of the cup at fracture (called the *Erichsen index*) is a measure for the stretchability and a reference value for the material.

The deep-drawing tests were carried out using a deep-drawing benchmark test. The limits of deep drawing are affected by the geometry of the tools, the material properties of the blank and the process parameters (Tschaetsch, 2006).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

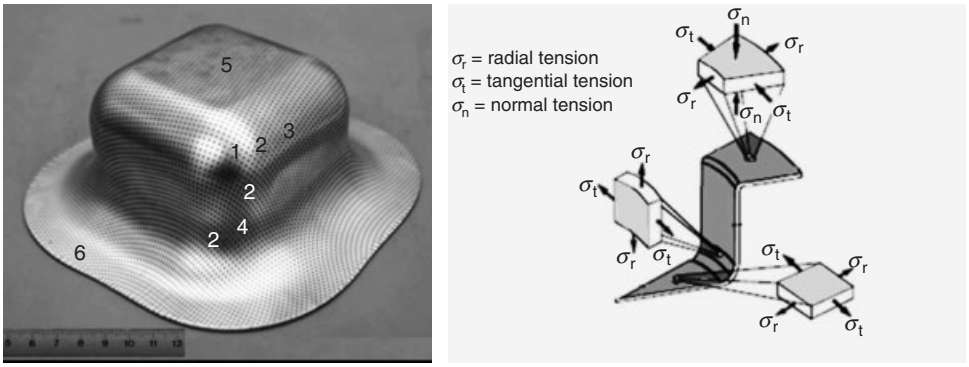


Figure 7. Forming zones of a square/rectangular cup (a) and typical stress states during deep drawing (b); 1, "corners"—punch or die rounding; 2, transition region "corner/edge"; 3, straightness of edges; 4, "edge" region (depends on forming degree); 5, "cup head" region; and 6, "flange" region.

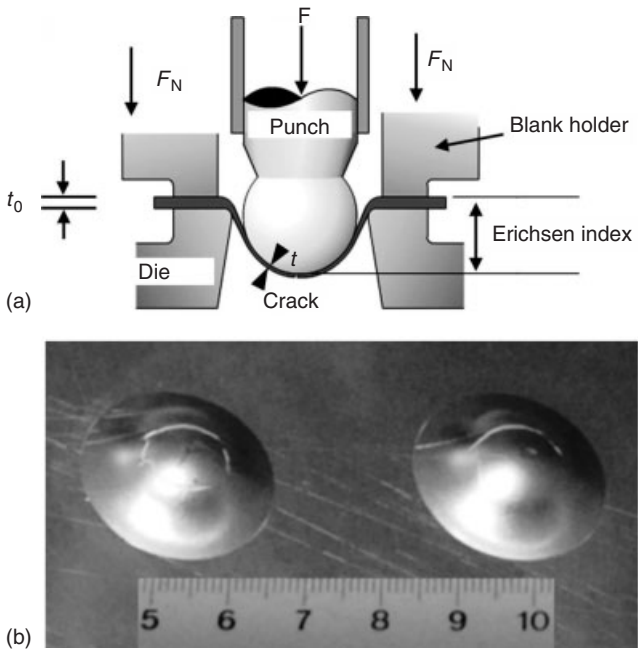


Figure 8. (a) Experimental setup for Erichsen test and (b) SMS failure (crack) in Erichsen test.

Owing to the tangential tension (σ_t), wrinkling in the flange and fracture of the cup wall can occur (Doege and Behrens, 2007). Investigations were carried out on SMS as well as on industrially manufactured materials (H400/PP/H400) (ThyssenKrupp Stahl, 2003) and pure metal sheets using the testing parameters: 225 mm^{ø} circular blank, 83 kN blank clamping force, 25 mm punch corner radius, 25 mm punch border radius, and a 25 mm drawing ring radius. The side length was 100 mm.

The results of the Erichsen tests performed show a slightly improved behavior of the laboratory-produced SMS (1.6 mm) compared to the industrially produced ones of the same layer thicknesses (H400, 1.6 mm), and come close to the value of the monolithic material (316 L, 1 mm thickness), see Figure 9. Investigations using optical light microscopy confirm that no delamination occurred.

The results of the deep-drawing tests show the same tendencies as those previously discussed for the stretching behavior. The approximately 74 mm deep-drawing depth for the SMS is comparable with that for the mono material and is slightly higher compared to the industrial material (Figure 9).

As mentioned earlier, the weakness of hybrid materials combining metals with polymers is thermal and mechanical joining. A possible solution for overcoming this weakness is to embed inlays (Figure 5) into the sandwiches. Therefore, the effect of different solid and mesh steel inlays on the formability and the flow of the material were investigated (Sokolova, Carradò, and Palkowski, 2011a). Some results are depicted in Figure 10, showing that using inlays in central position, failure by cracking occurs earlier compared to nonreinforced SMS, and that meshed inlays show an improved behavior compared to the solid REs. The SMS with 60 mm^{ø} solid REs failed by delamination caused by the small ratio of the supporting area (bonding between the single layers) to the inlay area, approaching is almost close to zero.

It can be concluded that the size as well as the geometry and the position of the inlays within the sample influence both variables (Sokolova, Carradò, and Palkowski, 2011b).

To investigate the local load conditions due to the change in thickness of the SMS, a photogrammetry method was used. This method is also known as *digital image*

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

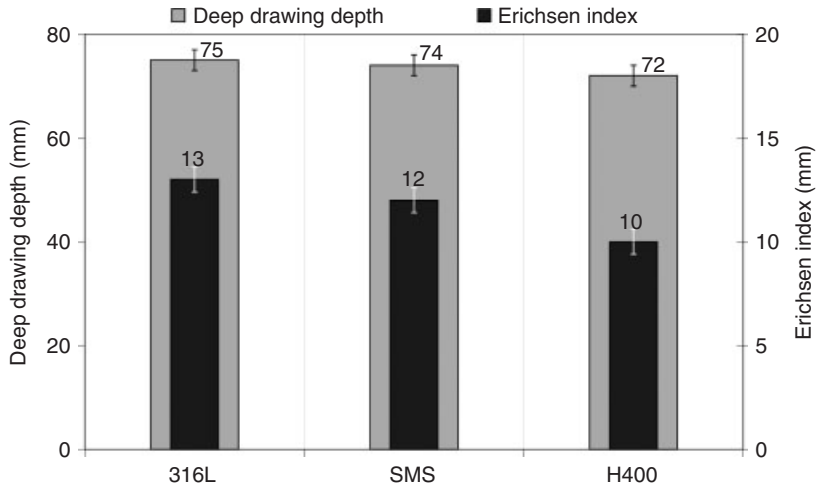


Figure 9. Results of deep drawing and Erichsen tests for 316L, SMS, and H400 (industrial sandwich), averaged maximum index; data: mean ± SD ($n = 5$).

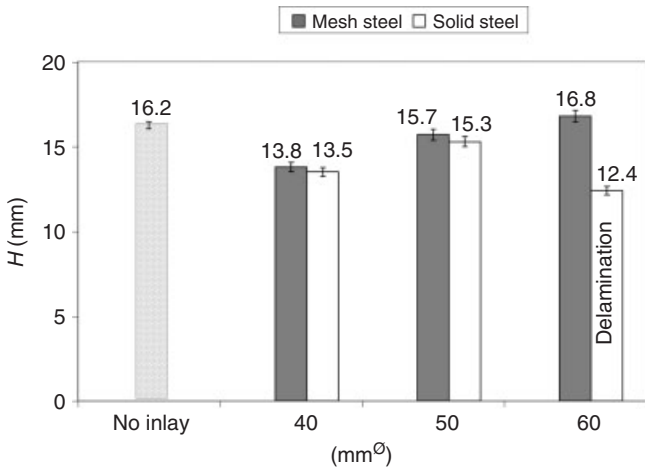


Figure 10. Cup test up to fracture for different inlay sizes. Sample: 68 mmø, punch: 33 mmø. *H*: cup height. Mean values ± SD ($n = 5$).

correlation (*DIC*) method. As an example, Figure 11 shows the thinning of the metal on the outer surface of a circular cup for SMS with and without reinforcing inlays. The effect of the reinforcement is clearly to be seen. Maximum thinning occurs in the nonreinforced cup in the area of the punch radius and the lowest reduction is given for the solid inlay that best strengthens the sample in the critical area.

7.3 Determination of forming limit curves

Following Keeler and Backofen (1963), the forming limit diagram (FLD) presents the ultimate ductility a sheet material may demonstrate under various strain conditions and a given boundary criterion, such as failure or the onset of necking. The FLD is determined to predict the extent to which sheet materials can be formed by deep drawing, stretch forming, or other load paths. In literature, several measures are described to determine FLDs. One of these is the Nakajima test (DIN 12004-2, 2008), based on the principle of deforming metal sheets with different waist geometries using a hemispherical stamp until fracture. By varying the sample section's width, different load paths can be adjusted and plotted on a true strain diagram; from deep-drawing conditions to uniaxial tension and plane-strain up to pure biaxial strain.

To determine the principle deformations and their directions, DIC was applied to measure the displacements of selected points on the sample following stamping. The principle strain values provide one point within the FLD. The combination of the points defines the FLD. Figure 12a shows the principle load paths in a major–minor strain diagram and the different geometries used for determining the forming limit curve (FLC) (Figure 12b).

The failure of a 1.6 mm thick SMS is subjected to different load conditions and was studied using the Nakajima test as well as 1 mm thick 316L sheets for comparison. Figure 13 shows the curves for both materials. It can be concluded that the mono material possesses slightly improved properties in the region of plane strain, but shows the same level for deep-drawing and biaxial loading conditions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

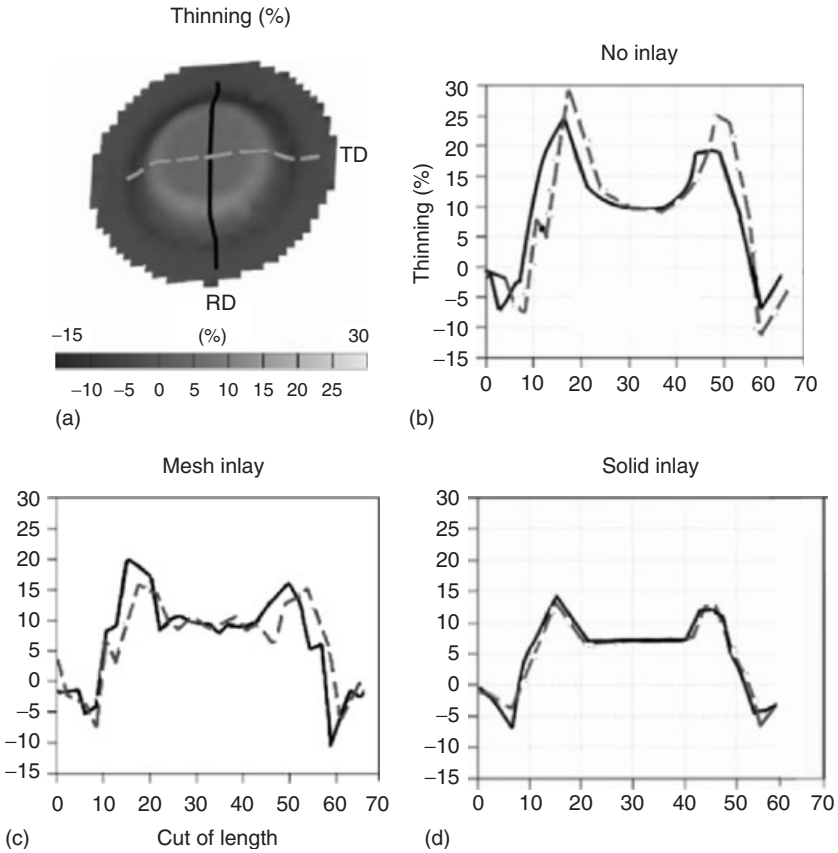


Figure 11. Thickness reduction along sections of SMS with and without inlays using photogrammetry (DIC); (a) blank 55 mm^{ø}, (b) punch 33 mm^{ø}, (c) RE 36 mm^{ø}, and (d) drawing depth 10 mm.

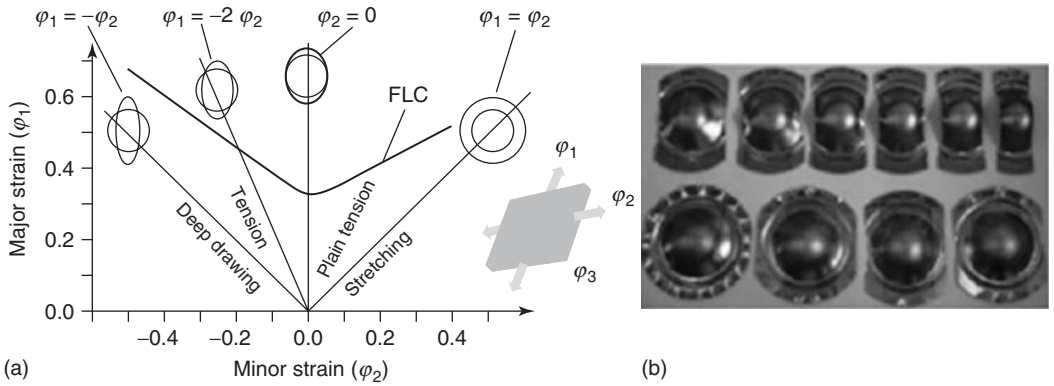


Figure 12. (a) Strain states in FLD; (b) geometry of the samples.

8 OUTLOOK

As could be shown, the development of hybrid materials is one of the most successful examples within the group of composites. Hybrid systems demonstrate considerable potential for designed properties following the specific demands of multifunctional parts and offer properties not achievable with mono materials. The combination of metals with fibers, particles, polymers, ceramics, and other materials, materials that are normally "incompatible," permit "designed" materials to be produced possessing the combined properties of the mono materials such as

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

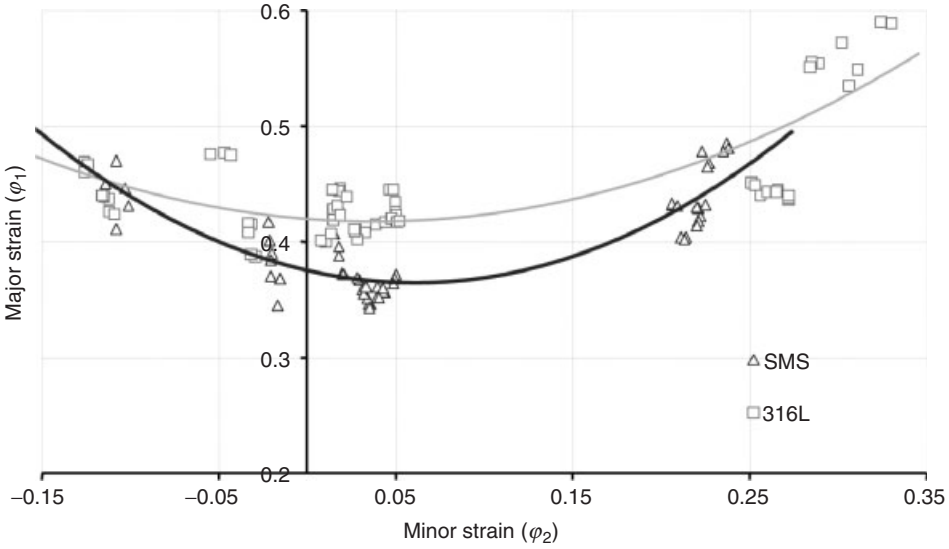


Figure 13. FLC of 316L sheet (1.0 mm) and SMS (0.5/0.6/0.5 mm).

high strength and deformability and are at the same time lightweight or have high damping, insulating, and corrosion-resistant properties.

Many of the composites currently on the market are nondeformable and are, therefore, only usable for batch production to produce final parts. For sheet parts in the automotive industry, continuous production is essential for significantly reducing the production costs. Deformable metal–polymer combinations can be a solution here.

Focusing on these types of sandwich materials, it can be concluded that compared to the mono metal, they have comparable deformation properties and show improved performances concerning lightweight and damping properties. Their weakness regarding thermal or mechanical joining can be eliminated by introducing reinforcing inlays, placed at those areas where vaporization could occur or where preloads are to be maintained under dynamic loading. Knowledge in this field is currently quite small but offers an enormous potential.

ACKNOWLEDGMENT

The authors are grateful for the support of the German Research Foundation DFG, the DAAD (D/0707603), EGIDE (PHC PROCOPE No. 17895XK), and APIC-DEU-PICS (No. 5245) for funding these investigations. Moreover, we thank ThyssenKruppNirosta for supplying us with the steel.

REFERENCES

Abdullah, S. and Fahruddin, A. (2009) Fatigue crack growth modelling of fibre metal laminate (FML) composite. *European Journal of Scientific Research*, **35** (1), 43–53.

Alcan composite GmbH (2009) Alukobnond Broschüre. 1–13.

Alcan composite GmbH (2010) Dibond Broschüre. 1–28.

Almeida, R.S., Damato, C.A., Botelho, E.C., *et al*. (2008) Effect of surface treatment on fatigue behavior of metal/carbon fiber laminates. *Journal of Materials Science*, **43**, 3173–3179.

ArcelorMittal (2011) Sollight AC. Vibration damping steel creates the sound of silence. ArcelorMittal Flat Carbon Europe. 1–2, http://www.arcelormittal.com/fce/repository/Brochures/SollightAC_datasheet_EN.pdf.

Ayari, F., Lazghab, T., and Bayraktar, E. (2009) Parametric finite element analysis of square cup deep drawing. *Computational Material Science and Surface Engineering*, **1** (2), 106–111.

Baldan, A. (2004) Adhesively-bonded joints and repairs in metallic alloys, polymers and composite materials. *American Journal of Materials Science*, **39**, 1–49.

Barnes, A.T. and Pashby, R.I. (2000) Joining techniques for aluminum spaceframes used in automobiles, part II—adhesive bonding and mechanical fasteners. *Journal of Materials Processing Technology*, **99** (1–3), 72–79.

Bayraktar, E., Isac, N., and Arnold, G. (2005) An experimental study on the forming parameters of deep-drawable steel sheets in automotive industry. *Journal of Materials Processing Technology*, **162–163**, 471–476.

Bischof, C., Bauer, A., Possart, W., *et al*. (1989) Zur Adhäsion in Metall-Polymer-Grenzschichten und ihrer praktischen Nutzung. *Acta Polymerica*, **40** (3), 214–221.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Bishopp, J. (2005) in *Handbook of Adhesives and Sealants—Basic Concepts and High Tech Bonding* (ed. P. Cognard), Elselvier, Oxford, vol. 1, pp. 163–214.

Boesemann, W., Godding, R., and Huette, H. (2000) Photogrammetric measurement techniques for quality control in sheet metal forming. *International Archives of Photogrammetry and Remote Sensing*, **33** (Part B5), 291–298.

Botelho, E., Silva, R., Pardini, L., and Rezende, M. (2006) A review on the development and properties of continuous fiber/epoxy/aluminum hybrid composites for aircraft structures. *Materials Research*, **9** (3), 247–256.

Bozhevolnaya, E. and Lyckegaard, A. (2005) Structurally graded core inserts in sandwich panels. *Composite Structures*, **68** (1), 23–29.

de Bruyne, N.A. (1939) Nature of adhesion. *Aircraft Engineer (London)*, **18** (12), 51–54.

Burchitz, I., Boesenkool, R.S., van der Zwaag, S., and Tassoul, M. (2005) Highlights of designing with Hylite—a new material concept. *Materials Science and Design*, **26**, 271–279.

Burianek, D.A. and Spearing, S.M. (2001) Delamination growth from face sheet seams in cross-ply titanium/graphite hybrid laminates. *Composites Science and Technology*, **61**, 261–269.

Campbell, F.C. (2010) *Introduction to Composite Materials* Chapter 1, Structural Composite Materials, ASM International, Materials Park.

Carradò, A., Sokolova, O., Ziegmann, G., and Palkowski, H. (2010) Press joining rolling process for hybrid systems. *Key Engineering Materials*, **425**, 271–282.

Carradò, A., Faerber, J., Niemeyer, S., *et al*. (2011a) Metal/polymer/metal hybrid systems: towards potential formability applications. *Composite Structures*, **93**, 715–721.

Carradò, A., Sokolova, O., Donnio, B., and Palkowski, H. (2011b) Influence of corona treatment on adhesion and mechanical properties in metal/polymer/metal systems. *Journal of Applied Polymer Science*, **120**, 3709–3715.

Chan, C.M. (1994) *Polymer Surface Modification and Characterization* Chapter 7., Hanser Gardner Publ., Cincinnati, USA.

Chen, Y.S., Hsu, T.J., and Chen, S.I. (1991) Vibration damping characteristics of laminated steel sheet. *Metallurgical and Materials Transactions A*, **22** (3), 653–656.

Chung, T.C. (2002) Synthesis of functional polyolefin copolymers with graft and block structures. *Progress in Polymer Science*, **27** (1), 39–85.

DIN 12004-2 (2008) Metallische Werkstoffe—Bleche und Bänder—Bestimmung der Grenzformänderungskurve—Teil 2: Bestimmung von Grenzformänderungskurven im Labor. ISO 12004-2: 2008. Deutsche Fassung EN ISO 12004-2.

DIN 50101 (1979) *Erichsen Test with Sheets and Belts* (ed. Deutscher Normenausschuss), Beuth Verlag GmbH, Berlin.

DIN 50114:1981–08: Prüfung metallischer Werkstoffe; Zugversuch ohne Feindehnungsmessung an Blechen, Bändern oder Streifen mit einer Dicke unter 3 mm, Beuth-Verlag GmbH, Berlin Germany., 1981.

Doege, E. and Behrens, B. (2007) *Handbuch Umformtechnik*, Grundlagen, Technologien, Maschinen, p. 915.

El-Dessouky, H.M. and Lawrence, C.A. (2013) Ultra-lightweight carbon fiber/thermoplastic composite material using spread tow technology. *Composites: Part B*, **50**, 91–97.

Engel, B. and Buhl, J. (2011) Metal forming of vibration damping composite sheets. *Steel Research International*, **82** (6), 626–631.

Euroinox (2012) Forming potential of stainless steel. Materials and Application Series, vol. **8**. The European Stainless Steel Research Association. 1–28.

Fan, Z., Tsakiropoulos, P., and Miodownik, A.P. (1994) A generalized law of mixtures. *Journal of Materials Science*, **29**, 141–150.

Grote, K.-H. and Antonsson, E.K. (2009) *Springer Handbook of Mechanical Engineering*, vol. **10**, Springer, Heidelberg.

Guo, Z.X. and Derby, B. (1993) Microstructural characterization in diffusion-bonded SiC/Ti–6Al–4V composites. *Journal of Microscopy*, **69** (2), 269–277.

Harris, B. (1991) A perspective view of composite materials development. *Material Design*, **12** (5), 259–272.

Hill H. (1994). *Introduction to the self-pierce riveting process and equipment*. IBEC'94 Body Assembly and Manufacture, Warren, **8**, 1–9.

Huck, W.R., Bosshard, B. 2007. Adhesive Bonding Concepts for Add-on Module in Assembly Lines. *Proceedings of the Car Body Hangons—8th European Automotive Meeting*. Bad Nauheim Frankfurt.

Hull, D. and Clyne, T.W. (1996) *An Introduction to Composite Materials*, Cambridge University Press, UK.

Ibarra-Castanedo, C., Avdelidis, N.P., Grinzato, E.G., *et al*. (2011) Delamination detection and impact damage assessment of GLARE by active thermography. *International Journal of Materials and Product Technology*, **41** (1), 5–16.

Jackson, K.P., Allwood, J.M., and Landert, M. (2007) Incremental forming of Sandwich panels. *Key Engineering Materials*, **344**, 591–598.

Johnson V.V.S. (1986) Impact and Residual Fatigue Behavior of ARALL and AS6/5245 Composite Materials. NASA Technical Memorandum 89013.

Jones, R.M. (1999) *Mechanics of Composites*, 2nd edn, Taylor & Francis, Philadelphia.

Kaftanoglu, B. and Alexander, J.M. (1961) An investigation of the Erichsen test. *Journal of the Institute of Metals*, **90**, 457–470.

Kawai, M. and Hachinohe, A. (2002) Two-stress level fatigue of unidirectional fiber–metal hybrid composite: GLARE 2. *International Journal of Fatigue*, **24**, 567–580.

Keeler, S. and Backofen, W. (1963) Plastic instability and fracture in sheets stretched over rigid punches. *ASM Transactions Quarterly.*, **56**, 25–48.

Kickelbick, G. (2007) in *Hybrid Materials. Synthesis, Characterization, and Applications* (ed. G. Kickelbick), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

Kim, K.J., Kim, C., and Choi, B.I. (2008) Formability of aluminium 5182-polypropylene sandwich sheet for automotive application. *Journal of Solid Mechanics and Materials Engineering*, **2** (4), 574–581.

Kurt, B., Wallenhorst, K., Siegfried, S. *et al*. (2004) Verfahren zur Herstellung einer Verbundplatte. Alcan Technology and management Ltd. Europäische Patentanmeldung. Int Cl. B29c 44/32. August 11, 2004.

Lahti, J. and Savolainen, A. (2004) The role of surface modification in digital printing on polymer-coated packaging boards. *Polymer Engineering Science*, **44**, 2052–2060.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Lange, K. 1993, Umformtechnik. Band 4: Sonderverfahren, Prozesssimulation, Werkzeugtechnik, Handbuch für Industrie und Wirtschaft. Produktion.

Lange, G., Carradò, A., and Palkowski, H. (2009) Tailored Sandwich Structures in the focus of research. *Materials and Manufacturing Processes*, **24** (10), 1150–1154.

LCVTP (2010) Low Carbon Vehicle Technology Project, http://www2.warwick.ac.uk/fac/sci/wmg/mediacentre/wmgcorporatebrochures/lcvtp_brochure_final.pdf.

Librescu, L. and Hause, T. (2000) Recent developments in the modeling and behavior of advanced sandwich constructions: a survey. *Composite Structures*, **48**, 1–17.

Liston, E.M. (1991) Plasma modification of polymer surfaces in *Polymer-Solid Interfaces* (eds J.J. Pireaux, P. Bertrand, J.L. Bredas), Institute of Physics, Belgium, pp. 429–442.

Liu, L. and Wang, J. (2004) Modeling springback of metal-polymer-metal laminates. *Journal of Manufacturing Science and Engineering*, **126**, 599–604.

Mackwood, A.P. and Crafer, R.C. (2005) Thermal modelling of laser welding and related processes: a literature review. *Optics & Laser Technology*, **37**, 99–115.

Michalos, G., Makris, S., Papakostas, N., *et al*. (2010) Automotive assembly technologies review: challenges and outlook for a flexible and adaptive approach. *CIRP Journal of Manufacturing Science and Technology*, **2**, 81–91.

Milton, G.W. (2002) *The Theory of Composites*, Cambridge University Press, UK.

Milton, S. and Grove, S.M. (1997) Composite Sandwich Panel Manufacturing Concepts for a Lightweight Vehicle Chassis. *Proceedings of the 30th International Symposium on Automotive Technology and Automation (ISATA)*, Florence, Italy.

Minzari, D., Møller, P., Kingshott, P., *et al*. (2008) Surface oxide formation during corona discharge treatment of AA 1050 aluminium surfaces. *Corrosion Science*, **50**, 1321–1330.

Mortimer, J. (2005) Jaguar roadmap rethinks self-piercing technology. *Industrial Robot an International Journal*, **32** (3), 209–213.

Narasimhan, K., Miles, M.P., and Wagoner, R.H. (1995) A better sheet-formability test. *Journal of Materials Processing Technology*, **50** (1–4), 385–394.

Oh, J., Cho, M., and Kim, J.S. (2005) Dynamic analysis of composite plate with multiple delamination based on higher-order zigzag theory. *International Journal of Solids and Structures*, **42** (23), 6122–6140.

Olsen, T.Y. (1920) Machines for ductility testing. *Proceeding of the American Society of Materials*, **20**, 398–403.

Olsen, E. (2007) Friction stir welding of high-strength automotive steel. Thesis at Brigham Young University.

Palkowski, H. and Lange, G. (2007) Creation of tailored high-strength, hybrid sandwich structures. *Advanced Materials Research*, **22**, 17–26.

Palkowski, H. and Lange, G. (2008) Production of tailored high strength hybrid sandwich structures, materials technology. *Steel Research International*, **79** (3), 27–36.

Palkowski, H., Giese, P., Wesling, V., *et al*. (2006) Neuartige Sandwichverbunde - Herstellung, Umformverhalten. *Fügen und Korrosionsverhalten. Mat.-wiss. u. Werkstofftechnik*, **377**, 605–612.

Pascual, M., Balart, R., Sánchez, L., *et al*. (2008) Study of the aging process of corona discharge plasma effects on low density polyethylene film surface. *Journal of Materials Science*, **43** (14), 4901–4909.

Pickin, C.G., Young, K., and Tuersley, I. (2007) Joining of lightweight sandwich sheets to aluminium using self-pierce riveting. *Materials & Design*, **28** (8), 2361–2365.

Possart, W. (2005) *Adhesion: Current Research and Applications*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

Rajput, R.K. (2007) *A Textbook of Manufacturing Technology: Manufacturing Processes*, Firewall Media, Laxmi publication pvt ltd New Delhi.

Reinhart, F.W. (1954) Nature of adhesion. *Journal of Chemical Education*, **31** (3), 128.

Roque, C.M.C. and Thomson, O.T. (2005) Modeling of composite and sandwich plates by a trigonometric layer wise theory and multiquadrics in *Sandwich structures 7: Advancing with sandwich Structures and materials*, Springer, Berlin, pp. 231–240.

Rosato, D.V., Rosato, D.V., and Rosato, M.G. (2000) *Injection Molding Handbook*, Springer, Heidelberg.

Rothon, R.N. (2003) *(Ed.) Particulate-Filled Polymer Composites*, 2nd edn, Rapra Technology, Shawbury, UK.

Rudd, C.D., Long, A.C., Kendall, K.N., and Mangin, C.G.E. (1998) *Liquid Moulding Technologies*, Woodhead Publishing Limited, Cambridge, England.

Ruokolainen, R.B. and Sigler, D.R. (2008) The effect of adhesion and tensile properties on the formability of laminated steels. *Journal of Materials Engineering and Performance*, **17** (3), 330–339.

Schindel-Bidinelli, E.H. and Gutherz, W. (1998) *Konstruktives Kleben*, VCH, Weinheim.

Sellin, N., Campos, J., and Sinézio de, C. (2003) Surface composition analysis of PP films treated by corona discharge. *Material Research*, **6** (2), 163–166.

Shirzadi, A.A. and Wallach, E.R. (1997) Temperature gradient transient liquid phase diffusion bonding: a new method for joining advanced materials. *Science and Technology of Welding and Joining*, **2** (3), 89–94.

Sieben, M. and Brunnecker, F. (2009) Welding plastic with lasers. *Nature Photonics*, **3**, 270–272.

Sokolova, O., Carradò, A., and Palkowski, H. (2010) Production of customized high-strength hybrid sandwich structures in *Creation of High-Strength Structures and Joints*, vol. **137** (eds H. Palkowski and K.-M. Rudolph) Advanced Materials Research, Trans Tech Publications, Switzerland, pp. 81–128.

Sokolova, O., Carradò, A., and Palkowski, H. (2011a) Metal-polymer-metal sandwiches with local metal reinforcements: a study on formability by deep drawing and bending. *Composite Structures*, **94** (1), 1–7.

Sokolova, O., Carradò, A. and Palkowski, H. (2011b) Adhesion and formability of thin steel/polymer/steel hybrid sandwich composites. Steel Research International, ICTP, pp. 435–440.

Strobel, M., Lyons, C.S., and Mittal, K.L. (1994) *Plasma Surface Modification of Polymers: Relevance to Adhesion*, VSP BV Publ., Utrecht, NL.

Sun, C., Zhang, D., and Wadsworth, L.C. (1999) Corona treatment of polyolefin films—a review. *Advanced Polymer Technology*, **18**, 171–180.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ThyssenKrupp (2008) ThyssenKrupp Steel Europe. Deep-Drawing Steels DD, DX and DC. Steels for Complex Forming Requirements. Deep Drawing Steels. 1–17.

ThyssenKrupp Stahl (2003) Bondal® Körperschalldämpfender Verbundwerkstoff, 5/2003. Bestell - Nr. 2060.

Timings, R. (2008) *Fabrication and Welding Engineering*, Newnes, Elsevier Linacre House, Jordan Hill, Oxford, UK.

van Tooren M.J.L. 2004. Airbus composite aircraft fuselages—next or never. A New Aircraft Material in Context. Engineering, Metallic Materials, Characterization and Evaluation Materials and Automotive and Aerospace Engineering, 145–157.

Tschaetsch, H. (2006) *Metal Forming Practise, Processes–Machines–Tools*, Springer-Verlag, Berlin Heidelberg.

Tsoukantas, G. and Chryssolouris, G. (2006) Theoretical and experimental analysis of the remote welding process on thin, lap-joined AISI 3004 steel. *International Journal Advanced Manufacturing Technology.*, **35**, 880–894.

Tsoukantas, G., Salonitis, K., Stavropoulos, P., *et al*. (2005) On optical design limitations of generalized two-mirror remote beam delivery laser systems: the case of remote welding. *International Journal Advanced Manufacturing Technology*, **32**, 932–941.

Varis, J.P. (2003) The suitability of clinching as a joining method for high-strength structural steel. *Journal of Materials Processing Technology*, **132** (1–3), 242–249.

Vermeren, C.A.J.R., Beumler, T.H., De Kanter, J.L.C.G., *et al*. (2003) Glare design aspects and philosophies. *Applied Composite Materials*, **10**, 257–276.

Vinson, J.R. (2005) Sandwich structures. Present, past and future in *Sandwich Structures 7: Advancing with Sandwich Structures and Materials*, Springer, Berlin, pp. 3–12.

Vogelsang, L.B. (1983) Development of a new hybrid material (ARALL) for aircraft structure. *Industrial and Engineering Chemistry Product Research and Development*, **22**, 492–496.

Weiss, M., Dingle, M.E., Rolfe, B.F., and Hodson, P.D. (2007) The influence of temperature on the forming behaviour of metal/polymer laminates in sheet metal forming. *Journal of Engineering Materials and Technology*, **129**, 530–537.

Wertheimer, M.R., Martinu, L., Klemberg-Sapieha, J.E., and Czeremuszkin, G. (2002) Plasma treatment of polymers to improve adhesion in *Adhesion Promotion Techniques* (eds K.L. Mittal and A. Pizzi), Marcel Dekker Inc. Publ., New York.

Wijskamp, S. (2005) Shape distortions in composites forming. PhD thesis, University of Twente, Enschede, NL.

Xiao, X., Huang, L., Liao, Y., *et al*. (2011) Numerical simulation of deep-drawing processes of square cup under bilateral constrained conditions. *Advanced Materials Research*, **189–193**, 2892–2896.

Zenkert, D. (1997) in *The Handbook of Sandwich Construction* (ed. D. Zenkert), Engineering Materials Advisory Services Ltd. (EMAS), UK.

Zhang, D., Sun, G., and Wadsworth, L.C. (1998) Mechanism of corona treatment on polyolefin films. *Polymer Engineering Science*, **38**, 965–970.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto163
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Recycling of Polymers and Composites

Bassam J. Jody, Edward J. Daniels, and Jeffrey S. Spangenberger

Argonne National Laboratory, Argonne, IL, USA

1 Introduction and Background 1
2 Impact of Recycling Automotive Polymers on Sustainability of the Industry 2
3 Technologies for Recycling Automotive Polymers 4
4 Technologies for the Separation of Polymers from Shredder Residue as a Concentrate 5
5 Technologies for the Recovery of Polymers from the Polymer Concentrate 5
6 Chemical Processes for Recycling Plastics 6
7 Utilization of Automotive Polymers as Energy Sources 7
8 Recycling of Polymer Composites 8
9 Landfilling of the Nonrecyclable Materials 9
10 Conclusions 9
Acknowledgments 9
References 9

1 INTRODUCTION AND BACKGROUND

The worldwide automobile industry is a growing material- and energy-intensive industry (Figure 1). The growth is most rapid in developing countries, such as China and India (Figure 2). In 2001, the vehicles on the road in the United States contained 5.3% of all steel and 13.8% of all aluminum in use in the United States (USGS, 2006a, 2006b).

Figure 3 shows the changes in the composition of a typical car in the United States during 1995–2009. The newer cars contain a higher percentage of polymers and composites. This use of materials is likely to continue as the automotive industry strives to reduce the weight of vehicles and as more efficient and economical composites are developed (Bain, 2012). Despite the use of lighter materials, the average weight of a vehicle has increased from about 1675 kg in 1995 to 1775 kg in 2009 (Davis, Diegel, and Boundy, 2011). This increase is due to the increase in the number of larger sport utility vehicles (SUVs). As the new CAFE (Corporate Average Fuel Economy) regulations take effect, automobile manufacturers will be using more of the lighter materials, including high strength steels, aluminum (Al), and polymers and composites.

At present, end-of-life vehicles (ELVs) are profitably processed for materials and parts recycling by an existing market-driven infrastructure that includes dismantlers, remanufacturers, and shredders. Dismantlers process the ELVs and recover usable parts. The shredders recover ferrous and nonferrous metals from the remaining auto "hulk." Recycling the polymers and composites is limited by the lack of (i) commercially proved technologies to economically recover them as marketable products and (ii) stable post-use markets. Therefore, most of the nonmetals are landfilled as part of the shredder residue generated by the shredders that wastes the value of these materials in addition to increasing the waste volume in the landfills.

Future vehicles will also contain different power trains (electric and hybrid vehicles). All of these factors will impact the recyclability of the ELVs, including their polymers. Replacing iron and steel, which are recyclable, with lightweighting materials will reduce the recycling rate even if the lightweighting metals are recycled at the same rate

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

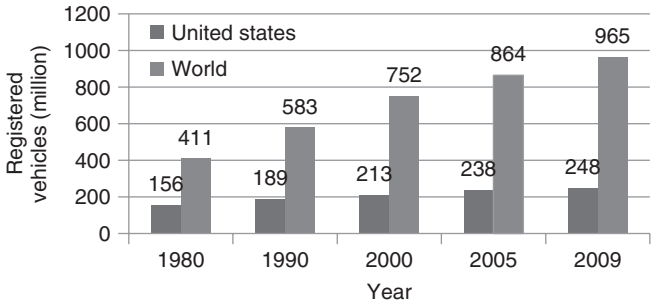


Figure 1. Registered cars, buses, and trucks. (Data taken from Davis *et al*., 2011.)

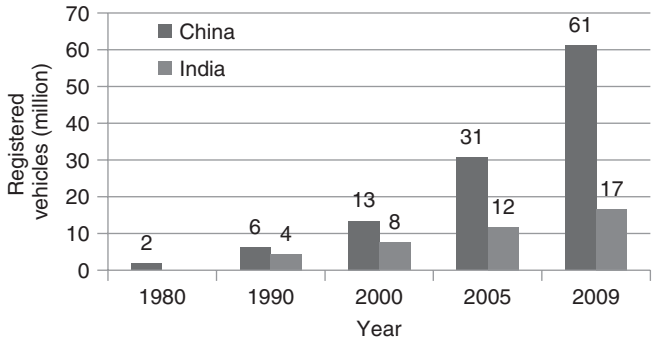


Figure 2. Registered cars, buses, and trucks in China and India. (Data taken from Davis *et al*., 2011.)

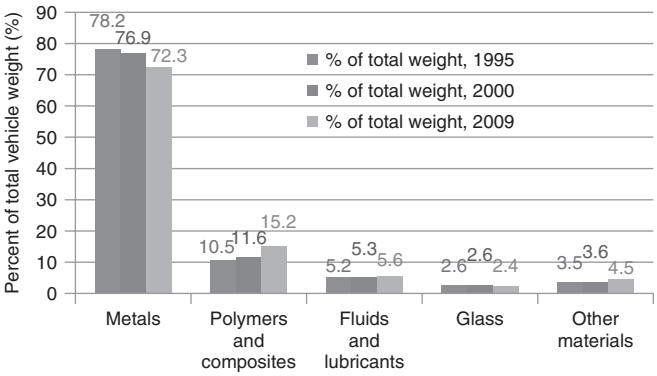


Figure 3. Material content of an average vehicle in the United States. (Data taken from Davis *et al*., 2011.)

as steel and iron because the relative weight of the metals in the vehicle will decrease. Therefore, to maintain high vehicle-recycling rates, it is important that the polymers be recycled. Recycling the polymers and composites is also necessary in order to comply with emerging regulations. For example, in 2000, the European Union (EU) legislators issued a (draft) directive [Directive 2000/53/EC (September 18, 2000); Sander *et al*., 2002; Konz, 2009; Decicions, 2010; Parliament, 2010] that called for reuse and recovery rates of materials from the ELVs to be at least 85% by January 1, 2006, and 95% by January 1, 2015.

2 IMPACT OF RECYCLING AUTOMOTIVE POLYMERS ON SUSTAINABILITY OF THE INDUSTRY

Polymers and composites are energy-intensive materials that are made primarily from petroleum and can last hundreds of years in landfills. Technology is needed to recycle them economically and in an environmentally responsible manner. The focus is on recycling the thermoplastics, which constitute about 80% of the automotive polymers, and they could be remolded into new products [polypropylene (PP), polyethylene (PE), polystyrene (PS), acrylonitrile–butadiene–styrene (ABS), nylons, polycarbonate (PC), PC-ABS, and polyvinyl chloride (PVC)].

Composites consist of a polymer matrix reinforced with fibers. Teijin Ltd. (in Tokyo, Japan) announced on November 30, 2011, that it will *establish the world's first pilot plant for fully integrated production of carbon fiber-reinforced thermoplastic (CFRTP) automotive components* (Teijin, 2012). Composites have many attractive features in automotive applications, including high strength-to-weight ratio and corrosion resistance, and can be molded into complex parts, which reduce manufacturing costs (Table 1).

Optimizing the recycling of automotive materials requires an integrated approach (Reuter *et al*., 2004) by the auto companies and their suppliers, dismantlers, and shredders. ELVs start their final journey at a dismantling site where usable parts are recovered for resale. Direct reuse of a part conserves the materials and energy that would otherwise be required to produce the part from raw materials, as well as the manufacturing energy required to stamp the part and assemble the component. Some parts,

Table 1. Estimated mass and tooling investment reductions when composites are used to replace conventional automobile components.

Part	Mass Reduction (%)	Reduction in Tooling Investment (%)
Composite hoods	30–40	60–70
Fenders	25–35	55–65
Deck lids	25–35	50–60
Floor pans	30–40	Up to 60
Trunk compartments	50	70
Bumper beams	30–40	Not applicable

Data taken from Plastemart, 2011.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

such as starters, alternators, engines, and transmissions, may require rebuilding. Others, such as thermoplastic olefin (TPO) bumpers, require only cosmetic repairs. The plastics in these parts are sometimes overlooked when estimating the recycling rates, even though they represent "primary" recycling of the polymers (Duranceau, 2009).

Most automotive polymers end up in the shredder residue after the "hulks" are shredded and their metals are recovered. More than 10 million vehicles annually are scrapped in the United States (Daniels *et al*., 2004). A recent European report reported the number for the 25 EU countries to be over 14 million in 2010 compared to 12.7 million in 2005 (Parliament, 2010).

2.1 Economic viability of recycling polymers and composites from ELVs

The economics of the recycling process is governed by a number of factors, including

- Value of the recovered materials. This depends on the type and quantity of material recovered, its purity, and the consistency of its properties and composition.
- Cost of separating, purifying, and removing contaminants from the recovered materials.
- The value of recycled plastics is generally between 50% and 75% of the value of their virgin counterpart. For example, for PP, which is the most used automotive plastic, during the fourth quarter of the year 2011, the price of clean regrind or flakes was about \$1.48–1.74 per kg, and the price of virgin injection grade PP was between \$2.18 and \$2.67/kg (Plastics News Magazine, 2011).
- A big portion of the cost of recovering the automotive polymers can be the cost of purification or upgrading of the material after it is separated from the source stream. For example, polyolefins can be easily separated by sink/float techniques. However, the separated polyolefins may contain wood, rubber, foam, and other residual materials. Removal of these residual materials from the target material may cost more than the cost of separating the olefins from the source stream.

2.2 Energy value of recycled automotive polymers

Over 6 million metric tons of thermoplastics worth about \$12 billion can be recovered from obsolete vehicles annually worldwide. These numbers are determined using the following assumptions: (i) 5% of the 950 million vehicles on the road reach the end of their useful life annually, (ii) the average weight of a vehicle is 1770 kg, (iii) 10% of the weight of an ELV is polymers and composites, (iv) 80% of the polymers are thermoplastics, (v) 90% of the polymers end up in landfills (which is the case at present), and (vi) recycled polymers and composites are assumed to be worth \$1/kg.

Figure 4 summarizes the embodied energy data (energy required to produce the material or the product starting from its raw materials) for some automotive plastics. Except for some disagreements with the older Oak Ridge National Laboratory (ORNL) (ORNL, 1994) data for the olefins, the data are in good agreement. Figure 4 shows that the embodied energy for most of these plastics is between 70,000 and 100,000 kJ/kg with an average of about 85,000

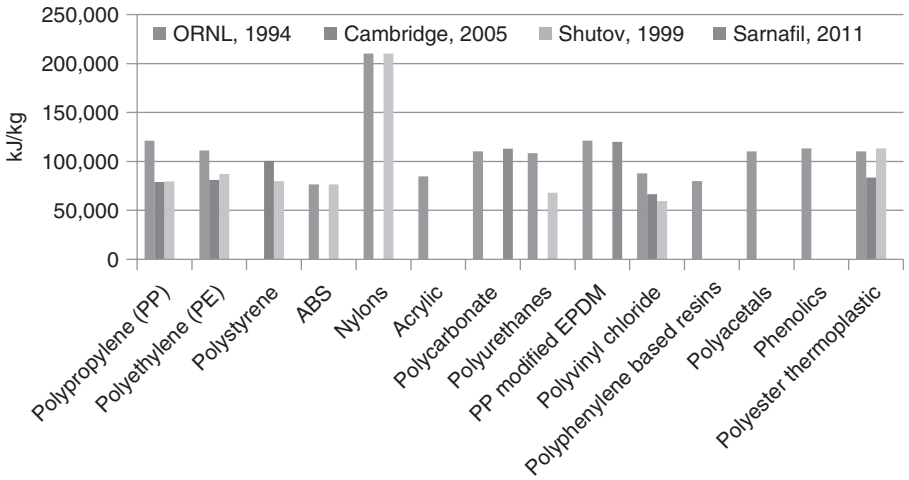


Figure 4. Embodied energy of plastics from different data sources. (Data taken from ORNL 1994, Cambridge 2005, Shutov 1999, and Sarnafil 2011.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

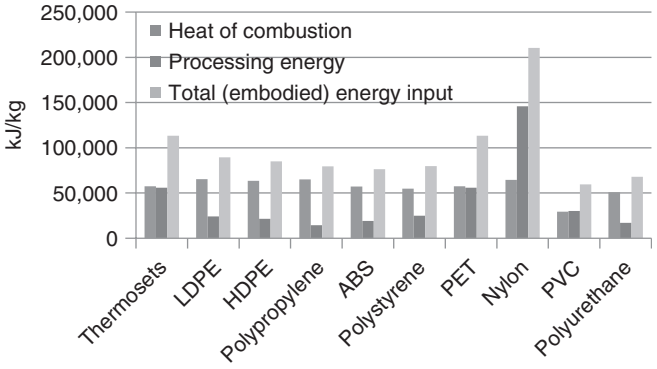


Figure 5. Breakdown of the embodied energy of polymers. (Data taken from Shutov, 1999.)

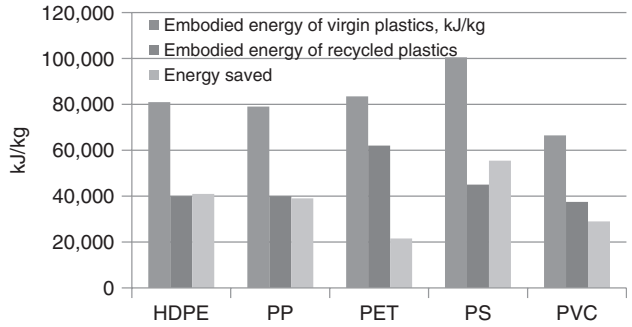


Figure 6. Embodied energies of virgin and recycled plastics. (Data taken from Cambridge, 2005.)

kJ/kg. At present, the olefins constitute about two-thirds of the plastics in vehicles. The data suggest that the average embodied energy for the automotive plastics is about 75,000 kJ/kg. Figure 5 shows the breakdown of the embodied energy of some polymers.

Figure 6 shows the approximate energy savings that can be realized by recycling some automotive plastics. Overall, the savings are about 37,500 kJ/kg of plastics recycled. As stated earlier, there are about 950 million registered vehicles in the world, and each contains about 227 kg of plastics. Recycling these plastics can result in energy savings exceeding 7.5×10^{15} kJ. The total world energy consumption in 2008 was about 475×10^{15} kJ.

2.3 Environmental impact of recycling

Recycling has many positive environmental impacts, including (i) reduction in demand for and depletion of natural resources; (ii) protection of the ecosystem by reducing the amount of waste and pollution; and (iii) energy savings, which will result in less pollution, including a reduction in the emission of greenhouse gases (GHGs).

Most plastics are derived from oil or natural gas, and processing the oil or gas to produce the monomers and polymerization of the monomers are energy-intensive processes. Assuming that 0.07 kg of CO_2 equivalent is emitted per million kilojoule of energy used, about 2.8 kg of CO_2 equivalent emissions will be avoided for every metric ton of plastics recycled. A life-cycle analyses study (Gallon and Binder, 2006) was conducted on automotive polymer recycling processes, including mechanical and chemical separation of the plastics for reuse as plastics and conversion to "fuel" or incineration to recover their heat value. The analyses concluded that both types of recycling resulted in environmental benefits. However, the environmental benefits are higher when the plastics are recovered for reuse as plastics. The maximum benefits can be obtained by combining both processes where materials that are recyclable as materials are first separated for reuse and the residual is then used to produce fuels.

3 TECHNOLOGIES FOR RECYCLING AUTOMOTIVE POLYMERS

Because most polymers and polymer composites end up in shredder residue, emphasis will be on separating and recovering them from shredder residue.

3.1 Shredder residue

Shredder residue contains polymers, composites, moisture, wood, metals, glass, sand, dirt, automotive fluids, fabrics, and fibers. Researchers van Schaik and Reuter (2004) have suggested that shredder residue data should be reported within statistical limits when used to calculate the recycling rate. Pineau, Kanari, and Menad (2005) have also estimated that a minimum sample size of 140 kg is required for the sample to be 90% representative. Table 2 shows the differences in the composition of two shredder residues processed by Argonne's process for recycling polymers from shredder residue. Table 3 compares the composition of the polymer concentrates separated from the two residues. Tables 2 and 3 represent the averages from processing over 11,000 kg of shredder residue 1 and 15,000 kg of residue 2 over a period of 6 months. Over 90% of the polymers were recovered as a "polymer concentrate."

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Gross composition of two shredder residue samples.

Material	Weight % in Shredder Residue 1	Weight % in Shredder Residue 2
Fines (<6 mm)	24	60
Polymer concentrate	36	14
Oversized heavies (metals and rocks)	08	02
Lights rejects and oversized foam	11	03
Metals-rich fraction	05	06
Loss (moisture, dust, and sweeps)	16	15

Table 3. Composition of the polymer concentrates described in Table 2.

Material	Weight % in Shredder Residue 1	Weight % in Shredder Residue 2
Mixed rubber	50	30
Polyolefins	17	30
Polystyrene	2	3
ABS	5	8
Polycarbonate	3	3
Nylon	2	1
Polyvinylchloride	2	3
Unknown polymers (rich in thermosets)	13	7
Other (including metals and wood)	6	15

3.2 Separation of polymers from shredder residue as a concentrate

Many of the polymers in shredder residue have overlapping properties. Rasshofer and Schomer (2003) have stated that except for ferrous metals, it is unlikely that a material can be recovered at sufficient purity from shredder residue in a single step or by mechanical means only. Several steps are normally required: (i) separation of the polymers from shredder residue, as a polymer concentrate, (ii) separation of the plastics from the polymer concentrate, (iii) separation of wood, rubber, and inert materials from the plastics, (iv) purification of the separated plastics, and (v) cleaning of the recovered materials. Processes for the recycling of materials from shredder residue are discussed in a number of review articles (Jody *et al*., 2010, Lundqvist *et al*., 2004; Zevenhoven and Saeed, 2003).

4 TECHNOLOGIES FOR THE SEPARATION OF POLYMERS FROM SHREDDER RESIDUE AS A CONCENTRATE

The Argonne Process. Argonne National Laboratory developed and tested a process for isolating the polymers as a concentrate (Jody, *et al*., 2010). The plant recovers >90% of the polymers in the shredder residue and >90% of the metals.

The Nimco Process. This process has four subsystems: (i) sizing and separation of shredder residue, (ii) grit processing, (iii) polymer recovery, and (iv) polyurethane foam separation and cleaning (Baker, Woodruff, and Naporano, 1994).

The Witten Process. This process was developed for producing an organic-rich fraction to be used as a fuel in cement kilns (Competitive Analysis Centre, Inc. and Economic Associates Inc., 1998). The process separates the ferrous metals, size reduces the remaining material, and removes fines. Air classifiers are used to separate fibrous from nonfibrous material, which contains the polymers.

The Sortec Process. This process produces an organic fraction from shredder residue (Competitive Analysis Centre, Inc. and Economic Associates Inc., 1998). The process consists of (i) separation of fines and recovery of large metal pieces; (ii) size reduction of the remainder to <2 cm; (iii) separation of ferrous metals; and (iv) drying, screening, and air classification. The light fraction is the organic fraction, which contains the polymers.

The Wesa-SLF Process. This process (Sattler and Laage, 2000; Zevenhoven and Saeed, 2003) separates the residue by size into three fractions and then size reduces the oversized material and recovers the ferrous metals. The material is then reduced to <7 mm and combined with the middling fraction, dried, air classified, screened, and then separated from the organic fraction, which contains the polymers.

In addition, many in the shredder industry have also experimented with such similar technologies mostly for increasing their metals recovery from shredder residue.

5 TECHNOLOGIES FOR THE RECOVERY OF POLYMERS FROM THE POLYMER CONCENTRATE

Over the past decade, major advances have been made in process development for separating and recovering polymers from shredder residue. Some of these advances are discussed in the following.

The Argonne Process for the Separation and Recovery of the Plastics. Argonne developed a froth flotation process (Jody *et al*., 2011; Jody, Pomykala, and Daniels, 2003b). The process consists of the following steps: (i) separate the granulated polymer concentrate using conventional sink/float techniques into more manageable polymer groups, (ii) recover targeted plastics from the concentrated groups, and (iii) purify the targeted species to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

increase purity. Argonne built a pilot plant and a 3 tons/h module and processed about 10,000 kg of polymer concentrate. In addition, a 20 tons of shredder residue per hour plant was built at a shredder site based on this process.

The MBA Polymers Process. MBA polymers built plastics recycling plants in China and Austria and a plant in England to recover automotive plastics from shredder residue. The plant has a design capacity of 60,000 tons/year (MBA, 2012).

The Recovery Plastics International (RPI) Process. This process uses a plasticizer to modify the surface of plastics. The process was applied for the recovery of plastics from shredder residue (Wielgat, 2002).

The Salyp Process. This process used infrared (IR) heating to soften and separate the plastics. However, when this process was applied to mixtures of plastics, the purity was low because the softening temperatures of the species overlapped.

The Volkswagen-SiCon Process. Construction of the first large-scale Volkswagen-SiCon plant was in 2005, and it had an annual design capacity of 100,000 tons of shredder residue. The first stage of a smaller plant based on Volkswagen-SiCon technology was commissioned in Belgium in 2005(Guschall *et al*., 2005). SiCon launched its first own industrial plant for the treatment of shredder residue in Antwerp, Belgium. The plant has a design capacity of up to 20,000 tons/year in a three-shift operation (Recycling International, 2008).

The Galloo Process. This process is for recycling thermoplastics from end-of-life products including ELVs and produces about 25,000 ton of recycled materials per year (Recycling International, 2012). The process produces several fractions, including a plastics fraction that is processed in a wet separation system. The remaining fraction is processed in a heavy-media separation plant to recover the metals (Galloo, 2012a). It is reported that the Galloo sink/float process is able to "make density separations accurate to within two points to the third decimal place" and is continually expanding its operation (Galloo, 2012b).

Rapid Identification and Separation Processes. These include IR and near-infrared (NIR), X-ray, UV fluorescence, Raman spectroscopy, and color sorting separation equipment. Some can produce >90% pure products from some mixtures. They can be combined with color sorters to achieve greater separation. Technologies for detecting dark plastics that have high content of carbon black by IR separators are still under development. Rapid identification of polymers can be achieved using NIR spectroscopy because many polymers have unique spectra in the 1350–1800 nm region (McDermott, Jenner, and Crocombe, 2007). X-ray separation is applied mostly to PVC because the chlorine atoms in PVC generate an easy-to-detect peak in the X-ray spectrum. The UV fluorescence method has been applied mostly to the separation of mixtures containing nylons/PC and PC/polymethyl methacrylate (PMMA), because PC fluoresces and the others do not. Some color sorters can distinguish between very close colors, such as light blue and clear, and can detect more than one color simultaneously. However, many plastics exist in more than one color. Raman spectroscopy uses characteristic peaks of specific chemical groups to identify the polymers.

Many types of such equipment are commercially used to sort plastics. A few of the many organizations that developed such equipment are TITECH (Titech, 2012a, 2012b), MSS, Inc. (MSS, 2012), and NRT (NRT, 2012).

6 CHEMICAL PROCESSES FOR RECYCLING PLASTICS

6.1 Methods using organic solvents

Selective dissolution of plastics from a mixed stream can recover highly pure plastics (PIA, 1980; Tesoro, 1987; Nauman and Lynch, 1993, 1994; Jody *et al*., 1990; Jody, Daniels, and Bonsignore, 2001, Daniels, Jody, and Bonsignore, 1990). Solvent-based processes are in practice in Europe—for example, the Solvay Chemicals Company's "Vinyloop" batch process and the Delphi process.

The Solvay Process. The Solvay "Vinyloop" process recovers PVC from PVC-containing scrap. The PVC is dissolved in methyl ethyl ketone, and the solution is pumped into a recovery tank, where steam is injected into the solution to precipitate the PVC.

The Delphi Process. This process removes PVC from automotive wire harnesses using solvents to swell and soften the PVC, but not dissolve it, so that it can be separated from the copper wires by centrifugation.

The Rensselaer Polytechnic Institute Process. In this process (Nauman and Lynch, 1993, 1994), a solvent (or solvents) is used to dissolve plastics from a mixture and then flash devolatilization is used to precipitate the dissolved plastics.

The Argonne Process. Argonne's process (Jody, Daniels, and Bonsignore, 2001; Jody, Daniels, and Pomykala, 1996; Daniels, 1994; Jody *et al*., 1994) uses multiple solvents to extract plastics. The process produces ABS, PVC, and PP/PE at high purity (>98%).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6.2 Polymer depolymerization to produce monomers

Several depolymerization processes to produce monomers have been developed (American Plastics Council, 1999). Polymers such as polyethylene terephthalate (PET) and polyurethanes can be depolymerized by glycolysis, methanolysis, and hydrolysis. Nylon 6 can be reacted with high temperature steam to produce caprolactam, which is used for making nylon 6. Mixtures of nylon 6 and nylon 66 can be reacted with ammonia at elevated temperatures to produce the monomers of both nylons (American Plastics Council, 1999). InfiChem Polymers (InfiChem, 2012) commercialized a process for converting polyurethane foam to polyols, which could be used for making new foam.

7 UTILIZATION OF AUTOMOTIVE POLYMERS AS ENERGY SOURCES

The energy value of polymers can be recovered by (i) burning the polymers and recovering their heat of combustion and (ii) converting the polymers to liquid and/or gaseous fuels.

7.1 Burning with heat recovery

The heating value of shredder reside is about half that of coal, and the heating value of the polymer concentrate is more than one and a half times that of coal. Shredder residue is also low in sulfur but high in ash and moisture. Burning shredder residue can reduce its mass and volume by more than 50% and 75%, respectively. However, it is not widely accepted because landfills are less expensive in many places and because shredder residue contains substances of concern (SOCs).

The US Department of Energy (DOE) conducted a study where shredder residue was burned in a rotary kiln (Hubble, Most, and Wolman, 1987). Incineration tests were also conducted where small percentages of shredder residue were co-fired with municipal solid waste (MSW) (Keller, 2003; Mark and Fisher, 1998, 1999). No major operational problems were reported. However, dioxins/furans and Cd, As, Pb, and Zn increased, but the gas cleanup system was able to handle it. The ash showed increased levels of Zn, Cu, Sb, Ni, Pb, and Sn (Mark and Fisher, 1998, 1999).

The Competitive Analysis Centre, Inc. and Economic Associates, Inc. (1997) conducted a study to evaluate the potential use of the organic fraction of shredder residue as an energy source and as a source of reducing agents in blast furnaces. The polymers were used to replace the coke or coal supplied to the blast furnace.

7.2 Pyrolysis

Pyrolysis of polymers has been demonstrated (Banks, Lusk, and Ottinger, 1971). Problems encountered (Leidner, 1981) include (i) longer residence times because plastics have poor heat transfer properties, (ii) carbon residue sticks to the reactor walls, and (iii) plastics produce viscous material that is difficult to pump.

Keller (1999a) reported the following observations concerning pyrolysis of shredder residue: (i) iron makes up to 25% of the solid product and copper up to 5%, (ii) the concentration of polycyclic aromatic hydrocarbons (PAHs) was reduced by more than 90% and polychlorobiphenyls (PCBs) were reduced by more than 99%, and (iii) the solids could not be landfilled because of the heavy metals. Day, Cooney, and Shen (1996) described a process for pyrolyzing shredder residue at about 500°C. The combined nitrogen- and sulfur-containing compounds constituted >10% of the weight of the oil. The oil also contained a very small amount of chlorinated compounds. Pyrolysis tests were also conducted at subatmospheric pressures (Roy and Chaala, 2001; Chaala, Ciochina, and Roy, 1999). The products were 52% solids, 28% oil, 13% water, and 7% gas.

Several large-scale processes pyrolized shredder residue. Siemens-KWU developed a process followed by combustion of the products for steam production. The plant conducted one trial using 30 tons of shredder residue. The Batrec process involves pyrolysis of the shredder residue followed by mechanical separation of the metals (iron and copper) from the solids. Keller (1999a, 1999b) conducted in the plant pyrolysis tests in a 400-kg/h reactor at 600°C using shredder residue. The Takuma process was tested in a 90-ton/day plant in Japan. The shredder residue is first pyrolized and then the residual solids are sorted to recover metals. The char is combusted along with the gases produced by the pyrolysis process, and the heat is recovered in a steam boiler. The citron oxyreducer process (Brüggler, 2002) is in operation; the product gases, which are rich in CO and H_2, are used to reduce the metal oxides and hydroxides.

7.3 Gasification

Gasification is a high temperature process in which organic materials react with steam and limited amounts of oxygen to produce H_2, CO, and CO_2. Kondoh *et al*. (2001) compared gasification with other thermal methods for processing shredder residue. Some of processes that were tested for shredder residue are described in the following.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

VOEST-ALPINE Process. Tests were conducted in which the shredder residue was blended with mixed plastics, waste oils, and fuel oil (Schmitt, 1990). The product gas is considered a low-energy industrial fuel gas, which would have to be used on-site.

TwinRec Process. This process uses fluidized-bed gasification with ash melting (Selinger, Steiner, and Shin, 2003; Kummer, 2003; Ando *et al*., 2002). The process produces steam for power generation. It also produces ferrous and nonferrous metals. The fine inert residue is vitrified in the process. Zinc and fly ash from the process are sent to the zinc industry.

SVC Process. This process has recycled shredder residue for many years (Obermeier and Markowski, 2002). Tests were conducted using a ratio of 30% shredder residue/70% other solid and liquid wastes. Buttker *et al*. (2005) reported that a test was conducted using over 900 tons of shredder residue and at gasification/partial oxidation temperatures of 1300–1600°C and a pressure of 25 bars.

Plasma Processes. Plasma gasification (~5000–10,000°C) has been tested for shredder residue (Circeo, 2008; Leal-Quiros, 2004; Sawyer, 2009; Tellini *et al*., 2007; Westinghouse, 2002). The organic material is converted to a gas rich in CO and H_2. The inorganic residue is melted into a slag. Many companies have built and tested plasma systems for waste treatment.

PyroArc Technology. This technology incorporates a gasifier followed by plasma combustion of the gasifier off-gases. The temperature in the first stage is about 1000°C, and in the plasma reactor, the temperature is can be up to 5000°C (Gustavsson *et al*., 2005).

RESHMENT® Process. The material is size reduced to <5 mm and fed to the furnace, which is at ~2000°C. Oxygen is provided to assist in oxidizing the decomposition products of the organics. In the process, the metal oxides are reduced and the energy is used to produce electricity (Sauert *et al*., 2001, 2005).

Thermoselect Process. This process employs a fixed-bed oxygen-blown gasifier. It is used mostly for processing MSW (Stahlberg *et al*., 2002; also see Drost *et al*., 2004). A trial, using over 900 tons of shredder residue mixed with MSW, was conducted.

A plant for gasifying 350,000 tons of shredder residue to produce 40 MW of electricity is being developed by innovative environmental solutions (IESs)-a joint venture between European metal recycling (EMR) and Cranford, New Jersey (Waste Management World, 2012). The plant uses the "RODECS" recycling and gasification technology. The gasification process operates in the temperature range of 450–600°C. The generated gas is then heated and conditioned at temperature between 1200 and 1400°C to produce an "ultra-clean synthetic gas" (Waste Management World, 2012).

8 RECYCLING OF POLYMER COMPOSITES

8.1 Overview

Polymer composites can be made with a thermoset or thermoplastic polymer. The two major fiber composites used in the automotive industry are glass-fiber- and carbon-fiber-reinforced polymer composites. Carbon-fiber-reinforced polymer-matrix composite (PMC) has high strength-to-weight ratios. Therefore, they are ideally suited for use in the automotive applications. Because these composites are made to meet very strict properties, recycling of these materials and particularly their valuable carbon fibers should avoid significant degradation of their properties. Aggressive thermal or chemical treatment can degrade the properties of the fibers and destroy the polymer (Sunderland, 2001; Teodorescu, *et al*., 2008). Generally, the value of the polymer is small compared to the value of the carbon fibers. The recycling processes that can be used are:

- direct recycling methods,
- thermal pyrolysis to liberate the carbon fibers, and
- chemical or solvent methods to dissolve the thermoplastic resin and recover both the fibers and the resin.

One type for direct recycling involves shredding the carbon-fiber-reinforced PMC scrap to a fine powder and using the fine powder in the making of polymeric compounds such as sheet molding compounds.

Pyrolysis methods were developed by several organizations. ELG Carbon Fiber Limited acquired Recycled Carbon Fibre Ltd. (RCF) is claimed to be the first company worldwide to build and operate a commercial plant for recycling carbon-fiber-reinforced plastics (ELG, 2013). The value of the polymer in carbon-fiber composites is insignificant compared to the value of the fibers. Processes for recycling carbon fibers from composites have been developed. Some are discussed later.

Argonne National Laboratory developed a process that involves heating scrap under different environments to degrade the polymer. The process was tested in a batch oven and in a continuous thermal reactor (Jody, Pomykala, and Daniels, 2003a; Jody *et al*., 2004). Over 90% of the polymer can be removed in a few minutes. However, removing the remaining 1% or 2% requires additional treatment at different temperatures and oxygen environments. Samples of carbon fibers recovered from known "control" panels

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

were evaluated. The results showed that the carbon concentration on the surface of the recovered fibers is essentially the same as that for the treated virgin surface. The mechanical properties of the recovered fibers were also comparable to the mechanical properties of virgin fibers from the same lot. No significant change in the relative concentration of functional groups on the surface was observed.

Adherent Technologies Incorporated developed wet chemical processes that degrades the polymer and produces a "99+% clean fiber without degrading the fiber strength." Adherent Technologies has a pilot plant that can process about 45 kg of composite scrap per batch (Adherent, 2012, 2013). In this process, the PMC is broken down in a liquid under milder conditions than those used in pyrolysis to free the fibers. The process uses off-the-shelf chemical equipment (Adherent, 2013). It is also reported that the process is suitable for all fiber-reinforced composite materials. However, its economic competitiveness is established only when processing carbon-fiber-reinforced polymer composites.

Boeing and BMW reported that they have agreed to jointly work on developing process for recycling carbon fibers and to share information about carbon fiber materials and manufacturing (Boeing, 2012). It was also reported that the University of Leeds, United Kingdom, and their industrial partners have developed a combined pyrolysis/physical separation process for recycling PMC waste materials. In this process, the fibers are recovered and the polymers are converted to oil (m.com, 2013).

9 LANDFILLING OF THE NONRECYCLABLE MATERIALS

One hundred percent recycling of ELV polymer and composite materials is not technically or economically attainable. Some left over material will have to be landfilled. Worldwide regulations are calling for more recycling and less landfilling. In addition to the environmental issues associated with landfills, cost of landfilling is rising and will reduce the profitability of some metal recycling operations that result in large amounts of residue that has to be landfilled.

On the positive side, landfills contain huge amounts of polymers, metals, and other materials that could be recovered once cost-effective and environmentally friendly technology is available (American Recycler, 2008).

10 CONCLUSIONS

Shredder residue is a complex mixture of materials, and the cost-effective separation and recovery of materials and polymers from it has been a challenge. Nonetheless, through continued research, several technologies have reached advanced stages of development. The two areas that received the most attention are (i) recovery of the polymers for reuse as polymers and (ii) conversion to fuels and energy. Most of the technologies start by separating the polymers from shredder residue as a concentrate before technologies to separate individual polymers can be recovered. Processes to recycle polymer composites have been developed. These focus primarily on recovering the fibers, which are more valuable.

Many of the processes for recovering the polymers also recover the residual metals in shredder residue. The value of the residual metals improves the economics of the recycling process. Researchers have tested several separation technologies at a relatively large scale. Polymer separation technologies have successfully separated and recovered the polyolefins and engineered plastics (such as ABS) from the polymer concentrate. Dry and wet processes have also succeeded in separating and recovering a mixed-rubber fraction. Gasification and pyrolysis processes have proved that "fuels" can be produced from the automotive polymers.

Despite recent technical advancements, most of shredder residues are still disposed of in landfills or by incineration. Developing economic technology to remove the SOCs is another milestone that must be accomplished before technologies can be commercialized. An efficient and economical solution to recycling shredder residue is likely to be an integrated system of many technologies to produce quality products at the lowest cost.

ACKNOWLEDGMENTS

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a US Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The US Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

REFERENCES

Adherent (2012) http://www.adherent-tech.com/ (accessed 03 December 2013).

Adherent (2013) http://www.adherent-tech.com/recycling_technologies (accessed 03 December 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

American Plastics Council (1999) *Municipal Recycling*, http://www.plasticsresource.com (accessed 26 October 2013).

American Recycler (2008) High Commodity Prices Boost Landfill Mining Prospects, http://www.americanrecycler.com/0408/high.shtml (accessed 03 December 2013).

Ando, G., Steiner, Ch., Selinger, A., and Shin, K., *et al*. (2002) Automobile Shredder Residue Treatment in Japan—Experience of 95'000 t ASR Recycling and Recovery Available for Europe through TwinRec. *Proceedings of the International Automobile Recycling Congress*, Geneva, Switzerland, March 13–15.

Bain, D. (2012) *BASF Building Better Cars through Chemistry, Tourque News*, http://www.torquenews.com/397/basf-building-better-cars-through-chemistry (accessed 5 January 2012).

Baker, B.A., Woodruff, K.L., and Naporano, J.F. (1994) Automotive Shredder Residue (ASR) Separation and Recycling System. U.S. Patent 5443157.

Banks, M.E., Lusk, W.O., and Ottinger, R.S. (1971) New chemical concepts for utilization of waste plastics. U.S. Environmental Protection Agency Report SW-IGC, Washington, D.C.

Boeing (2012) Boeing, BMW to Partner on Carbon Fiber Recycling, Environmental Leader, http://www.environmentalleader.com/2012/12/13/boeing-bmw-to-partner-on-carbon-fiber-recycling/ (accessed 03 December 2013).

Brüggler, M. (2002) ASR Recycling in Europe. *Proceedings of the International Automobile Recycling Congress*, Geneva, Switzerland, March 13–15.

Buttker, B., Giering, R., Schlotter, U., *et al*. (2005) Full-scale industrial recovery trials of shredder residue in a high temperature slagging gasifier in Germany. Plastics Europe report # 8043/GB/06/2005.

Cambridge (2005) *The ImpEE Project, Recycling of Plastics, The Cambridge-MIT Institute, University of Cambridge, 2005*, http://www-g.eng.cam.ac.uk/impee/topics/RecyclePlastics/files/Recycling%20Plastic%20v3%20PDF.pdf (accessed 03 December 2013).

Chaala, A., Ciochina, O., and Roy, C. (1999) Vacuum pyrolysis of automotive shredder residue, use of the pyrolytic oil as a modifier for road bitumin. *Resources, Conservation and Recycling*, **26**, 155–172.

Circeo, L.J. (2008) *Plasma Arc Gasification of Municipal Solid Waste*, presentation by Georgia Tech, http://www.energy.ca.gov/proceedings/2008-ALT-1/documents/2009-02-17_workshop/presentations/Louis_Circeo-Georgia_Tech_Research_Institute.pdf.

Competitive Analysis Centre, Inc. and Economic Associates, Inc. (1997) Automotive Shredder Residue, Its Application in Steel Mill Blast Furnaces, prepared by Competitive Analysis Centre, Inc., and Economic Associates, Inc., for the American Plastics Council (APC) and the Environment and Plastics Industry Council (EPIC) of CPIA, October.

Competitive Analysis Centre, Inc. and Economic Associates Inc. (1998) Automotive Shredder Residue, Review of ASR—Blast Furnace Concept with European Interests, prepared by Competitive Analysis Centre, Inc., and Economic Associates, Inc., for the American Plastics Council (APC) and the Environment and Plastics Industry Council (EPIC) of CPIA, March.

Daniels, E.J., Jody, B.J., and Bonsignore, P.V. (1990) Automotive Shredder Residue: Process Development for Recovery of Recyclable Constituents. *Presented at and published in the Proceedings of the 6th International Conference on Solid Waste Management and Secondary Materials*, Philadelphia, PA, December.

Daniels, E.J. (1994) Separation and Recovery Process R&D to Enhance Automotive Materials Recycling. *Proceedings of the 3rd International Conference on Automobile Recycling, Auto Recycle Europe '94*, Milano, Italy, May 5–6.

Daniels, E.J., Carpenter, Jr., J.A., and Duranceau, C., *et al*. (2004) Sustainable end-of-life vehicle recycling: R&D collaboration between industry and the U.S. DOE *Journal of Metals*, **56** (8), 28–32.

Davis, S.C., Diegel, S.W., and Boundy, R.G. (2011) *Transportation Energy Data Book*, 30th edn, Oak Ridge National Laboratory, Oak Ridge, TN.

Day, M., Cooney, J.D., and Shen, Z. (1996) Pyrolysis of automobile shredder residue: an analysis of the products of a commercial screw kiln process. *Journal of Analytical and Applied Pyrolysis*, **37**, 49–67.

Decicions (2010) Commission Decisions of 23 February 2010 amending Annex II to Directive 2000/53/EC of the European Parliament and of the Council on end-of-life vehicles, (notified under document C(2010) 972), (2010/115/EU).

Drost, U., Eisenlohr, F., and Kaiser, B., *et al*. (2004) *Report on the Operating Trial with Automotive Shredder Residue (ASR)*, http://www.thermoselect.com/news/2004-03-10,%204th%20IARC,%20Geneva,%20THERMOSELECT2.pdf.

Duranceau, C. (2009) Picking up speed *Recycling Today*, 38–43. http://www.recyclingtoday.com/plastics_automotive_feature_december_2009.aspx (accessed 26 October 2013).

ELG (2013) http://www.jeccomposites.com/news/composites-news/elg-invests-recycling-carbon-fibre (accessed 03 December 2013).

Gallon, N. and Binder, M. (2006) Life cycle inventory (LCI) of Argonne's process for recycling shredder residue. Final report, PE Europe GMBH-Life Cycle Engineering, August.

Galloo (2012a) http://www.gallooplastics.eu/GB/galloogb.html (accessed 03 December 2013).

Galloo (2012b) *Dense Medium Separation: Plastics, ESR International*, http://www.esrint.com/pages/pages/plastic.html (accessed February 2013).

Guschall, H., *et al*. (2005) End-of-Life Vehicle Recycling: The Volkswagen-SiCon Process® for the Recycling of Shredder Residue. Presented at the 5th International Automobile Recycling Congress, Amsterdam, Netherlands, March 9–11.

Gustavsson, B., *et al*. (2005) The PyroArc® Process—An Economically and Environmentally Friendly Process for Treatment of Car Fluff. Presented at the 5th International Automobile Recycling Congress, Amsterdam, Netherlands, March 9–11.

Hubble, W.S., Most, I.G., and Wolman, M.R. (1987) Investigation of the energy value of automobile shredder residue. U.S. Department of Energy Report DOE/ID-12551, Washington, D.C.

InfiChem (2012) http://infichempolymers.com/applications.php (accessed 13 February 2012).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Jody, B.J., Daniels, E.J., Bonsignore, P.J., and Dudek, F.J., (1990) Recycling of Plastics in Automobile Shredder Residue. *Proceedings of the 25th Intersociety Energy Conversion Engineering Conference*, Vol. 5, Reno, NV.

Jody, B.J., Daniels, E.J., Bonsignore, P.J., and Dudek, F.J., (1994) Recovering recyclable materials from shredder residue *Journal of Metals*, **46** (2), 40–43.

Jody, B.J., Daniels, E.J., and Pomykala, Jr., J.A. (1996) Progress in Recycling of Automobile Shredder Residue. *Proceedings of the Extraction and Processing Division*, TMS Annual Meeting, Anaheim, CA.

Jody, B.J., Daniels, E.J. and Bonsignore, P.V. (2001) Process to recycle shredder residue. U.S. Patent Number 6,329, 436 B1.

Jody, B.J., Pomykala, Jr., J.A., and Daniels, E.J. (2003a) A process to recover carbon fibers from polymer matrix composites. SAE 2002 Transactions—Journal of Materials & Manufacturing, SAE Paper 2002-01-1967, ISBN Number 0-7680-1289-9.

Jody, B.J., Pomykala, Jr., J.A., and Daniels, E.J. (2003b) Process for the recovery and separation of plastics. U.S. Patent 6,599,950.

Jody, B.J., Pomykala, J.A., Jr, Daniels, E.J., and Greminger, J.L. (2004) A process to recover carbon fibers from polymer-matrix composites in end-of-life vehicles *Journal of Metals (JOM)*, **56** (8), 43–47.

Jody, B.J., Daniels, E.J., and Duranceau, C.M., *et al*. (2010) End-of-life vehicle recycling: state of the art in resource recovery from shredder residue. Argonne National Laboratory Report ANL/ESD/10-8.

Jody, B.J., Spangenberger, J.S., and Daniels, E.J., *et al*. (2011) Process and apparatus for separating solid mixtures. U.S. Patent Number 7,954, 642 B2.

Keller, C. (1999a) The Swiss Way of Handling Plastics in Cars, Objectives, Concepts and Recent Developments. *Proceedings of the Identiplast Congress, APME Publications*, April 26–28.

Keller, C. (1999b) Thermal Treatment of Auto Shredder Residue (ASR) under Reducing Conditions, Experiences on a Technical and Laboratory Scale, Vol. II, pp. 187–192. A. Barrage and X. Edelmann *Proceedings of the R'99 Recovery Recycling Re-integration*, Geneva, Switzerland, February, www.environmental-expert.com/events/r2000/r2000.htm (accessed 26 October 2013).

Keller, C. (2003) Optimized disposal of automotive shredder residue in *Municipal Solid Waste Management—Strategies and Technologies for Sustainable Solutions* (eds C. Ludwig and S. Hellweg), Springer-Verlag, Heidelberg, Germany, pp. 294–307.

Kondoh, M., *et al*. (2001) Study of gasification characteristics of automobile shredder residue. *JSAE Review*, **22**, 234–236.

Konz, R.J. (2009) The End-of-Life Vehicle (ELV) Directive: The Road to Responsible Disposal. 18 Minn. J. Int'l L. 431 (2009). http://www.law.umn.edu/uploads/BX/fw/BXfwZTM0VoxN2BtOQ7E2Vg/Konz-Final-Online-PDF-03.30.09.pdf (accessed 26 October 2013).

Kummer, B. (2003) Problems with Automobile Shredder Residue (ASR) and Solutions for ASR. *Presented at the International Automobile Recycling Congress*, Geneva, Switzerland, March 12–14.

Leal-Quiros, E. (2004) Plasma processing of municipal solid waste *Brazilian Journal of Physics*, **34** (4B), 1587–1593.

Leidner, J. (1981) *Plastics Waste; Recovery of Economic Value*, Marcel Dekker, Inc., New York.

Lundqvist, U., *et al*., 2004, Design for Recycling in the Transport Sector—Future Scenarios and Challenges, Department of Physical Resource Theory, Chalmers University of Technology, Goteborg University, Goteborg, Sweden.

Mark, F.E. and Fisher, M.M. (1998) *Energy Recovery from Automobile Shredder Residue through Co-combustion with Municipal Solid Waste*, Report 8026, APME Publications, Brussels, Belgium.

Mark, F.E. and Fisher, M.M. (1999) Energy Recovery from Automobile Shredder Residue through Co-combustion with Municipal Solid Waste, Vol. II, pp. 46–53. A. Barrage and X. Edelmann *Proceedings of the R'1999 Recovery Recycling Re-Integration*, Geneva, Switzerland.

MBA (2012) http://www.mbapolymers.com/home/mba-polymers-uk-ltd (accessed 24 January 2012).

McDermott, L.P., Jenner, R.K., and Crocombe, R.A. (2007) Identification of Recyclable Polymers with a Handheld Near-Infrared Spectrometer, http://www.americanlaboratory.com/914-Application-Notes/35120-Identification-of-Recyclable-Polymers-with-a-Handheld-Near-Infrared-Spectrometer/ (accessed 03 December 2013), American Laboratory.

MSS (2012) http://www.magsep.com/ (accessed 26 October 2013).

m.com (2013) New Technology Paves the Way for Recycling Polymer Matrix Composites, m.com February 11, 2013. http://www.azom.com/article.aspx?ArticleID=1987 (accessed 03 December 2013).

Nauman, B.E. and Lynch, J.C. (1993) Polymer recycling by selective dissolution. U.S. Patent 5,198,471.

Nauman, B.E. and Lynch, J.C. (1994) Polymer recycling by selective dissolution. U.S. Patent 5,278,282.

NRT (2012) http://epa.gov/ncer/sbir/success/pdf/national_recovery_success.pdf (accessed 03 December 2013).

ORNL (1994) Recent Trends in Automotive Recycling: An Energy and Economic Assessment. T.R. Curlee, S. Das, C.G. Rizy, and S.M. Schexnayder, Oak Ridge National Laboratory report ORNL/TM-12628.

Obermeier, T. and Markowski, J. (2002) Gasification of shredder residue at SVZ Schwarze Pumpe. *Proceedings of the International Automobile Recycling Congress*, Geneva, Switzerland, March 13–15.

Parliament (2010) Directorate General for Internal Policies-Policy Department A: Economic and Scientific Policy Environment, Public Health and Food Safety. End of life vehicles: legal aspects, national practices and recommendations for future successful approach, IP/A/ENVI/ST/2010-07 October 2010, PE 447.507.

Pineau, J.L., Kanari, N., and Menad, M. (2005) Representativeness of an automobile shredder residue sample for a verification analysis *International Journal of Integrated Waste Management, Science and Technology*, **25** (7), 737–746.

PIA (1980) *Maximizing the Life Cycle of Plastics, Final Report*, Plastics Institute of America, Inc., Lowell, MA.

Plastemart (2011) *Polymer Composites Have Great Potential in Hybrid Vehicles—A Slow Growing Market*, Plastemart.com, http://www.plastemart.com/Plastic-Technical-Article.asp?LiteratureID=1634&Paper=polymer-composites-have-great-potential-in-hybrid-vehicles-a-slow-growing-market (accessed 20 December 2011).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Plastics News Magazine (2011) October–December issues.

Rasshofer, W. and Schomer, D. (2003) European and national ELV regulations of car recovery, their impact on polyurethane applications in the automotive industry, and a proposal by the German plastics industry to solve the plastics recovery issue. SAE Paper 2003-01-0644.

Recycling International (2008) *BST Treats Shredder Residue with SiCon Plant*, Recycling International, April 25, 2008 by Editorial Staff, http://www.recyclinginternational.com/recycling-news/4896/other-news/archiv/bst-treats-shredder-residue-sicon-plant (accessed 03 December 2013).

Recycling International (2012) *Breakthrough Moment Features—Processing Point, Auto Shredding*, Brian Taylor, May 15, 2012, http://www.recyclingtoday.com/rtge0512-auto-shredder-residue-debate.aspx (accessed 03 December 2013).

Reuter, M.A., *et al*. (2004) The optimization of recycling: integrating the resource, technological, and life cycles *Journal of Metals*, **56** (8), 33–37.

Roy, C. and Chaala, A. (2001) Vacuum pyrolysis of automotive shredder residue *Resources, Conservation and Recycling*, **32**, 1–27.

Sander, K., Schilling, S., Zangl, S., and Lohse, J. (2002) *Rules on compliance with Article 7.2 of Directive 2000/53/EC, Final Report, September 2002, Report compiled for the Directorate General Environment, Nuclear Safety and Civil Protection of the Commission of the European Communities under Contract No B4-3040/2002/335823/MAR/A2*, http://ec.europa.eu/environment/waste/studies/elv/compliance_art7_2.pdf (accessed 26 October 2013).

Sarnafil (2011) *Sika, Solutions and Products*, http://gbr.sika.com/en/solutions_products/roofing_services/sarnafil/performance/environmental_assessment/embodied_energy.html (accessed 12 December 2011).

Sattler, H.P. and Laage, B. (2000) ASR—From Waste to Products. *Proceedings of the R'2000 Recovery, Recycling, Reintegration Conference*, Toronto, Canada.

Sauert, F., *et al*. (2001) RESHMENT®—an ASR process for maximized recycling, reuse, and recovery. SAE Paper 2001-01-3757.

Sauert, F., *et al*. (2005) RESHMENT®—The Shredder Residue Solution for Switzerland. *Presented at the 5th International Recycling Congress*, Amsterdam, Netherlands, March 9–11.

Sawyer, A. (2009) Plasma gasification system. Patent application No. 20090133407, May 28.

van Schaik, A. and Reuter, M.A. (2004) The optimization of end-of-life vehicle recycling in the European Union. *Journal of Metals*, **56** (8), 39–42.

Schmitt, R.J. (1990) Automobile shredder residue: the problem and potential solutions. Center for Metals Production (CMP), CMP Report No. 90–1, Carnegie Mellon Research Institute, Pittsburgh, PA.

Selinger, A., Steiner, C., and Shin, K. (2003) TwinRec—Bridging the Gap of Car Recycling in Europe. *Presented at the International Automotive Recycling Congress*, Geneva, Switzerland, March 12–14.

Shutov, F. (1999) Effective energy and gas emission savings using plastics waste recycling technologies. Report, Expert Group Meeting on Industrial Energy Efficiency, Cogeneration and Climate Change, Environmental Sustainability Division, Kyoto Protocol Branch, United Nation Development Organization (UNIDO), prepared in cooperation with The International Cogeneration Alliance (ICA) and The International Institute of Energy Conservation (IIEC), http://www.unido.org/fileadmin/import/userfiles/ploutakm/shutov.pdf (accessed 26 October 2013).

Stahlberg, R., *et al*. (2002) The THERMOSELECT High Temperature Recycling Technology for Automotive Shredder Residue—Results and Perspectives. *Proceedings of the International Automobile Recycling Congress*, Geneva, Switzerland, March 13–15.

Sunderland, P. (2001) Recycling of Polymer Matrix Composites, Encyclopedia of Materials: Science and Technology, 2001, Sunderland, pp. 7396–7399.

Teijin (2012) http://www.teijin.co.jp/english/news/2011/ebd110309.html (accessed 6 February 2012).

Tellini, M.G., *et al*. (2007) Automobile shredder residue (ASR) destruction in a plasma gasification reactor, *International Journal of Environmental Technology and Management*, **7**,(1/2), 21–38.

Florin, T., *et al*. (2008) *On the recycling of carbon fibers reinforced polymer matrix composites*. (eds F. Teodorescu, H. Teodorescu, G. Stanca, *et al*.,) 4th edn IASME/WSEAS International Conference on Energy, Environment, Ecosystems and Sustainability (EEESD'08), Algarve, Portugal, June 11–13, 2008.

Tesoro, G. (1987) Recycling of Synthetic Polymers for Energy Conservation—The State-of-the Art, *Polymer News* 12.

Titech (2012a) http://www.titech.com/news/titech-technology-first-20772 (accessed 03 December 2013).

Titech (2012b) http://www.titech.com/sorting-equipment/titech-autosort-4-21677 (accessed 03 December 2013).

USGS (2006a) Science for a changing world. Fact Sheet 2005–3144, U.S. Geological Survey.

USGS (2006b) Science for a changing world. Fact Sheet 2005–3145, U.S. Geological Survey.

Waste Management World (2012) *40 MW Gasification Plant to Recycle ELV Shredder Fluff in Midlands*, July 16, 2012, http://www.waste-management-world.com/articles/2012/07/40-mw-gasification-plant-to-recycle-elv-shredder-fluff-in-midlands.html (accessed 03 December 2013).

Westinghouse (2002) Westinghouse Plasma Coal Gasification & Vitrification Technology. *Presented to the Electric Power Generation Association, Power Generation Conference*, Hershey, PA, October 16–17, http://www.epga.org/2002conference/Westinghouse.pdf (accessed 26 October 2013).

Wielgat, A. (2002) *Recycling Process Cuts Shredder Residue Sent to Landfills*, http://findarticles.com/p/articles/mi_m3012/is_5_182/ai_87105902/ (accessed 13 February 2012), RPI.

Zevenhoven, R. and Saeed, L. (2003) Automotive shredder residue (ASR) and compact disc (CD) waste: options for recovery of materials and energy. Final Report #TKK-ENY-14, Helsinki University of Technology, Helsinki, Finland.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Please note that the abstract and keywords will not be included in the printed book, but are required for the online presentation of this book which will be published on Wiley Online Library (http://onlinelibrary.wiley.com/). If the abstract and keywords are not present below, please take this opportunity to add them now.

The abstract should be a short paragraph of between 150–200 words in length and there should be 5 to 10 keywords

Abstract: Each year, more than 25 million vehicles reach the end of their service life throughout the world, and this number is rising as the number of vehicles on the road is increasing. In the United States, over 95% of the more than 10 million vehicles scrapped annually enter a comprehensive recycling infrastructure that includes auto parts recyclers (dismantlers), remanufacturers, and material recyclers (shredders). Over 75% of the automotive materials, primarily metals, are profitably recycled. Automobiles are the largest source of recycled ferrous scrap for the iron and steel industry. The scrap processors recover metal scrap from automobiles by shredding the obsolete automobile hulks, along with other obsolete metal-containing products and recovering the metals from the shredded material. The nonmetallic fraction that remains—commonly called *shredder residue*—constitutes about 25% of the weight of the vehicle, and most of it is landfilled.

In the past two decades, research and development has been undertaken to enhance the recycle rate of end-of-life vehicles (ELVs), including improving dismantling techniques and remanufacturing operations. However, most of the effort has been focused on developing technology to separate and recover the polymers from shredder residue. To make future vehicles more energy efficient and to meet the new Corporate Average Fuel Economy (CAFE) regulations, more lightweighting materials—primarily polymers and polymer composites, high strength steels, and aluminum—are being used by the automobile industry. The nonmetallic materials increase the percentage of shredder residue that must be disposed of, compared with the percentage of metals that are recovered. Therefore, recycling of automotive polymers will be more important in the future. In addition, recycling the polymers and the composites has the following additional benefits: (i) conserves valuable resources of materials and energy, including the energy consumed in manufacturing automotive materials; (ii) provides lower cost materials to industry; (iii) reduces the amount of waste going to the landfill; and (iv) reduces greenhouse gas (GHG) emissions.

The future will also bring an increase in hybrid and electric vehicles. This will introduce new materials for disposal, including batteries. Recycling the new materials presents technical and economic challenges to the existing automotive recycling infrastructure. New technologies will be required to sustain and maximize the ultimate recycling of the vehicles.

Keywords: polymers; composites; recycling; separation; technology

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is a US government work and is in the public domain in the United States of America. Copyright © 2014 John Wiley & Sons, Ltd. in the rest of the world.
DOI: 10.1002/9781118354179.auto164
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Automotive Environmental Life Cycle Assessment

Lynette W. Cheah

Singapore University of Technology and Design, Singapore

1 Introduction 1
2 An Automobile's Life Cycle Stages 2
3 Life Cycle Assessment (LCA) Steps 4
4 Automotive LCA Results 8
5 Summary and Discussion 10
Appendix A 11
Related Articles 12
References 12

1 INTRODUCTION

The automotive industry is one of the world's largest manufacturing sectors, whose products are widely pervasive in our daily lives. The world vehicle population topped one billion in 2010 (Sousanis, 2011), and is expected to continue growing. Cars and trucks enable personal mobility and provide economic opportunity for billions of people. However, they also contribute to complex urban planning, environmental, and health challenges. The production of motor vehicles requires significant amounts of energy and material resources. Their popular use leads to traffic congestion, urban air and noise pollution, global warming, and increases the demand for oil. Worldwide, automobiles are estimated to be responsible for around 20% of carbon dioxide releases, a greenhouse gas that contributes to global warming, and 60% of the world's oil consumption (Figure 1) (IEA, 2010).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

The notion of sustainability is undoubtedly an important consideration for the automotive industry. Automakers are facing pressures in the form of rising material and oil prices, more stringent air emissions regulation, and a push for greater fuel efficiency in vehicles. Addressing these challenges will require new vehicle designs and technologies. One useful tool or approach to evaluate the environmental impacts of these new innovations is life cycle assessment (LCA).

LCA is a systematic environmental accounting methodology, which helps quantify and evaluate the environmental impacts associated with a specific service, manufacturing process, or product over its entire life cycle. For a product such as an automobile, the assessment considers its complete life cycle, from the extraction of raw materials needed until the point at which all residuals are returned to the environment. The environmental impacts are considered over each of the product's life cycle phases, or from "cradle to grave".

An assessment of an automobile's environmental performance should consider the impact over its entire life cycle. This accounts for not only the effects of operating an automobile but also other upstream and downstream effects in its life cycle. Adopting such a life cycle perspective is more holistic and can assist in optimizing the environmental performance of an automobile. It can also be used to compare different vehicle technologies and determine the more environmentally favorable alternatives.

In this chapter, we will first review the key life cycle phases of an automobile as shown in Figure 2. We will then elaborate on the steps needed to carry out a full automotive LCA. The next section will report the results of the LCA, before we arrive at the key conclusions on the life cycle impact of automobiles.



Figure 1. Urban traffic congestion in Beijing, China. (Reproduced from Wikimedia, 2010. © Australian cowboy/Wikimedia Commons/Public Domain.)

2 AN AUTOMOBILE'S LIFE CYCLE STAGES

2.1 Raw material acquisition and processing

Let us begin by exploring each of an automobile's life cycle stages in further detail. To begin to build an automobile, mineral resources are first mined from the earth and processed into usable materials that are needed. Current automobiles are made of a mix of materials, mostly metal (around 75% by mass).

Table 1 shows the material breakdown of a typical compact gasoline car, which weighs around 1230 kg (scaled down from data reported in (Ward's Communications, 2010)). Dominant metals used are steel, iron, and aluminum. Other key automotive materials include plastics and polymer composites, rubber, and glass.

2.2 Component manufacturing

In the next stage of an automobile's life cycle, automotive parts are fabricated using semifinished materials and energy resources. For metallic components, forming operations such as stamping, extruding, casting, or machining are applied to derive the final shape and properties. For example, the engine block in an automobile is usually sand-casted out of iron or aluminum alloys. Plastic components, such as the automobile's instrument panel and interior trim panels, are molded (Figure 3).

2.3 Vehicle assembly

Next, automotive components are put together within an assembly plant to make a finished car. This stage also

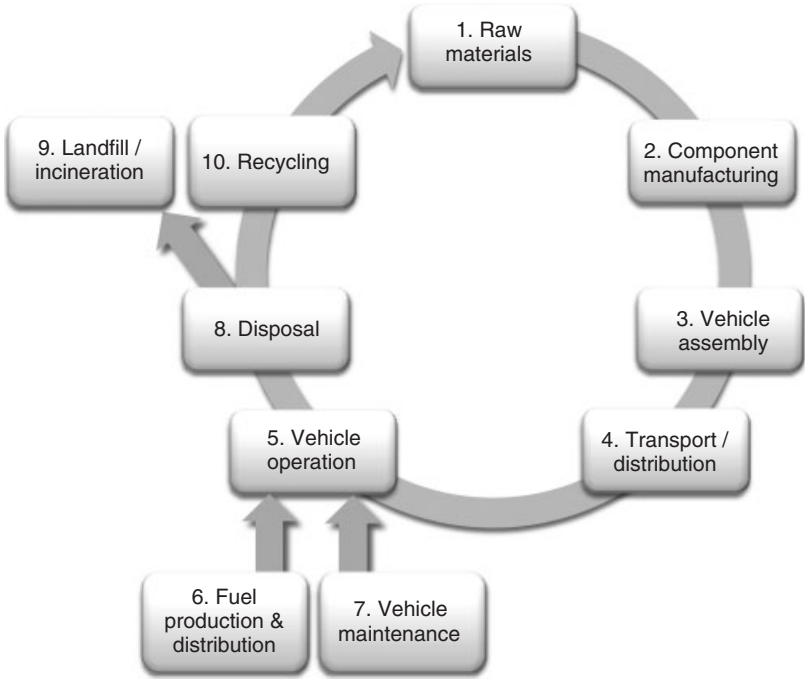


Figure 2. Automotive life cycle phases or stages.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Material breakdown of an average compact gasoline car.

Material	Mass (kg)	Mass (%)
Regular steel	501	40.6
High strength steel	175	14.2
Other steels	33	2.7
Iron	69	5.6
Aluminum	108	8.8
Magnesium	4	0.3
Copper and brass	21	1.7
Lead	15	1.2
Zinc	3	0.2
Powder metal parts	14	1.1
Other metals	2	0.1
Plastics and composites	128	10.4
Rubber	71	5.7
Coatings	11	0.9
Textiles	18	1.4
Glass	31	2.5
Other materials	30	2.4
Total	1235	100.0

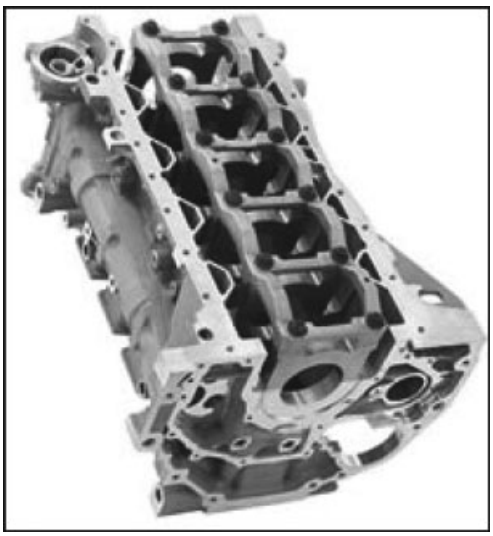


Figure 3. An aluminum engine block.

includes the processes of body welding and painting. Figure 4 portrays the set of activities that are covered during this assembly stage as well as the manufacturing stage of the automobile's life cycle.

2.4 Transportation/distribution

Transportation involves the movement of parts, components, subassemblies, or complete automobiles in between each of the automobile's life cycle stages. There are transport needs to supply the thousands of automotive parts and

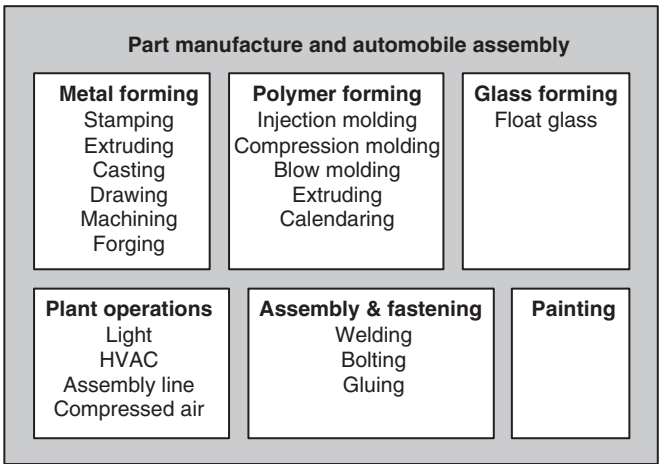


Figure 4. Activities in automobile's manufacture and assembly stages. (Reproduced from Sullivan, Burnham and Wang, 2010).

components to automotive manufacturers. Newly built automobiles are transported by rail, freight truck, or shipped to distributors, and are then forwarded to auto dealerships in order to reach end consumers. After the automobile's use phase, retired automobiles are eventually hauled to the scrappage and recycling facility (junkyard).

2.5 Use or operation

During its operation, an automobile is driven and refueled by the user to provide mobility. The total distance driven over its operation stage varies depending on the user. In the United States, this lifetime distance is estimated to be around 245,000 km (NHTSA, 2006) and has increased over the years. A typical automobile has a long service lifespan of around 10–15 years.

2.6 Fuel production and distribution

This stage of the life cycle relates to the production and supply of fuel needed over the automobile's use phase. It includes feedstock recovery, processing, storage, and transportation, followed by fuel production, transportation, storage, and distribution. The most common feedstock is crude oil (petroleum) and it is used to produce gasoline (petrol) and diesel fuels, among others. Alternative transportation fuels include compressed natural gas (CNG) and biofuels. Electricity is also an option for electric vehicles (EVs). For EVs, this stage will include the upstream activities of generating, transmitting, and distributing electricity used to charge them. Figure 5 shows various feedstock and fuel pathways that may be used to derive transportation fuels.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

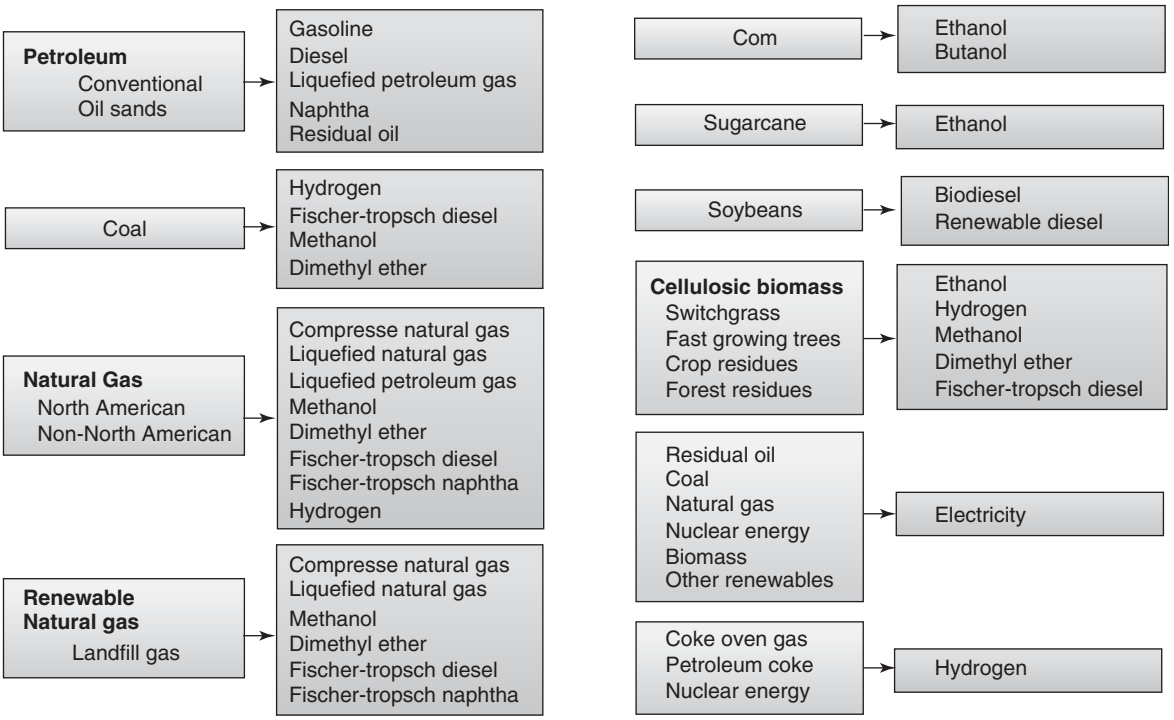


Figure 5. Possible fuel pathways, indicating fuels and electricity produced from various energy feedstock sources. (Reproduced from UChicago Argonne LLC, 2011.)

2.7 Vehicle maintenance

While the automobile is in use, maintenance processes and replacement parts are required to keep the automobile in safe, running condition and to maintain operability. These include replacing spark plugs, windshield wipers, tires, engine oil, and brake fluids, among others.

2.8 Disposal (end-of-life)

When the automobile arrives at the end of its service life, it is sent to a dismantler to remove parts that may be reused. Operating fluids and the battery are removed for recycling or proper disposal. The remaining automobile hulk, which refers to the remains of the end-of-life vehicle that is inoperable, is shredded, and metals are separated out and recycled. The remaining mix of materials is called *automotive shredder residue* (*ASR*), which is disposed of by way of landfill or incineration. Most of the automobile, being metallic, ends up being recycled. In Europe, there is an end-of-life vehicles (ELV) directive that targets 85% of an automobile's mass to be reused or recycled by year 2015.

3 LIFE CYCLE ASSESSMENT (LCA) STEPS

With an understanding of the automobile's life cycle stages, one can now begin to analyze its overall environmental impact. The International Organization for Standardization (ISO), a worldwide federation of national standards bodies, has standardized the evaluation framework for carrying out a LCA study under the series ISO 14040. According to this standard, an LCA involves four general interlinked steps:

1. Goal and scope definition;
2. Life cycle inventory (LCI) analysis;
3. Life cycle impact assessment; and
4. Interpreting the results.

Let us look into each of these steps for an automotive LCA (Figure 6).

3.1 Goal and scope definition

To begin, the LCA's purpose or study objectives should first be articulated. The project scope and system boundary are to be defined. This involves specifying the geographical and temporal scope of the project—the location and timeframe

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

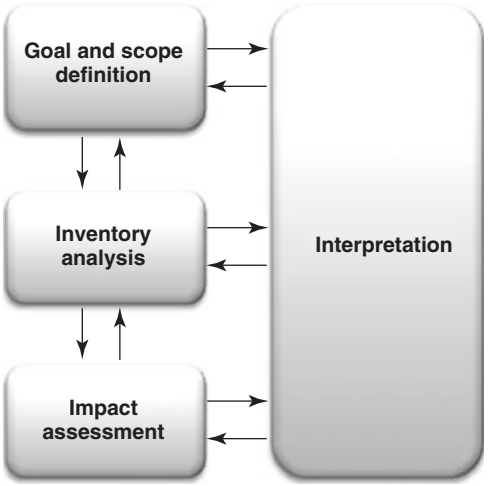


Figure 6. General life cycle assessment steps. (Reproduced from ISO, 2006. Permission to reproduce extracts from British Standards is granted by the British Standards Institution (BSI). No other use of this material is permitted.)

Table 2. Various boundaries for automotive LCA studies.

LCA Scope	Boundary Definition
Cradle to grave	Full life cycle assessment, includes all phases of the automobile's life cycle (#1–10 in Figure 2)
Cradle to gate	Partial life cycle assessment from resource extraction until the car factory gate, before it is transported to customers (#1–3 only)
Well to tank	Upstream impacts of fuel production and distribution (#6 only)
Tank to wheel	Impacts associated with driving the automobile only (#5 only)
Well to wheel	Also called the *fuel cycle*. Includes "well-to-tank" as well as "tank-to-wheel" impacts (#5 and #6)

of the analysis—as well as the system boundary, or consideration of what is or is not included in the study. Some common boundaries for assessing automobiles are shown in Table 2.

The functional unit for the LCA study should also be specified. The functional unit is the quantified performance of the automobile which is used as a reference unit in the study. One typical functional unit in an automotive LCA study would be distance traveled, say, in kilometers. So, the environmental impacts of the automobile are expressed per kilometer of travel. Other possible functional units could be simply impact per single automobile, or impact for an entire automobile fleet.

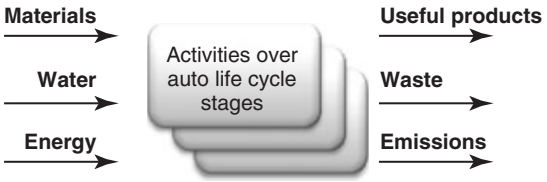


Figure 7. Life cycle inventory (inputs and outputs).

3.2 Automotive life cycle inventory (LCI)

The next step of an LCA is to compile a comprehensive inventory of the energy and raw materials inputs, as well as environmental releases (outputs or burdens) to the air, land, and water throughout the vehicle's life cycle phases of interest. All energy and material input and output flows of interest, based on the predefined study scope, are to be identified and quantified. The LCI is akin to an environmental balance sheet (Figure 7).

To illustrate the compilation of LCI, Table 3 shows the energy and emissions inventory for making stamped steel sheets that are used in automobiles (Burnham, Wang, and Wu, 2006). This inventory arises from steelmaking, as well as component fabrication. The LCI data are usually expressed per unit mass of material, in this case, per kilogram of steel product. Similar information is to be collected and organized for all other components within the automobile.

During the automobile's use phase, one aspect of the LCI is associated with the production of fuel. This "well-to-tank" (WTT) LCI depends on the fuel type and pathway

Table 3. Energy and emissions related to stamped steel production.

Life Cycle Inventory of Steelmaking	
Energy use: MJ per kg of steel product	
Total energy	37.41
Fossil fuels	35.65
Coal	15.20
Natural gas	19.51
Petroleum	0.94
Total emissions: grams per kg of steel product	
VOC	0.33
CO	39.92
NO*x*	2.85
PM10	6.10
PM2.5	2.47
SO*x*	2.73
CH_4	5.15
N_2O	0.03
CO_2	3005
CO_2-eq (VOC, CO, CO_2)	3050
Greenhouse gases	3188

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

taken to get from raw source (e.g., crude oil for fossil fuels) to the refueling station. Figure 8 shows the WTT greenhouse gas (GHG) intensity for different types of fuels that may be used in California (California Air Resource Board, 2009). This is expressed in grams of carbon-dioxide-equivalent GHG released during the production of fuel needed to deliver one megajoule of energy (gCO_2-eq/MJ). In the instance of gasoline, this includes the GHG emissions that occur over the processes of recovering and transporting crude oil to a typical refinery in California, gasoline storage, and transportation.

Over the automobile's long service life, the use of fuel forms a significant part of the automobile's LCI. One simple way of estimating the amount of fuel used is by taking the product of the automobile's average fuel consumption, usually expressed in liters per 100 kilometers (L/100 km), and the distance traveled over its service life. Another common reciprocal metric used to rate the fuel efficiency of the automobile is its fuel economy, expressed in mileage achieved per gallon of fuel (miles per gallon, or MPG). Depending on the type of automobile—size, segment, powertrain—and how it is driven, its fuel consumption would vary. Figure 9 shows the ranges of rated fuel consumption (in gasoline-equivalent liters) for various types of automobiles available in the US market, based on a representative drive cycle that includes a combination of urban and highway driving. In general, larger and heavier automobiles consume more fuel. Automobiles that run on

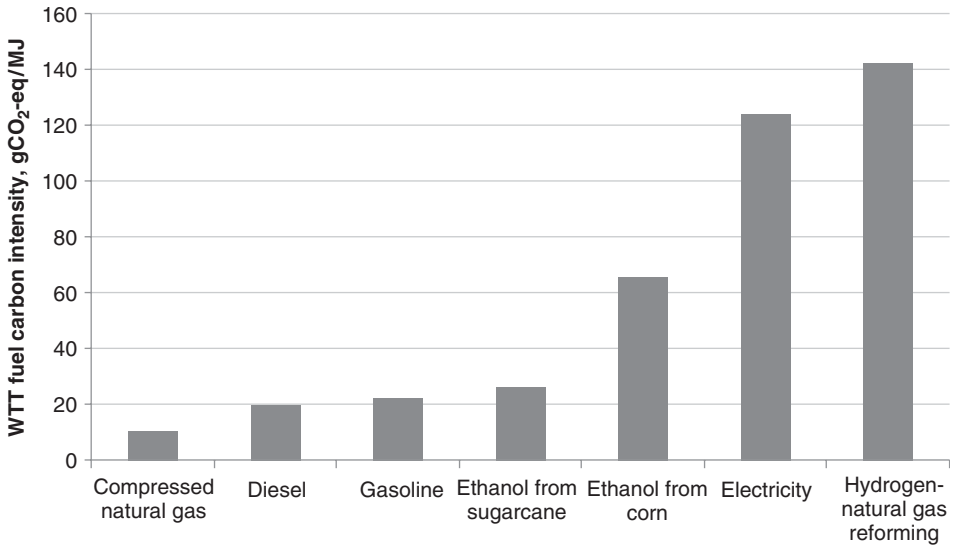


Figure 8. Well-to-tank greenhouse gas intensity of transportation fuels used in California. (Data source: CARB, 2009.)

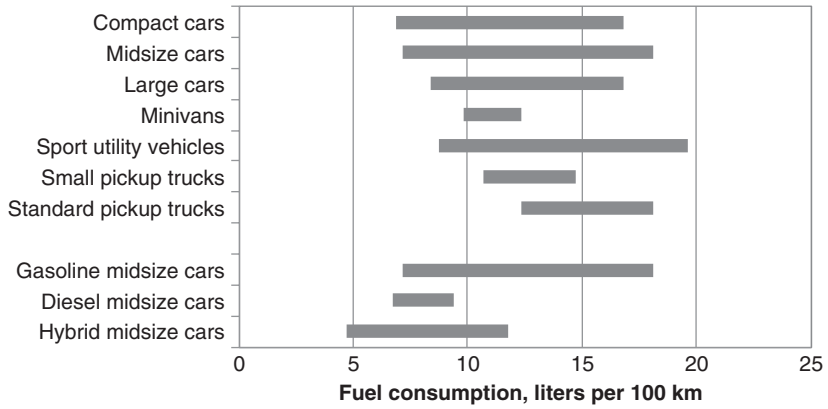


Figure 9. Ranges of the rated fuel consumption for model year 2012 vehicles sold in the United States. (Based on data from EPA, 2013.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

diesel would consume 20–25% less fuel than comparable gasoline models. Hybrid electric vehicles, which employ a battery in addition to an internal combustion engine to provide motive force, also consume less fuel. Assumptions made on the automobile's fuel consumption and the total distance traveled over the automobile's lifetime should be explicitly stated in the LCA study.

During the automobile's use phase, other items that contribute to the LCI are fuel and air-conditioning refrigerant losses due to evaporative leakage, as well as replacement parts and fluids that are needed to keep the automobile in running condition. The intensity of resource or material use for automobile maintenance depends very much on the individual customer, and data logging must be carried out, or assumptions made to determine the resulting inventory.

Table 4 below shows an estimated inventory of consumables needed to maintain an automobile over its useful life.

As one may perceive from the above illustrations, the collection of accurate and precise LCI data can be quite onerous. In lieu of collecting data at the source for each automobile, which can be resource- and time-intensive, one can turn to LCI databases. There are several LCI databases available that contain information on a large set of materials and processes. These databases provide proxy data that can help streamline the LCA effort.

For example, the Swiss Centre for Life Cycle Inventories has developed a comprehensive LCI dataset called *ecoinvent* for use in any LCA study. In the United States, the National Renewable Energy Laboratory (NREL) leads an effort to maintain a US LCI database for producing materials, components, or assemblies in the United States. The US Argonne National Laboratory has also made available another Greenhouse Gases, Regulated Emissions, and Energy Use in Transportation (GREET) model, which was specifically developed to assess automobiles. The GREET model includes the energy and emissions inventory for producing various components used in automobiles, which was derived from the literature. It also includes the effects of automobile assembly and eventual disposal or recycling.

There are, however, inherent challenges with using LCI databases. The geographical and temporal context or scope of the study might differ from that compiled in the database. Uncertainties in LCI data also exist: precise information on material use in automobiles is lacking, given the large number (thousands) of automotive components and suppliers. The use of materials in automobiles also differs across models and is constantly evolving. An acknowledgment of LCI data uncertainties in any automotive LCA study is therefore important. Part of this LCA step includes referencing the use of databases, if any, and clearly stating all assumptions made in estimating the inventory.

Table 4. Automobile maintenance inventory—estimated replacement parts.

Replacement Part or Fluid	Quantity	Unit
Brake fluid	3	Liters
Engine coolant fluid	22.2	Liters
Engine oil	78.1	Liters
Transaxle fluid	28	Liters
Windshield cleaner fluid	44	Liters
Air filter	4.3	Pieces
Battery	1.7	Pieces
Brake pads	1	Sets
Drive belt	2	Pieces
Lamp bulbs	3.5	Pieces
Exhaust system	1	Sets
Oil filter	15.7	Pieces
PCV valve	2	Pieces
Shock absorbers	1	Sets
Spark plugs	16	Pieces
Tires	2	Sets
Transaxle fluid filter	1	Pieces
Windshield	1	Pieces
Windshield wiper blades	18.7	Pieces

From Sullivan, *et al.* 1998. Copyright © 1998 SAE International. Reprinted with permission.

3.3 Life cycle impact assessment (LCIA)

After the automobile's life cycle inventory is compiled and documented, the environmental burdens and/or human health impacts associated with the resource inputs and environmental releases are assessed. On the basis of the study scope, the impact categories should have already been defined. These could be single metrics such as the automobile's net life cycle energy use, or emission amounts. Alternatively, quantitative results of the auto LCA may be presented as aggregated values within environmental themes, also known as *categories*. Examples of these categories are the automobile's impact on human toxicity, terrestrial toxicity, ozone depletion, aquatic eutrophication, aquatic acidification, or global warming. There are also various impact indicators that attempt to represent the automobile's impacts in composite scores, such as Eco-indicator 99, or Impact 2002+.

Impact 2002+, for instance, is a LCIA methodology that interprets life cycle inventory results to four "damage" categories—human health, ecosystem quality, climate change, and resources (Jolliet *et al*. 2003). The overall framework for Impact 2002+ is shown in Figure 10, which links LCI results via "midpoint" categories to "damage" categories. The damage units are also indicated in this figure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

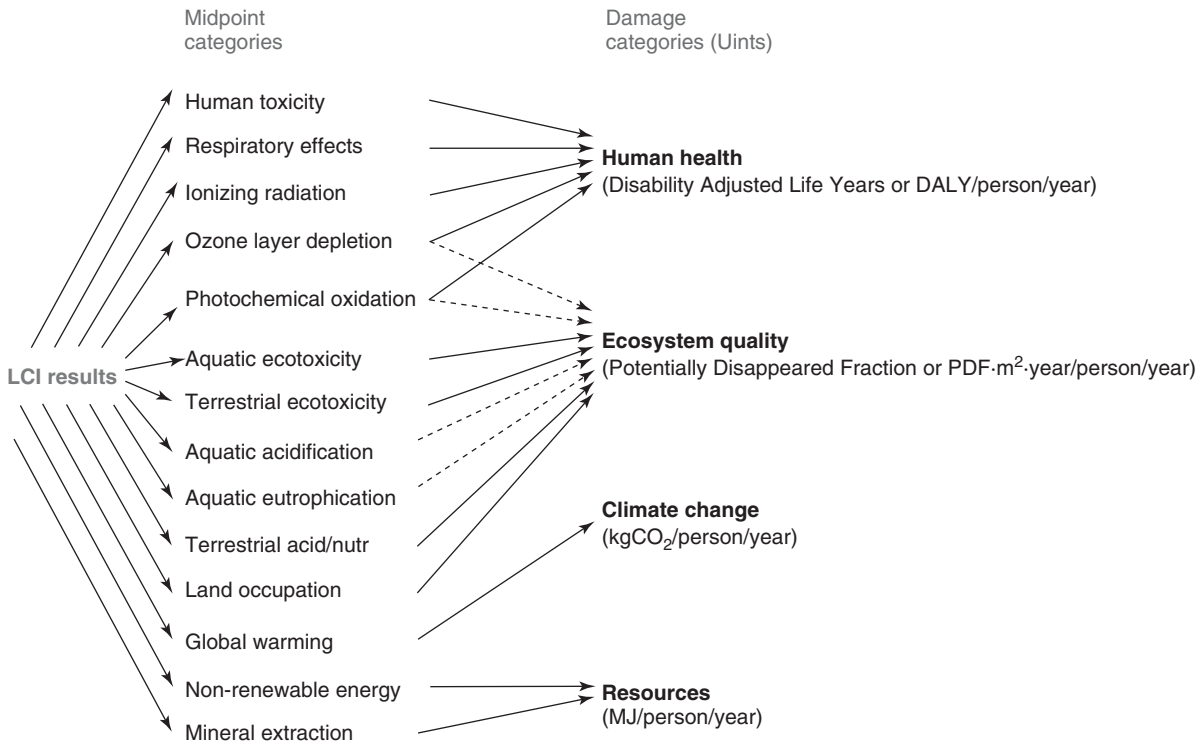


Figure 10. Framework for Impact 2002+, a life cycle impact assessment (LCIA) method.

4 AUTOMOTIVE LCA RESULTS

In an LCA study, the final step of interpretation accompanies the results of the inventory and impact assessment, in relation to the objectives of the study. This involves reviewing, analyzing, and reporting the results. Explaining assumptions and limitations of the study, as well as generating recommendations are also part of this step.

Let us now consider two simple case studies to illustrate the application of LCA to automobiles and interpretation of the results. The cases are: (i) assessing the life cycle GHG emissions of a conventional gasoline car, followed by (ii) a comparative assessment of electric cars in the United States.

Various LCA modeling software titles are available to help practitioners carry out LCA studies. These include SimaPro, GaBi, Quantis Suite, Umberto, and openLCA, among others. These software packages usually integrate the use of LCI databases and impact assessment methods to help users derive results and generate reports more readily. In the following two case studies, none of the mentioned LCA software packages were used, however. The datasets are instead compiled and results prepared using spreadsheets.

4.1 Life cycle greenhouse gas emissions of a gasoline car

Let us first examine an average compact gasoline car that is driven in the United States. Being specific about the location where the automobile is used would provide indication of its use characteristics, such as lifetime distance driven and average fuel consumption. All of the previously discussed automobile's life cycle stages (shown in Figure 2) will be included within the "cradle-to-grave" scope of this case study. In addition, the following key assumptions will be made:

- The car's curb weight, or weight without passengers or cargo, is 1235 kg;
- The material breakdown of the car is as shown in Table 1;
- Life cycle GHG emissions inventory data are obtained primarily from the GREET database;
- The car has an average fuel consumption of 7.8 L/100 km, or 30 miles per gallon (MPG); and
- Distance traveled over its service life is 245,000 km (152,000 miles).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

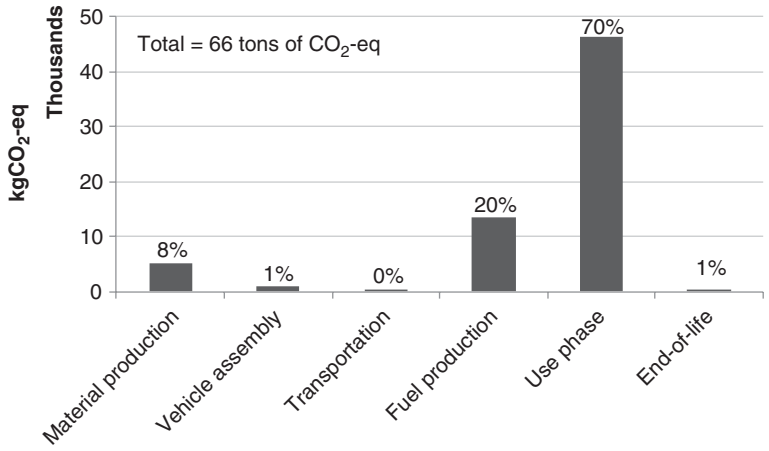


Figure 11. Life cycle greenhouse gas emissions of an average automobile, by life cycle stage.

Other assumptions relating to this LCA case study are detailed in the Appendix A.

The resulting GHG emissions by life cycle stage from an average US compact gasoline car are shown in Figure 11. The emissions are reported in the form of kilograms of carbon-dioxide-equivalent (kg CO_2-eq). Over its life span, an automobile is estimated to emit 66 metric tons of CO_2-eq. While GHG emissions arise at every stage of an automobile's life, combustion of fuel (gasoline) during automobile use accounts for the greatest share of emissions. The automobile's long use phase thus dominates (70%) its life cycle impact. The next most dominant stage is fuel production and distribution (known as *fuel cycle* or "*WTT*" *stage*, 20%), followed by material production (8%). The GHG emissions arising from the other life cycle stages—component manufacturing and vehicle assembly, transportation, maintenance, and end-of-life disposal stages, are modest (2%).

4.2 Life cycle greenhouse gas emissions of electric cars

Unlike a conventional gasoline car that relies on an internal combustion engine in its powertrain, a fully electric vehicle (EV) uses an electric motor as well as an onboard energy storage device, such as a battery, to provide motive force. Without a need to combust fuel within the vehicle, EVs have zero "tailpipe" emissions during their use phase and are generally more energy efficient.

Assessing the environmental impact of EVs is more complex and underscores the importance of a life cycle perspective. Several more inputs are required to understand the electric car's impact, in particular, the user's charging pattern and the emissions associated with generating, transmitting, and distributing electricity used to charge the automobile.

Let us now compare the life cycle impact of two similar electric cars, again in the United States. One will be charged and driven primarily within the State of California, and the other in the US Midwest. The scope will be as before—all of the previously discussed automobile's life cycle stages will be included, and the following assumptions will be made:

- The electric car is a compact car with a 100-mile range;
- Its curb weight is 1535 kg, which includes a 300-kg lithium ion battery;
- The material breakdown of the electric car is as shown in Table A.1;
- When driven, it consumes energy at an average rate of 760 kJ/km;
- Distance traveled over its service life is 245,000 km (152,000 miles);
- The GHG emissions intensity of the California electricity grid is 124 gCO_2-eq/MJ. In the Midwest, this is 2.5 times higher at 314 gCO_2-eq/MJ, due to greater reliance on coal-fired power generation; and
- The electric car's battery will be replaced once over its lifetime.

On the basis of these inputs, the total life cycle GHG emissions from the two electric cars are shown in Figure 12, and benchmarked against that previously estimated for the gasoline car. It is found that electric cars emit more GHG during their material processing and manufacturing phases compared to a conventional gasoline car, due to the additional resources and materials required in the battery and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

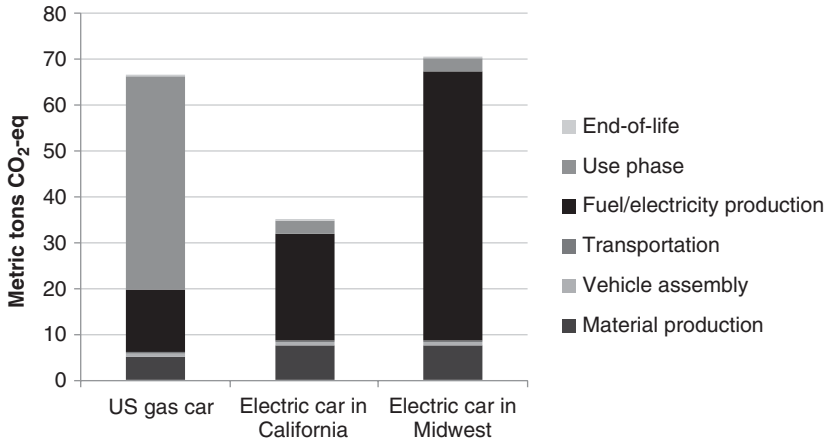


Figure 12. Life cycle greenhouse gas emissions of gasoline versus electric cars in the United States.

electric motor components. As before, emissions associated with vehicle assembly, transportation, maintenance, and end-of-life disposal are trivial.

The upstream or "WTT" phase of producing electricity used to charge these vehicles is observed to dominate their life cycle emissions. This can be significant, depending on the fuel mix used to generate electricity. In the Midwestern states in the United States, the grid relies predominantly on coal to generate electricity, which is carbon-intensive. As a result, an electric car in the Midwest would tend to emit more GHG over its life span compared to a comparable conventional gasoline car of similar size. In contrast, California's electricity generation mix features natural gas, nuclear, hydroelectric power, and renewable sources of energy, which do not emit as much GHG. In this case, driving an electric car does offer GHG savings over the benchmark gasoline automobile.

Using LCA, this case study illustrates how the net GHG emissions associated with electric vehicles depend very much on the nature of the regional electricity grid where they are located and can vary significantly.

5 SUMMARY AND DISCUSSION

LCA is a data-intensive, systematic evaluation of the environmental consequences of manufacturing, driving, and disposal of an automobile. It quantifies the environmental impact in each life cycle stage, from the extraction of raw materials, until the automobile's disposal and waste treatment. This method involves defining the scope of the study, compiling the inventory of energy and material inputs and environmental releases, assessing the environmental impact associated with the inventory, and interpreting the results.

Applications of Automotive LCAs

- Vehicle design for environment
- Material selection
- Environmental performance tracking
- Comparing new vehicle technologies

LCA can be used to assess the environmental performance of new vehicle design concepts, or the impact of new automotive material choices. This methodology can also be used to establish a baseline and track the environmental performance of vehicle models over time. One other application of automotive LCAs is to assess the net impact that result from alternative vehicle technologies and compare them with existing solutions. For example, we compared gasoline versus electric cars in one of the two case studies examined earlier in this chapter.

Using LCA, it has been found that for conventional fossil-fuel-driven automobiles, their life cycle impact is dominated by the impacts arising over their long use phase. Around 70% of total life cycle greenhouse gas emissions are estimated to result from combusting fuel needed to drive the automobile. Improving fuel efficiency is therefore an important focus area, in order to reduce the automobile's energy and environmental footprint. For electric cars, their life cycle impact is dominated by the impacts arising from electricity generation. Depending on the source of energy used to generate the electricity, their environmental performance would vary.

As we consider the automobile's lifespan and the entire set of activities, it is clear that its environmental impact is not controlled by a single company or consumer. The environmental profile of the automobile depends on the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

environmental performance and decisions made by companies involved across its supply chain, from component manufacturers to oil companies, original equipment manufacturers (OEMs), and even highway administration, which affect the way an automobile is driven. Not only is LCA a tool to identify vehicle design aspects to focus on, it also helps identify relevant stakeholders that can influence the environmental performance of automobiles on our roads. Using findings from LCA studies to generate recommendations and engage stakeholders is thus a valuable follow-on activity.

APPENDIX A

Case study assumptions

Additional assumptions to assess the life cycle GHG emissions from an average gasoline car:

- For the raw materials processing stage,
  - The GREET database is used to derive the life cycle inventory data (Table A.2).
- For the automobile assembly stage,
  - 889 $kgCO_2$-eq is emitted (Sullivan, Burnham, and Wang, 2010).
- For the transportation stage,
  - Details on the distances, modes of transportation, and GHG intensity are shown in Table A.3. Only the following links are accounted for:
    - Transportation of iron and steel from domestic suppliers to a domestic auto manufacturer;
    - Transportation of automotive components from global suppliers to a domestic auto manufacturer;
    - Transportation of finished car from assembly plant to dealership;
    - Transportation of retired car from consumer to junk yard.
- For fuel processing,
  - The WTT impact is 22 gCO_2-eq/MJ of gasoline delivered;
  - The heating value of gasoline is 32 MJ/L.
- For the use stage,
  - The tank-to-wheel impact is 73 gCO_2-eq/MJ of gasoline used.
- For automobile maintenance (included in use stage), only the following processes are accounted for:
  - Replacement of fluids and lubricants;
  - Two battery replacements;
  - Two sets of tire replacements;
  - This additional inventory leads to 1755 $kgCO_2$-eq of emissions.
- For eventual automobile disposal,
  - 314 $kgCO_2$-eq is emitted (Burnham, Wang, and Wu, 2006).

For the case study involving electric cars, similar assumptions have been made. In addition:

- For the raw materials processing stage, the material breakdown of the electric car is shown in Table A.1;
- For automobile maintenance (included in use stage), only the following processes are accounted for:

Table A.1. Material breakdown and GHG intensity of a pure electric car.

Material	kg	$kgCO_2$-eq/kg Material
Iron	66	0.54
Regular steel	423.39	3.19
	156.95	3.19
High strength steel	179	5.35
Aluminum	91.86	6.89
	76.30	6.89
Copper and brass	123	8.48
Nickel	8	7.91
Cobalt	8	7.91
Lithium oxide (LiO_2)	16	7.91
Manganese	8	8.83
Graphite/carbon	32	14.87
Plastics	149	4.75
Rubber	23	3.19
Glass	35	1.56
Fluids and lubricants	84	7.60
Electrolyte	26	1.75
Other materials	31	7.60
Total mass, kg	1535	—
Total GHG, $kgCO_2$-eq	—	7589

Table A.2. GHG intensity of automotive material processing.

Material	$kgCO_2$-eq/kg Material
Regular steel	3.19
High strength steel	5.35
Other steels	3.19
Iron	0.54
Aluminum	6.89
Magnesium	28.59
Copper and brass	8.48
Lead	0.63
Zinc	8.83
Powder metal parts	6.17
Other metals	6.17
Plastics and composites	4.75
Rubber	3.19
Coatings	7.60
Textiles	4.75
Glass	1.56
Other materials	7.60

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table A.3. Case study: details of automobile's transportation stage.

Item	Origin–Destination	Distance (km)	Mode	GHG Intensity ($kgCO_2$-eq/ton-km)
Iron and steel	Pennsylvania–Michigan	1020	Rail	0.05
All other	China–Michigan	11,810	Ocean freight	0.01
materials/parts		3670	Rail	0.05
Finished car	Michigan–California	3670	Truck	0.21
Retired car	Household–junkyard	100	Truck	0.21

- One Li-ion battery replacement;
- Two sets of tire replacements;
- This additional inventory leads to 3049 $kgCO_2$-eq of emissions.

RELATED ARTICLES

Overview of Electric, Hybrid and Fuel Cell Vehicles
Exhaust Emissions
Customer Expectations
Recycling of Polymers & Composites
Lightweighting Approach: A Historical Perspective

REFERENCES

Burnham, A., Wang, M., and Wu, Y. (2006) *Development and Applications of GREET 2.7 - Transportation Vehicle-Cycle Model*, Argonne National Laboratory, Argonne, Illinois.

California Air Resource Board (2009) "LCFS Lookup Tables as of December 2009." *Low Carbon Fuel Standard Program*. December http://www.arb.ca.gov/fuels/lcfs/121409lcfs_lutables.pdf (accessed 15 January 2014).

EPA, U.S. Environmental Protection Agency (2013) "Light-duty Automotive Technology, Carbon Dioxide Emissions, and Fuel Economy Trends: 1975 through 2012."

IEA (2010) *Key World Energy Statistics*, International Energy Agency, Paris.

ISO (2006) "ISO 14040:2006 Environmental management – Life cycle assessment – Principles and framework." Note: British Standards can be obtained in PDF or hard copy formats from the BSI online shop: www.bsigroup.com/Shop or by contacting BSI Customer Services for hard copies only: Tel: +44 (0)20 8996 9001, Email: cservices@bsigroup.com.

Jolliet, O., Manuele, M., Raphael, C., *et al*. (2003) IMPACT 2002+: a new life cycle impact assessment methodology *The International Journal of Life Cycle Assessment*, 324–330.

NHTSA, United States National Highway Traffic Safety Administration (2006) "Vehicle Survivability and Travel Mileage Schedules."

Sousanis, J. (2011) "World Vehicle Population Tops 1 Billion Units." *WardsAuto.com*, 15 August (accessed 15 January 2014).

Sullivan, J.L., Burnham, A., and Wang, M. (2010) *Energy-Consumption and Carbon-Emission Analysis of Vehicle and Component Manufacturing*, Argonne National Laboratory, Argonne, Illinois.

Sullivan, J., Williams, R., Yester, S. *et al*. (1998) "Life Cycle Inventory of a Generic U.S. Family Sedan Overview of Results USCAR AMP Project." *Total Life Cycle Conference and Exposition*. Graz, Austria: SAE International.

UChicago Argonne LLC (2011) GREET Model. 22 February http://greet.es.anl.gov/ (accessed 15 January 2014).

Ward's Communications (2010) "Ward's Motor Vehicle Facts and Figures." Detroit, MI.

Wikimedia (2010) Commons. Retrieved from http://commons.wikimedia.org/wiki.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto165
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Intelligent Transport Systems: Overview and Structure (History, Applications, and Architectures)

John C. Miles
Ankerbold Consulting, Nailsworth, Gloucestershire, UK

1 Introduction 1
2 Origins of ITS 2
3 Development of ITS Practice 5
4 The Structure of ITS 8
5 Conclusions 14
Acknowledgments 15
Related Articles 15
References 15
Further Reading 16

1 INTRODUCTION

Intelligent Transport Systems (ITS) is a generic term for the integrated application of communications, control, and information processing technologies to the transportation system. Computers, electronics, satellites, and sensors are applied to the performance of transport operations and for traffic management, journey planning, and vehicle control. The benefits are various, the most important being greater safety, efficiency, and convenience. Successfully applied, ITS can bring time, cost, and energy-efficiency gains. ITS can also help to inform travelers about their journeys, improve interconnections between different transport modes, and contribute directly and indirectly to environmental sustainability.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

> ITS covers all modes of transport and consider all elements of the transportation system – the vehicle, the infrastructure and the driver or user, interacting together dynamically. . . . The definition encompasses a broad array of techniques and approaches that may be achieved through stand-alone technological applications or as enhancements to other transportation strategies. (Chen and Miles, 2004)

The applications of ITS are often grouped into bundles, based largely on the acronyms used in the early days of ITS:

- Advanced Traffic Management Systems (ATMS)
- Advanced Traveler Information Systems (ATIS)
- Advanced Vehicle Control Systems (AVCS)
- Commercial Vehicle Operations (CVO)
- Advanced Public Transport Systems (APTS)
- Electronic Payment Systems (EP)
- Security and Safety Systems (SSS)

ITS in all these forms have become central to efficient transport operations and the relief of road traffic congestion. Computer systems for traffic management are widely used to ensure maximum efficiency of the road network: for example, the coordination of traffic signals to minimize delays or by detecting and managing incidents on the highway network. Electronic payment, access control, and enforcement systems also rely on ITS technologies and are applied to automatic tolling, road pricing, and congestion charging. People experience the benefits of ITS through satellite navigation and a variety of travel and traffic information services. In-vehicle and portable navigation devices (*SatNavs*) give door-to-door route guidance with remarkably detailed turn-by-turn instructions to the driver. Web sites are available for door-to-door

journey planning with advice on travel mode and route choices. There are roadside electronic dynamic message signs to provide important safety and traffic information to drivers; automated traffic alerts are sent by text message or digital coding to smartphone handsets and car radios.

The information on traffic and transport is becoming ever more immediate and ubiquitous. Real-time travel information will include coverage of current point-to-point journey times, current and forecast traffic and weather conditions, and information on service disruptions. Traffic camera images at congestion hot spots can be viewed remotely on smartphones and mobile handsets over the Internet. All these are examples of ITS-based information services that bring benefits though better journey planning, improved navigation, or more efficient fleet management and logistics.

Other applications of ITS include:

- traffic signal and speed enforcement using camera-based systems for vehicle license plate/number plate recognition;
- vehicle safety systems, including adaptive speed control, collision detection, automatic collision notification, and cooperative vehicle-highway systems (CVHS);
- systems that make public transport travel more attractive by providing real-time information at bus and tram stops;
- electronic ticketing systems that are more convenient for passengers and allow the operators to offer flexible, integrated ticketing;
- systems that give priority to public transport vehicles at traffic signals to reduce journey times, improve reliability and punctuality;
- automatic vehicle location systems for fleet managers to monitor service reliability and just-in-time deliveries, with vehicle tracking for stolen vehicle recovery;
- car and truck parking information systems that show availability in real time; and
- air-quality monitoring and management, such as pollution detection and prediction and implementation of strategies to ease air-quality problems.

This chapter traces the development of ITS from its early beginnings through to the present day. It looks at the structure of ITS applications that has evolved and the important role that system architecture has played in providing a framework for these developments.

2 ORIGINS OF ITS

Much of modern ITS technology was originally developed for use on road networks, with computerized urban traffic signal control systems such as SCOOT (Split Cycle Offset Optimization Technique) and SCATS (Sydney Coordinated Adaptive Traffic System). The history of developments that took place in Japan, Europe, and North America will show the origins of ITS and introduce some of the principal components.

2.1 Japan

In Japan, research and development in urban traffic control can be traced back to the initial work on the Tokyo traffic control system, begun in 1967. There followed a steady program to provide signal control across all of urban Japan through a series of 5 year plans, starting in 1971. By 1985, the number of urban traffic control centers operated by the police had reached 74. Separately the Japan Ministry of Construction, with responsibility for road construction and maintenance, had invested extensively heavily in traffic detection and surveillance equipment on the motorways. By 1990, each of the heavily congested metropolitan expressways in Tokyo and Osaka had its own control center. These urban and expressway control centers and the associated investment in traffic monitoring provided an important foundation for the development of ITS in Japan.

Japan also invested heavily in traveler information services. On the urban expressways in Tokyo automated dynamic message signs to warn of congestion first appeared in 1973. Roadside radio began in 1983, broadcasting traffic information with automatically generated voice messages using low power transmitters. Highway advisory radio was widely adopted for traffic information services on the urban expressways and busy downtown areas, even for news about parking space availability.

Along with early investment in traffic control and information systems, Japan was researching navigation systems using ground to vehicle data communications. Development of a Comprehensive Automobile Control System (CACS) began in 1973 and featured an interactive route guidance system using inductive loop antennae and an in-vehicle display unit. This was followed by the Automobile Traffic Information and Control System (ATICS), starting in 1978. ATICS culminated in a demonstration of route guidance systems known as the *Tsukuba experiment* to coincide with the Tsukuba Science Exposition in 1985 (Davis *et al*., 1991).

The next round of development involved the Japanese automotive and electronics industry in two projects to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

develop car navigation systems: RACS (Road/Automobile Communication System) and AMTICS (Advanced Mobile Traffic Information and Communication System). RACS, begun in 1984 and sponsored by the Ministry of Construction, used a dedicated beacon-based communications system on the expressways. The two AMTICS demonstrations in metropolitan Tokyo and Osaka and initiated by the National Police Agency used commercially available mobile data communications systems. In parallel, the Ministry of Construction was working on development and exploitation of the Japan Digital Roadmap (Kawashima, 1994).

Following the success of RACS and AMTICS, an important initiative emerged to promote in-vehicle navigation as a consumer product, combining route guidance with traffic information. A consortium of Japanese automotive and electronics companies, known as the *VICS Promotion Council* (*Vehicle Information and Communication System*), was formed in October 1991 with the intention of developing a national system to collect and distribute traffic information in real time (Harvey, 1999).

2.2 Europe

Developments in Europe follow similar lines to that in Japan. For example, work on vehicle navigation and route guidance started in the late 1970s in Germany with the ALI system (Autofahrer Leit- und Informationssystem) from Bosch/Blaupunkt, which like CACS in Japan used inductive loops in the roadway to communicate with the vehicles. In addition, in Germany, Siemens developed the AUTO-SCOUT system using infrared communications. The two companies collaborated on ALI-SCOUT, later called *EUROSCOUT* (Sodeikat, 1994), which used infrared beacons for a two-way roadside-vehicle data link. An important feature of ALI-SCOUT was the way equipped vehicles could be used as traffic probes to report their journey times between intersections to a central database, which could be used in real time for route planning to avoid congestion. A large-scale demonstration of ALI-SCOUT was installed in Berlin in 1988 to test these concepts. An operating company COPILOT was formed with the main shareholders Siemens and Bosch to commercially market and roll out the system. Trials began in Stuttgart but the main barrier for success was the necessity to negotiate details of infrastructure and system operation with each city's authority independently. Another barrier was the emerging mobile communication, which allowed for a more flexible communication architecture and (theoretically) faster roll out.

SCOOT, the computerized urban traffic control system that monitors and responds to traffic conditions in real time, was developed at the Transport Research Laboratory in England during the 1980s. Other game-changing research and development (R&D) programs in Europe were PROMETHEUS (1985–1993), which engaged the European automotive industry in a collaborative program of research and DRIVE (1989–1991). Results from both programs were presented at the DRIVE conference held in Brussels in February 1991 with the title *Advanced Telematics in Road Transport* (European Commission, 1991). The significance of DRIVE and PROMETHEUS is considered elsewhere in this encyclopedia (see Evolution and Future Trends).

The ALI-SCOUT technology from Germany formed the basis of the London Autoguide project, first mooted in 1986 and the subject of UK legislation to provide the necessary licensing framework. Legislation was sought because of concerns about the system sending volumes of traffic on inappropriate routes such as residential streets (so-called rat-running) this being a hot political issue in London especially for heavy good vehicles at night. The legislation permitted the Autoguide license to provide "positive-and-dynamic" route guidance to their customers, subject to terms and conditions to be specified in the operating license (Catling, Miles and Harris, 1999).

In the event, the Autoguide project was abandoned in the early 1990s because of a combination of factors, among them a failure to agree satisfactory commercial terms for the selected license. Nevertheless, there was one successful taker for a UK license to provide real-time traffic information on a commercial basis, namely TrafficMaster. An initial pilot scheme was set up in 1990 on the London orbital motorway (M25) using traffic speed sensors mounted on over-bridges at strategic intervals along the motorway, each with a telephone landline to transmit data back to the center. Data for sites with slow-moving traffic (below 50 km/h) were automatically transmitted to subscribers by telephone paging. The TrafficMaster terminal (Figure 1) presented this data to the driver in the form of a monochrome schematic map display, allowing the driver to assess the traffic conditions ahead. Over the next 10 years, some 6000 sensors were installed by this private company across the UK motorway and trunk road network to provide a national coverage. Although modest in its use of technology, TrafficMaster is significant because it was the first example of a fully commercial ITS-based real-time traffic information service.

2.3 North America

Compared with Japan and Europe, North America was fairly late in exploiting the potential of computers to assist

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 1. Philips "Socrates" and TrafficMaster "YQ" in-vehicle units demonstrated at the G7 Conference on the Information Society, Brussels 1995. (Reproduced by permission of the Author. © John Miles.)

traffic management on city streets. Computerized urban traffic signal control was not introduced to Los Angeles until 1984, just before the Olympic Games. It was not until the 1990s that computer control of urban traffic signals began to take hold, with New York modernizing its traffic control system and an ambitious program in Los Angeles to cover all 4000 signalized intersections. A demonstration in 1993 of the SCOOT adaptive traffic control system in Toronto, Canada, showed an average reduction in vehicle delays of 14% over the existing fixed-time plans (Greenough and Kelman, 1994).

By comparison, traffic management on urban freeways in North America had a much higher profile. Electronic ramp metering was first tried in Chicago in 1963, with Los Angeles following in 1968. By 1995, ramp meters were in operation in 23 metropolitan areas of North America. Improvements in mainline speeds of 16–62% and accident reductions of 24–50% were reported. The downside in peak periods was that traffic might be held on the access ramp for a considerable time, up to 20 min, before gaining access to main line. Priority lanes for high occupancy vehicles (HOV lanes) were another way to improve the use of road space and give them priority.

The main metropolitan areas of the United States frequently experience serious freeway congestion with accidents causing long delays. Through the 1980s and 1990s, downtown freeways in cities such as Chicago, Detroit, Houston, Los Angeles, San Francisco, and Toronto were equipped with loop detectors, closed-circuit TV cameras (CCTV), and text-based electronic dynamic message signs at the roadside to alert drivers to lane closures and traffic accidents. Traffic control centers were built for staff to monitor the cameras, handle emergency calls, and operate incident management procedures in liaison with the police, breakdown trucks, and emergency services. In the Chicago area, special emergency traffic patrols with heavy-duty breakdown equipment were deployed for rapid response, known appropriately as the *Illinois Minutemen*. Automatic incident detection based on loop detection and artificial vision was another development.

The year 1992 was a turning point in North America, with three significant events. One of the first field tests of ITS technologies in the United States was launched in Oakland County on the outskirts of Detroit. Called FAST–TRAC (an acronym for *Faster and Safer Travel through Traffic Routing and Advanced Control*) it brought together the Australian system of adaptive signal control responding to real-time traffic conditions (SCATS) with the German Ali-Scout system of dynamic route guidance. Twenty-eight intersections were brought under SCATS control and a dedicated network of infrared beacons installed to transmit optimal route directions to drivers based on the current traffic situation. Vehicles were equipped with an infrared transponder and a dashboard unit that displayed turn-by-turn guidance for the driver.

In the same year, 1992, the US Department of Transportation (USDOT) and General Motors launched another landmark project in Orlando, Florida, known as *TravTek*, to demonstrate navigation and route guidance. A precursor to the ubiquitous satellite navigation consumer products of today, TravTek coverage was 10,000 miles of road with 75,000 intersections over an area of 1200 square miles. A database of yellow pages and tourist information was included in the package (USDOT, 1996).

The third significant event of 1992 was the publication by the Intelligent Vehicle-Highway Society of America, (now *ITS America*) of a strategic plan for Intelligent Vehicle-Highway Systems in the United States (Deen *et al*., 1992). This 295-page report discussed organizational roles and responsibilities and legal and institutional matters in addition to detailing the main technical opportunities. In response, the USDOT funded a 3- year study to examine the requirements for implementing ITS across America. This major piece of work embraced 29 ITS-based user services nominated by USDOT (some already in place) and was carried out by defense industry contractors with experience of planning and designing large and complex systems. It took a top-down approach to look at the potential co-dependencies, data sharing, and communications requirements that could make these services happen and matched this with an evaluation of costs and benefits. The project included extensive outreach to major stakeholders

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

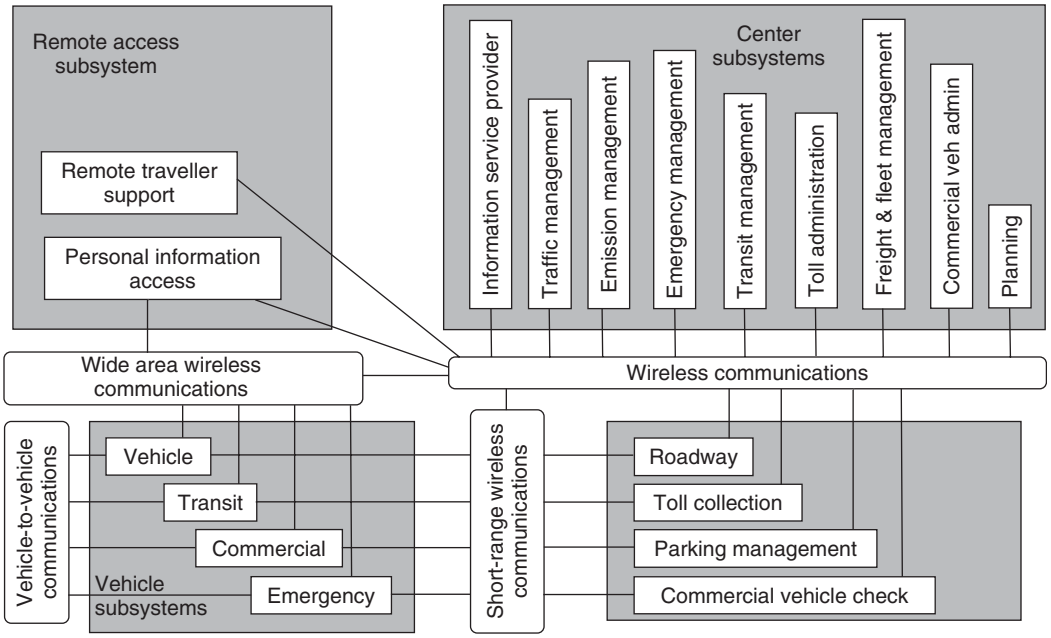


Figure 2. US National ITS architecture (1996): subsystems and communication elements. (Reproduced by permission of US Department of Transportation.)

through a series of regional and national consultation meetings. The outcome was a proposed architecture for ITS with 19 major subsystems gathered into four groups as shown in Figure 2 (Chen and Costantino, 1999).

Measured by its breadth and comprehensiveness, the US National ITS Architecture study was breaking new ground. It did a great deal to put ITS on the map in North America by creating awareness among transportation professionals of the potential benefits of ITS and the value of taking a systems approach. It spawned an intensive program of work to develop national and international standards and protocols that would promote the widespread use of ITS technology with the aim of ensuring interoperability to the maximum extent practicable. Since 2001, it has been a matter of policy in the United States that ITS projects that benefit for Federal funds shall demonstrate compliance with the national or regionally adapted ITS Architecture requirements. Canada has followed with its own variant of ITS Architecture that recognizes the need for cross-border interoperability.

3 DEVELOPMENT OF ITS PRACTICE

3.1 Emergence of ITS

With a US strategic plan for ITS in the making and commercial applications coming onstream in Europe and Japan, a world congress on *Applications of Transport Telematics and Intelligent Vehicle-Highway Systems* took place in Paris in 1994. This landmark event was organized by ERTICO, the European Road Transport Telematics Implementation Organization, now also known as *ITS Europe*. The event brought together researchers and practitioners from Europe, Japan, and the United States and laid the foundations for a series of congresses that continues today. The ITS World Congress is held every year, successively in Europe, Asia-Pacific, and North America. It provides an important platform for debating deployment issues and establishing ITS practice.

Hard on the heels of the Paris congress there were developments at a political level. In February 1995, a conference on *The Information Society* took place in Brussels for the leaders of the Group of Seven leading industrial nations. The objective was to raise awareness of the revolution in digital technologies that was occurring. Hosted by the European Commission, it included a showcase of demonstrations on how information and communications technologies were being applied to automobile engineering and transport. (Figure 1.)

That same year a second congress took place in Yokohama, organized by VERTIS (*ITS Japan*) under the banner *World Congress on Intelligent Transportation Systems*. This title has remained and with it the widely used acronym ITS. Japan paraded state-of-the-art traffic control centers in Yokohama and Tokyo with video walls that greatly

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

impressed visitors from around the world. The VICS system, mentioned earlier, was also demonstrated at the congress.

The following year in April 1996 the VICS center real-time traffic information service commenced, connecting with transmitters along major highways between Tokyo, Nagoya, and Osaka. VICS continues today as a thriving public–private partnership with a nationwide data transmission network. VICS transmits real-time traffic data that is decoded by the in-car VICS unit for display on digital maps.

The true utility and benefit of VICS is hard to quantify directly. However, simulations using models suggested that if 20% of the drivers on the metropolitan highway use VICS the congestion will be reduced by 10% (Chen and Miles, 2004). It has a reputation as one of the world's leading dynamic traffic information provision systems. Cars with receivers can get up-to-the minute information on traffic, weather, and parking. The huge sales of the VICS units in Japan attest to its popularity and the success of the VICS strategy as a model for ITS deployment. As of March 2011, there were over 30 million VICS units installed in car navigation systems (see Applications—Intelligent Vehicles: Driver Information).

South Korea and Singapore were two countries quick to follow Japan's lead. Singapore conducted trials with nonstop electronic tolling as early as 1994 and Korea hosted the fifth World Congress in Seoul in 1998. By then, Korea was advancing with operational tests of ITS and had started work on a national ITS architecture study.

One of the first examples of nonstop electronic payment was the Trondheim toll ring in 1991, set up to improve traffic conditions in the city center and to fund a package of major transport infrastructure improvements. The system was one of the first in Europe to use short-range communications between the vehicle and roadside equipment as a method of charging. In 2001, it was upgraded to as part of a national strategy to create an interoperable system for all tolls in Norway known as *AUTOPASS*. Canada followed Norway's lead in 1997 when Highway 407, one of the world's first all-electronic nonstop toll roads, opened in Toronto.

3.2 Characteristics of ITS

The movement of people, vehicles, and goods brings with it a need to determine geographical position with a known level of accuracy and with a location reference that has meaning for the user—the answer to the question: "Where am I?" President Regan's decision in 1983 to open up Global Positioning Satellites (GPS) for civilian use opened up the possibilities. A further step was taken in 2000 when selective availability of the GPS signals was turned off. This meant that positioning could be done with a level of accuracy previously only available to the military, improved from 100 m (330 ft) down to 15–20 m (66 ft). The use of differential GPS allows for further refinement, in some cases down to 10 cm (see Technologies—Positioning: GNSS). However, accurate coordinates are only part of the ITS story. Also needed is a representation of the transport network with a level of detail and accuracy to match.

Navigation databases have their origin in the mid-1980s. The Dutch provider of digital maps, *TeleAtlas*, was founded in 1984. Japan established the world's first National Digital Road Map database standards in 1988. About the same time, the Navigation Technologies Corporation (Nav-Tech, now known as *NavTeq*) began mapping the San Francisco Bay Area. High resolution digital maps drawn from transport network models and navigation databases have a very fundamental part to play in ITS. With differential GPS matched with "always on" communications and very detailed digital mapping, it becomes possible to keep track of a vehicle's position on the road in real time with a high degree of confidence (see Tracking and Navigation for Goods and People). The combination of detailed digital maps and accurate positioning systems, linked with data from in-vehicle sensors and control systems, provides a powerful platform for computing direction indications, warnings, and, if required, an automated driving response (see Applications—Intelligent Vehicles: Autonomous Vehicles).

Although these technological capabilities are impressive, our understanding of intelligence in ITS is not so well advanced. The adjectives *smart* and *intelligent* are very over-worked. Data and information are gathered, stored, and processed but does this amount to intelligence? Information is not the same as intelligence, which requires an understanding of purpose, context, and previous knowledge. Writing in 2004, Joseph Sussman, one of the authors of the seminal 1992 Strategic Plan of ITS America, observed that creating the kind of automated network management that was originally envisaged is elusive. He judged that many of the ITS-based systems developed are more correctly described as decision support systems (Sussman, 2004). It is fair to say that the whole subject of Artificial Intelligence in transport is still very much in its infancy (Miles and Walker, 2006).

By the mid-1990s, the basic set of automated vehicle controls, such as adaptive cruise control, collision warning, and automated lane keeping were sufficiently well developed to mount a demonstration of automated driving. The US National Automated Highway System Consortium (NAHSC) was formed in response to legislation passed by

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the US Congress in 1991. In August 1997, using dedicated car-pool lanes on Freeway I-15 in San Diego California, the consortium presented a convincing demonstration of seven different automated driving scenarios using some 20 fully automated vehicles. Automated highways were seen as the next evolutionary step in highway transportation systems bringing benefits to all classes of traffic (NAHS Review Committee, 1998). Similar development work was taking place in Europe with electronic coupling of trucks through the CHAUFFEUR project (Schulze *et al*., 1998) and in Japan the Advanced Cruise-Assist Highway Systems concept (MLIT, 2002) (Bishop, 2001).

Greater automation brings with it a second important question: what happens if the "intelligent" system fails though mishap or misadventure? A system failure, however caused, demands the notion of *graceful degradation* as a basic principal of traffic system design. It means that if an important component of the system drops out the failure should not be catastrophic. The system should be resilient and have the capability to adapt. In the 1980s, graceful degradation was built-in as a design feature of SCOOT and the Dutch motorway control system and has continued to inform the deployment of dynamic traffic systems ever since. System design issues such as these become critical where public safety is an issue, not least because of worries about manufacturer's liability, which is a prominent factor in the case of advanced vehicle control systems. Therefore, although vehicles began to incorporate an increasing amount of sensor technology, such as front anti-collision radar, focus area side radar, video-based road recognition sensors, GPS positioning, and navigational maps, this was usually in the form of driver support, not to take over vehicle control functions (see Driver Assistance).

3.3 Growth and consolidation

The first meeting of the PIARC (*World Road Association*) Technical Committee on Intelligent Transport took place in 1996. This PIARC committee was responsible for publishing the *ITS Handbook 2000* (PIARC, 1999) and in doing so the committee did much to spread knowledge about ITS deployments and make ITS accessible to the mainstream of highway and transportation planning practitioners, including those from countries with emerging economies. A revised edition of the handbook published in 2004 has been translated from English into French, Spanish, and Chinese (Chen and Miles, 2004).

In the United States, work continued to establish and refine the costs and benefits of public-sector ITS deployments following on from the program of field operational tests and the National ITS Architecture project. The USDOT established a database on costs and benefits, which it still maintains today (USDOT, 2013). The subject of ITS benefits and costs was the theme of a one-day workshop organized by ITS America at the seventh World Congress held in Turin, Italy, in November 2000. Out of this initiative grew IBEC, the influential International Benefits, Evaluation, and Costs group which was formally established at the 2002 ITS World Congress in Chicago (IBEC, 2013).

Meanwhile, steps were being taken in Europe to foster the deployment of cross-border ITS and services. This was in addition to the established Framework Program for Research and Technical Development, which is described by Vits (see Evolution and Future Trends). A set of proposals was published in 1997 identifying five priority areas (European Commission, 1997):

- development of Europe-wide traffic information services using RDS-TMC (the Radio Data System-Traffic Message Channel);
- traffic data exchange and information management, to progress technical standards and operating protocols;
- electronic fee collection, working toward European interoperability of systems;
- safety of onboard devices, through a code of practice for the design of the human–machine interface (HMI); and
- system architecture, with the aim of defining an open system architecture for Europe.

Similar efforts at harmonization of ITS Technologies were taking place on other continents. Among the countries that have followed the US lead and sought to adopt a reference architecture for planning and designing their ITS are Australia, Austria, Canada, Czech Republic, France, Finland, Hungary, Italy, Japan, the Netherlands, Norway, Romania, Taiwan, Slovenia, South Korea, Spain, and Switzerland.

Urban traffic management continued to develop in most European countries. The United Kingdom in 1997 began work on open systems for Urban Traffic Management and Control (UTMC). In the same year, the Swedish Parliament advanced a long-term goal known as *Vision Zero* with the intention that no one should be killed or seriously injured within the Swedish road transport system. In Japan, the National Police Agency sponsored work on driving safety support systems (DSSS) as part of their Universal Traffic Management System (UTMS).

The beginning of the twenty-first Century saw the consolidation of ITS and services regionally, which had until then existed in isolated examples, trials and demonstrations. A group of experts from Canada, the European Union, and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the United States was formed to review experiences of successful deployment of ITS on both sides of the Atlantic (Miles *et al*., 2002). There followed a period of rapid expansion. Numerous case studies of successful ITS projects were compiled by PIARC (Chen and Miles, 2004; PIARC, 2007 and 2011).

Systems for "Active Traffic Management" of heavily trafficked and congested sections of European motorways were developed involving a combination of ITS applications. The United States put in place arrangements for a national travel information system using the 511 code. In the United Kingdom, the government sponsored the development of *Transport Direct*, a national multimodal journey planning portal. The rollout of third-generation mobile phone networks with broadband data capabilities made possible the mobile Internet. Traveler information systems took advantage of this broadband capability and the rapid take-up of "Smart" telephone handsets.

In Europe, the use of ITS-based payment and enforcement systems gathered momentum. London introduced its congestion charge in 2003 with the object of reducing congestion in the central area. It uses video enforcement based on vehicle license plate number recognition and compares the result against a database of permitted vehicles. Prepayment of the congestion charge puts the vehicle onto the permitted list for that day. Postpayment is possible up to midnight on the day. A different approach using GPS technology was adopted in Germany on the autobahns for heavy goods vehicles starting in 2005. An onboard unit tracks the vehicle's movement. Charges are calculated based on the distance driven, the number of axles, and the emission category of the vehicle. Austria also introduced a scheme for charging heavy goods vehicles but with different technology, having installed a multilane, free flow open toll system across the entire Austrian motorway network.

France illustrates the safety potential of ITS. Starting in 2003, the country invested heavily in a national automatic enforcement and penalty system for speed violations. The French system uses digital technology with over 3000 devices in operation throughout the road and highway network. The system was developed with reference to the French national ITS architecture ACTIF and allows for near-complete automation of the detection and penalty procedure covering the entire enforcement chain, from detecting offences to paying fines including the legal processes. The effect has been positive with reductions in average speed of 7.5% since 2004 and a very marked reduction in accident fatalities.

4 THE STRUCTURE OF ITS

4.1 Technologies and user services

Communications are at the heart of ITS. The emergence of the Internet in the latter part of the 1990s along with the roll out of digital cell phone networks with capabilities for data transmission proved to be highly significant. The prospect of a mobile information society came into focus that would encompass an "intelligent" transport infrastructure, intelligent strategies for operating that infrastructure, advanced vehicle technologies and widespread dynamic traffic and travel information services operating in real time. However, to achieve this vision required the development of a wealth of common systems embracing among other things: mobile communications, data models, data dictionaries, data exchange, multi-service smartcards, digital mapping, navigation databases, and location referencing standards.

Table 1 illustrates the range of technologies that are involved. They work together in various combinations but some technologies are absolutely fundamental. For example, digital communications in various forms provide the backbone of ITS (see Technologies—Communication: Mobile and Technologies—Communication: Wireless LAN-based Vehicular Communication). Data is another key element (Table 2) in ITS (see Technologies—Data Acquisition: Data fusion) and is the subject of an encyclopedia article in its own right. Data generated by ITS may provide real-time information about current conditions on a network, or on-line information for journey planning, enabling highway authorities and agencies, road operators, passenger transport, and commercial transport providers and individual travelers to make better informed, safer, more coordinated, and "smarter" use of transport networks (see Applications—Intelligent Vehicles: Driver Information).

The specific applications of ITS are usually identified by the service on offer more than by the technologies that are utilized. The colloquial *SatNav* is an exception that combines a reference to the technology along with the function. A service-based view of ITS represents what ITS applications can deliver that will benefit the user. The concept of user services allows a system or project to be developed that will address identified transport problems and user needs. The International Standards Organization (ISO) has reached broad agreement on a classification of ITS services drawing on experience from the European Union, Japan, the United States and elsewhere (ISO Technical Committee 204, 2007). ISO categorizes ITS services in 11 different service domains. For each service domain the ISO standard provides a listing of ITS

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Some of the principal enabling technologies for Intelligent Transport Systems.

Network data	*Roadway sensor technologies*
Network modeling	Traffic detectors (various types)
Tree-building algorithms	Weigh in motion
Geographical Information Systems	Weather monitoring
Network condition databases	Vehicle sensors/ traffic probes
• Real time	Spot speed detection (Radar)
• Historic	Average speed detection/Journey time monitoring (tag based and ANPR)
Location referencing systems	*Vehicle sensor technologies*
• Global positioning (Lat/Long)	Position sensors
• Mobile phone triangulation	Speed and braking sensors
• Link and node systems (Alert C, etc.)	Vehicle proximity sensors
Digital communications	Occupant comfort (sleepy driver) sensors
Fixed link (fiber, cable, and microwave)	Near-distance obstacle detection
Dedicated Short-Range Communications	Far-distance obstacle detection
• Passive microwave (tag and beacon)	*Signs, in-vehicle units, and handsets*
• Active microwave (IEEE 802.11/WAVE)	Public information points and kiosks
• Infrared	Roadside electronic variable message signs
• Wi-Fi	• Text based
• Bluetooth	• Graphics boards
3G and 4G cell phone networks	In-vehicle units:
Digital radio (DAB; FM digital side band)	• dashboard displays and audio
The Internet	• head-up displays
Navigation	• voice recognition
Global Navigation Satellite Systems	Nomadic devices:
Navigation databases	• Smartphones
Location-based services databases	• Tablet computers
Dynamic maps (local, regional, and national)	• Personal navigation devices
Digital image processing	*Electronic payment*
Automatic vehicle identification	DSRC-based systems
Artificial vision (image recognition)	GNSS-based systems
Automatic incident detection	Permitted list using ANPR (pre- and postpayments)
Vision-based sensors	Electronic ticketing
White line detection	Smartcard ticketing
	Mobile phone payment
	Contactless smartcards

services and service groups that fall into that category (Table 3). Although the services and service domains are made distinct, there are significant interdependencies and co-dependencies to be considered, sometimes across different service domains.

4.2 ITS Architecture

4.2.1 What is ITS architecture?

The amalgamation of technologies to perform advanced ITS functions is based on the principles of systems engineering. Control systems for road traffic were installed initially to provide just one or two services that worked independently as separate autonomous systems. In contrast, the application of information and control technologies to transport systems today requires complex data gathering, information management, and control systems that need careful design. As the use of ITS grows so do the possibilities for synergy between different systems, often involving data sharing beyond established organizational boundaries. System architecture provides a planning and design framework that will make the most of the available synergies.

The US National ITS Architecture was the first comprehensive vision of a future integrated ITS. From 1996 onward, the US National Architecture has been maintained

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Features of the ITS information supply chain (see Technologies—Data Acquisition: Data fusion and Applications—Intelligent Vehicles: Driver Information).

1. Data Acquisition	*4. Transmission and distribution*
Vehicle and roadway sensor data	Multiple communication channels
CCTV and webcams	Timeliness and latency of transmission
Image processing	Security and integrity of transmission
Tag readers	Geographical coverage
Event and incident data	*5. End-user interfaces*
• Automatic detection	Travel news broadcasts
• Journalistic sources	Highway advisory radio
• Crowd sourcing	Roadside VMS
Transaction processing	Internet and mobile internet
Data mining	Social media
2. Data Processing	In-vehicle unit audio and displays
Data formatting	Nomadic devices
Location referencing	Public information points and kiosks
Data dictionaries	*6. Supply & ownership issues*
Data registries	Public-sector data
Data exchange	Commercial and proprietary data
Data fusion	Information service providers
Quality controls	Information branding
3. Information management	Value-capture, revenue, and payment
Data analysis	
Contextual relevance	
Content detection	
Content processing	
Formatting for service	
User needs/preferences	

and developed and is now in its seventh version (USDOT, 2012a). It includes a number of user services that were not even thought of when the original version was published.

A national, regional, or city-wide ITS architecture is a formal statement of the preferred approach to ITS and is the first step on the way to creating detailed system designs. The term *architecture* is used here to describe a reference model to guide the implementation of ITS by different parties so that the whole can function efficiently. It is the conceptual design that defines the structure and/or behavior of an integrated ITS. The architecture will describe the following:

- the specific activities and functions that are required to deliver the ITS service (e.g., "gather traffic information" or "request a route");
- the physical entities or subsystems where these functions reside (e.g., at the roadside or with the vehicle); and
- the information flows and data flows that connect together these functions and physical subsystems to create an integrated system.

In essence, the ITS architecture is an analytical framework that shows conceptually and in some detail how different component systems and technologies should fit together.

A national, regional, or city-wide ITS architecture is by definition strategic: it provides a framework for implementation rather than a definitive system design. The key to establishing a reference architecture at this level is the selection and prioritization of user services. All the main stakeholders and actors should be involved in this process, making architecture development an opportunity for consensus building. Failure to achieve an adequate level of stakeholder involvement and buy-in has sometimes been a barrier to progress.

When developing their own ITS architecture, organizations have to make decisions about the form of architecture that they create, which may be one of three types:

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Classification of ITS services by the International Standards Organization (ISO Technical Committee 204, 2007).

Service Domain	Service Groups
Traveler information	Pre-trip information
	On-trip information
	Route guidance and navigation pre-trip
	Route guidance and navigation on-trip
	Trip planning support
	Travel services information
Traffic management and operations	Traffic management and control
	Transport-related incident management
	Demand management
	Transport infrastructure maintenance management
	Policing/enforcing traffic regulations
Vehicle services	Transport-related vision enhancement
	Automated vehicle operation
	Collision avoidance
	Safety readiness
	Precrash restraint deployment
Freight transport	*Administrative functions*:
	Commercial vehicle preclearance
	Commercial vehicle administrative processes
	Automated roadside safety inspection
	Commercial vehicle onboard safety monitoring
	Commercial functions
	Freight transport fleet management
	Intermodal information management
	Management and control of intermodal centers
	Management of dangerous freight
Public transport	Public transport management
	Demand responsive and shared transport
Emergency	Transport-related emergency notification and personal security
	After theft vehicle recovery
	Emergency vehicle management
	Emergency vehicle pre-emption
	Emergency vehicle data
	Hazardous materials and incident notification
Transport-related electronic payment	Transport-related electronic financial transactions
	Integration of transport-related electronic payment services
Road-transport-related personal safety	Public travel security
	Safety enhancements for vulnerable road users
	Safety enhancements for disabled road users
	Safety provisions for pedestrian using intelligent junctions and links
Weather and environmental conditions monitoring	Environmental conditions monitoring
Disaster response management and coordination	Disaster data management
	Disaster-response management
	Coordination with emergency agencies
National security	Monitoring and control of suspicious vehicles
	Utility or pipeline monitoring

- framework architecture, that is, an architecture consisting of user needs and a functional viewpoint but omitting the physical viewpoint. The FRAME architecture for Europe is an example of this type (FRAME Architecture Consortium, 2011). It is only really suitable for architectures at the national or regional level but it can be used to create architectures of the following two types.
- mandated architecture, which has a physical, communications, and other viewpoints, plus the other required outputs. It can be used at the national, regional, and local levels to provide close specification of what is required for ITS deployments. The contents of the physical viewpoint may be fixed, or there may be a limited range of options.
- service architecture, which is like the mandated architecture, but only supports a particular service, for example, Traveler Information and Public Transport Management.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The operating assumptions that lay at the foundation of ITS architecture need to be expressed in laymen's terms and will explain, for the benefit of the agencies directly involved, what is required. Central are the data-gathering and data-processing activities that will be involved in creating each ITS-based service and the interactions that are needed between the different agencies. Currently, in the light of emerging cooperative vehicle-highway systems (sometimes called *C-ITS*) even the traveler, himself or herself, can be regarded as an element of the ITS architecture.

Achieving a consensus view of what is required is important to inform the determination of functional requirements for the system. Sometimes this is done in a planning document called the *concept of operations* that describes the activities that are components of the total system so that everyone involved have a clear understanding of what is being planned and the interdependencies, whether overt or implicit. The concept of operations will explain the functional requirements for each service component, the character of any data or information to be exchanged, expectations of accuracy and timeliness, and any other dependencies.

4.2.2 ITS architecture viewpoints

ITS architecture is concerned with information exchange and the allocation of control functions between systems at various levels of abstraction. Even if the operational concepts are clear and well documented, there is no single view of a complex system that can convey all aspects of the system in an understandable manner (Belinova, Bures and Jesty, 2010). Instead, multiple viewpoints, depicting different aspects of the system and different types of information are needed. These viewpoints might include: a functional or logical architecture describing how various data should flow and be processed; a physical architecture allocating the functions to physical subsystems; an organizational architecture to assign functions and responsibilities to providers and recipients of selected ITS services; and a communications architecture that specifies what forms of communication are required.

Functional (logical) architecture depicts the processes and data flows that are needed to meet the functional requirements described in the concept of operations. The common ground between the various user requirements and ITS services is examined so that shared functions and common requirements can be grouped within the same set of processes. An important part of the logical architecture is to describe how the system would deal with abnormal circumstances. All failure modes need to be considered for possible safety hazards, and logical steps need to be described to achieve graceful degradation under abnormal conditions, consistent with the higher level concepts described in the concept of operations.

Physical architecture: In the context of systems engineering, the physical architecture allocates the processes defined by the logical architecture into physical subsystems, which the hardware and software will deliver. The top-level diagram of the US national physical architecture in Figure 2 (earlier in this chapter) shows a highly simplified, strategic example. The design of the physical subsystems will be based on the functional requirements, the process specifications, and the interdependencies and will be affected by whether the functions are to be performed in one or more locations. Thus, the physical architecture allocates specific processes to physical subsystems, considering the institutional responsibilities.

Organizational architecture considers ownership and business issues: who owns what, who manages what and the business/contractual relationships between the various parties involved.

> The Organization Viewpoint is usually a derivative of the Physical Viewpoint. It is used to show the organizations that will own, and/or operate, and/or maintain the Sub-Systems and Modules in the Physical Viewpoint. (FRAME Architecture Consortium, 2011)

Communications architecture determines the way in which information is exchanged and describes the kinds of communications that will be required. An example is the European cooperative ITS architecture (ETSI, 2010). The architecture is organized into four principal domains as shown in Figure 3 (COMeSafety Consortium, 2008). It covers all types of communications to and from moving vehicles: vehicle-to-vehicle (V2V), vehicle-to-roadside (V2R and R2V), and vehicle-to-infrastructure (V2I and I2V). The intention is to take advantage of multiple communications media to achieve seamless communication as the vehicle changes location. The hope is that the architecture will give sufficient flexibility for applications to be independent of communications with a degree of future proofing.

4.3 ITS standards

Technology standards have particular significance in the ITS world, firstly because of the global market for automotive products and secondly to promote interoperability of systems across national jurisdictional boundaries, especially important in Europe. Standards provide a specification for the important communications, hardware and software components, to assure that they are compatible one with

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

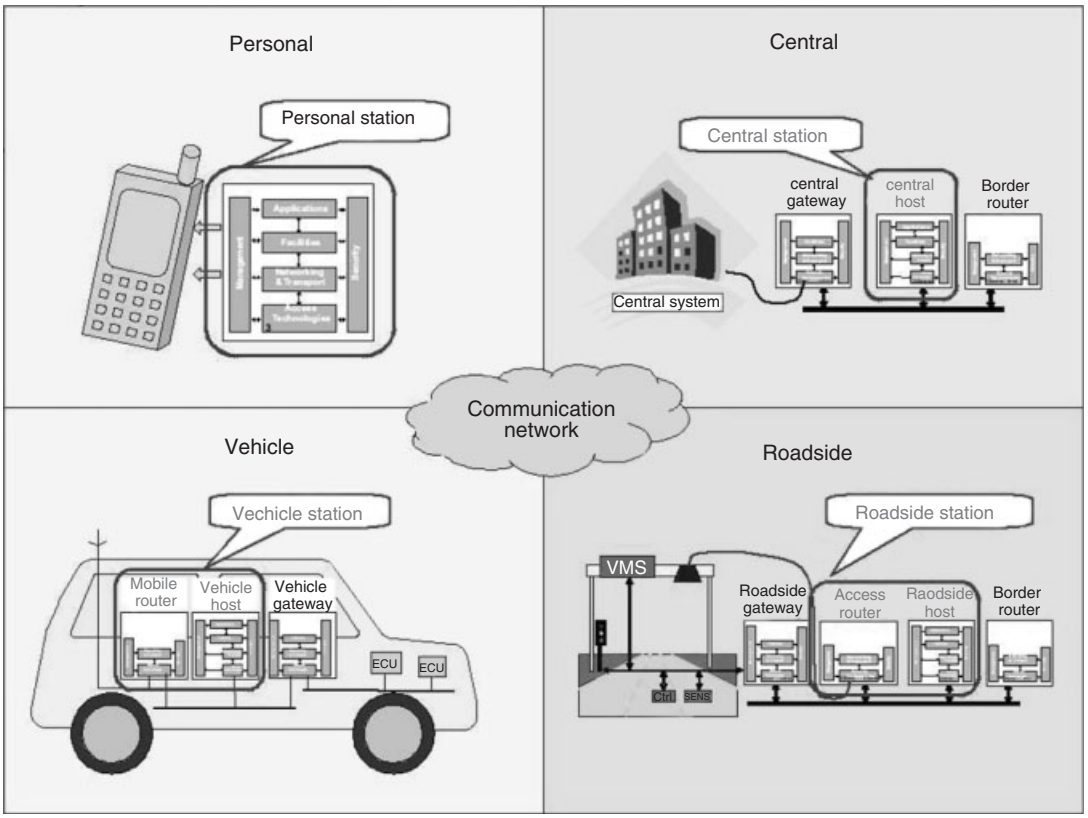


Figure 3. European ITS communication architecture subsystems. (Reproduced by permission of BMW Forschung und Technik GmbH on behalf of the COMeSafety Consortium.)

another, even when they are from competing vendors. They play an important role in bringing ITS products and services to market, lower prices for the users, and an integrated transportation system that appears seamless to the traveler.

International efforts to establish a standardized framework for ITS can be traced back to 1992 when the ISO convened Technical Committee 204 (ISO TC 204). In the same year, the Comité Européen de Normalization (CEN) established its technical committee on road transport and traffic telematics (CEN TC 278). The two committees are entirely devoted to ITS standards and work in a complementary way where possible. Japan and the United States each have their own national ITS standards programs that feed into the ISO standards consensus-building process.

There are different taxonomies of ITS standards. First and foremost, standards are needed in protocols and message sets to allow smooth data flow and information exchange among the subsystems. Protocols, such as TCP/IP for the Internet, give the formulae for passing messages, specify the details of message formats, and describe how to handle error conditions. Standard message sets, usually defined in data dictionaries, are also needed to allow accurate exchange of information between subsystems. For example, information exchange related to traffic incidents will need a coding standard for unambiguous description of the incident location (e.g., a road segment number) and the type of incident (e.g., fire and body injury). If wireless communication is needed, standardization of the frequency and modulation technique is also implied.

ITS standards may be established at the local, regional, national, international, and global levels. Some standards may be needed only up to a certain level. For example, for most commercial vehicle operations, international standards may be needed for a given continent (e.g. Europe or North America) but global standards are not required as trucks generally do not travel across continents. In contrast, standards for cargo identification should be global in order to facilitate freight identification, security checks, and movement between continents.

Like any other type of standards, ITS standards may be *de facto* standards, when everyone follows the standards set by the dominant supplier, or *consensus* standards, which are arrived through procedures established by standard setting organizations such as the ISO. Standards may also be set by government in the form of a regulation, usually as a last resort.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For the rapidly developing field of ITS, the timing of standard setting is important. Premature standards may stifle innovation. Standard setting often runs into difficulties with equipment and software suppliers who see a commercial advantage in promoting their own proprietary standards or who have an established market position and do not wish to change their products. Even when standards are agreed, practical considerations require an acceptable migration path for existing systems to move toward the new standards over a reasonable period of time. Users who have already invested in specific systems may be reluctant to switch to new standards before they have realized a reasonable return on their investment.

Those needing further information on ITS standards are referred to the *comprehensive reference book* on this subject based on the work of the ISO technical committee on ITS, TC 204 (Williams, 2008). The Japan Society of Automotive Engineers produces a very helpful introduction to the work of the committee in English (Japan SAE, 2011).

5 CONCLUSIONS

A great deal has been achieved in the two decades since ITS first emerged out of the laboratory, off the test track, and on to the open road, but much remains to be done. The intelligent algorithms and automated systems that were slow to develop in the early days of ITS are more of a reality today but the institutional challenges of ITS loom just as large as they ever did. Technology itself is not a barrier. Progress depends on the business case for ITS services and establishing operations with sufficient flexibility to cope with rising user expectations as well as technological change.

In 2005, returning to the theme of advanced highway systems, the United States launched the Vehicle Infrastructure Integration (VII) initiative bringing together a coalition of Federal and State departments of transportation and automobile manufacturers. The objective was to evaluate the technical, economic, and social/political feasibility of deploying vehicle-to-vehicle and vehicle-to-roadside communications dedicated to ITS services. This is work still in progress. What is proposed is a "connected transportation environment" through vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications and applications. The technological, institutional, and policy challenges associated with moving from research prototypes to full-scale deployment are now being considered (USDOT, 2012b). The intention is to finalizing by late 2013 the initial set of V2V technologies and safety applications.

The European Union has acted to put in place a legal framework of cooperation between the 27 EU countries to support the coordinated deployment and use of ITS. In future, a limited group of ITS applications will have to comply with European specifications made under the

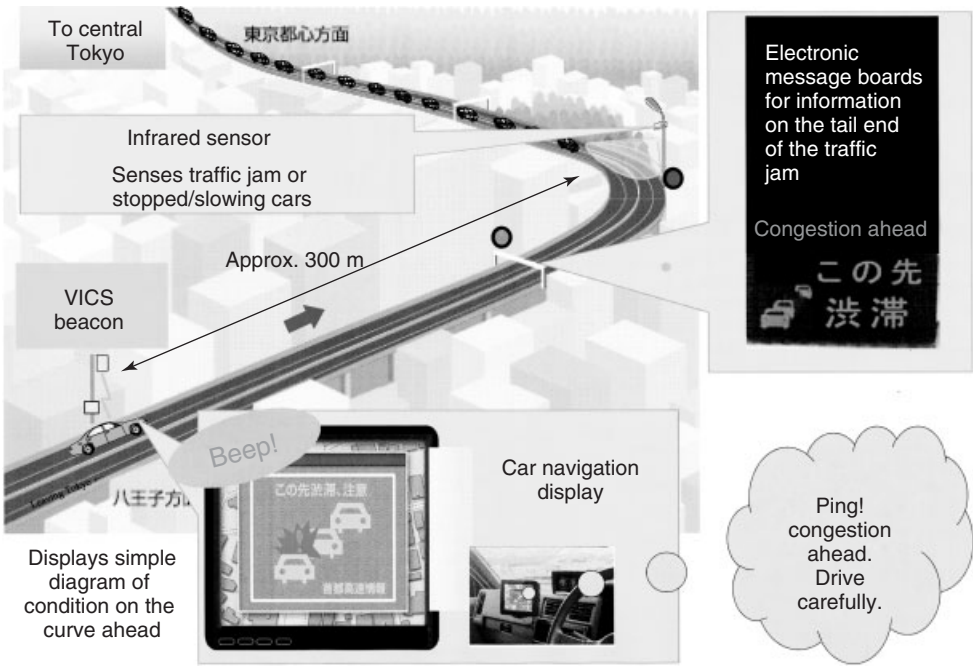


Figure 4. Japanese "Smartway" blind corner warning system, Tokyo. (Reproduced by permission of the National Institute of Land and Infrastructure Management, Japan.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

so-called *ITS Directive* (European Parliament and Council, 2010). The European Standards Organizations, CEN and ETSI, will develop the necessary standards to support this legal framework, in particular to enable interoperability of cooperative ITS (European Commission, 2011).

Japan has developed its own comprehensive package for ITS deployment branded as *Smartway* based on a communications architecture that *inter alia* uses the infrastructure installed for electronic toll collection (ETC) on the expressways. In addition to ETC, there is multiple functionality through Smartway communications with the vehicle: multipurpose payment for parking and other fees; probe data collection from cars; systems for road management and operation; systems providing drivers with information and an internet connection; bus location systems, etc. (MLIT, 2007). By way of example, Figure 4 shows a blind curve warning device that has been installed on the Tokyo metropolitan expressway to provide voice warning of queue ahead to the driver. The Smartway system also has potential for other applications such as communications for vehicle diagnostic applications, drive-through establishments, entry/exit control, and payment on ferries.

Future vehicle control systems will give increased driver support and greater automatic control but will that create a dependency that would be crippling in the event of failure? Will constant monitoring of vehicle component condition, driver condition and driver performance affect a step change in road safety toward the ultimate goal of Vision Zero? The fourth-generation mobile phone networks have ultrafast broadband capability and Internet Protocol (IP)-based communications. How will this affect the data and information we send and receive? What businesses and business models will be able to feed off the mobile internet and exploit the potential of connected vehicles? The answers to these and many other questions will keep ITS developers on their toes for many years to come.

ACKNOWLEDGMENTS

Parts of this chapter draw on material included in the PIARC Handbook on Intelligent Transport Systems second Edition (Chen and Miles, 2004). Material is reproduced with permission from the World Road Association. The author is grateful for the advice of colleagues from the Transport Associates' Network (www.transport-associates.net) during the preparation of this chapter.

RELATED ARTICLES

Intelligent Transport Systems: Market and Policies
Evolution and Future Trends
Technologies—Communication: Mobile
Technologies—Communication: Wireless LAN-based Vehicular Communication
Technologies—Communication: Broadcast
Technologies—Positioning: GNSS
Technologies—Positioning: Optical positioning
Tracking and Navigation for Goods and People
In-vehicle sensors
Data Acquisition by Roadside Detection
Technologies—Data Acquisition: Data fusion
Applications—Intelligent Vehicles: Driver Information
Driver Assistance
Applications—Intelligent Vehicles: Autonomous Vehicles
Applications—Intelligent Roads and Cooperative Systems: Urban Traffic Management
Advanced Highway Management Systems
Road Traffic and Travel Information (RTTI)
Logistics and Fleet Management
Tolling, Mobility Pricing
Applications—Further Applications: Emergency Services, eCall
Quality Management in ITS
IT-Security for Communicating Automotive Systems
Driver Distraction
Conclusions, Visions, Upcoming Standards

REFERENCES

Belinova, Z., Bures, P., and Jesty, P. (2010) Intelligent Transport System architecture different approaches and future trends. in Data and Mobility Advances in Intelligent and Soft Computing, vol. 81 (eds J. Düh, H. Hufnagl, and E. Juritsch), Springer, Heidelberg, pp. 115–125.

Bishop, R. (2001) *Whatever Happened to Automated Highway Systems (AHS)?* http://faculty.washington.edu/jbs/itrans/bishopahs.htm (accessed 12 October 2013).

Catling, I., Miles, J.C., and Harris, R. (1999) ITS in Europe. in *Advances in Mobile Information Systems* (ed J. Walker), Artech House, Boston, London, pp. 311–336.

Chen, K. and Costantino, J. (1999) ITS in the United States. in *Advances in Mobile Information Systems* (ed J. Walker), Artech House, Boston, London.

Chen, K. and Miles, J.C. (2004) *PIARC ITS Handbook*, 2nd edn, Route 2 Market, Swanley Kent, UK on behalf of PIARC, World Road Association, Paris.

COMeSafety Consortium (2008) *European ITS Communication Architecture Overall Framework*, BMW Forschung und Technik GmbH, Munich.

Davis, P., Ayland, N., Hill, C., Rutherford, S., Hallenbeck, M.E., and Ulberg, C.G., (1991) *Assessment of Advanced Technologies for Relieving Urban Traffic Congestion: NCHRP Report 340*, Transportation Research Board, Washington D.C.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

IVHS America (1992) *Strategic plan for intelligent vehicle-highway systems in the United States*. Report No: IVHS-AMER-92-3. IVHS America, Washington, D.C.

ETSI (2010) Intelligent Transport Systems (ITS) Communications Architecture. EN 302 665 V1.1.1 European Telecommunications Standards Institute, Sophia Antipolis, France.

European Commission (1991) Advanced road transport telematics. *Proceedings of the DRIVE Conference*, Elsevier, Amsterdam, New York, Oxford, Tokyo.

European Commission (1997) Community strategy and framework for the deployment of road transport telematics in Europe and proposals for initial actions. COM(97) 223 final, European Commission publication, Brussels.

European Commission (2011) *Intelligent Transport Systems in Action*, Publications Office of the European Union, Luxembourg.

European Parliament and Council (2010) Directive 2010/40/EU 7 July 2010, Official Journal of the European Union L207/1.

FRAME Architecture Consortium (2011) European Intelligent Transport System (ITS) Framework Architecture, http://www.frame-online.net/ (accessed 12 October 2013).

Greenough, J.C. and Kelman, L. (1994) *Metro Toronto SCOOT: Traffic Adaptive Control Operation*, Urban Traffic Engineers Council, Institute of Transportation Engineers, June 1994, Washington, D.C.

Harvey, S. (1999) ITS in Japan. in *Advances in Mobile Information Systems* (ed J. Walker), Artech House, Boston, London, pp. 337–365.

IBEC (2013) International Benefits, Evaluation and Costs (IBEC) Working Group, https://sites.google.com/site/ibecits/ (accessed 12 October 2013).

ISO Technical Committee 204 (2007) Intelligent transport systems—reference model architecture(s) for the ITS sector, Part 1: ITS service domains, service groups and services. Standard TS14813-1:2007, International Standards Organization, Geneva.

Japan SAE (2011) *Standard Developing Activities of ISO/TC204*, Japan Society of Automotive Engineers, Tokyo. http://www.itsa.org/industryforums/isotc204/isotc204-public-documents (accessed 12 October 2013.)

Kawashima, H. (1994) Overview of Japanese development and future issues. in *Advances in Mobile Information Systems* (ed J. Walker), Artech House, Boston, London, pp. 289–314.

Miles, J.C., Chen, K., White, C., Johnson, W., Abdulhai, B., Dussutour, I., Morello, S., Rupprecht, S., Hopkin, J., Harris, R., Zografas, K., Crompton, P., and Catling, I. (2002) The ATLANTIC Network of Excellence, http://www.ankerbold.co.uk/projects/atlantic-network/ (accessed 12 October 2013).

Miles, J.C. and Walker, A.J. (2006) The potential application of artificial intelligence in transport. *IEE Proceedings Intelligent Transport Systems*, **154** (3), 183–198.

MLIT (2002) *ITS Handbook Japan 2002-2003*, Road Bureau, Japan Ministry of Land, Infrastructure and Transport, Tokyo. http://www.mlit.go.jp/road/ITS/index/indexHBook.html (accessed 12 October 2013.)

MLIT (2007) *ITS Introduction Guide: Shift from Legacy Systems to Smartway*, Japan Ministry of Land, Infrastructure and Transport, Tokyo.

NAHS Review Committee (1998) National Automated Highway System Research Program—a Review. Transportation Research Board Special Report 253, National Academy Press, Washington, DC.

PIARC (1999) *ITS Handbook 2000: Recommendations from the World Road Association (PIARC)*, Artech House, Boston, London.

PIARC (2007 and 2011) Road Network Operations, http://road-network-operations.piarc.org/ (accessed 12 October 2013).

Schulze, M., Schwarz, J., Sonntag, J., Benz, T., Braun, A., and Ulken, U., (1998) CHAUFFEUR—TR 1009 Report on Safety Analysis of System Components and Hazard Analysis, http://cordis.europa.eu/telematics/tap_transport/research/projects/chauffeur.html (accessed 12 October 2013).

Sodeikat, H. (1994) Cooperative transport management with EURO-SCOUT. in *Advances in Mobile Information Systems* (ed J. Walker), Artech House, Boston, London, pp. 65–78.

Sussman, J. (2004) *What We Know Now That We Wish We Knew Then About Intelligent Transportation Systems: A Retrospective on the ITS 1992 Strategic PlanTransportation Research Record*, National Academy Press, Washington, DC.

USDOT (1996) TravTek global evaluation and executive summary. Publication No. FHWA-RD-96-031, Turner-Fairbank Highway Research Center, March 1996.

USDOT (2012a) The National ITS Architecture, http://www.its.dot.gov/arch/ and http://www.iteris.com/itsarch/ (accessed 12 October 2013).

USDOT (2012b) Transforming Transportation through Connectivity *FHWA-JPO-12-019 ITS Strategic Research Plan, 2010-2014 Progress Update 2012*. US Department of Transportation Research and Innovative Technology Administration, Washington DC.

USDOT (2013) Intelligent Transportation Systems Joint Program Office Knowledge Resources, http://www.itsbenefits.its.dot.gov/ and http://www.itscosts.its.dot.gov/ (accessed 12 October 2013).

Williams, B. (2008) *Intelligent Transport Systems Standards*, Artech House, Boston, London.

FURTHER READING

Catling, I. (1994) *Advanced Technology for Road Transport: IVHS and ATT*, Artech House, Boston – London.

Chen, K. and Miles, J.C. (2004) *PIARC ITS Handbook*, 2nd edn, Route 2 Market, Swanley Kent, UK on behalf of PIARC, World Road Association, Paris.

Hepworth, M. and Ducatel, K. (1992) *Transport in the Information Age: Wheels and Wires*, Belhaven Press, London, New York.

McQueen, B. and McQueen, J. (1999) *Intelligent Transportation Systems Architecture*, Artech House, Boston, London.

Walker, J. (1998) *Advances in Mobile Information Systems*, Artech House, Boston, London.

World Road Association (PIARC) and FISITA (2012) *The Connected Vehicle, PIARC Report 2012ROEN, World Road Association*, World Road Association, Paris.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto166
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Intelligent Transport Systems: Markets and Policies

Steven E. Shladover
University of California, Berkeley, Richmond, CA, USA

1 Introduction 1
2 Benefits of ITS Deployment 1
3 Constraints to ITS Deployment 3
4 Regional Contrasts in ITS Deployment 5
5 Future Directions in ITS 7
Abbreviations 8
References 8

1 INTRODUCTION

Since their introduction more than a century ago, road vehicles have had minimal direct interactions with each other or with the roadway infrastructure on which they operate. Indeed, virtually the only interaction between the vehicles and the infrastructure was the physical connection at the tire contact patch, and the only direct vehicle–vehicle interaction was unintentional, when they crashed into each other.

With the advent of intelligent transport systems (ITS), active exchange of data among vehicles and between the vehicles and the infrastructure assumes central importance. Indeed, not only the vehicles and the infrastructure but also the people and goods being transported in the vehicles become elements of an integrated road transport system, connected by the exchange of information, as shown in Figure 1. It is well known in systems science that it is possible to achieve much higher performance when all the relevant subsystems of a complex system of systems are explicitly treated in analysis and optimization. ITS therefore offers great opportunities to improve all the major measures of effectiveness of road transport systems: mobility, efficiency, safety, and costs (capital and operating, for both the vehicles and the infrastructure).

The concept of integrating vehicles and the infrastructure is straightforward on its face, but the implementation is challenging for a variety of reasons that will be explained in this chapter. These ITS have developed in different directions and at different rates in Europe, the Americas, and the Asia/Pacific region, based on regional differences in land development, transportation system characteristics, government structures, and the market for private vehicles.

In the remaining sections of this chapter, the opportunities and challenges associated with ITS deployment are explained. The opportunities are represented by the benefits that ITS can provide to the transportation system, the motivation for vehicle and infrastructure owners and operators to invest in the ITS market. The challenges represent a combination of institutional, technical, and economic constraints on the ability of these actors to act in favor of ITS solutions. The contrasts among the major economic regions of the world with regard to ITS deployment are then explained.

2 BENEFITS OF ITS DEPLOYMENT

ITS is a broad collection of information technology applications that can confer a wide range of benefits to the transport system and its users. Different ITS applications provide benefits in different categories, so no single ITS application or product provides the full suite of benefits.

2.1 Mobility enhancements

The original motivations for ITS were primarily associated with mobility improvements. The heavily industrialized

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto167
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

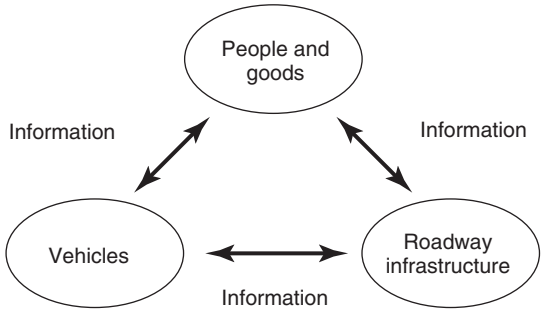


Figure 1. ITS enabling road transport to function as an integrated system.

and urbanized countries recognized in the 1980s that they could not afford to continue to expand the civil infrastructure of their roadway networks at a fast enough rate to stay ahead of the growing demand for road transport services, so they would need to use that infrastructure more efficiently. The information technology of ITS makes this possible in several ways.

2.1.1 Relieving traffic congestion

Traffic management systems on freeways can help the traffic flow more smoothly and at a higher capacity by metering the entry of vehicles at the on-ramps to avert traffic flow breakdowns or advising drivers to reduce speeds upstream of bottlenecks. Real-time traveler information systems can help travelers adjust their travel times and routes away from the worst congestion. Intelligent traffic signal control systems can help urban traffic flow more smoothly, minimizing wasted green signal time and spillback of queues to upstream intersections. Enhanced detection of incidents can enable emergency responders to reach the sites of crashes more quickly, so that the impediments to traffic can be removed and full capacity restored sooner. Integrated freeway/arterial corridor management systems can promote the most efficient use of the available roadway capacity in a corridor, especially when an incident temporarily reduces the capacity of the freeway or the arterial, promoting the shifting of traffic to the parallel facility. In the longer term, automated driving of vehicles in close-formation platoons will make it possible for each lane of a limited-access highway to carry two to three times the traffic volume (Michael *et al.*, 1998) of today's conventional highway lanes.

2.1.2 Improving public transport service quality

Public transport systems have had competitive disadvantages relative to private automobiles in all but the highest density urban locations, limiting their opportunities to attract drivers out of their cars. These can be mitigated using ITS. Transit signal priority systems enable buses to gain priority for green signal time over other classes of vehicles, providing some travel time savings (Smith, Hemily, and Ivanovic, 2005). Multimodal traveler information systems can provide potential transit riders with comprehensive information to relieve anxiety about bus arrival times and connections, so that they will be more comfortable depending on a public transport system. Fleet management systems enable bus fleet operators to make more efficient use of their assets, so that they can be more economically viable. Automatic guidance and precision docking systems enable buses to provide a quality of service to their passengers that would otherwise only be achievable on much more costly rail transit systems and to operate on busways that occupy narrower rights of way than conventional highway lanes, so that they can be implemented more easily in crowded urban environments (Shladover, 2000). In the longer term, automated buses or automated guideway transit systems can save driver labor costs, improving the economic viability of transit services.

2.1.3 Facilitating goods movement

ITS can provide specialized support to address the particular needs of truck operators, so that they can make their operations more efficient. Real-time routing advice can incorporate restrictions specific to trucks, such as limited overhead clearances and limited weight-bearing capacity of bridges. Trucks can be given priority at traffic signals, so that they are less likely to have to stop and restart than light-duty vehicles, saving time, energy, and emissions. By means of data communications with the roadside, the regulatory processes specific to truck operations can be executed without requiring the trucks to stop at every weigh station or border crossing, saving even more time, energy, and emissions. With accurate information about road grades and curves, truck operators can be given advice about the best speed profiles to follow to minimize energy consumption on each trip and to explicitly optimize the trade-offs among energy use, emissions, and travel time.

2.2 Efficiency improvements

Efficiency and mobility improvements are closely related to each other, such that improvements in one are likely to have similar benefits for the other. However, some additional efficiency improvements do not necessarily

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto167
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

affect mobility directly. For example, ITS can smooth out the transient disturbances in traffic, reducing acceleration and deceleration maneuvers, thereby saving energy and emissions (Shladover, 2011). Using ITS technologies to enable vehicles to automatically drive safely at shorter than normal gaps, aerodynamic drag can be reduced, saving significant energy in higher speed driving.

2.3 Safety improvements

Many ITS applications can improve traffic safety by warning drivers about unsafe conditions, by assisting drivers to control their vehicles more accurately and reliably, or by taking over the driving responsibilities. Although most of these applications are vehicle centered, some are based on the roadway infrastructure or involve cooperation between the vehicles and the roadway.

2.3.1 Infrastructure-based safety systems

Detection systems installed in or near the roadway can detect the presence and movements of vehicles and vulnerable road users such as bicyclists or pedestrians. On the basis of the detection information, traffic signal changes can be adjusted to minimize potential conflicts, for example, when a vehicle is predicted to run through a red signal (Zhang *et al.*, 2012). Detected hazards can be brought to drivers' attention by variable message signs, displaying information such as variable speed limits to slow traffic approaching the upstream end of a congestion queue.

2.3.2 Vehicle-based safety systems

Sensors on vehicles can detect many driving safety hazards, such as imminent lane departures, or vehicles or other obstacles in the lane ahead; vehicles in the blind spot; the proximity of obstacles at parking spaces; or driving speeds that exceed the local speed limits or the ability of the vehicle to maintain safety. Audible, visual, or haptic displays can convey this hazard information to the driver, so that he or she can take appropriate corrective action.

2.3.3 Cooperative safety systems

The latest generation of safety systems is cooperative, using wireless communications with the vehicle to take advantage of data sources in any other vehicle- or roadway-based sensor to warn the driver about an imminent unsafe condition. A wireless local area network can provide the drivers of all vehicles with the same information about all detected hazards (Sengupta *et al.*, 2007).

2.4 Cost reductions

Although ITS technologies are not free, in some cases investments in ITS can produce savings in infrastructure construction costs or vehicle operating costs. Most dramatically, when ITS operational improvements increase the effective capacity of a transport facility, they help avoid the costs of physical expansions to the facility, which would be substantially more expensive.

3 CONSTRAINTS TO ITS DEPLOYMENT

Although a few of the ITS technologies fit the typical automotive industry model of entirely in-vehicle systems, most of them involve significant interactions with the roadway infrastructure and/or other vehicles, which need to be equipped with compatible technology. This makes implementation of ITS technologies generally more complicated from an institutional and policy perspective than implementation of other automotive technologies, which can typically be done as market-based decisions between vehicle developers and their private market customers.

3.1 Challenges of public–private collaboration

As the roadway infrastructure is generally the responsibility of public agencies while vehicles are bought and sold in the private market and operated by private individuals, systems that integrate the vehicles with the infrastructure require collaboration between the public and private sector domains. Neither sector can make its decisions in isolation from the other, because the viability of the system as a whole depends on ensuring technical compatibility between the vehicle and roadway elements and their simultaneous availability in the same locations.

The public and private sectors need to work together from the start on developing the ITS operational concepts and designs, to ensure that the needs and constraints of both sectors are satisfied. This should extend to the development of technical standards for compatibility and interoperability of the vehicle and infrastructure elements. They also need to cooperate on implementation plans, so that the rollouts of the vehicle and infrastructure systems are coordinated.

These types of collaboration are challenging because of the cultural differences between public and private sector organizations in most countries. They tend to have different priorities, different economic interests, different decision-making processes and timescales, and different incentive structures. If they do not have some significant history of collaboration in other endeavors, they may not even know

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto167
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

how to work together and may lack the mutual understanding and trust needed to form effective partnerships.

3.2 Chicken-and-egg dilemma of vehicle–infrastructure cooperation

The need for close public–private cooperation is most acute for cooperative systems that require vehicle and infrastructure elements to work together. These systems, despite their great potential benefits, are hampered by a classic "chicken-and-egg" dilemma of which element goes first. Can the public sector justify deploying its cooperative infrastructure elements before there are any vehicles on the market that can interact with them? Can a private vehicle manufacturer or purchaser justify investments in new vehicle-based devices before the public sector has deployed the needed cooperative infrastructure? The answers are generally *no* in both cases, which means that they need to work together on a coordinated deployment strategy. This strategy, despite the best of intentions, can still be undermined by surprises that prevent either sector from following through on its commitment to the other.

3.3 Incompatible decision horizons of infrastructure, vehicle, and information technology industries

Public infrastructure agencies are accustomed to designing and building systems such as bridges and road networks that are expected to have usable lifetimes of multiple decades. Vehicle manufacturers design and build their vehicles to last for multiple years, but information technology companies expect their devices to become obsolete in months. Each of these organizations has decision processes and business models suitable for its respective product's functional lifetime, but with ITS, they are required to work together on a system that combines all three types of element. The ITS, as a composite of elements with lifetimes of months, years, and decades, can become an awkward hybrid that is difficult for all three industries to grasp and manage.

3.4 Introduction of information technology into more traditional capital-intensive industries

There are few industries more capital intensive or more focused on development of capital facilities than the roadway infrastructure industry. The automotive industry is also very capital intensive because of the size and complexity of the facilities needed to build motor vehicles. By contrast, the information technology industry, with its emphasis on software and small devices, is much less capital intensive and therefore more agile in adapting to change. This contrast affects organizational culture and economic models, making it more difficult for these industries to work together effectively.

3.4.1 Ratio of capital to operating and maintenance costs

When transport agencies develop new facilities, their dominant focus is on the capital costs and how to minimize them, because the annual operating costs are only a tiny fraction of the capital costs. With motor vehicles, the capital and operating costs are more evenly balanced, but with information technology, the operating and maintenance (including periodic update) costs are more likely to be dominant. This introduces tension into the design of an ITS that must be shared among the three industrial sectors, each with different economic models.

3.4.2 Development of understanding and comfort with new technologies

Public transport agencies have historically been dominated by civil engineers, whereas the automotive industry is dominated by mechanical engineers. ITS requires the integration of these technologies with electrical engineering and computer science, fields that are generally foreign to the leaders of the transport system's industries. They need to gain enough understanding of these newer technologies and develop sufficient staff expertise in these fields within their organizations to be able to manage and operate systems based on ITS technology.

3.5 Politicians' needs for fast implementation and ribbon-cutting ceremonies

The roadway infrastructure operates in the public political arena, where funds for construction, operation, and maintenance must be allocated by elected officials. These officials are always thinking about what will help them get re-elected at the next election, when they need to show the general public that they have accomplished something tangible by their investments of tax revenues. This leads them to favor construction projects, where the tangible results can be shown very publicly in ribbon-cutting ceremonies and structures can be named in their honor. It is more difficult to gain the support of politicians and the general public for ITS operational improvements, which are not nearly as visible. Improved operations of traffic signal systems or freeway ramp metering to reduce travel delays are much harder to explain and demonstrate to the general public,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto167
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

even though they are likely to produce significantly higher ratios of benefits to costs.

3.6 Limited capital and staff resources to invest in new initiatives

Public investment decisions are based not only on economic valuation criteria such as the ratio of benefits to costs (or net present value) but in many cases they are severely constrained by resource limitations. Even though a new ITS project may be expected to be highly beneficial and cost effective, it cannot necessarily be funded because there may be no funding available for new initiatives. If the existing transport infrastructure is deteriorating and needs extensive maintenance to prevent further deterioration and to keep it in operational condition, that is likely to take priority over any new initiative.

4 REGIONAL CONTRASTS IN ITS DEPLOYMENT

The degree of ITS deployment varies widely from country to country, even within the advanced industrialized countries. Many factors influence this propensity to deploy ITS, beyond national wealth and general technological sophistication.

4.1 Ease of public–private cooperation and attitudes toward industrial policy

Because of the need for public–private cooperation to implement vehicle-infrastructure cooperative ITS, as explained in Section 3.1, the ease of achieving that cooperation is a dominant factor in determining how widely ITS is deployed. The industrialized countries of east Asia (Japan, South Korea, Singapore, and now China) have particularly well-established cooperative relationships between their public sector agencies and large private companies. This makes it possible for them to agree on technical standards and coordinated investments to deploy both vehicle and infrastructure elements of ITS (Ezell, 2010). It is no coincidence that these are also the countries that are most intent on using public–private cooperation to enhance the international competitiveness of their industries, so that they can increase national wealth by exporting their products to other countries.

The European countries have lagged somewhat behind their Asian counterparts in the extent of their public–private cooperation on transport and economic development policies, but they remain far ahead of their North American counterparts. The American example is near the opposite extreme from the Asian example, because there is substantial mistrust between the public and private sectors, especially in the United States. They do not have a significant history of cooperation, and indeed, there is strong ideological opposition among a large part of the political establishment to industrial policy, by which the government would favor specific industries.

4.2 Centralization of public agency transport decision-making

Public agency decisions about transport investments and operations are handled in diverse ways in different countries. This has a major influence on the decisions about deployment of ITS. When the decisions are centralized in a national authority, it becomes possible for that authority to make a uniform decision for the entire country, so that both the technical characteristics and the scheduling of the deployment can be made consistent and predictable for the private industry partners. As the decisions become more distributed, there is a heightened risk of incompatible technical implementations and spotty scheduling, so that the availability of the ITS services may not be consistent for travelers and fleets who need to drive through multiple jurisdictions.

The east Asian countries have the most centralized transport investment decision-making processes. One national ministry can decide to deploy a specific technology or product throughout the country and can allocate the funding to ensure that it happens according to a well-specified schedule (by region or in a logical sequence from urban to rural or rural to urban). This provides the most predictable interface for private industry partners, who can plan for the introduction of the in-vehicle systems based on a known plan for the infrastructure systems.

The European countries have diverse approaches to transport investment decision-making, with some being more centralized (France) whereas others are more decentralized (Germany). Some countries have privately operated motorway networks, which add another dimension to the infrastructure decision-making.

The North Americans have the most decentralized transport decision-making processes, with the national governments having no direct responsibility for designing, building, owning, operating, or maintaining any of their roadway networks. The primary highway network is the responsibility of the states or provinces, whereas the secondary networks are under county and city governments. Within the United States alone, there are 50 states, 3000 counties, and more than 30,000 municipalities, which

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto167
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

mean that border crossings are frequent and unavoidable. Adjacent cities and counties may have very different priorities, resources, and capabilities, so the ITS deployments could vary greatly when crossing the invisible borders between these jurisdictions.

4.3 Private vehicle market willingness to pay

Automobile purchasers in different countries have different priorities with regard to the features that they are interested in having and how much they are willing to pay for them, so the vehicle manufacturers adapt their offerings to the local market interests. This applies as much to the in-vehicle ITS functions as it does to other features such as vehicle and engine sizes, transmission types, and navigation systems.

The Japanese and Korean markets tend to be the most inclined to favor new electronic gadgetry, even before the technology is mature and robust, so these markets have tended to be the earliest adopters of new ITS features. For example, the limited-edition Mitsubishi Diamante of 1995 was the first production vehicle to offer adaptive cruise control (but only in Japan). The European luxury car market, as well as the Japanese and Korean markets, has a substantial number of buyers who buy ITS features by choosing to purchase their vehicles "fully loaded" (with all options selected). Some of these buyers do not actually learn how to use all the features on their vehicles and may never use some of them. In each of these cases, the selection of a premium vehicle with premium options is more of a statement of personal pride and prestige than an explicit desire to use all the features. Other car buyers study the choice of options carefully and decide exactly which features they want to have.

The North American vehicle market tends to be more skeptical of new technology and extra-cost options. Car buyers are less inclined than their Asian or European counterparts to choose all the options, and they tend to be careful about how much they are adding to the cost of their vehicle. Furthermore, most new cars are purchased off the dealer's lot, where they have been ordered with prepackaged combinations of options. The ITS features, especially the collision warning and avoidance options, are generally only included in the "fully loaded" vehicles, so it can be hard to find them in moderately priced vehicles.

4.4 Public agency financial and technical resources

The attitudes toward public investments in transport infrastructure vary widely from country to country, which means that the resources available to the responsible public agencies vary similarly. In countries where improvements to transport infrastructure are regarded as important investments in future economic productivity, the resource allocations tend to be generous. Similarly, where transport infrastructure development and enhancement are seen as economic stimulus opportunities when economic activity is sluggish, governments are likely to maintain their transport infrastructure investments at a sustained level regardless of economic cycles.

Within the industrialized world, Japan and the United States represent the opposite ends of the spectrum. Japan assigns high priority to transport infrastructure for both long-term economic strength and near-term stimulus, whereas the United States is more likely to view this as an avoidable expense rather than an investment. Consequently, the average annual investment per kilometer of roadway infrastructure (from design to construction, operations and maintenance) at all levels of government is approximately four times as large in Japan as it is in the United States, even though the roads are more heavily utilized on average in the United States. The other Asian countries and Europe are between these two extremes, and China, with its rapid expansions to its roadway infrastructure, may well be exceeding Japan in its investments per kilometer of existing roadways.

The technical expertise available within the public agencies follows the same general trend as the financial investment. Where resources are ample, it is easier to hire people with a higher level of expertise and there is more of an incentive for highly capable and expert people to work in transport where they will have sufficient resources to do new and exciting things.

4.5 Traffic and roadway network conditions

ITS is not an end in itself but is a collection of technological tools that can be used to help solve transport problems. It is needed most urgently in the locations that have the worst transport efficiency, congestion, and safety problems but is not nearly as important in locations where traffic flows freely. Consequently, the levels of ITS deployment tend to be higher in the most congested urban areas and the locations with the highest density of heavy truck activity. The exception is the megacities of the developing world, which tend to have the worst transport problems of all but lack the financial and technological resources to implement ITS solutions. These are also the locations where the technical challenges are the greatest for ITS safety systems because of the chaotic traffic conditions and great diversity of road users.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto167
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.6 Motivations to save energy and CO_2 emissions

As discussed in Section 2.2, ITS can improve the operating efficiency of road vehicle operations, enabling savings of energy and CO_2 emissions (Shladover, 1993). This can be a strong motivation for deployment of ITS in countries that are seriously concerned about reducing energy use and CO_2 emissions, because of either high energy costs or strong commitments to the Kyoto Protocol on CO_2. ITS is a particularly attractive set of strategies for saving energy and CO_2 because it can provide these benefits without necessarily adding inconvenience or diminishing the quality of life.

Japan has created the Energy ITS research program to study these opportunities for CO_2 reductions (Horiguchi *et al.*, 2010), the European Commission has initiated several analogous programs under more than one of its Directorates General, and the United States has the Applications for the Environment: Real-Time Information Synthesis (AERIS) program to explore how ITS can save energy, CO_2, and criteria pollutants. The Japanese and European programs started earlier and have had larger funding, reflecting the stronger commitments that their governments have made to implement the Kyoto Protocol.

5 FUTURE DIRECTIONS IN ITS

ITS changes rapidly, based on the rate of change in the information technologies on which it is based. Within the first decade of the twenty-first century, the dramatic changes in wireless technologies, smart phones, and the Internet have already had significant impacts on ITS and its use in automotive vehicles. Most of the major international automotive companies have established technical offices in California's Silicon Valley in order to get closer to the source of the newest innovations and to understand the next generations of technology as they are developed. In some quarters, the vehicle is already being seen as little more than a package for delivery of future information technologies. The younger generation is less interested in driving than in social networking and Internet use, to the extent that they would rather be a connected passenger than an unconnected driver. These trends have profound implications for the future of the automobile.

Two of the major technological trends in ITS that will affect the automotive future are the growth in connected vehicle technologies and applications and in-vehicle automation.

5.1 Connected vehicles

Although drivers now take it for granted that they can have a voice connection with people outside their vehicle at any time using mobile phones, the more significant growth will be in wireless data connections to and from vehicles. These may or may not involve the driver directly and may connect with other vehicles or with roadside or centralized infrastructure. Many of the ITS services depend on effective wireless communications with vehicles, but the technology of mobile wireless communications, combined with accurate positioning, is not yet fully mature.

Cooperative collision warning systems (Sengupta *et al.*, 2007) are based on the notion that each vehicle can identify its own position, heading angle, and speed very accurately and broadcast that information to all other vehicles in its immediate vicinity (a few hundred meters). On the basis of these broadcasts, each vehicle can know the state of all of its neighbors and can use that knowledge to predict potential collisions or other hazards and then warn its driver. This concept depends on highly reliable and low latency local communications (using an ad hoc wireless local area network formed using a technology known as *dedicated short-range communication*, DSRC), combined with very accurate positioning. The positioning needs to be within a meter accuracy in order to correctly associate vehicles with the correct lanes, but this accuracy stretches the capabilities of current Global Navigation Satellite Systems (GNSSs) in all but their most costly incarnations.

DSRC can also be used to support vehicle-infrastructure cooperative safety systems for intersection collision warnings and can enable each vehicle to serve as a traffic data probe (Shladover and Kuhn, 2008). The traffic data probe functions and other non-time-critical services such as medium-range safety warnings, traveler information, and traffic management can be implemented using the fourth-generation (4G) cellular data technologies that are already on the market and growing rapidly. As these wireless technologies become more pervasive and affordable, newer connected vehicle applications will be conceived and developed to provide mobile services that have not even been imagined yet.

5.2 Vehicle automation

Until now, road vehicles have always been operated under the direct control of a human driver. As long ago as the 1930s, futurists were already conceiving the notion of an automated vehicle that could be driven without human intervention. Research on this topic has been conducted

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto167
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

intermittently since the late 1940s (using vacuum tube analog electronics at that time), motivated by the opportunity to overcome some of the limitations of human drivers (Shladover, 1995). Under automatic control, vehicles could be driven more accurately and closer together to increase highway capacity and their speed profiles could be smoothed out to reduce energy use and pollutant emissions. As the automated control systems do not get fatigued or inattentive, there is also a potential to reduce crashes, although there are still significant challenges in software safety and fault tolerant control in order to ensure that automation failures do not create new categories of crashes (Shladover, 1999).

The enabling technologies for vehicle automation have been advancing and becoming more affordable based on the market introduction of other vehicle capabilities (ranging sensors for collision warning and adaptive cruise control systems, electromechanical steering and brake actuators for electric and hybrid vehicles). The interest of the public and the industry has been piqued by the autonomous automated vehicle demonstrations sponsored by the Defense Advanced Research Projects Agency (DARPA) and the entry of Google into this field. Ultimately, safe and high performance automated vehicles are likely to require combining the sophisticated sensing and software technologies of the autonomous automated vehicles with the connected vehicle wireless technologies, to produce cooperative automated vehicles that are capable not only of perceiving their surroundings but also of communicating actively with their peers and with their host infrastructure (Shladover, 2009).

ABBREVIATIONS

DSRC	dedicated short-range communications
GNSS	Global Navigation Satellite System
ITS	intelligent transport system

REFERENCES

Ezell, S. (2010) *Explaining International IT Application Leadership: Intelligent Transportation Systems*, Information Technology and Innovation Foundation, Washington, DC, 58 pp.

Horiguchi, R., Hanabusa, H., Kuwahara, M., *et al*. (2010) *Validation Scheme for Traffic Simulation to Estimate Environmental Impacts in Energy-ITS Project*. 17th ITS World Congress, Busan, Korea.

Michael, J.B., Godbole, D., Lygeros, J., and Sengupta, R. (1998) Capacity analysis of traffic flow over a single-lane automated highway system. *Intelligent Transportation Systems Journal*, **4**, 49–80.

Sengupta, R., Rezaei, S., Shladover, S.E., *et al*. (2007) Cooperative Collision warning systems: concept definition and experimental implementation. *Journal of Intelligent Transportation Systems*, **11** (3), 143–155.

Shladover, S.E. (1993) Potential Contributions of Intelligent Vehicle/Highway Systems (IVHS) To Reducing Transportation's Greenhouse Gas Production. *Transportation Research*, **27A** (3), 207–216.

Shladover, S.E. (1995) Review of the state of development of advanced vehicle control systems (AVCS). *Vehicle System Dynamics*, **24**, 551–595.

Shladover, S.E. (1999) Why We Should Develop a Truly Automated Highway System. *Transportation Research Record No. 1651*, Transportation Research Board, Washington, DC.

Shladover, S.E. (2000) Bus rapid transit and automation—opportunities for synergy. *Proceedings of Seventh World Congress on Intelligent Transport Systems*, Turin, Italy, November.

Shladover, S.E. (2009) Cooperative (rather than autonomous) vehicle-highway automation systems. *IEEE Intelligent Transportation Systems Magazine*, **1** (1), 10–19.

Shladover, S.E. (2011) Challenges to evaluation of CO_2 impacts of intelligent transportation systems. *IEEE Forum on Integrated Sustainable Transportation Systems (FISTS)*, Vienna, Austria, June.

Shladover, S.E. and Kuhn, T.M. (2008) Traffic Probe Data Processing for Full-scale Deployment of Vehicle-infrastructure Integration. Transportation Research Record No. 2086, Transportation Research Board, Washington DC, pp. 115–123.

Smith, H.R., Hemily, B., and Ivanovic, M. (2005) Transit Signal Priority—A Planning and Implementation Handbook, Intelligent Transportation Society of America, Washington, DC, 212 pp.

Zhang, L., Wang, L., Zhou, K., and Zhang, W.-B. (2012) Dynamic all-red extension at a signalized intersection: a framework of probabilistic modeling and performance evaluation. *IEEE Transactions on Intelligent Transportation Systems*, **13** (1), 166–179.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto167
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Evolution and Future Trends

André J. Vits
Directorate-General Information Society and Media, European Commission, Brussels, Belgium

1 Introduction 1
2 Navigation/Telematic Services 3
3 Traffic Management 7
4 Future Trends 9
5 Conclusions 9
Abbreviations 10
Endnotes 10
References 10
Further Reading 10

1 INTRODUCTION

The development over recent few decades of information and telecommunication technologies (ICT) has changed dramatically the transportation sector and how it operates.

The step from mainframe computers to microprocessor technology, personal computer, and handheld devices has provided the transportation and traffic engineering community with an almost unlimited toolset for on-site real-time processing of traffic data, complex modeling, and visualization of results in a highly interactive way, image processing, on-board processing units and network architectures, visualization devices, and so on.

Communication technology development became the second pillar on which basic services could be build. Mobile handsets (GSM, Global System for Mobile Communications) became not only the main voice communication channel but is today the main carrier for data communication (3G and 4G GSM generations). The necessary infrastructure for new "telematic" services became available to a wide range of potential service providers.

The demand for mobile internet has been the driver for the development of new communication protocols, such as WiFi and WiMax.

But also in other frequency bands, new opportunities became available: 5.8-GHz band became the carrier for automatic debiting, 24-GHz band for short-range radar applications (automatic cruise control), and is moving to the use of the 79-GHz band. In line with the developments worldwide, the 5.9-GHz band has been made available in Europe for (safety-related) vehicle-to-vehicle (V2V) and more general V2X communication.

While development during recent decades has been mainly driven by safety and system performance on the one side and market competition on the other side, energy efficiency and climate change have brought new challenges to be addressed. However, in all cases, it has become obvious that information technology and telecommunication services will remain the key drivers in vehicle design and the management of the transportation systems.

In the following discussions, a few aspects are elaborated that demonstrate the impact that these developments had and how this will determine our future transportation systems.

1.1 Safety and energy efficiency

Looking at the impact that ICT had, and still has, on innovations in the automotive sector, one has to recognize the importance of national and international research programs in this domain. In Europe, the (institutionalized) cooperation between original equipment manufacturers (OEM),

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto168
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

suppliers, and research institutes has boosted the development of new and performing equipment and has helped to keep the European automotive industry competitive on the global markets.

1.1.1 Setting the scene

Perhaps the first comprehensive European initiatives of the kind have been the PROMETHEUS project (Program for European traffic with highest efficiency and unprecedented safety) under the EUREKA (http://www.eurekanetwork.org/) framework.

EUREKA was created in 1985 under a multilateral agreement between European countries, with the objective to stimulate research initiatives from industry and research institutes in the participating countries. The countries committed themselves, based on a cost-shared model, to finance the efforts of their own organization(s) according to the research priorities agreed between the State Heads.

PROMETHEUS (Glathe, 1993) has probably been the most prominent from the early transport projects. The project led by the European automotive industry [BMW, Daimler, Porsche, Volkswagen, and MAN (Germany), Fiat and Matra (Italy), PSA and Renault (France), Saab and Volvo (Sweden), and Jaguar and Rolls-Royce (the United Kingdom)] involved many suppliers and research institutes.

PROMETHEUS can be considered to be the project that has set the vision for the next decades. Many of the applications elaborated in the project are on their way for broad implementation nowadays. The main objective of the project was the development of a "co-pilot," nowadays called *ADAS* (advanced driver assistance systems) that would provide the driver with a set of integrated systems to enhance his or her safety performance. The work was structured around three vertical lines of activities: PRO-CAR, PRO-NET, and PRO-ROAD and four horizontal activities: PRO-ART, PRO-CHIP, PRO-COM, and PRO-GEN (Figure 1). The project ran till 1993 with the production of a large set of test vehicles, aiming at demonstrating the feasibility of advanced in-car safety systems and smart traffic management systems.

Although the pilots were able to do so, the technology was not yet sufficient advanced to provide the necessary processing power and allow adequate integration. Much more work was needed for the development and design of industrial applications requiring more computing power, major reduction is size of equipment, fail-proved operation, and, last but not least, low production costs.

1.1.2 Major milestones

Almost simultaneously with the establishment of EUREKA, the European Union initiated his or her research policy instrument with the adoption of the EU framework program (FP) for research and technological development. In 1988, an application program was launched (DRIVE[1]) dedicated to the use of ICT in the transport sector, thus providing additional funds in the elaboration on the ideas originated not only from the research community but also from the PROMETHEUS project. The program covered not only

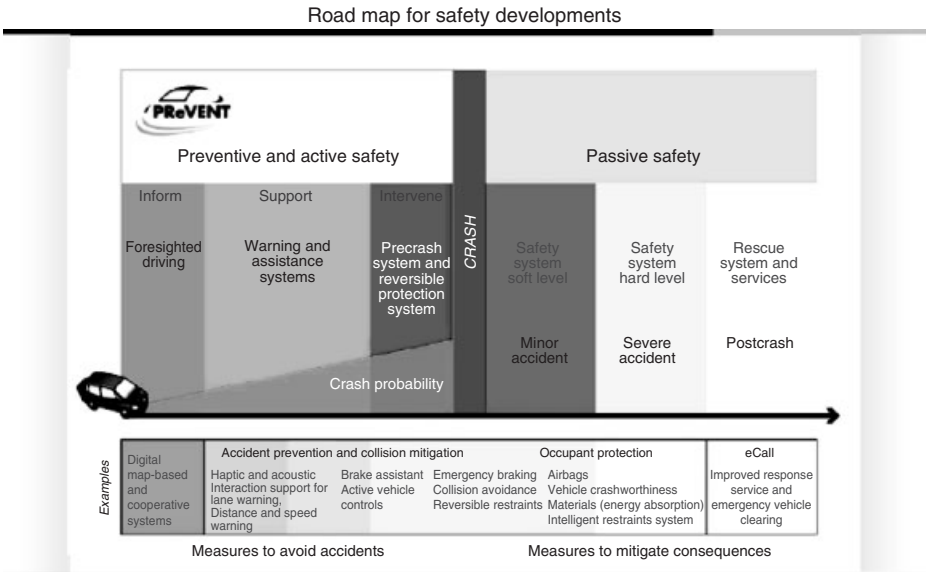


Figure 1. Functional mapping of preventive and active safety systems. (Reproduced from Schulze *et al*., 2008 © European Commission, PReVent project.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto168
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

automotive applications but also telematic services and traffic management (see further).

In 2004, another milestone was taken by the project PREVENT (Schulze *et al*., 2008), focusing on preventive safety systems (Figure 1). By the end of the project, an impressive set of automotive safety systems had been integrated in the vehicles at a development stage meeting the level of precompetitive development. An international event took place in Versailles in 2007. Several of the applications demonstrated have in the mean time found themselves into commercial products, for example, intelligent cruise control, brake assist and emergency braking, pedestrian collision avoidance, and driver drowsiness control.

Considering that a modern car has more than 100 processing units on board, it is obvious that the on-board electronic architecture is of major concern to the OEM. But also the integration of several functions, for example, sensors and processing units, into one single unit is a key priority in design, both from a safety and cost perspective. A major project on this issue, InterActive (Etemad, 2010), is on-going and will present quite soon its results.

The demand for increased performance and cost reduction has made that many of the mechanical and/or hydraulic systems are replaced by electrical/electronic devices. This has paved the way for further automation of functions and potentially will some of the early ideas on "automated driving" be available for mass production quite soon. The project HAVE-IT (Hoeger *et al*., 2011) has brought this vision to reality with its concept of highly automated vehicles for intelligent transport. Main outputs of the project have been demonstrated the late 2011 and include the design of the task repartition between the driver and the codriving system (ADAS) in the joint system. A failure tolerant safe vehicle architecture including advanced redundancy management has been developed paving the way for the next generation of ADAS directed toward higher level of automation as compared to the current state of the art (see Driver Assistance).

An overview of the progress in the different regions is best illustrated by listing the major demonstration events over recent 25 years, showing evidence that the focus has moved from technical validation tests, to precommercial prototype testing, and to large-scale user acceptance testing and social cost/benefit assessment programs (Table 1).

In parallel with the development of in-car safety systems, car industry has a major interest in responding to consumers demand for comfort systems. Broadband connection is nowadays considered as a must and access to a wide range of services is part of the attractiveness of the brand. Most OEM have therefore launched this connectivity as part of their company policy, the so-called connected car. However, this is not only about comfort but also about safety. Research programs around the world have boosted the development of "cooperative systems" (the European Union), IVI—intelligent vehicle initiative (the United States), and advanced cruise-assist highway system (AHS) (Japan). Main objective is to enable the exchange of safety relevant information between vehicles and between infrastructure and vehicle using standardized messages over the most appropriate communication channel, RDS-TMC (radio data systems-traffic message channel), GPRS (general packet radio service), and short-range communication (IEEE 802.11p standard). The SIM-TD (Safe and Intelligent Mobility—Test Field Germany)project (DE) (Weiss, 2008) is using 150 probe vehicles for testing the performance of the system. Several other test sites will provide more insight on user acceptance and business potential models, for example, at the ITS (Intelligent Transport Systems) America annual Meeting 2012.

The energy crisis and in particular the need to reduce substantially green gas emissions (EU objective: −50% by 2050) has boosted the demand for electrical cars. Major research is ongoing worldwide, in particular with regard battery lifecycle and production costs (see Advanced batteries for vehicle applications). The consequences of the increase of electricity consumption by cars will have major impact on the management of the power grid. As electricity production from natural sources is increasing, vehicle batteries are considered as a potential temporary storage facility for surplus production (see Communication of Electric Vehicles). This scenario will require extensive data communication between the vehicle and the energy provider, thus giving another argument to speed up the development of a generic telematic platform in all cars.

As conclusion, one can state that the automotive industry has been able to progress along a relative clear roadmap. The large number of partners involved in the process makes progress complex and lengthy. Systems and services resulting from cooperative research is finding its way into product development and markets. Although benefits to the driver and society are obvious, the wide-scale introduction remains strongly dependent on the economical environment, further technical progress and user acceptance.

2 NAVIGATION/TELEMATIC SERVICES

2.1 Navigation systems—comfort and safety benefits

Perhaps one of the most early "telematics" systems has been the navigation devices, as paper maps are replaced by digital maps and the driver is assisted in his driving task through route guidance and turn-by-turn information.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto168
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Major demonstration events worldwide.

Year	Europe	The United States	Asia
1986–1989	LISB (Leit und Informations system Berlin—DE)—Autoguide (London—UK) (information and navigation system)—V2I		RACS–AMTICS (JA) (traffic information services—V2I)
1994	**Prometheus demo** (collision avoidance, cooperative driving, and autonomous vehicles)		AHS–ASHRA (JA) (automated vehicle platooning)
1995–1996	STORM (Stuttgard—DE) Integrated traffic management system (several services)		AHS Demo, Komoro (JA) (advanced cruise assist)
1997		AHS, San Diego (automated cars, trucks, and bus on a segregated lane)	
1998	Automated vehicle guidance demo, Leiden (NL) (FP4 project results)		South Korea demo (KR) (platooning by four passenger vehicles)
1999	CHAUFFEUR I demo (FP4—truck platooning)	Demo '99 Launch of the IVI program	
2000	Torino ITS world conference (IT) (advance driver assistance, etc.)		Demo 2000 Ministry of Land, Infrastructure, and Transport (JA) (automated vehicles, crash avoidance at traffic lights, and curbs) Japan, Ministry of Economics, Trade and Industry (JA) (platooning)
2002	Demo 2002, Versailles (FR) (a wide range of applications)		
2003	CHAUFFEUR II (automatic platooning)	National IVI demo, FHWA (forward collision warning, intersection-collision warning, etc.)	
2004	Intelligent vehicles conference, Parma (IT)-IEEE (EU research projects) Cybercars Demo, Antibes (full automated public transport)		
2005	Road of the future (NL) (showcase of ADAS systems)	Innovative mobility showcase, ITS world conference, San Francisco (cooperative driving, 5.9 GHz DSRC, ADAS, etc.)	Smart demo, ITS Australia (active safety, crash notification, driver-assistance systems, etc.), AVS-3 demo, Hokkaido (JA)
2006			Demo 2006 (JA) (near-term deployment of 5.8 GHZ DSRC—several functions)
2007	PREVENT, Versailles (FR) (safe speed and safe following, lateral support, intersection safety, vulnerable road users, and collision mitigation—25 cars)		
2010	Cooperative mobility showcase 2010 (NL) (ITS App Store)		
2011	HAVE-IT, final conference (SE) (driver assistance systems—highly automated driving)		

(Reproduced by permission of UKiP Media & Events.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto168
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

From the early development stage, two different approaches have been pursued: the first infrastructure based and the second using autonomous system approach.

Under the leadership of Siemens (von Tomkewitsch, 1986), an innovative system for route guidance was developed, known as *ALISCOUT* (1986) or *EUROSCOUT* (1989). After entering his destination, the driver receives from the Traffic Operation Center routing information when passing beacons installed at strategic intersections. The system used infrared communication technology and gave turn-by-turn guidance. Major field trials took place in West-Berlin (LISB) 1989–1990 and also in London (Autoguide). As new technologies emerged (i.e., satellite positioning—microwave communication), the technology was abandoned.

As positioning though satellite systems was not yet available for public use, early autonomous navigation systems used odometer and gyroscope data to calculate the trajectory of the vehicle and map matching algorithms to position the vehicle on the map. Early 1990, Japan had already a sizable market for navigation systems. The development of the US GPS (global positioning system) (1978–1995) and its availability for public use (May 2000) has resulted in an explosive growth of the sector.

In Europe, major development was made through EUREKA research projects: DEMETER (Digital Electronic Mapping of European Territory) and CARMINAT, involving several OEM and equipment suppliers. The high cost of in-vehicle integrated navigation systems and the poor quality of the available maps made market deployment very low for a long time.

The mass production of personal digital assistant (PDA), integrating general-purpose processing unit and a GPS receiver, provided the platform for the implementation of navigation systems. As of May 2000, GPS signals available for public use could profit from a high accuracy and navigation equipment gave substantial improvement in user acceptance. Prices have since fallen to consumer levels and navigation systems have become a mass market item.

Worldwide the number of turn-by-turn navigation systems is estimated at 200 million devices, of which some 120 million are personal navigation devices (PNDs) and 40 million in-car integrated systems. The prospects for navigation-enabled mobile phones, estimated at 40 million in 2011, are expected to rise up to 200 million by 2015 (Berg Insight, 2011) (Figure 2).

As the GPS constellation came to completion, the deployment of the GALILEO system will allow to increase the position accuracy to a few centimeters (see Technologies—Positioning: GNSS). This will allow an additional number of potential functionalities, for example, lane departure warning and additional support in maneuvering will be provided on navigation systems. In addition, the inclusion of safety-related features in the map data base gives a number of interesting and potentially important functionalities. As part of the PREVENT (MAPS & ADAS) project, the potentials have been investigated for several applications, for example, for advanced cruise

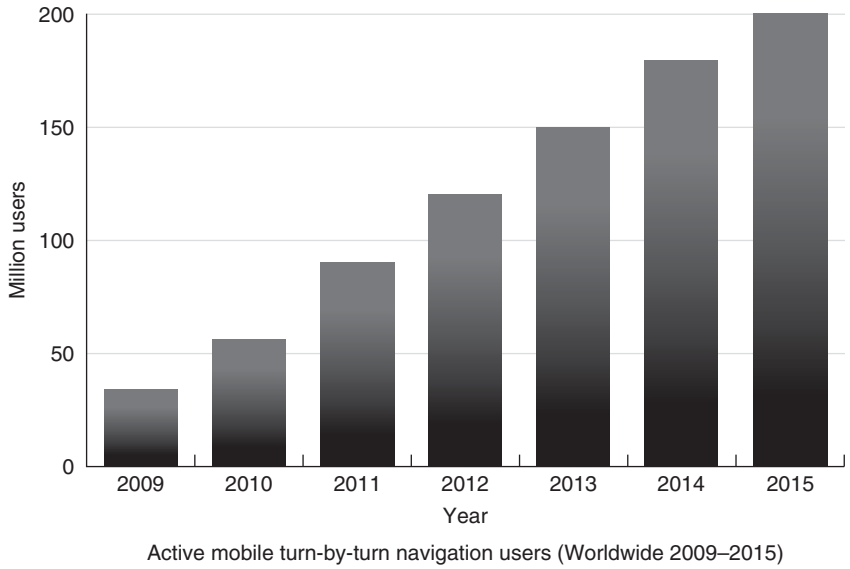


Figure 2. Evolution of number of turn-by-turn navigation system users. (Reproduced with permission from Berg Insight Mobile Navigation Services, 2011. © Berg Insight AB.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto168
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

control (ACC) and stop-and-go, where the digital map acts as a predictive sensor.

2.2 Traffic information services

Already in the early 1970s, the ARI (Autofahrer Rundfunk Information) broadcasting system was developed using the FM band. This system was the basis for the radio data system (RDS), which specifications were completed in 1984 by the EBU (European Broadcasting Union). Experiments took place in Germany and in several European countries. A further important building block for the development of telematic services has been the adoption at an early stage of the ALERT protocol (Davies, 1991), defining coding rules for a large set of messages.

Given the limited possibilities for location coding in the ALERT protocol, a new set of specifications was defined, called *TPEG* (transport protocol experts group) (Kopitz and Marks, 2003), and is on its way for deployment. These new coding specifications will enable a wider range of services to a vast range of users and devices. It allows a wider range of applications not only for road traffic information services but also for public transport information, parking information, and so on.

Each TPEG message comprises three key parts: message management, the application event, and the location referencing. The location referencing itself is partitioned into many components allowing all types of client devices to localize an event and display the localization as appropriate to the client.

Traditionally, road operators have been the main sources of traffic information as part of their responsibilities for safe operation of the network and efficient management of the service. Owing to public demand, public transport operators have become more active in this area, in particular when dealing with traffic disturbances and emergency situations. Nowadays, most operators offer direct access to timetable information and ticketing over smart phones. Bilateral agreements between operators allow access to multimodal trip information and will allow the creation of pan-European traveler information systems. Issues remain the access to real-time data and updating of traffic information, as to provide optimal support to travelers during their trip (see Applications—Intelligent Vehicles: Driver Information).

More and more service providers appear on the market collecting and broadcasting travel and traffic information. These private service providers can build their services on the data collected and made available by the operators and also from own data. The very wide deployment of mobile telephony (GSM) has allowed telecommunication operators and service providers to monitor traffic flows using the handover messages from mobiles as they travel from one cell to another. This allows the calculation of average speed and by extrapolation length of queues. A similar technique is expanding using WiFi connected devices. Although the range is much smaller, it has potential for the monitoring of traffic in urban areas.

As the number of mobile devices with GPS receiver is growing very fast, uploading location data is becoming a very attractive method to collect (real-time) traffic data.

2.3 Client feedback

Quite early, navigation system providers have integrated RDS-TMC messages into the map attributes, thus providing almost real-time traffic information (when and where available).

As map data, and in particular white page information, is very sensitive to changes, regular updates of the map database is a requirement for optimal performance, but it is also very sensitive in customer appraisal.

Triggered by the success of the social network models, customer feedback has become part of the business model, for example, TOMTOM (http://www.tomtom.com/?Lid=1) and Coyote (http://shop.coyotesystems.com/). The customer becomes part of the production and updating process. The registered user benefits from his or her participation through higher data quality. This trend will further expand and act as a model for other telematic services as data transmission costs will further decrease.

2.4 Open platforms

Smart phones have become an important alternative for OEM integrated telematic platforms, due to the price and their flexibility. A highly integrated board, including GSM, GPS modules, accelerometer, and more to come, provides a platform for an almost unlimited number of applications. Major issue remains the fact that road accidents are in the majority of cases attributed to driver distraction and mobile devices are of the kind to take driver attention for periods longer than is considered safe.

The flexibility of smart phones is challenging for the OEM as technology cycles are extremely quick and product lifetimes are very short compared with the average lifetime of a car. To provide a comparable level of applications, two different development scenarios are on the drawing table.

1. The on-board system interfaces with the portable device and limits functionalities to those considered safe while driving. This is based on the consideration

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto168
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

that most drivers own a smart phone and want to use its functionalities and stored personal data.

2. The other builds on the belief that the future generation of vehicles will have an integrated platform as a basic feature that provides connectivity to its environment. This is not only for comfort and entertainment purposes but mainly for safety purposes [C2C (car-to-car)–C2I (car-to-infrastructure)—cooperative driving] and vehicle monitoring (proactive maintenance, power train, and battery control). A further area for potential integration is related to payment systems, for example, tolling.

In both scenarios, the human machine interface (HMI) becomes an extremely important part of the overall design of the car. The challenge is to provide customer demand while ensuring that safety and security requirements are met. This need has been recognized by a large number of stakeholders and led to the creation of the HUMANIST VCE (Virtual Center of Excellence) (Médevielle, 2010). HUMANIST acts as a network of research centers, universities, and industry with the common objective to get a better understanding of human behavior and optimal design of the man–machine interface (see Driver Distraction).

3 TRAFFIC MANAGEMENT

A third area of great importance for the development and the future deployment of telematics is related to the operation and management of the road network.

The explosive growth of road traffic has put public authorities over recent decennia for the challenge to maintain reasonable traffic flow conditions. While in the past, growth could be compensated by new infrastructures, this becomes more and more difficult because of urbanization and environmental constraints.

Similar to other domains, the large-scale availability of high performing processing devices has changed the environment of the road manager. From mainly construction and maintenance technologies, ICT has taken today an important part of the investments and operation and will continue to do so. Although the advantages of wide-scale deployment of telematics/ITS are obvious, road operators are concerned about the high rate of technology changes and continuity of service. Progress in standardization and drafting of technical specifications is too slow to cope with technological developments (see Applications—Intelligent Roads and Cooperative Systems: Urban Traffic Management and Advanced Highway Management Systems).

3.1 Data acquisition and variable message signs

Perhaps the first priority of road operators is to have access to real-time data from its network. Detectors and image processing have therefore been the forerunners. Speed, volume, and vehicles classes are in real-time available and are feeding automatic incident detection (AID) algorithms. They provide the necessary statistical data for tactical decisions and calibration of simulation models.

Considering the deployment of embedded telematics in cars, many development scenarios look at the potential for the vehicle to act as sensor in the network, sending its (vehicle) data to the operators network, thus potentially reducing the cost of road infrastructure and expanding the traffic monitoring capacities to the whole network.

Most traffic control operators require a visual confirmation of the correctness of the data or the emergency call. Cameras (in the past closed-circuit TV) have become an important instrument in the traffic control centers to reduce the need to send police to the site before taking remedial action. Digital technology has made the cost of equipment and transmission (in most cases over dedicated communication network) to decrease sharply. Nowadays, on-site equipment, including cameras, can be connected to the control center through wireless links, thus making cabling requirements less stringent.

The same applies for variable message signs (VMS), where major progress has been achieved with the availability of high performance LED technology. Although low power consumption and IP-based communication are cost-efficient, the physical infrastructure on which the panels are mounted remain expensive and form a potential obstacle in case of an incident. As the "connected car" finds its way into the market, local traffic signs (messages) can be displayed on the vehicle display using he or she same communication channel as the one used for V2V communication. As shown from experiments, driver attention is increased and a safer driving behavior achieved.

3.2 Electronic toll collection

Tolling exists since the ancient times and has been a tool to finance large infrastructures, as construction and operation has been outsourced to public–private initiatives. In most cases, charging was introduced as the infrastructure became available (bridge, tunnel, and motorway). For many years, toll was collected at large toll plazas at toll booths manually and by credit cards.

From the early beginning of the European research programs, short-range communication has been a topic of research, looking at the most appropriate frequency range

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto168
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and communication protocols. After long and difficult negotiations, a standard was adopted by CEN (EN 12253:2004, etc.) using microwave transmission at 5.8 GHz. This technology has been deployed on all major toll motorways in Europe, allowing free-flow tolling, and thus reducing the land use needed for toll plazas. The European Union Directive 2004/52 EC on the interoperability of electronic road toll systems in the community (EFC, electronic fee collection) requires the operators to arrive to an interoperable system and single charging agent (one invoice) (European Commission, 2004).

The interoperability of toll services in Europe is considered as a prerequisite for the further implementation of road charging for trucks (>3.5 Ton).

Several other technologies have been used for tolling, or congestion charging (see below), but mainly license plate recognition (e.g., London and Stockholm). Charging is based on access time and duration. These schemes have been evaluated in depth and from the results it is clear that user acceptance is good if the charging fees are adapted to the specific size of the city, the alternative transport facilities, and the modal interfaces.

The progress achieved, both in navigation technology and mobile communications (GSM), led to the development of the toll collect (http://www.toll-collect.de/en/home.html) system. Toll collect has been developed in Germany for truck tolling on motorways and became fully operational as from 1 January 2005. The total network covered by toll collect is 12,600 km of motorways and trunk roads.

3.3 Mobility pricing

From an economical view, the (road) operator provides a service and it is logical that the customer (the road user) pays for this service. The price of this service should basically be depending on demand and supply. This basic principle in transport economics has been difficult to implement, due to already existing taxes and levies imposed on the road user. Many cities have implemented schemes in which price is dependent on time of day, type of vehicle (environmental zones), and so on. Poor traffic conditions and unpredictability of journey times (both private and professional) lead to the fact that more and more (road) users are inclined to accept the need to have a more flexible pricing system.

In principle are the different technologies, that is, image processing, transponders, and so on, suited for the implementation of (local) pricing schemes. If a scheme is implemented that covering a wide(r) network and compatible over different cities and countries, the GSM/GPS-based tolling system seems to be the most appropriate way forward. This requires standardized on-board-equipment and a competitive market across the European Union for service providers (see Tolling, Mobility Pricing).

3.4 Junction control

Traffic light controllers are perhaps the most sophisticated on-site equipment nowadays. Over the years, these controllers have evolved from electromechanical switching equipment to fully electronic devices, running complex control algorithms. Most cities have installed over the years a central coordination of the traffic lights, triggering the local controllers according to optimal offsets between junctions, aiming at optimizing flow on main corridors (green waves) or minimizing total delays in the network (e.g., using simulation models such as TRANSYT7).

Nowadays, distributed architectures are put in place, using mesh technology for communication between traffic controllers.

Junction control could well become one of the early applications of vehicle-to-infrastructure (V2I) communication, warning the driver on the remaining time to end-of-green, recommended speed, and so on. The high accident rate at signalized junctions could be the main rational to add this functionality into the traffic controller. RITA, the Research and Innovative Technology Administration of the United States, has made intersection control one of its priorities within the VII, also called *Connected Car*, research program.[2] Several traffic controller manufacturers are testing the application of V2I technology in close cooperation with the European efforts on "cooperative driving" through the EU research program.

Together with pricing is full network management definitely the next step to achieve acceptable levels of service, for example, not only on the highway network but also on the feeding network. As demand for mobility will further increase and environmental requirements will become more stringent, optimization across transport modes is nonavoidable. In this process, due account must be taken of the complex nature of behavior and choice of transport mode (car–train/metro–bus/tram walking). As urban centers will further grow and transport intensity increasing, it is essential that modal transfers are optimally managed. The multitude of operators, complexity in layout, and modal characteristics make this a major challenge for the future operation of our transport systems. This requires at urban and local level strong teams with specialized skills in transport management, willing to work together with a common objective.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto168
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4 FUTURE TRENDS

Looking at the technological progress made over recent 20 years, technology has fundamentally changed our view on mobility and our choices of travel modes. This choice is not any more exclusively the private car, but depending on the motive of the travel, the most appropriate mode or combination of modes to perform the trip in the most convenient way (walking, bike, tram–bus, and metro/train/airplane).

On the vehicle side, both energy supply and environmental issues will dominate the design of the next generation vehicle fleets. The electrification of the car is clearly on its way and a sizable market penetration of hybrid and full electric vehicles in the near future is realistic, in particular in the market segment of household second car and shared-car schemes. These segments are well suited as usage is characterized by short distance travel. Further deployment will depend on progress in battery technology, the use of low weight materials and size reduction.

Traffic safety needs to remain a major priority. As new designs will emerge on the market, in particular small and light vehicles, the composition of traffic flow fundamentally changes. Interaction between vehicles with significantly different dynamic characteristics could have a negative impact on road accidents. In such a case, appropriate measures of segregation of vehicles become necessary with important impact on network design and operation.

However, many issues remain which are not necessarily technological but much more relate on how society will look like in a number of years and how it will impact on the mobility of people and goods.

In this perspective, the further urbanization of society is a major driver on the evolution of our transportation systems. Owing to the concentration of activities in city centers, further restriction of individual transport will be needed to achieve environmental acceptable conditions. Even under the hypothesis of large-scale deployment of electric cars, congestion will be a sufficient reason the restrict access. This is only possible if highly performing public (collective) transport systems are available and transfer between modes is optimally organized.

The further integration of the European Union and its economy depends on an efficient and highly efficient transportation system. As these networks are used by both local and long distance traffic, traffic conditions are highly dynamical and unpredictable.

In both cases, telematic services are extremely important and will be able to show their full benefits. Although each operator runs today its own portal with traffic information (mainly roads), timetables, and booking facilities, they tend to be reliable only under normal traffic situation and perceived by the user as static data. The real challenge is to provide real-time data and, perhaps most important, information on the disruption of service and possible alternatives across modes. Within the eSafety Initiative (http://ec.europa.eu/information_society/activities/esafety/index_en.htm, 2003), traveler information has been one of the key issues addressed by stakeholders and has been taken up as one of the actions under the EC ITS Action Plan (http://ec.europa.eu/transport/themes/its/road/action-plan/, 2008).

As technology changes and our mobility needs is evolving, the role of governments and public authorities becomes critical in creating the conditions necessary to be met, both in terms of regulations and of new infrastructures. Many platforms exist addressing the transportation issues from a strategy level, for example, the United Nations and the International Transport Forum. In parallel with the ITS Action Plan, the European Union adopted a Directive on the deployment of ITS (2010/40/EU) (European Union Directive, 2010), putting in place a framework in support of the coordinated and coherent deployment and use of (ITS) within the Union, in particular across the borders between the Member States, and sets out the general conditions necessary for that purpose. The Directive also provides for the development of specifications for actions within the priority areas referred to in Article 2, as well as for the development, where appropriate, of necessary standards. One can argue if this is the appropriate instrument to address these issues. It has the benefit however that the need for common specifications and standards is brought at the attention of Member States and a platform for discussion and consensus is established.

To help the decision process, both an EU, national and regional levels, the need exists to create, or improve existing simulation and support tools able to provide a clear and objective view on the impact of policies.

5 CONCLUSIONS

Although the technology has made huge progress, large-scale deployment depends on a number of external factors such as economic growth, stakeholders' cooperation, and user acceptance. Typical for transport telematic services is their complex nature and a value chain that involves many organizations. As internet has provided the user an almost unlimited access to data and information free of charge, identifying the correct business case for new telematic services is difficult and not always successful. All innovations however demonstrate that data V2X communication has huge potentials in supporting major societal needs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto168
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Technological breakthrough in ICT has been the main drivers for the development of transport telematics; GSM, GPS, smart phones, and social network software are the building blocks for telematic services. It is therefore extremely important that transport experts keep an open mind for developments in other sectors and take advantage of emerging technologies to provide high quality telematic services.

ABBREVIATIONS

ACC	Advanced cruise control
ADAS	Advanced driver assistance system
AID	Automatic incident detection
CEN	European Committee for Standardization
EBU	European Broadcasting Union
EFC	Electronic fee collection
EUREKA	Cooperative research and technological development initiative
FP	Framework program for cooperative research and technological development
Galileo	European satellite system for positioning and rescue services
GPS	Global positioning system
GSM	(Groupe spécial mobile) mobile communication standard
HMI	Human–machine interface (interaction)
ICT	Information and telecommunication technologies
ITS	Intelligent transport systems
PDA	Personal digital assistant
PND	Personal navigation device
PROMETHEUS project	Program for European traffic with highest efficiency and unprecedented safety
RDS	Radio data system
RDS-TMC	Digital traffic message information broadcasting over FM
TPEG	Standard for coding traffic informationV2V–V2I: vehicle-to-vehicle and vehicle-to-infrastructure communication
VMS	Variable message signs
VCE	Virtual centre of excellence
VII	Vehicle infrastructure integration
Wi-Fi–WiMax	Wireless network standards for internet access

ENDNOTES

1. Traffic, Proceedings of the DRIVE Conference, Brussels. Elsevier.
2. RITA. Federal Highway Administration (http://www.its.dot.gov/).

REFERENCES

Berg Insight, Mobile Navigation Services (2011) http://www.berginsight.com/Default.aspx?m_m=1.

Davies, P. and Klein, P. (1991) RDS-Alert—Advice and Problem Location for European Road.

Etemad, A. (2010) Interactive—Project Website. http://www.interactive-ip.eu/ (accessed 2 July 2013).

European Commission (2004) Directive 2004/52/EC. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:166:0124:0143:EN:PDF (accessed 2 July 2013).

European Union Directive (2010) 2010/40 of 7 July 2010 on the Framework for the Deployment of Intelligent Transport Systems in the Field of Road Transport and for Interfaces with Other Modes of Transport.

Glathe, H.-J. (1993) PROMETHEUS common European demonstration: a tool to prove feasibility. IVHS America 1993 Annual Meeting, pp. 174–179.

Hoeger, R. *et al*. (2011) HAVE-IT Final Report. http://haveit-eu.org/LH2Uploads/ItemsContent/24/HAVEit_212154_D61.1_Final_Report_Published.pdf (accessed 5 July 2013).

Kopitz, D, Marks, B. (2003) TPEG what is it all about. TPEG Project Office, 2003-7, EBU http://tech.ebu.ch/docs/other/TPEG-what-is-it.pdf (accessed 2 July 2013).

Medevielle, J.-P. (2010) Activities Report 2009–2010. http://www.humanist-vce.eu/.

Schulze, M., Mäkinen, T., Irion, J., *et al*. (2008) Final Report. http://ec.europa.eu/information_society/activities/esafety/doc/rtd_projects/fp6/PReVENT_Final Report.pdf.

von Tomkewitsch, V. (1986) Ali-scout—a universal guidance and information system for road traffic. Second International Conference on Traffic Control, IEE, London.

Weiss, C. (2008) Safe and Intelligent Mobility—Test Field Germany. http://www.simtd.de/.

FURTHER READING

European Commission, eSafety Compendium (2006) http://www.icarsupport.eu/esafety-forum/ (accessed 5 July 2013).

European Commission Satellite Navigation. http://ec.europa.eu/enterprise/policies/satnav/index_en.htm; http://www.humanist-vce.eu/humanist-vce/objectives.html (accessed 2 July 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto168
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Cellular Mobile Networks

Friedhelm Ramme[1], Gero Fiege[2], and René Rembarz[1]

[1] *Ericsson GmbH, Eurolab R&D, Herzogenrath, Germany*
[2] *Vodafone GmbH, Düsseldorf, Germany*

1 Introduction 1
2 Cellular Mobile Networks 1
3 System Architecture 3
4 End-to-End Procedures 12
5 Standards and Regulatory Frameworks 20
6 Outlook 22
Glossary 23
Endnotes 27
References 27

1 INTRODUCTION

Being connected while on the move has become a common customer expectation. This expectation goes hand in hand with a change in the observed user behavior. Nowadays, apps and connected devices have become part of people's everyday life. Cellular networks are expected to provide the required connectivity whenever and wherever needed, which is a big change compared to the situation just two decades ago. A further trend with impact on society and major economic and environmental implications is the ever-growing demand for personal and goods mobilities. A steady increase in road traffic, particularly in developed or rapidly developing counties, is spurring the demand for real-time traffic information, safety applications, and a broad set of connected car services. Hence, the underlying expectation on ubiquitous cellular connectivity is becoming increasingly important for the mobile society and the just-in-time economy. A changing perception of the usefulness of the deployed cellular networks and recent technology advancements have led to a more balanced thinking when arguing for additional large-scale investments in dedicated vehicle communication infrastructure for certain driver or passenger services. In any case, we do not discuss anymore whether cars will become connected in the future—they will! The question is rather: what is the best option to realize tailored connectivity for automotive applications from a technology perspective, considering operational, economic, and timing conditions?

This chapter provides some basic information about cellular communication systems, required to properly address those questions. Major trends in cellular communications and resulting impacts to business and society are discussed in Section 2. The various generations of cellular wireless communication systems, their key characteristics, and prime network elements are introduced in Section 3. Section 4 illustrates selected use case procedures and how core network (CN) components interact with each other in an integrated system design. Special attention is put on advanced scenarios with particular relevance for automotive communication. Finally, Section 5 provides a short introduction on the standards and regulatory frameworks related to cellular communication. An outlook on latest developments and implications concludes this chapter.

2 CELLULAR MOBILE NETWORKS

Cellular mobile digital networks are one of the fastest growing and most rapidly developing technology segments of the past few decades. Starting with the establishment

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

of the global systems for mobile communication (GSM) standardization group in 1982, with the ambition to specify a pan-European mobile cellular radio system for primarily voice communication at 900 MHz, it has laid the foundation for connecting more than 5 billion users worldwide in 2010 (More than 50 billion connected devices, 2011; City life, 2012).

The pace of development is remarkable, constantly gaining speed and spread. The decision to specify a digital system (rather than a new analog one) in 1985 was followed by the selection of the time division multiple access (TDMA) method in 1987 and the promise of all early European mobile network operators (MNOs) to have the first interoperable GSM systems running by 1 July 1991.

Over the following two decades, GSM has evolved to a digital mobile communication system that connects the world, impacting society and businesses worldwide. Recently, another major transition is materializing, comparable to the step from analog to digital telephony. Originally developed for point-to-point (PTP) voice communication, circuit-switched (CS) communication networks are gradually being replaced by the so-called all-IP (all Internet protocol) data communication networks. This new breed of networks transports IP packets rather than switching voice circuits. In Q4 2009, the data traffic volume of these so-called packet-switched (PS) networks has already overtaken the aggregated traffic in CS networks. It doubled its volume by Q1 2011. Between Q3 2011 and Q3 2012, the mobile data traffic volume doubled again (Ericsson mobility report—on the pulse of the networked society, 2012). A major driver and accelerator for this trend was the introduction of Apple's iPhone in 2007. The subsequent rapid spread of powerful smartphones has in turn led to an explosion in data traffic volume in mobile networks worldwide. According to studies done by Ericsson (Ericsson mobility report—on the pulse of the networked society, 2012), mobile data traffic is expected to grow with a compounded annual growth rate of around 50% (2012–2018), driven mainly by video. This entails growth of around 12 times by the end of 2018. In that period, data traffic will be split fairly equally between devices such as smartphones on the one hand, and PCs and tablets on the other (Figure 1).

To accommodate such tremendous communication network demands, entirely new radio technologies and new communication system designs had to be developed and deployed within just a few years, responding to capacity and bandwidth challenges. The additional demands, added by the transition of CS voice communication to voice-over Internet protocol (VoIP), are small compared to the always-on desire of the twenty-first century society, expecting Internet access and interconnected

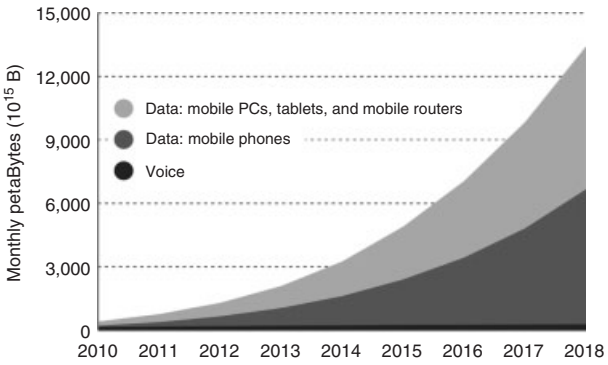


Figure 1. Mobile traffic—voice and data, 2010–2017. (Reproduced by permission of Ericsson.)

media applications to work seamlessly from all kinds of connected devices at any time.

The rapidly changing habit in human communication has deeply impacted business and business processes. With modern cellular network technologies, many types of digital devices can be connected to automate business processes and implement novel system functions in a virtually connected way. Experts predict that by 2020, there will be more than 50 billion connected devices (More than 50 billion connected devices, 2011; City life 2012). This enormous growth and requirements from entirely new application segments have put demanding new challenges to network suppliers and operators. Beside a substantial increase in network capacity and cellular communication bandwidth, also demand for a much simplified network operations and business systems management has been expressed, together with the need for communication latency reductions and a reduction in the connection establishment time in cellular networks.

Since its early GSM days, data communication speeds have been boosted from 9.6 kbps to more than 100 Mbps with long-term evolution (LTE) technology today—an increase by a factor of 10,000 in just 15 years. Latency time for data transmission has reduced from many seconds with GSM to <50 ms with LTE Release 10 in 2012 (Next generation LTE, LTE-Advanced, 2010). This breakthrough in the capabilities of cellular mobile communication technology has in turn led to a new evaluation of latest developments in connected vehicle designs, real-time telematics services, and just-in-time logistics.

The following sections will, therefore, provide a brief summary of the cellular mobile digital communication systems evolution, to the extent relevant for automotive engineering. System functions specific to CS voice communication or those being of little relevance in the automotive engineering context (such as wireline communication) will be only described on a very basic structure level.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3 SYSTEM ARCHITECTURE

3.1 System families and their evolutions

The appearance of digital radio networks, with its various technology choices, has led to a number of different network families. For simplicity reasons, their evolution steps have been mapped to a commonly used "generation numbering scheme." Analog communication is referred to as *first generation*, the first digital systems are called *second generation* (*2G*). Systems, fulfilling the requirements of the International Telecommunication Union's (ITU), the so-called IMT-2000 (International Mobile Telecommunications 2000) system (Rappaport, 2001), are counted as third generation (3G). Systems fulfilling the newer IMT-Advanced requirements are called *4G (fourth generation) systems*. Most of these system families have evolved in various steps and phases since its beginning at the mid-1990s. Figure 2 illustrates how the various cellular network families and their various versions relate to each other.

Some of the first 2G systems were only developed for just one or a few countries, such as personal digital cellular (PDC), which is used exclusively in Japan. Others gained wider acceptance, such as the IS-136 and IS-95 systems, both standardized within the Telecommunications Industry Association (TIA). IS-136 is a 2G system from the early 1990s, also called the *North American Time Division Multiple Access (TDMA)*, which has been widely used in North America, Canada, and South America. IS-95, commonly called *cdmaOne*, is already using a modern code division multiple access (CDMA) wireless access, a precursor to the radio access used in many state-of-the-art systems. cdmaOne emerged in the mid-1990s and is a popular 2G technology in the Americas, Asia, and Eastern Europe.

Originating from Europe, the 2G GSM uses TDMA, comparable to IS-136, but with a higher system capacity. GSM has first been standardized in 1990 by the European Telecommunications Standards Institute (ETSI) and is the most widely used 2G system. With the establishment of the third-generation partnership project (3GPP), the GSM standardization has been transferred from ETSI to 3GPP.

All these systems were, however, primarily designed for voice communication. The first system primarily designed for data communication, also referred to as *packet-switched communication*, was the general packet radio services (GPRSs), which is also known as *2.5G*. GPRS was later enhanced by the enhanced data-rates for global systems for mobile communication evolution (EDGE) upgrade. Although it was never considered a 3G system, it is interesting to note that EDGE even fulfills the basic requirements for IMT-2000 systems, such as peak data rates of at least 200 kbps (Rappaport, 2001).

The first true 3G system was the universal mobile telecommunications system (UMTS), standardized in 3GPP and first offered in 2001. It was primarily used in Europe, Japan, China (however, with a different radio interface), and other regions predominated by existing GSM 2G system infrastructure. Broadcast and multicast capabilities have been added through the multimedia broadcast

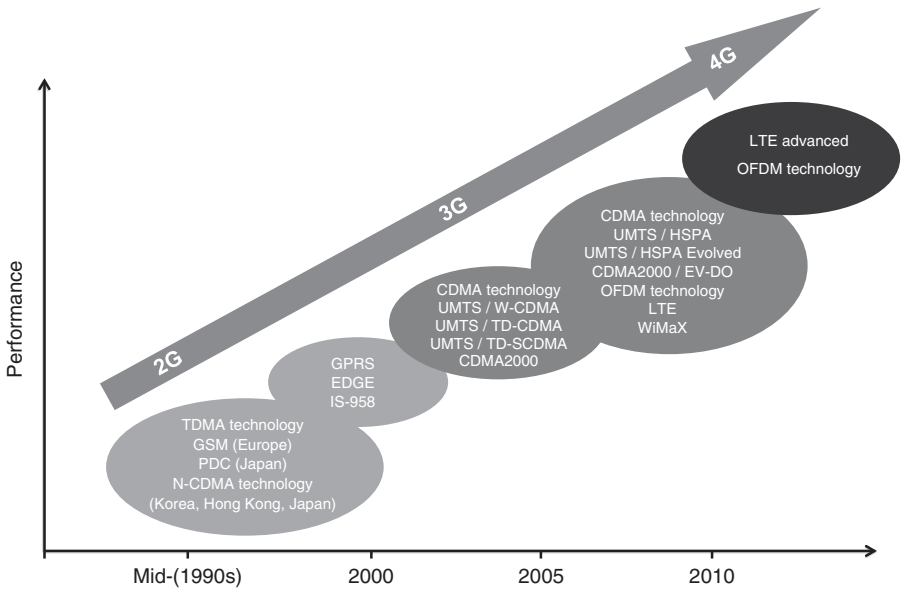


Figure 2. Evolution of cellular networks.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

multicast service (MBMS). Seamless interworking between GSM and UMTS was already part of the initial standards, so that 3G phones are typically UMTS and GSM hybrids. UMTS was further improved with the high speed downlink packet access (HSDPA) and high speed uplink packet access (HSUPA), often jointly referred to *HSPA* (*high speed packet access*). These evolution steps primarily added higher transmission speeds, but also other improvements such as faster round-trip times required for interactive applications.

UMTS is, however, not the only system fulfilling the IMT-2000 requirements, one other popular example being the CDMA2000 standard, which is the next generation of the IS-95-based cdmaOne systems. CDMA2000, first offered in 2002, is standardized by the "third-generation partnership project 2" (3GPP2) and has been used especially in the Americas, Asia, and Eastern Europe, mostly because of the installed base of 2G cdmaOne systems. The base system called *CDMA2000-1x* has symmetric bandwidth for up- and downlinks, the later CDMA2000-1x EV-DO (where DO is for data only) has introduced an asymmetric channel model optimized for data transmission. The later CDMA2000-3x systems support larger frequency bands, allowing for higher throughput. Similar to the GSM/UMTS systems, mobile phones are typically CDMA2000 and IS-95 hybrids.

The 4G standard follows the ITU's IMT-Advanced requirements. The most prominent example of a 4G system is LTE. LTE is the 4G wireless data communications technology of the 3GPP family that already standardized GSM and UMTS. The world's first publicly available LTE service was launched by TeliaSonera in the Scandinavian capitals Oslo and Stockholm in 2009. The LTE specification provides downlink peak rates of 300 Mbps, uplink peak rates of 75 Mbps, and a novel scheduling mechanism enabling round-trip times of <50 ms (Astély *et al*., 2009; Larmo *et al*., 2009). LTE has the ability to manage fast-moving mobiles and supports multicast and broadcast streams using "evolved multimedia broadcast multicast service" (eMBMS, Section 4.2.1). LTE supports scalable carrier bandwidths, from 1.4 to 20 MHz and supports both frequency division duplexing (FDD) and time division duplexing (TDD). Later releases of LTE have the name LTE-Advanced.

According to current trends, LTE is leading the convergence of various cellular network standards. For networks with an installed 3GPP base, LTE is set as default evolution path. But also network operators with existing CDMA2000 networks have a migration path toward LTE, including a solution for seamless handovers between CDMA2000 systems and LTE. Operators in North America and Asia, with a deployed CDMA system base, have already taken steps to migrate to LTE. Correspondingly, LTE is anticipated to become the first truly global communication standard for cellular mobile networks.

Mobile WiMAX (worldwide interoperability for microwave access), based on the IEEE 802.16 family, that fulfills the ITU's IMT-Advanced requirements is, therefore, formally also a 4G communication system. It can provide bit rates of up to 50 Mbps with its 2005 version. Despite the addition of mobility features, with the IEEE 802.16e enhancements, WiMAX has, however, not gained ground as a true mobile network. Its usage is seen primarily in fixed and nomadic data network installations. Figure 3 illustrates these technology convergence market trends, together with the corresponding technology standardization bodies.

An overview of most prominent cellular mobile data communication technologies is provided in Figure 4. Peak data rates have drastically improved with each generation's technology advancements. For instance, LTE Release 10 (which includes first LTE-Advanced enhancements) can support up to 1.2 Gbps peak rates under ideal channel

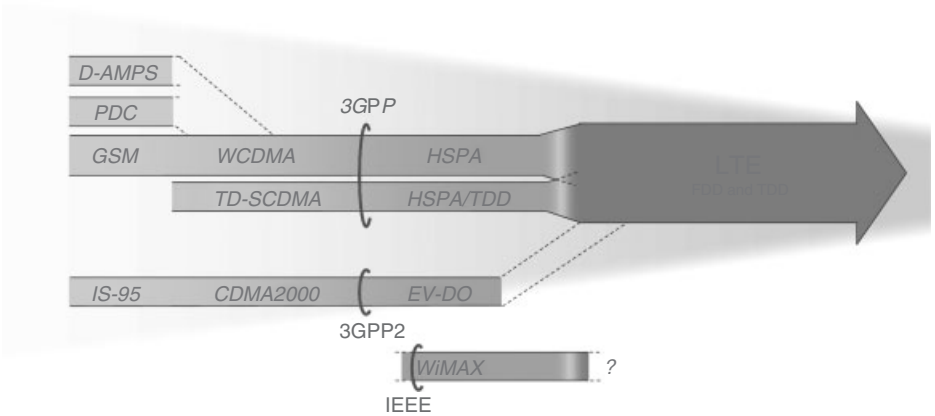


Figure 3. Technology convergence and related standardization bodies.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

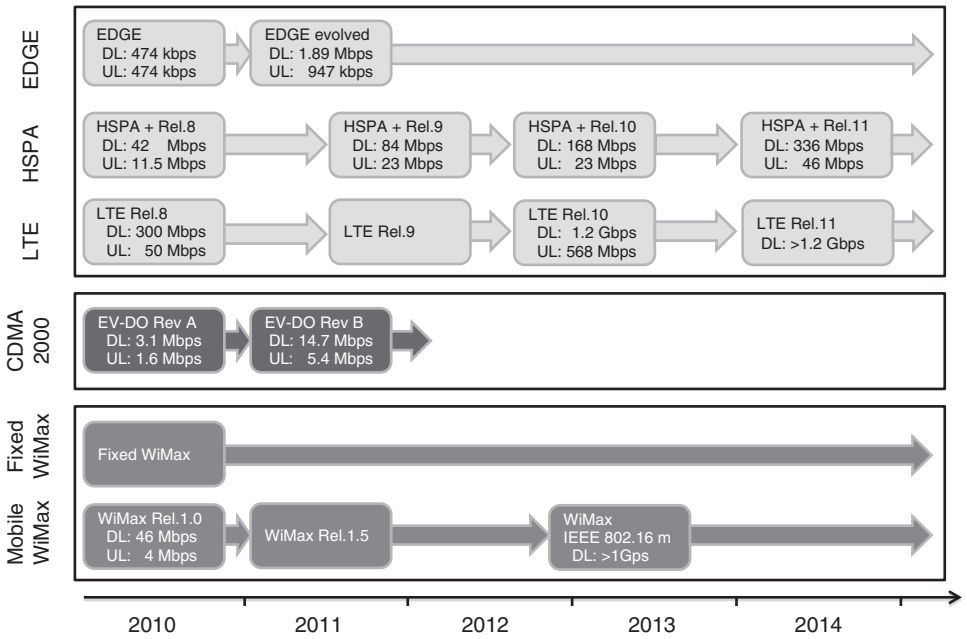


Figure 4. Yearly peak rate evolution for various wireless technologies.

conditions, nearly 2500 times higher than what EDGE technology offers. Under such conditions, LTE-Advanced provides more than three orders of magnitude higher data rates than EDGE provided only a decade ago.

We believe that with the rapid global roll out of LTE and the already confirmed technology improvements coming with the introduction of LTE-Advanced, LTE will be the main technology track for coming 4G systems. Consequently, we will focus on LTE as 4G technology base in the following sections. To stay focused, we limit the following discussion on engineering aspects and only briefly touch upon legal and regulatory aspects. The following sections provide a systematic overview of the core elements of a 3GPP cellular mobile system, its essential interfaces, and the corresponding business support functions.

3.2 Architecture overview

Core elements of an overall system blueprint are indicated in Figure 5. To help the reader with a basic understanding of a fairly complex system landscape, we provide a walk-through of these core elements in the following paragraphs. Some basic end-to-end procedures, involving several of the elements, are explained in Section 4.1. Section 4.2 details selected scenarios with particular relevance to automotive engineering. Regulatory and standardization aspects are outlined in Section 5. Section 6 concludes with an outlook on trends and developments.

Given the several cellular mobile network families, their various evolutionary stages and generations, and the enormous changes in system and service requirements, it is very challenging to provide a concise, yet comprehensive, view on corresponding architectural principles. Nevertheless, an effort is being made as follows, with a few disclaimers to be considered:

- Within each cellular communication system family, a number of gateway functions have been introduced to provide backward compatibility, interworking, or fall-back capabilities. For simplicity reasons, these gateway functions and protocol converters are largely omitted. Some operators of cellular and wireline communication networks had different technology families in operation at the same time. Corresponding interworking functions and aspects related to specific adaptations of network operation components or business support functions are omitted.
- Functionalities dealing with country-specific legal or regulatory concerns, or the so-called market adaptations, are not presented or discussed. The same holds for circuited switched voice specifics going beyond typical needs in automotive engineering. Intelligent network (IN) features, integrated services digital network (ISDN) characteristics, and multimedia messaging service (MMS) messaging capabilities are being of less relevance in the automotive engineering context and hence omitted.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

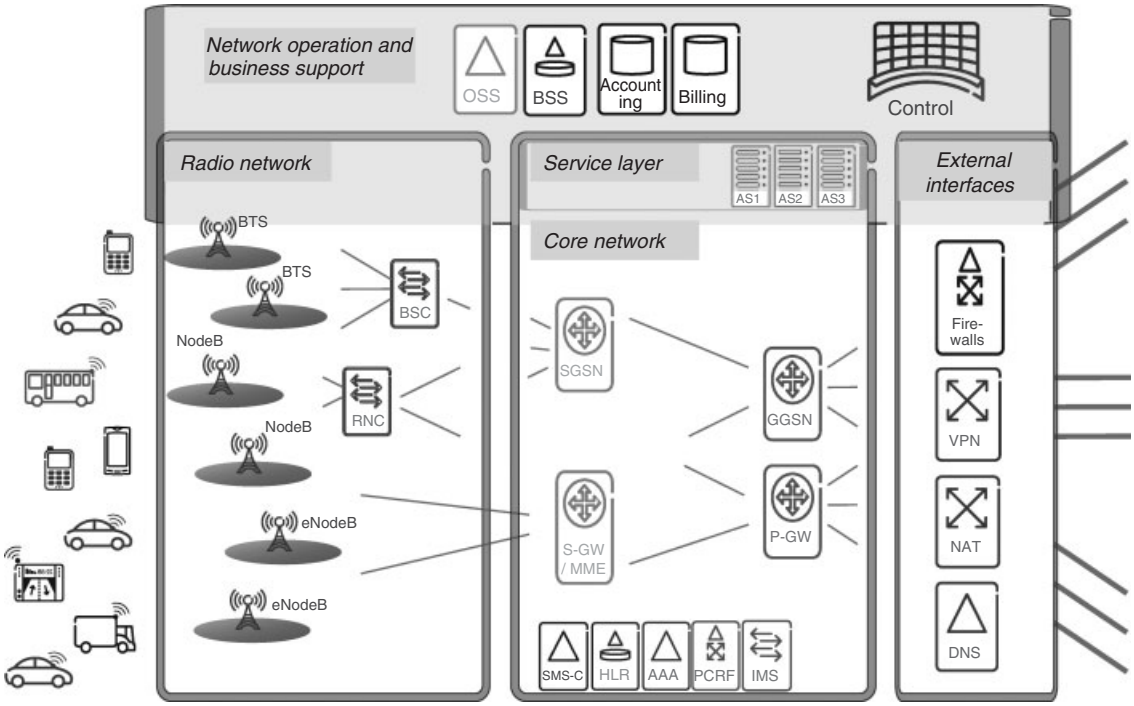


Figure 5. Basic domains of a reference cellular network.

- On the basis of the technology trends discussion in Section 3.1, we will use the 3GPP system family as architecture reference. System functions and terminology have changed many times when moving from 2G to 4G designs. For the sake of simplicity and without any claim of consistency, we will use common system component terminology for explaining the principal interactions and just refer to related names at other technology generations.

From a high level logical abstraction point of view, a cellular network architecture can be considered as a tree-like consolidation and control structure, with the geographically spread out mobile cellular terminals (Section 3.2.1) on the one end and a virtually central access node, for all incoming and outgoing data and voice communication to or from a particular mobile network, on the other end (Section 3.2.3).

The basic domains of a cellular mobile network, in a single operator instantiation, are illustrated in Figure 5. The radio sites, with its base stations (BSs), antenna systems, BS control notes, and the corresponding transmission networks, constitute the radio access network (RAN). The remaining domains are forming the core of the operational network, with its service layer extensions, interfaces to national and international connect, and operational support systems (Section 3.2.5). The latter is a cross-section function, spanning from the RAN nodes to the external interfaces (Section 3.2.4). These domains will be described in turn.

3.2.1 Mobile terminal

From an end user perspective, the most tangible part of a cellular communication system is the device that terminates the connection on his or her end. These devices come as a smartphone, a USB (universal serial bus) stick, connected navigation device, embedded vehicle module, or in many other forms. Their common denominator is the capability to interact with the mobile network according to the agreed standards. We will focus on this common aspect, and will, therefore, also consistently use the term *user equipment* (*UE*) throughout the following sections, which is common in 3GPP standardization.

The UE is identified through two main identifiers: The international mobile equipment identity (IMEI) and the international mobile subscriber identity (IMSI). While former identifies the physical device, the latter identifies its current user, also referred to as *subscriber*. The format of the IMSI is internationally harmonized (International Telecommunication Union Recommendation E.212, 2011) and is used in the same way in all interoperable mobile networks around the world. It consists of a three-digit mobile country code (MCC), followed by a two-digit or

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

three-digit mobile network code (MNC) and an identification number of the subscriber within the network. MCC and MNC together uniquely identify a mobile network and, therefore, also enable connecting to a user's home network when abroad (Section 4.1.4). Note, however, that the home network does not necessarily need to be physically located in the country indicated by the MCC, which is, for example, useful for complex multicountry operators. There is also a dedicated "world-wide" MCC (901) for global systems, which is, however, mainly used for satellite communication networks.

In 3GPP networks, the IMEI is hardcoded in the UE, whereas the IMSI is stored in a small chip card, commonly called *subscriber identity module* (*SIM*). Technically, this chip card is called *universal integrated circuit card* (*UICC*), which can hold different so-called applications. One of them is the SIM for 2G GSM; others are the USIM for 3G UMTS networks or the ISIM for the Internet protocol multimedia subsystems (IMSs). UICCs also come in various form factors, including surface-mounted versions that can directly be soldered onto a printed circuit board (PCB). Along with the IMSI, the SIM stores various other information such as the credentials that are used in conjunction with a personal identification number (PIN) to authenticate the UE to the network. A major advantage of the SIM is that, unless a soldered version is used, the IMSI can easily be transferred from one device to another, for example, when buying a new phone. In CDMA2000 networks, the IMSI is traditionally hardcoded on the UE, which makes hardware changes more cumbersome. Lacking a SIM and hence the ability to securely store credentials, CDMA2000 is using a username/password authentication schema to authorize the network access. CDMA2000 devices, however, increasingly use the UICC that contain a special code division multiple access-SIM (CSIM) application, so that also here the IMSI is no longer tied to the device.

In addition to the SIM, various other software components may be needed to complement the UE. Examples are a client for the IMS or common enablers such as the GeoMessaging client proxy (Jodlauk, Rembarz, and Xu, 2011) used for geographically targeted distribution of unicast messages. In CDMA networks, additional software may also be required to terminate certain communication protocols such as the point-to-point protocol (PPP) or mobile IP.

To manage radio resources, all relevant mobile networks implement different communication states for the UEs. The two most important ones are the idle state, that is, when the UE does not have an active data transmission and the active state that is used when actively transmitting data. All 3GPP systems have the notion of states. In addition, CDMA2000 networks implement such concepts,

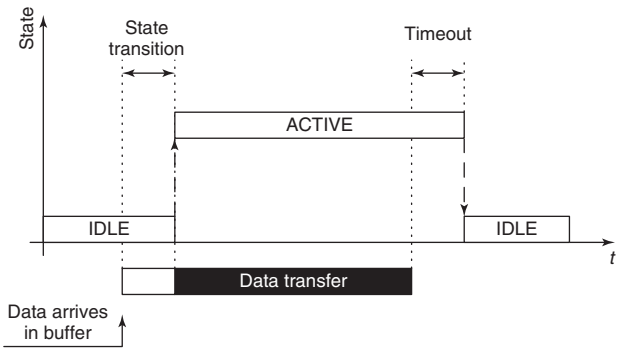


Figure 6. Transmission states.

for example, the dormant state. Most modern systems use a state model with more than two states, for example, the discontinuous reception (DRX) mode.

Although very different in terminology and implementation, all systems share a basic operating principle that is illustrated in Figure 6. When a UE first wants to transmit data, it needs to go through a certain procedure to become active, for example, to synchronize its timing with the network or to allocate network resources. After this state transition, the data transfer can commence. Following a certain inactivity timeout, the UE falls back into an idle mode that releases network resources and saves battery. A practical implication of the state model is that a certain time has to elapse before a first data transmission can start. Transitioning from idle to active mode required several seconds in UMTS and HSPA systems, with LTE this time has been reduced to <100 ms.

The UE is integrated into automotive communication in different manners:

- Fully embedded systems have an on-board unit (OBU) with integrated UE. These devices do not rely on any other device to establish a network connection. A prime advantage of this approach is its optimal connection to the on-board vehicle infrastructure, such as roof-mounted antennas or in-car communication buses.
- Carry-in communication concepts allow a user to provide the connectivity using a user-owned mobile phone. A cell phone, for example, can be tethered to the car-integrated communication network, for example, using a cable, Bluetooth, or Wi-Fi connection. One option is to have the application logic reside in the car, so that the mobile phone only operates a modem. Another option is to execute the applications on the phone and transfer the content to the in-vehicle systems using well-defined interfaces (Car Connectivity Consortium homepage). One major advantage of a carry-in concept is that no UE needs to be integrated into the car,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

which saves cost and makes upgrades to new network features easier. In addition, the connection costs are paid by the owner of the phone.
- Hybrid systems combine both approaches and split the communication between an on-board UE and a carry-in device. The on-board connectivity is typically used for car-centric services, such as remote door unlock, whereas the carry-in connectivity is used for driver-centric services, such as infotainment. As a variation of a hybrid system, a car may also offer a dedicated slot to insert a user-owned SIM only to be used for driver-centric services.

It is common that the communication modules that are finally built into vehicles support multiple of the network technologies from the various network generations (Figure 2), explained in the following sections. It should, however, be noted that not in all cases all possible frequency bands for a certain technology may be available in a given module. Especially, when a large number of frequencies are standardized for a certain technology, it is a common practice to introduce localized versions of the modules that support a subset of frequencies commonly used in the given region.

3.2.2 Radio access network

From an end-user perspective, the RAN is the most visible element of a cellular wireless communication network. UE, that is, cell phones, wireless modems or modules (Section 3.2.1), or the corresponding parts of a vehicle communication platform, constitute the mobile leaves of a large logical tree structure, as indicated in the introduction.

Once activated, the UE attaches to the so-called radio base station (BS). In 3GPP, the BSs (Figure 5) have different names, depending on the system generation: in a 2G GSM system architecture, they are called *BTS* (*base transceiver station*); in a 3G UMTS system, the BS is called *NodeB*; in 4G LTE, it is called *evolved NodeB* or *eNodeB*. The main task of the BS is to terminate the wireless radio link to the UEs, manage the radio resources in its coverage area, and facilitate UE mobility, for example, handling the communication hand-over procedures when a UE is moving from one serving BS to another. The preconditions and the more detailed procedures for a successful attach of a UE to a certain BS are detailed in Section 4.1.1.

Further on, a BS is the most remote part of a cellular communication network, that is, it represents the connectivity end point to the CN (Section 3.2.3) as illustrated in Figure 5. In 2G and 3G networks, additional active network nodes were introduced between the BS and the CN. In 2G systems, these nodes are called *base station controllers* (*BSCs*). In 3G systems, these nodes are named *radio network controllers* (*RNCs*). They serve as an additional control level for mobility management and radio resource management. Owing to the tree topology of the transmission network, voice and data traffic within a geographic area is aggregated at the sites where these elements are placed. With the 4G LTE network architecture simplifications, those nodes disappeared. Corresponding system functions were moved into the eNodeBs or were taken over by 4G CN nodes.

The connectivity between the elements of the radio network and the CN has been implemented through various transmission network components. These so-called backhaul networks were traditionally based on copper cables. Nowadays, these connections have largely been replaced by fiber-optic cables. In case of very remote BS sites, directional microwave transmission networks are used. With the enormous technology development in recent years, it is anticipated that future 4G and 5G radio networks will even be able to use their own primary radio interface for connecting itself to the CN infrastructure (self-backhauling).

3.2.3 Core network

The CN is the central data handling and control domain of a cellular mobile network. It is surrounded by a number of CN support functions, shown as network support systems in Figure 5, and a service layer, with its various application servers (ASs). The CN is complemented by an overarching networks and business operation support domain (Section 3.2.5) and the external interfaces (Section 3.2.4).

While the CN support functions constitute typical elements for network control, naming, and addressing schemas, the complementing service elements provide functionality for subscriber and UE management as well as service layer extensions, such as databases, look-up services, authentication and authorization controls, or messaging functions. Some of them are briefly described in Sections 3.3 and 4.

The strictly hierarchical structure, indicated at the introduction to Section 3.2, continues toward the CN. Incoming data traffic (seen from a BTS or RNC perspective) is handled by SGSNs (serving general packet radio service support nodes). Consolidated outgoing data traffic is handled by the GGSN (gateway general packet radio service support node).

In 2.5G (GPRS) or 3G (UMTS) systems, the GGSN (Figure 5) represents the termination point for all external IP data traffic to the CN and for IP traffic initiated at UEs and tunneled through the interim nodes and links between a UE and the GGSN. In terms if IP routing, it is the last

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

IP hop before the UE. Consequently, the (virtually single) GGSN node of a cellular mobile network provides a port for injecting IP data traffic that potentially can reach any mobile terminal within the entire cellular network of a particular operator. On the other way round, all IP data traffic originated from a UE attached to the cellular network must pass through the GGSN node before leaving the mobile network. In other words, the GGSN represents a distinguished connection port for data consolidation, traffic inspection, access control, and charging purposes.

With the introduction of 4G LTE systems, a number of more drastic changes have been introduced to the CN design and its RAN counterparts. The innovations have started on the radio side. LTE provides more efficient use of the spectrum with wider spectral bands reserved than its predecessors. This results in greater system capacity and performance. At the same time, the CN needed to change to provide a more efficient structure capable of handling high data throughput. This goal has been reached with the so-called *EPC* (*evolved packet core*), first specified in the 3GPP Release 8 standards that were finalized in 2009. It features a very much simplified and improved flat all-IP network architecture, spanning from mobile handsets and other terminal devices with embedded IP capabilities, over IP-capable eNodeBs (LTE BSs), a renewed and capacity-enhanced backhaul connectivity network (Section 3.2.2), across an EPC network, throughout the application domain—with and without IMS extension (Poikselka *et al*. 2006; Camarillo *et al*. 2011)—to the external interfaces, illustrated in Figure 5.

Four key elements of the EPC are the Serving Gateway (S-GW), the Mobility Management Entity (MME), the Packet Data Network Gateway (PDN-GW, or P-GW), and the Policy and Charging Rules Function (PCRF). The S-GW and the MME are typically co-deployed and thus illustrated as "S-GW/MME node" in Figure 5. The S-GW is a data plane element whose primary function is to manage user-plane mobility and acts as termination point of the packet data CN interface to the 4G RAN. The MME performs the signaling and control functions to manage the network access of a UE, the assignment of network resources, and the management of UE mobility states to support tracking, paging, roaming, and handovers.

Analog to the S-GW, the PDN-GW, or P-GW, in Figure 5, is the termination point of the external interfaces toward the packet data CN. Being a further anchoring point for external data network sessions (Section 3.2.4), the PDN-GW supports policy enforcement features (applies operator-defined rules for resource allocation and usage), packet filtering (e.g., deep packet inspection for application type detection), and charging support [e.g., per URL (uniform resource locator) charging]. Thus, the PDN-GW is comparable to the GGSN in a GPRS/UMTS CN. It should be noted that even though the transport network has become fully IP-capable, the IP traffic to the UE is still encapsulated in tunneling protocols in order to support mobility. From an external interface perspective, both nodes, therefore, appear as high capacity IP routers that constitute the last IP hop before the UE.

The fourth node introduced with the EPC is called *policy and charging rules function* (PCRF). The generic 3GPP policy and charging control model supports service data flow detection, policy enforcement, and flow-based charging. It provides means required for dynamic policy and/or charging control. 3GPP Release 8 enhanced the scope of the initial PCRF functionality to facilitate also policy and charging control functions for non-3GPP access to the network (i.e., Wi-Fi) or for fixed IP broadband access.

The network elements described so far are assisted by a number of network support functions. These functions ensure end-to-end quality of service (QoS) communication characteristics, implement policy control features, provide network security means, handle address name resolutions with domain name services (DNS), perform network address translation (NAT), and support IPv4 and IPv6 address allocations. The PCRF is considered part of these network support functions.

With UMTS Release 5 and with the accelerating development toward an all-IP CN design (Olsson *et al*. 2009), a call and session control system called the *IP multimedia subsystem* (IMS) (Section 3.3.2) was introduced. IMS supports IP-based multimedia and voice applications and is designed for both wireless and wireline networks. It is based on IETF[1] protocols such as the Session Initiation Protocol (SIP) or Diameter, which have been extended for the use in 3GPP networks. IMS simplifies the establishment and handling of end-to-end IP sessions, across multiuser shared radio and wireline links. As a side effect, and with particular relevance for automotive engineering, it supports service-oriented billing and QoS assignments by relating different service groups (such as telematics safety services and infotainment services) to different IMS sessions.

IMS-initiated QoS requests are further processed and implemented by PCRF node, described earlier. Depending on current system load, established QoS requests, QoS classes, and priority schemas, the PCRF responds to the requesting system node. Most modern PCRFs are multitenancy capable so that it can support 3GPP cellular networks with and without IMS extensions at the same time.

The IMS system itself consists of a number of functions, linked by standardized interfaces, and grouped to form one IMS administrative network. The call session control function (CSCF) is used to process SIP signaling packets in the IMS subsystem. There are three types of CSCF:

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

proxy call session control function (P-CSCF), serving call session control function (S-CSCF), and interrogating call session control function (I-CSCF). Each CSCF node is an SIP proxy with routing capabilities, linked to control and charging functions of its underlying cellular network. The P-CSCF is the initial SIP signaling contact point for an IMS-capable UE. It is responsible for forwarding SIP registration messages from the UE to the I-CSCF and subsequent call setup requests and responses to the S-CSCF. The P-CSCF maintains the mapping between logical subscriber SIP URI address and the physical UE IP address and its security association, for both authentication and confidentiality.

As the CSCFs are pure control functions, they rely on network elements such as media gateways or media resource functions (MRFs) to handle the actual data traffic, for example, for on-the-fly media processing or further service enrichments such as multiparty calls.

Application servers (ASs) host and execute system or external services, and interface with the S-CSCF using SIP. The MRF provides media-related functionality such as media manipulation (e.g., voice stream mixing) and playing of tones and announcements. A breakout gateway control function (BGCF) is a SIP server that includes routing functionality based on telephone numbers. See (Poikselka *et al*., 2006; Camarillo *et al*., 2011) for a deeper description of the IMS system, additional functions and architecture variants.

For the network access, the UE can connect to an IMS network in various ways using IP. An IMS-capable UE can register directly on an IMS network, even when they are roaming in another network or country (the visited network). The only requirement is that the UE can use IP (typically IPv6, but also IPv4 in early IMS deployments), can run IMS-compatible SIP user agents, and has the corresponding IMS identities provisioned. All modern IP-capable access networks are supported. Older systems, such as "plain old telephone service" (POTS—the old analog telephones), H.323, and non-IMS-compatible VoIP systems, are supported through gateways.

The service layer, indicated in Figure 5, with a multitude of ASs, is complementing the CN with subscription management, access right validation, UE location functionality, and much more. Note that some nodes described here as part of the service layer can also regarded as being CN components, depending on the context. In this context, we focus on selected nodes with central coordination roles.

The home location register (HLR) acts as centralized database node, providing most static information about all subscribers and subscriptions, within the whole cellular network. Thus, it constitutes the central database of a 2G cellular network for subscription-related information. The HLR is typically colocated with the authentication centre (AuC). The AuC is a function to authenticate each SIM card that attempts to connect to the GSM CN (e.g., when the UE gets powered on and initiates an access request to a cellular network). Once the authentication is successful, the HLR is allowed to manage the SIM and the services registered with the corresponding subscription. An encryption key is generated that is subsequently used to encrypt all wireless communications [voice, SMS (short messaging services), etc.] between the mobile phone and the 2G GSM CN. For more information about the attach/detach procedures, see Section 4.1.1.

Similarly to the hierarchical transport connectivity structure, indicated in Figure 5, the HLR is assisted by a set of connected databases for temporary subscriber data, such as the currently attached BTS and corresponding serving CN nodes. These associated databases are called *VLRs* (*visitor location registers*). With the introduction of 3G UMTS, the HLR, AuC, and some further functions got consolidated into the home subscriber server (HSS).

Together with additional service-layer security means, the authentication, authorization, and accounting (AAA) functions provide one of the most value-adding features in the 3GPP network architecture. This set of functions, complemented with a dedicated legislative framework and country-specific regulatory requirements, provide the foundation for cellular MNOs becoming a kind of trusted service providers (SPs). Some operators have leveraged that position and even applied for a banking license to provide even mobile payment solutions.

In the context of cellular networks, one can distinguish two groups of services: first of all, there is a wide array of services that are either placed on the Internet or only available within the system domain of a certain network operator. A second group, however, is the so-called standardized telecommunication services. The latter are available in almost all cellular networks, work with any suitable device, and are interoperable between different cellular mobile networks.

Examples of standardized telecommunication services are: to put a voice call on hold, to have the ability of forwarding a voice call, or to get a call-waiting indication. However, also modern IP-based services for presence, chat, and video conversations fall into the category of standardized services that are interoperable and supported across network operator boundaries. Furthermore, messaging services such as the MMS, rich communications services (RCS), and SMS are standardized services. Only the latter is of bigger relevance to the implementation of telematics services and thus is described in more details in Section 4.2.1.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.2.4 External interfaces

The prime integration layer for external data communication services in Figure 5 is indicated by the external interfaces domain. Despite some shared functionalities, such as DNS, NAT, network access control, logging or monitoring functions, and dedicated firewalls, one can distinguish three types of interfaces to external SPs or communities:

1. Type 1: Connections to network elements of other cellular or fixed-line operator networks.
2. Type 2: Connections to specific enterprise networks.
3. Type 3: Connection to the public Internet.

Connections of type 1 are typically linked via special very high speed international network connections (GSMA IR.67, 2012). There are dedicated operators providing these high speed long distance and inter-Atlantic connectivity capabilities, an important precondition to provided roaming capabilities (Section 4.1.4) between different operator networks. Two groups of data protocols are used with type 1, namely IP and non-IP protocols. The latter are standardized lower-layer protocols to interconnect primarily 2G cellular systems and wireline telephony systems of various kinds. The IP links within type 1 are tailored to support end-to-end QoS across network operator domains, enable network congestion control on global levels, and link the various accounting and control functions, being engaged in multioperator data and service scenarios.

The external interfaces in types 2 and 3 are predominantly IP based. Type 2 is a value-added service that cellular MNOs provide primarily to enterprise customers (GSMA IR.67, 2012). APN configurations are provided to these enterprises that make it possible to connect to a virtual connection point at a specific GGSN or P-GW (Figure 5).

The concept of virtual private networks (VPN) can be used to provide end-to-end secured network connections from, for example, an enterprise network to a UE or vice versa. A VPN user typically experiences a VPN-connected UE as if being directly connected to a specific enterprise network.

From an external network's point of view, the GGSN and the P-GW are high capacity routers to a subnetwork, because they "hide" the further CN infrastructure from the external network. When the GGSN (or P-GW) receives data addressed to a UE (respectively to a specific user), it checks if the UE is active. In that case, the GGSN (or P-GW) forwards the data to the SGSN (or S-GW), which is serving the UE. UE-originated packets are routed to the target destination network by the GGSN (P-GW) routing function.

Access to the public Internet in type 3 is rather similar to type 2, considering the public Internet as a kind of specific enterprise network. VPN technology can be used to establish secured connections through the public Internet. This advancement has mostly replaced the need to provision and maintain expensive dedicated leased-line telecommunication circuits that used to be a common practice in wide-area network installations just a few years back.

3.2.5 OSS and BSS support systems

This topmost domain in Figure 5 is a cross-sector function with two complementing duties. The operations support system (OSS) most frequently describes "network systems" that cellular operators use for supporting processes such as maintaining network inventory, provisioning of services, configuring network components, and managing faults. The complementary term *business support systems* (*BSS*) typically refers to "business systems" dealing with customers, supporting processes such as taking orders, processing bills, and collecting payments. The two systems together are often abbreviated OSS/BSS.

Naturally, only a few elements of the OSS/BSS domain are described in this context. Elements mostly relating to the OSS subarea are control and monitoring functions for network elements and system nodes. Communication network related data traffic inspection, data volume monitoring and accounting components, QoS supervisions, network policy controls, and network security monitoring elements may add to the OSS.

The BSS system is encompassing all functionalities related to service accounting and billing, the processing of charging data records (CDRs), such as CDR mediation functions and real-time rating functions for prepaid subscribers. It also contains self-provisioning portals for enterprise customers and all functions dealing with inter-operator key performance indicator (KPI) supervision, roaming-cost related data capturing and processing, as well as data back-up handling, or country-specific business and operation adaptations.

3.3 Network security

The basic security principals of today's mobile networks have been included in the 2G GSM network already, using authentication mechanisms and encryption based on the SIM (Section 3.2.1). In 3G systems, UMTS, and later in LTE, these basic principles remain and have been further developed. Especially, in LTE and IMS on service layer, Internet and Web security mechanisms [HTTP (hypertext transfer protocol), AKA (authentication and key agreement), and internet protocol security (IPSec)] are included as well.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.3.1 Basic telecommunication security principles

Providing data security and tapping protection is a key requirement for communication networks. As it is practically impossible to protect all network nodes and links between these nodes against unauthorized physical or logical access, it must be ensured that the information itself is protected. The requirement to be no more vulnerable to eavesdropping than fixed phones was considered in the design of GSM. It addresses these goals by providing user-related security features for authentication, confidentiality, and anonymity.

The authentication feature is intended to allow a network operator to verify the identity of a user such that it is practically impossible for someone to make fraudulent calls by masquerading as a genuine user. Confidentiality protects the user's traffic, both voice and data, and sensitive signaling data, such as dialed telephone numbers, against eavesdropping on the radio path. An anonymity feature was designed to protect the user against someone who knows the user's IMSI from using this information to track the location of the user or to identify calls made to or from the user by eavesdropping on the radio interface.

A well-known security feature, introduced with GSM, is the use of a chip card, the SIM (Section 3.2.1). The SIM contains all the identification and security-related data that the subscriber needs to make or receive a call. It is, in effect, a portable security module, personalized for the subscriber. The SIM can be used to access services in any network with which the subscriber's home network has a roaming agreement (Section 4.1.4).

During roaming, the subscriber's home network provides all the data needed by the serving network to operate the security features without revealing any of the sensitive security data stored in the subscriber's SIM.

UMTS and LTE security builds on the success of GSM by retaining the security features that have proved to be needed and that are robust. As in GSM, a smart card is used in UMTS and LTE to store all the identification- and security-related data that the subscriber needs in order to make or receive a call. In UMTS and LTE, the GSM security features have been improved, for example, extending the key length from 64 to 128 bit. This key is used for the generation of a ciphering key that is needed for the symmetric encryption of the user data.

3.3.2 IMS security

On top of the network-level security mechanisms described in Section 3.3.1, IMS (Sories, Huschke, and Phan, 2008) as service layer management system with an underlying more general security architecture (as specified in 3GPP TS 33.20331) can be used.

An IMS subscriber will have a IP multimedia private identity (IMPI), which is authenticated and stored beside all other relevant subscriber data in the central repository of the IMS system, the HSS (Section 3.2.3). During user registration, the subscriber data is transferred from the HSS to the S-CSCF. Hence, upon request by a user, the S-CSCF can match this request with the subscriber profile before access is granted, such that the home network can control access.

Overall, IMS provides five security associations in order to secure the communication between all relevant nodes of the home or a visited cellular network. The authentication and key agreement for IMS is based on the same mechanism as in UMTS.

An additional feature relevant in this context is the generic bootstrapping architecture (GBA) (3GPP TS 33.220, 2012) that makes it possible for services to reuse the existing security association between terminal and network. GBA can create special session keys during the standard authentication procedure that can later on be used to authenticate the user to a service.

As a conclusion, it can be noted that mobile networks feature extensive security mechanism that are already part of the standardized and commercial available networks. Mechanisms such as the GBA even facilitate the reuse of these proven security associations for automotive services and for corresponding end-to-end solutions.

4 END-TO-END PROCEDURES

This section provides some sample end-to-end procedures, illustrating the interworking of the various cellular network elements introduced in Section 3.2.

4.1 Basic procedures

This section delivers an overview of the most important procedures in a 3GPP cellular network. We focus on the general operating principle that is common to most systems. Detailed descriptions are available in (Dahlman, Parkvall, and Sköld, 2011; Camarillo *et al*., 2011; Olsson *et al*., 2009).

4.1.1 Registering to a network

When a UE (Section 3.2.1) is first switched on, it has to attach to the network in order to be able to establish and receive calls and connections. The procedures slightly differ depending on the exact network used, but the principles are very similar. As a very first step, the UE detects available

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

radio networks. For this purpose, all radio BSs constantly broadcast a defined set of system information, among them the identification of the mobile network and the current cell and area identifier. Different area identifiers exist, which are, however, conceptually very similar. Here, the example of the tracking area (TA) or location area (LA) code is used, which will become important later on. Using this broadcast information, the UE can determine which mobile networks are available. It will likely see more than one radio BS per network. In the simplest case, the UE will receive a signal from its home operator and will subsequently contact the BS with the strongest signal to initiate the attachment and registration procedures.

After some initial steps to synchronize the radio communication, an authentication phase follows (Section 3.3.1). The UE uses the credentials stored on the SIM card to authenticate against a corresponding set of credentials stored in HLR or HSS. The user identifies to the network using the subscriber identity (IMSI). Upon successful completion of the authentication, the user is marked as reachable in the network databases (Section 3.2.3), along with the TA and LA it is currently in.

The connection setup in 3G and 4G networks typically also involves establishing the so-called default bearer and obtaining an IP address. This bearer can then be used for basic IP communication. When the UE is IMS-capable (Poikselka *et al*., 2006; Camarillo *et al*., 2011), the IP connectivity is also used to register the user to the IMS. In addition to the default bearer, dedicated bearers with specific characteristics can be established at a later stage (Section 4.2.2).

When no data or voice transmission is required after this phase has been successfully completed, the UE turns into a power-saving idle mode (Section 3.2.1) in which it only monitors a special set of broadcast channels.

4.1.2 Call and session handling

For any kind of connection establishment, two basic cases have to be distinguished: connections initiated by the UE (mobile originated) and connections initiated by another party (mobile terminated).

For mobile-originated connections, the UE contacts the mobile network using a signaling channel. For 2G and 3G systems, dedicated signaling channels and protocols are available. For all-IP networks, for example, based on 3G or 4G, typically, the IP-based IMS (Poikselka *et al*., 2006; Camarillo *et al*., 2011) is used. The mechanisms in the mobile network locate the other party or parties, and assign the network resources to transport the actual data. In 2G GSM systems, this includes assigning the TDMA timeslot to be used. For IMS systems, this involves setting up a dedicated bearer for the IP communication related to the connection. Such a bearer is mapped to an appropriate QoS class (Section 4.2.2), which, for example, ensures that interactive voice or video conversations get priority handling in the network. The signaling mechanisms in IMS are flexible enough to set up various data sessions for virtually any application; they are not limited to setting up voice, video, or similar. A possible automotive use case would be to set up a generic data connection with special requirements, for example, with prioritized delivery for time-sensitive messages of a cooperative ITS system. This integrated process that includes resource reservations in the network is one of the major differentiators to purely Internet-based voice or video services.

Mobile-terminated calls generally follow a similar pattern. Before the procedure described earlier can, however, start, the network may need to locate the UE. When the UE has an ongoing, active connection, the network knows exactly where and how the UE can be contacted. If it is in an idle state (Section 3.2.1), the TA or LA stored in network databases (Section 3.2.3) are used to determine the approximate area in which the UE has last reported its position. As this area typically involves a larger number of radio networks cells (Section 3.2.2), a paging message is sent over the broadcast channels of all cells belonging to the area. The UE monitors these channels, receives the paging message, and establishes a connection to the radio network. The remaining procedures are carried out similarly to the mobile-originated cases. Note that also advanced features such as special sessions involving dedicated bearers can be initiated by the calling party, so that, for example, an AS in the network (Section 3.2.4) can proactively set up connections without waiting for the UE to initiate them.

4.1.3 Mobility management

Being able to maintain a connection while moving between cells (Section 3.2.2), or staying reachable despite constantly changing the locations, is a crucial functional aspect of cellular networks. The process of dealing with changing UE (Section 3.2.1) locations is referred to as *mobility management*; the actual procedure to update the location is commonly called *handover*. Two main cases have to be distinguished: handovers while the mobile is actively transmitting data and handovers while the mobile is in a power-saving idle mode (Figure 6).

The latter is by far the easier variant. In this state, the network does need to know the exact position of the UE, so that signaling every cell change is not required. Rather, the UE listens to the broadcast channels of the cells it passes, and only contacts the network when it notices that

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the TA or LA code has changed. This procedure, which can sometimes be heard when the transmissions from the UE cause interference, for example, in a car radio, ensures that the network can track the location of the UE on a sufficiently high accuracy while minimizing the signaling load and the UE's battery consumption.

When the UE is active, the network uses a sophisticated set of procedures to keep the location changes transparent to the user. The central paradigm is "make before break," that is, the handover is prepared well in advance so that no data is lost in the process. Tunneling protocols encapsulate the data stream between UE and network, so that the actual data transmission is not impacted by the handover. This is also the reason why all user data flows through the GGSN or PDN Gateway (Section 3.2.3). These nodes constitute anchor points that remain stable despite any kind of location changes. Without such an approach, handovers may impact the end-to-end data flows, as IP addresses change or the data stream suddenly needs to be routed to another access point. Such behavior would almost always lead to an at least temporary disruption in the communication, and would likely not be accepted by users. Thus, when arguing for making the IP communication accessible closer to the radio BSs, there is always a trade-off with the ability to perform transparent handovers.

Anonymized location data can also be used to observe and predict the movement patterns of the UEs. Solutions exist that correlate UE cell movement characteristics to derive road traffic pattern, so that the mobility signaling data can be used to detect road traffic congestion.

4.1.4 Roaming

Besides the ability to perform transparent handovers, another success factor of modern cellular networks is that through standardized technology, it is possible to stay connected and reachable even outside the coverage area of the home network operator. This process is commonly known as *roaming*.

From a functional perspective, the network elements used in a roaming case are jointly provided by the home network operator and the so-called visited network operator. The radio network (Section 3.2.2) and some other network elements (Sections 3.2.3 and 3.2.4) are used in the visited network, whereas central functions such as the subscriber database are still located in the home network. Thus, when registering to a visited network, the UE (Section 3.2.1) will connect to the local RAN of the visited network operator. The visited network detects that the registering UE is connecting from a foreign network and hence directs the authentication to the user database of the home network (Sections 3.3.1 and 4.1.1). Upon successful registration, the user can use the visited network. There are different options on which network elements are used in the visited networks, a particularly important choice being whether the GGSN or PDN Gateway (Section 3.2.3) in the visited network is used. Nowadays, it is still a common practice in many networks to tunnel all user traffic back to the home network, as illustrated in the simplified example in Figure 7.

In such a setup, the PDN Gateway in the home network is still the termination point for all IP traffic. This ensures that all possible services and features work just as if being connected directly to the home network operator. With the increasing usage of data connections, this, however, becomes more and more of a burden, so that using a local breakout point in the visited network will probably increasingly become the standard system setup in the future, at least for normal Internet data traffic.

Physically, the data exchanged between the home and the visited networks is transferred through interconnect networks dedicated to MNOs (Section 3.2.4). To avoid PTP connections between each and every operator, major network SPs such as Deutsche Telekom or Telia Sonera (Ala-Luuko, 2006) operate the so-called IP Packet eXchange (IPX) networks (GSMA IR.67, 2010; GSMA

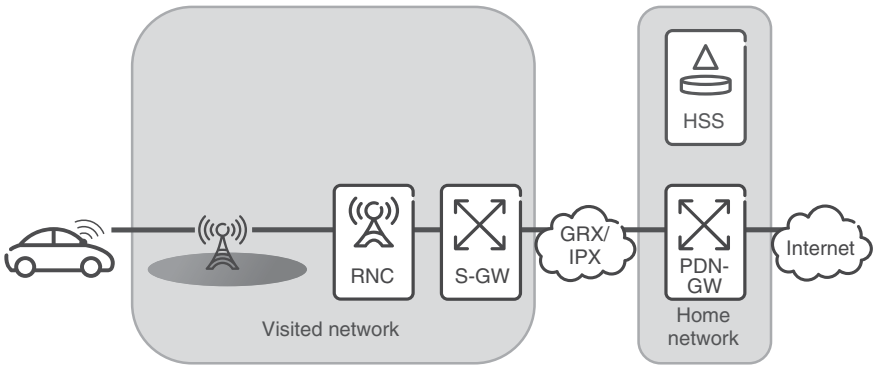


Figure 7. Simplified roaming network architecture.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

IR.34, 2012), an evolution of the earlier general packet radio service roaming exchange (GRX). These networks ensure interconnection under secure conditions and with telecommunication-grade quality. Network operators connect to one or more of these SPs, which again interconnect their IPX networks. Using IPX interconnections guarantees the high network connectivity quality enterprise customers are used from telecommunication networks, but of course also constitute a cost factor.

When roaming, a user is allowed to use the network even though the user does not have a direct contractual relation with the currently visited network operator. The standardized charging and billing mechanisms ensure that the network usage-monitoring is compatible to the system of the home network operator who will finally invoice the usage fees. There are additional fees in roaming cases, which are relevant especially for data connections. In addition, in many automotive use cases, roaming cost is a substantial cost factor in the cost calculation for telematics and infotainment services.

A prerequisite for roaming is that a user's home network operator has a roaming agreement with the network operator of the visited network. Nowadays, there is no practical limitation anymore, as almost all major network operators have agreements with at least one operator in all relevant countries around the globe. Besides these international roaming cases, it is also possible from a technical perspective to roam into another mobile network in the home country, for example, in case of poor coverage by the current cellular home network operator. While the benefit to the user would be obvious, this so-called national roaming is, due to regulatory restrictions, nowadays only allowed in exceptional cases. Typically, the regulators want to avoid the case that operators concentrate their network build-out on few lucrative areas and leave the coverage of the less business-attractive areas to other operators. In such an environment, there would be little incentive for anyone to build out good network coverage also in areas with little data or voice traffic density, such as rural areas. National roaming has, however, been permitted in cases where network sharing has helped providing better coverage in the network build-up phases.[2] Another constellation, where national roaming is usually not controversial, is in very large countries such as Russia or certain parts of Asia, where licenses are granted for subregions and where it is assumed that network operators will not be present in all those regions.

4.2 Advanced scenarios for vehicle communication

In the following section, a set of advanced capabilities, mechanisms, and functionalities of modern cellular networks are described. The scenarios detailed in Sections 4.1.1–4.2.4 are of specific relevance to automotive applications. It should be noted that the generalized capabilities are representing just a subset of a much wider portfolio, supporting a broad set of use cases and applications.

4.2.1 Cellular messaging capabilities

4.2.1.1 SMS and instant messaging services. In terms of messaging, SMS is one of the most prominent cellular services with over 8 trillion messages sent in 2011. The corresponding system control node is the short messaging services service center (SMS-C) in the service layer (Figure 5) of the CN (Section 3.2.3). The SMS-C implements a store-and-forward for SMS' to be sent. Under certain operation conditions or at certain popular events (such as New Year or a popular TV show voting), an enormous SMS messaging traffic occurs. As SMS is generally not designed to be a fully reliable service, such situation may lead to SMS-C overload situation with some messaging being dropped or becoming heavily delayed before delivering. A single SMS message size is restricted to 140 bytes or 160 text characters; however, concatenating messages is possible. Special SMS message tags can be used to make the SME appear directly on the main screen of the receiving UE (flash SMS) or to trigger auto-configuration actions. SMS can also be associated with premium services and special payment actions. With minor adaptations, the SMS service can also be made fully reliable, so that it can be used for delivering emergency notifications (emergency SMS), for example, directly to a public safety answering point (PSAP) in an emergency call scenario.

Owing to its relatively low cost and its built-in flexibility, SMS messaging capabilities were frequently used in early telematics and particularly in logistics solutions. With the global deployment of 2.5G GPRS technology (Section 2) and cross-border IP communication capabilities (Sections 3.2.4 and 4.1.4), many of these early telematics solutions were migrated to run over GPRS. With 4G LTE being an all-IP technology that does not natively support SMS, a migration wave is expected toward native IP-based messaging technology. An example for user-centric messaging is the enhanced rich communication suite standard (RCS-e), also known as *Joyn*, which shall be as interoperable as the SMS service, but add many features, such as video or group chat. As a backwards compatibility solution, the global systems for mobile communication association (GSMA) has lately released a SIP-encapsulation for SMS as part of the voice-over LTE standard (GSMA IR.92, 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.2.1.2 Broadcast and multicast services. The messaging service described in Section 4.2.1.1 is based on a PTP communication. The end points are either two telephones or, in the case of machine-to-machine (M2M) applications such as telematics services, a client–server connection.

Broadcast and multicast, on the other hand, are synonyms for point-to-multipoint (PTM) communication, where data packets are simultaneously transmitted from a single source to multiple destinations. The term *broadcast* refers to the ability to deliver content to all users. Known examples are radio and TV services, which are broadcasted over the air (OTA) (either terrestrial or via satellite) and cable networks. *Multicast*, on the other hand, refers to services that are solely delivered to users who have joined a particular multicast group. Ordinarily, a multicast group is a group of users interested in a certain kind of content, for example, traffic information. A multicast-enabled network ensures that the content is solely distributed over those links that are serving receivers that belong to the corresponding multicast group. This is thus a very resource-efficient way of delivering services to larger user groups. Multicasting was first introduced as Internet communication service.

Broadcast and multicast service capabilities for cellular networks were defined in 3GPP as the so-called multimedia broadcast and multicast service (MBMS) and in 3GPP2 as broadcast and multicast service (BCMCS). (Bakhuizen and Horn, 2005), for example, provide a comprehensive MBMS overview.

With MBMS, a new network element, the broadcast/multicast service center (BM-SC), has been introduced, providing a set of functions for controlling the broadcast/multicast delivery service.

The BM-SC serves as an entry point for content-delivery services that want to use MBMS. It sets up and controls broadcast or multicast data flows in the mobile CN (Section 3.2.3) via MBMS transport bearers and schedule and deliver efficient radio bearers (Section 4.2.2) for PTM radio transmission within a cell. The BM-SC also provides service announcements to end-devices (Section 3.2.1). These announcements contain all necessary information (such as multicast service identifier, IP multicast addresses, time of transmission, and media descriptions) that a UE demands in order to join an MBMS service. The BM-SC can be used to generate charging records for data transmitted from the content provider. It also manages the security functions specified by 3GPP for multicast mode.

In MBMS, different transmission scenarios are defined:

Broadcast mode: A pure broadcast mode provides a unidirectional PTM transmission. The broadcast mode cannot guarantee error-free reception and does not provide billing functionality. When a MBMS broadcast service starts, MBMS data is sent to all nodes of the UMTS network, from the SP down to all RNC nodes (Section 3.2.2) and to cells in the broadcast service area.

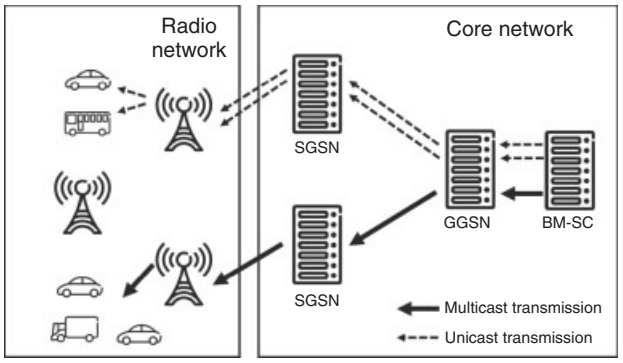


Figure 8. Unicast and multicast transmission principles.

Multicast mode: The multicast mode provides service only to groups of subscribed users. The serving RNC gets the data only once from the BM-SC by the SGSN/GGSN (Section 3.2.3). Billing functionality is included. When the MBMS multicast service starts, MBMS data is sent only to those network nodes and cells with subscribed users and its UE.

Figure 8 illustrates the differences between unicast and multicast. In unicast operation mode, each UE requests content for the same service from the BM-SC. The unicast transmission is visualized in Figure 8 as dashed arrows. From the UE toward the BM-SC (in uplink direction), the number of bearers to be handled by higher network entities increases. This is a consequence of the network's architecture hierarchy (Section 3.2) and the definition of unicast. The multicast transmission mode tries to serve the other way around: a single MBMS bearer is created at the BM-SC and is transmitted to each lower network entity requesting the associated MBMS service.

MBMS capabilities have been enhanced for LTE networks to meet additional requirements for the IP-based evolution of the cellular network architecture. Those enhancements are summarized with the term *eMBMS* (Dahlman, Parkvall, and Sköld, 2011). Correspondingly, MBMS data in LTE networks are sent from the BM-SC to the eNodeBs (Section 3.2.2) via IP multicast, a method of forwarding IP packets to multiple receiving network nodes in a single transmission. Therefore, MBMS is a resources-saving technology for the radio and the transmission network.

4.2.1.3 Location management and GeoMessaging. The cellular messaging mechanism described in Sections 4.2.1.1 and 4.2.1.2 enables the implementation of a wide variety of services, from PTP text, video, or

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

data transmission, to PTM transmission to all or a group of subscribed users, in a certain region. As MBMS was developed for the distribution of mass media services such as mobile TV, the intended distribution areas, the so-called MBMS service areas, are relatively large and static, for example, as the today's distribution areas of local radio and TV networks.

The GeoMessaging functionality (Jodlauk, Rembarz, and Xu, 2011; ETSI ITS TR 102 962, 2012; ETSI ITS TS 103 084, 2012) complements these messaging mechanism by making it possible to address recipients in a specific geographic area, which can be defined in a fast and flexible way with very fine granularity down to the size of a fraction of a single radio cell. A possible use case could be the need to inform all subscribers of a traffic information service about the traffic situation of a certain highway segment within a very short time. The basic aim of the GeoMessaging service (ETSI ITS TR 102 962, 2012) is, therefore, to deliver data from a source to all UEs within a specific geographic region. A basic prerequisite to this is that all UEs' locations are known, for example, UEs being equipped with a positioning system [such as GPS (global positioning system)], in order to determine if these nodes are located in the target geographical zone.

Owing to the hierarchical concept of cellular systems, the communication path from an AS toward a cellular-equipped car, as mobile node, is always passing several nodes of the cellular communication system (Figure 5). In order to realize GeoMessaging functionalities in cellular networks, a second basic requirement is to establish a centralized network node as dedicated GeoMessaging AS. This node receives and handles all locations and location updates from the clients (vehicles) and steers in a smart way the frequency of the location updates, such that the cellular network's signaling load, generated by localization updates, is minimized and the required precision of targeted distribution area is still maintained. Furthermore, this node distributes messages to the involved BSs (Section 3.2.2) in the concerned geographical area.

In order to realize vehicle-to-vehicle communication in cellular networks (Dietz, 2009; Gehlen *et al*. 2007; Sories, Huschke, and Phan, 2008; Phan, Rembarz, and Sories, 2011), the GeoMessaging AS may consolidate incoming messages to redistribute replication-free content back to the other vehicles in the proximity. Depending on the number of recipients in the cell, unicast or multicast transmission can be used as actual message delivery mechanism.

4.2.2 Quality of service mechanisms

Both the IP-based Internet and the PS domain work on the principle of "best effort," which means all services and applications have equal access to bandwidth, regardless of their actual needs. For example, downloading an updated navigation map is given the same priority as an Internet radio stream. The problem is that when downloading a file, the user will not experience any instant quality issues if the available bandwidth varies during the download, but when listening to Internet radio, having sufficient, constant bandwidth is crucial to the audio quality.

In the transition of telecommunications networks design from CS speech networks to all IP-based networks (PS domain), it is important that speech calls continue to be available at the same or a higher quality level than the so-called best effort services. More importantly, it must be possible to prioritize emergency calls and traffic hazard warning services to ensure that even if the networks are heavily loaded, for example, at New Year's Eve, these calls execute with a sufficient quality. This is not possible with a best-effort-only network. Therefore, the telecommunications industry defines mechanisms to ensure a specific QoS according to the individual service requirements. The QoS mechanisms defined for 3G networks were advanced and further improved in the specifications of LTE (Ekström, 2009).

Network resources can be adapted dynamically. For example, it is possible to reduce the allocated bandwidth of the wireless connection for a service if system capabilities of a vehicular user are temporary not able to cover the premium level of service. On the other hand, the provided QoS level of the network can also be reduced for nonessential services if bandwidth needs to be freed up for an emergency situation, for example, an accident involving multiple cars on a small stretch of a highway.

QoS handling in LTE networks is network controlled. Requests for altering the QoS levels can, however, be made by both the network servers and the UE. Network-initiated QoS is considered superior to UE-requested QoS. UEs can request a bearer with a specific QoS from the network; however, this requires a vendor-specific QoS API (application programming interface). By having to specify the QoS information for the bearer, it is required that the client application is "access QoS aware," that is, access-specific information is required in the signaling. UE-initiated QoS also prevents the network from modifying the existing bearer QoS settings if required.

Network-initiated QoS allows the client application to be "access QoS agnostic"; no special QoS API is required at the UE and the network can establish and modify the bearer when necessary. For example, a deep-packet inspection (DPI) function can be used to identify specific services. The required QoS parameters can be assigned and then carried over a standardized interface (Rx and/or Gx) (3GPP TS 23.203, 2012) to the network. However, in most cases,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

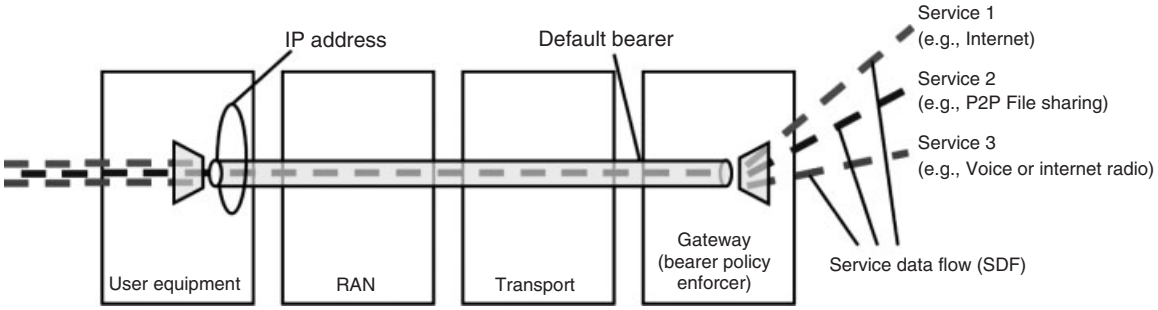


Figure 9. 3G and LTE quality of service support by dedicated bearers.

the client application is aware of the QoS that is needed in order to negotiate with the network the appropriate QoS class via application-layer signaling such as SIP (Rosenberg *et al*., 2002) and real-time streaming protocol (RTSP) (Schulzrinne *et al*., 1998). However, it should be noted that there is no access technology specific information in the signaling.

The implementation of the QoS mechanisms in the network depends on the specific cellular network technology. The QoS concept standardized in 3GPP Release 8 (Ekström, 2009) is leveraging the so-called "bearer" mechanism (Figure 9), which uniquely identifies the packet flows having the same QoS parameters.

The initial, basic connectivity is always established using a default bearer, which has no guaranteed bit rate (non-GBR). This bearer can experience issues such as packet loss because of congestion. Additionally, the UE can have one or more dedicated bearers. These can be non-GBR or GBR. A dedicated bearer with a GBR does not experience congestion-related packet loss. It is usually established "on demand," as it needs to reserve transmission resources to ensure the specific bit rate.

Each bearer, whether default or dedicated, GBR or non-GBR, is assigned a quality of service class identifier (QCI). This parameter references node-specific parameters (preconfigured by the network operator) regarding the user plane treatment (packet forwarding), for example, priority, or packet-error-loss rate. Additionally, an allocation and retention priority (ARP) parameter specifies how the setting up and retaining of bearers on the control plane is handled.

For dedicated GBR bearers, parameters such as GBR and MBR (maximum bit rate) are important for defining their priority within the network.

4.2.3 Flexible charging and billing

On the basis of the concept of IMS sessions (Poikselka *et al*., 2006; Camarillo *et al*., 2011) and dedicated bearers

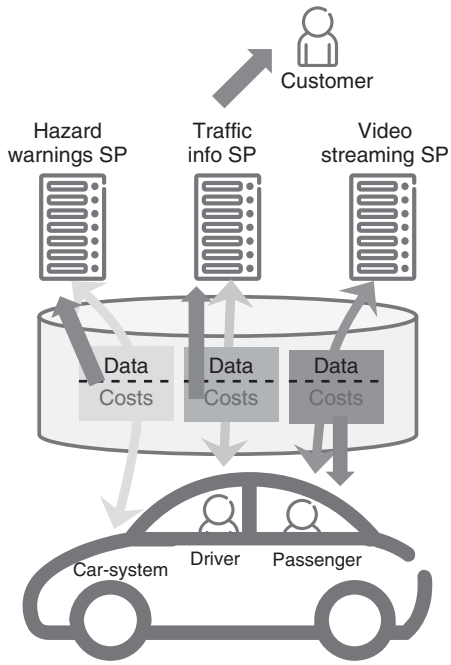


Figure 10. Flexible charging concept.

(Section 4.2.2) for each service, it is also possible to flexibly charge customers (consumers or enterprises) for the service use. The basic principle is explained in Figure 10.

Data streams, which are called *bearers* in the IMS context, are being separated within the cellular communication network (Figures 9 and 10). Thus, the charging-related parameters can vary across these data streams. In the example of having a "hazard warning" data stream, a "traffic info" data stream and a "video" data stream directed to different SPs; it is possible to charge the costs related to the data communication to different parties: the consumer (driver or passenger), an enterprise, or an SP. The enterprise entity, for example, will then be responsible for clearing the charges to their end-customers.

The outlined technical solution gives the flexibility to support a wide range of business models and differentiate

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

between regional markets and customer segments in order to address the ongoing discussion on suitable business models for cooperative vehicle services.

4.2.4 *Lifecycle management*

The provisioning process of mobile communication subscriptions is directly connected to the lifecycle of cellular-connected vehicles (Figure 11). Hence, it will be different to the provisioning of subscriptions to cell phones. Cell phones are typically produced and delivered to dealer shops without any subscription. At the time the phone gets sold to the consumer, or some time later on, the customer inserts a SIM (Section 3.2.1) and the technical relation between the phone and the cellular network, corresponding to the business relation between the SIM holder and the network operator, gets established (Section 4.1.1). In this context, it is important to mention that the SIM is part of the network architecture. It is typically owned by the cellular network operator. The SIM is the key element in the provisioning process and in terms of security (Section 3.3.1). It is used to authenticate the phone subscription to the network and to encrypt the communication link. In case a corresponding connected device becomes tightly integrated with the vehicle electronics (e.g., soldered to a PCB of a car communication unit), the traditional provisioning process may no longer be feasible.

Figure 11 illustrates a sample lifecycle of vehicle and related online services. After production, test, transport, and sale of the car, car-centric services such as eCall, remote diagnostics, software- or map updates may be activated. At the same time, first customer centric services, such as Internet access, could be subscribed to by the first owner of the car. Additional online services (such as the Web radio in Figure 11) may be used by the first owner during a shorter period of time. All of these customer-centric service subscriptions shall end after first resale of the car. Following the suspension phase, the second owner of the car should be able to subscribe to other set of services. All of these services shall be terminated when the end of life of the vehicle has been reached and the car gets scrapped.

From a technical point of view, a wide variety of provisioning mechanisms and methods are feasible to accommodate such a procedure. Several of them are implemented nowadays. The following description exemplifies two such advanced provisioning schemas.

One approach is to use the advanced provisioning process based on IMS capabilities (Poikselka *et al*., 2006; Camarillo *et al*., 2011) of modern mobile communication networks (Sections 3.2.3 and 4.1.2). The key capability is the ability to extend and alter subscription and session identity data at the SIM card, installed within the vehicle, without physically replacing a SIM card or module. It is even possible to mount SIMs on a sustained basis during the car production process and keep that SIM throughout the entire lifecycle of the car.

During the initial phase of the vehicle production, test, transport, and sales (Figure 11), default subscription identities relating to the vehicle manufacturer (OEM IDs, original equipment manufacturer) are used. Hence, at that moment, the OEM itself appears as owner and subscriber of all its newly produced cellular-connected cars to the network operator, to which the OEM has engaged in a contractual relation. No end-customer relation is in place at this phase of the production lifecycle. The OEM can now access the car systems remotely using cellular networks, for example, to check where and when a car has been shipped. After the car has reached the dealer and is being sold to a customer, a relation among the new car owner, its service packages, and the car itself gets established in the cellular network. Therewith, the car is represented by its vehicle OEM ID and the IMEI (Section 3.2.1) of the communication module of the car. The user data is known at the dealer shop. Car services or services packages, being sold to the user with the new car, get specific service package IDs (i.e., IMS IDs, Section 4.1.2). The first time the new car owner is starting his or her preconfigured car, a connection to the cellular network is established. The car owner's identity and service package IDs are provisioned to the vehicle OTA. From

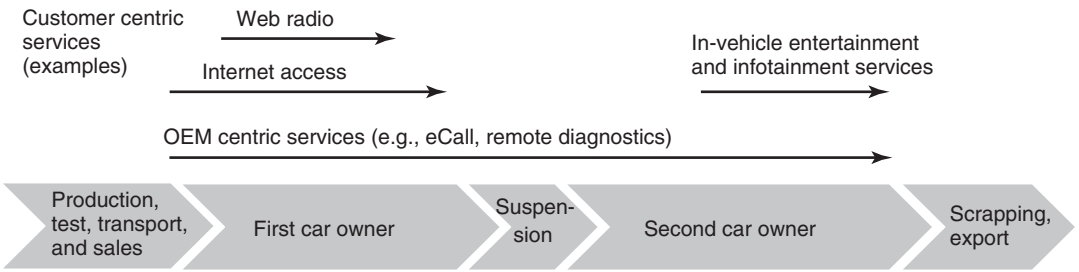


Figure 11. Sample lifecycle of a vehicle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

now onwards, service invocations and service-related data traffic get tagged with its corresponding service package IDs. Hence, cost can be clearly assigned, for example, to the car OEM for its "OEM-centric services" (Figure 11), or to the car owner for driver-centric services such as "Internet access," as shown in Figure 11. This flexible provisioning also allows reprovisioning of service IDs or cost items, without exchanging any hardware, when the car is sold to another owner.

Another approach is, utilizing the so-called "embedded SIM" concept that defines new methods for the remote (OTA) provisioning of access credentials of embedded SIMs and for managing the change of subscription credentials (Sections 3.2.1 and 3.3.1) from one cellular network operator to another. The GSMA, the association of 800 MNOs and 200 companies in the broader mobile ecosystem, defined in February 2011 requirements and use cases for the embedded SIM. New work items were established in several standards-developing organizations, such as ETSI and 3GPP (Section 5.1). The goal is to better support life-cycle management needs in long-lasting M2M businesses, likewise cellular-connected vehicles.

5 STANDARDS AND REGULATORY FRAMEWORKS

Much has been said on the previous sections on how the technical systems operate that constitutes a modern, cellular network. Mobile communication, however, has also an organizational and a regulatory dimension, which has a strong influence on technical solutions and its further development. In this section, we briefly outline three key areas, serving as examples for the complex environment that is influencing cellular network operation and its system evolution.

5.1 Standardization approach

Communication in cellular networks, as we know it today, would not be possible without standards. Nowadays, it is taken for granted that every UE (Section 3.2.1), purchased from any device vendor, can connect without problems to the cellular network of every operator, whereas the network itself is being delivered and installed by any network vendor. Harmonized numbering plans make it possible to have a truly global communication. Network operators rely on standardized interfaces, so that components purchased from different network equipment manufacturers can interact seamlessly. Finally, a major contributor to the decreasing price levels for mobile cellular communication is that research and development is more efficient if it can focus on widely accepted standards that enable reusing a product in many markets.

On the highest level, the standards enabling global communication are made by the ITU. Examples of ITU standards are global schemes for telephone numbers (International Telecommunication Union Recommendation E.164, 2010) and subscriber identities (IMSI), as well as the regulation of spectrum usage (Section 5.2). In addition, the requirement framework for the IMT-2000 that UMTS belongs to and the IMT-Advanced framework, which LTE falls into, has been set by the ITU. Only technology that fulfills these high-level requirements, defined by the ITU, can use the radio spectrum globally allocated for such technology.

ITU standards are, however, typically only the smallest common denominator among many basic principles. The detailed specifications are agreed by national and regional bodies. Over time, this work has become more and more global. The GSM standards initially developed in ETSI were the first cross-country standards that have also been adopted in many countries outside Europe. With the standardization for UMTS, the standards were already made on a global basis, but still split between the 3GPP and its counterpart 3GPP2. The next generation of systems has solely been developed in 3GPP, leading to the LTE standard, which can be seen as the first truly global cellular communication standard. A body such as 3GPP has participants from all relevant parts of the industry, most notably the cellular network operators and the equipment manufacturers. The decision process is typically consensus based, and decisions are documented in the so-called technical standard (TS) in the 3GPP case. A detailed description of the standards process is provided in Sesia *et al*. (2011).

5.2 Spectrum allocation

There is one critical natural resource that is required to operate a cellular network: the frequencies that the system uses to wirelessly transmit data. The license to use a certain frequency band in a certain country or region is one of the central assets of an MNO. Given the exponential growth in mobile data transmission, it can be expected that despite efforts to use spectrum more efficiently, additional frequency bands will be required in the foreseeable future to keep up with user demands (Ericsson mobility report—on the pulse of the networked society, 2012; More than 50 billion connected devices, 2011; City life, 2012). This section briefly describes the regulatory mechanisms that eventually lead to the allocation of frequency ranges, also referred to as *frequency spectrum* or just *spectrum*, to a mobile cellular network operator.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Cellular networks try to use low frequencies as much as possible, as radio waves travel longer distances at low frequencies and penetrate buildings more easily. Spectrum at these desirable low frequency ranges is, however, already fully allocated, so that only two alternatives remain: sharing a frequency band is a common practice, a popular example being the 802.11b/g/n Wireless LAN system operating at 2.4 GHz. For a commercial service such as a cellular network, it is, however, common to rely on spectrum with primary usage rights. This makes refarming, that is, the re-allocation of existing usage rights, the only viable option to obtain additional radio spectrum for future mobile networks. Spectrum comes in two variants: FDD systems use separate frequency bands for up- and downlinks; TDD systems share one frequency band by splitting it into time slots. Both options are actively used, depending on the region. In Europe, predominantly FDD systems are used.

On the highest level, the ITU is responsible for the allocation of spectrum. The ITU Radio Regulations govern the use of spectrum worldwide, separately for three major geographic regions (Europe, Middle-East, and Africa—America—Asia-Pacific). These regulations are revised by the World Radio Conference (WRC) that takes place approximately in 3-year intervals. Their agenda is agreed by the ITU Council 2 years ahead of the conference. The basic scope of an agenda is, however, often established even 4–6 years in advance. On the conference itself, proposals from countries (or groups of countries) are discussed and negotiated. The goal is a consensus between all participating parties. If the consensus cannot be reached, voting is possible, although in practice it is more likely that a decision is postponed to the next conference. The WRC held in February 2012 gathered participants from 153 countries.

Once a decision is taken at the WRC, it has to be translated to local regulation in the respective countries, for example, by the Bundesnetzagentur in Germany, OFCOM in the United Kingdom, or the Federal Communications Commission (FCC) in the USA. In countries such as the USA, the local regulator can directly adopt the decision and translate it into a local regulation. In other regions, such as the densely populated Europe, additional coordination is required. In the example of Europe, the European Conference of Postal and Telecommunications Administration (CEPT) coordinates the local policy-making across the members states and tries to harmonize the spectrum allocation policy among the members states. Once an agreement has been reached and national regulation is in place, a typical mode of distributing licenses for spectrum usage is an auctioning process, as recently practiced in many countries for the 3G and 4G spectra.

This short overview makes it apparent that spectrum allocation is a long-term process that is subject to discussions and consensus building on many levels. New frequency bands will not become available shortly. Corresponding processes have to be started well in advance. A recent example of successful spectrum refarming is the reallocation of digital dividend frequencies, that is, TV frequencies for analog TV broadcasting that were not needed anymore because of the higher spectral efficiency of digital transmission technology. These frequencies below 1 GHz have been allocated to cellular communication in many countries, and are, for example, used to provide broadband coverage in rural regions. In some regions of the world, mainly Asia, there are also cases where older cellular networks are switched off to use the frequencies for newer generations of cellular networks.

A second observation is that the situation around the world is becoming more and more fragmented. The usage of frequencies is different from country to country, and harmonization is not always possible. The 2G system was standardized for three main frequencies that are in use worldwide. For LTE, 25 FDD bands and 11 TDD bands had to be standardized in order to safeguard coverage for the most relevant regions around the world. In addition, the fragmented spectrum landscape makes it increasingly difficult to obtain larger contiguous blocks of radio spectrum. This has been considered for the development LTE-Advanced, which can support noncontiguous spectrum allocations.

5.3 Privacy aspects

Privacy means different things to different people, one definition can be found in Article 8 of the European Convention on Human Rights, namely the right to respect for private and family life: "*everyone has the right to respect for his private and family life, his home and his correspondence*." The issue of privacy is a good example of local regulation that also has to be fulfilled by every communication network. Privacy regulation is generally defined by the national legislation of each country. In Europe, it has been harmonized to some extent on EU level by European Directives. In this context, data security or information security plays an important rule. It is possible to provide security without privacy, but not to provide privacy without security. Therefore, protecting and controlling personal information is what the European legal concept of "data protection" is concerned with. The European Directives have in common that they represent only the minimum standard of data protection in the respective field that shall be transposed into national law. The fundamental basis is defined in the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

European Data Protection Directive. This Directive regulates the processing of personal data within the European Union and is an important provision in terms of privacy and human rights law. All members states have enacted their own data protection legislation. In the field of telecommunications, the European Data Protection Directive is accompanied by the more specific directive on privacy and electronic communications.

For determining legal effects of the processing of personal data, it is important to also consider the combined benefits of data processing for the end-user and for the bigger society. It is, for example, possible to process location data of vehicles in order to provide drivers with local relevant near real-time traffic information (including information to enhance safety) that can have effect for navigation and vehicle guidance. This purpose of data processing can be seen as a "value-added service" for the end-user. For such value-added services, the processing of location data is admitted but regulated: "*. . . such data may only be processed when they are made anonymous, or with the consent of the users or subscribers to the extent and for the duration necessary for the provision of a value added service. The service provider must inform the users or subscribers, prior to obtaining their consent*, . . . " (Article 9 of Directive on Privacy and Electronic Communications); see also (European Commission).

6 OUTLOOK

The worldwide geographical availability of cellular mobile networks as well as their technical performance has increased continuously over the last few decades. Bandwidth and latency improvements have jumped orders of magnitude in just a few years. Especially, a growing amount of spectrum below 1 GHz allocated for IMT technologies will lead to further geographic coverage and capacity improvements of high speed cellular networks. The ongoing market trend toward technology convergence, combined with the spectral efficiency gains of LTE and LTE-Advanced, is paving the way toward the first truly global communication standard for cellular mobile networks.

The global uptake is also clearly visible in numbers: as of September 2013, more than 213 cellular network operators have launched commercial LTE services in 81 countries, providing high speed mobile broadband access to 10% of the world's population by the end of 2012. More than 440 operators have publicly committed to LTE technology across 130 different countries with a large number of LTE trials currently in operation.[3] GSA confirmed September 2013 that 1064 LTE user devices have been announced by 111 manufacturers, representing around 150% annual growth.

This rapid worldwide deployment, the technological advances, and the tight integration with business processes (as briefly described in this chapter) will pave the way toward a networked society (City life, 2012) with an anticipated 50 billion connected devices by 2020 (More than 50 billion connected devices, 2011). How many of those will be cellular connected vehicles? And how many cooperative vehicle services will benefit for cellular network capabilities?

Like in the decades before, technological enhancements and innovations in key areas such as energy efficiency and processing power of the UEs will keep up the pace of development also in the future. For example, maximum data rates enhanced from 384 kbps of first UMTS networks, introduced in the 1990s, to 42 Mbps of HSPA networks, deployed just one decade later.

First LTE networks based on Release 8 of the 3GPP standard were commercially launched in 2011. While the availability of LTE-capable UEs and networks have steadily increased, the standard for the next evolutionary step, LTE-Advanced, was already finalized (Third Generation Partnership Project (3GPP)). With the defined technology improvements, LTE-Advanced targets maximum data rates of 1 Gbps and improved network performance at the cell edges.

On the other hand, enhanced connected car services such as cooperative traffic management, driver information and assistance applications, or infotainment services are gaining importance (see GSMA mAutomotive, 2012, 2025: every car connected: forecasting the growth and opportunity; GSMA mAutomotive, 2013). The authors are convinced that future car drivers and passengers expect connected car services, resulting in an increasing market demand. New trends and innovations such as in the area of electric vehicles will lead to an even faster growing need for connectivity in the car. Efficient energy management, vehicle monitoring, and services such as automatic billing at local charging stations will require communication among driver, charging stations, and utility companies.

With the geographical coverage and performance of cellular networks constantly increasing, more and more advanced connected car services will become available. A fruitful interaction between the telecommunication and the automotive industry will be key for turning the connected vehicle vision into reality; fast and in an economically viable manner. Relying on widely accepted, deployed, and open standards has been a cornerstone of the success of the global telecommunication industry. A similar approach may lead the way forward for the connected vehicle vision, with all its definite advantages to the broader society.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

GLOSSARY

Term	Name	Description
2G	Second generation	Synonym for the second-generation cellular telecommunications systems.
3G	Third generation	Synonym for the third-generation cellular telecommunications systems.
3GPP	Third-generation partnership project	Standards body developing responsible for standardization of GSM, UMTS, LTE, IMS, and other systems.
3GPP2	Third-generation partnership project 2	3GPP2 is the standardization group for CDMA2000, the set of 3G standards based on earlier 2G CDMA technology. 3GPP2 is not a part of 3GPP, but rather a separate organization.
4G	Fourth generation	Synonym for the fourth-generation cellular telecommunications systems.
AAA	Authentication, authorization and accounting	A set of core network functions and system nodes dealing with subscriber authentication, service and access permissions, and charging data processing.
APN	Access point name	A configurable identifier used by a mobile device when connecting to a cellular network to indicate which external network or service the user wishes to connect to via the GGSN/PDN Gateway.
AuC	Authentication center	Used in GSM networks to authenticate SIMs that attempt to connect to the network.
BM-SC	Broadcast/multicast service center	A system control node, serving as entry point for content delivery services utilizing MBMS capabilities.
BSC	Base station controller	A 2G system node, responsible for controlling the base transceiver stations (or base stations for short) that are connected to it.
BSS	Business support systems	Supports the business operation of network operators toward the customer with functions such as billing and customer and ordering management.
BTS	Base transceiver station	The radio base station of a 2G system, responsible for termination of the wireless radio link to the UEs and manages the radio resources in its coverage area.
CDMA	Code division multiple access	A multiplexing scheme based on a coding method that assigns a specific code to each transmitter in order to allow multiple users to access the same physical channel.
CN	Core network	A set of access-technology independent nodes and functions, comprising, among others, user databases and interfaces to other networks.
CS	Circuit switched	The traditional implementation of a telecommunications network, in which two network nodes establish a dedicated end-to-end communications channel (circuit) through the network.
CSCF	Call session control function	A system node family within the IMS subsystem, used to process SIP signaling.
EDGE	Enhanced data rates for GSM evolution	Evolution of GPRS, uses advanced coding and modulation to achieve higher data rates.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

eNodeB	evolved node B	The radio base station in LTE networks, successor to the NodeB of a 3G UMTS system.
EPC	Evolved packet core	Core network architecture introduced with LTE featuring a simplified all-IP network design.
ETSI	European Telecommunications Standards Institute	Standardization organization for telecommunication technologies in Europe. ETSI is a partner of the 3GPP.
eUICC	Embedded universal integrated circuit card	Embedded version of the UICC, e.g., to be soldered on a printed circuit board.
FDD	Frequency division duplexing	Frequency division duplex systems use separate frequency bands for up- and downlinks (the so-called paired spectrum).
GGSN	Gateway GPRS support node	A core network node, responsible for the interworking between the GPRS network and external packet-switched networks such as the Internet.
GPRS	General packet radio service	A cellular communication technology that introduced IP packet transport to the 2G core network, sometimes also called *2.5G*.
GRX	GPRS roaming exchange	A dedicated type of interconnect network between mobile network operators, e.g., used for connecting home network and visited network in a roaming case.
GSA	Global supplier association	Representing the leading GSM/EDGE, WCDMA-HSPA, 4G/LTE suppliers worldwide.
GSM	Global system for mobile communications	2G cellular communication system. Primarily used for circuit-switched voice communication and SMS.
GSMA	GSM association	Official industry trade group representing GSM network operators worldwide.
HLR	Home location register	A central database that contains a record of each mobile phone subscriber that is authorized to use the GSM core network.
HSPA	High speed (downlink/uplink) packet access	Evolution of UMTS with many enhancements, among them advanced coding and modulation to achieve higher data rates in downlink (HSDPA) and uplink (HSUPA).
HSS	Home subscriber server	A core network node manages subscription-related information, includes HLR, AuC, and some further functions of the 2G system architecture.
IMEI	International mobile equipment identity	Globally unique identity of a cellular communication device. The IMEI identifies all physical devices of GSM, UMTS, and LTE networks using SIM cards.
IMS	IP multimedia subsystem	The IP multimedia subsystem is an architectural framework to deliver multimedia services such as voice-over IP services.
IMSI	International mobile subscriber identity	Globally unique user identification in mobile networks, associated with all GSM-, UMTS-, and LTE-capable mobile phones or modules. It is stored as a 64-bit field in the SIM inside the phone or module and identifies the user to the network (as opposed to the IMEI that identifies the device). The IMSI is used in internal network authentication and authorization procedures, for example, in the HLR or HSS.
IP	Internet protocol	A networking protocol that is the basis for all packet-switched communication, originally developed for Internet communication.
IPX	IP Packet Exchange	An IP-based evolution of the earlier GPRS roaming exchange (GRX) network.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ITU	International Telecommunication Union	The International Telecommunication Union is an organization of the United Nations and is responsible for basic standards, coordination, and regulation for information and communication technologies.
KPI	Key performance indicator	Typically a numeric value associated with a performance or progress related monitoring process.
Leased Line	Leased communication line	A dedicated network connectivity link, connecting two end points in a secure and QoS-ensured manner. Although in the past leased lines had been implemented as dedicated physical connections, nowadays more and more of them get implemented as QoS-managed VPN connections over high capacity telecommunication networks.
LTE	Long-term evolution	4G cellular communication system standardized by 3GPP.
M2M	Machine-to-machine	Umbrella term for communication that is not person-to-person (e.g., phone call) or person-to-machine (e.g., Web surfing), but connecting a machine of any kind to another machine.
MB	Megabyte	1024×1024 bytes (=2^20 bytes, ≈10^6 bytes). One byte consists of 8 bits, so that there is roughly a factor of 8 between Mb and MB. MB is typically used for data volumes and file sizes.
MBMS	Multimedia broadcast multicast service	MBMS systems add broadcast capabilities to mobile networks, i.e., the capability to save transmission capacity when transmitting the same data from a single source to multiple destinations.
NodeB	Node B	Radio base station in UMTS, successor to the base station of a 2G GSM system (see BTS).
OFDM	Orthogonal frequency division multiplexing	OFDM is a multiplexing scheme where a large number of closely spaced but uncorrelated (orthogonal) subcarriers are used to carry data of multiple users over the same physical channel.
OSS	Operations support system	Supports the operations of a telecommunication network, provisioning of network or telecommunication services, configuring network components, and managing faults.
OTA	Over the air	Umbrella term for end-to-end procedures carried out over a wireless connection without considering specific network segment characteristics (e.g., network load and radio resources) and not using network-assisted functionalities.
PDN-GW	PDN gateway	A system node of cellular networks, serving as entry and exit points for IP data traffic and as termination point for tunneled-IP data traffic in roaming cases. The PDN GW is the successor to the GGSN in SAE-based networks.
PS	Packet-switched	A digital networking communications paradigm that transmits any kind of data in the form of suitably sized blocks, the so-called packets.
PSTN	Public-switched telephony network	The global communication network for circuit-switched wireline communication.
QoS	Quality of service	Umbrella term for different technical measures to prioritize and differentiate traffic in a communication network in order to ensure a given network quality for a certain service.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

RAN	Radio access network	The part of a cellular communication network that implements the radio communication technology in order to connect a mobile terminal to the core network.
RNC	Radio network controller	A 3G system node, responsible for controlling the NodeBs that are connected to it.
SDP	Service delivery platform	A service enablement extension, complementing network communication and control functions.
SGSN	Serving GPRS support node	A core network node responsible for the delivery of data packets from and to the mobile stations within its geographical service area. Among other tasks, it is responsible for the mobility management (attach/detach and location management) and logical link management to a mobile station (or terminal) when moving from one serving base station to another.
SIM	Subscriber identity module	A SIM is an integrated circuit embedded into a removable SIM card that securely stores network-specific information used to authenticate and identify subscribers such as the IMSI, personal identification numbers (PIN), and authentication keys.
SLA	Service level agreement	A binding agreement between parties to safeguard the delivery of a certain service. Often expressed through a number of KPIs together with monitoring principles and information handling duties.
SMS	Short messaging service	A short text message which in size is restricted to 140 bytes or 160 characters.
TDD	Time division duplexing	Time division duplex systems share one frequency band (the so-called unpaired spectrum) by splitting it into time slots for up- and downlinks.
TDMA	Time division multiplexing access	A multiplexing scheme; a physical channel by subdividing it in the time domain.
UE	User equipment	A term originally used in 3G standardization, referring to the device that terminates the cellular connection on the user side. These devices come as smartphones, USB sticks, connected navigation devices, embedded vehicle modules, or in many other forms.
UICC	Universal integrated chip card	Multifunctional chip card for mobile devices, e.g., used as SIM card.
UMTS	Universal mobile telecommunications system	UMTS is the 3G mobile cellular system in the standards family developed and maintained by 3GPP.
VLR	Visitor location registers	A set of connected databases used for storing temporary subscriber data to support mobility management and handover procedures.
VPN	Virtual private networks	VPNs establish isolated, secure connections between individual users and a remote network or between multiple networks.
WRC	World Radiocommunication Conference	The WRC is organized by the ITU in basically 3-year intervals in order to revise the regulation for spectrum usage worldwide, separately for three major geographic regions (EMEA, Americas, and Asia-Pacific)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ENDNOTES

1. http://www.ietf.org/
2. European Commission Decision of 16 July 2003 relating to a proceeding under Article 81 of the EC Treaty and Article 53 of the EEA Agreement (Case COMP/38.369: T-Mobile Deutschland/O2 Germany: Network Sharing Rahmenvertrag), OJ 2004, L 75/32.
3. Data from GSA September 2013 (http://www.gsacom.com/), and from Ericsson statistics.

REFERENCES

GSMA mAutomotive 2025 (2012) Every car connected: forecasting the growth and opportunity. GSMA White Paper, April 2012, http://www.gsma.com/connectedliving/gsma-2025-every-car-connected-forecasting-the-growth-and-opportunity (accessed 13 August 2013).

GSMA mAutomotive (2013) Connected car forecast next five Years, June 2013, http://www.gsma.com/connectedliving/connected-car-forecast-next-five-years.

3GPP TS 23.203 (2012) Policy and charging control architecture, V11.7.0, September 2012.

3GPP TS 33.220 (2012) Generic authentication architecture (GAA); generic bootstrapping architecture, v11.4.0, September 2012, http://www.3gpp.org/ftp/Specs/html-info/33-series.htm (accessed 13 August 2013).

Ala-Luuko, S. (2006) IPX—a key enabler for IP communication ecosystem, September 2006, http://www.teliasoneraic.com/iccons/groups/public/documents/webdocument/ts020933.pdf (accessed 13 August 2013).

Astély, D., Dahlman, E., Furuskär, A., *et al.* (2009) LTE: the evolution of mobile broadband. *IEEE Communications Magazine*, **47** (4), 44–51.

Bakhuizen, M. and Horn, U. (2005) Mobile broadcast/multicast in mobile networks. Ericsson Review 01/2005, http://www.ericsson.com/ericsson/corpinfo/publications/review/2005_01/files/2005015.pdf (accessed 13 August 2013).

Camarillo, G. and García-Martín, M. (2011) *The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds*, 3rd edn, John Wiley & Sons, Ltd, Chichester.

Car Connectivity Consortium homepage, http://www.mirrorlink.com/ (accessed 13 August 2013).

City life. Ericsson Report, May 2012, http://www.ericsson.com/res/docs/2012/city_life.pdf (accessed 13 August 2013).

Dahlman, E., Parkvall, S. and Sköld, J. (2011) *4G—LTE/LTE-Advanced for Mobile Broadband*, Academic Press, New York.

Dietz, U. (ed.), *Adaptive and Cooperative Technologies for Intelligent Traffic—Cooperative Cars*, 2009, http://www.aktiv-online.org/english/downloads/CoCar_D04_%20public.pdf (accessed 13 August 2013).

Ekström, H. (2009) QoS control in the 3GPP evolved packet system. *IEEE Communications Magazine*, **47** (2), 76–83.

Ericsson (2012) Ericsson mobility report—on the pulse of the networked society. November 2012, http://www.ericsson.com/ericsson-mobility-report, http://www.ericsson.com/res/docs/2012/ericsson-mobility-report-november-2012.pdf.

ETSI ITS TR 102 962 (2012) Intelligent transport systems (ITS); framework for public mobile networks in cooperative ITS (C-ITS), V1.1.1, February 2012.

ETSI ITS TS 103 084 (2012) Geomessaging enabler, current draft version, June 2012. http://www.etsi.org/deliver/etsi_tr%5C102900_102999%5C102962%5C01.01.01_60%5Ctr_102962v010101p.pdf.

European Commission "Regulatory framework for electronic communications": http://europa.eu/legislation_summaries/information_society/legislative_framework/l24216a_en.htm and "Current general legal framework": http://europa.eu/legislation_summaries/information_society/legislative_framework/ (accessed 13 August 2013).

Gehlen, G., Ramme, F., Sories, S. and Jodlauk, G. (2007) *Cooperative cars—using cellular communications for co-operative automotive applications*. ITS World Congress 2007, Beijing, China, October 2007.

GSMA IR.34 (2013) Inter-service provider IP backbone guidelines, version 9.1, May 2013, http://www.gsma.com/newsroom/wp-content/uploads/2013/05/IR.34-v9.1.pdf.

GSMA IR.67 (2012) DNS/ENUM guidelines for service providers & GRX/IPX providers, version 7, May 2012, http://www.gsma.com/newsroom/wp-content/uploads/2012/06/IR6770.pdf.

GSMA IR.92 (2013) IMS profile for voice and SMS, version 7.0, March 2013, http://www.gsma.com/newsroom/wp-content/uploads/2013/04/IR.92-v7.0.pdf.

International Telecommunication Union Recommendation E.164 (2010) The international public telecommunication numbering plan, November 2010.

International Telecommunication Union Recommendation E.212 (2011) International operation—maritime mobile service and public land mobile service, version 5.3, June 2011.

Jodlauk, G., Rembarz, R. and Xu, Z. (2011) *An optimized grid-based geocasting method for cellular mobile networks*. ITS World Congress 2011, Orlando, USA, October 2011.

Larmo, A., Lindström, M., Meyer, M., *et al*. (2009) The LTE link-layer design. *IEEE Communications Magazine*, **47** (4), 52–59.

More than 50 billion connected devices. Ericsson White Paper, February 2011, http://www.ericsson.com/res/docs/whitepapers/wp-50-billions.pdf (accessed 13 August 2013).

Next generation LTE, LTE-Advanced. Ericsson Review 2/2010, December 2010, http://www.ericsson.com/res/thecompany/docs/publications/ericsson_review/2010/next-generation-lte.pdf (accessed 13 August 2013).

Olsson, M., Sultana, S., Rommer, S., *et al*. (2009) *System Architecture Evolution (SAE): Evolved Packet Core for LTE, Fixed and Other Wireless Accesses*, Academic Press, New York. http://www.amazon.de/System-Architecture-Evolution-SAE-Wireless/dp/0123748267#reader_0123748267.

Phan, M., Rembarz, R. and Sories, S. (2011) A capacity analysis for the transmission of event and cooperative awareness messages in LTE networks. ITS World Congress 2011, Orlando, USA, October 2011.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Poikselka, M. and Mayer, G. (2006) *The IMS: IP Multimedia Concepts and Services in the Mobile Domain*, vol. 2nd, John Wiley & Sons, Ltd, Chichester.

Rappaport, T.S. (2001) *Wireless Communications—Principles and Practice*, 2nd edn, Prentice Hall, Upper Saddle River.

Rosenberg, J., Schulzrinne, H., Camarillo, U.G., *et al.* (2002) SIP: session initiation protocol, RFC 3261, June 2002.

Schulzrinne H., Rao A., and Lanphier R. (1998) Real time streaming protocol (RTSP), RFC 2326, April 1998.

Sesia, S., Baker, M., and Issam, T. (2011) *LTE—The UMTS Long Term Evolution From Theory to Practice*, John Wiley & Sons, Ltd, Chichester.

Sories, S., Huschke, J. and Phan, M. (2008) Delay performance of vehicle safety applications in UMTS, 15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting, November 2008.

Third Generation Partnership Project (3GPP) Standard Release 10, http://www.3gpp.org/Releases (accessed 13 August 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto169
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Technologies—Communication: Wireless LAN-Based Vehicular Communication

Andreas Festag[1], Hannes Hartenstein[2], and Jens Mittag[2]

[1]*NEC Europe Ltd., Heidelberg, Germany*
[2]*Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany*

1 Introduction 1
2 System Architecture 3
3 Physical Layer and Medium Access 3
4 Communication Protocols 6
5 Security and Privacy 8
6 Radio Resource Management and Decentralized Congestion Control 9
7 Standardization 10
8 Outlook 11
Glossary 12
References 13
Further Reading 13

1 INTRODUCTION

Wireless communication to, from, and between vehicles allows vehicles to announce their current status to other traffic participants as well as to receive information that is beyond the vehicle's sensing ability and the driver's direct perception. This article discusses the use of WLAN techniques for the purpose of exchanging information among vehicles and between vehicles and the roadside infrastructure as a basis for active safety, green mobility, and traffic-efficiency applications.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Various notions have been used to address WLAN-based vehicular communication: car-to-X communication (C2X), vehicular ad hoc networks (VANET), dedicated short-range communication (DSRC), cooperative systems, intelligent transportation systems (ITS), vehicular area network, or inter-vehicular communications. While these notions themselves represent "technology-neutral" terms, their use typically refers to WLAN-based techniques, in particular to the IEEE 802.11p amendment to the IEEE 802.11 standard, in a vehicular environment. WLAN-based vehicular communication is often seen as an alternative or a complement to mobile communication based on cellular networks (see Technologies—Communication: Mobile).

Although the envisioned scenario of vehicles that exchange information using WLAN-based techniques appears intuitively convincing, the technical implementation is not straightforward. Indeed, inter-vehicle communication networks are challenged by requirements as well as by constraints and limitations that exist either because of the inherent characteristics of the considered scenario or because of the selected communication technology has not been designed for usage in such an environment in the first place.

Independent of the selected communication technology, inter-vehicle communication networks, or systems should fulfill the following key requirements.

- The communication system should allow for a transmission range of at least a few hundred meters when sending with maximum power. The received signal strength typically weakens by at least $1/d^2$ where d denotes the distance to the sender. Hence, the coverage range of a sender is limited.

- The communication system needs to be standardized to allow communication between vehicles of all makes and brands.
- The communication system needs to fulfill application requirements, for example, in the case of safety applications, it needs to guarantee low data dissemination delays and a high reliability.
- The communication system needs to be adaptive and robust in order to deal with all relevant system dynamics: the mobility of vehicles and corresponding changes in network topology, as well as the wide range of scenarios in which inter-vehicle communication networks will be deployed.
- The communication should be secure and should preserve privacy, thus, two sometimes contradictory or opposing goals need to be balanced.

Additional technology- and standardization-specific challenges are introduced with the characteristics of IEEE 802.11p standard specification and the system architecture that is built upon IEEE 802.11p:

- There exists one common control channel for safety- and signaling-related information. WLAN-based vehicular communication is intrinsically of broadcast nature, as a warning or status message sent out by a vehicle should be received by all vehicles within the transmission range.
- No central control or operator schedules and organizes the various transmissions of all vehicles. Instead, access to the wireless channel is coordinated in a distributed and cooperative fashion, that is, fundamental issues of self-organizing networks are adopted.
- The system has to support at least two types of safety messages: periodic awareness messages, which are broadcasted by any vehicle to inform neighboring vehicles about the own presence and status, as well as event-driven alert messages, which are sent out in case of an emergency situation that requires immediate attention. Periodic messages are envisioned to be only broadcasted for one hop (i.e., there is no retransmission of the identical message by another vehicle) and termed *cooperative awareness message* (CAM), *basic safety message* (BSM), or simply *beacon*. Event-driven messages (also called *decentralized environmental notification message*, DENM) may be disseminated over more than one hop. In general, the communication system should allow for prioritizing and differentiating types of communication.

General issues and limitations of wireless communications challenge WLAN-based vehicular communications in the 5.9 GHz frequency band even further (see, e.g., Molisch, 2010 for fundamentals on wireless channels and wireless local area networks):

- Owing to the 5.9 GHz carrier frequency, shadowing effects and reflections will be very prominent in inter-vehicle communication networks. As a result, severe multi-path propagation effects have to be expected, which, together with the high mobility of the vehicles, lead to a time- and frequency-selective fading of the channel impulse response. At the same time, communication "around the corner" or "through buildings" will be limited.
- As the coverage range is limited, every vehicle has a different perspective on who is currently transmitting and who is not. A distributed coordination mechanism is therefore challenged by the hidden terminal problem. As a result, simultaneous transmissions that lead to packet collisions at receiving nodes will not be an exception. Channel fading increases the severity of the hidden terminal problem.
- The capacity of the wireless channel is limited. While this is not an issue if only a few vehicles are participating in the network, it will become an issue if every vehicle is equipped. The communication system should therefore be able to support both extremes and should scale with the number of vehicles in the network, in particular with respect to resource and bandwidth allocations.

While each individual aspect alone can easily be solved separately, their combination poses a challenge for the design of an optimal or even acceptable solution. In short: the overall challenge lies in the support for applications with strong reliability requirements over a highly unreliable communication medium. For the application engineer, it is important to keep in mind that even while the field of vehicular communication advances to provide an increased level of reliability of the communication, there is always the chance that some information is not sent or received in a certain time frame.

In the following, we first present the general system architectures for WLAN-based vehicular communication systems that emerged in standardization activities in Europe and the US (Section 2). We present the basics of physical layer and medium access issues (Section 3) as well as communication protocols for WLAN-based vehicular communication, including the use of single-hop protocols and multi-hop protocols utilizing geographical positions (Section 4). Aspects of security and privacy are discussed (Section 5) and the role of radio resource management,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

particularly congestion control, as a means to deal with scalability and reliability is outlined (Section 6). We conclude this article with a view toward standardization (Section 7) and an outlook (Section 8) as well as offer references and hints for further reading.

2 SYSTEM ARCHITECTURE

The system architecture for WLAN-based vehicular communication is primarily seen as communication architecture and includes not only vehicles, but also roadside and central infrastructure that can be connected to legacy equipment, such as road sensors and traffic lights. In the context of ITS, the system can be divided into vehicle, personal (representing mobile personal devices), roadside (infrastructure), and central (back-end systems and traffic management centers) subsystems. In this architecture, WLAN-based vehicular communication is mainly in the vehicle and roadside subsystem. A central element of the system architecture is the ITS station, which is common to all subsystems. A generic protocol stack of an ITS station is shown in Figure 1 and put into relation with the ISO OSI protocol stack.

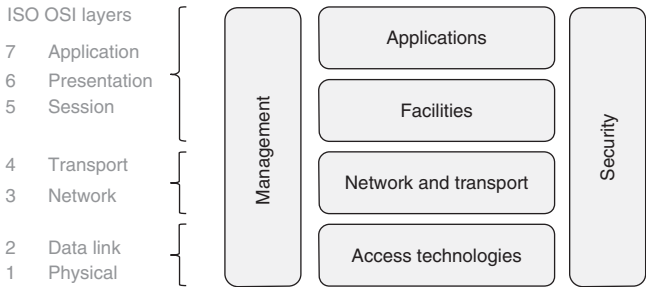


Figure 1. Generic protocol stack.

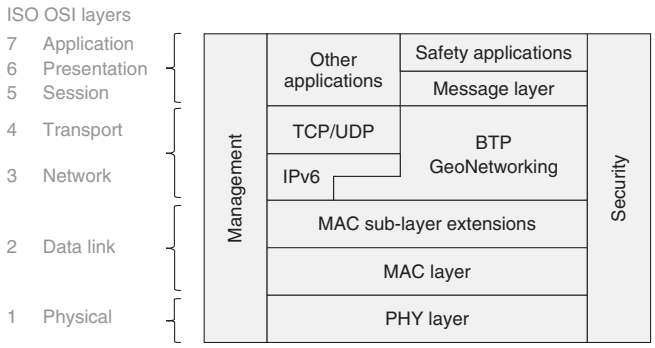


Figure 3. Protocol stack in Europe.

From this generic protocol stack, two stacks can be derived that are developed in the US (Figure 2) and Europe (Figure 3). The stacks are detailed in Sections 4 and 7.

3 PHYSICAL LAYER AND MEDIUM ACCESS

In 1999, the Federal Communications Commission in the US assigned 75 MHz of spectrum for DSRC services (Figure 4). The frequency band ranges from 5.850 to 5.925 GHz and is commonly referred to as the *5.9-GHz band*. The 5.9-GHz band is divided into one 5-MHz guard band and seven 10-MHz channels. One of the channels (channel 178) is the control channel, to which all vehicles must listen to. To make use of the other channels, either a channel-switching scheme or a multi-channel operation based on several radios have been proposed and discussed.

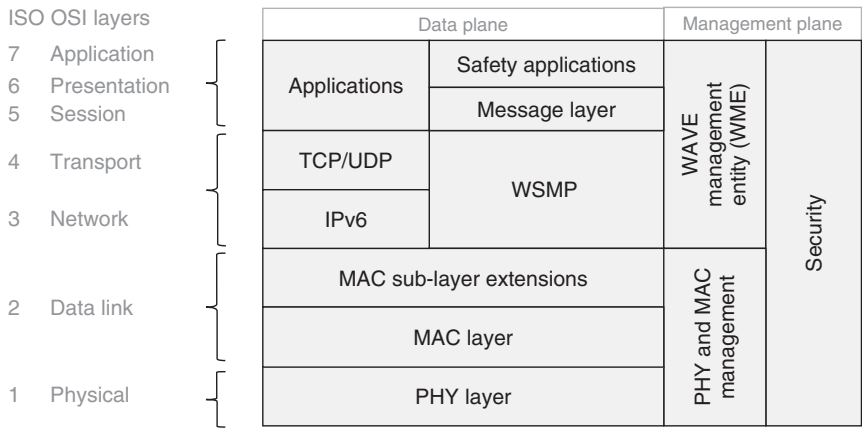


Figure 2. Protocol stack in the US.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In Europe, in August 2008 the EU allocated 30 MHz of spectrum for safety-related ITS communication. The frequency band ranges from 5.875 to 5.905 GHz and is referred to as the *ITS-G5A band*. In addition, a frequency band ranging from 5.855 to 5.875 GHz is foreseen for non-safety applications. This band is called *ITS-G5B*. ITS-G5A and ITS-G5B are divided into channels of 10 MHz. Furthermore, a frequency band called *ITS-G5C* in the range 5.470–5.725 GHz can be used for communication between fixed stations and vehicles, but not for vehicle-to-vehicle communication. Approaches for multi-channel operations, particularly dual-receiver concepts, were being discussed at the time of writing.

In Japan, a frequency band for ITS applications ranges from 5.770 to 5.850 GHz; another one, which ranges from 715 to 725 MHz, was assigned in 2007 and is dedicated to road safety applications.

Many other countries have spectrum allocated to WLAN-based vehicular communications, typically in the 5.8 and 5.9 GHz ranges. The maximum power allowed by a sender is typically restricted to 33 dBm equivalent isotropically radiated power (EIRP). To reduce inter-channel interference, regulations define transmit-spectral masks. An overview on 5.9 GHz field trials in a variety of physical environments is presented in Alexander, Haley and Grant (2011).

When modeling the corresponding radio channels, one has to consider the relatively high carrier frequency as compared to current cellular networks. The small signal wave-length of approximately 4 cm implies that any object larger than 4 cm acts as a reflector (or scatterer) and generates one (or multiple) copies of the transmitted signal that arrives at a receiver as a time-delayed echo. This phenomenon is typically called *multi-path propagation*, and, in the case of inter-vehicle communications, leads to a narrowband fading of the channel. As transmissions "through buildings" are not possible and antennas mounted only 1.5 to 2.5 m above ground, communication "around corners" (non-line-of-sight communication) will be limited. As small changes in the geometry of the vehicle also affect the radiation patterns of antennas significantly, their proper placement is a challenge as well. Likewise, signal propagation is influenced by surrounding vehicles and their mobility, as well as by the type of environment, that is, the number of buildings or trees next to the street, or the current weather conditions. Developed channel models are therefore typically classified into highway, urban, and suburban scenarios (e.g., Acosta-Marum and Ingram, 2007 and Mecklenbräuker *et al.*, 2011).

The physical layer of IEEE 802.11p and, accordingly, of the European ITS-G5 interface is based on orthogonal frequency division multiplexing (OFDM) with 52 subcarriers in order to deal with the narrowband fading characteristics of the channel. Four of these 52 subcarriers are used as pilots to track the variation of the channel over time and with respect to the frequency domain; the remaining 48 subcarriers are used for data transmission. In order to support different data rates from 3 to 27 Mbps, the data bits to be transmitted are encoded using a binary phase shift keying (BPSK), a quadrature phase shift keying (QPSK), or a quadrature amplitude modulation (QAM) with 16 or 64 constellation points. A convolutional encoder with a rate of 1/2, 2/3, or 3/4; further enables error correction capabilities at a receiver. While lower data rates are to be preferred because of their robustness against channel fading, they reduce the capacity of the network. Hence,

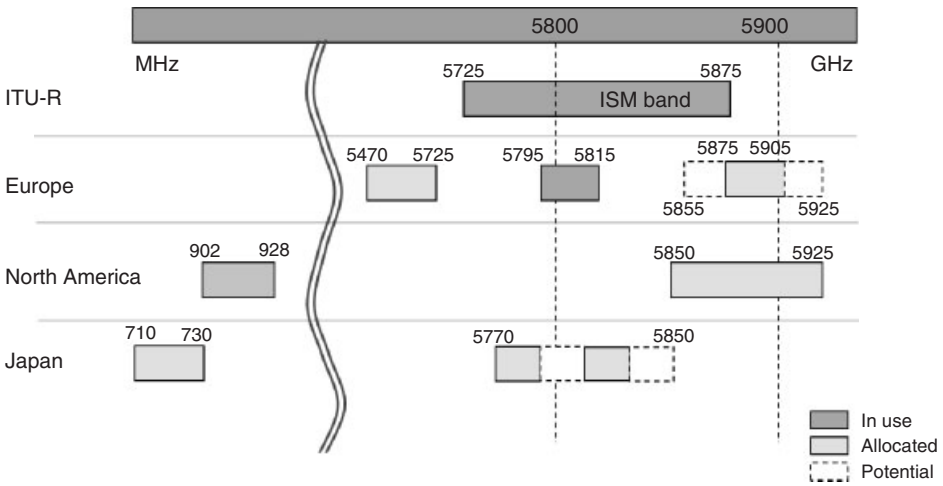


Figure 4. Frequency usage for vehicular communication worldwide.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

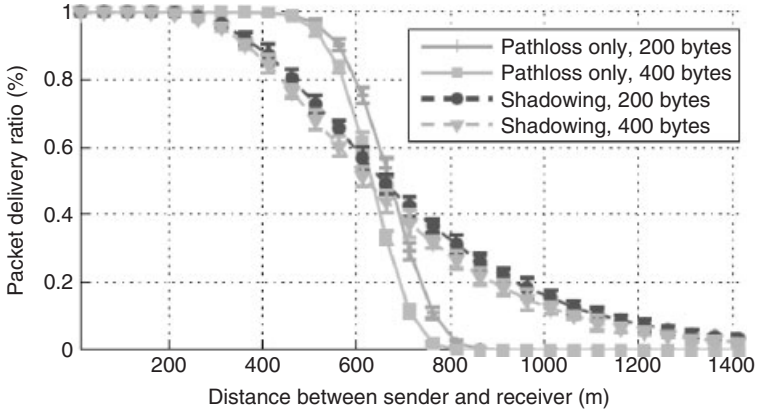


Figure 5. Packet delivery ratio (i.e., proportion of successfully received beacon messages) with respect to the distance between a single sender and a single receiver. The radio channel is assumed to follow a deterministic power-law, either including a large-scale log-normal shadowing effect or not.

a trade-off between robustness and scalability needs to be made. According to Jiang, Chen and Delgrossi (2008), a bit rate of 6 Mbps has shown to provide an optimal balance of these two aspects in typical vehicular environments. Figure 5 depicts typical packet delivery ratios with respect to the distance between a single sender and a single receiver for such a 6 Mbps configuration. The plotted ratios are based on the results of detailed physical layer simulations.

Access to the medium (medium access control, MAC) is governed by a distributed version of the carrier sense multiple access with collision avoidance (CSMA/CA) mechanism, also known as the *distributed coordination function* (DCF) in the context of IEEE 802.11 networks. With CSMA, a station listens to the medium before transmitting data. If the channel appears to be idle, the station transmits. If the channel appears to be busy, that is, someone else is currently transmitting and the corresponding signal is stronger than the clear channel assessment (CCA) threshold, the station will defer its access and waits until the channel becomes idle again. Then, the station will use a countdown-based random access to select a slot for transmission. If the channel becomes busy again before the countdown process is finished, the station interrupts its count-down for as long as the channel is busy. Packet collisions can occur on a receiving side when two stations select the same time slot accidentally or because they are not able to detect their transmitted signals (the latter case is commonly referred to as *hidden terminal problem*). While a method based on ready-to-send (RTS) and clear-to-send (CTS) signals was established for unicast communication in order to counter the hidden terminal problem, communication within vehicular networks is primarily of a broadcast nature. Hence, the benefit of the RTS/CTS approach cannot be exploited. A

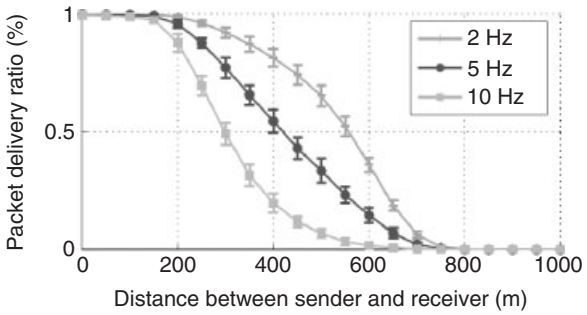


Figure 6. Expected packet delivery ratios (i.e., proportion of successfully received beacon messages) with respect to the distance between sender and receiver in a highway scenario of about 80 vehicles/km. The radio channel is assumed to follow a deterministic power-law model.

diagram showing typical packet delivery ratios with respect to the distance between sender and receiver (in this case in a highway scenario with 80 vehicles/km) is given in Figure 6. In this set-up, all vehicles transmit several beacon messages of 400 bytes/s using a 20 dBm transmission power. The radio channel is assumed to follow a deterministic power-law model. As can be seen, the packet delivery ratios decrease significantly if the beacon transmission rate increases from 2 to 10 Hz because of the increasing load on the channel. An in depth analysis on the effectiveness of CSMA in vehicular scenarios can be found in Mittag, 2012.

To allow for prioritization of certain frames over others, the enhanced distributed channel access (EDCA) can be used as part of the IEEE 802.11-2012 standard. With EDCA, the waiting times of a station between assessing whether a channel is idle or selecting a slot to start

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

transmissions vary with respect to the importance of the packet being transmitted.

While IEEE 802.11p is very close to other wireless local area networks in the IEEE 802.11 standard with respect to physical and MAC layers, it deviates quite a bit from office environments, where mobile stations and access points form so-called basic service sets (BSS) through association procedures. All these association procedures are skipped in IEEE 802.11p, and 802.11p stations are said to operate "outside the context of a basic service sets" (OCB). Therefore, other mechanisms are required for authentication and security; see Section 5.

When the number of vehicles per kilometer or square kilometer is very high, MAC based on CSMA/CA might become unstable and transmission efficiency might decrease because of packet collisions. Thus, it is important to keep the channel load below a critical value by means of congestion control approaches (Section 6). Alternatives to a CSMA-based MAC were discussed and analyzed, for example, self-organizing time division multiple access (STDMA), which is already used in the automatic identification system (AIS) for communication between ships. While reserving slots to access the channel can reduce the probability of collisions, the mobility of the vehicles and the fading characteristics challenge consistent slot reservations.

4 COMMUNICATION PROTOCOLS

The communication protocols ensure the transport of data from the source to the destinations in the VANET over the radio link. This section covers communication protocols at the network and transport layers and the application message/facilities layer.

VANET are self-organizing communication networks that operate without the need for a central control device or operator. As such, a VANET represents a special type of mobile ad hoc network with certain characteristics: (i) a potentially high number of nodes, (ii) nodes that can move at high speeds despite the constraint of road topology, (iii) frequent changes in the network topology, and (iv) the need to transmit data packets rapidly without prior signaling.

For the data in a mobile ad hoc network, two main types of communication protocols exist: single-hop protocols exchange data among nodes that are in direct communication range. With multi-hop protocols, data can be forwarded using relay nodes on the path from the source to the destination.

Routing refers to the process of finding routes and of forwarding packets at the network layer. For conventional mobile ad hoc networks, topology-based routing uses the information about the links that exist in the network in order to forward a packet. Two main classes of multi-hop, topology-based routing protocols exist: proactive routing protocols maintain routing information about the available paths even if these paths are not used. Reactive routing protocols maintain only the routes that are currently in use, typically based on a route discovery prior to packet transmission. Protocols of both classes do not cope with frequent changes in the network topology, as the route maintenance can occupy a major part of the bandwidth.

The specific characteristics of VANET have led to the development of dedicated network protocols in two main directions: (i) vehicular network protocols are optimized for single-hop communication and minimize the packet overhead, and (ii) multi-hop routing protocols for vehicular communication make use of geographical positions. The latter eliminate some of the limitations of topology-based proactive and reactive routing protocols and introduce geographical addressing for safety and traffic-efficiency applications.

On top of the network and transport protocols, additional application-related protocols, also referred to as *message protocols* or *facilities*, carry application-specific contents and provide direct support to the application processes. For safety and traffic-efficiency applications, in principle periodic and event-driven messages can be distinguished. While periodic messages convey status information, typically at a high frequency, the event-driven message is triggered upon detection of a safety-critical or safety-related situation. Periodic and event-driven messages are handled differently by the protocol stack in order to meet the application requirements on dissemination delay and reliability and the system requirements for stability and data congestion control.

Among the variety of research proposals for communication protocols, mainly two sets of communication protocols have prevailed; one in the US and standardized in IEEE and SAE, the other in Europe and standardized in ETSI and CEN. It is common to both protocol sets that they use the Internet Protocol (IP) for non-safety applications. For the safety-related column of the stack, a process to harmonize the standards is ongoing.

4.1 WSMP and BSM (US)

The US set of communication protocols include the WAVE short message protocol (WSMP) as a network protocol and the basic safety message (BSM) plus other messages as application message protocols, see Kenney (2011). They represent the left branch in the protocol stack in Figure 2.

WSMP is a network protocol that supports single-hop communication and is optimized for a minimum size of its packet header. A network packet, that is, a WAVE short

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

message (WSM), has a variable header length and variable payload length. Mandatory header fields (minimum of 5 bytes) include protocol version, provider service identifier (PSID), WAVE element identifier (WEI), and a field describing the length of the packet payload. The PSID identifies the service with which the payload is associated. The WEI indicates the format of the packet payload. Optional extension fields of variable length carry information about the channel number, data rate, and transmission power level. For operation in a multi-channel environment, the wireless access in vehicular environments service advertisement (WSA) is defined. This is a management message that informs about services offered in the vicinity of the sender. Upon reception of a WSA message, a receiver may decide to tune its transceiver to another channel.

BSM is a periodic message that is carried in the payload of the WSM. It transmits core status information of a sending vehicle, for example, position, dynamics, and size. The information carried in the BSM is shared by different safety applications, which makes the definition of application protocols unnecessary. The BSM has two parts. Part I is mandatory and must be sent by every vehicle. Part II is optional and includes fields that are required only for some safety applications, such as event flags (e.g., hard break), path history for lane-level information, path prediction for the intended driving direction, and GPS correction data. Part II allows for flexibility: data fields can be included with a lower frequency than the overall BSM rate and company-specific data fields can be used.

BSM is the most important message of the communication protocols, but a variety of other messages have been defined in SAE 2735, such as emergency vehicle alert message, signal phase and timing (SPAT) message for intersections and probe vehicle data.

4.2 GeoNetworking, BTP, CAM, and DENM (Europe)

In Europe, GeoNetworking and basic transport protocol (BTP) are the network and transport protocols standardized in ETSI. Application message protocols are the CAM and the DENM. The set of communication protocols comprise the left column of the protocol stack in Figure 3 in Section 2.

GeoNetworking is an ad hoc network layer protocol for single-hop and multi-hop communication that supports geographical positions for two main reasons: (i) it allows for addressing geographical areas in which a packet should be distributed, more precisely to address nodes in an area without using their node identifier, (ii) geographical positions are utilized in the forwarding process in order to make the packet routing efficient and scalable. GeoNetworking supports multiple transport modes: GeoUnicast for the communication between two network nodes, GeoBroadcast for the packet distribution to all nodes in a geographical area, as well as GeoAnycast to reach any (one) node in the area, see Festag, Papadimitratos, and Tielert (2010).

CAM is a periodic message with the same purpose as the BSM, that is, to transmit status information, and has a format similar to the BSM. A CAM can be sent by different types of vehicles—distinguished by profile—for which the CAM format has common data and profile-specific data. The message is structured in a header and several containers (Figure 7). Common data present in every CAM include the common header (fields for protocol version, message ID) and the basic container (time, type, etc.). In general, a CAM from a vehicle has a basic container for dynamic data (position, vehicle, and path) and for static data (vehicle size). Other containers can be added for special vehicles,

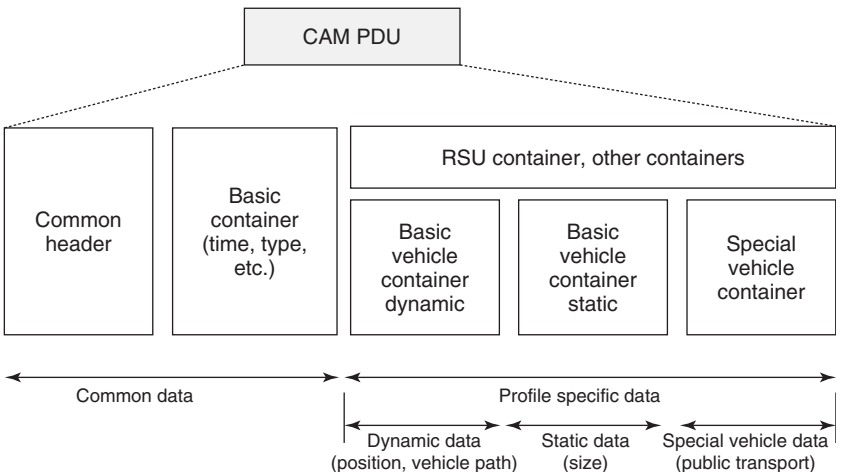


Figure 7. CAM structure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

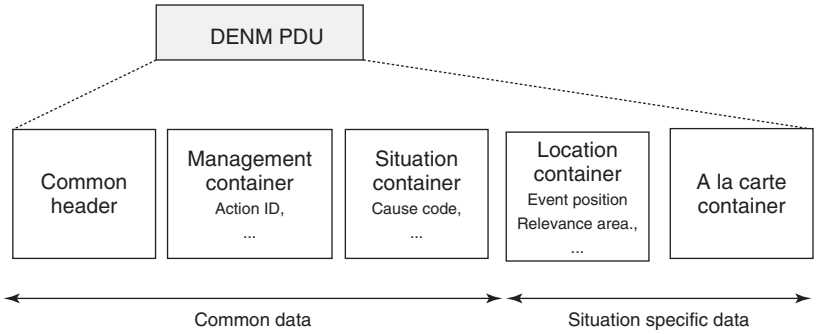


Figure 8. DENM structure.

such as for public transport or emergency vehicles). The concept of profiles ensures a flexible message format that can be adapted to the needs of the sending vehicle.

DENM is an event-triggered message, where an event gets assigned a unique action ID. As opposed to the single-hop CAM, a DENM targets a defined relevance area and can be transported using multi-hop, typically through GeoBroadcast at the network layer. Its message format has common data and situation-specific data. It is organized in a DENM header with protocol information and several containers (Figure 8). Containers for common data are the management container (e.g., fields for action identifier and data version) and the situation container (e.g., field for cause code). In the situation-specific data, the location container carries fields for the event position, the relevance area, and others. An a la carte container can be used to transmit application-specific contents. The DENM protocol handles the event lifecycle: an event with a specific action ID can be triggered and updated by the originator of the DENM, where event updates are distinguished by increasing data versions. An event can also be canceled by the originator or negated by a third station. A DENM can also be repeated by the originator, typically at a lower frequency than a CAM, to ensure that a message is received by vehicles that enter the relevance area.

Further application messages are also being standardized, such as for intersection topology (TOPO) and signal phase and timing (SPAT), which transmit static and dynamic information (e.g., traffic light phases) of intersections.

5 SECURITY AND PRIVACY

Security is a fundamental requirement for vehicular communication (Raya and Hubaux, 2007). In an example attack scenario, an adversary may publish fake traffic jams or repudiate existing ones, resulting in a chaotic road traffic situation. In a worse scenario, an adversary may fake an emergency brake message, which causes other drivers to brake hard, resulting in rear-end collisions. More systematically, the vulnerabilities of the system can be classified in: (i) jamming, (ii) forgery, (iii) in-transit traffic tampering, (iv) impersonation, (v) privacy violation, and (vi) on-board tampering. In order to cope with the vulnerabilities, five main security objectives can be identified (see ETSI TR 102 893, 2010 for more details):

- Confidentiality: messages should not be revealed to unauthorized parties. Also, unauthorized parties should not be able to deduce past or future locations, routes, and identities from communication means.
- Integrity: messages should be protected against unauthorized or malicious modifications or manipulations during transmission.
- Availability: communication should not be prevented by malicious activity.
- Accountability: audit all changes to security parameters and applications (updates, additions, and deletions).
- Authenticity: unauthorized stations should not be able to pose as authorized ones.

The security objectives can be achieved using different cryptographic means, that is, asymmetric, symmetric, and hybrid (asymmetric and symmetric) schemes. Among these, the asymmetric scheme based on digital signatures is favored (Papadimitratos *et al.*, 2008). Here, all stations obtain digital certificates and sign all their messages with their private key. The recipient verifies the integrity of the sender and data using signature verification. Optionally, depending on the case, the message payload can be encrypted to ensure confidentiality. The asymmetric scheme that is going to be used is elliptic curve cryptography (ECC). Compared to more common RSA-based methods, ECC offers moderate key and signature lengths for a reasonable duration of time for signature generation and verification.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

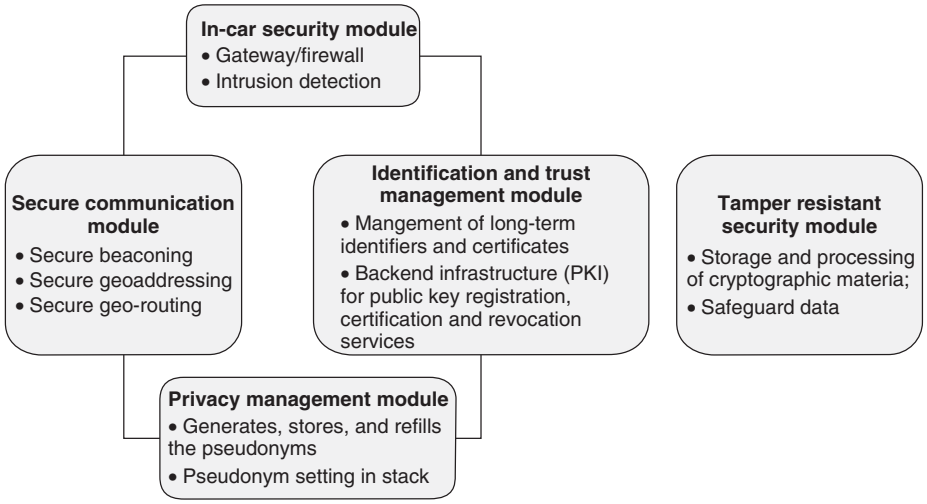


Figure 9. Security core elements. (Reproduced by permission of Kung *et al*., 2008. © Antonio Kung.)

In order to ensure confidentiality, messages can be encrypted. As encryption does not allow for location privacy, pseudonymity is applied: every station obtains a set of short-lived certificates with pseudonyms that belong to the same station and changes them frequently. If the changes occur often enough, the tracked data cannot be assigned to a single station. In order to prevent linking of new and old pseudonyms, addresses at all communication layers, including MAC addresses, must be changed simultaneously.

The cryptographic protection and privacy support using pseudonyms is complemented by plausibility checks. Such tests ensure that received data, such as position and speed values, are plausible according to system assumptions and physical rules, for example, the maximum communication range and timestamps.

An overall security architecture comprises five main components (see Kung *et al.*, 2008, baseline architecture for secure communication developed by the SeVeCom project) for secure communication, identification and trust management, in-car security, privacy management, and a tamper-resistant box to safeguard credentials (Figure 9).

6 RADIO RESOURCE MANAGEMENT AND DECENTRALIZED CONGESTION CONTROL

While it will be challenging to achieve a sufficient penetration rate in the beginning of the deployment stage, it will be even more challenging to deal with the large number of participants once every vehicle is finally equipped. As the capacity of the wireless channel is limited, congested channel situations, and therefore reduced channel access opportunities as well as increased packet collision probabilities, have to be expected. Consequently, the reliability and performance of inter-vehicle communication networks will drop significantly if no counter measures are implemented.

In order to deal with congested channel conditions, several mechanisms can be implemented: first, stations can knowingly "ignore" packet transmissions that arrive with very low signal strength and block a sender from sending. This strategy is typically referred to as an *adaptation of the CCA* threshold of the CSMA-based MAC layer, and increases the spatial reuse of the channel. Second, stations can reduce the time required to transmit a packet, for example, by adapting the modulation scheme and the coding rate used at the physical layer. While the usage of a higher order modulation scheme and coding rate will reduce the period of time that the channel is considered busy, the robustness against fading signal strengths is reduced as well. Hence, such an adaptation is not necessarily the best choice. As a third option, stations can reduce the amount of data traffic that is generated and transmitted to neighboring stations, for example, by an adjustment of the transmission power and/or the packet generation rate at the application layer (e.g., Huang *et al.*, 2010 and Tielert *et al.*, 2011). While a reduced transmit power increases the spatial reuse of the channel and limits the range within which a transmission can be detected and successfully decoded, an adaptation of the packet generation rate reduces the amount of data that may be transmitted, similar to how the transmission control protocol (TCP) of the internet protocol stack reacts to congested network conditions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Mechanisms that aim to reduce or limit the congestion on the wireless vehicular channel are usually called *congestion control protocols*. In relationship to the principles of distributed control theory, such protocols implement two core tasks: congestion detection and congestion control. As the impact of a station's transmission behavior (e.g., the used transmission power, modulation scheme, coding rate, and the average packet generation rate) cannot be determined by the station itself, but only by its neighbors, congestion detection requires the provisioning of feedback among neighboring stations, for example, through the exchange of locally measured channel congestion levels. On the basis of the collected feedback from all neighbors, the controller can then derive an appropriate reaction, for example, determine whether it should react at all, and if yes, how significantly. Similar to other control algorithms, oscillation and fairness issues have to be considered.

When considering congestion control, that is, the restriction of the load in the network, a uniform assignment of transmit power, packet generation rate, data rate, and CCA threshold values among neighboring stations can be considered as a fair solution to the scalability problem. Indeed, such an assignment would ensure that every station gets the same share of the channel. However, when considering a safety application implemented on top of the communication system, this might not be fair. A uniform allocation of channel resources might lead to different levels of safety in certain scenarios, for instance in an intersection scenario in which the awareness requirements of a vehicle that is approaching the intersection are probably higher than the awareness requirements of a vehicle that has already passed through the intersection. In order to ensure that minimum awareness (or performance) requirements are met, awareness control protocols are used in vehicular networks (Sepulcre *et al.*, 2011). Such awareness control protocols adjust the transmission parameters with respect to the current traffic situation and the corresponding safety application requirements. As those protocols each have a different objective target, their control decisions can conflict with the actions proposed by congestion control approaches.

As there are multiple ways to deal with congested channel conditions, radio resource management is clearly a cross-layer issue and cannot be solved on one layer alone. Further, the control algorithms applied to the individual layers need to consider the configuration parameters used by all other layers in order to be able to derive proper actions. It is therefore necessary to include a cross-layer configuration management in the protocol stack.

An information-theoretic formulation of the capacity of WLAN-based vehicular communication systems, considering the specific features of the system, was not available at the time of writing, as the optimal control of coding and modulation schemes, CCA, transmit power, and packet rates is further complicated by multiple and parallel application requirements, detection of radio environment, and game-theoretic considerations. Simulation results estimating the capacity for highway scenarios with two lanes in each direction are presented in Schmidt-Eisenlohr and Hartenstein (2010).

7 STANDARDIZATION

Standards enable the interoperability of devices for WLAN-based vehicular communication implemented by different manufacturers. Many standards have already been published or are still in the drafting stage. Involved standards development organizations (SDO) include global and regional/sector-based SDOs as well as industry consortia (Figure 10).

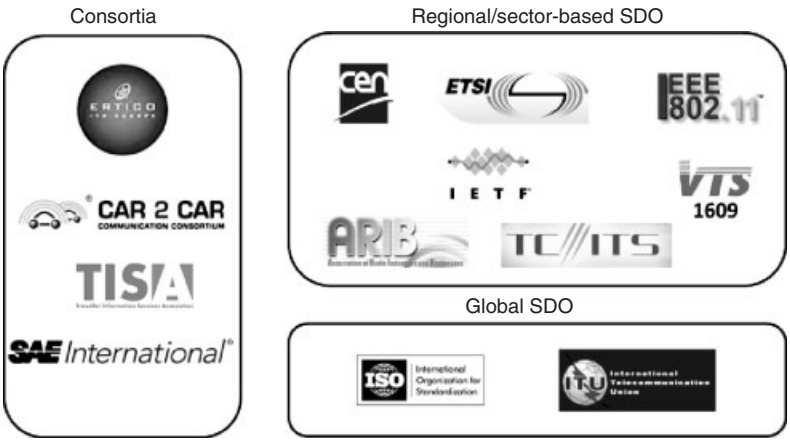


Figure 10. Standardization landscape for vehicular communication.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

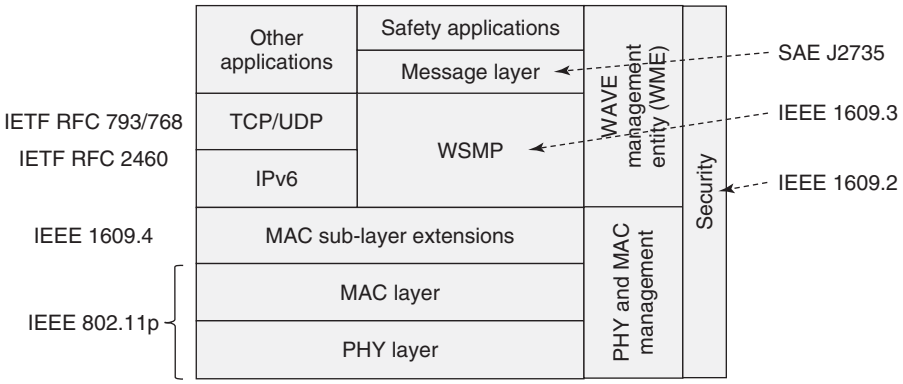


Figure 11. Standards related to the US protocol stack.

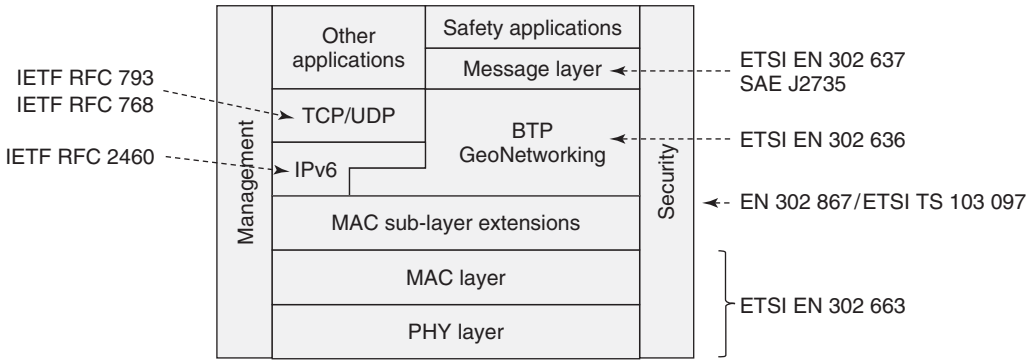


Figure 12. Standards related to the European protocol stack.

Following the main protocol stacks described in Section 4, two sets of standards are specified in the US (Figure 2) and Europe (Figure 3), respectively. At the physical and MAC layers, both stacks rely on IEEE 802.11p, referred to as *wireless access in vehicular environments* (WAVE). It is noted that IEEE 802.11p was merged into IEEE 802.11-2012, together with other variants of the same standards family. The European norm EN 302 663 standard represents a European profile of IEEE 802.11p. Also, both stacks have two branches, one for safety applications and the other for non-safety applications. Non-safety applications make use of the well-known IP protocols, specified in RFC 768, 793, 2460, and other RFCs.

The core part of the US set of standards is the IEEE 1609 family of standards: 1609.4 for channel switching, 1609.3 for networking services, in particular the WSMP and 1609.2 for security services (Figure 11). The message layer at the top of the protocol stack beneath the safety applications is specified in SAE J2735, with the WSM being the most important message.

In the European set of standards, the middle part is made up of the BTP and GeoNetworking specified in EN 302 636 and the CAM and DENM protocols in EN 302 637 (Figure 12). Other messages, such as TOPO and SPAT, follow basically the SAE J2735 standard that is also used in the US stack. Furthermore, EN 302 867 is a European profile of IEEE 1609.2. An overview of relevant standards is given in Table 1.

Beyond the main two standard sets, other standards exist, such as the ISO CALM family of standards developed by ISO TC 204 and the DSRC standards developed by the Japanese standard development organization ARIB.

8 OUTLOOK

Research and development for WLAN-based vehicular communication over the last decade has created a solid technical basis. Standards as a prerequisite for deployment are being developed. The system and technical components are currently being tested in field trials under realistic conditions in order to assess the maturity of the system and the positive impact of the system on road safety and traffic efficiency. The automotive and road infrastructure industry,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Core communication standards.

Number	Short title
IEEE 802.11-2012	Wireless LAN medium access control (MAC) and physical layer (PHY)
IEEE 802.11p	Wireless access in vehicular environments, amendment to IEEE 802.11
IEEE 1609.1	WAVE resource manager
IEEE 1609.2	WAVE security services for applications and management messages
IEEE 1609.3	WAVE networking services
IEEE 1609.4	WAVE multi-channel operation
SAE J2735	DSRC message set dictionary
SAE 2945	DSRC minimum performance requirements
ETSI EN 302 663	Access layer specification for ITS operating in the 5 GHz frequency band
ETSI EN 302 636	ITS GeoNetworking
ETSI EN 302 637-2	ITS cooperative awareness basic service (CAM)
ETSI EN 302 637-3	ITS decentralized environmental notification basic service (DENM)
ETSI TS 101 539-1	ITS road hazard signalling (RHS) application requirements
ETSI EN 302 867	Security, stage 3 mapping for IEEE 1609.2
ETSI TS 102 941	Security identity, trust, and privacy management
TS 103 097	Security header and certificate format

together with governments and consortia, are working on the system's introduction and sustainable deployment. For the initial deployment phase the following applications are foreseen:

- emergency vehicle warning
- emergency brake light
- stationary vehicle warning, V2X rescue signal
- traffic jam ahead warning
- in-vehicle signage (speed management)
- hazardous location warning
- contextual speed limit
- road work warning (stationary and moving)
- signal violation warning
- green light optimal speed advisory.

Additional infrastructure-related applications are considered:

- traffic info and recommended itinerary
- decentralized floating car data
- signal phase and time
- secure inter urban parking zone
- interactive bus corridor
- automatic access control—parking management.

GLOSSARY

Ad hoc communication	Communication without a coordinating infrastructure
BSM	Abbreviation of basic safety message; periodic message that is carried in the payload of a WSM
BTP	Abbreviation of basic transport protocol
CAM	Abbreviation for cooperative awareness message; periodic message that is carried in the payload of a GeoNetworking packet
Control channel	Common control channel for safety- and signaling-related information
DENM	Abbreviation for distributed environmental notification message; event-triggered message that carries event-related information
DSRC	Abbreviation for dedicated short-range communication; synonym for WLAN-based vehicular communication
EDCA	Abbreviation for enhanced distributed channel access; allows prioritization of the medium access of data frames as part of the IEEE 802.11 standard
Event-driven message	Message, typically non-periodic, triggered by the detection of an event and transmits event-related information
Facilities	Layer in the protocol stack of an ITS station that provides application-related functionality
GeoNetworking	GeoNetworking is an ad hoc network layer protocol for single-hop and multi-hop communication that utilizes geographical positions
ITS station	Core element of the ITS architecture with instantiations for vehicles, roadside, and central stations
Multi-hop protocols	Exchange of data, where the data can be forwarded using

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

	relay nodes on the path from the source to the destination
OCB	Abbreviation for outside the context of a BSS; simplified ad hoc communication mode in IEEE 802.11p
Periodic messages	Awareness messages that are broadcasted by any vehicle about own presence and status, such as BSM and CAM
Pseudonymity	Functionality for privacy support using short-lived pseudonym
SAE 2735	Core standard for the definition of facilities messages, such as BSM
Single-hop	Exchange of data among nodes that are in direct communication range
WAVE	Abbreviation for wireless access in vehicular environments, synonym for WLAN-based vehicular communication
WSMP	Abbreviation for WAVE short message protocol; network protocol that supports single-hop communication and is optimized for a minimum size of its packet header

REFERENCES

Acosta-Marum, G. and Ingram, M.A. (2007) Six time- and frequency-selective empirical channel models for vehicular wireless LANs. *IEEE Vehicular Technology Magazine*, **2** (4), 4–11.

Alexander, P., Haley, D., and Grant, A. (2011) Cooperative intelligent transport systems: 5.9-GHz field trials. *Proceedings of the IEEE*, **99** (7), 1213–1235.

ETSI TR 102 893 (2010) Intelligent Transport Systems (ITS); Security; Threat, Vulnerability and Risk Analysis (TVRA).

Festag, A., Papadimitratos, P., and Tielert, T. (2010) Design and performance of secure geocast for vehicular communication. *IEEE Transactions on Vehicular Technology*, **59** (5), 1–16.

Huang, C.-L., Fallah, Y.P., Sengupta, R., and Krishnan, H. (2010) Intervehicle transmission rate control for cooperative active safety system. *IEEE Transactions on Intelligent Transportation Systems*, **12** (3), 645–658.

Jiang, D., Chen, Q., and Delgrossi, L. (2008) Optimal Data Rate Selection for Vehicle Safety Communications. *Proceedings of the Fifth ACM International Workshop on VehiculAr Inter-NETworking* (VANET '08). ACM, New York, NY, pp. 30–38.

Kenney, J.B. (2011) Dedicated short-range communications (DSRC) standards in the United States. *Proceedings of the IEEE*, **99** (7), 1162–1182.

Kung, A. (ed) (2008) Security Architecture and Mechanisms for V2V / V2I. Sevecom Project, Deliverable D2.1 v3.0.

Mecklenbräuker, C.F., Molisch, A.F., Karedal, J., *et al.* (2011) Vehicular channel characterization and its implications for wireless system design and performance. *Proceedings of the IEEE*, **99** (7), 1189–1212.

Mittag, J. (2012) Characterization, avoidance, and repair of packet collisions in inter-vehicle communication networks. Dissertation. Karlsruhe Institute of Technology.

Molisch, A.F. (2010) *Wireless Communications*, 2nd edn, Wiley. ISBN: 978-0-470-74186-3.

Papadimitratos, P., Buttyan, L., Holczer, T., *et al.* (2008) Secure vehicular communications: design and architecture. *IEEE Communications Magazine*, **46** (11), 100–109.

Raya, M. and Hubaux, J.-P. (2007) Securing vehicular ad hoc networks. *Journal of Computer Security*, **15** (1), 39–68.

Schmidt-Eisenlohr, F. and Hartenstein, H. (2010) Simulation-based Capacity Estimates for Local Broadcast Transmissions. *Proceedings of the Seventh ACM International Workshop on VehiculAr InterNETworking* (VANET '10). ACM, New York, NY, pp. 21–30.

Sepulcre, M., Mittag, J., Santi, P., *et al.* (2011) Congestion and awareness control in cooperative vehicular systems. *Proceedings of the IEEE*, **99** (7), 1260–1279.

Tielert, T., Jiang, D., Chen Q., *et al.* (2011) Design methodology and evaluation of rate adaptation based congestion control for vehicle safety communications. *Proceedings of IEEE Vehicular Networking Conference (VNC)*, Amsterdam, Netherlands, pp. 116–123.

FURTHER READING

Dressler, F., Kargl, F., Ott, J., *et al.* (2011) Research challenges in Inter-vehicular communication: lessons of the 2010 Dagstuhl seminar. *IEEE Communications Magazine*, **49** (5), 158–164.

Hartenstein, H. and Laberteaux, K. (eds) (2010) VANET—Vehicular Applications and Inter-Networking Technologies, Wiley, ISBN: 978-0-470-74056-9.

Kosch, T., Schroth, C., Strassberger, M., and Bechler, M. (2012) *Automotive Internetworking*, Wiley. ISBN: 978-0-470-74979-1.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto170
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Technologies—Positioning GNSS

Klaus P. Jaschke[1,2]

[1] *German Aerospace Center, Institute of Transportation Systems, Braunschweig, Germany*
[2] *Thales Transportation Systems GmbH, Berlin, Germany*

1 Introduction 1
2 Global Navigation Satellite Systems 1
3 Outlook 7
Abbreviations 7
References 7
Further Reading 8

1 INTRODUCTION

Global navigation satellite systems (GNSSs) are commonly used to detect the absolute position and support the navigation of objects such as cars, trucks, buses, trains, or even pedestrians and cyclists. A GNNS receiver calculates autonomously the position, velocity, time, and acceleration of the object. The two global operational and running GNSSs are the global positioning system (GPS) set up by the US and the Russian globalnaja nawigazionnaja sputnikowaja sistema (GLONASS).

GNSS is widely used in various applications to detect the position of an object; however, there are several constraints because of the physical principle of the system. There are signal runtime errors because of the ionosphere and troposphere. Additionally, biases are caused by offsets of the satellite clocks or errors in the calculated ephemerids. Some of these errors could be mitigated by the so-called augmentation systems; others such as multipath errors can only be detected and corrected with special (high budget) receivers or special antenna designs. The problem of signal shadowing, for example, in urban canyons or tunnels cannot be solved without the use of other sensors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto172
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 GLOBAL NAVIGATION SATELLITE SYSTEMS

An autonomous system for the absolute detection of the position of objects is the satellite-based navigation system. The first applications using this system are navigation systems that detect the position of the vehicle and calculate the routing to a defined destination that is based on the underlying map data. Additionally, positioning information is used to fuse it with other sensor data and achieve a better availability and accuracy.

2.1 Existing systems

Currently, there are two GNSSs in use. One is the US system GPS, also known as *NAVSTAR* (*navigation system for timing and ranging*), that was fully operable in 1995; the other is the Russian system GLONASS. GLONASS was fully set up and operational 1 year later (Mansfeld, 2004). Nowadays, both systems are renewed, so the second generations will be fully operable in 2014. Additionally, China is extending their local BeiDou system to the so-called compass navigation system to become a global system as well. In Europe, the European Community is setting up the Galileo positioning system that is intended only for civil use and will be compatible with the US system. Therefore, the combination of GPS and Galileo will lead to a better coverage and availability because of more satellites in sight.

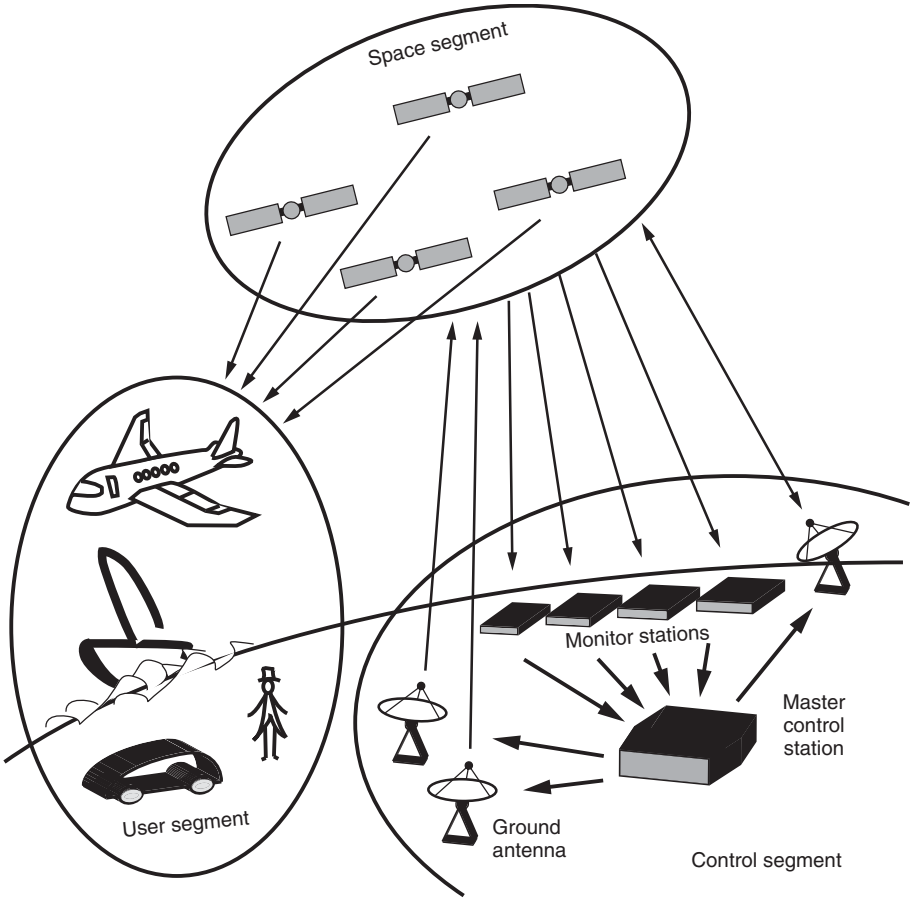


Figure 1. GPS segments.

The structure of GPS—as well as of the other systems—consists of three segments: the space, the control, and the user segment (Figure 1).

The space segment includes the constellation of the satellites and the satellites themselves. The six circular orbits are inclined at an angle of 55° relative to the equator and are separated from each other by multiples of 60° right ascension; see Figure 2 (Grewal *et al*., 2007). The orbits are arranged so that at least six satellites are always within the line of sight from almost everywhere on the Earth's surface (Kaplan and Hegarty, 2006)—while having no shadowing due to tunnels, buildings, or vegetation. The non-geostationary orbits are approximately circular, with radii of 26,560 km and orbital periods of one-half sidereal day (approximately 11.967 h).

The control segment consists of five base stations: the master control station in Colorado Springs that controls the whole operation of the system and the GPS time, and the monitoring stations and ground control station on Hawaii, Ascension Island (Atlantic Ocean), Diego Garcia (Indian Ocean), and Kwajalein (Pacific Ocean). The major tasks are the prediction of the satellite orbits, the monitoring of the satellite clocks, and the transmission of the navigation messages to the satellites (Kaplan and Hegarty, 2006). The ground control stations are the communication interface to the satellites from where the ephemerides and clock parameters are transmitted, while the monitoring stations are equipped with highly precise cesium clocks and multiple GPS receivers. These monitoring stations collect data, process these data, and correct them with ionosphere and meteorological data before transmitting them to the master control station.

The user segment consists of receivers specifically designed to receive, decode, and process the GPS satellite signal. Receivers can be stand-alone, or integrated with or embedded into other systems. GPS receivers can vary significantly in design and function depending on their application for navigation, accurate positioning, time transfer, surveying, and attitude reference (US Coast Guard, 1996).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto172
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

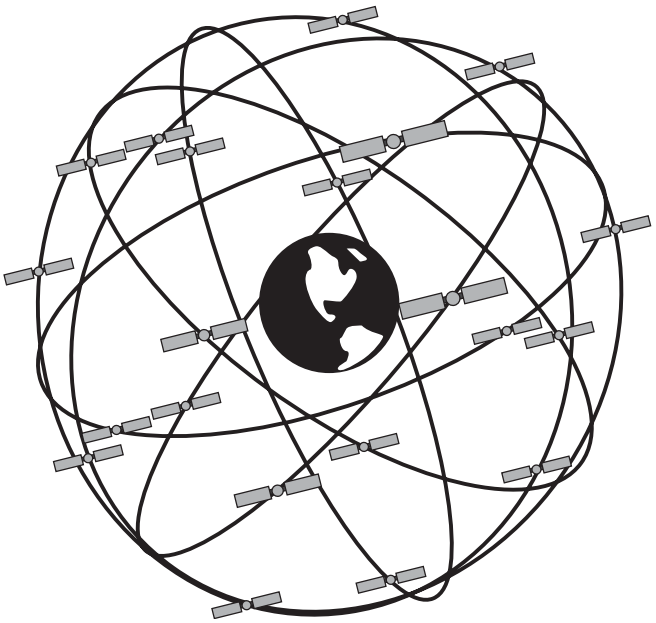


Figure 2. GPS—constellation.

2.2 General principle

For satellite-based positioning, the calculation of the position is based on a runtime measurement of the signal between the satellites and the receiver. The position of the satellites could be taken as well known, and the signal transmission speed could be approximated with the speed of light. Furthermore, the clocks of the satellites and receiver are assumed as highly precise. The principle to detect the runtime is the use of pseudo-random noise codes that are produced at the same time within the transmitter (satellite) and the receiver. Therefore, all transmitted satellite signals are stored in a reference data store. The time shift between the signals could then be derived by shifting the two signals of Δt until they match, see Figure 3 (Meyer zu Hörste and Lemmer, 2009). This can be achieved by finding the maximum cross-correlation of these two signals.

The satellites transmit the following information (Mansfeld, 1993; Grewal *et al*., 2007):

- *Satellite almanac data*: Each satellite transmits orbital data called *the almanac*, which allows the user to determine which satellites are visible. It helps the user to calculate the approximate location of every satellite in the GPS constellation at any given time at a given location.
- *Satellite ephemeris data*: Ephemeris data are only broadcast by satellites for their own position. It is a much more accurate satellite position—analogous to the almanac data—that is needed to convert signal propagation delay into an estimate of the user position. The data is valid only for several hours.
- *Signal timing data*: The 50-bps (bits per second) data stream includes time tagging. This data is needed to determine the satellite-to-user propagation delay for ranging.
- *Ionospheric delay data*: Ranging errors due to ionospheric effects can be partially canceled, while ionospheric delay correction data is broadcast in the data stream.
- *Satellite health message*: The datastream also contains information regarding the current health of the satellite, so that the receiver can ignore that satellite if it is not operating properly.

Considering two satellites, two possible intersections exist considering the signal runtime; however, only one position is real. Assuming the speed of light $c_0 \approx 2.99 \times 10^8$ m/s as the propagation speed, which is not the real propagation speed, already tiny clock errors result in big errors in the position calculation. To avoid these

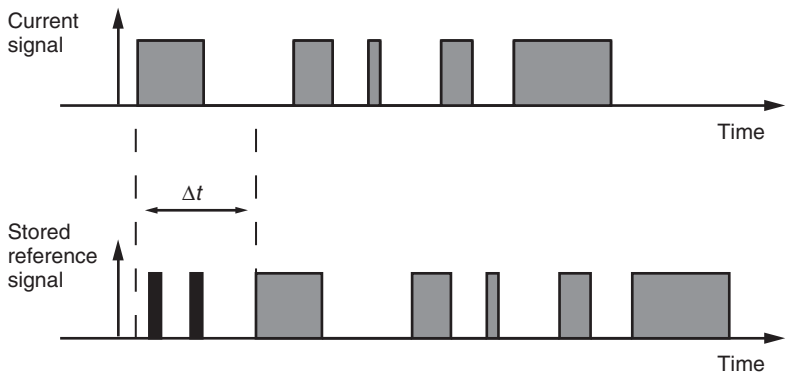


Figure 3. Time shift of transmitted and received signal.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto172
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

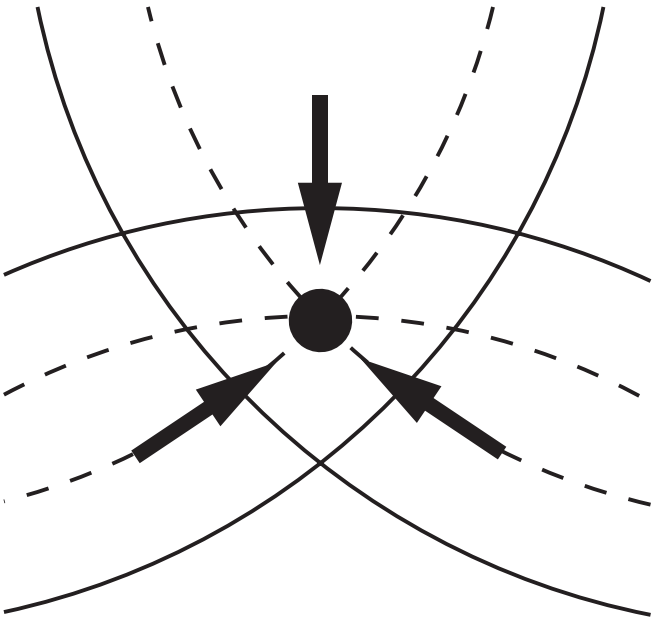


Figure 4. Triangulation—position detection in a plane using three satellites (two dimensional).

errors, compensations based on models for propagation not in vacuum and clock error compensation should be considered. The last errors could be mitigated by the use of a third satellite. On the basis of the principle of triangulation, the position can be calculated, see Figure 4.

Considering also the z-position (altitude), four satellites should be used for the calculation, see Figure 5.

Mathematically, the estimated distance between the receiver and the satellites (the so-called pseudo range ρ_r) could be expressed with the known satellite positions x_i, y_i, z_i, the radius of the earth r, the clock errors N_{rr}, and the unknown position of the receiver X, Y, Z:

$$\rho_r = \sqrt{(x-X)^2 + (y-Y)^2 + (z-Z)^2} \tag{1}$$

$$\rho_r^2 = (x^2 + y^2 + z^2) - r^2 = N_{rr} - 2Xx - 2Yy - 2Zz \tag{2}$$

$$\begin{bmatrix} \rho_{r1}^2 - x_1^2 + y_1^2 + z_1^2 - r^2 \\ \rho_{r2}^2 - x_2^2 + y_2^2 + z_2^2 - r^2 \\ \rho_{r3}^2 - x_3^2 + y_3^2 + z_3^2 - r^2 \\ \rho_{r4}^2 - x_4^2 + y_4^2 + z_4^2 - r^2 \end{bmatrix} = \begin{bmatrix} -2x_1 & -2y_1 & -2z_1 & 1 \\ -2x_2 & -2y_2 & -2z_2 & 1 \\ -2x_3 & -2y_3 & -2z_3 & 1 \\ -2x_4 & -2y_4 & -2z_4 & 1 \end{bmatrix} \times \begin{bmatrix} X \\ Y \\ Z \\ N_{rr} \end{bmatrix} \tag{3}$$

As seen from Equations 1–3, the errors derive from the inaccuracy of the position of each satellite. The satellite coordinates are available via ephemerides that are calculated on ground-based observation of the satellites (Mansfeld, 2004).

Owing to the biases with different error sources described in the next section, the position accuracy depends on the available satellites and geometrical conditions. Typically, a precision of ± 15 m could be achieved without any

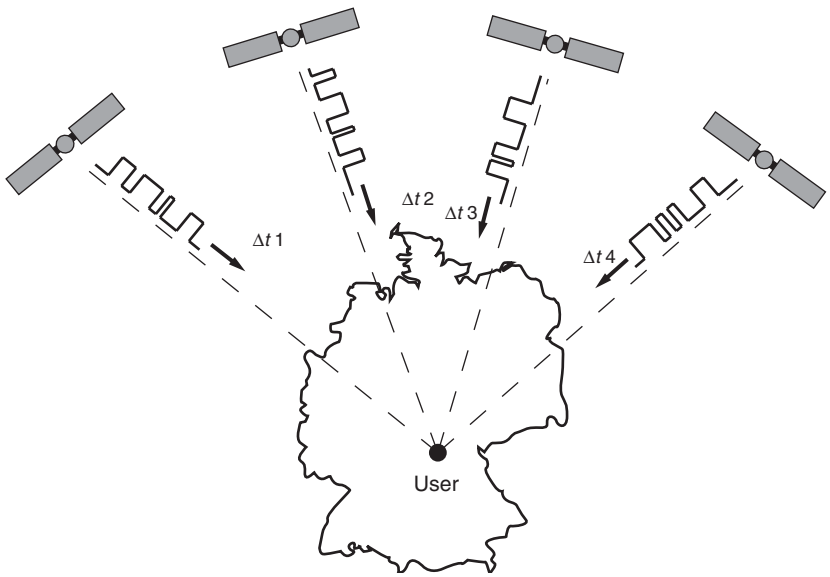


Figure 5. Detection with four satellites (three-dimensional).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto172
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

augmentation systems and ±1 to ±3 m with augmentation systems, depending on the environmental conditions, for example, open sky and urban canyons.

2.3 Error sources

As already shown in the previous section, one major bias comes for clock errors, which can be corrected by using additional satellites for the calculation of the position. Additional signal runtime delays are caused while traveling through the ionosphere. Radio signals are refracted, delayed, or scattered. Extreme travel delay may cause range errors of up to 100 m. Plasma turbulences may cause rapid fluctuations of the signal strength, well known as *ionosphere radio scintillations* (Jakowski, 2008).

Figure 6 shows the possible error sources for the positioning:

- Clocks: error between 0 and 1.5 m.
- Ephemeris error < 3 m: it becomes smaller as satellite tracking technologies improve.
- Ionosphere delay: error up to 30 m during the day and up to 6 m by night.
- Troposphere error: 0–10 m, also depends on satellite elevation angles.
- Multipath and reflection: can lead to errors up to 50 m and more, typically 0.1–3 m.

For GPS, a former problem was also the selective availability (SA), where the civilian GPS signal was disturbed to reduce the accuracy such that this signal could not be used by enemies, for example, to steer weapon systems. In 2002, President Bill Clinton ordered SA to be turned off at midnight of May 1.

Moreover, the shielding of satellite signals is a problem, which occurs in tunnels, urban canyons, or other obstacles such as trees (especially wet foliage) because of no detection at all or a reduction of visible satellites and weak detection of signals. Owing to the physical principles, this problem can be solved only by independent additional sensors, such as, for example, inertia platforms or wheel sensors and an adequate sensor data fusion.

Multipath errors such as the ones mentioned earlier could be minimized by using special antenna designs or other means. These mitigations are currently mainly used for stationary measurements and will not overcome the challenge for automotive applications (moving vehicles or vulnerable road users). Here, it has to be seen how new low cost receivers might minimize these errors by, for example, using the second frequency with the military code—which

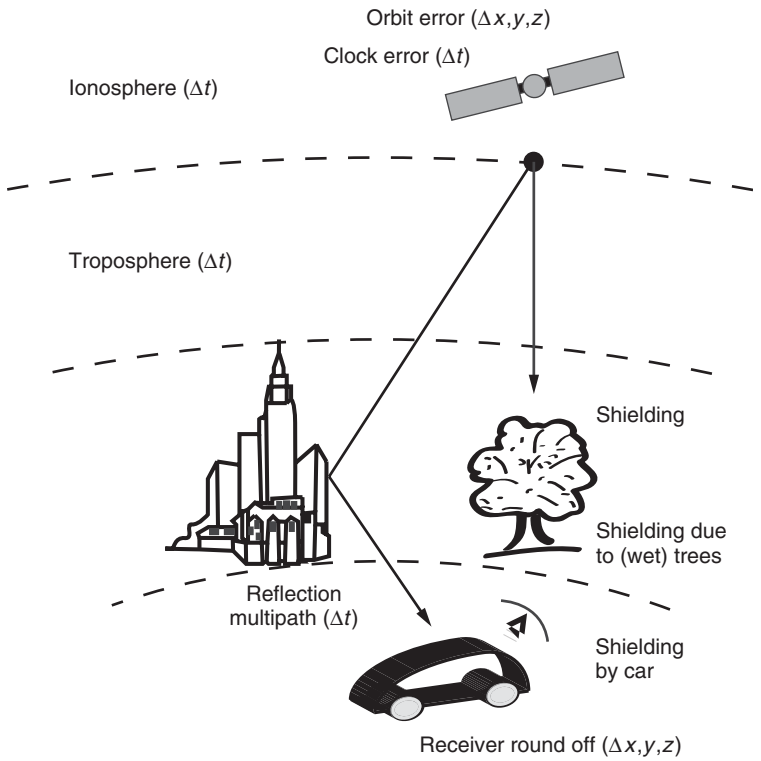


Figure 6. Errors and EGNOS.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto172
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

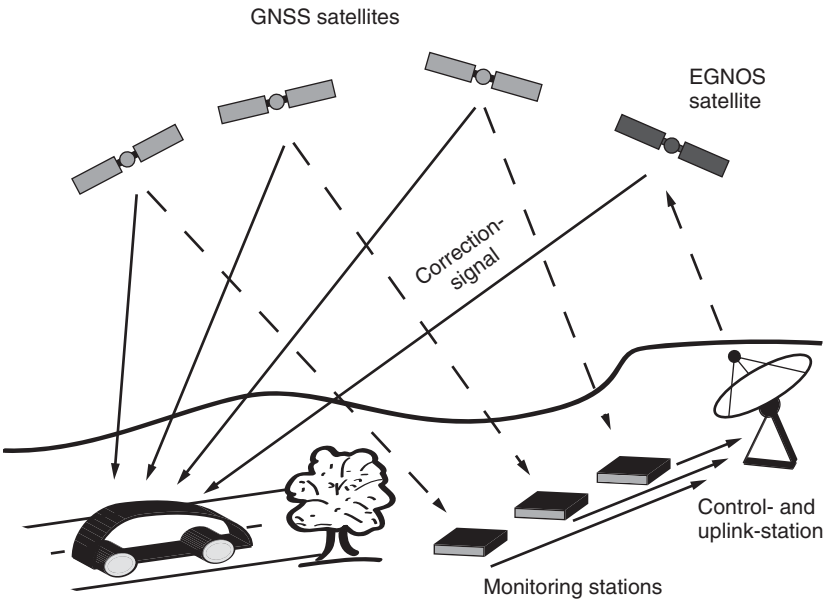


Figure 7. Satellite navigation with satellite-based augmentation, for example, EGNOS.

is not available for civilians—to estimate multipath errors or establish other algorithms for mitigation. The current way forward is still the use of other sensor data and data fusion with the GNSS position fixes.

2.4 GNSS augmentation

Some of these errors can be corrected by GNSS augmentation systems. Very well known is the local differential global positioning system (DGPS). Here, a highly precise reference position is known and is compared to a recent measurement. The difference (delta) between the reference and the measurement is used to calculate the correction data that is sent to receivers in an area very close to the reference point. Now the receivers can use this correction signal to enhance the precision of the current position calculation. DGPS belongs to the so-called ground-based augmentation system (GBAS). GBAS can be used to cover an area of approximately 20 km around the reference station.

An alternative is satellite-based augmentation systems (SBASs). Here, various ground stations exist whose positions are well known and highly precisely measured. These stations calculate with the satellites in view correction data that is then provided via satellites for the users. One of these SBASs is the European Geostationary Navigation Overlay Service (EGNOS), where correction data for the clock and orbit errors as well as corrections for the ionosphere are provided (Figure 7). For North America, the Wide-Area Augmentation System (WAAS) as air navigation aid is well known, with the goal to enhance the accuracy (lateral < 2 m, vertical < 3 m) and the availability $> 99\,\%$ of time (Grewal *et al*., 2007).

2.5 Applications

In the beginning, the use of GNSS was mainly focused on air, space, and maritime navigation and positioning systems. As the receivers (integrated or stand-alone) have become cheaper over the years and are now available as low budget off-the-shelf receivers, the market for other applications will grow tremendously.

For land applications of GNSS, it started with surveying and geodesy applications while the receivers were still of high cost. Owing to their massive cost reduction, vehicle tracking and navigation have become an emerging market, as well as hand-held receivers for hiking and for data collection of other recreational pursuits (Jacobson, 2007). While the receivers are now available as single-chip solutions, they are also integrated into smartphones and other hand-held devices to achieve precise timing and positioning at very low costs.

Looking at vehicle navigation systems, today most drivers use in-built navigation systems of the manufactures or stand-alone systems (hand-held systems or even cellular/smartphones) for navigation. In parallel with the receiver development and their cost reduction, the availability of more and more precise maps for these navigation techniques and the optimization of tracking and mapping algorithms additionally have pushed the market. With the arrival of car-2-car and car-2-infrastructure applications for

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto172
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

cooperative driver assistance, highly accurate positions of the cars are needed to calculate and predict the right interactions and also to broadcast these positions via cooperative awareness messages (CAMs) and decentralized environmental notification messages (DENM). However, there are also other applications that will use GNSS positioning, such as eCall or tolling application. Owing to the error sources mentioned earlier, most systems will not use a standard GNSS system to detect the position, but will fuse these GNSS data with other sensor data such as wheel sensors and inertia platform. Additionally, links via cellular or wireless to the Internet and other back-ends can bring further advantages and business cases in combination with GNSS.

3 OUTLOOK

Owing to the physical principles and the errors described earlier, it is still a challenge to mitigate these errors to achieve better availability and accuracy with GNSS systems only. As seen in some recent research projects, see for example, GSC (Lindholm, 2011), DemoOrt (Meyer zu Hörste and Lemmer, 2009), the low cost receivers have become more and more powerful, thereby increasing the accuracy of the position fixes. Even comparing just the latest to the last receiver generation brings added value to the accuracy and availability of the GNSS. However, the problem of no satellites being in sight due to shielding, for example, in tunnels, remain and cannot be solved by GNSS only.

Looking at the augmentation system, EGNOS and also the upcoming European Galileo system will bring advantages for accuracy in fixing the position. As EGNOS is now fully operational, at least two geostationary satellites are available to receive the correction signals. Owing to the placement of these satellites, a good coverage even in urban areas is available that could be used for the mitigation of errors. Additionally, it could be expected that new GPS receivers will be upgraded with the opportunity to use the EGNOS corrections. On the other hand, it can be foreseen that the new Galileo system will bring better coverage and availability in combination with the existing GPS. Owing to the same signal setup, receivers can use all visible GPS and Galileo satellites to calculate the position.

For cooperative drive assistance systems and other mobility-related applications, it has to be seen how the future deployment of these systems and their integration will evolve. Already, today it can be seen that a lot of indication applications are available on smartphones or other mobile devices that could be mounted on vehicles. For safety-relevant applications, especially with, for example, active interaction with the steering or breaking system of a car, GNSS alone will not be sufficient to fulfill the positioning and availability criteria. Therefore, even if augmentation systems are operational, additional sensors are needed to achieve an overall trustworthy and robust system.

ABBREVIATIONS

CAM	Cooperative awareness message
DENM	Decentralized environmental notification message
DGPS	Differential global positioning system
EGNOS	European geostationary navigation overlay service
GBAS	Ground-based augmentation systems
GLONASS	Globalnaja nawigazionnaja sputnikowaja sistema
GNSS	Global navigation satellite system
GPS	Global positioning system
SBAS	Satellite-based augmentation system
WAAS	Wide-area augmentation system

REFERENCES

Grewal, M., Weill, L., and Andrews, A. (2007) *Global Positioning Systems, Inertial Navigation, and Integration*, Wiley-Interscience, New Jersey.

Jacobson, L. (2007) *GNSS Markets and Applications*, Artech House, Boston.

Jakowski, N. (2008) *Space Weather and Ionosphere*, http://www.dlr.de/kn/en/desktopdefault.aspx/tabid-2204/3257_read-9150 (accessed 4 July 2013).

Kaplan, E. and Hegarty, C. (2006) *Understanding GPS: Principles and Applications*, Artech House, Norwood, MA.

Lindholm, R. (project coordinator) (2011) *European Funded Project: Galileo Service Convergence (GSC)*, http://www.ertico.com/unlocking-the-potential-of-galileo-and-egnos (accessed 4 July 2013).

Mansfeld, W. (1993) *Funkortungs- und Funknavigationsanlagen*, Hüthig, Heidelberg.

Mansfeld, W. (2004) *Satellitenortung und Navigation*, Vieweg, Wiesbaden.

Meyer zu Hörste, M. and Lemmer, K. (2009) *Entwicklung eines Demonstrators für Ortungsaufgaben mit Sicherheitsverantwortung im Schienengüterverkehr—DemoOrt, Berichte aus dem DLR-Institut für Verkehrssystemtechnik*.

US Coast Guard Navigation Center (1996) *NAVSTAR GPS User Equipment Introduction*, http://www.navcen.uscg.gov/pubs/gps/gpsuser/gpsuser.pdf (accessed 26 July 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto172
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

FURTHER READING

Hofmannn-Wellenhof, B., Lichtenegger, H., and Wasle, E., (2008) *GNSS Global Navigation Satellite Systems*, Springer.

Zhao, Y. (1997) *Vehicle Location and Navigation Systems*, Artech House, Norwood, MA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto172
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Tracking and Navigation for Goods and People

Moritz Kessel, Martin Werner, Florian Gschwandtner, and Claudia Linnhoff-Popien

Ludwig-Maximilians-University Munich, Munich, Germany

1 Introduction 1
2 Fundamentals of Positioning Systems 2
3 Positioning Methods 4
4 Technologies for Indoor Positioning 6
5 Conclusion 10
Related Articles 11
Glossary 11
References 11

1 INTRODUCTION

Classically, the aim of positioning is mainly to give orientation and navigational aids. As the rise of the global positioning system (GPS) as an inexpensive positioning technology, many applications have been designed that make daily life easier. Nowadays, the complexity of navigating a car in foreign environments has been completely moved into the digital world. The problem is reduced to having the right map, enough electrical power, and a concrete point of interest. This is one source of interest in indoor positioning services: namely, provide a high quality orientation service inside buildings. From this perspective, it would of course be optimal to have a seamless positioning system that works inside and outside buildings equally well based on the same technology. However, such a system does not yet exist.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Sometimes, it seems that the aim for indoor navigation is a bit pointless, as often buildings are used by small, closed groups (except public buildings such as airports, shopping malls, etc.) and the problem is seldom orientation and localization. However, another reason for interest in indoor positioning comes from the areas of security management and control of production quality. High quality indoor positioning can allow activity recognition and can enable a production system to determine whether a given set of activities has been performed. As a very good example for the automotive domain, the paper by Zinnen, Wojek and Schiele (2009) explains how to use high quality ultrawideband positioning to do some type of body-model-derived primitive selection to determine, in high dependability, whether a given set of quality checks has been done to a car. The only source of information in this case comes from the 3-D localization tags mounted to the quality checkers themselves.

The main barrier to a wide adoption of such systems lies, of course, in the area of privacy: that is, does the system give enough advantage over a classical checklist, such that the worker accepts that technology keeps himself under permanent surveillance?

A third area of application is, of course, the tracking of goods and people inside buildings. For goods, it can be very important to have a clear understanding of where and when the goods have been moved inside a warehouse. The tracking of people can be important and accepted in high-risk environments, where at any point in time a rapid evacuation might become essential. A common application domain where human beings need permanent surveillance inside buildings is the domain of ambient-assisted living. The aim of ambient-assisted living is to allow elderly people to live in their own homes for a longer time using digital surveillance where nowadays living in a nursing home is

mandatory. Positioning and activity recognition are then used to detect situations where the residents need help immediately.

From this diversity of application domains and scenarios, many different positioning and navigation techniques have been developed, all of which have their strengths and weaknesses. In the following, we explain, in Section 2, the fundamentals of positioning systems and their relation to map information and different notions of position. Section 3 discusses the algorithms that can be used to infer location from observations at a high level of abstraction. In Section 4, we give a broad overview of the existing positioning systems. This section is organized along the physical sizes used for position determination and how the algorithms of Section 3 have been adopted to specific environments and systems. Section 5 concludes this article with a hint on research perspectives.

2 FUNDAMENTALS OF POSITIONING SYSTEMS

Position is possibly the most important source of context for mobile context-aware systems. However, position does not make sense without environmental information that can be used to infer some interpretation of location. This is essentially important for indoor positioning, as the determination of positions inside buildings is error-prone and hence the interpretation of positioning results becomes more complex.

2.1 Modeling of indoor locations

While for outdoor positioning, navigation, and guidance very simple maps containing a graph of roads interconnected by junctions would suffice, the problem of navigation cannot be solved by having a graph of "possible ways" inside a building. Imagine a large hallway with infinite possibilities of pedestrian movement, not only restricted to one-dimensional lines, for example, the edges of a navigation graph, but also free movement in a two-dimensional area. Providing step-by-step guidance or utilizing map matching with techniques known from outdoor car navigation is not possible. In addition, the determined positions are often not accurate enough to map the location to one single edge in a navigation graph.

As a solution to these problems, it is common practice to use more advanced environmental models than interconnected networks of points. These models are often tailored to the quality of the positioning system, the available map data, and the service demands of the intended location-based service.

Following Hightower and Boriello (2001), these models can be best understood from a classification of positioning algorithms. The authors define three types of such positioning algorithms that are described in detail in Section 3: triangulation, in which distances and angles are used to infer a position; proximity, in which the nearness to some known points is measured; and scene analysis, in which a set of observations, which vary with location, is used to infer a location of a mobile device. In the cited work, the authors limit scene analysis to a view from a particular point inside the navigation space; however, nowadays especially signal-strength patterns of existing wireless infrastructure are often used to infer positions with some method of machine intelligence that we want to include in the term "scene analysis."

On the basis of these three types of position inference, the type of position is completely different: triangulation approaches typically lead to numeric coordinates in some reference coordinate systems; proximity detection typically limits the possible set of locations to a smaller area; and scene analysis typically calculates some kind of probability distribution of location.

Hence, environmental models should be able to deal with these types of location. In consequence, environmental models inside buildings should be able to model geometric coordinates of course, but also to model symbolic coordinates such as room names. This is because, often, a scene analysis method reaches poor performance unless it is used with symbolic coordinates between which the measurements change significantly. Think for example of images taken with a camera. Two images of some object inside a room, which have been taken from different points, are still very similar, while the same movement distance between two other points might lead to fundamentally different images, because the semantic location has changed (e.g., leaving a room typically changes the complete appearance of the scene).

Consequently, indoor environmental models typically model location in a hybrid form and allow translating between symbolic and geometric coordinates. Furthermore, positioning is possibly based on symbolic coordinates, and hence on areas rather than points. To be able to calculate shortest ways, an environmental model must provide a sensible meaning of distance, spatial containment, and reachability even for symbolic places.

While the need of geometric coordinates leads to a spatial representation of the model in form of a coordinate system, where the building and relevant objects are assigned two- or three-dimensional coordinates, there are several ways to maintain symbolic coordinates. Becker and Dürr (2005) differentiate between set-based, hierarchical, and graph-based approaches. The set-based model consists of subsets of the set of all symbolic coordinates. The subsets can

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

be used to define overlapping locations, but a set-based model directly supports neither distance nor reachability queries. Hierarchical models directly model containment of symbolic locations, for example, the containment of rooms in floors, but provide no information about topological interconnections between locations. In the graph-based approach, symbolic locations are modeled as the nodes of a graph that are connected by an edge if a direct real-world connection exists. This means that two rooms (being symbolic locations) are usually connected by an edge if there is a door in between. Graph-based models explicitly describe the reachability and enable the calculation of distances, but have no means to describe spatial containment. In conclusion, an indoor environmental model should not only be hybrid in form of symbolic and geometric coordinates but also have a graph-based and a hierarchical (or set-based) representation of symbolic coordinates.

2.2 Different aspects of positioning systems

Before different methods and systems are explained in detail, we want to focus on some aspects that can be important in choosing the right systems and for which, in contrast to the situation outside buildings where most systems rely on a single source of location such as GPS, different positioning systems are favorable.

The first general consideration is about mobility: if the targets to be localized are humans, they can move freely. If the target is a machine, its movement can be limited: a car cannot move sideways. If the mobility of the target is passive, for example, given by conveyor, the possible movement is completely known. Of course, this should influence the choice of the positioning system. A simple and inexpensive radio frequency identification (RFID) detector gives a pretty exact location in time and space for a conveyor-based production system using proximity. RFID for localization of people would lead to very expensive systems, as the functionality of RFID is limited to a very small area and hence a sensor network of RFID readers would have to be deployed and maintained at high costs.

Another important general consideration is among the scaling parameter. Often, there is a correlation among cost, accuracy, and coverage. If the positioning system has to provide a highly accurate position to a single mobile entity, an expensive inertial sensor unit can be the best choice. On the other hand, systems that have to provide a coarse location to a multitude of objects will not use expensive sensors mounted on to the objects. They should rely on infrastructure-based positioning, optimally on existing infrastructure, as is the case with wireless local area network (WLAN) localization.

Cost calculations and their scaling with respect to accuracy, coverage, and the number of targets can be very important for the right choice of indoor positioning systems. From this viewpoint, one typically differentiates between terminal-based positioning, where the sensing and position calculation is carried out by the mobile entity itself, infrastructure-based positioning, where the position of a mobile item is determined by some infrastructure possibly without any communication with the mobile entity, and terminal-assisted positioning, where the position of a mobile entity is calculated by some infrastructure, but the sensing of the parameters for position estimation is done by the mobile entity.

The use of dedicated infrastructure in general leads to installation and maintenance costs that scale with the area of localization. A modern highly accurate indoor positioning system based on ultrawideband technology typically uses four or more sensors per room, all of which need dedicated power and communication cables and induce costs, of course. The use of existing infrastructure such as WLAN, however, does not incur any (additional) costs. However, indoor localization using WLAN is much less accurate than indoor localization using dedicated wideband signals.

For terminal-based positioning systems, costs basically scale with accuracy, electric power demands, and the number of items to be localized. Every mobile item has to be able to calculate its own position and basically needs a computation unit and a sensing unit. If not only the mobile entity itself is interested in the position, one additionally needs a communication unit. Nevertheless, the effort has one advantage: terminal-based positioning offers privacy, which might be a desirable goal for the sensitive location information of a human user.

For terminal-assisted positioning systems, costs basically scale with all these parameters: the area of coverage where the infrastructure has to be installed and maintained, the number of mobile objects that have to be localized, and the desired accuracy and precision of location. However, in some cases, no additional costs occur at all. Think of using existing mobile phones, which already have a WLAN interface, along with an existing WLAN infrastructure and an Internet-based localization service, which takes signal strength information and returns location. In this situation, no additional cost is generated, and truly ubiquitous and cheap indoor localization becomes possible, although limited in accuracy, as WLAN was never designed to be used for localization.

Thus, when thinking of installing an indoor positioning system, one has to carefully consider the design principles and weigh the cost against the desired accuracy and coverage. Another basis for the right choice of system is given in form of the positioning methods used, because

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

they also have an impact on the deployment, the infrastructural needs, and even the required physical properties of the site. Some systems, for example, only work in line-of-sight conditions between the positioning infrastructure and the target.

3 POSITIONING METHODS

Localization and tracking of people and goods inside buildings is much more difficult as compared to positioning outside buildings. The main reason is that radio-based methods have difficulties dealing with attenuation and multipath effects. Hence, many methods that are dedicated to deal with these circumstances and are able to even exploit these propagation complexities have been developed.

As described in the previous paragraph, a simple way to organize positioning methodology (following Hightower and Boriello (2001)) is into the three categories of triangulation, proximity detection, and scene analysis. We want to follow this organization, but add a fourth type of position determination called *dead reckoning*, in which an initial position is updated according to measurements concerning the acceleration, movement, and heading of a mobile entity.

3.1 Triangulation

The methods of triangulation are defined to be using the geometry of a plane triangle to deduce position information. One such method is lateration, in which distances between known points and a mobile entity are measured. Another such method is angulation, in which angles between different reference positions and a mobile entity are measured. Combinations are also possible: for example, the measurement of an angle and a distance from a known point resulting in a position.

3.1.1 Lateration

The most common method of triangulation is multilateration in which multiple distance measurements from multiple reference points with known locations are being used to find the position. For undisturbed measurements, a single distance estimation limits the locus of the object to be tracked to a circle centered at the given reference position with a radius given by the distance measurement. As a result, at least three reference positions (which must not be collinear) are needed to be able to uniquely identify the location of the target. In practice, such circles will not intersect in a common point because of measurement errors. It is common practice to use Gaussian least-squares regression on a linearization of the circle equation given by a Taylor series expansion. As a characteristic of such extensions, this results in a correction vector for an initial position, which could, for example, have been chosen as the middle point of all reference points. This correction vector is then used to update an initial position until this process converges to the position of minimal least-squares residuum.

A special form of lateration is given by hyperbolic lateration, where only the distance difference of a mobile entity to pairs of reference stations is known. As all points that have the same distance difference to two fixed points give the definition of a hyperbola, this is known as *hyperbolic lateration*. The method of dealing with disturbances is the same as for classical lateration; only the circle equations have to be replaced by the hyperbolic equations for the Taylor expansion.

3.1.2 Angulation

For position determination by angulation (Figure 1), the angles between given reference positions are measured, for example, by antenna arrays, and the location of the mobile entity is given by the intersection of rays starting at each reference location into the measured direction. For this intersection to take place, the orientation of the reference measurement units has to be known. Again, the method of Gaussian least-squares regression along with Taylor linearization leads to an iterative location determination algorithm. The measurement of angles at known reference locations is very dominant in practice, but it is important to mention the possibility of measuring all angles at the mobile entity. This typically leads to complexity in mobile entities, but the consistency between angles is automatic.

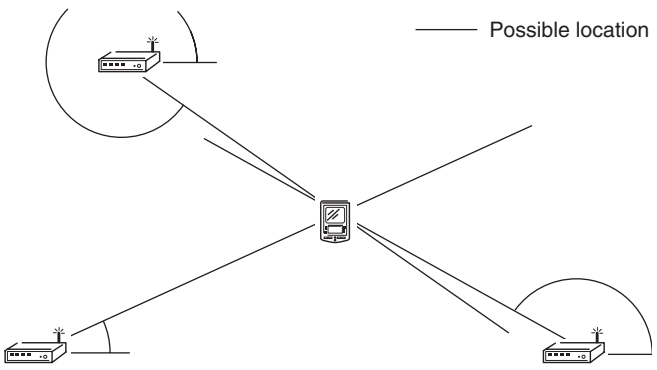


Figure 1. Location determination by angulation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Inside buildings, it is difficult to infer a distance between two points. Consequently, lateration in general leads to erroneous results inside buildings. Nevertheless, in areas where the propagation characteristics of signals are known (typically free space for sound, light, and radio waves), lateration is used. For the determination of distances, there are two main possibilities: either a measurement of signal strength that leads to a distance estimation based on expected path loss, or a measurement of time. For time measurements, the following three main methods can be distinguished: time of arrival, time difference of arrival, and roundtrip time of flight.

3.1.3 Time of arrival

In the case of time of arrival, pilot signals are sent at a known time at known locations. The time difference between sending and receiving of a pilot signal can often be used to infer the distance from the propagation speed of the signal (e.g., speed of light, speed of sound). These distances are called *pseudo-ranges*, as they can be quite different from the actual distance because of time synchronization errors, reflection, scattering, shadowing, and fading. All entities have to be time-synchronized, and one pseudorange estimation leads to a circle of possible locations. The position is then given by the intersection of those circles (Figure 2). This is in effect the same method as used by GPS.

3.1.4 Time difference of arrival

For the case of time difference of arrival, pilot signals are emitted at the reference locations at equal times, and the mobile entity records the time difference between receiving those pilot signals. This gives a good basis for hyperbolic multilateration. The most important advantage over the classical time-of-arrival method lies in the fact that the mobile entity needs no time synchronization with the infrastructure at the reference locations. As all points that have the same distance difference to two different positions lie on a hyperbola, each distance difference estimation leads to a hyperbola and the position is given by the intersection of these hyperbolas for more pairs of reference locations.

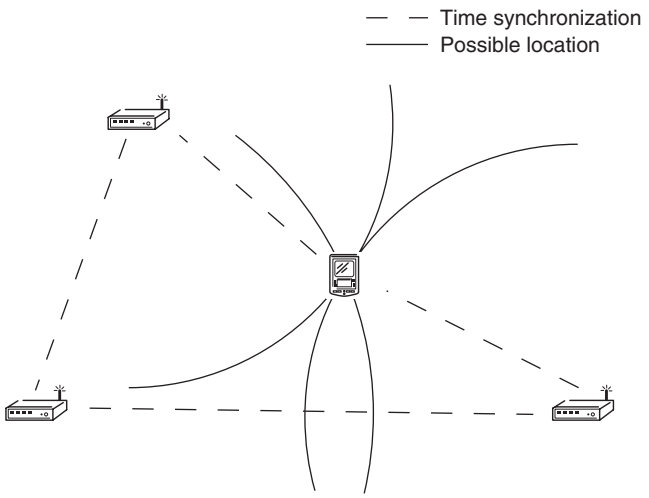


Figure 2. Time-of-arrival positioning.

3.1.5 Roundtrip time of flight

In scenarios of roundtrip time of flight, (typically) the mobile entity initiates a measurement by sending a signal that is mirrored back either by physical objects in space or even by active stations located at known reference locations. The mobile entity measures the time difference between sending the initial signal and receiving the reflected signal. Half of this time, possibly reduced by a processing time for active reflectors, gives the time that the signal needed to travel from the mobile entity to the reflector, which can be translated into a pseudo-range using the signal propagation speed (Figure 3).

3.1.6 Angle of arrival/angle of emission

For angulation, two main types of angle determination can be distinguished. The first one is associated with the term "angle of arrival," in which the angle of an incoming signal is determined, for example, by an antenna array. The second variant is called *the angle of emission* or *theta coding*, in which a signal is sent out, containing digital or analog information identifying the angle of emission. In both cases, the angle information is used along with the known reference locations to infer a position.

3.2 Dead reckoning

Dead reckoning is a method of calculating subsequent positions out of a fixed initial position. As such, dead reckoning is often based on the measurement of initial sensors giving acceleration and heading. All these sensors have in common that they measure a first or second derivative of the location (e.g., acceleration, acceleration in rotation, velocity). The location at a given time is then given as the integral of these accelerations. For discrete measurements, this integral is given by a sum. The most important problem of this type of positioning is the fact that measurement errors add up, leading to an unrealistic

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

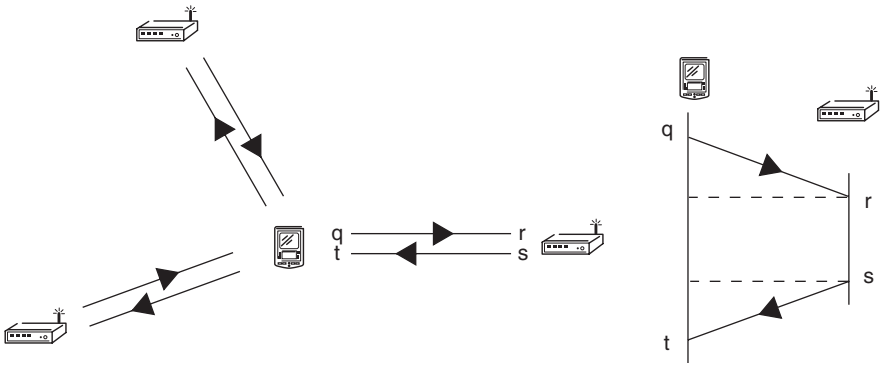


Figure 3. Roundtrip time-of-flight positioning.

movement state (e.g., nonzero speed for a stationary object) and location. Hence, dead reckoning is typically used as an intermediate method for the time between two subsequent position fixes of a slow positioning system or using high-quality, expensive sensors in a short timeframe.

3.3 Presence detection

Presence detection can be seen as a special form of lateration in which it is only known that the distance between a mobile entity and some reference point is below a fixed threshold (e.g., given by the area of coverage of a wireless network). In these cases, the position is often given as the mean of the positions of the stations that detect presence or as a symbolic coordinate.

3.4 Scene analysis

The technique of scene analysis uses methods of pattern matching and artificial intelligence to recall long-living environmental properties of given locations. A classical approach called *fingerprinting* is to divide the navigation space into cells that can be characterized by some statistics of measurable values such as the signal strength distribution of an existing wireless infrastructure. This type of positioning is often tied to a specific application and does not allow for a general description. You will find many different examples throughout the next section. In general, scene analysis is in some sense orthogonal to triangulation, as disturbances of propagation can be very characteristic for a specific place, and pattern-matching can be much more precise for highly perturbed signal propagation scenarios where the performance of geometric location determination techniques degrades. Scene analysis techniques are successfully used on almost any environmental feature that can be measured by mobile entities and varies with location (Figure 4).

4 TECHNOLOGIES FOR INDOOR POSITIONING

With the high number of positioning methods, it is no surprise that there also exist a multitude of technologies that can be used to infer the position of a target in indoor environments. Often, each technology supports multiple methods and each method can be applied with a variety of technologies. In the following, we give an overview of those technologies and the corresponding methods that are used to infer positions, accuracy levels, and application scenarios.

4.1 High sensitive GNSS, pseudolites

Signals of global navigation satellite systems (GNSSs), for example, GPS, suffer from attenuation effects and multipath mitigation in indoor areas. Therefore, those systems offer inaccurate or no position information at all, which can be compensated by senders on ground, for example, in or near the buildings, where position information is needed.

One kind of such senders, called *pseudolites*, emit GPS-like signals that can be used for lateration based on the time of arrival. Similar to satellite positioning systems, these senders need to be strongly time-synchronized and their position must be precisely known (Cobb, 1997). The signals can be used either together with existing GNSS to enhance the visibility of signals in degraded environments or to provide additional senders for more reliable and accurate positioning, or independent of any satellite system for the purpose of indoor positioning without reception of satellite signals. While pseudolites can offer very high accuracies of few centimeters or even millimeters (Cobb, 1997), they do not work well in non-line-of-sight conditions. Owing to their high accuracy, pseudolites can be used in automated industrial production.

A system similar to pseudolites is the Locata system. Locata offers a time-synchronized network of LocataLites,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

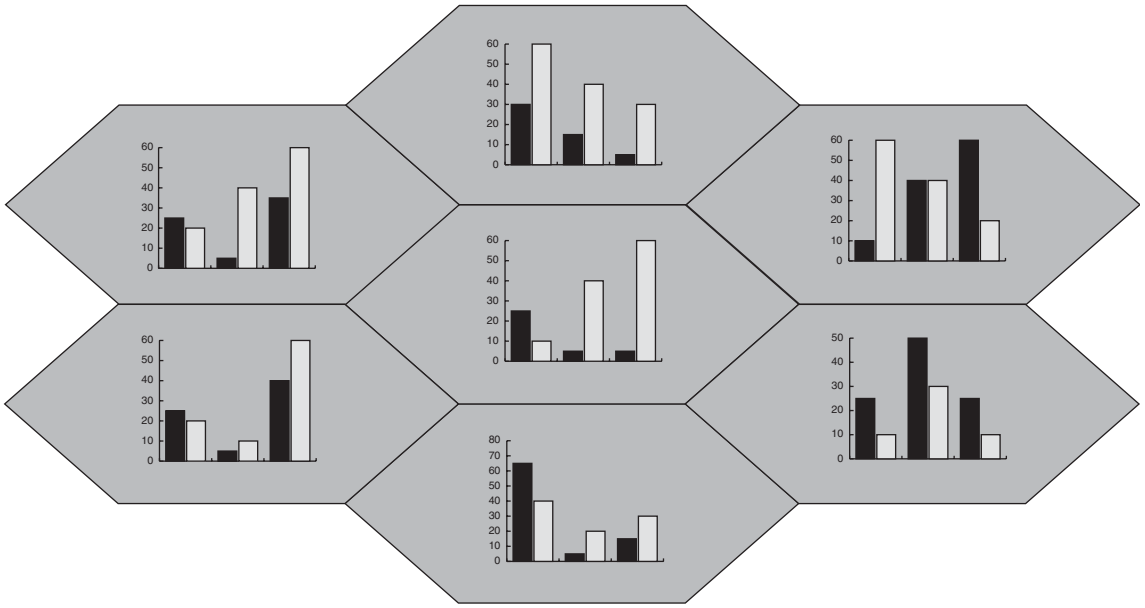


Figure 4. Scene analysis and position fingerprinting.

which transmit GPS like signals in another frequency spectrum [e.g., the license-free ISM (industrial, scientific, and medical) band at 2.4 GHz]. These signals can be used to calculate the position of a component called *Locata*. This component is able to receive standard GPS signals as well as signals from LocataLites and, therefore, is able to calculate its own position with centimeter-level accuracy even when no or too few GPS signals are available for position calculation, as long as enough LocataLites are visible (Rizos *et al*., 2010).

Despite the complex signal propagation in buildings, there are approaches to provide indoor positioning with highly sensitive GPS or GNSS receivers. Even if the signals are weak and might have been reflected, they can be detected with the help of highly sensitive receivers and utilized for position calculation. Experiments demonstrate the ability of such systems to offer positioning capabilities in indoor areas, but the accuracy is generally lower than in outdoor areas.

4.2 Light-based systems

A very promising technique for indoor localization is the use of special light. As light interacts with most indoor material in a much more deterministic way as compared to radio signals, it is possible to calculate distances to objects using the methodology of roundtrip time of flight. These systems are called *LiDAR* (light detection and ranging) systems. The term *LiDAR* reflects the fact that the basic working principle is the same as the one used by radar outside.

Travis, Simmons, and Bevly (2005) were able to detect depth information in a field of view spanning 180° with high frequency, allowing positioning and even continuous navigation using existing maps. The same system can also be used to generate three-dimensional map information using techniques of "simultaneous localization and mapping" (SLAM). Such systems can basically be enriched by the application of detectable landmarks and mirrors for the laser detection. With the use of a few such landmarks, variations in the depth image (due to existing or missing objects) do not harm the positioning process anymore. Tracking of special natural features inside buildings (such as edges, corners, and other regularities) is also possible by application of some image processing and analysis to a LiDAR-acquired depth image, possibly in combination with a camera image (Adams and Kerstens, 1998).

The same type of information can be generated by a much cheaper technology that does not rely on the measurement the roundtrip time of flight of a laser signal. By sending out a regular pattern of light, it is possible to calculate a depth image by observing the deformation of the pattern due to obstacles. The best-known example of this type of system is the Kinect, which is used by the Microsoft XBox to track movements of players. Unfortunately, the working range of such systems is bounded by the resolution of the camera used to detect the pattern.

Another class of light-based systems relies on presence detection by generating modulated light inside the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

navigation space. This modulated light can be generated by an infrastructure and detected by mobile devices allowing for self-contained positioning, or the mobile entity can send an identification code to a network of sensors (Want *et al*., 1992).

Though the use of color image information detected by a digital camera is of course a kind of light-based systems, we describe camera-based systems in a separate section because the techniques are quite different from active light-based positioning systems.

4.3 Camera-based systems

Another technology often found in indoor positioning applications is a camera. The field of camera positioning can be subdivided according to two different approaches. In the first approach, cameras are attached to a moving object with an unknown position that should be determined. This is called *camera egomotion* in the following. The other approach relies on stationary cameras with known positions, which are used to estimate the position of targets moving through the cameras' views (Mautz and Tilch, 2011).

In camera egomotion, there are again different methods for position determination. One method relies on scene analysis, where distinctive features, objects, or landmarks are extracted from the camera view. In the case of natural features, they are compared with a database of previously recorded images, where the position of recording is known. Depending on the size of the database and the resolution of the images, the matching may take some time, but offers highly reliable position information. Werner, Kessel and Marouane (2011) were able to obtain submeter-level position accuracy with natural feature matching. Problems arise from different points of view concerning the database images and the image used for positioning, which leads to scaled and rotated variations of the captured scene. Therefore, scale- and rotation-invariant descriptions of scenes, for example, generated by the well-known SIFT (scale-invariant feature transform) algorithm (Lowe, 2004), offer high reliability. Instead of natural features, artificial distinctive markers such as barcodes can be distributed in the environment. When the position of the marker is known or can be stored inside the marker itself, the problem of scene analysis can be reduced to the detection of markers in an image. This problem is much easier to solve and faster to calculate than natural features, but the markers need to be set up carefully and are prone to partial occlusion by other objects.

Another method for position estimation often encountered in camera egomotion is the use of time-domain information from consecutive images, called *optical flow*. The technologies of SLAM, for example, calculate the movement of a camera projection between adjacent frames by solving a point-set configuration problem that comes from marker or natural feature comparison between frames. This type of system suffers from accumulation of measurement errors over time, as the next position is always calculated from the previous defective position. For recalibration, SLAM often relies on a technique called *loop closure*, where the mobile entity returns to an already mapped location. By identifying the scenery, errors in the trajectory can be corrected and the map quality heavily enhanced. However, this type of algorithm soon gets very complex, as it depends on the complete history, and the point-set configuration problem in itself is very hard and often only solved using a randomized Monte Carlo method.

There are some other systems that use commercial optical mouses (or similar techniques by taking a video of the ground) for dead reckoning (similar to optical flow). These systems are, for example, found in low-cost robot localization systems, where the sensor data of the mouse is used to compensate slip effects of the robot's wheels.

Stationary cameras detect targets moving through the captured scene. As the position of the camera and the position of objects in the field of view can be calibrated, it is easy to retrieve the position of a target. However, stationary cameras are faced with the task of identifying targets in the scene to assign the calculated position to the right target. Furthermore, it can be expensive to provide full coverage for the whole area of positioning and might imply privacy problems, as photos or videos of individuals could be recorded to calculate position estimates and map them to users. In addition, image segmentation needs to be performed carefully in the case of partial occlusion when multiple users are in front of the camera. Therefore, stationary cameras are often used for high accuracy positioning of robots or construction components in automated production scenarios.

4.4 Radio-based systems

Many systems for indoor positioning are based on radio signals. However, there are many different methods and technologies, and how radio information is utilized to deduce the position of a target. There are approaches based on timing, for example, time of arrival, time difference of arrival, and roundtrip time of flight, based on signal strength, either for lateration or for scene analysis (here often called *fingerprinting*), based on angulation, or even presence detection. A similar diversity can be observed concerning the radio technologies. There exist systems based on cellular networks for mobile communication

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

such as GSM (global system for mobile communications) or CDMA (code division multiple access), personal area networks such as Bluetooth, WLAN, RFID, or ultrawideband. Even radio or television signals can be used for indoor positioning.

Utilizing radio signal for positioning is not a new idea, as in outdoor areas the common positioning technologies GPS and cellular positioning rely on radio signals. The latter also work in indoor areas, but suffer from increased inaccuracy because of the multipath propagation, fading, and attenuation. Furthermore, the physical characteristics of cellular communication enable the signals to easily penetrate walls, making it hard for fingerprinting techniques to distinguish between adjacent rooms.

Owing to the lower range (less transmit power) and the physical characteristics of the frequency band (higher frequency), WLAN-based techniques are often able to provide more accurate position estimates as compared to cellular positioning. When using fingerprinting (which is considered to be more accurate than timing approaches), the estimated position in indoor areas often lies within a few meters of the real position. Furthermore, many buildings have already an infrastructure of wireless access points installed. One of the first methods for WLAN positioning was based on fingerprinting (Bahl and Padmanabhan, 2000). However, fingerprinting requires a time-consuming calibration phase with the need for recalibration when structural changes alter the signal propagation characteristics. Much research has been done on the subject of calibration. There are approaches dealing with the simulation of propagation of WLAN signals, allowing automatic calculation of the expected signal strength [e.g., by using a building model, a propagation model, and counting walls between the known position of an access point and a certain reference position (Bahl and Padmanabhan, 2000)], automatic calibration techniques using crowd sourcing approaches, additional measurement stations, or mobile robots (Ocana *et al*., 2005). Another field of extensive research is dedicated to the algorithms for position estimation. Machine learning algorithms such as *k*-nearest neighbors, naïve Bayes or Bayesian networks, support vector machines, and neuronal networks have been proposed and extensively utilized for positioning (Liu *et al*., 2007). WLAN is often the basis of cheap pedestrian positioning systems in environments where a WLAN communication infrastructure already exists. However, a trend toward a combination of WLAN fingerprinting with additional positioning technologies such as inertial sensors exists.

While WLAN-based positioning techniques seldom acquire submeter-level accuracy, methods based on ultrawideband do. Those systems usually calculate the position based on short pulsed signal bursts from a target, which are received and evaluated in a time-synchronized infrastructure of receivers. One of the most successful commercial systems, Ubisense, combines time difference of arrival and angle of arrival for 3-D positioning with centimeter-level accuracy (Steggles and Gschwind, 2005). Owing to limitations in transmit power, most wideband-based systems typically are restricted to approximate line-of-sight conditions, that is, do not provide coverage through walls. As ultrawideband-based systems are comparatively expensive, they are often set up in industrial production scenarios in large factory halls, where the benefit compensates the expenses.

Another radio-based technology, RFID, is mainly used for presence detection in positioning scenarios. RFID readers are placed throughout the building, especially in corridors or at doors, and their position is stored according to a reference system. Whenever a target comes near such a reader, the system is able to receive a short-range signal from the target's RFID tag and can therefore deduce that it is near the position stored for that reader. While the main use of RFID is the identification, a coarse location of an identified item can also be retrieved when the location of the reader is known. This is often the case in industrial settings, for example, when RFID is used to locate and identify objects on a conveyer.

Finally, any other radio technology can be used to infer the position of a target using one of the described methods. The reason why WLAN is popular at the moment comes from the distribution of WLAN and the capability of mobile devices used for positioning. Some 20 years ago, a system would use infrared as a near-field communication and presence detection technology, while this can nowadays be achieved by near-field communication or Bluetooth. The latter could also substitute WLAN positioning, but as there often is no fixed Bluetooth infrastructure, it is seldom used for positioning.

4.5 Inertial navigation

Inertial sensors measure physical effects independent of any infrastructure. Examples are accelerometers sensing acceleration and gravity, gyroscopes measuring rotation, magnetic field sensors (i.e., compass) for orientation, barometers measuring the air pressure that can be used to deduce the altitude of the sensor, and odometers measuring wheel rotation to calculate the traveled distance. Some of these sensors provide absolute values such as the direction of a compass or the altitude of a barometer, whereas some offer relative changes such as the rotation of a gyroscope or the acceleration as a change of velocity. In general, inertial sensors do not provide a position directly, but can

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

only report changes of position. Thus, inertial measurement units (IMUs) rely on dead reckoning techniques for position estimation and generally need to be supported with an initial position. However, the drift of IMUs leads to an accumulation of the position error over time.

For indoor positioning, inertial sensors are often combined with other positioning technologies to compensate for the drift, and to offer a better accuracy, coverage, or continuity. This integration is achieved by sensor fusion mechanisms such as Kalman (Kalman, 1960) or particle filters (Arulampalam *et al*., 2002). While the Kalman filter is an approach for modeling linear dynamic systems with a Gaussian distribution, particle filters are sequential Monte Carlo methods where a continuous probability distribution is approximated by a point cloud of particles. For indoor positioning, however, both work with a state vector and two phases called *prediction* and *update*. The state vector represents all relevant information of the observed system, that is, the estimated position, speed, and orientation of the target. This vector is altered in the prediction phase according to a system model that is often based on IMU data and then corrected using a measurement model based on absolute but noisy position measurements of some other technology. From a statistical point of view, the prediction and correction can be understood as the prior and the posterior distribution in a Bayesian approach.

In the field of indoor positioning, IMUs have been investigated mainly for pedestrian positioning, utilizing accelerometers as pedometer by counting steps or to measure the speed by integrating the acceleration in combination with a compass or gyroscope for the direction of movement (Woodman and Harle, 2008). However, there is also ongoing research for robot localization and SLAM techniques.

4.6 Audio-based systems

Using audio signals for indoor positioning is one of the early approaches for indoor positioning, but has continued as an active field of research up until now. The early systems such as ActiveBat (Ward, Jones and Hopper, 1997) used ultrasonic signals emitted by a moving target, which were captured by a dense infrastructure of receivers. Those approaches offered centimeter-level accuracy at high expenses, meaning that the infrastructure was expensive and therefore was only installed in small areas such as meeting rooms. The positioning method used for these systems is often multilateration based on time of arrival or time difference of arrival.

Many other approaches, such as microphone arrays capturing the main direction of pulsed audio signals or the utilization of the acoustic background spectrum in different rooms, have also been investigated in recent years. Microphone arrays are used together with angulation techniques, where either an infrastructure of senders with known positions emit pilot signals (beacons) and the target determines its position with the help of the microphone array (by exactly measuring the time of reception or signal strength of the signal at each microphone in the array), or the target transmits audio signals and the infrastructure of microphone arrays captures the signal and calculates the position of the target.

The technique based on the acoustic background spectrum was investigated for pedestrian indoor localization with smartphones in a university building (Tarzia *et al*., 2011). The authors were able to distinguish between several rooms at different times of day, although the chatter of multiple people impeded the use of their system.

4.7 Pressure-based systems

Finally, there exist some very specialized indoor positioning systems that measure the pressure created by a target moving over the ground. One system, the Smart Floor (Orr and Abowd, 2000), is based on a network of pressure sensors in the ground to detect location and identity of pedestrians using the user's footfall signature. The system was trained with a small number of 15 persons and was able to locate and identify more than 90% of the trained persons correctly. Smart Floor has its application area in a smart home environment, where only few different users need to be tracked and identified.

5 CONCLUSION

With this article, we have given a recent overview of the topic of indoor positioning with a strong focus on the task of positioning. Positioning is very complex inside buildings and applications range from navigation over production control to activity inference and business process automation.

Obviously, a position is meaningless without a sense of position, which is typically given by an indoor map. The topics of map creation, map modeling, and navigation semantics have been described very coarsely. A good source of information is Becker and Dürr (2005), which explains very clearly why indoor maps have to be different from outdoor maps and what they have to provide. We expect that standardization will make the generation and exchange of indoor maps favorable soon. Unfortunately, at this point in time, there is neither a standard nor a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

common sense how navigational information for indoor environments can be modeled.

RELATED ARTICLES

Interfaces between Sensors and ECUs
Applications of radio wave technologies to vehicles
Applications of image recognition technologies to vehicles
Car navigation
Telecommunications
Active safety, pre-collision safety and other safety products (millimeter wave, image recognition, laser)
Evolution and Future Trends
Cellular Mobile Networks
Technologies—Communication: Wireless LAN-based Vehicular Communication
Technologies—Communication: Broadcast
Technologies—Positioning: GNSS
Technologies—Positioning: Optical positioning
Data fusion

GLOSSARY

Angle of arrival	A method for determining the angle between two entities in which the angle of an incoming signal is measured, for example, by an antenna array
Angle of emission	A method for determining the angle between two entities in which the angle of an outgoing signal is coded into the signal and can be decoded by the receiver
Dead reckoning	A method for position determination based on a previously known position and measuring the change in position, for example, by measuring speed and direction of movement
Infrastructure-based positioning	Describes positioning systems that measure signals from and computes the position of a target without the need of a communication channel
Proximity detection	A method for localization based on the finite propagation distance of signals (i.e., the distance at which a signal still can be recognized), where the position can be assumed to be near the known position of the sender of the signal
Roundtrip time of flight	A method for determining the distance between an entity and a target that relies on the total traveled time of a signal whose propagation speed is known from the entity to the target and back
Scene analysis	In the context of indoor positioning, describes a method of localization by matching received signal pattern with known pattern, where the position of occurrence is known
Simultaneous localization and mapping	Stands for a technique where no initial map or reference data is available for positioning but is created on the fly and localization is performed with respect to the already generated incomplete map data
Terminal-assisted positioning	Describes positioning systems in which the mobile terminal measures signals and sends the gathered information to some fixed infrastructure, where its position is computed
Terminal-based positioning	Describes positioning systems in which a mobile terminal measures all signals and computes the position of itself without the need of a communication channel to any infrastructure
Time difference of arrival	A method of determining the distance between entities that relies on the time difference between two signals sent at the same time at different locations to identify the location of a mobile entity
Time of arrival	A method of determining the distance between entities that relies on the time of arrival of a signal to identify the time of flight of a specific signal whose propagation speed is known
Triangulation	Stands for localization techniques that make use of triangular geometry including lateration, angulation, and their combinations

REFERENCES

Adams, M.D. and Kerstens, A. (1998) Tracking naturally occurring indoor features in 2-D and 3-D with lidar range/amplitude data *The International Journal of Robotics Research*, **17** (9), 907–923.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Arulampalam, M.S., Maskell, S., Gordon, N., and Clapp, T. (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, **50** (2), 174–188.

Bahl P. and Padmanabhan V.N. (2000) Radar: an in-building RF-based user location and tracking system. *IEEE Inforcom 2000*, vol. 2, 775–784.

Becker, C. and Dürr, F. (2005) On location models for ubiquitous computing. *Journal Personal and Ubiquitous Computing*, **9** (1), 20–31.

Cobb, H.S. (1997) GPS pseudolites: theory, design, and applications. Ph.D. Thesis. Stanford University.

Hightower, J. and Boriello, G. (2001) Location systems for ubiquitous computing. *Computer*, **34** (8), 57–66.

Kalman, R.E. (1960) A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering*, **82** (D), 35–45.

Liu, H., Darabi, H., Banerjee, P., and Liu, J. (2007) Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, **37** (6), 1067–1080.

Lowe, D.G. (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, **60** (2), 91–110.

Mautz, R. and Tilch, S. (2011) Optical Indoor Positioning Systems. *Proceedings of the 2011 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Guimarães, Portugal.

Ocana, M., Bergasa L.M., Sotelo, M.A., *et al*. (2005) Indoor Robot Localization System Using WiFi Signal Measure and Minimizing Calibration Effort. *Proceedings of the IEEE International Symposium an Industrial Electronics*, Dubrovnik, Croatia, pp. 1545–1550.

Orr, R.J. and Abowd, G.D. (2000) The Smart Floor: A Mechanism for Natural User Identification and Tracking. *Proceedings of the 2000 Conference on Human Factors in Computing Systems*, The Hague, The Netherlands, pp. 275–276.

Rizos C., Roberts, G., Barnes, J., and Gambale, N. (2010) Locata: A New High Accuracy Indoor Positioning System. *Proceedings of the 2010 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Zürich, Switzerland.

Steggles, P. and Gschwind, S. (2005) The Ubisense Smart Place Platform. *Adjunct Proceedings of the Third International Conference on Pervasive Computing*, vol. 191, pp. 73–76.

Tarzia, S.P., Dinda, P.A., Dick, R.P., and Memik, G. (2011) Indoor Localization without Infrastructure Using the Acoustic Background Spectrum. *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, Bethesda, Maryland, USA, pp. 155–168.

Travis, W., Simmons, A.T., and Bevly, D.M. (2005) Corridor Navigation with LiDAR/INS Kalman Filter Solution. *Proceedings of Intelligent Vehicles Symposium*, pp. 343–348.

Want, R., Hopper, A., Falcão, V., and Gibbons, J. (1992) The active badge location system. *ACM Transactions on Information Systems*, **10** (1), 91–102.

Ward, A., Jones, A., and Hopper, A. (1997) A new location technique for the active office. *IEEE Personal Communications*, **4** (5), 42–47.

Werner M., Kessel M., and Marouane, C. (2011*)* Indoor Positioning Using Smartphone Camera. *Proceedings of the 2011 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Guimarães, Portugal.

Woodman, O. and Harle, R. (2008) Pedestrian Localization for Indoor Environments. *Proceedings of Tenth International Conference on Ubiquitous Computing*, Seoul, South Korea, pp. 114–123.

Zinnen, A., Wojek, C., and Schiele, B. (2009) Multi Activity Recognition Based on Bodymodel-Derived Primitives. *Proceedings of the 4th International Symposium on Location and Context Awareness (LoCA)*, Tokyo, Japan, pp. 1–18.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto174
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In-Vehicle Sensors

Frank Niewels, Steffen Knoop, Rüdiger Jordan, and Thomas Michalke

Robert Bosch GmbH, Stuttgart, Germany

1 Introduction 1
2 Vehicle Dynamics Sensors 1
3 Environment Sensors 5
Glossary 24
Endnote 26
References 26
Further Readings 27

1 INTRODUCTION

In the last decades, a growing number of vehicle systems got electrified. By this, they changed from originally mechanical systems to the so-called mechatronical systems. This process still holds on. A major part in this process is played by sensors. Their task is to transform a physical or chemical quantity into an electrical quantity, a control unit can calculate with. Figure 1 gives an overview over a large variety of sensors that exist already nowadays in series vehicles (Reif, 2012).

All these sensors will surely have a relevance from the telematics point of view. This can be engine sensors for diagnosis purposes, temperature or rain sensors for weather or road condition purposes, and lots of others.

As it is not possible to mention all of them, this chapter focuses on vehicle dynamics sensors on the one hand (Section 2) and environment sensors on the other hand (Section 3).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 VEHICLE DYNAMICS SENSORS

2.1 Introduction

Hand in hand with the introduction of different electronic vehicle dynamics control systems into series passenger vehicles such as the antilock braking system (ABS) (1978), the traction control system (TCS) (1986), and the electronic stability control (ESC) or electronic stability program (ESP) (1995), a large variety of corresponding vehicle dynamics sensors entered series vehicles (Kost *et al*., 2004). Nowadays, they are a kind of quasi-standard in series vehicles even more than any other driver assistance system (DAS) sensor.

Their main purpose is to measure and describe the vehicle motion state in three-dimensional (3D) space (e.g., velocities, accelerations, and turn rates) or adjacent variables such as wheel speeds, steering angle, and braking pressure.

There is a large variety of physical measurement principles (e.g., inductive vs Hall-wheel speed sensors) and technological building forms (e.g., surface-micromechanical vs. bulk-silicon-micromechanical acceleration sensors). The focus of this section is not to give a detailed comparison of all these technologies or latest layouts, but to give an overview of the relevant sensors and physical quantities. The corresponding main measurement principles as well as the relevance for vehicle dynamics systems and future telematics use are described.

2.2 Inertial sensors

The most important ESP sensors that are based on inertial measurement principles are acceleration and rotation-rate sensors.

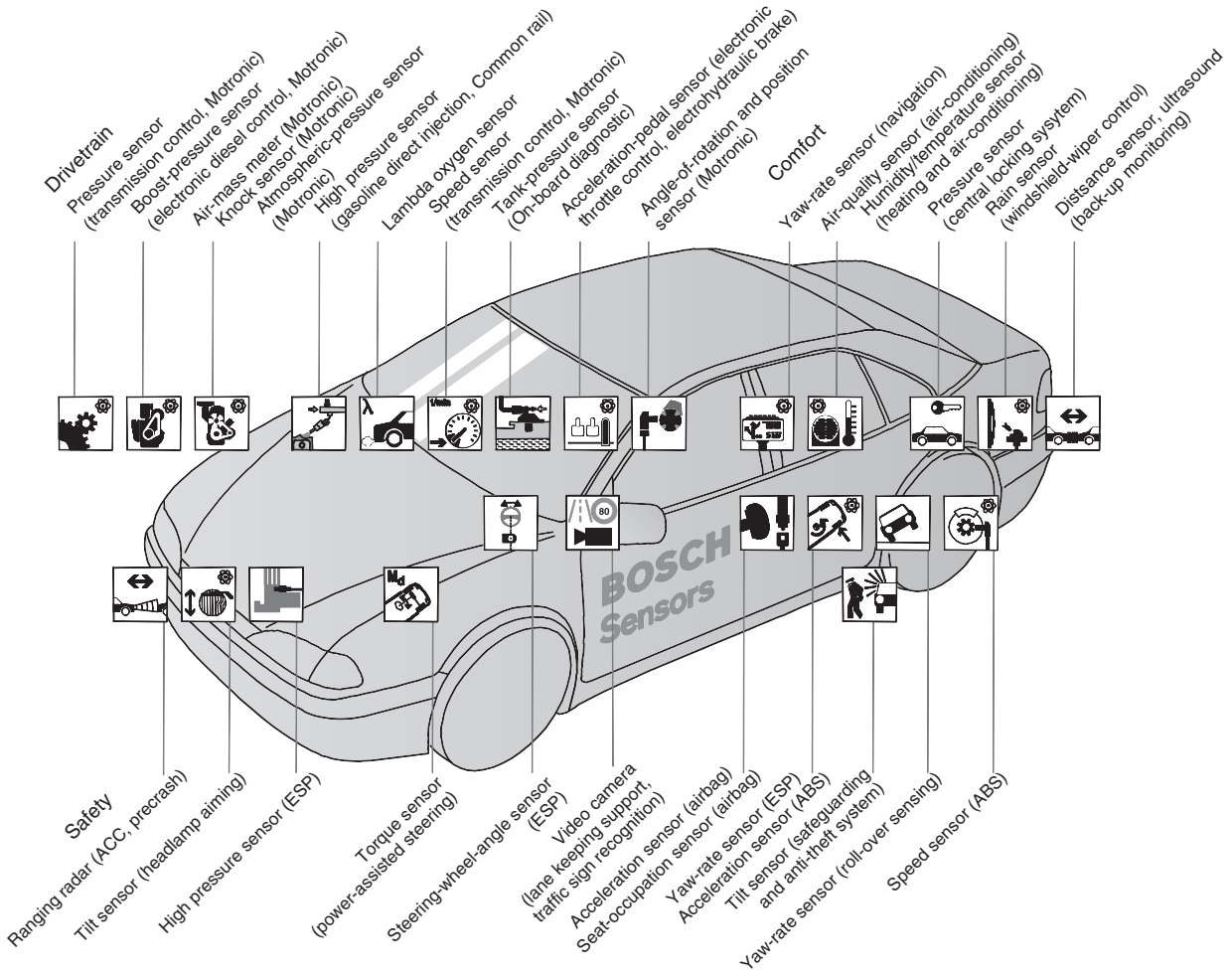


Figure 1. Data acquisition: in-vehicle sensors. (Reproduced by permission of Bosch GmbH.)

2.2.1 *Acceleration sensors*

The task of an acceleration sensor is to measure the acceleration a_B of a (vehicle-)body, it is attached to, in a predetermined measurement direction (MD). It does so, by measuring the force F on a seismic mass, which is an effect of the acceleration a_B or the acceleration a of the seismic mass, respectively. Figure 2 shows the measurement principle. Owing to the inertia of the seismic mass m, a force F results, if the (vehicle-)body is accelerated. This force is transferred to m by the two springs, which results in a corresponding elongation of the springs. Thereby, this elongation is a direct quantity for the acceleration a and with good approximation for the relevant acceleration a_B. It is measured electronically.

Figure 3 shows a surface-micromechanical realization of the principle described earlier (Kost *et al.*, 2004). The electronic measurement of the springs' elongation is

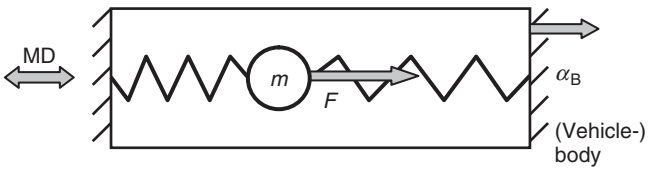


Figure 2. Measurement principle of an inertial acceleration sensor.

realized by measuring a variation of the resulting capacities C_1, C_2, and C_M.

With respect to the dimensioning of the acceleration sensor, different sensing ranges are realizable. Roughly, these ranges are separated into

- low-g, for example, ESP sensors with $a \approx 1$ g;
- high-g, for example, airbag sensors $a \approx 35 \ldots 100$ g

with $g = 9.81$ m/s^2.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

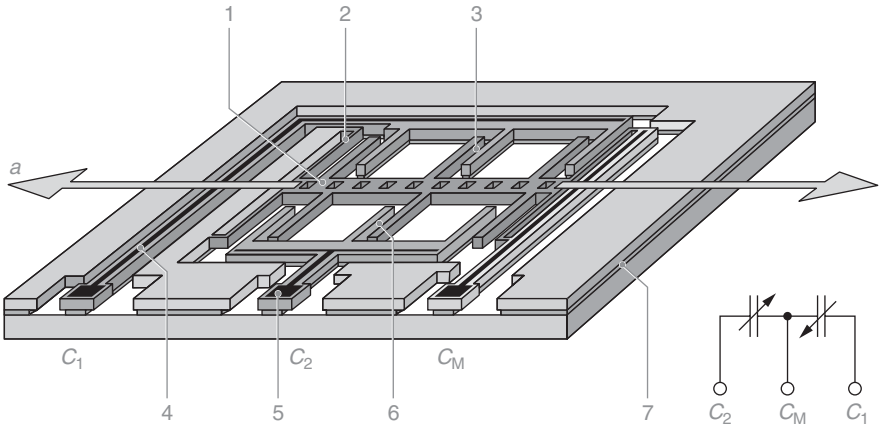


Figure 3. Surface-micromechanical realization of an acceleration sensor 1: seismic mass m; 2: spring; 3: capacity C_1; 4, 5: electric contact points; 6: capacity C_2; and 7: chassis linked to (vehicle) body. (Reproduced by permission of Bosch GmbH.)

Depending on the MD of the sensor in 3D space, an acceleration sensor measures also parts of the gravitational acceleration. For example, a sensor, mounted in z-direction (Figure 6) in a vehicle that stands still, shows $g = 9.81$ m/s^2.

2.2.2 *Rotation-rate sensors*

The task of a rotation-rate sensor is to measure the rotation rate Ω of a (vehicle-)body, it is attached to, with respect to a predetermined rotation axis (RA) of this body.

Figure 4 shows the measurement principle: an acceleration sensor element, similar to the one described earlier, is stimulated to small oscillations (OSC) relative to the (vehicle-)body, it is attached to, and vertical to its MD. If a rotation Ω around RA occurs, the seismic mass m is influenced by a Coriolis force F_C, because of the radial oscillation OSC. This force results once more in an elongation of the depicted springs, which can be measured electronically (see above). Finally, the Coriolis force—and thereby the elongation—depends only on the oscillation OSC, which is known, and the rotation rate Ω, which should be measured.

Figure 5 shows a micromechanical realization of the principle described earlier (Kost *et al.*, 2004). It contains two acceleration sensor elements, which are oscillating on

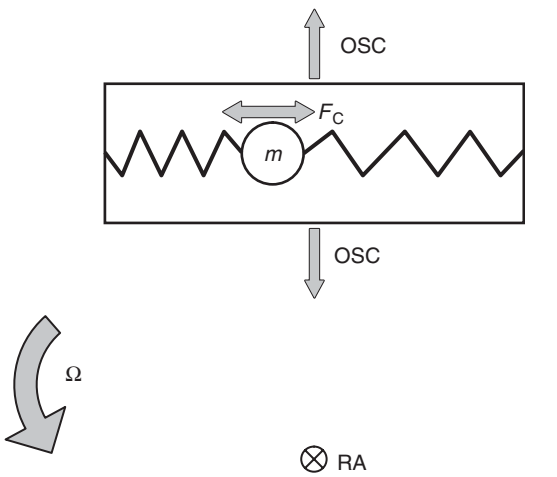


Figure 4. Measurement principle of a rotation-rate sensor.

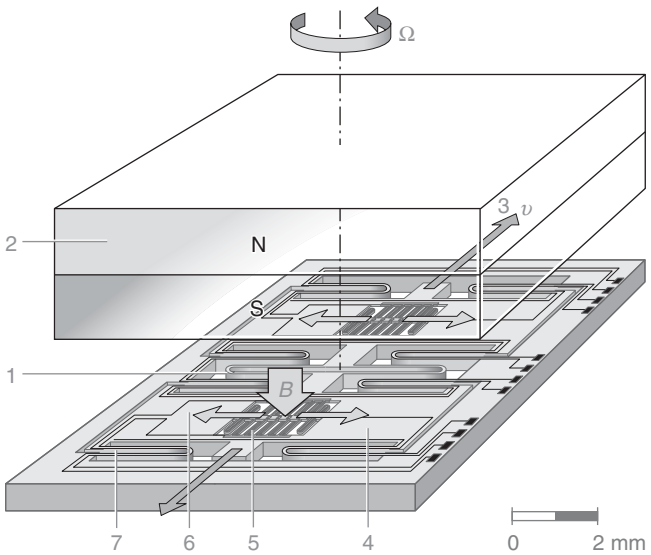


Figure 5. Micromechanical realization of a rotation-rate sensor. 1 and 7: springs influencing the stimulated oscillation OSC; 2: permanent magnet creating a magnetic field B; 3: direction of stimulated oscillation OSC; 4 and 5: acceleration sensor element, detecting the Coriolis force F_C; and 6: direction of Coriolis force F_C. (Reproduced by permission of Bosch GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

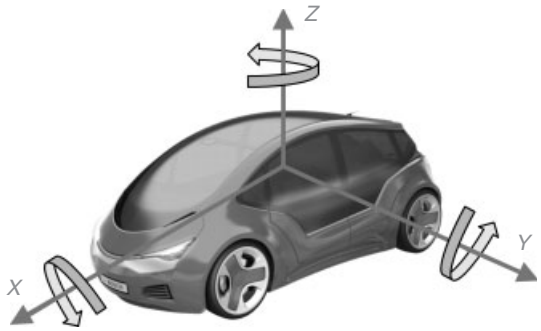


Figure 6. Vehicle attached coordinate system in 3D space. Reproduced by permission of Bosch GmbH.)

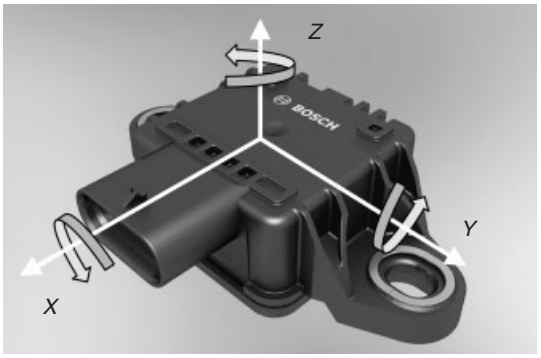


Figure 7. Sensor cluster with integrated inertial sensors for all 6 DOF of the vehicle body (see Figure 6, also). (Photo: Bosh. Reproduced by permission of Bosch GmbH.)

different sides of the RA. The oscillation is stimulated by an electric current. Owing to the magnetic field B, this creates a Lorentz force and thereby makes the relevant sensor parts move, that is, oscillate. The magnetic field B is created by the poles N and S of a permanent magnet.

2.2.3 *Measurement directions and clustering of sensors*

Figure 6 depicts the six degrees of freedom (DOF) of a rigid body—representing a vehicle—in 3D space. Depending on the mounting direction of the sensors relative to the vehicle, the sensors described earlier can measure these movements: acceleration in longitudinal (x), lateral (y), and vertical (z) directions and roll-, pitch-, and yaw-rates of the vehicle around these axes.

In principle, for each of these measurements, a separate sensor element is required. Nevertheless, several of these elements can usually be combined to one sensor cluster.

Common ESP systems come along with at least a lateral acceleration sensor and a yaw-rate sensor. Suspension control systems and modern lightening systems often make use of additional vertical acceleration sensors or pitch- and roll-rate sensors. If you have even more than this, for example a 5D configuration, very precise velocity and turn-angle information plus additional quatities can be computed from this (Dissanayake *et al*., 2001). With a full 6D sensor (Figure 7), this can be done even without using any vehicle model, only using the kinematic and kinetic differential equations of a rigid body (Reim *et al*., 2008), well known, for instance, from navigation tasks in aviation.

2.3 Additional ESP sensors

Besides the inertial sensors described earlier, there are some other sensors, important for ESP systems.

2.3.1 *Wheel speed sensors*

As all forces from the road are transferred to the vehicle via the wheels. The wheel speed sensors are one the most important sensors for vehicle dynamics control systems right from the beginning with ABS. Originally, they are for measuring the rotational speed of each of the four wheels of a vehicle. From this, other important quatities can be computed, such as the wheel (rotational) acceleration, the speed of the overall vehicle, each wheel's slip, and even additional quantities such as a loss of tire pressure or—primarily in braking or accelerating cases—the friction coefficient.

In early ABS systems, wheel speed sensors were mainly inductive sensors, the so-called passive wheel speed sensors. Owing to demands concerning the accuracy and the minimum measurable wheel speed when introducing ESP, this changed to active wheel speed sensors. They use either magnetoresistance ICs (integrated circuits), that is, elements that change their electrical resistance depending on a magnetic field, or Hall ICs. A Hall element lies over an impulse wheel either permanently magnetized (Figure 8) or made of steel with an additional permanent magnet (not shown here). This impulse wheel is mounted on the vehicle's wheel, whereas the Hall IC is mounted to the vehicle's chassis. When turning the wheel—and thereby the impulse wheel—the Hall element detects the permanent variation of the magnetic field using the Hall effect and can derive the turning speed from this (Kost *et al*., 2004).

2.3.2 *Steering-angle sensors*

The purpose of a steering-angle sensor is to measure the angle of the steering wheel, usually in an interval from

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

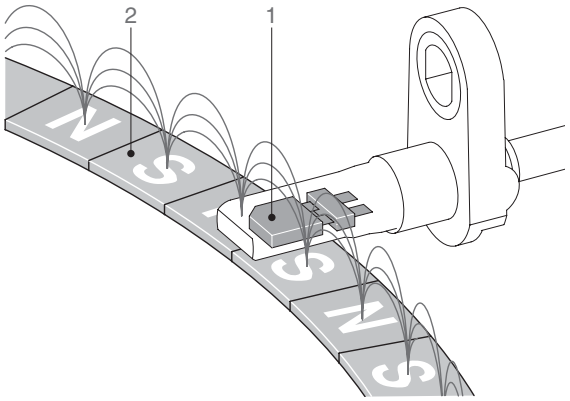


Figure 8. Active wheel speed sensor with a Hall IC and a magnetic multipole impulse wheel. 1: Hall sensor and 2: impulse wheel permanently magnetized. (Reproduced by permission of Bosch GmbH.)

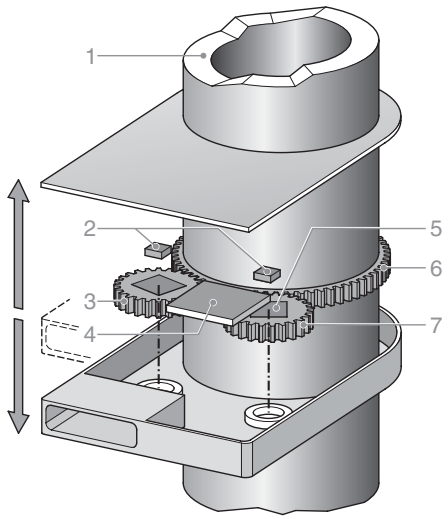


Figure 9. Steering-angle sensor. 1: steering column; 2: two magnetoresistance sensors, detecting the position of 5 and thereby of 3; 3 and 7: smaller cogwheels, with different number of cogs; 4: electronics; 5: permanent magnet; and 6: larger cogwheel. (Reproduced by permission of Bosch GmbH.)

-720° to $+720^\circ$. This indicates the intended driver's course to the system. Figure 9 shows the usual measurement principle (Kost *et al*., 2004): a cogwheel is mounted on the steering column. This larger cogwheel drives two smaller cogwheels. These smaller cogwheels have a different number of cogs. By this, it is possible to nonambiguously derive the position of the larger cogwheel (and thereby of the steering column) in an interval from -720° to $+720^\circ$ when knowing the combination of the positions of the two smaller cogwheels only in a 360° interval. These two positions are measured directly by two magnetoresistance sensors (Section 2.3.1).

2.3.3 Brake sensors

As the (wheel individual) braking system is the main actuator of an ESP system, there are different sensors measuring the state of the braking system. These are mainly sensors

- measuring instantly the actuation of the brake pedal by the driver, for example, the brake light switch;
- measuring the internal pressure state of the hydraulic system.

The latter—the use of pressure sensors—can be helpful at different positions in the hydraulic system. Basic ESP systems use a pressure sensor to measure the pressure of the brake master cylinder. To enlarge comfort or functionality, additional pressure sensors could be used, for example, to measure the pressure of each brake circuit individually.

2.4 Applications overview

Although Section 2 deals with sensors for vehicle dynamics, there are a lot of applications arising from the telematics point of view. In the following, there are given only some short hints, as there is a special chapter for this purpose in this encyclopedia.

Detection of

- friction coefficient;
- road surface conditions;
- abrupt braking maneuvers;
- accident of ego-vehicle;
- traffic jams.

Support of

- mapping;
- navigation by odometry.

3 ENVIRONMENT SENSORS

3.1 Introduction

Environment sensors are used to detect objects in the surrounding of the vehicle. The information is used by DASs to increase comfort and safety for the driver. In Figure 10, some of the most popular sensor types are displayed. Their measurement principles are described in Sections 3.2. to 3.4.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

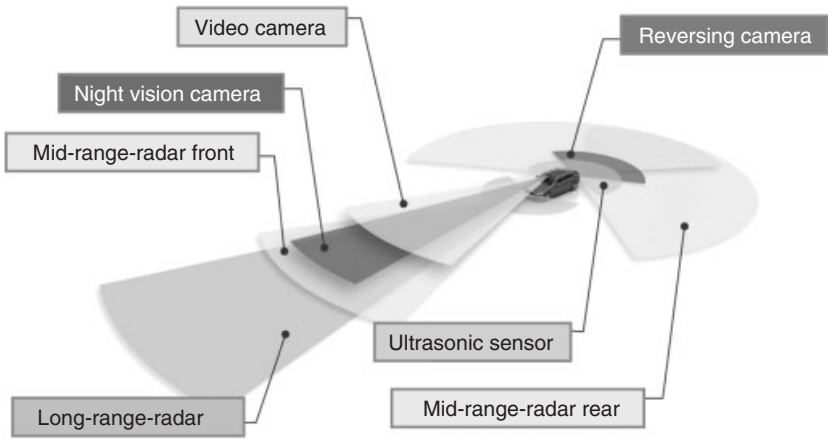


Figure 10. Environment sensors. (Photo: Bosh. Reproduced by permission of Bosch GmbH.)

3.2 Ultrasonic sensing

3.2.1 Introduction

Ultrasonic sensing, often subsumed with the acronym SONAR (sound navigation and ranging), is a measurement technique that uses sound propagation to measure objects and distances. A widely known application field for SONAR is underwater obstacle measurement, for example, for marine navigation. Animals such as whales use related techniques for communication over large distances with water as sound propagation medium. Bats use the SONAR principle even for precise navigation and for hunting their prey.

SONAR systems can be categorized in active and passive systems. Passive systems analyze sound signals that are generated by the environment or other objects. Active systems generate a sound signal, for example, a sound pulse, measure the reflected echo(es), and thus extract information about the environment.

Automotive DASs use the active SONAR principle for measuring distances to obstacles around the host vehicle. These systems help the driver especially in low speed maneuvers, such as parking, for example, by signaling any obstacles and remaining distance in the driving direction. Typically, ultrasonic sensors are applied to detect objects in the near field of the host vehicle. Even fully automated parking maneuvers can be conducted based on state-of-the-art ultrasonic sensors, if the vehicle allows control of steering angle, throttle, and brake.

3.2.2 Fundamentals: measurement principle

Active ultrasonic sensors in the vehicle generate a sound pulse. Objects in the measurement field reflect this sound wave. The reflected wave can be measured, and the signal delay between generated and reflected pulses is used for computation of the distance between sensor and the object. Figure 11 shows a sketch of this setup.

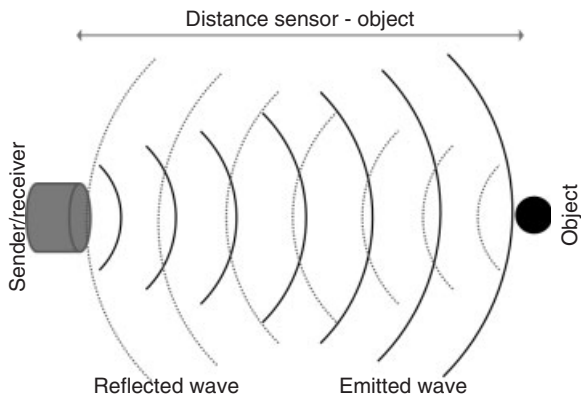


Figure 11. The generated sound pulse of the active ultrasonic sensor is marked in black. The wave is reflected by an object, and the reflected wave (light gray, dotted line) can be measured by the sensor.

Generally, the reflected sound signal contains different kinds of information about the target. This is obviously the radial distance of the object, which is coded in the signal delay. In addition, the reflected signal is affected by relative velocity between sensor and object as well as surface reflectivity of the target.

As the wavelength of the ultrasonic signal exceeds typically the roughness of objects (walls, other vehicles, etc.) in the close environment of the host vehicle, surrounding objects act as a mirror for ultrasonic signals. Thus, incident and reflection angles are equal, and only surfaces

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

perpendicular to an emitted pulse reflect the signal toward the sender.

3.2.2.1 Range measurement. Common ultrasonic sensors for automotive applications use only the distance information, as the signal is digitized directly in the receiver with a delay-dependent threshold.

With the information of the delay t between sending and receiving the sound pulse, the distance x between the sensor and the object is computed as follows:

$$x = \frac{c \cdot t}{2} \tag{1}$$

where c is the speed of sound. For the speed of sound, the following approximation can be used, which models the dependency from the temperature of the medium (Linder, 1999):

$$c = 331.5 + 0.6 \cdot \theta \quad (^{\circ}) \tag{2}$$

For a standard temperature of 20°C, this results in $c = 343.5$ m/s.

3.2.2.2 Trilateration. As mentioned, the delay between signal generation and receiving of the reflected response in the same sensor enables computation of the radial distance of the object. Together with the sensor opening angle, this yields a segment of a circle as possible object position. For parking assistance applications, it is necessary to achieve higher precision especially in the lateral dimension and around the vehicle corners. Measurement errors are decreased by using multiple sensors and the trilateration principle. Therefore, the detection areas of adjacent sensors overlap. An object located in the detection area of two sensors is measured by both sensors independently (direct echo). In addition, the reflected signal of one sensor sound pulse is measured in adjacent sensors (cross echo).

With at least two measurements of the same target in different sensors, the location of the target in the sensor plane can be computed geometrically. This yields, for distance a between target object and vehicle (Figure 12):

$$a = \sqrt{c^2 - \frac{(d^2 + c^2 - b^2)^2}{4d^2}} \tag{3}$$

As can be seen from Figure 12, the trilateration principle relies on the relative position of the ultrasonic sensors within the vehicle, which must be known.

An important assumption for trilateration to be valid is that both sensors measure the identical spot on the target, that is, all measured echoes originate from the same location on the target. The accuracy of the result depends on the validity of this approximation. In the example of Figure 12, the reflection centers of both sensors on the target object diverge, which results in a small error for the estimation of distance a.

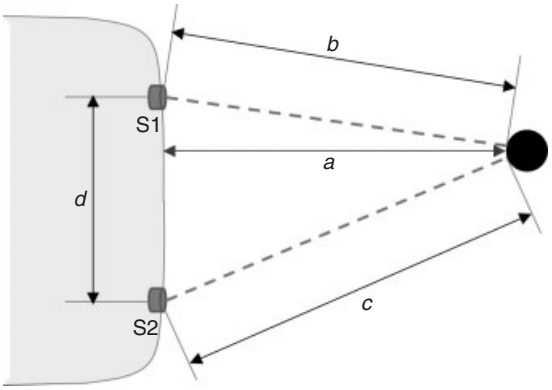


Figure 12. Trilateration: the required distance a between a vehicle and an object can be computed by combining several sensors. Measurements in different sensors.

3.2.2.3 Doppler. In addition to range measurement, the Doppler shift can be used to also measure the relative velocity between sensor and target (Raj *et al*., 2012). In static environments, the frequency is identical between emitted and received signals. In dynamic environments, when sensor and/or target object is moving, the signal frequency is shifted by the relative velocity between sensor and target. The Doppler shift caused by the reflection is approximately

$$f_{\mathrm{d}} = \frac{2vf}{c - v} \tag{4}$$

with f_{d} as observed Doppler shift, v the radial component of the target velocity, f the emitted frequency, and c as speed of sound.

3.2.3 Sensor design and measurement

Ultrasonic sensors for automotive applications consist typically of an electroacoustic transducer, a controller, and the body, including mounting and connector (Figure 13). The acoustic part consists of a cup-shaped alloy body, with the flat floor as the active area. Inside, a piezoceramic element is used to stimulate oscillation of the alloy transducer at its resonant frequency, and to measure oscillation induced by reflected echoes. The connection to the control unit comprises power supply and one signal line that is used bidirectionally for trigger and measurement.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

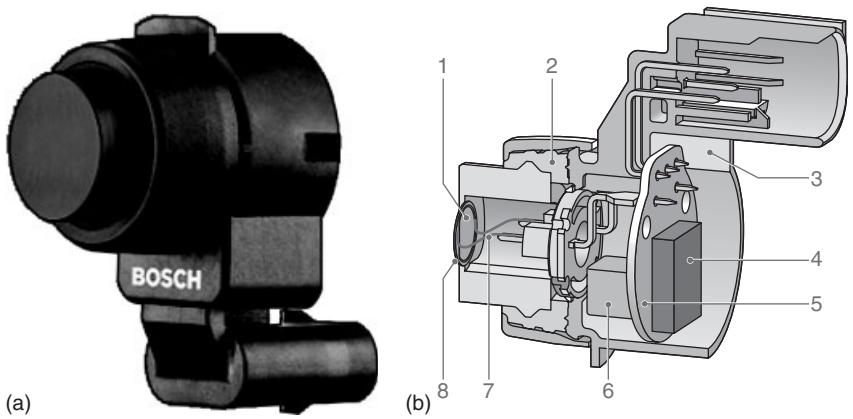


Figure 13. (a) Picture of a Bosch ultrasonic sensor fourth generation. (b) Cutaway view of the sensor (Reif and Knoll, 2012). 1: Piezoceramics, 2: decoupling, 3: housing, 4: ASIC, 5: electronics, 6: transducer, 7: bonding wire, and 8: diaphragm. (Reproduced by permission of Bosch GmbH.)

3.2.3.1 Transducer principle. Ultrasonic sensors typically work at frequencies of 40–50 kHz. This has proven to be a good compromise between damping of acoustic pressure, which increases with higher frequencies, and signal noise ratio (SNR), which is reduced by higher background noise at lower frequencies. With

$$c = \lambda \cdot f \tag{5}$$

this results in a wavelength of approximately 7 mm.

The transducer acts as both an emitter and a receiver of acoustic pulses. The diaphragm is, therefore, stimulated to emit an acoustic pulse, with a duration of 300–600 μs. After emitting the signal, the sensor acts as a microphone and is stimulated by the reflected ultrasonic pulse, which is again transformed into a digital signal that depicts the signal delay. After sending the pulse, it takes up to 900 μs for the mechanical oscillation to decay. During this period, it is not possible to discriminate a reflected pulse from the stimulated oscillation. This results in a minimum detection range 20–30 cm. To further reduce the decay time, it is possible to use active damping techniques. These are currently not used in automotive sensors.

Figure 14 shows signal generation and reception with two adjacent sensors. The pulse is generated in sensor 1. The reflected signal is received first in the emitting sensor, then in a second sensor, which is in listen-only mode. With both signal delays, it is possible to determine the location of the reflecting object within the sensor plane using the trilateration principle.

3.2.3.2 Detection range. The maximum detection range is limited on the one hand by damping effects and on the other hand by interference of different reflections. There

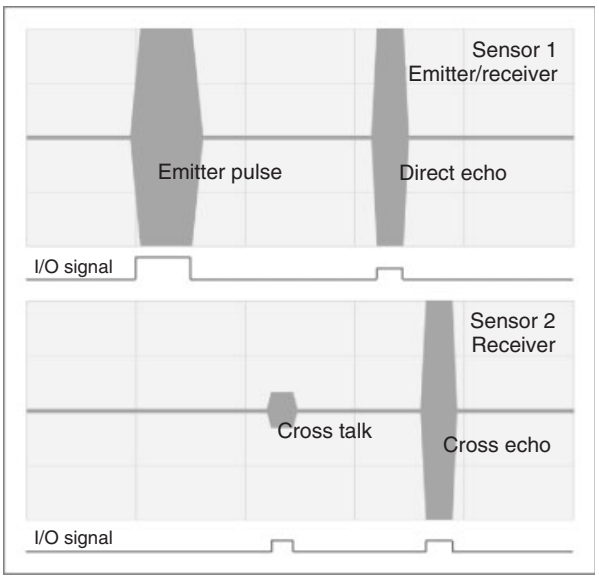


Figure 14. Pulse generation and reception of reflected signal in emitting and one adjacent sensor.

are three main damping effects: atmospheric damping, geometric damping, as given by the wave type, and, during reflection, the damping of the reflecting object. Signal reflections interfere with reflections of the ground, typically the street, which also reduces the SNR.

The detection range of ultrasonic sensors is limited to 0.25–2.5 m; new systems for extended parking assistance have a maximum range up to 5 m. To minimize ground reflections in the received echo while still detecting a maximum of obstacles, the detection characteristic is asymmetrically formed. The vertical opening angle covers ±30°, while the horizontal opening angle sums up to ±60°.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

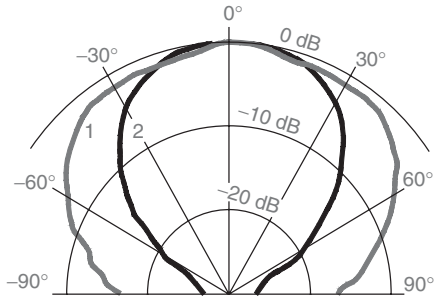


Figure 15. Antenna diagram of ultrasonic sensor, with horizontal (1) and vertical (2) intensity curves, see (Reif and Knoll, 2012). (Reproduced by permission of Bosch GmbH.)

Figure 15 shows a typical antenna diagram of an ultrasonic sensor in horizontal and vertical axes.

3.2.3.3 System design. An ultrasonic assistance system consists of the sensors mounted around the vehicle in a plane parallel to ground surface and one central processing unit (CPU) that coordinates and evaluates the sensor outputs. Figure 16 shows the block diagram of the sensor and the processing unit.

The processing unit triggers pulse generation in the sensors. In listen mode, the sensors return received echo pulses to the processing unit, which then computes the distance to measured objects from all returned sensor signals. Typically, each sensor's pulse generation is triggered consecutively to avoid interference between different pulses.

3.2.4 Sensor configurations

An ultrasonic measurement system for parking assistance consists of a set of typically at least four sensors on one side of the vehicle, up to 12 sensors around the vehicle's contour. The sensors are mounted roughly in one plane, at a height of approximately 50 cm in the bumper.

The applied configuration for a certain vehicle depends on the applications implemented in the vehicle. Generally, the configurations can be categorized according to the sensor mounting position.

Most common configuration is a setup consisting of four sensors in the rear bumper, also called a *four-channel system* to detect and measure objects behind one's vehicle. Analogically, the sensors can be mounted in the front bumper to measure the distance to obstacles in front of the vehicle. Sensors on each side of the car typically serve for detecting parking spaces. Side sensors can also be used for detection of vehicles during lane changes, often referred to as *blind spot detection*. Therefore, side sensors are often configured to higher measurement distances.

3.2.5 Applications overview

The major domain of automotive applications for ultrasonic sensors is parking assistance, because of their comparatively low detection range and high field of view. The most common parking assistance system, often called *park pilot*, measures the distance to the nearest obstacle during parking maneuvers and signals the remaining distance to the driver via optical and acoustical feedbacks, often using a beeping sound with decreasing pulse repetition time while approaching the obstacle. Depending on the number and mounting position of the sensors, obstacles behind and/or in front of the vehicle are detected. These systems exist since 1991.

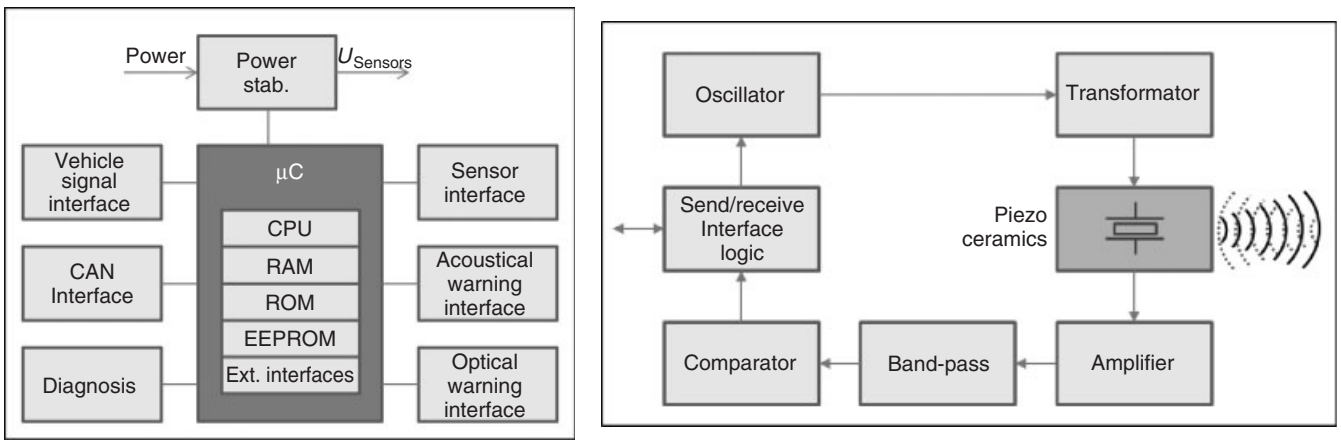


Figure 16. Block diagram of the central processing unit and of one ultrasonic sensor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In addition to the park pilot, other assistance systems exist. Sensors located at the vehicle side, often in the front fender, are used to detect and signal parking spaces. This is done by measuring the distance to side obstacles while tracking the host vehicle with respect to these objects based on its internal motion measurements. Thus, gaps between side obstacles are detected. Their length is estimated based on the host vehicle's motion and compared to the host vehicle's size. If the gap is large enough to serve as parking space, this is signaled toward the driver. This means that the driver always has to first pass a parking space at least with the sensor location (front fender, as mentioned) before the parking space can be signaled by the system. As parallel parking typically requires reversing, there is no critical system restriction.

Advanced parking assistance functions automatically compute parking trajectories from measured parking spaces. These trajectories can be applied in two different ways. "Park steer information" (PSI) systems inform the driver of the best parking trajectory, and the driver conducts parking action as advised by the system. During the maneuver, the trajectory needs to be recomputed continually to adapt to deviations. "Park steer control" (PSC) systems control the steering wheel according to the computed trajectory. Throttle and brake are typically operated by the driver, who still has full responsibility for the vehicle movements, even if the car almost automatically reaches its parking position.

3.2.6 Summary

Ultrasonic sensors provide a mechanism for measuring obstacles in the near field of the vehicle. This serves as a sensorial basis for DASs that support the driver in low speed maneuvers, especially parking.

An ultrasonic sound pulse is generated, and by measuring the time delay between pulse generation and its echo, the distance to reflecting objects can be computed.

By applying several sensors (e.g., four sensors in the rear bumper) and by evaluating the cross echo (Section 3.2.2), the measurement accuracy and resolution are increased. A CPU triggers and synchronizes the sensors, receives sensor measurements, and computes obstacle distances, which are then communicated to the driver.

Applications vary from simple acoustical distance indication to automatic parking. Future applications will even use ultrasonic side sensors at higher speeds for assisting the driver during lane changes, and by warning the driver of other vehicles in adjacent lanes.

3.3 Radar

3.3.1 Introduction

Originally developed for military use, radar (radio detection and ranging) sensors can be found in civilian vehicles since several years. The measuring principle is based on the transmission of an electromagnetic wave in the gigahertz band. The radar wave is reflected especially at metallic objects and can be detected by a radar receiver.

The distance of a target object can be determined by evaluation of the time delay between sending and receiving the radar signal. Through evaluation of the frequency shift caused by the Doppler effect, also the object's velocity can be determined. Furthermore, there exist several methods to determine the azimuth or elevation angle. By preprocessing the received radar signals with electronic circuits and a digital signal processor, the raw measurement values are generated. Typically, a radar measurement consists at least of the attributes radial distance d_r, relative speed v_r, azimuth angle α, and the received signal power P_R. The raw measurements are further processed by a tracking algorithm. This is done to reduce the measurement noise and to estimate signals that cannot be measured directly, such as the object's acceleration and signal uncertainties (e.g., signal variances).

Automotive long-range radar (LRR) sensors have a detection range of more than 200 m, which is currently not exceeded by any other type of comparable sensor. In automotive applications, radar sensors are particularly suitable for distance and speed control, as the required quantities are measured directly without any interpretation. Furthermore, collision warning or automatic emergency braking systems are often based on radar sensors (Figure 17).

3.3.2 Fundamentals of radar technology

3.3.2.1 Antenna gain. Usually, radar antennas are designed to send most of the radiated power toward a preferred direction. The antenna gain G describes the ratio of the maximum power density S_r of a directional antenna and the power density S_i of an (imaginary) isotropic radiator. The same principle is valid for receiving a signal. The calculation of the antenna gain is given by equation 6:

$$G = 10 \cdot \log_{10}\left(\frac{S_r}{S_i}\right) \tag{6}$$

In Figure 18, a basic antenna diagram is shown with one main and several side lobes. The half-value angle of the antenna is a measure to describe the width of the main lobe. It describes the angle area beside the direction of maximum radiation until a power drop of 3 dB.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 17. Automotive LRR sensor. (Reproduced by permission of Bosch GmbH.)

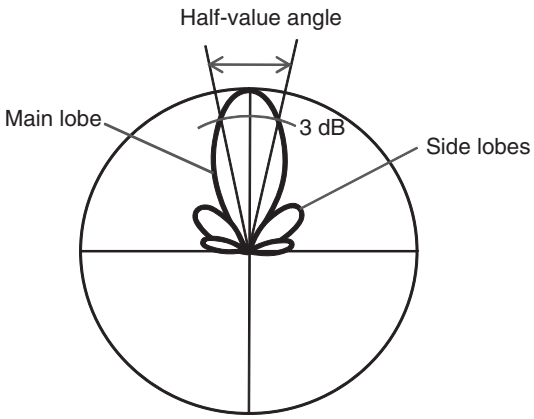


Figure 18. Basic antenna diagram.

3.3.2.2 Radar cross section. The radar cross section (RCS) describes the backscattering behavior of a single target, that is, how much of the sent-out power is reflected to the receiver. The used model is that the same signal power would be received if the specific target was replaced by a metal sphere of the given cross section. The measure is usually given in the unit square meter.

Type of reflector	Dimensions	RCS
Sphere		$\frac{\pi \cdot d^2}{4}$
Flat mirror		$4\pi \cdot \frac{(a \cdot b)^2}{\lambda^2}$
Dihedral corner reflector		$8\pi \cdot \frac{a^2 \cdot b^2}{\lambda^2}$
Triple-mirror		$\frac{4}{3}\pi \cdot \frac{a^4}{\lambda^2}$

Figure 19. RCS calculation of typical radar reflectors.

The RCS value σ is dependent on the geometry of the target object and the wavelength λ of the radar signal. In Figure 19, some examples of typical radar reflectors and their maximum RCS value are shown.

3.3.2.3 Radar equation. With the help of the radar equation, it can be calculated how much of the sent out radar power will be received after reflection. For a better understanding, the equation is derived step by step in the following. Figure 20 illustrates the mentioned variables.

If the radar transmitter has an antenna gain G and emits pulses with the power P_T, the power density at the distance r will be S_T:

$$S_T = \frac{P_T \cdot G}{4\pi \cdot r^2} \tag{7}$$

The reflected power P_B depends on the RCS σ and becomes accordingly:

$$P_B = \sigma \cdot S_T \tag{8}$$

Therefore, the power density at location of the receiving radar antenna is:

$$S_R = \frac{P_B}{4\pi \cdot r^2} \tag{9}$$

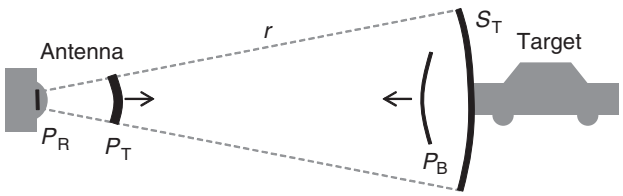


Figure 20. Derivation of the radar equation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

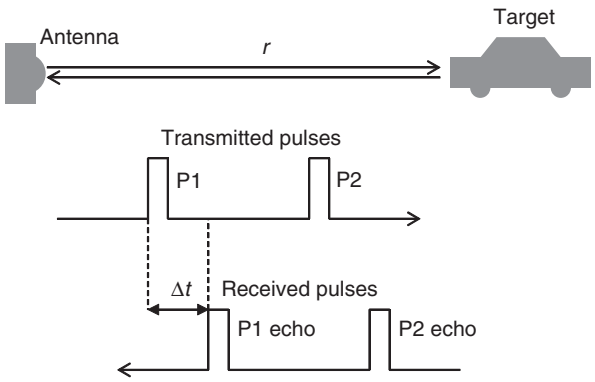


Figure 21. Distance measurement with pulse radar.

The antenna can absorb only a fraction of P_B, which is due to its limited extension. The effective antenna area A_{eff} depends on the antenna gain G and the wavelength λ:

$$A_{eff} = \frac{G \cdot \lambda^2}{4\pi} \tag{10}$$

Combining all the equations, we can derive the so-called radar equation 11 that denotes which quantities influence the received radar power:

$$P_R = S_R \cdot A_{eff} = \frac{P_T \cdot G^2 \cdot \lambda^2 \cdot \sigma}{(4\pi)^3 \cdot r^4} \tag{11}$$

3.3.3 Techniques for distance and velocity measurement

To measure the position and velocity of objects with a radar sensor, different techniques can be used, which vary in complexity, the achievable measurement accuracy, and the error rate. The following sections provide a brief overview of some well-known techniques.

3.3.3.1 Pulse-Doppler radar. The main idea of a Pulse-Doppler radar is to emit short signal pulses and to evaluate the time until the back-scattered signal is detected by the receiver (refer to Figure 21).

The time delay Δt is proportional to the distance to the reflecting object. Taking the velocity of light c, the distance r can be calculated directly:

$$r = \frac{c \cdot \Delta t}{2} \tag{12}$$

To allow the measurement of multiple objects, the received signal is sampled consecutively many times. Each sample corresponds to a particular distance, the so-called range gate.

To be able to separate objects close to each other, the duration of a pulse has to be short enough, so that two separate pulses are received (compare Figure 22). If the pulse duration would be chosen two long, the two pulses would be merged together and only one echo pulse would be received.

The distance separability Δr can be calculated as follows:

$$\Delta r = \frac{c \cdot \tau_p}{2} = \frac{c}{2 \cdot B} \tag{13}$$

where

$$\tau_p \cdot B \approx 1 \tag{14}$$

Here, c is the velocity of light, τ_p the pulse duration, and B the pulse bandwidth.

Furthermore, the time between two consecutive transmitted pulses has to be long enough that no reflections from objects positioned far away would be received in the next transmission cycle. With a given pulse repetition time T_R, the maximum distance r_a without range ambiguities is:

$$r_a = \frac{c \cdot T_R}{2} \tag{15}$$

If there is a large radar reflector at high distances, it cannot always be avoided that the reflected signal is received in the following transmission cycle. One method to exclude range ambiguities is to change the pulse repetition frequency. While true objects are measured at the same distance every time, ghost objects change the position in dependence of the pulse-repetition frequency.

The relative speed of the object is determined by evaluation of the Doppler shift. From moving objects, a slightly

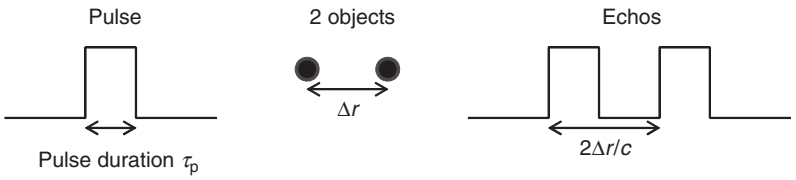


Figure 22. Distance separability.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

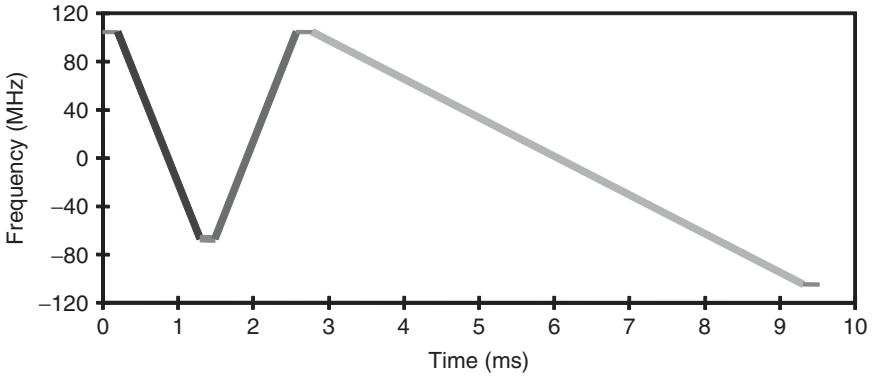


Figure 23. FMCW modulation.

changed frequency is received compared to the emitted signal. To detect this, a large number of successive samples of one range gate is collected and transformed into the frequency domain. By doing this, also the SNR is increased, as the received power of many individual pulses is integrated. The dominant frequency of the spectrum is proportional to the object's velocity.

A main advantage of the pulsed Doppler technique is that the calculation of distance and speed is made independently, which leads to a relatively low error rate. The disadvantage is, however, that sophisticated capabilities of signal processing are needed: the received signal has to be sampled at a very high frequency, and for each range gate, a frequency analysis must be performed.

3.3.3.2 Frequency-modulated continuous wave. The frequency-modulated continuous wave (FMCW) technique is characterized by a continuously transmitted radar wave, whose frequency is varied linearly over time (Figure 23). The radar wave is reflected by metallic objects and received again by the sensor. The time delay and the Doppler effect cause a frequency shift of the received signal compared to the currently emitted signal. Hence, the difference frequency is dependent on both the distance and the relative velocity of the object.

The received difference frequency f_r can be calculated from Equation 16:

$$f_r = 2\frac{s_T}{c} \cdot d_r + \frac{f_T}{c} \cdot v_r \tag{16}$$

Here, f_T is the mean transmission frequency, s_T the slope of the frequency ramp, c the speed of light, and d_r and v_r the distance and the relative velocity of the measured object, respectively. By solving Equation 16 to the relative velocity v_r and plotted over d_r, we obtain a straight line in the so-called dv-diagram (Figure 25). The slope of the emitted ramp corresponds to the slope of the line in the dv-diagram. To calculate the distance and velocity of an object, at least a second ramp with different slope has to be modulated. This way, two crossing lines are created in the dv-diagram. The intersection point describes the distance and relative velocity of the measured object. In multitarget scenarios, intersections in the dv-diagram can occur that do not represent a real target. Therefore, at least three consecutive ramps with different slopes have to be used, so that a real object is described by an intersection point of three lines in the dv-diagram.

In Figure 23, the transmitted radar signal of a 3-ramp modulation is shown. In Figure 24, the received frequency spectrum corresponding to the lines in the dv-diagram of Figure 25 can be seen. As illustrated here, two real targets (numbers 1 and 2) and one ghost target (number

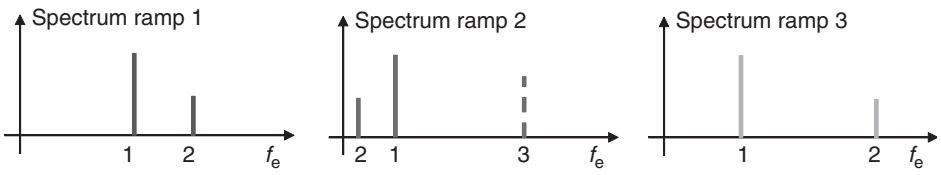


Figure 24. Frequency spectrum of received FMCW signal.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

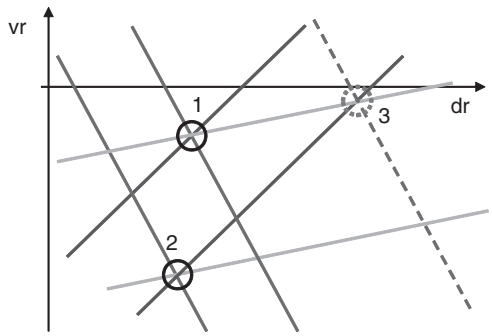


Figure 25. Frequency matching in dv-diagram.

3) can be determined. The ghost target arises from the fact that some detections incidentally generate a three-line intersection point in the dv-diagram. These wrong matches can be generated by detections of real objects or false detections because of noise or clutter (see dashed dark gray line). The probability is growing with increasing number of detection of the frequency spectrum. As the occurrence of wrong matches is almost random, ghost objects do not show a reasonable movement and thus can be refused by a tracking algorithm. Another countermeasure is the use of even more consecutive frequency ramps, which clearly lowers the probability of wrong matches.

3.3.3.3 Frequency-shift-keying. The frequency-shift keying (FSK) technique is related to the FMCW technique. Here also a continuous radar wave is transmitted and received simultaneously. In contrast to the FMCW technique, the transmission frequency remains constant for a certain time. Therefore, a frequency shift of the received signal is generated only by the Doppler effect. By performing a frequency analysis of the difference frequency between emitted and received signals, the velocities of all measured objects can be defined. To determine the corresponding distance of the objects, two blocks of slightly different frequency are emitted (Figure 26). The Doppler shift is almost the same (Figure 26) in each of the two blocks, so that a measured object is mapped to a peak at the same position in the frequency spectrum. However, the phasing of the two signals is different. From this phase shift between the two blocks, the distance of the measured object can be calculated.

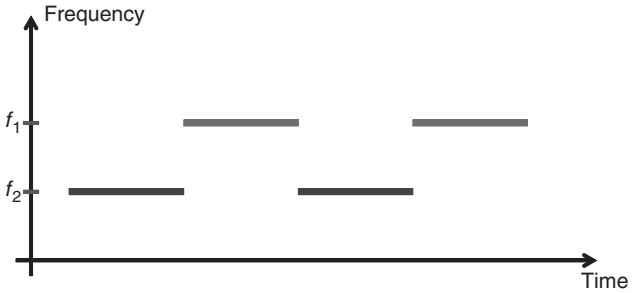


Figure 26. FSK modulation.

The hardware of an FSK radar can be realized very cost-effective, because the demands on the linearity of the transmission signal are lower compared to an FMCW radar. A disadvantage is, however, the fact that the accuracy of distance measurement is dependent on the SNR. Therefore, the distance can be determined less accurate if an object either has a low RCS or is located at a high distance. If two objects have the same relative speed, they cannot be separated by an FSK radar, because they share the same peak in the frequency spectrum. This is the case for all stationary objects. Therefore, a different radar technique (e.g., Pulse-Doppler or FMCW) should be used if the application needs to consider stationary objects.

3.3.4 Techniques for angle measurement

3.3.4.1 Monopulse. To determine the azimuth angle of an object with a single radar pulse or FMCW ramp, at least two partially overlapping radar beams are needed. In a monostatic[1] radar system, there are horizontally arranged antenna elements, which sent out, simultaneously, a coherent radar wave. Each element has its own reception channel. The angle of the measured object is calculated by comparing the received amplitude and phase information with a known antenna pattern (also called *antenna diagram*). The position in the antenna diagram, which fits best to the received signal, is most likely the azimuth angle of the measured object. In Figure 27, an example of such an antenna pattern is shown. Displayed are the amplitude values of a four-beam radar as a function of the horizontal angle.

The described beam-matching procedure is not applicable to separate two objects that are located at the same distance with same relative velocity. In this case, the determined angle would be erroneous. Nevertheless, there exist some more advanced methods to separate two objects in angle, even with a fixed-beam radar.

3.3.4.2 Mechanical scanning. The scanning principle is well known from air surveillance and sea traffic. To determine the azimuth angle of an object, a highly focused antenna is mechanically rotated or moved from side to side. Thus, the desired viewing area is scanned over time. This simple method allows a precise angle measurement in a large angular field of view.

As the angle sectors are sampled consecutively, the available measurement time for a single target object is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

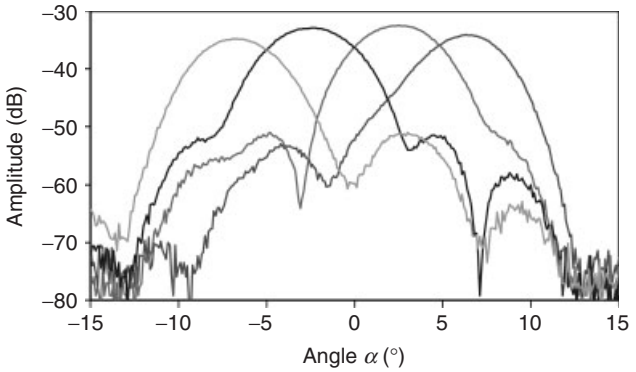


Figure 27. Antenna diagram of a fixed beam radar. (Reproduced from Bosch GmbH with permission.)

limited, which especially could have negative impacts on the accuracy of the velocity determination. Because the sensor installed in a vehicle is exposed to vibrations, it is challenging to guarantee the durability of the moving sensor parts.

3.3.4.3 Beam forming. An alternative to a mechanical scanning is a beam-forming method. The advantage is that no mechanical moving parts are required, although a focused radar beam can be moved over the viewing area within certain limits.

The basis of an electronic beam-forming radar is an antenna array consisting of a large number of antenna elements. If the antenna elements are supplied with signals of different amplitudes and phases in a systematic way (Figure 28), the center of the main lobe can be steered to different directions.

A disadvantage of this method is that the hardware of the sensor has to be equipped with complex and also expensive high-frequency components.

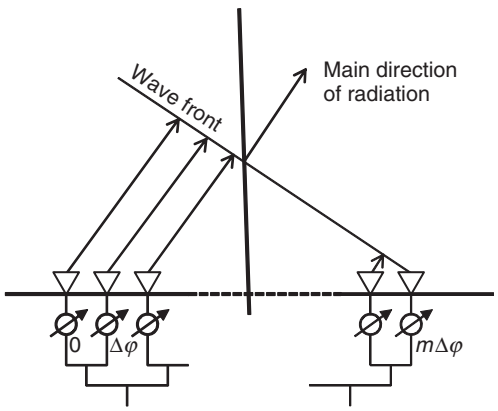


Figure 28. Electronic phase steering.

In contrast, a digital beamforming (DBF) radar supplies all antenna patches with an identical signal comparable to the monopulse method. The radar beam steering is not done physically but instead with the help of digital signal processing: after digitalization, the received signals are multiplied with complex weights. This way, a virtual radar beam can be steered to the desired direction. As the used algorithms are computationally complex, a powerful digital signal processor is needed, which increases the system costs.

3.3.5 Radar applications

Radar sensors are used for DASs since several years. One of the first applications that went into series production was the comfort system adaptive cruise control (ACC). It is an extension of the classic speed control. The vehicle automatically keeps the user-selected velocity as long as the road is clear. If a slower preceding vehicle appears, the velocity is reduced and the ego-vehicle follows in a constant distance. If the preceding vehicle disappears, for example, because of a lane change, the velocity is increased again to the initially chosen velocity. ACC takes over the task of accelerating and decelerating within comfort limits, and only the steering task remains to the driver. In situations where the area of comfortable driving has to be exeeded, for example, for a strong braking maneuver, the driver is acoustically informed by the system and he or she has to take over the longitudinal control of vehicle again. For ACC, usually a LRR sensor is used, which is capable of detecting objects at distances over 200 m.

A second class of radar applications concern safety systems that help to avoid accidents. As a first step, the driver is optically, acoustically, or haptically warned if there is a high risk of a collision with moving or stationary objects. If the driver does not react in time, a partial braking maneuver is started. As a last step, a fully automatic emergency brake is triggered in case there is no other driving maneuver left to avoid the collision. Even if the collision could not be avoided by the system, the reduction of velocity helps to mitigate the severity of the crash.

The lane change assistant (LCA) is a warning function to prevent collisions with vehicles on neighbor lanes while doing a lane change maneuver. Typically, two mid-range radar sensors observe the area behind and besides the own vehicle. The sensors are typically located behind the bumper at the left and right corners of the vehicle's rear end. The system is activated when the driver turns on the direction indicator. In case there is an approaching vehicle on the neighbor lane, a warning signal is giving to the driver, for example, a red flashing lights at the exterior mirror.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

As a last example, short-range radar sensors are used for the parking assistance. The sensors are mounted behind the front and rear bumpers and inform the driver about obstacles in the near surrounding of the vehicle, when the vehicle is maneuvered at low speeds. Usually, optical and acoustic signals are used to indicate the remaining distance to other objects.

3.3.6 Summary and outlook

Radar sensors are established in automotive applications since several years. Its main advantages are the precise and reliable measurement of objects' distance, relative velocity, and azimuth angle under all weather conditions. There exist several measurement techniques, which differ in complexity and performance. Automotive radar sensors can be produced in a compact housing and can be mounted invisible, which is an important aspect for many car manufacturers.

As the sensor costs continuously decrease, radar-based DASs become attractive to the customer even in medium-class and small cars. It can be assumed that the number of radar sensors a car is equipped with will increase in the future. Radar sensors can help to improve the comfort of driving and play an important role for reducing the number of fatalities in traffic accidents.

3.4 Video

3.4.1 Introduction

A very popular and cost-efficient sensor system for capturing the features of light is the (digital) video camera—hereafter, simply called *camera*. Especially, the technological progress in the field of semiconductors has helped introducing cameras for driver assistance. The major advantage of camera-based driver assistance is its functional multipurpose character. No other automotive sensor in mass production offers the information density and hence functional potential of video signals. With stereo video, a novel sensor technology has reached the required maturity for the automotive market. Stereo video shifts known sensory constraints in terms of accuracy and availability of traffic-relevant measurands. Owing to that, customers can expect numerous novel driver assistance functions for further improving comfort and safety in driving.

The remaining section is organized as follows: in Section 3.4.2, the video measurement principle is illustrated, after which in Section 3.4.3, the focus will be on the basic design of video sensors. Section 3.4.4 elaborates on typical installation locations for cameras in vehicles. Before summarizing the chapter, in Section 3.4.5, some telematics-related automotive applications for video cameras are introduced.

3.4.2 Fundamentals: measurement principle

As for all sensor devices, a nonelectric physical quantity is transformed into an electric signal in order to allow its application in a technical system. For cameras, the driving physical principle for the required transformation is the so-called photo effect. An inner and outer photo effect can be distinguished. The latter (also called *photoelectric effect*) designates the emission of electrons from a (semi) conductor when absorbing radiation (e.g., visible light). The former (also called *photoconductive effect*) is based on a change of conductibility of a semiconductor (caused by the creation of electron–hole pairs) that is proportional to the intensity of light. The inner photo effect is exploited in (digital) cameras.

It is known that light can be interpreted as consisting of inseparable basic elements—the so-called light quantum (or photon). The notion of photons gives a very intuitive understanding for the model of a major source for noise in images. By absorbing photons, each pixel of a camera imager (a semiconductor with a photosensitive surface) gathers electrons, whose amount is proportional to the level of radiation. The current of photons (the number of gathered electrons per time interval) changes when repeatedly capturing the same scene. The number of absorbed photons is a stochastic process. When determining the relative frequency of measured pixel gray values (these are proportional to the number of gathered electrons), a specific probability density function (PDF) results. The PDF can be modeled by the Poisson distribution, which is defined as follows (with λ as the current of photons and Δt as the exposure time of a measurement cycle) (Jaehne, 2005):

$$f(\lambda \cdot \Delta t) = \frac{\mathrm{e}^{-\lambda \cdot \Delta t}(\lambda \cdot \Delta t)^n}{n!} \text{ (with } \quad n \geq 0) \qquad (17)$$

Mean $\mu = \lambda \cdot \Delta t$ and variance $\sigma^2 = \lambda \cdot \Delta t$.

A typical imager pixel will gather more than 10,000 electrons per measuring cycle Δt, which in 68% of cases lead to a relative error σ/μ for the photon noise of $<1\%$. When analyzing σ/μ for the Poisson distribution, it can easily be derived that the relative influence of photon noise is growing with a decreasing exposure time Δt (Figure 29). Put differently and hence more intuitively, in a short exposure time Δt, a small amount of electrons is gathered. The resulting low voltage level has to be largely amplified, boosting the signal and the noise.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

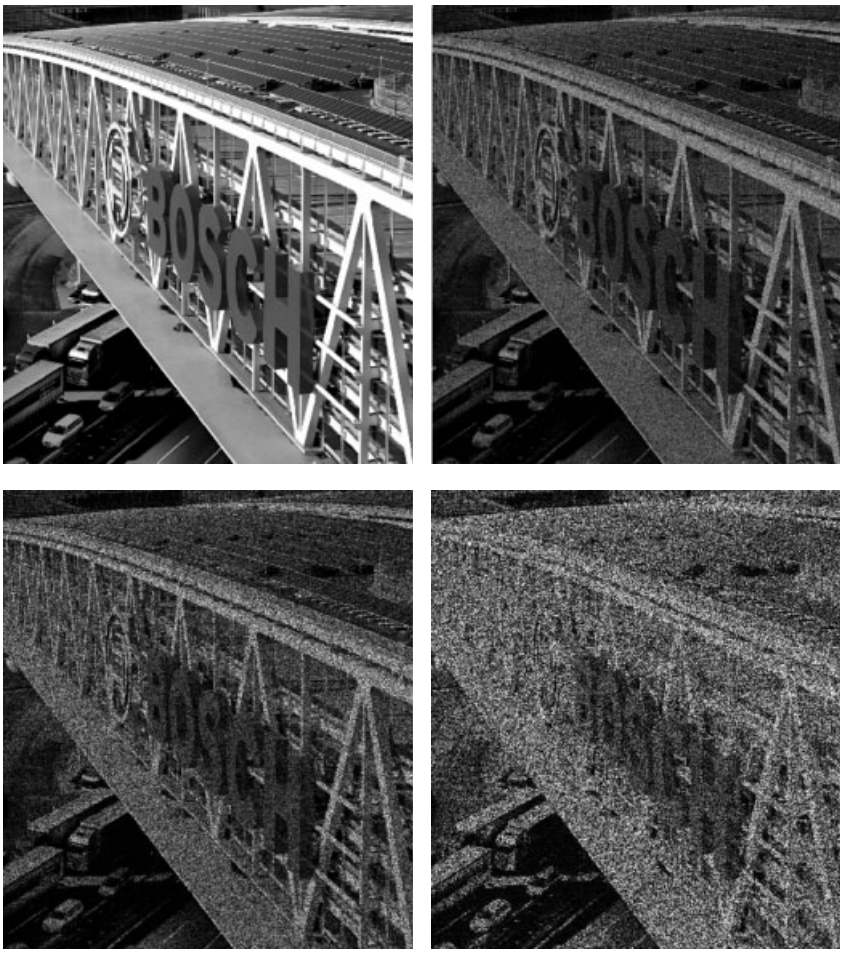


Figure 29. Increasing (synthetically generated) photon noise with decreasing exposure time $\Delta \mathbf{t}$. (Photo: Bosh. Reproduced by permission of Bosch GmbH.)

It is important to note that on moving platforms, long exposure times lead to the motion blur effect, marked by a lack of sharpness of object edges induced by fast motion. As a result, there is a trade-off between motion blur and photon noise.

As the current of photons λ reaching an image pixel linearly depends on the size of the pixel surface, it can be derived that in terms of photon noise, the effect of a growing number of image pixels can be compensated by an equally growing imager size.

3.4.3 Sensor design

In this section, after describing the general setup and specific automotive requirements to a camera, various basic design characteristics of cameras will be described. More specifically, the most common imager technologies, lens types, imager morphology, and camera types are highlighted.

3.4.3.1 Basics. A typical automotive video camera (Figure 30) consists of the following basic components:

- optics (high quality lens system);
- imager (CMOS, complementary metal oxide semiconductor);
- camera body (metal frame for dissipating heat);
- processing units (multicore CPU and/or hardware for graphic acceleration);
- communication device [CAN (controller area network), Flexray].

Different from consumer electronics, vehicle cameras require a high sensitivity to light and a higher dynamic range. These requirements lead to large imager sizes with comparatively small resolution (the number of pixels). Furthermore, vehicle cameras dispose of high quality optical lenses without an analog (hardware-based) zoom. In general, an analog zoom could be helpful in automotive

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

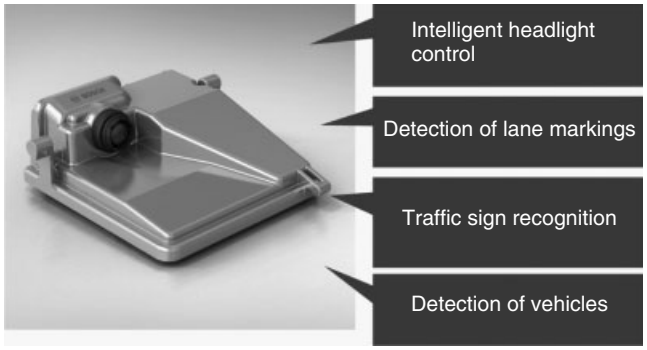


Figure 30. Multipurpose camera manufactured by Bosch. (Reproduced by permission of Bosch GmbH.)

applications; however, all moving system parts are prone to defects and have to be avoided. As opposed to that, cameras in consumer electronics typically have a higher resolution (as a central measure for product marketing) with typically low quality optical systems.

3.4.3.2 Imager technology. The most prevalent video imager technologies are CCD (charge-coupled device) and CMOS sensors. Both CCD and CMOS sensors are based on the inner photo effect. The captured light-inherent energy leads to the creation of electrons and holes in the semiconductor material. An applied voltage prevents a recombination of the charge carriers. The electrons of each pixel are gathered in a potential well in order to determine the resulting charge at the end of an exposure cycle. More specifically, a charge-to-voltage conversion takes place. The resulting voltage level is digitized and stored. It is important to note that an electronic shutter is required that assures that the image pixels will stop gathering electrons at the end of the exposure time interval.

The major distinguishing characteristic of CCD is that all charges are transferred to one charge-to-voltage converter. This resembles a bucket chain for charges that are moved toward the named converter. CCD sensors tend to be susceptible to the so-called blooming effect (overflow of pixels into neighboring pixels in case of an overexposure to light). They are highly sensitive to light, which is helpful for capturing images of sufficient quality at night or in tunnels. Typical applications for the CCD technology are passive (meaning without active system interference), camera-based DASs, that display surround information (e.g., visualization of the vehicle's back region in parking maneuvers).

As opposed to that, the major distinguishing characteristic of CMOS is that each pixel disposes of its own charge-to-voltage converter. Instead of transferring charges, the voltage levels and hence signals are transferred. The initial application for CMOS sensors was in low-cost consumer products. Nowadays, various technical improvements allow their application in driver assistance. The frame rate of CMOS imagers is high (up to 30 Hz in modern CMOS sensors with rolling shutter), which allows for short exposure times, minimizing the motion blur. Motion blur results from the technical restrictions in reading the charge form all pixels at the same time (called *shutter* as a reminiscence to analog photography). Depending on the shutter technique, a specific motion blur results (e.g., the rolling shutter effect). CMOS offers a very high dynamic range up to 100 dB (as a comparison, humans dispose of 105 dB). CMOS imagers are marked by less power consumption (inherent to its functional principle) and a higher frame rate compared to CCD. The blooming effect is reduced, as the gathered charges are not transmitted (a transistor as charge-to-voltage converter is connected in parallel to each light-sensitive pixel). Current CMOS have still a comparatively low sensitivity to light, as the processing electronics is integrated on the chip, which results in a lower filling factor (share of light-sensitive surface related to the overall surface of the imager chip). On the other hand, the possibility of integrating the processing units on the same chip as the imager itself leads to lower costs and a tight packaging. Typical application for CMOS imagers is cameras for active DASs that require high frame rates together with a high dynamic range. It can be expected that the CMOS technology will outrun CCD in most automotive applications, especially when CMOS reaches a sufficient sensitivity in low-light situations.

It is important to note that CMOS imagers can be constructed which are sensitive to the wavelength of light which is below the visible light [the so-called infrared (IR)]. Although not visible, humans can sense IR radiation as heat. IR-sensitive imagers are able to deliver gray value images at night. On the basis of the requirement of illuminating devices, active and passive IR sensors can be distinguished. Short wave or near infrared (NIR) sensors require a small amount of low power LEDs (light-emitting diodes) that illuminate the scene to be captured. As opposed to that, long wave or far infrared (FIR) are passive sensors without any illuminating device. It is important to note that NIR sensors typically are superior in terms of performance and availability.

3.4.3.3 Lens types. Mainly application-driven, two major lens types are used in vehicles: (i) perspective lenses and (ii) wide-angle lenses.

Very common in driver assistance are perspective lenses (monocular cameras typically for comfort systems and stereo cameras for safety systems). Cameras with perspective lenses are typically installed behind the windshield

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

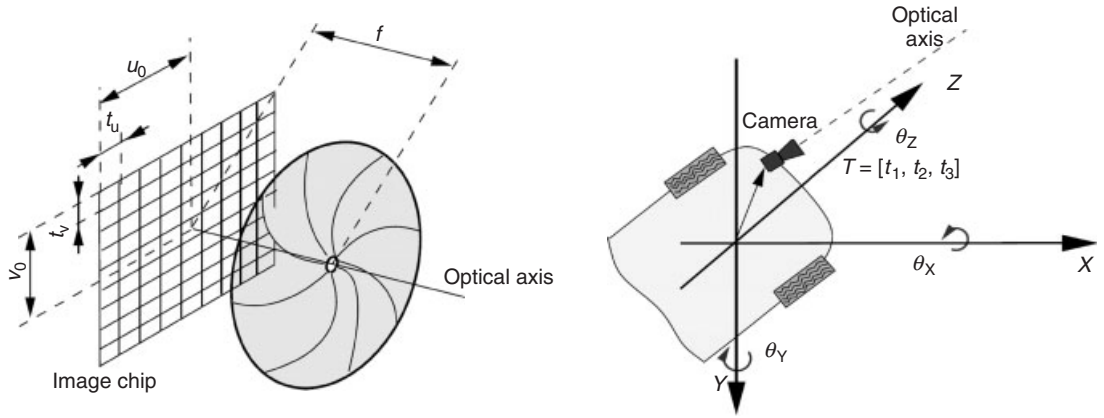


Figure 31. (a) Pinhole camera model with its intrinsic camera parameters. (b) Camera coordinate system and extrinsic camera coordinates. (Reproduced with permission from Michalke, 2010. © VDI-Verlag.)

capturing images, for example used for intelligent headlight control or lane marking detection. A perspective projection can be expressed by the so-called pinhole camera model (refer to Figure 31a), thereby allowing a very intuitive formal description (focal length f pixel size t_u, t_v).

More specifically, a 3D world position (X, Y, Z) (refer to Figure 31b for the coordinate system) can be transformed to a 2D pixel position (u, v) using a pinhole camera model that contains all intrinsic (i.e., camera internal) and extrinsic (i.e., camera external) parameters (in detail, these are the three camera angles θ_X, θ_Y, and θ_Z, which are aggregated in the rotation matrix R, three translational camera offsets t_1, t_2, and t_3, the horizontal and vertical principal points u_0 and v_0, and the normalized horizontal and vertical focal lengths $f_u = f/t_u$ and $f_v = f/t_v$), refer to Equations 18 and 19.

$$u = -f_u \frac{r_{11}(X - t_1) + r_{12}(Y - t_2) + r_{13}(Z - t_3)}{r_{31}(X - t_1) + r_{32}(Y - t_2) + r_{33}(Z - t_3)} \tag{18}$$

$$v = -f_v \frac{r_{21}(X - t_1) + r_{22}(Y - t_2) + r_{23}(Z - t_3)}{r_{31}(X - t_1) + r_{32}(Y - t_2) + r_{33}(Z - t_3)} \tag{19}$$

$$R(\theta_X, \theta_Y, \theta_Z) = R_X R_Y R_Z = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \tag{20}$$

with :

$$r_{11} = \cos(\theta_Z)\cos(\theta_Y)$$

$$r_{12} = -\sin(\theta_Z)\cos(\theta_X) + \cos(\theta_Z)\sin(\theta_Y)\sin(\theta_X)$$

$$r_{13} = \sin(\theta_Z)\sin(\theta_X) + \cos(\theta_Z)\sin(\theta_Y)\cos(\theta_X)$$

$$r_{21} = \sin(\theta_Z)\cos(\theta_Y)$$

$$r_{22} = \cos(\theta_Z)\cos(\theta_X) + \sin(\theta_Z)\sin(\theta_Y)\sin(\theta_X)$$

$$r_{23} = -\cos(\theta_Z)\sin(\theta_X) + \sin(\theta_Z)\sin(\theta_Y)\cos(\theta_X)$$

$$r_{31} = -\sin(\theta_Y)$$

$$r_{32} = \cos(\theta_Y)\sin(\theta_X)$$

$$r_{33} = \cos(\theta_Y)\cos(\theta_X)$$

Equations 18 and 19 can also be expressed in homogenous coordinates, which decrease the computational demands considerably. The required intrinsic and extrinsic parameters have to be determined in a calibration procedure usually applying a reference pattern (e.g., a checkerboard) that is captured from different view angles [see Heikkila and Silven (1997) for more details on the determination of camera parameters].

A second lens type of major importance is the wide-angle camera (also called *fisheye*).

Typically, for the wide-angle 3D-world-to-image projection, the so-called equidistant projection model is applied [refer to Abraham and Förster (2005) for a comparison of different projection models]. In Equations 21 and 22, the wide-angle image coordinates (u_{wide}, v_{wide}) are computed from the horizontal and vertical focal lengths given in pixels ($f_{u,\text{wide}}$, $f_{v,\text{wide}}$), the horizontal and vertical principal points ($u_{0,\text{wide}}$, $v_{0,\text{wide}}$) and the normalized wide-angle image coordinates (u^*_{wide}, v^*_{wide}).

$$u_{\text{wide}} = u^*_{\text{wide}} \cdot f_{u,\text{wide}} + u_{0,\text{wide}} \tag{21}$$

$$v_{\text{wide}} = v^*_{\text{wide}} \cdot f_{v,\text{wide}} + v_{0,\text{wide}} \tag{22}$$

The normalized wide-angle image coordinates (u^*_{wide}, v^*_{wide}) are computed from the 3D world position X_C, Y_C, Z_C (camera coordinate system) based on Equations 23 and 24

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

resembling an equidistant projection model.

$$u^*_{\text{wide}} = \frac{X_C/Z_C}{\sqrt{X_C^2 + Y_C^2}}\theta \tag{23}$$

$$v^*_{\text{wide}} = \frac{Y_C/Z_C}{\sqrt{X_C^2 + Y_C^2}}\theta \tag{24}$$

The variable θ represents the (undistorted) projection angle and is computed by Equation 25:

$$\theta = \arctan \frac{\sqrt{X_C^2 + Y_C^2}}{Z_C} \tag{25}$$

In the last step, the camera-related world coordinates $X_C = [X_C, Y_C, Z_C]$ have to be computed from the 3D world coordinates $X_W = [X, Y, Z]$ that are relative to the vehicle coordinate system (with R as the rotation matrix defined in Equation 20 and the translational world position of the camera $T = [t_1, t_2, t_3]$:

$$X_W = R \cdot X_C + T \tag{26}$$

It is important to note that (any kind of) projection model should always be applied on images that are free from lens distortions. For compensating the image distortion (also called *undistortion*), a lens distortion model is required. Such models dispose of lens-specific parameters that have to be determined offline [refer to Heikkila and Silven (1997) for details on a well-known lens distortion model for perspective lenses and Michalke, Stein, and Franke (2011) for an application using wide-angle lenses].

On the basis of the above projection equations for perspective and wide-angle lenses, image processing algorithms can be derived, for example, estimating the existence, size, and world position of traffic-relevant objects (refer to Section 3.4.5 for more details on applications for image processing).

3.4.3.4 Imager morphology. The imager pixels can have different shapes. For example, in consumer electronics, also alveolar pixel shapes exist (e.g., the super CCD sensor by Fujifilm). However, the typical pixel shape for automotive cameras is square.

Camera imagers can deliver monochrome (gray value) or RGB color images. In automotive cameras, RGB color images are realized by adding a repetitive pattern of three color filters to all pixels. Thereby, three images of reduced resolution are delivered. Pattern-specific interpolation operations (demosaicing) are used to receive a full-resolution image in three colors. The most prominent color pattern is the Bayer filter that consists of a repetitive pattern of two green, one red, and one blue pixel (Figure 32). Green was selected to fill half on the imager, in order to maximize light sensitivity. Other color pattern of inferior importance exists (e.g., RGBE for red/green/blue/emerald or CYGM for cyan/yellow/green/magenta).

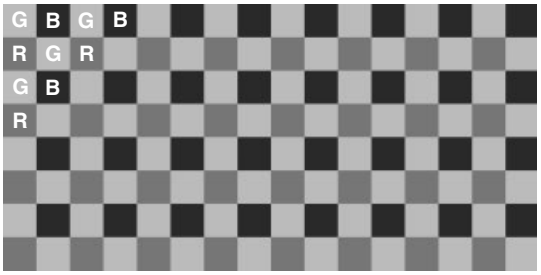


Figure 32. Bayer pattern (filter that covers the camera imager, R=red/G=green/B=blue).

For the sake of completeness, it should be mentioned that there are imagers that dispose of a pattern of one red color-filtered pixel among three unfiltered pixels. Thereby, the red-color channel can be measured together with a gray value image. In driver assistance, the color red is very important for the robust classification of traffic signs, traffic lights, and lane markings in construction sites.

3.4.3.5 Sensor system morphology and technology. Camera-based sensor systems can be distinguished into monocular (one camera) and stereo (two cameras) systems.

Until now, in automotive applications, monocular cameras are predominant. Monocular cameras are mainly used for object classification and for visualizing the close surroundings of the vehicle in parking support systems. As the driver is interested in the remaining space (e.g., in a narrow parking lot) distances and hence 3D data is required. Monocular cameras are not able to directly deliver 3D data. However, under some relaxing assumptions, it is possible to derive 3D information.

A typical relaxing assumption is the so-called flat plane assumption, that is, it is assumed that the measured world is flat. Under this assumption, the height Y can be set to 0 for all measured pixels (i.e., $Y = 0$ in Equations 18, 19, and 26, which simplifies the 3D to 2D conversion considerably). The described procedure is called *inverse perspective mapping* (refer to Broggi, 1995).

As an exemplary result, the monocular image in Figure 33a is mapped to the X–Z plane depicted in Figure 33b. It is important to note that all nonflat objects violate the initial assumption and hence are stretched to infinity, thereby leading to a mapping error.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

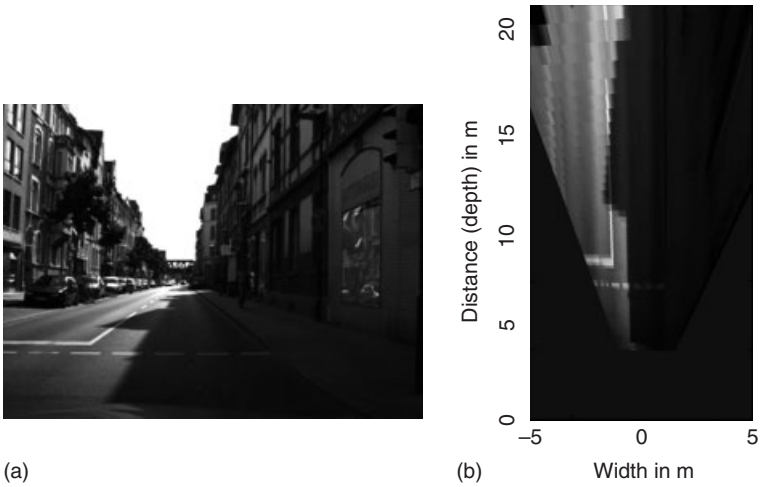


Figure 33. Inverse perspective mapping: (a) inner-city scenario (fulfilling the flat plane assumption); (b) Bird's-eye view. (Reproduced with permission from Michalke, 2010. © VDI-Verlag.)

Stereo vision is an emerging technology in cameras for consumer electronics. Also for automotive applications, stereo cameras are beginning to spread. Different from monocular cameras, using stereo cameras, it is possible to measure the 3D position of all measured image pixels without requiring any relaxation. An image pixel having a 3D position is often also called *voxel* in literature. For stereo vision, two cameras (with typically perspective lenses) are required whose relative position and orientation have to be determined based on a calibration step. More generally, when measuring data of N dimensions, a sensor system delivering $N+1$ dimensions is required.

In the following, the basic theoretical background for stereo computation is given. For computing 3D data, some mandatory preprocessing steps are required (Figure 34).

As described earlier, an undistortion step is required in order to compensate the lens distortion. After that, the images are rectified, meaning that the orientation of the optical axes of both cameras is aligned. This step is required in order to simplify the following correspondence search between the left and the right camera images. Finding correspondences is the key to computing the 3D position of pixels. In order to give some intuition to the role of correspondences in stereo processing, imagine an object being very close to two cameras that are positioned in parallel. The object will be projected on a different horizontal position at both camera imagers. As opposed to that, an object that is far away at the horizon will be positioned at virtually the same spot on both image planes. Hence, the smaller the distance of an object to the stereo camera is, the larger will be the relative horizontal shift of the object position. The horizontal shift is be called *disparity* $D(u, v)$ in the following. For computing the disparity, different approaches exist. Among the most prominent (also in the automotive domain) are correlation-based approaches (Willert *et al*., 2006), approaches relying in the so-called semi-global matching (see Hirschmüller, 2008, for some fundamental background), and approaches

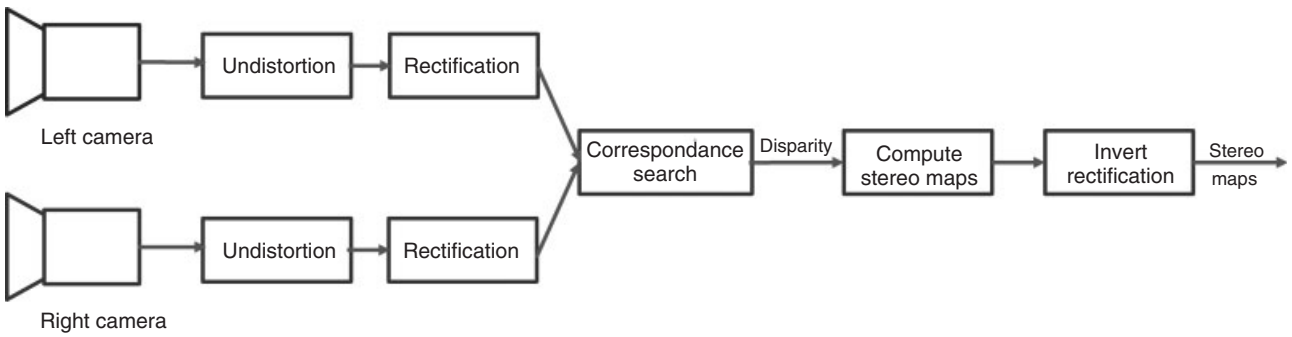


Figure 34. Stereo processing steps.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

using the census transformation (see Zabih and Woodfill (1994), for details). Following the computation of the disparity $D(u, v)$, Equations 27–29 are applied for computing the pixel 3D positions (stereo maps) relative to one of the cameras:

$$Z_{\text{stereo}} = \frac{f_u B}{D(u, v)} + t_3 \tag{27}$$

$$Y_{\text{stereo}} = \frac{Z_{\text{stereo}}(v - v_0)}{f_v} + t_2 \tag{28}$$

$$X_{\text{stereo}} = \frac{Z_{\text{stereo}}(u - u_0)}{f_u} + t_1 \tag{29}$$

In Equation 27, B is the basic distance between the left and the right camera's principal points.

Figure 35b–d visualizes the 3D positions of all image pixels for the exemplary image in Figure 35a. Please note that in the map values above the ground are negative, which is due to the chosen coordinate system (Figure 31b). Following Figure 34, as a last processing step, the stereo maps can be remapped inverting the image rectification step. Thereby, the 3D data match the undistorted input data. Thereby, for example, classification algorithms can run directly on the unrectified image raw data improving algorithmic performance (Figure 35).

3.4.4 Sensor configuration

In the automotive domain, the available installation locations are sparse. Typical locations for video sensors are as follows (Figure 36):

- Behind the windshield close to the rear mirror: NIR camera, stereo or monocular camera.
- Behind a plastic shielding at the radiator grill: FIR camera.
- Side mirrors: wide-angle camera.
- Trunk lid: wide-angle camera.

FIR sensors have to be installed at the front of the vehicle in order to avoid interference with motor heat and because the windshield would block the FIR-relevant wavelength, thereby weakening the signal. The active NIR sensors can be installed at the vehicle windshield, as the sensor can be tuned to the wavelength of the illuminating LED and because the windshield does not block the NIR-relevant wavelength.

3.4.5 Applications

DASs in general can be separated into comfort systems and safety systems. Comfort systems support drivers in standard, mainly tiresome tasks helping to center the driver attention on the remaining driving tasks. Different

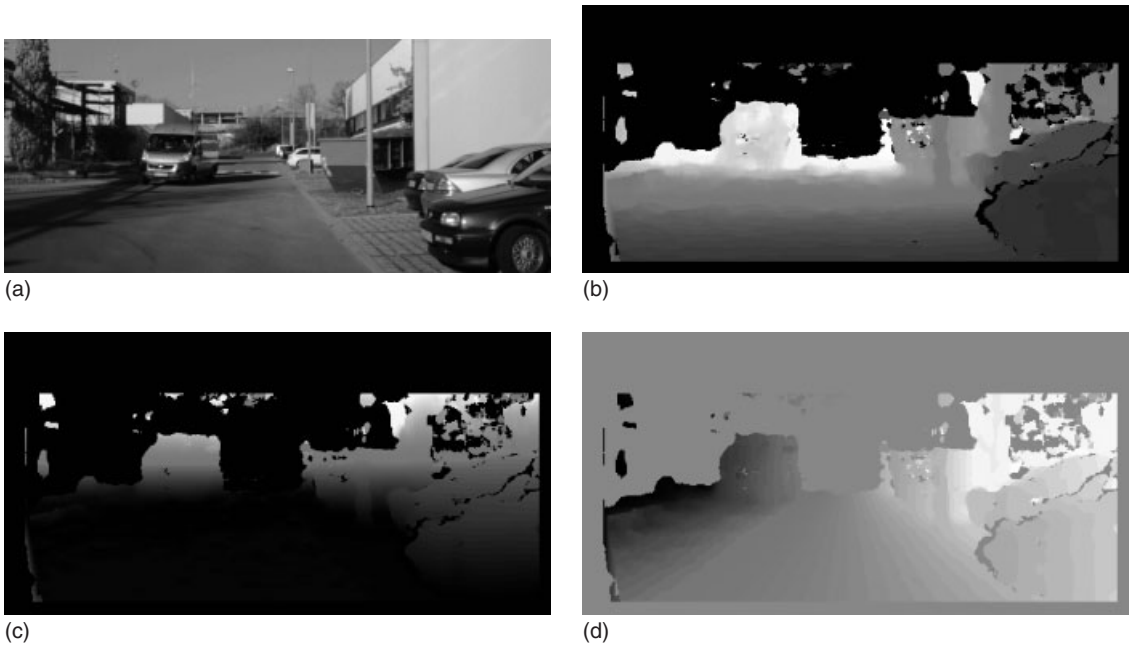


Figure 35. 3D world data from stereo video camera (dark represents small values, bright high values): (a) left input image, (b) distance map Z_{stereo}, (c) height map Y_{stereo}, and (d) horizontal position map X_{stereo}.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

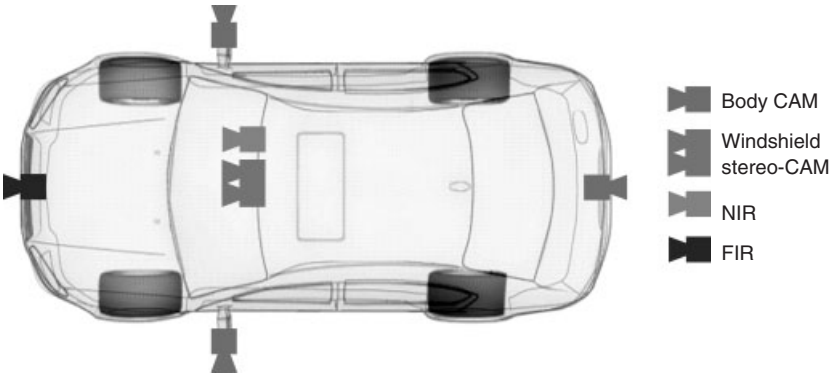


Figure 36. Typical installation locations of automotive cameras.

from safety systems, comfort systems are never activated automatically without the direct intension of the driver. Safety systems actively support the driver in dangerous driving situations and thereby help to prevent or mitigate accidents.

Camera-based comfort systems typically support the driver by offering information of the vehicle surrounding. Camera-based safety systems allow the automatic activation of safety measures such as braking and in the future also steering. In the following, different telematics-related, camera-based DASs are described.

Currently cameras are used in comfort systems for:

- detecting traffic signs. Traffic signs are projected in the head-up display or in the command unit of the vehicle. Furthermore, a warning can be triggered in case the driver surpasses a detected speed limit.
- supporting the driver in the night. More specifically, the visibility range and object resolution can be improved. Specifically, the visibility of vulnerable road users (VRU) can be improved. The gray value IR image can be used for classifying objects and displaying them to the driver.
- supporting the driver during parking maneuvers by displaying objects that are close to the vehicle rear end, or in more recent systems, the entire 360° vehicle surroundings.

Currently, cameras are used in safety systems for:

- detecting traffic participants and estimating their motion parameters in order to predict their future course. Thereby, a collision risk can be estimated. In case of an imminent collision, automatic avoidance or mitigation measures are taken (e.g., automatic emergency braking).
- achieving sensory redundancy and assuring plausibility of performance-critical sensor signals.



(a) (b)

Figure 37. Camera-based driver assistance functions: (a) rear view camera offering support in parking maneuvers; (b) traffic sign detection. (Reproduced by permission of Bosch GmbH.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

When gathering the individually detected signals and information in a central database, numerous innovative telematic applications could be realized, such as:

- real-time adaptation of map data;
- real-time traffic flow estimation;
- enhanced dynamic route navigation;
- support in finding parking spaces in inner city (Figure 37).

3.4.6 Summary and outlook

Section 3.4 gave insight into automotive-specific requirements of camera sensors. Different basic design characteristics of cameras were described (e.g., imager technologies, lens types, imager morphology, and camera types). Furthermore, the theoretical background required for designing camera-based driver assistance algorithms was given as well as some important references to explain specific topics. Section 3.4 specifically focused on stereo vision as an important future technology, driving the development of novel DASs.

GLOSSARY

Adaptive cruise control (ACC)	Bosch system (other trade names exist) based on a radar sensor setup that allows the system vehicle to automatically slow down when approaching another vehicle ahead and accelerate again to a preset velocity.
Antenna gain	It is a key performance figure that combines the directivity and electrical efficiency of an antenna. In case of a transmitting antenna, the gain defines the ability of the antenna to convert input power into electromagnetic waves in a specified direction. For a receiving antenna, the gain defines how well the antenna can convert the electromagnetic waves arriving from a specified direction into electrical power.
Application-specific integrated circuit (ASIC)	ASIC is an integrated circuit that is designed for a specific task or application (e.g., data compression with a specific compression algorithm). Today's ASIC can be system-on-chip architectures with complex, heterogeneous structures.
Bayer pattern	A repetitive color pattern that is set in front of a light-sensitive imager of a digital video camera. The captured single channel image can be decomposed into a multichannel color representation (e.g., RGB).
Beamforming	Approach that dynamically adapts the antenna characteristic of a radar sensor (direction and emitted power of electromagnetic waves is changed). Digital beamforming realizes the named adaptation without relying on moving parts.
Bird's-eye view	A term that is used in image processing designating an approach that transforms a perspective camera image (under some assumptions) into a view from above.
Controller area network (CAN)	An automotive microcontroller communication protocol that allows the communication between different electronic components in a vehicle.
Charge-coupled device (CCD)	A well-known technology for constructing transistors among other things on light-sensitive video imagers.
Complementary metal oxide semiconductor (CMOS)	A well-known technology for constructing transistors among other things on light-sensitive video imagers.
Correspondence search	Basic principle for computing disparity data from two or more camera images. More specifically, a pattern (representing, for example, an object) present in one image is searched in the second image.
Central processing unit (CPU)	Central computing hardware in a computer, which conducts instructions of a software program, thereby performing basic logical operations on input data and generating output data.
Disparity	Typically horizontal shift (in pixels) between one and the same objects displayed in the rectified images (with parallel optical axes) of a stereo camera setup. The bigger the disparity is, the nearer is the object to the camera.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Doppler effect	Designates a change in frequency of an (electromagnetic) wave in case the source is moving relative to an observer. Hence, from the frequency change, the object velocity can be computed.
Driver assistance system (DAS)	An electronic system installed in vehicles to assist drivers in certain driving situations. The main goals are not only the increased safety and driving comfort but also the improvement of economics.
Extrinsic camera parameters	All external parameters of a camera setup (3D position and orientation relative to the world).
Far infrared (FIR)	Related to the wavelength of electromagnetic wave, the spectral band above the visible light is called *infrared*. With a growing wavelength, the infrared spectral band can be subdivided into near-, mid-, and far-infrared.
Fisheye camera	Camera relying on a specifically formed lens that allows capturing wide-angle images (up to 180° field of view).
Flexray	An automotive microcontroller communication protocol that allows the communication between different electronic components in a vehicle.
Frequency-modulated continuous-wave technique (FMCW)	Radar technique that is characterized by a continuously transmitted radar wave, whose frequency is varied linearly over time. The time delay and the Doppler effect of the radar reflections lead to a frequency shift that can be used to derive the object's position and velocity.
Frequency-shift keying technique (FSK)	Similar to the FMCW, a continuous radar wave is transmitted. In contrast to FMCW, the transmission frequency remains constant for a certain time. Therefore, a frequency shift of the reflected signal is generated only by the Doppler effect. By emitting two different frequencies consecutively, the position and velocity of a reflecting object can be derived.
Hall effect	An electric voltage measurable at a conductor positioned in a stationary magnetic field, while being traversed by an electric current. A change in the magnetic field results in a change in voltage.
Intrinsic camera parameters	Internal sensor parameters that cannot be directly measured from outside (e.g., focal length or pixel size for video imagers).
Lane change assistant (LCA)	A driver assistance system that warns the driver if he or she intends to do a lane change maneuver, but there is a vehicle approaching on the neighbor lane.
Light-emitting diode (LED)	A light source consisting of semiconductor material.
Long-range radar (LRR) sensor	A radar sensor used to detect objects at far distances. In automotive applications, these are distances of over 200 m.
Monopulse	Radar technique that allows measuring the azimuth angle of an object with a single radar pulse or FMCW ramp.
Mechanical scanning	As opposed to digital beamforming, the antenna characteristic is changed relying on moving mechanical parts (typically, rotating cylinders) in the radar sensor.
Near infrared (NIR)	Related to the wavelength of electromagnetic wave, the spectral band above the visible light is called *infrared*. With a growing wavelength, the infrared spectral band can be subdivided into near-, mid-, and far-infrared.
Pinhole camera model	Simplified perspective camera model that allows a straightforward mapping between the 3D world and the 2D image. Its application requires undistorted images.
Piezo element	A piece of specific crystal or ceramic that reacts to mechanical pressure by accumulating a charge (or vice versa).
Poisson distribution	A well-known statistical distribution that is used to model various physical phenomena.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Probability density function (PDF)	Mathematical model of a continuous, statistical distribution. The PDF describes the relative likelihood of a continuous random variable to take on a specific value.
Pulse-Doppler radar	Radar sensor that emits signal pulses to evaluate the time until the backscattered signal is detected by the receiver (in order to derive the object distance) as well as the Doppler frequency shift (in order to derive the object velocity).
Radio detection and ranging (RADAR)	Radar is a technique to detect objects and determine its position and relative velocity. This is achieved by sending out an electromagnetic wave and evaluating the time delay and frequency shift of the received signal that was backscattered by an object.
Radar cross section (RCS)	Radar-related measure that describes the backscattering behavior of a single target, that is, how much of the sent out power (of an electromagnetic wave) is reflected back to the receiver.
Rectification	Compensation of relative differences in the extrinsic parameters (especially orientation) of the cameras in a stereo setup. Thereby, the optical axes of the cameras are virtually aligned (made parallel), allowing a simplified correspondence search. Some authors see the undistortion step as a subpart of the rectification step. In this contribution, rectification and undistortion are handled separately.
Semi-global matching	Widely known principle for realizing correspondence search in stereo image processing.
Signal noise ratio (SNR)	A measure that compares the level of a desired signal to the level of (e.g., thermal) noise. Typically, it is defined as the ratio of signal power to the power of noise.
Sound navigation and ranging (SONAR)	A measurement technique that uses sound propagation to measure distances and velocities of objects.
Transducer	A sensor that acts as both an emitter and a receiver of an active measuring pulse.
Trilateration principle	Approach that allows calculating the 3D world position of an object measured by multiple sensors (placed at different positions) relying on triangulation.
Undistortion	Model-based compensation of image-distorting lens effects in image processing.

ENDNOTE

1. In a monostatic radar system, the same antenna is used for sending and receiving. Unlike, in a bistatic radar system, the antenna for sending is different from the antenna for receiving.

REFERENCES

Abraham, S. and Förster, W. (2005) Fish-eye-stereo calibration and epipolar rectification. *ISPRS Journal of Photogrammetry and Remote Sensing*, **59** (5), 278–288.

Broggi, A. (1995) Robust Real-Time Lane and Road Detection in Critical Shadow Conditions. *Proceedings of IEEE International Symposium on Computer Vision*, Parma.

Dissanayake, G., Sukkarieh, S., Nebot, E.M., and Durrant-Whyte, H.F. (2001) The aiding of a low-cost strapdown inertial measurement unit using vehicle model constraints for land vehicle applications. *IEEE Transactions on Robotics and Automation*, **17** (5), 731–747.

Heikkila, J. and Silven, O. (1997) A four-step camera calibration procedure with implicit image correction. *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 1106–1112.

Hirschmüller, H. (2008) Stereo processing by semi-global matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30** (2), 328–341.

Jaehne, B. (2005) *Digital Image Processing, 6.* Springer-Verlag, Auflage.

Kost, F., Koch-Dücker, H., Niewels, F., and Schuh, J. (2004) *Fahrstabilisierungssysteme*, Robert Bosch GmbH, Stuttgart.

Linder, H.-J. (1999) *Physik für Ingenieure*, Fachbuchverlag Leipzig im Carl-Hanser-Verlag, München.

Michalke, T. (2010) Task-dependent scene interpretation in driver assistance. PhD dissertation. VDI-Verlag.

Michalke, T., Stein, F., and Franke, U. (2011) *Towards a closer fusion of active and passive safety: optical flow-based detection of vehicle side collisions*. IEEE Intelligent Vehicles Symposium, Baden.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Raj, B., Kalgaonkar, K., Harrison, C., and Dietz, P. (2012) Ultrasonic Doppler sensing in HCI. *IEEE Pervasive Computing*, **11** (2), 24–29.

Reif, K.[Hrsg.] (2012) *Sensoren im Kraftfahrzeug*, Vieweg+Teubner (Bosch Fachinformation Automobil), Wiesbaden.

Reif, K. and Knoll, P. (2012) in *Sensorik für Fahrzeugrundumsicht, Fahrstabilisierungs-systeme und Fahrerassistenzsysteme* (ed. K. Reif), Vieweg+Teubner, pp. 130–145.

Reim, A., Klier, W., Hillenbrand, S. and Otterbein, S. (2008) *Design and implementation of a vehicle observer for sideslip angle computation without vehicle and tire models*. 8th Stuttgart International Symposium "Automotive and Engine Technology", Stuttgart.

Willert, V., Eggert, J., Adamy, J., and Koerner, E. (2006) Non-Gaussian velocity distributions integrated over space, time and scales. *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics*, **36** (3), 482–493.

Zabih, R. and Woodfill, J. (1994) Non-parametric Local Transforms for Computing Visual Correspondence. *Proceedings of the 3rd European Conference on Computer Vision*.

FURTHER READINGS

Bouguet, J. (2010) *Camera Calibration Toolbox for Matlab*, http://www.vision.caltech.edu/bouguetj (accessed 01 August 2013).

Reinhold, K. (2010) *Analyse kurzer akustischer Signale zur Bestimmung der Relativgeschwindigkeit zwischen Objekten*, Universität Karlsruhe (T.H.), Karlsruhe.

Richards, M.A. and Scheer, J.A. (2010) *Principles of Modern Radar, Basic Principles*, SciTech Publishing, North Carolina.

Skolnik, M. (2010) *Radar Handbook*, 3rd edn, McGraw-Hill, New York.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto175
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Data Acquisition by Roadside Detection

Christoph Roth
Siemens AG, Munich, Germany

1 Introduction to Data Acquisition 1
2 Data Types 1
3 Use Cases 4
4 Detector Technologies 6
Glossary 13
References 13
Further Reading 13

1 INTRODUCTION TO DATA ACQUISITION

The management of sustainable, efficient, and competitive mobility requires innovative and reliable traffic data acquisition systems since the efficiency of many intelligent transport systems depends on the accuracy of the traffic data supplied by the roadside detection systems. High traffic flows, short travel times, and low air pollution can only be achieved with systems based on modern detection technologies. Many traffic signals, for instance, optimize the use of the road by calculating the phases for red and green based on the traffic demand. Line control systems on highways prevent congestions by reducing the speed limit in case of high traffic flow.

Many different technologies are used for the various intelligent transport systems in order to meet their specific requirements. The choice of the detection system depends on the kind of traffic data and the accuracy that best meets the demands of the overall system. For some systems, pure presence information is sufficient to achieve a high efficiency, but other systems need advanced traffic data with detailed information about the vehicle mix on the road.

This section begins with a description of the types of traffic data provided by current detection systems ranging from simple presence information to advanced traffic data, followed by an overview of use cases for traffic data acquisition systems, and the final part covers the technologies used for detection of certain traffic parameters.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 DATA TYPES

2.1 Vehicle presence

Many detectors such as inductive loops provide information about the vehicle presence in a certain detection zone. When the vehicle enters the detection zone, the output of the detector switches on and the output is turned off once the vehicle leaves the zone. Many detectors used for intersection control such as inductive loops, video, or radar are vehicle presence detectors. Even though the information content of presence data is rather low, many systems require a very high accuracy. At intersections, small detection errors can deteriorate the quality of the signal control considerably. The detectors need to be sensitive enough to have a high detection rate even for small vehicles, motorbikes, and bicycles but they must not be disturbed by environmental conditions or vehicles on the neighboring lane. This limits the use of detectors to the ones that are specifically designed for traffic applications and that can meet these strict requirements.

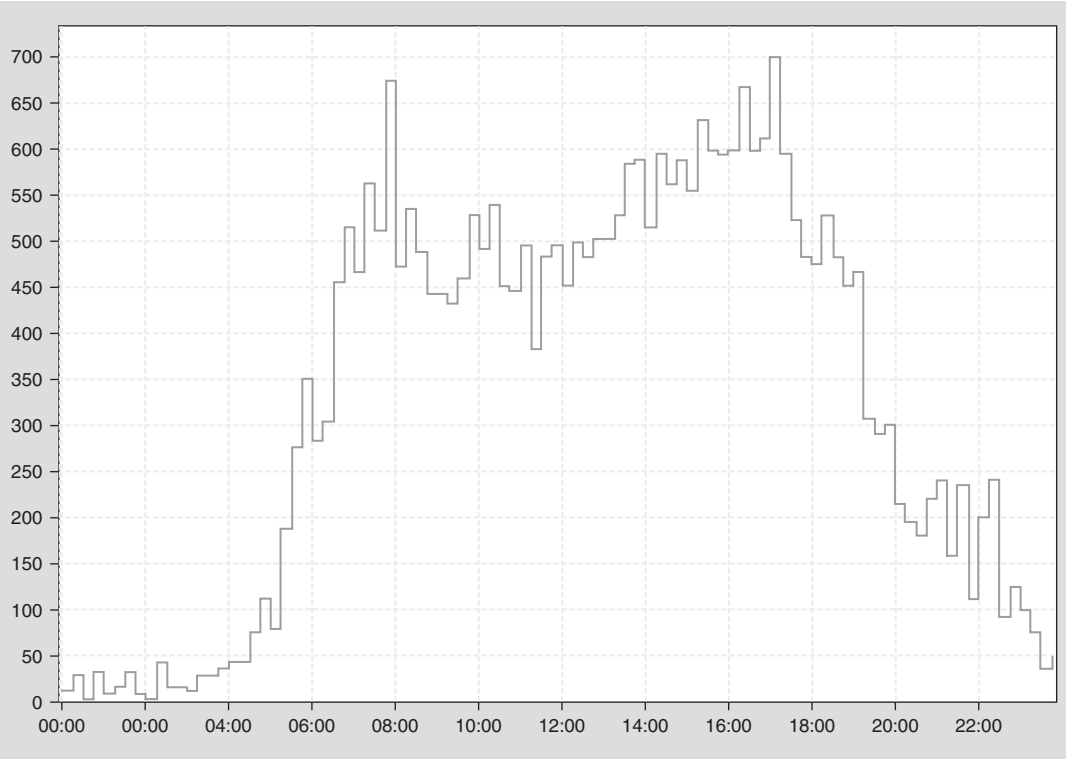


Figure 1. Typical traffic volume (vehicles per hour) with peak traffic around 8 a.m. and 5 p.m.

2.2 Traffic volume

The traffic volume is equal to the number of vehicles per time interval—usually 1 h. Nearly all traffic detectors can provide traffic volume. Even from the presence detectors described earlier, traffic volume data is derived by counting the number of output pulses in the signal controller. Figure 1 shows a typical graph of the traffic volume plotted versus the time of day in units of vehicles per hour. The distribution shows the peak traffic between 8 a.m. and 5 p.m.

Detectors for traffic volume need to be able to distinguish two consecutive vehicles from a vehicle with a trailer, as in most traffic applications a vehicle with trailer is regarded as one vehicle.

Many detectors have a detection zone that extends one or a few meters in the direction of travel. This might reduce the count accuracy in applications with frequent queues, as the spacing between successive vehicles might be smaller than the detection zone in which case the detector will not be able to separate the vehicles and count them correctly. In applications where high count accuracy is required in queued traffic detectors with small detection zones are used.

2.3 Speed

Vehicle speed is one of the main kinds of traffic data used by many intelligent transport systems. The accuracy of the speed measurement depends on the technology and often it is not the speed of individual vehicles but the average speed in a certain time interval, which is used by the overall system.

Speed can be measured either directly, for instance, with Doppler Radar technology or indirectly by two presence detectors separated by a few meters in the direction of traffic flow, for instance, double loop systems. In this case, both detectors are triggered by the vehicle and from the time difference of the trigger points the speed can be derived (Markovic *et al*., 1996).

2.4 Occupancy

The occupancy is defined as the time a certain point of the road is occupied by a vehicle divided by the length of the time interval considered. The most common unit of occupancy is percent, but sometimes it is reported in time units, provided the length of the reference interval is known.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Occupancy is often used for the assessment of the traffic condition. In free-flow traffic, the occupancy is usually in the low single digit percentage range and it increases considerably in case of congestion.

Even though the definition relates to a point detector in all practical applications, the occupancy is measured by detection zones, which have an extension of one or a few meters in the direction of traffic flow. For loop detectors, for instance, the occupancy is measured as the percentage the loop is occupied. The difference to the occupancy of a point detector is uncritical, as the absolute values of the occupancy are not important but the relative changes related to changes in the traffic condition.

2.5 Vehicle classification

Vehicle classification is used to determine the types of vehicles the traffic consists of. Worldwide, there are many different regional definitions for vehicle classes. One of the most common classification schemes is the distinction between passenger cars and trucks, where the weight or the maximum allowed weight is used for the differentiation. Examples for a more detailed classification are motorcycles, passenger cars, passenger cars with trailer, vans, trucks, trucks with trailer, semitrailer trucks, and busses.

Many detectors can classify vehicles into two classes based on the vehicle length. In case the detector can measure the length of the vehicles, the classification can easily be done by introducing a limit length below which the vehicle is classified as passenger car and above which the detector outputs a truck. In many cases, the limit length is adjustable and is set in the range between 5 and 6 m.

For more than two vehicle classes, a simple length classification is not sufficient. The detector needs to retrieve more advanced information about the vehicle for instance the height profile or the inductive signature from a loop detector. The actual classification is done by sophisticated software algorithms, which derive the information about the vehicle type from the height profile or inductive signature (Gajda *et al*., 2001).

2.6 Level of service

The level of service (LOS) describes the actual traffic situation on a certain stretch of the road network. It is a measure of the effectiveness of the road meaning the capacity to support a certain traffic flow. Many traffic management systems use three LOS levels: free flow, slow traffic, and congestion.

The LOS is derived from basic traffic data, in many cases from traffic volume and speed. At free-flow traffic, the traffic volume is low and the speed is high—usually close to the local speed limit. As traffic volume increases, the speed will not change until a critical volume is reached where the speed drops. This is the point where the traffic volume reaches the maximum capacity of the road. A further increase in traffic will result in stop and go traffic, which has even lower speed and a decreasing traffic volume. At complete congestion, the traffic volume and speed are close to zero.

The LOS levels are areas on a two-dimensional graph of traffic volume and speed (Figure 2). The location of the boundaries between the different LOS levels depends on the type of road, the maximum allowed speed limit, and the number of lanes.

2.7 Travel time measurement

While all traffic parameters mentioned earlier can be measured by local detectors, which observe a certain point in the road network, travel time measurement requires at least two detectors at different locations. The travel time is measured by reidentification of vehicles at two or more locations and comparing the time stamps of the two detection events. For reidentification, certain criteria such as the vehicle's license plate are used.

One interesting aspect of travel time measurement is the fact that a medium or low detection rate is often sufficient for a good assessment of the traffic situation. Any traffic is characterized by the fact that the speed of vehicles is highly correlated at short distance, which means that all cars have a speed that is similar to the cars in their vicinity. Therefore, it is sufficient to detect a small sample, a few percents, of the vehicles for the determination of the actual travel time. This relaxes the requirements for the detection systems, as they do not need to detect every vehicle with high accuracy.

It might be that two vehicle classes travel with different speed like on highways, where the trucks often have a different speed limit than passenger cars. But even in this case, the vehicles within one class travel with similar speed and it is sufficient to capture a sample within each vehicle class.

Several different technologies are used for travel time measurement systems ranging from cameras for automatic number plate recognition (ANPR), magnetic detectors to Bluetooth tracking. All systems are described in more detail later.

The accuracy is defined by the detection rate and the reidentification rate. The first defines the number of vehicles detected correctly by the first detection system at the beginning of the travel time section. How many of them are then reidentified correctly at the end of the section

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

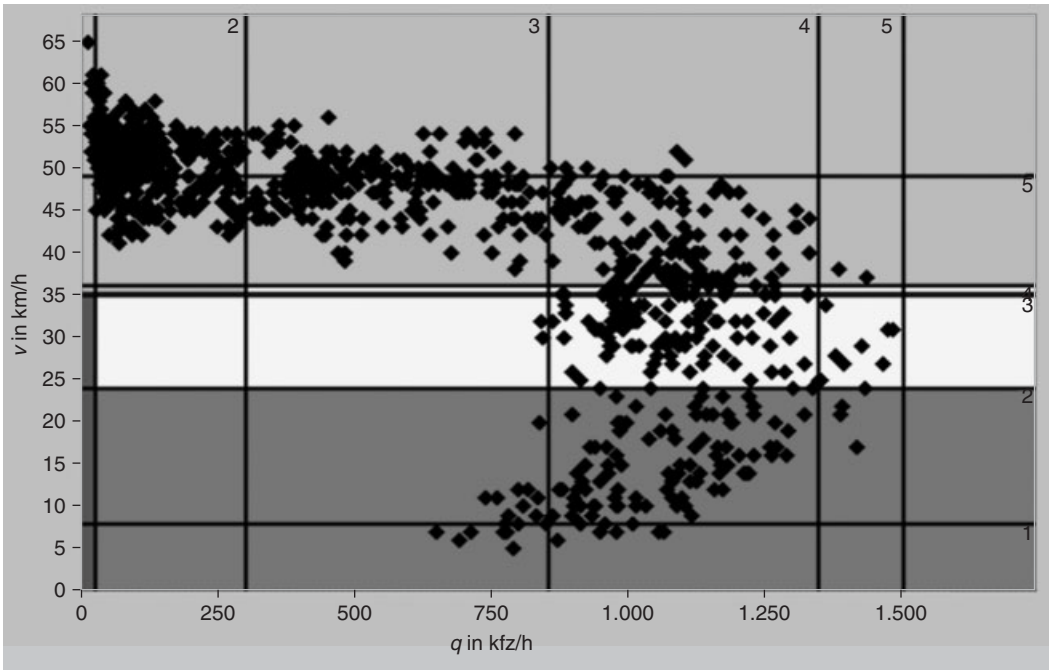


Figure 2. Example of traffic data divided into three LOS levels. Number of vehicles per hour on the *x*-axis and speed in kilometer per hour on the *y*-axis. Each data point represents a 5-min interval. (Reproduced with permission from VMZ Berlin Betreibergesellschaft mbH. Copyright VMZ Berlin Betreibergesellschaft mbH.)

is defined by the reidentification rate. Magnetic detectors, for instance, have a high detection rate; however, as the magnetic signatures are not unambiguous, the reidentification rate is usually around 50%. This means that nearly all vehicles are detected at the first location but only about every other vehicle can be correctly reidentified at the second detector location.

Bluetooth tracking systems, as another example, have a low detection rate, as only a small number of vehicles contain mobile Bluetooth devices that are switched on. Nevertheless, the reidentification rate is close to 100%, as all vehicles that are detected at the beginning of the section are with high certainty also detected at the end of the section (for more details, see Road Traffic and Travel Information (RTTI)).

2.8 Other traffic data

Many intelligent traffic systems rely on the basic traffic data such as presence information, traffic volume, speed, occupancy, and classification. Nevertheless, in some special cases, other data is used. A few examples are the headway that is defined as the temporal space between two vehicles. Specifically, the headway is the time that elapses between the front bumper of a first vehicle and the front bumper of the second vehicle. Headway is similar to the gap that relates to the time between the first vehicle's rear bumper and the front bumper of the second vehicle (University of Idaho). Average values for headway and gap can be derived from traffic volume, speed, and occupancy data.

The queue length is an interesting parameter, which is sometimes used for intersection control. It is defined as the distance between the stop line and the last vehicle that stopped at the end of the queue in meters.

3 USE CASES

Traffic data is used by many traffic systems and most of them are described in more detail, for instance, in Applications—Intelligent Roads and Cooperative Systems: Urban Traffic Management, Advanced Highway Management Systems, and Road Traffic and Travel Information (RTTI). A short summary of some of the basic use cases for roadside detectors in intelligent traffic systems are discussed in the following sections.

3.1 Vehicle detection at intersections

Detection at intersections is used to adjust the green and red phases according to the traffic demand. The traffic situation in urban environment changes often depending

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

on the time of day, type of day (weekday, weekend, or holiday), and special events such as sports, games, or fairs. Furthermore, traffic always shows large fluctuations in respect to nearly all traffic parameters. In order to handle the changes especially in traffic volume, vehicle-actuated traffic signals are equipped with detectors to measure the traffic parameters and use this information to adapt the signal phases according to the actual traffic situation.

As an example, consider an intersection of a main through-road and a side road with very small traffic volume. A fixed time traffic signal would have to give green time to the side road—and at the same time, stop the traffic on the main road—in every cycle. This is likely to cause congestion on the main road during peak traffic. A vehicle-actuated traffic signal would give a green phase to the side road only when a vehicle is detected to be waiting at the stop line. This increases the green time on the main road and, therefore, increases the traffic volume this intersection can handle.

Another example is the approach detection that is often used to handle intersections of two roads with similar high traffic volume. The approach detectors are usually located 30–60 m in front of the stop line. During the red signal phase, a queue of vehicle builds up the length of which usually exceeds the distance from the stop line to the detector. When the signal switches from red to green, the queue starts to dissolve and the detector is used to find out whether the main part of the queue has already dissolved or is about to pass the intersection. If the time gap between the vehicles exceeds a certain length—usually 2–3 s—it is assumed that the queue has dissolved and a free-flow traffic situation is present at the location of the detector. This is the time when the green time for this approach will be stopped in order to give the valuable green time to the other approach where a queue has built up in the meantime. This kind of traffic control ensures that green time is used efficiently for the dissolution of queues and not wasted for free-flow traffic situations where the actual traffic volume is small.

Depending on the size and layout of the intersection, the traffic volume, and the requirements of the road operators, many different systems for the traffic actuation of the signals are used. In all cases, the detector data is used as an input for the calculation of the optimum length of the signal phases.

Nearly all intersection control systems use the detection of vehicle presence only. The detectors have one output for each detection zone, which is activated by the vehicles. The data processing such as vehicle counting or the measurement of the time gap between vehicles is done in the controller.

3.2 Adaptive traffic control system

The vehicle actuation of traffic signals increases the throughput of the intersection but it is not sufficient for urban environments with a high density of signalized intersections. For these applications, adaptive traffic control systems are required that optimize the relative timing of the phases of all traffic signals in a certain area. The simplest implementation of traffic control systems is the so-called green wave, where the timing of the green phases is adjusted such that most of the vehicles can pass all traffic lights without a stop. In many cases, more advanced systems are used, which can adapt to the actual traffic situation and optimize all traffic signals for highest traffic flow, lowest travel time, or least number of stops at the intersections.

All adaptive traffic control systems rely on accurate traffic data in order to generate efficient signal plans for all traffic lights in their area. The data of the detectors that are used for the local intersection control can sometimes be used for the adaptive control system as well. However, in many cases, additional detectors need to be installed, either because not all intersections are equipped with a sufficient number of detectors or because the adaptive control system needs additional data.

An example for adaptive traffic control systems is SCOOT (Split Cycle Offset Optimization Technique), which is used in many large cities worldwide. (Scoot)

3.3 Traffic management

While intersection control and adaptive traffic control systems are designed to optimize the traffic flow in a road network, traffic management is related to information about the actual traffic situation and the means to communicate this information to the road users. Many large cities have traffic management systems, which display the actual traffic situation in the Internet, where the city map is shown and for the main roads a color code is used to indicate LOS. In many cases, three levels are used: green for free-flow traffic, yellow for slow traffic, and red for congestion.

Most traffic management systems use detectors in the free-flow traffic, which are located at a distance of several hundred meters from an intersection. They measure traffic volume, speed, occupancy, and often a low level vehicle classification from which the LOS can be calculated in the traffic management center. The detectors at intersections can usually not be used for traffic management, as the traffic queues on every red phase. Only detectors in a large distance from the intersection can differentiate well between the different levels of service.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Recently, travel time systems have become available, which measure the traffic situation on a complete section of the road network. They are often used as additional data sources for modern traffic management systems, as they cover a much larger area than the local detectors which measure the traffic situation at one particular spot on the road.

Besides the display of the actual traffic situation in the Internet, the information is often used by other media as well; for instance, on variable message signs on the arterial roads or via traffic news in the radio. The traffic information is also used by the navigation systems to find the best route for the driver depending on the actual traffic situation.

4 DETECTOR TECHNOLOGIES

4.1 Inductive loops

Inductive loops are the most widely used technology for vehicle detection. The technology has been developed many years ago and still today it is estimated that more than 50% of all detectors are based on inductive loop technology. The main advantages are high data accuracy and reliability in all environmental conditions, which can only, in a few cases, be surpassed by other detection technologies. Inductive loops achieve high detection rates and low false detection rates in all environmental conditions.

Inductive loops are wire loops placed in the road surface. A rectangular trench is cut in the asphalt, a wire is laid in the trench with several turns—usually 3 or 4—and the trench is sealed with a sealing compound. The wires are extended to a loop detector card, which is usually located in the traffic signal controller or motorway outstation. The size of the loop depends on the application. Often, the width is slightly smaller than the traffic lane and the length—in the direction of traffic flow—extends over a few meters.

Applying a voltage to the inductive loop creates a magnetic field, which is vertical within the loop as schematically shown in Figure 3. Any metallic object close to the loop will modify the strength and direction of this magnetic field because eddy currents are induced in the object that oppose the loop field according to Lenz's law. In most cases, an AC field is induced in the loop at a frequency in the range of 30–100 kHz. An electrical resonance circuit is used to drive the loop at the respective frequency and a vehicle is detected by measuring the deviation of the loop frequency from the initial value as shown in Figure 4. The inductance of the loop changes in the presence of vehicles and causes the frequency of the system to change—usually in the order of a few percents—which can be measured by the loop detector card.

Standard in the industry are multichannel detector cards, which can drive several—often four—loops. Most often, a multiplex mode is used, which drives and measures the loops sequentially in time intervals of several milliseconds. A time resolution of about 10 ms is sufficient for most

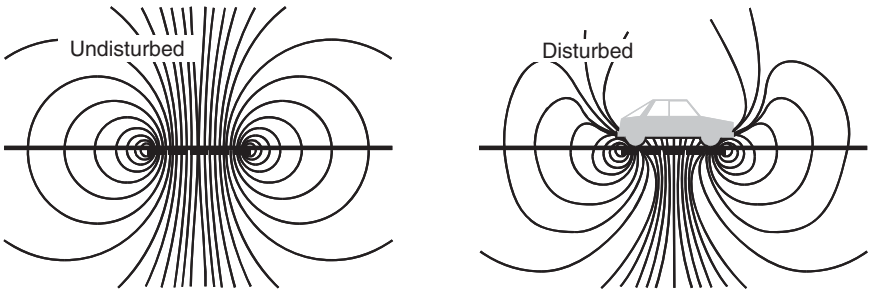


Figure 3. Magnetic field of an inductive loop detector and its change caused by vehicle presence.

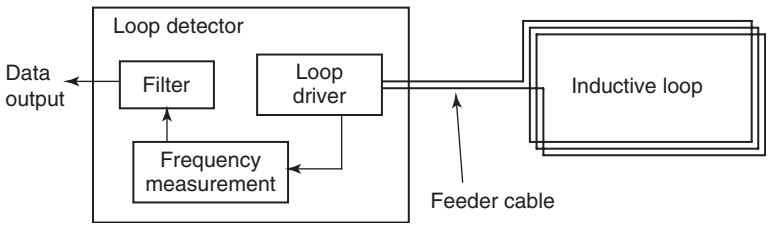


Figure 4. Functionality of an inductive loop detector.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

applications, as even a car at 200 km/h travels only about 55 cm in this time. The frequencies of all loops connected to one detector card can be set to the same value, as they are operated in a multiplex mode, which means that only one loop is activated at any time. In case more than one detector card is used, the loop frequencies need to be set to different values in order to prevent cross talk between the loops.

The loop frequency changes with temperature, which must be compensated for by the detector card. For moving traffic, this drift compensation can be implemented easily, as the frequency changes induced by vehicles are of the order of seconds, whereas the temperature drifts have a time constant of several hours. For the detection of vehicles in parking spaces where vehicles do not move for several hours or days, an advanced signal processing is necessary, which can distinguish clearly between the signal of the vehicle and the temperature drift.

Single inductive loops are very often used for vehicle presence detection at intersections. For additional data such as speed, vehicle length, and classification, double loops are used, where two loops are installed in the road with a spacing of one or a few meters in the direction of the traffic flow. The speed is calculated from the time delay of the signals a vehicle induces in both loops, and the vehicle length can be derived from the temporal length of the loop signal. Loop detectors achieve advanced vehicle classification with high accuracy by analyzing the inductive signature, which is the complete signal a vehicle induces in the loop. From the length and the shape of this signal, advanced software algorithms can derive accurate information about the vehicle type. Compared to other detection technologies in a similar price range, loop detectors have currently the best vehicle classification accuracy.

4.2 Radar

Radar, or microwave technology as it is sometimes called, is widely used in many application areas ranging from air traffic control to antimissile systems or speed enforcement in road traffic. Radar systems detect objects by emitting microwave radiation and detecting the part that is reflected from the objects. One of the main advantages of radar systems is the fact that water in any form, such as fog, rain, snow, or sleet, is transparent for microwave radiation and the weak signals of heavy rain can be filtered out easily in the detector. This means that radar systems are well suited for outdoor applications like vehicle detection in road traffic.

Several frequency bands are allocated for the license-free use of radar systems. In road traffic, 12 and 24 GHz are most frequently used. Internationally, they are specified by the ITU; however, most countries have additional regulations, which define the available frequencies and the maximum bandwidth and radiated power in these bands. In nearly all countries, radar detectors need to be tested and certified before they are allowed to be used in outdoor systems.

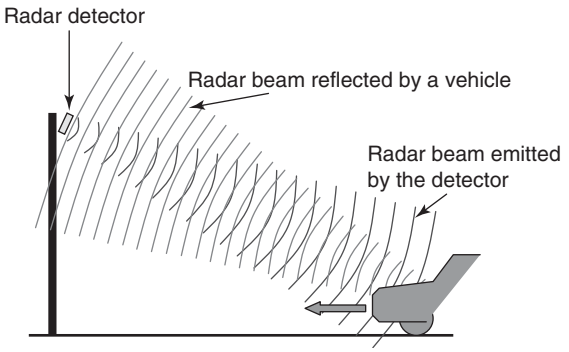


Figure 5. Principle of radar detection.

Most of the radar detectors used in road traffic use the Doppler radar technology, the principle of which is shown in Figure 5. It relies on the fact that moving objects change the frequency of microwaves, which means that the emission from the radar detector and the reflection from the vehicle have different frequencies. The effect is similar to the sound of cars, which has a higher pitch when they drive in our direction and a lower pitch when the drive away from us. Doppler radar is a quite mature technology that can be implemented on cost-effective hardware systems. The downside of Doppler radar detectors is their disability to detect stationary vehicles.

As the Doppler technology relies on the movement of the objects to be detected, stopped vehicles have a zero Doppler signal. The signal often drops below the noise level for vehicles traveling at a speed below 3–4 km/h. As many traffic signal controllers do not require the detection of stopped vehicles, but rely on the data of the moving cars, Doppler radar detectors are widely used for intersection control or traffic data acquisition.

The detection of stationary objects is possible with the FMCW (frequency-modulated continuous wave) technology. By modulating the frequency of the emitted radar beam and detecting the frequency shift of the reflected beam, the distance from the detector to the object can be determined. With FMCW radar detectors, it is, therefore, possible to detect all vehicles regardless of whether they stop in the detection zone or not, which allows a more advanced intersection control compared to the still more frequently used Doppler radar detectors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 6. Typical installation of radar detectors.

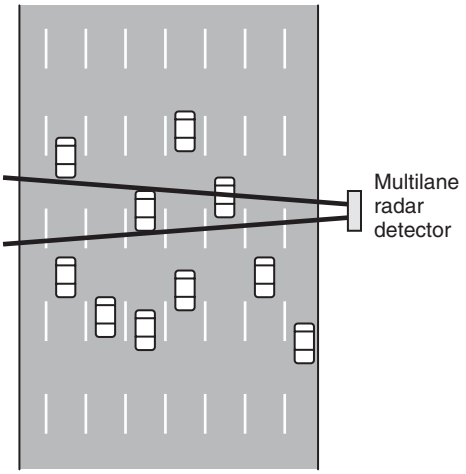


Figure 7. Example of a side-mounted multilane radar detector.

At intersections, radar detectors are one-spot detectors meaning that each detector covers one detection zone. They are installed on a pole or gantry above or on the side of the road (a typical example is shown in Figure 6) and are aligned in the direction to the location where vehicles shall be detected, usually close to the stop line or in the approach to the intersection. Depending on the detector type, they are lane selective or they cover multiple lanes. Vehicle detection at the stop line is often done lane selectively, especially when the right-turn, straight, and left-turn lanes have separate signals. For approach detection, multilane detectors can be used unless the vehicle actuation software in the controller requires a lane-selective detection.

Compared to the vehicle detection at intersections, a rather different technology is used for traffic data acquisition on interurban roads. Multilane radar detectors are mounted on the side of the road at a height of about 4 m or more and use a high functionality radar technology, which allows them to retrieve traffic data from multiple lanes (Figure 7). They are based on a two-beam technology: one radar system is used to generate two FMCW beams, such that there is a horizontal angle between them of around 10°. The detector is installed such that the two radar beams are directed almost perpendicular to the direction of traffic.

By measuring the distance of the radar reflection from the vehicles, the detector can determine on which lanes the vehicles drive. With the two radar beams, the time difference from the vehicle entering the first and the second beams can be measured; from this data, the speed of the vehicle can be calculated. The length of the vehicle in meters is derived from the duration of the signal in seconds and the calculated speed. Overall, the detector can measure all relevant traffic data: count, speed, vehicle length, and a classification based on the vehicle length. The accuracy might not reach the high level of inductive loop detectors but the ease and low cost of installation on a rather short pole besides the road without the need for a road closure are attractive advantages of these multilane radar detectors.

With advanced signal processing, even vehicles that are not in the direct line of sight can be detected. In case a truck on the first lane blocks a passenger car on one of the next lanes, the weak radar signal, which is redirected by the truck and which arrives at the detector, can be used for the detection of occluded vehicles.

Another application of radar detectors is based on their ability to measure the speed of objects with high accuracy and their rather low manufacturing cost. Radar detectors are often used in enforcement systems to measure vehicle speed with an accuracy which is able to stand up in court. Always when a ticket for exceeding the speed limit is generated, there is a good chance that the vehicle speed was measured by a radar detector.

4.3 Video detection

Video detectors for road traffic consist of a camera and a video image recognition system. A camera is directed to the area of the road where the vehicles shall be detected. It creates a video stream, which is fed into an automatic image recognition system where the images are analyzed

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 8. Image of a video detector with three detection zones.

in respect to the presence of vehicles. The software is designed to automatically find the vehicles in the video images and be able to detect their presence once they are in the designated zone. The locations where the vehicles shall be detected are defined during set up of the video detector by adding the detection zones as overlay on the video images. Figure 8 shows an example of a video image with three detection zones. Once a vehicle enters the detection zone, the software recognizes its presence and changes the output state for this zone to "occupied." The output returns to "not occupied" immediately after the vehicles left the zone. Depending on the type of detector, additional information such as vehicle speed, length, or classification are available.

The detectors need to work reliably at day and night, which requires the image recognition software to adjust to the outside lighting conditions. During daytime, it is mainly the contrast of the vehicle in respect to the road surface, which is used as a basis for the detection; however, at nighttimes, the detector looks for the headlights, as the body of the vehicle is dark and often difficult to recognize in the image. In most installations, the video detectors are directed opposite to the traffic flow and can, therefore, see the headlights of the vehicles. However, even in cases where the cameras look in the direction of the traffic flow, the tail lights are bright enough for the detection of the vehicles.

Even more important than the detection of a vehicle by the image recognition software is the suppression of disturbing effects. The normal lighting conditions of any outdoor scenery produce effects that are difficult to distinguish from vehicles by automatic systems. For instance, when the sun shines perpendicular to the direction of travel, each vehicle casts a shadow on the neighboring lane that has the size of a vehicle and has, under most conditions, a high contrast to the road surface in the sun light. For the image recognition software, it is quite difficult to distinguish this shadow from a real car. Other examples for disturbing effects are shadows from trees at the road side, which move in the wind, movement of the camera itself, or reflections of the headlights on the road surface at night especially when it is wet.

Many video detectors are used for presence detection at intersections, where they have a similar functionality like inductive loop detectors. However, they are also used for traffic data acquisition, where they measure speed and classification. One of the advantages of video detectors is that several detection zones can be placed in one detector. If the approach to an intersection has more than one lane, one video detector is able to detect the vehicles in all lanes separately. It often provides several outputs, one for each detection zone.

For some data types like vehicle length or speed the video detector needs to translate the coordinates in pixel space into real space coordinates. However, the length of an object in number of pixels in the image cannot directly be related to a real length in meters since they depend on the optical parameters of the camera. Most video detectors offer a semiautomatic calibration procedure during setup: a rectangle is overlaid on the video image and positioned over objects with known distance such as road markings. The size of the rectangle on the road is then input to the video detector and with additional information such as focal length of the lens and the size of the image sensor in the camera, the detector is able to translate pixel distances into real distances and measure speed and vehicle length.

Apart from vehicle presence detection and traffic data acquisition, video detectors are widely used for incident detection in tunnels and on highways. In order to increase the safety of tunnels or critical parts of highways, automatic incident detection systems based on video technology are used. Several cameras are installed at about equal distance, so that the complete stretch of the road is covered and the images of neighboring cameras overlap slightly. Like with the detectors for vehicle detection, an automatic image recognition software analyzes the video images and looks for possible incidents. The software mainly looks for stopped vehicles, as any incident or accident will result in one or more vehicles that are not able to move anymore. In free-flow traffic, it is rather easy to detect stopped vehicles; however, during congestions where all vehicles move at a low speed or stop, it is difficult to distinguish this traffic situation from an incident. Nevertheless, modern video systems have advanced algorithms, which achieve high detection rates for incidents in nearly all traffic situations.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

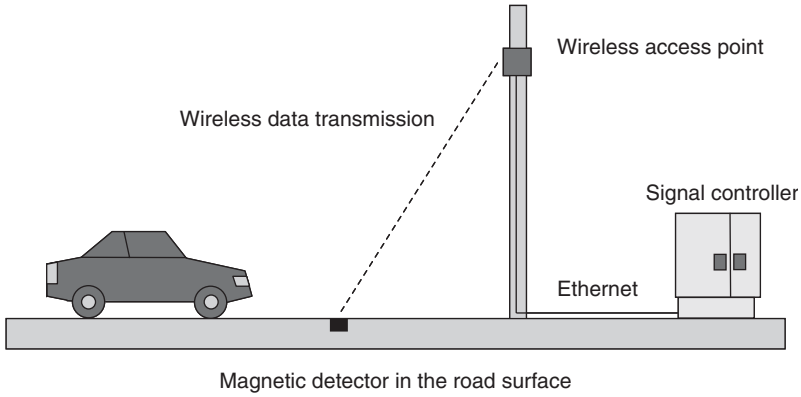


Figure 9. Schematic of a wireless magnetic detection system.

4.4 Magnetic field sensors

All vehicles or more generally all metallic or magnetic objects change the natural magnetic field of the Earth in their vicinity. Magnetic field sensors utilize this effect and measure the change of the magnetic field to detect vehicles. Most modern magnetic detectors are equipped with three magnetic sensors for the measurement of the field strength in all three spatial directions in order to detect all changes, which can possibly be caused by vehicles.

The strongest change in magnetic field can be observed close to the vehicles. With distance from the object, the signal amplitude decreases. The best results are achieved with detectors placed in the road surface, so that the vehicles drive directly above them. There are above ground detectors based on magnetic detection in the market but as the distance to the vehicles is rather large, their detection rate is limited.

Positioned in the road surface, magnetic detectors achieve high detection rates similar to inductive loop detectors. The sensitivity of the sensor is set such that the detection area has a diameter of about 2–3 m, which allows lane-selective detections and which means that only those vehicles are detected which drive on the respective lane. Vehicles on neighboring lanes are not detected.

One of the advantages of magnetic detectors is their small size and low power consumption. They can operate on battery power for 10 years or longer and combined with wireless data transmission, the detectors can be installed without any cables. This reduces the installation cost considerably and adds flexibility to the location of the detectors. A typical wireless magnetic detector system is shown in Figure 9, an example of the detector itself in Figure 10.

Similar to inductive loops, magnetic detectors detect only the vehicle presence. For additional data such as vehicle speed, length, and classification, two detectors are used, which are spaced within a distance of a few meters in the direction of traffic flow.

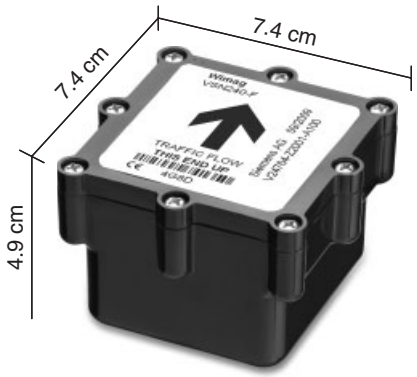


Figure 10. Example of a wireless magnetic detector.

Even though wireless magnetic detectors have not been long in the market, they are already widely used. One of the reasons is their reliability in all environmental conditions, as temperature, rain, snow, day, or night have hardly any effect on the magnetic signal.

4.5 Automatic number plate recognition

Automatic number plate reading cameras (ANPR cameras) are based on video technology, nevertheless they are described here in a separate section since they differ from the "standard" video detectors mentioned in Section 4.3. Many modern ANPR cameras are integrated devices including a video camera, illumination of the detection area (usually infrared LEDs), and a processing unit for the image analysis (Figure 11). The process of reading a license plate consists of several steps. The video camera generates

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 11. Example of an ANPR camera.

a continuous stream of images of the detection area. All images are analyzed by an image processing unit, which detects vehicles in the detection area. Once a vehicle is present in the respective location, a high resolution image is taken that is analyzed by the plate finding software, which looks for the location of the license plate in the image. This can be a difficult task when vehicles have more than one license plate, for instance, trucks, which often have private plates behind the wind screen. Once the license plate has been found in the image, the respective part of the image is transferred to the OCR (optical character recognition) module, which reads the plate. The challenge is to read the license plate under all conditions such as different angles depending on camera position, all lighting conditions, and the level of pollution on the plate. The result of the OCR is validated against the syntax of the plates meaning the position of the characters and numerals before it is released as the reading result. OCRs are usually optimized for plates of specific countries, as they are trained for the fonts used on the respective plates. The use in other countries often requires a new training of the OCR.

ANPR cameras are used in many applications ranging from tolling to travel time measurement or access control. The London congestion charge, for instance, is enforced with ANPR cameras within the congestion charge zone that read the number plates of all vehicles and check whether the car owner has paid the toll. If a license plate is found that is not registered as paid, the system issues a ticket to be sent to the vehicle owner.

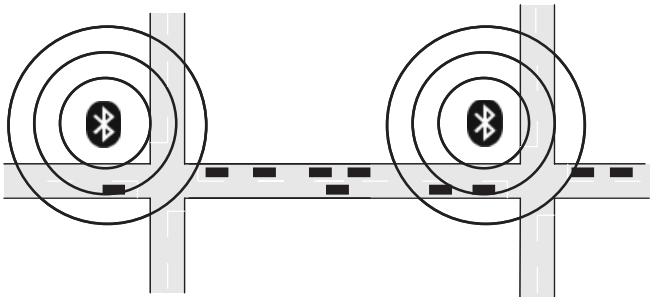


Figure 12. Bluetooth tracking system for travel time measurement.

Another application is the access control to company parking lots. All vehicles that are permitted to use the parking lot are registered in the system and when the ANPR camera reads one of these plates, the entrance barrier is opened and access to the parking lot is granted.

A rather new use case is section speed control for the enforcement of speed limits. At two locations within the speed limit zone, which are usually spaced by one or several kilometers, the license plates of the vehicles are read and sent to a central office with the respective time stamps. By subtracting the time stamps of the same plate reading at the two locations, the average speed of the vehicle can be calculated and in case it exceeded the local speed limit, the normal enforcement process is started.

4.6 Bluetooth tracking

Bluetooth tracking systems were developed with the objective to get travel time information in road traffic and to reduce the cost of the road side equipment. Travel time systems measure the average time of the vehicles driving between two points in the road network. Similar to floating car data, the travel time information is used for traffic management to determine the LOS.

Bluetooth tracking systems rely on the fact that many vehicles are equipped with mobile devices such as mobile phones, headsets, navigation systems, and handsfree systems. Bluetooth tracking systems are installed at the road side and scan the area for visible Bluetooth devices (Figure 12). Once they receive a reply, a time stamp is added to the MAC address of the device and sent to the central office. The traffic center looks for matches of the MAC addresses from different locations and can calculate the travel time from the respective time stamps.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

One of the main advantages of the Bluetooth travel time systems is the affordability and the ease of installation of the Bluetooth scanners at the road side.

The reidentification rate of Bluetooth scanners is very high: once a Bluetooth device has been detected, it can be reidentified at a second location with high certainty. The limits of the technology are related to the penetration rate of visible Bluetooth devices in the vehicles. In urban areas, around 20% are detected; however, on highways, the rate is often above 25%. This means that Bluetooth systems are well suited for travel time measurement on highways or on arterial roads with high traffic volume.

4.7 Floating car data

Conventional systems require detectors to be mounted at the road side or above the road to measure the traffic situation at this location. This means that the acquisition of traffic data requires investment in the infrastructure, as the detectors need to be purchased, installed, and maintained. A completely different approach is to use signals from the vehicles itself without any roadside equipment, which would reduce the cost of the data acquisition system considerably.

Some vehicles are equipped with navigation systems that always measure their location via GPS. In the floating car data system, these vehicles report their position via mobile data transmission (GPRS or UMTS) at regular intervals to the traffic centers, where this information can easily be used to derive the traffic situation, LOS, and actual travel times.

Even though the technology is available since many years, only a few systems have been implemented. The major advantage that no roadside equipment is needed is offset by the fact that the communication between the vehicles and the traffic center is not free of charge, which means that the operation of an FCD system incurs running costs. As they work best when many vehicles are included and the location information is transmitted in short intervals, a reduction in communication cost—which means lower number of vehicles or less frequent transmissions to the traffic center—degrades the performance of the system.

The commercial situation is less critical for vehicles that have already a wireless communication link to a central office. Taxis, for instance, are often equipped with on-board units for the communication to the taxi-office where the position information is used for an optimization of the fleet management. In these cases where the position information of the vehicles is available without additional cost, the information can be used as floating car data for the traffic management as well. Even though the data is available at low cost, the technology is limited by the fact that taxis often have different driving patterns compared to other vehicles. Taxis often know where congestions are and avoid this route, which means that the information about these traffic jams does not reach the traffic center.

Recently, some suppliers of car navigation systems introduced floating car systems, where some of the navigation devices report their position to a data center. The cost for the communication is paid for by the user who in return receives online traffic data and advanced routing based on the actual traffic situation.

Floating car data is used in some traffic management systems as additional data source to existing detectors. The accuracy of the data is currently limited by the low penetration rate of these systems, as only a small fraction of the vehicles are equipped with these devices. More detailed information about floating car data can be found in Road Traffic and Travel Information (RTTI).

4.8 Public transport priority systems

In order to increase the attractiveness of busses and streetcars, many operators want to give priority to these transport means, so that these vehicles move faster through the city and experience shorter travel times. Many larger cities have priority systems in operation, which accelerate the public transport traffic.

The basic technology of public transport priority systems is a communication between the bus or streetcar and the traffic signal controller. When approaching the intersection, the bus transmits a call to the traffic controller that adjusts the phases so that the bus or streetcar will experience no or only a small waiting time. The green phase for this approach will start earlier or will be extended depending on the approximate arrival time of the public transport vehicle at the signal.

The bus or streetcar needs to be equipped with a positioning system such as GPS or in older systems via infrared communication to roadside equipment. Once the position system indicates the correct location in the approach to the intersection, a wireless signal is sent to the controller. The main information are the current position of the bus relative to the intersection and an information whether it will turn right or left or go straight at the intersection. In many cases, the bus transmits its line number to the controller where a list exists with all information about the different bus lines. In many cases, there are bus stops close to the intersection, which have to be considered when designing the priority system at that intersection. The bus might send additional information to the controller when arriving at the bus stop or when opening and closing the doors. Once the bus has passed the intersection, it will send this information

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to the intersection controller in order for it to switch back to normal signal control.

Experience has shown that priority systems can shorten the travel time of public transport vehicles significantly. The actual reduction in travel time depends on many factors such as the number of signalized intersections on the route, overall traffic volume, and the flexibility of the signal control (upper and lower limits for the length of the phases).

There are many different technical implementations of priority systems. Modern systems rely on GPS positioning of the transport vehicle and use wireless communication from the vehicle to the intersection.

4.9 Other detection technologies

Many different technologies are used for vehicle detection and within the scope of this chapter; only the most frequently used ones can be described. Passive infrared, active infrared (laser), ultrasound, or acoustic detection are further examples of technologies used in vehicle detectors.

The accuracy of the traffic data can be increased by the use of several different sensor technologies in one detector. Combination detectors, for instance, with infrared, ultrasound, and radar sensors are often used as above ground detectors on highways. They are mounted on gantries above the road and achieve advanced vehicle classification with high accuracy.

GLOSSARY

Occupancy	Occupancy defines the percentage of time a certain point of the road is occupied by a vehicle.
Level of service	The LOS is a measure of the current traffic situation of a certain part of the road network. In many cases, several LOS levels are defined such as free-flow traffic, slow traffic, congestion, or stationary traffic.
Inductive loop detectors	Inductive loop detector systems consist of a wire loop in the road and a electronic unit that induces an alternating electromagnetic field in the loop and measures the electromagnetic field induces by metallic objects such as vehicles in the vicinity of the loop. Inductive loop detectors are used for vehicle detection.
Vehicle actuation	Vehicle actuation describes the use of traffic data for the calculation of the signal phases at a signalized intersection.
Traffic volume	Traffic volume is defined by the number of vehicles per unit time (usually 1 h) crossing a section of road.
Traveltime	Traveltime measures the time it takes a vehicle to travel from point A to B, where A and B are two points in a road network.
Radar	Radar is an object-detection system that uses radio waves. Radar is the acronym for radio detection and ranging.
Bluetooth	Bluetooth is a wireless technology standard for data transmission over short distance in the ISM band from 2400–2480 MHz. It is defined in the IEEE 802.15.1.

REFERENCES

Gajda, J., Sroka, R., Stencel, M., *et al*. (2001) A Vehicle Classification Based on Inductive Loop Detectors. *Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference*, 2001. IMTC 2001. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=928860 (accessed 4 July 2013).

Markovic, A., Kapkovic, J., Luehrs, C.-H., *et al*. (1996) "Geschwindigkeitsmessungen im strassenverkehr," Physikalisch-Technische Bundesanstalt, Expert Verlag Stuttgart, SRN, Tech. Rep.

Ryus Paul, P.E. *Highway Capacity Manual 2010*. Transportation Research Board. http://onlinepubs.trb.org/onlinepubs/trnews/trnews273HCM2010.pdf (accessed 15 January 2012).

Scoot SCOOT—*The World's Leading Adaptive Traffic Control System*. http://www.scoot-utc.com/ (accessed 4 July 2013).

VMZ Berlin Betreibergesellschaft mbH Tempelhofer Damm 1–7, 12101 Berlin, Germany.

FURTHER READING

University of Idaho *Traffic Flow Theory, Theory & Concepts*, http://www.webs1.uidaho.edu/niatt_labmanual/Chapters/trafficflowtheory/theoryandconcepts/TrafficFlowParameters.htm.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto176
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Data Fusion

Axel Leonhardt[1] and Álvaro Catalá-Prat[2]

[1]*PTV AG, Karlsruhe, Germany*
[2]*DLR, Braunschweig, Germany*

1 Introduction 1
2 Data Fusion Methods 2
3 Data Fusion for Object Recognition and Tracking with Camera and Laser Scanner 6
4 Data Fusion for Traffic State Estimation and Prediction 9
5 Summary 11
Glossary 13
References 13

1 INTRODUCTION

With the recent developments of intelligent transportation systems and intelligent vehicles, data fusion gains more and more attention as a technique to make optimal use of a wide range of in-vehicle and roadside sensors. Applications include various approaches for object and incident detection, state estimation, and state forecasting.

This chapter gives an overview of data fusion methods and applications from the vehicular as well as the roadside perspectives. It also gives an introduction of the main concepts of data fusion. Section 2 describes the commonly used methods. The main part consists of two sample applications, one detailed example from the area of vehicle automation (Section 3) and the other example from the area of traffic state estimation and prediction, which is described in Section 4. Finally, in Section 5, a summary and an outlook on sensor data fusion is given.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.

This article is © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

1.1 What is data fusion?

The term *data fusion* is widely used in the industry and there exist a wide range of definitions. In general, data fusion can be described as the intelligent processing and combination of information from many diverse sources. In particular, data fusion is applied to gather information by combining different sources that is of higher value than information from only one source. If only sensors are used as sources, often the term *sensor data fusion* is used.

There are different occurrences of data fusion.

Cooperative data fusion is applied on independent data sources to get information that would not be available from the single sensor. For example, using two 2D camera images, a stereo 3D image can be generated. Moreover, by the combination of two sensor perspectives, the whole object size can be detected. Another example is the derivation of vehicle densities on a stretch of road with upstream and downstream detectors.

Complementary data fusion means that data from different sources are combined in order to give a more complete picture of the observed phenomenon. For example, the front area and rear area sensors of a vehicle can be fused to an all-around detection system.

Competitive data fusion means that in a multisensor setting, each sensor delivers independent measurements of the same characteristic.

Data fusion is applied to reach one or several of the following advantages:

- Extension of the detection area: The detection area of two or more sensors can be fused into a common one.

- Extension of detectable features (see examples of cooperative data fusion earlier).
- Improvement of accuracy: By the fusion of two data sources with known accuracy, a better common accuracy can be reached. For example, by fusing many samples of a quantity, a better estimation of the quantity is possible.
- Increasing confidence: The combination of two detection sources can provide more confident detection rates. For example, the use of different data sources can help reduce false positive and false negative rates of an object detection system. A further example is the application of several incident detection algorithms in parallel.
- Higher availability: By means of sensor data fusion, the effects of temporary limitations of acquisition systems can be reduced. For example, a multisensor fusion system may be able to provide output even if one of the sensors is failing.
- Increased robustness: The fusion system is more able to cope with errors during execution, for example, errors due to data outliers, data noise, or sensor failure.

1.2 Sensor data fusion levels

Commonly, data fusion systems are based on one of the following abstraction levels:

Sensor level or low level fusion: The raw data of the sensors is processed closely together (usually in a cooperative way). These systems can reach very high accuracy and reliability. However, in case of failure of one of the sensors, the whole system may fail. The calculation of stereo-information and the projection of radar data on a camera image for target confirmation are examples of low level fusion.

Object level or mid-level fusion: In this case, the data of each sensor is preprocessed separately before it is fused (usually in a competitive way). The aim of the preprocessing is to reduce the amount of data to the features or objects to be fused, getting thus more abstract data. On the one hand, this means that a higher independence of the sensors used. Moreover, the system can usually reach high reliability and availability because it does not depend on all of the sensors to work. In case that one of the sensors has a short failure or malfunction (e.g., the camera is glared), the fusion system is still able to provide output based on the other one. The system offers high availability and data rate as well, which comes from the sum of both sensor rates.

Depending on the abstraction level of preprocessed data, different sublevels are defined:

1. feature level (image features, such as edges, can be fused with laser scanner features, such as reflections);
2. object level (object observations extracted from different sensors are fused to object hypotheses);
3. tracking level (each sensor carries out its own tracking and the tracks are then associated and fused).

Information level or high level fusion: This term is used to define data fusion systems in which the involved input sources already contain a complete data processing. These systems usually combine semantic information about the observed objects or scene. For example, in relation to a driving situation, information about vehicle properties, maneuvers, intentions, or dangers can be obtained.

1.3 Spatial and temporal scopes

For the system design, it is important to know the spatial and temporal scopes of the data fusion configuration. Spatial scope here means the location of the sensors. How the sensors are distributed (one detector station or several detectors along a road, sensors in one vehicle or in several vehicles) determines the requirements for communication channels. The temporal scope describes if only data from the current interval are used or also data from the past (last 15 min, past 2 years). This has an impact on the system design in terms of data storage capabilities.

2 DATA FUSION METHODS

Many data fusion methodologies have been proposed in the literature, including statistical approaches, artificial intelligence, and filter approaches. This section gives a brief overview of the commonly applied methods.

2.1 Logic-based data fusion

Logic-based methods, as described here, are rather straightforward and easy-to-implement data fusion methods. They are basically a set of IF, AND, and OR statements to compare and evaluate information from different data sources. In logic-based data fusion, there are rules and reasoning statements used to combine and utilize different information sources, for example, in the following form:

Rule: IF $X > a$ *AND* $Y < b$ *THEN* $Z > c$
Reasoning: IF $X > a$ *AND* $Y < b$ *AND* $Z < c \rightarrow Z$ *is implausible*

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where X, Y, and Z are measurements from different sensors and a, b, and c are thresholds of observables, for example, derived based on fundamental physical principles or based on a range of (plausible) observations.

In the field of traffic engineering, these rules can be applied, for example, to estimate whether detector data is plausible or not. For practical application, rule-based error checking methods for detectors have been implemented in national guidelines; for example, the German guidelines for the quality management of traffic detectors (FGSV, 2006) and the German guidelines for quality management of road weather and road surface condition sensors (FGSV, 2010).

An example for a plausibility check is shown in Figure 1, where data from a road visibility sensor and relative air humidity sensor are fused to identify implausible (or at least suspect) visibility measurements. The reasoning rule is simply *IF Visibility < Visibility_Threshold AND RelativeHumidity < RelativeHumidity_Threshold → Visibility is implausible*. The reasoning is based on two phenomena: (i) the physical or meteorological principle that a low visibility (fog) can be observed only if relative humidity is high and (ii) the experience that the relative air humidity measurement is in general much more reliable than a visibility measurement.

Fuzzy logic (or fuzzy reasoning) generalizes the traditional logic concept and allows not only binary outcomes (false/true) but also "degrees of truth." Fuzzy Logic is rather popular for data fusion, as it naturally allows to deal with incomplete information (e.g., data from one source is missing). In the data fusion context, the degrees of truth can be utilized, for example, to express the extent to which the different sources support the presence of a certain system state. The final state estimation is done based on a set of fuzzy rules (inference) and the so-called defuzzyfication process. An example is the automatic incident detection system proposed by Busch and Ghio (1994), which fuses different incident detection algorithms to get a more reliable incident detection.

2.2 Data mining methods

In the context of data fusion, data mining can be regarded as a group of methods that automatically learn input–output relationships based on training data. Hence, data mining deals with large collections of data and its main objectives are to automatically detect and process patterns. For detailed introductions, the reader is referred to the respective textbooks, such as Hastie, Tibshirani, and Friedman (2001), Sankar and Pabitra (2004), Takezawa (2005), and Berthold and Hand (1999).

Regression analysis as a subarea of pattern recognition deals with the estimation and prediction of numerical data and is widely used to facilitate sensor data fusion. Observations (training data) are used to learn functional relationships between an input pattern (independent variables) X, with $X = x_1, \ldots, x_n$, and an output (dependent variable) y. The goal of a regression model is to find a suitable function or a model that solves $y' = f(X)$, given a set

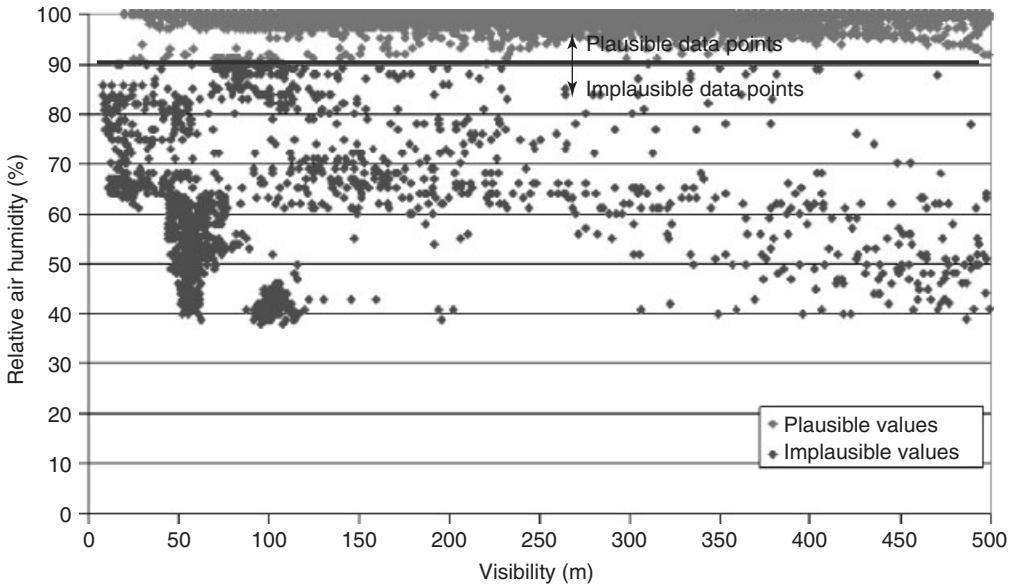


Figure 1. Rule-based assessment of the plausibility of road weather data (Dinkel, Leonhardt, and Piszczek, 2008). (Reproduced by permission of SIRWEC and Axel Leonhardt.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of observations (y, X). Models frequently used for regression are parametric regression methods, artificial neural networks (ANNs), adaptive expert systems, and nonparametric regression.

2.2.1 Parametric regression models

Parametric regression models describe input–output relationships by polynomial functions, with the polynomial coefficients to be estimated based on training data. A simple example is the well-known linear regression, where two coefficients have to be estimated. Parametric regression is applied to fuse data sources in the field of traffic state estimation by Kwon and Petty (2005). They apply a varying coefficient linear regression model to predict travel times on freeways segments based on local traffic data and an estimate for the current travel time. Sun *et al*. (2003) describe a method to predict travel times on freeways using a local linear regression model using past and the current observations of the variable to be predicted (travel time). Maier (2010) applies a segmented linear regression approach to estimate travel times observed with probe vehicles and local detector data.

2.2.2 Artificial neural networks (ANNs)

ANNs are flexible function approximators that consist of simple processing units (neurons) and emulate a human brain's learning and decision-making processes. In particular, their structure allows the application of efficient learning algorithms that optimize the connections between the neurons to map an input pattern to an output.

ANN are used extensively for data fusion in traffic engineering, especially for traffic state estimation and short-term prediction, in particular, as they are well suited to approximate highly nonlinear relationships between input and output patterns. For example, Park and Rilett (1998) propose a two-step approach. In the first step, historical link travel times are segmented using unsupervised learning. In the second step, one ANN is calibrated for each class, forming a system of situation-specific ANN. Dailey, Harn, and Lin (1996) give a comprehensive overview over ANNs and their applications in the area of intelligent transportation systems.

2.2.3 Adaptive expert systems

Adaptive expert systems are hybrid methods that incorporate expert knowledge in the form of prototypical rules, examples, or decision trees. These preformulated rules are optimized based on training data. Linauer *et al*. (2006) use an ANFIS (adaptive neuro-fuzzy inference system) for the prediction of travel speeds on routes in urban networks. Speed and traffic volume of one local detector on the respective route are used to predict the travel speed; training data to optimize the fuzzy membership function are derived from taxi probe data.

2.2.4 Nonparametric regression models

Instance-based learning (or nonparametric regression/k-nearest-neighbor methods) is a data mining method to solve regression problems, where the output variable is estimated based on the most relevant observations. It basically compares a present pattern X with its counterparts in the database with respect to a defined measure of similarity (distance metric) and calculates y' online based on the observed y that correspond to the k most similar patterns in the database.

2.3 Bayesian data fusion

Bayesian data fusion is based on Bayes' theorem, which basically allows deriving the probabilities of events given some a priori knowledge. The joint probability P of two events A and B occurring is the probability of event A, given that event B already has occurred:

$$P(A,B) = P(A|B) \cdot P(B) \tag{1}$$

Bayes' rule now states that the probability of event A given that event B already occurred is:

$$P(A|B) = P(B|A) \cdot \frac{P(A)}{P(B)} \tag{2}$$

or in case there are several events A_i,

$$P(A_i|B) = P(B|A_i) \cdot \frac{P(A_i)}{\mathrm{SUM}(P(B|A_i) \cdot P(A_i))} \tag{3}$$

It becomes clear that there is some a priori knowledge needed to successfully apply Bayes' rule. It is necessary to know $P(B|A)$ or $P(B|A_i)$ and the a priori probabilities of events $P(A_i)$.

One example for the application of Bayes' fusion is the classification of vehicles. There are two sensors (e.g., a radar sensor to measure the speed and an ultrasonic sensor to sense the vehicle's signature) that deliver signals, which can be used for vehicle classification (e.g., two classes: passenger car and passenger car with trailer). Furthermore, the a priori probabilities of a vehicle being in a certain class (e.g., based on a known fleet distribution) are given, as well as the individual probabilities that a signal belongs

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to a certain class. On the basis of initial probabilities and the individual probabilities, the probability that the vehicle is, for example, a passenger car can be estimated with higher certainty than based on one of the individual probabilities.

Further examples and applications can be found in the literature, for example, in El Faouzi (2006), for travel time estimation and Koks and Challa (2005) for sensor tracking.

2.4 Dempster–Shafer theory of evidence

The Dempster–Shafer theory can be understood as a generalization of the Bayesian method, as it employs a confidence interval of certainty instead of the single-point probability of the Bayesian method. Thus, it offers a way to combine uncertain information from different data sources and each sensor can contribute at a different level of detail, as described by Sarma and Raju (1991). Especially, Dempster–Shafer theory introduces two new concepts: support and plausibility. Support can be interpreted as a lower limit of uncertainty, whereas plausibility can be regarded as the upper limit of uncertainty.

Examples and applications can be found in the literature, for example, in El Faouzi (2006) and Dailey, Harn, and Lin (1996).

2.5 Filtering methods

The ground principles of data fusion can often be found in the area of estimation and tracking. In relation to these topics, a large amount of literature can be found, such as Bar-Shalom and Li (1995), Brooks and Iyengar (1998), Brown and Hwang (1997), and Maybeck (1979). This section aims at giving a short overview of theory and methods of estimation and tracking that are frequently used in the literature for the purpose of data fusion.

In automotive area, it often comes to estimating the state of a dynamic object, such as the position of the ego-car, or the position of other vehicles or pedestrians, which is often referred to as *tracking*. The input for estimation is given in the form of observations or measurements, coming from sensors or other data sources.

Many different approaches and methods can be applied in the task of tracking. Most of them are based on the concepts of linear regression (Section 2.2) and linear Bayes' estimation (Section 2.3), which are enumerated in the following.

State vector $x(t)$—containing all relevant information to estimate and from which all future states (predictions) can be derived. The system engineer decides the dynamic model for the tracking.

Error covariance matrix $P(t)$—representing the estimation accuracy or prediction accuracy as well as the correlation between its components.

Transition matrix $F(t)$—expressing the relationship between two consecutive states in time. It is defined according to the dynamic model as well.

Control input $u(t)$ and control input matrix $B(t)$—meaning an additional known input in the tracked system.

Process noise $w(t)$ and process noise matrix $Q(t)$—representing dynamic factors nonconsidered by the dynamic model. It is assumed to be white noise, that is, a random process with expected value 0 and autocovariance matrix $Q(t)$.

The relationship between the two state vectors in different instants is given as:

$$x(t_{k+1}) = F(t_{k+1}) \cdot x(t_k) + B(t_{k+1}) \cdot u(t_{k+1}) + w(t_{k+1}) \tag{4}$$

Observation or measurement vector $z(t)$—containing the data provided by the sensor or the data source.

Observation or measurement matrix $H(t)$—representing the linear relationship between the state vector and a corresponding observation.

Measurement noise $v(t)$ and measurement noise matrix $R(t)$—representing the accuracy of the measurement vector (i.e., the current accuracy of the sensor or the extraction system). It is assumed to be white noise too, characterized by the autocovariance matrix $R(t)$.

The relationship between a state vector and the corresponding observation is given by the measurement equation:

$$x(t_k) = H(t_k) \cdot x(t_k) + v(t_k) \tag{5}$$

A recursive, optimized estimator based on these principles is the well-known Kalman filter. The Kalman filter consists of two cyclical phases. In each filter step, the state vector is predicted to the time of the new measurement t_{k+1} and then it is updated by means of this measurement. A detailed description and its mathematical functions can be found in Kalman (1960).

Other nonlinear (and nonoptimal) solutions exist, such as the extended Kalman filter or the unscented Kalman filter. Another possibility for nonlinear systems consists of the particle filter, based on a Monte Carlo sampling of the estimate.

The pure estimation is usually not enough in order to get a proper tracking. There are multiple aspects, such as special noise, maneuvering targets, or data outliers, which make it necessary to define additional mechanisms for a robust tracking. Such mechanisms may include adaptive process noise (fudge factor), input estimation, variable state dimension, multimodel filter, and interacting multimodel

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

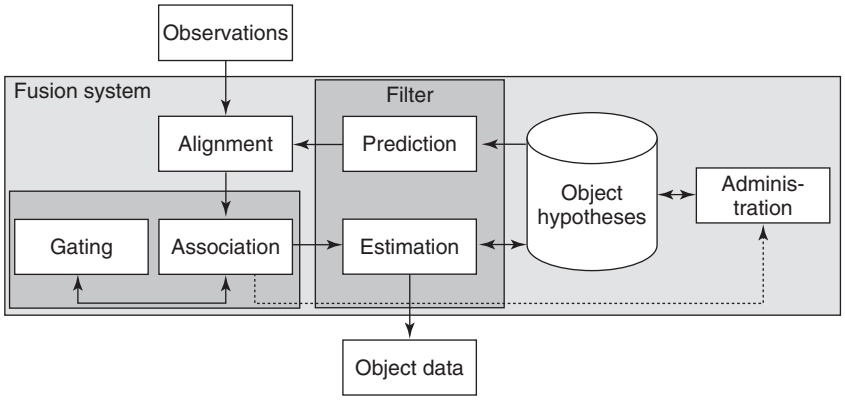


Figure 2. Multitarget tracking scheme.

(IMM). See the literature sources given at the beginning of this section for more details.

2.6 Multitarget-tracking

Object detection and tracking systems are frequently based on the well-known multitarget tracking scheme (Bar-Shalom and Li, 1995), as shown in Figure 2. This cyclical model has a list of observations as input at each cycle.

In a first step, the state of the object hypotheses already existing in the system is predicted to the current time-stamp of the observations according to the dynamic model.

Then, the incoming observations are aligned with the list of predicted objects. Alignment usually affects synchronization and coordinate transformation, in order to get the data in a common time axis and reference coordinate system.

After alignment, the association between observations and predictions takes place. For this step, a metric must be defined in order to evaluate each combination between an observation and a prediction. For example, direct distance, Mahalanobis distance, or similarity metrics are often applied. To decide the associations, different approaches can be used, such as the nearest-neighbor, the probabilistic data association filter (PDAF), or variations of them. The basics of these methods are common to the ones given in Sections 2.2, 2.3, and 2.4. In parallel to association, statistically or physically improbable combinations are excluded by a gating mechanism.

On the basis of associations set, a new estimation step takes place for each object hypothesis, for example, by means of the Kalman filter.

Additionally, some administrative tasks are carried out on the object list, such as initializing, deleting, and managing object hypotheses.

3 DATA FUSION FOR OBJECT RECOGNITION AND TRACKING WITH CAMERA AND LASER SCANNER

This section presents an application of sensor data fusion in the area of vehicular sensor data to sketch a more precise picture of sensor data fusion in practice. It consists of an object detection and tracking system based on a camera and a laser scanner.

3.1 Introduction

Object detection and tracking is a crucial component of advanced driver assistance and automation systems, such as adaptive cruise control (ACC), brake assistant, collision avoidance system, blind spot assistant, or maneuver planner. For the task of object detection, high reliability, availability, and accuracy are required. In addition, further important challenges appear, as schematically shown in Figure 3 and listed as follows:

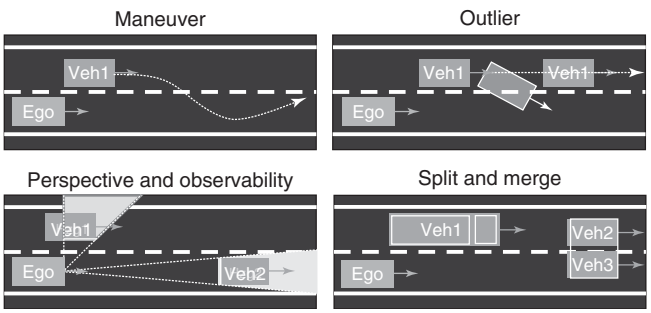


Figure 3. Selected challenges at object detection and tracking.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- uncertain sensor data and data outliers;
- maneuvering objects, whose movement differs from the expected dynamic model;
- partial observability due to sensor characteristics, perspective changes, and objects covering each other;
- splitting and merging of observations.

This example system, presented as well by Catalá-Prat and Köster (2011), provides a compact solution to address object detection and tracking, with special focus on the challenging situations mentioned earlier.

Concretely, it means that the system shall be able to detect objects near the ego vehicle. Their position shall be determined accurately and independently of the objects' constellation. In case of object occlusion, the observable parts of the object shall be modeled correctly. Moreover, objects shall be tracked in the course of time, irrespective of their movement. The tracking system must be able to follow maneuvering objects and deal with data outliers as well as split and merge effects.

All in all, high accuracy, reliability, availability, and robustness are desired. The system shall be kept modular, expandable, and sensor independent.

3.2 State of the art

Object detection and tracking systems are frequently aimed in the automotive area. These solutions can be classified according to their level of abstraction.

Aiming at a high performance, many systems are based on the sensor level fusion. An example of such systems is given by Kato, Ninomiya, Masaki (2002), who combine a radar and a camera by projecting the radar targets into the camera image. Another example is given by Vu, Aycard, and Appenrodt (2007), who apply an occupancy grid as a base for alignment and association between the different sensors.

Further solutions are based on the feature level. For example, Kaempchen *et al*. (2004) extract features from a camera and a laser scanner and use them as input to the fusion (feature level fusion). There, the alignment between observations and predictions requires a transformation corresponding to the sensor model. For tracking, they propose an IMM filter at object level.

At object level, Weiss, Stüker, and Kirchner (2003) use a model-switching mechanism instead. Stüker (2004) extracts object observations from each sensor, which are used as input for the fusion system. Moreover, he also explicitly considers the objects' partial observability and the adaptive measurement noise.

Finally, Thomaidis, Spinoulas, and Lytrivis (2010) fuse at tracking level, which means that objects are tracked for each sensor separately and are then integrated.

3.3 Approach

The modular structure of the proposed solution is shown in Figure 4. The central fusion takes place at object level and uses the scheme represented in Figure 2. The advantages of

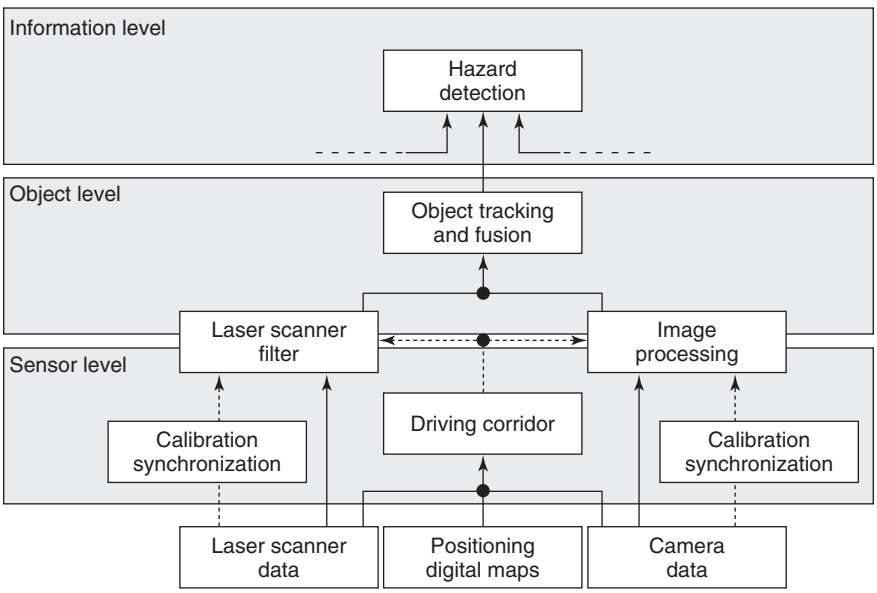


Figure 4. Proposed solution.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

object level fusion for this application are its competitive character, the higher independence from the sensors used, and the higher reliability and availability.

In order to feed the object fusion, object observations are extracted from the data of each sensor. Preliminarily, the sensors must have been accurately calibrated, referenced, and synchronized.

Camera image and laser scanner preprocessing are not in the focus of this chapter. Some details can be found in Catalá-Prat, Köster, and Reulke (2010). The extraction of object observations out of images is done based on texture-based image segmentation. Laser scanner preprocessing for object extraction is done by a fast shape indicator based on the internal angles of the detected object contours. Figure 5 shows an example of a situation with camera and laser scanner observations.

In order to filter and reduce the amount of data to fuse, the boundaries of the current driving corridor are determined and used as area of interest in the data processing. The driving corridor is understood as the road in which the vehicle is driving and other immediate neighboring vehicles can be found (e.g., one motorway direction, a road, or even a whole intersection). Determining the driving corridor is done by means of an early sensor data fusion (based on an occupancy grid) of digital map data, lane information from the camera, and raw laser scanner data.

At object level, the observations of both sensors are fused in a multitarget tracking system. As outcome, a robust list of detected and tracked objects is provided.

A central aspect of the multitarget tracking system is the object model, on which both the object extraction modules and the object fusion itself (and its output) are based. In this example application, the object model consists of a box, represented by its size, heading angle, and a reference point. The reference point, described by position and velocity, can be changed depending on observability issues. The chosen dynamic model consists of a straight constant velocity movement, which corresponds to a simple model and allows quick initialization and convergence of tracking. As central estimation filter, the information filter is applied (Rao, Durrant-Whyte, and Sheen, 1993). This filter is equivalent to the Kalman filter (Section 2.5) and allows parallel input data from different sources at a time.

In order to allow a robust object detection and tracking, the following fusion strategies have been developed:

- adaptive measurement and process noise, to simulate nonconsidered acceleration or yaw rate;
- changing logical reference, as mentioned earlier, in order to always represent an object by the optimally observable reference point (Figure 6);
- partial observability matrix, in order to only consider observable parts in the estimation;
- multiple association, to deal with split or merged observations;
- object hypothesis duplication by unclear association;
- object hypothesis unification by high overlap.

The two latter strategies are well suited to deal with different challenges such as split or merged object hypotheses, with maneuvering objects and data outliers (Figure 7).

Finally, after object detection and tracking, further modules use this outcome at information level. In this framework, an early hazard detection approach has been developed. The approach is based on the detection of atypical events or situations, which possibly can represent hazards. However, the information level is not at the aim of this publication.

3.4 Results

The presented methods and strategies have been exhaustively tested in different manners. On the one hand, in

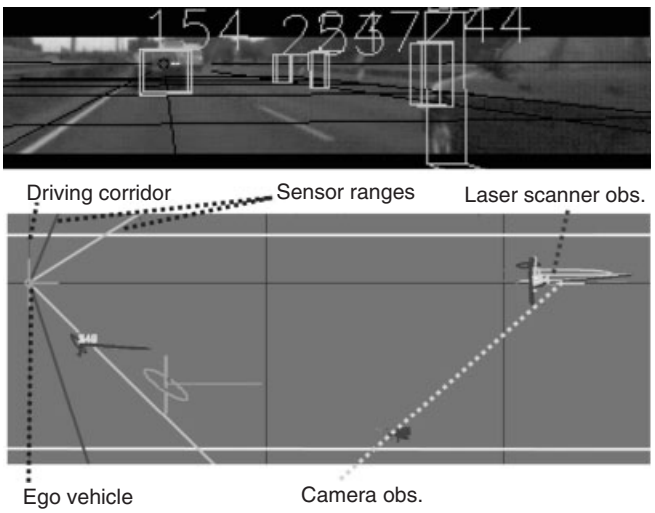


Figure 5. Examples of observations extracted from the camera and the laser scanner.

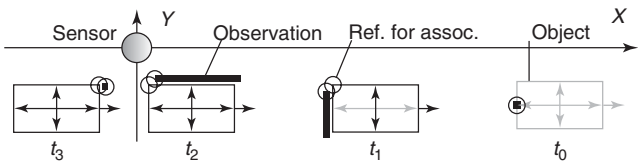


Figure 6. Example of changing object reference in the course of time.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

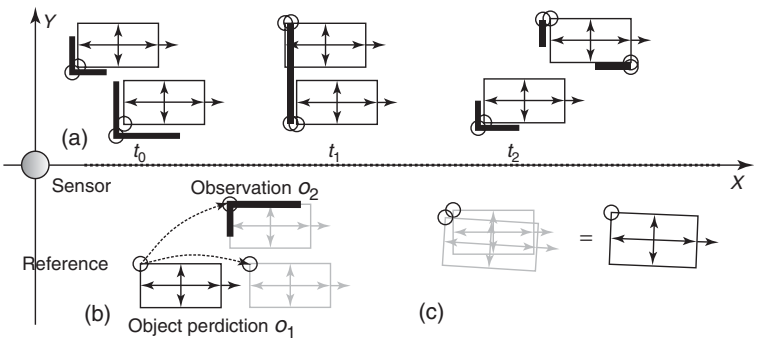


Figure 7. (a–c) Examples for avoiding split and merge of objects (see details in the text).

order to evaluate the methods, a simulation framework has been used. The advantages are the knowledge about the ground truth, the repeatability, and the testing of the system in extreme situations. The experiments consisted of motorway scenarios and contained different maneuvers, such as approaching, lane change, and overtaking. On the other hand, in order to validate the simulated results, real test drives have also been carried out in motorway scenarios with a test vehicle equipped with a camera and laser scanner sensors.

Comparing to the raw data, the object tracking has shown very promising results, such as a higher system availability, higher accuracy, lower false positive rates due to high confident tracking, and low false negative rates.

To sum up, the results show that objects were initialized very quickly, and that a smooth and uninterrupted tracking was possible because of the high availability of both sensors. Moreover, maneuvering objects, outliers, and split and merge did not have major negative effects on the object fusion and tracking. The focus of these first experiments has been set on the quality of the results. Thus, not all system components have been optimized for real-time yet. However, the structure of the strategies allows for a straightforward optimization.

4 DATA FUSION FOR TRAFFIC STATE ESTIMATION AND PREDICTION

In addition to the vehicular application presented in Section 3, this section sketches an instance-based learning method for the estimation and short-term prediction of traffic variables using spatiotemporal traffic patterns from multiple data sources. In the following, it is described how the methodology is applied for travel time estimation. Yet, the approach is generic and can be applied to other input–output data sets as well.

4.1 Introduction

Typically, in any urban network that exceeds a certain size and complexity, there are data from local traffic detectors available. The richness of the data varies, depending on the technology used, with at least traffic flow and occupancy (vehicle presence) being available.

Methods to measure travel times became more and more available in recent years. With the evolution of positioning devices and advances in wireless communication techniques, the use of vehicle positioning data has gained more and more attention. Travel time can be derived based on vehicle position and time stamp by applying some map matching and path reconstruction methods. Taxis are a particularly useful data source. They are equipped with GPS (global positioning system) devices for disposition and transmit their positioning data regularly to a central control system, where they can be accessed.

Today's IT infrastructures allow to gather and store huge amounts of data—and this holds true for traffic data as well. Given a setup with several local measurement devices and at least occasionally observed travel times, this can be used as a training data basis to derive a model that is able, based on the typically continuously available data, to estimate and predict the occasionally observable variable. Data mining methods are particularly suitable to be used for such a task (see Section 2.2 for an overview of different data mining methods).

4.2 State of the art

Friedrich, Matschke, and Heinig (2004) present a holistic approach on the integration of floating car data (FCD) in traffic state estimation (based on detector data) on different network levels (Figure 8). In particular, probe vehicle data are used to get samples for queue length at an intersection level, travel times on a link/route level, and origin–destination (OD) patterns on a network level.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

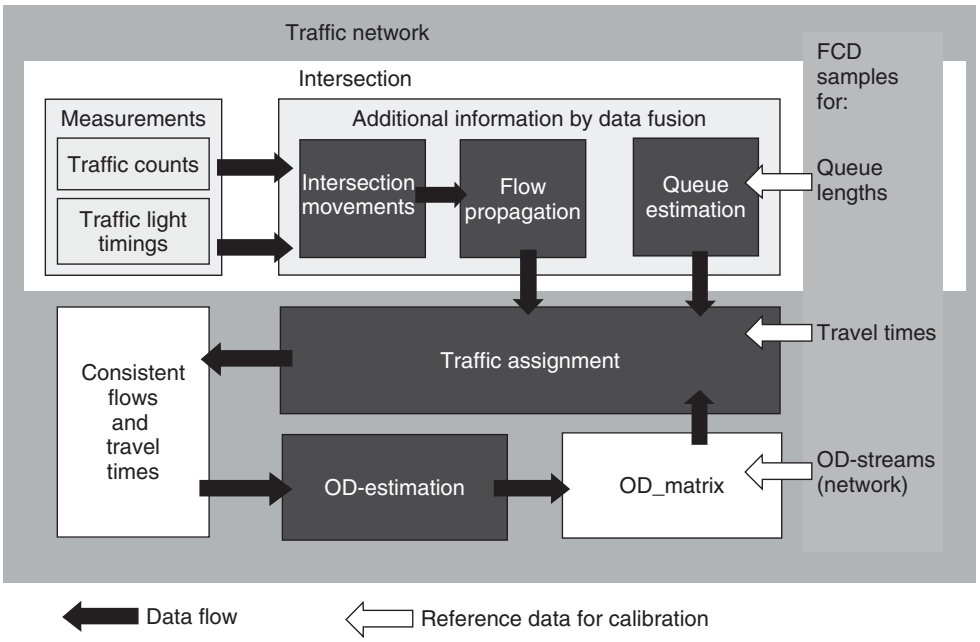


Figure 8. Data fusion in the context of traffic state estimation. (Reproduced with permission from Friedrich *et al*. 2004. © TRISTAN V).

Several approaches apply pattern recognition techniques to fuse data from different sources: Robinson and Polak (2005) propose a method to estimate travel times using *k*-nearest neighbors. Travel times are gathered from license plate readers, local occupancy and volume data from detectors on the respective routes are used as independent variables. The method has been applied to two routes and outperformed several benchmark algorithms, including simple regression models and ANN. Turochy and Pierce (2004) investigated *k*-nearest neighbors to predict local speeds on freeways for one time step. Davis and Nihan (1991) developed an algorithm to predict occupancy and volume based on occupancy and volume from several surrounding detectors from the previous interval. Smith and Demetsky (1997) found that the instance-based learning method they used outperformed an ARIMA (autoregressive integrated moving average) and an ANN for the prediction of traffic volumes. You and Kim (2000) use probe vehicle data for both, training and input variable. Data from the last hour is used; the *k*-nearest-neighbor method was applied in an urban network.

4.3 Approach

Figure 9 shows the overall framework from data collection through the processing steps down to the actual estimation or prediction corresponding to the example application described in the following.

Continuously available traffic data (here, local traffic data) is collected and used as independent variables X. The traffic data of interest needs to be observed (at least) occasionally (e.g., vehicle positions from a GPS-equipped vehicle) and processed to make up usable traffic variables y. These steps include data plausibility checking and smoothing of local traffic data time series, as well as necessary processing steps to, for example, derive travel times from vehicle position data.

The data is organized in a database with time stamp and location information, such that for each pair of data points, their spatial distance ds and their temporal distance dt can be determined.

If a prediction $y'_r(t+\Delta t)$ is queried (e.g., travel time on route r for a prediction horizon Δt), for all observations y_r in the database, the relevant features are identified to form the pattern X_r. X_r consists of traffic data (e.g., occupancy data) that are spatially and temporally close to y_r, with individual data points $x \in X$ to be weighted according to their spatial and temporal positions. In addition, the so-called calendar attributes are included into the pattern (day of week and time of day).

In the next step, the k most similar instances from the database are identified by applying a distance metric to compute the distance D. y'_r is estimated based on the observations $y_{r,k}$ that correspond to the identified instances, applying a Gaussian kernel to derive the relative weights w_k of the k-nearest neighbors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

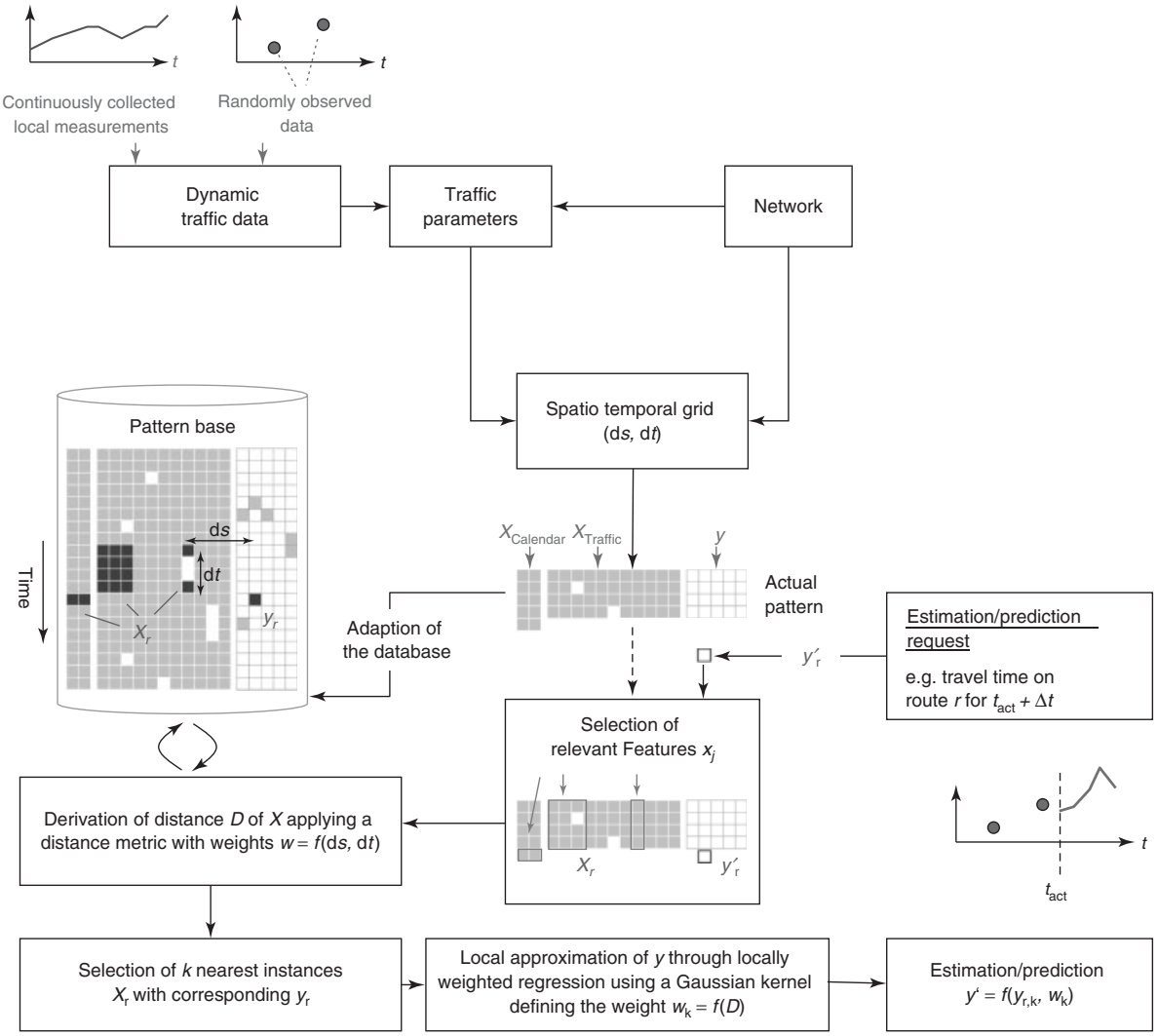


Figure 9. Overall framework of the estimation and prediction processes.

A detailed description of the method can be found in Leonhardt and Steiner (2012).

4.4 Results

Figure 10 shows some exemplary results of the method applied to estimate travel times on one route in Munich/Germany (a) and one route in Graz/Austria (b). Note that the "measured" travel times are only displayed as some sort of a reference time series and are not used to derive the estimation (light solid line). It can be seen that the method produced plausible results by exhibiting day-to-day dynamics and the extraordinary high travel times.

The instance-based learning method has been compared to three other methods. These comparisons indicate that

- the method outperforms the simple, calendar-based prediction (i.e., predicting the travel time based on day of week and time of day only without considering current measurements);
- the method performs comparable to different ANN topologies (multilayer feed architecture, trained with back propagation), yet yielding the advantage of being a rather transparent approach (not a "Black Box").

5 SUMMARY

In this chapter, an overview of sensor data fusion methods and applications has been given. Section 1 has described the most important terms related to data fusion. In Section 2, the most common data fusion methods are briefly introduced.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

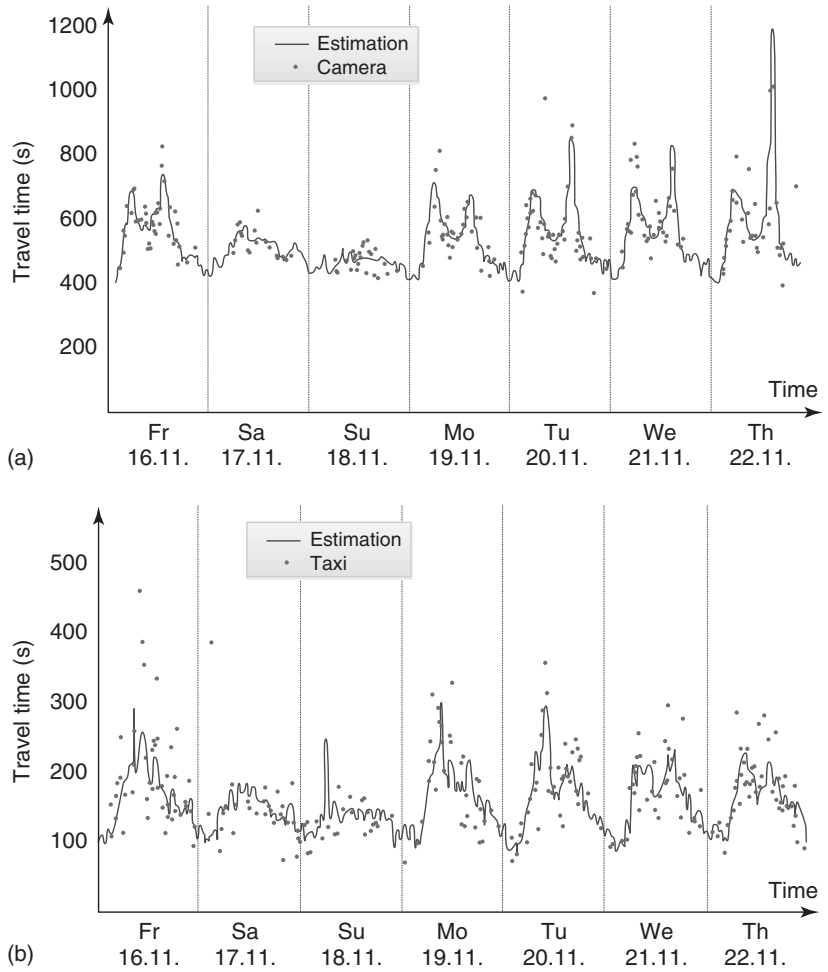


Figure 10. Estimated travel time and training data from vehicle (a) reidentification and (b) taxi probes.

In Section 3, an object tracking system based on an object level fusion and further strategies has been presented. The system has been exemplified with the extracted observations of a camera and a laser scanner. On the choice of object level, a high reliability and system flexibility can be reached. However, effects such as perspective changes, maneuvering objects, and split and merge effects must be treated apart. Thus, a variable logical reference is proposed, which might be adapted for alignment between observations and object hypotheses. Furthermore, a partial observability matrix is introduced, which allows keeping a general object model and tracking only observable model parts. Finally, strategies of hypothesis duplication and unification have been presented, which are used for both tracking maneuvering objects and dealing with sensor typical split and merge effects. These strategies have proven to bring an improvement of the quality of the test results. The methods presented have been prototypically implemented and tested. The results have shown an increase of the object data quality. Although not all components have been real-time optimized yet, a core of these strategies has already been adopted to the autonomously driving test vehicle FASCar (Löper *et al*., 2011), which has been successfully used within different demonstrations.

In Section 4, an instance-based learning method to fuse data from local detector stations and probe vehicles to estimate and predict travel times in road networks has been described. This application demonstrates the benefit of data fusion in case there is a very valuable information observed only occasionally (here, travel times) and other data (less valuable, but usable as independent variable) is available permanently. A potential application in traffic managements is the information of drivers with network-wide traffic state information, either in the form of tables, where the travel times for different routes can be displayed, or visually as "level of service maps."

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

GLOSSARY

Term	Definition
Sensor	Device to convert physical quantities into (analogical or digital) signals that can be further processed.
Sensor data fusion	Combining or processing information from multiple sensors with the objective of improving quality and/or quantity of information compared to only using a single sensor.
Estimation	Calculating an approximate value of a (physical) variable based on observation data (input data set), which may be incomplete or uncertain.
Prediction	Estimation of a (physical) variable as of a future time point.
Filtering	Processing a (sensor) signal with the objective of removing noise or other irrelevant parts in order to estimate the target variable.
Tracking	Estimating the state of a dynamic object in the course of time, such as the position of the ego-car or the position of other vehicles or pedestrians. Tracking contains normally a filtering effect and is thus often used as a synonym of it.
Object	Representation of a physical object (e.g., vehicle) in data processing.
Observation	Detection of objects or phenomena out of raw sensor data.
Hypothesis	Proposed or assumed explanation for a phenomenon. In relation to object detection and tracking, an object hypothesis consists of an object estimation.
Laser scanner	Sensor consisting of controlled laser beaming to measure distances, often with the objective to detect the environment, that is, objects.
Camera	Optical device for image (or image sequence) acquisition.
Artificial intelligence	Science that aims at replicating or using principles of human intelligence by means of computer programs.
Data mining	Process of discovering patterns and information in large databases using algorithms.
Training data	Set of observations used to detect relationships for estimation and prediction, for example, input data (independent variables) and output data (dependent variables).
Probe vehicle	Vehicles that are participating in the traffic flow and collect state data (e.g., position and speed). Data are stored or transmitted online for further processing.
Floating car data (FCD)	Data collected by means of probe vehicles.
Plausibility	Validity of an observation or measurement, possibly estimated using data fusion.
Smoothing	Process of finding an approximate function for a signal that reduces or eliminates noise while keeping or emphasizing the wanted signal. (Related to filtering.)
Time series	Sequence of data, in temporal order.
Cooperative data fusion	Using several independent data sources to get information that would not be available from the single sensors.
Complementary data fusion	Using several independent data sources to get a more complete picture of the observed phenomenon.
Competitive data fusion	Using several independent data sources to get independent measurements of the same characteristic.
Sensor level fusion (low level fusion)	The raw data of the sensors is processed closely together.
Object level fusion (mid-level fusion)	Data of each sensor is preprocessed separately before it is fused.
Information level fusion (high level fusion)	Data fusion systems in which the involved input sources already contain a complete data processing, in order to provide semantic information.

REFERENCES

Bar-Shalom, Y. and Li, X. (1995) *Multitarget-Multisensor Tracking: Principles and Techniques*, University of Connecticut, Storrs, CT.

Berthold, M. and Hand, D.J. (1999) *Intelligent Data Analysis: An Introduction*, Springer, New York.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Brooks, R.R. and Iyengar, S.S. (1998) *Multi-Sensor Fusion: Fundamentals and Applications with Software*, Prentice Hall PTR, Upper Saddle River, NJ.

Busch, F. and Ghio, A. (1994) Automatic incident detection on motorways by fuzzy logic. *Proceedings Traffic and Transport Solutions*, Amsterdam

Brown, R.G. and Hwang, P.Y.C. (1997) *Introduction to Random Signals and Applied Kalman Filtering: With MATLAB Exercises and Solutions*, 3rd edn, John Wiley & Sons, Inc, New York.

Catalá-Prat, Á., Köster, F., and Reulke, R. (2010) Image and laser scanner processing as confident cues for object detection in driving situations. *Proceedings of the ISPRS Commission V Symposium: Image Engineering and Vision Metrology*.

Catalá-Prat, Á. and Köster, F. (2011) Object level fusion and tracking strategies for modeling driving situations. *Proceedings of the IEEE Conference on Vehicular Electronics and Safety*.

Dailey, D.J., Harn, P., and Lin, P.-J. (1996) ITS data fusion. Final Research Report. Research Project T9903, Task 9. ATIS/ATMS Regional IVHS Demonstration.

Davis, G.A. and Nihan, N.L. (1991) Nonparametric regression and short-term freeway traffic forecasting. *Journal of Transportation Engineering*, **117**(2), 178–188.

Dinkel, A., Leonhardt, A., and Piszczek, S. (2008) Plausibility of road weather data. WIRELESSCOM (Hrsg.) *SIRWEC 2008, 14th International Road Weather Conference Prague—Abstract proceedings*, ISBN 978-80-87205-01-3.

El Faouzi, N.-E. (2006) Bayesian and evidential approaches for traffic data fusion: methodological issues and case studies. *Proceedings of the 85th Transportation Research Board Annual Meeting*, January 22–26, Washington, DC

FGSV (2010) Hinweise zur Erfassung und Nutzung von Umfelddaten in Streckenbeeinflussungsanlagen, FGSV-Nr. 306. (Recommendation paper published by the German Road and Transportation Research Association).

FGSV (2006) Hinweise zur Qualitätsanforderung und Qualitätssicherung der lokalen Verkehrsdatenerfassung für Verkehrsbeeinflussungsanlagen, FGSV-Nr. 386. (Recommendation paper published by the German Road and Transportation Research Association).

Friedrich, B., Matschke, I., and Heinig, K. (2004) Data fusion technique in the context of traffic state estimation. *Proceedings of the Triennial Symposium on Transportation Analysis TRISTAN V*, June 13–18.

Hastie, T., Tibshirani, R., and Friedman, J. (2001) The Elements of Statistical Learning. Data Mining, Inference and Prediction. Springer. http://www-stat.stanford.edu/~tibs/ElemStatLearn/ (accessed 27 July 2011).

Kaempchen, N., Weiss, K., Schaefer, M., and Dietmayer, K.C.J. (2004) *IMM object tracking for high dynamic driving maneuvers*. IEEE Intelligent Vehicles Symposium.

Kalman, R.E. (1960) A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, **82**(1), 35–45.

Kato, T., Ninomiya, Y., and Masaki, I. (2002) An obstacle detection method by fusion of radar and motion stereo. *Intelligent Transportation Systems, IEEE Transactions on*, **3**(3), 182–188.

Koks, D. and Challa, S. (2005) An introduction to Bayesian and dempster–shafer data fusion. Report DSTO-TR-1436, on behalf of the Australian Government, the Department of Defense.

Kwon, J. and Petty, K. (2005) Travel Time Prediction Algorithm Scalable to Freeway Networks with Many Nodes with Arbitrary Travel Routes. Transportation Research Record: Journal of the Transportation Research Board, No. 1935, Transportation Research Board of the National Academies, Washington, DC, pp. 147–153.

Leonhardt, A. and Steiner, A. (2012) Instance based learning for estimating and predicting traffic state variables using spatio-temporal traffic patterns. *Proceedings of the TRB 91th Annual Meeting*, Washington DC, January 21–26.

Linauer, M., Din, K., Asamer, J., *et al*. (2006) FUSION - Floating Car Daten und Sensordaten intelligent fusionieren. Final report as part of the "I2 - Intelligente Infrastruktur" research program sponsored by bmvit.

Löper, C., Catalá-Prat, Á., Gacnik, J., *et al*. (2011) Vehicle automation in cooperation with V2I and nomadic devices communication. *IEEE Intelligent Transportation Systems*, pp. 650–655.

Maier, F. (2010) Segmented Regression approach for Estimation of Traffic Characteristics—Application to Local Data, Section Data and Information derived from Position Reports. *13th International IEEE Conference on Intelligent Transportation Systems*. Funchal, Portugal, September 19–22.

Maybeck, P.S. (1979) *Stochastic Models, Estimation, and Control* vol. 1, Mathematics in Science and Engineering., Academic Press, New York.

Park, D. and Rilett, L.R. (1998) Forecasting multiple-period freeway link travel times using modular neural networks. Transportation Research Record, No. 1617, Paper No. 98-0743, pp. 163–170.

Rao, B.S.Y., Durrant-Whyte, H.F., and Sheen, J.A. (1993) A fully decentralized multi-sensor system for tracking and surveillance. *The International Journal of Robotics Research*, **12**(1), 20–44.

Robinson, S. and Polak, J. (2005) Modelling urban link travel time with inductive loop detector data using the k-NN method. *Journal of the Transportation Research Record*, **1935**, 47–56.

Sankar K. P. and Pabitra, M. (2004) *Pattern Recognition Algorithms for Data Mining*. Chapman & Hall/CRC, Boca Raton 158488457.

Sarma, V.S. and Raju, S. (1991) Multisensor data fusion and decision support for airborne target identification. *IEEE Transactions on Systems, Man and Cybernetics.*, **21**, 1224–1230.

Smith, B. and Demetsky, M. (1997) Traffic flow forecasting: comparison of modeling approaches. *Journal of Transportation Engineering*, **123**(4), 261–266.

Stüker, D. (2004) Heterogene Sensordatenfusion Zur Robusten Objektverfolgung Im Automobilen Straßenverkehr. PhD.

Sun, H., Liu, H.X., Xiao, H., *et al*. (2003) Use of local linear regression model for short-term traffic forecasting. Transportation Research Record, No. 1836, Paper No. 03-3580.

Takezawa, K. (2005) *Introduction to Nonparametric Regression*, 1st edn, Wiley-Interscience. ISBN: 0471745839

Thomaidis, G., Spinoulas, L., and Lytrivis, P. (2010) *Multiple Hypothesis Tracking for Automated Vehicle Perception*. IEEE Intelligent Vehicles Symposium (IV).

Turochy, R.E. and Pierce, B.D. (2004) Relating short-term traffic forecasting to current system state using nonparametric regression. *2004 IEEE ITS Conference*, Washington, DC, 3–6 October.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Vu, T.-D., Aycard, O., and Appenrodt, N. (2007) *Online localization and mapping with moving object tracking in dynamic outdoor environments*. IEEE Intelligent Vehicles Symposium.

Weiss, K., Stüker, D., and Kirchner, A. (2003) *Target modeling and dynamic classification for adaptive sensor data fusion*. IEEE Intelligent Vehicles Symposium, pp. 132–137.

You, K. and Kim, T. (2000) Development and evaluation of a hybrid travel time estimation model. *Transportation Research C*, **8**, 231–256.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto177
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Applications—Intelligent Vehicles: Driver Information

Bernd Rech, Stephan Glaser, and Christian Wewetzer

Volkswagen AG, Wolfsburg, Germany

1 Introduction 1
2 Technical Aspects 2
3 Value-Added Services 2
4 Strategical Driver Information: Navigation and Efficient Driving 4
5 Tactical Driver Information: Warnings and Recommended Actions 6
Glossary 8
References 8

1 INTRODUCTION

This chapter gives an overview about the current state and possible future of driver information applications. In the context of telematics, a driver information application can be regarded as a software component that wirelessly receives data, processes it, and presents it to the driver and/or the passengers with the goal of supporting long-term decisions (e.g., what route should the driver select) or short-term decisions (the driver should react soon because of a certain event) in driving, or providing other information relevant to the trip (e.g., historical information on nearby landmarks) or the comfort of the driver and passengers (e.g., a personalized entertainment program). In this sense, three classes of driver information applications can be distinguished:

- strategical applications with a horizon of several kilometers (route choice, planning the trip);
- tactical applications with a horizon of a few hundred meters (notification/recommended actions regarding events right ahead);
- value-added services/entertainment (making a trip entertaining, informative, and convenient)

From a system architecture point of view, a driver information application is a part of a driver information system. This system can be split up into two domains: a vehicle domain and an infrastructure domain. Concerning the vehicle domain, a driver information system consists of a communication unit for wirelessly sending and receiving data, a processing unit for evaluating data, and a human–machine interface to interact with driver and/or passengers. The infrastructure domain can, in short, be regarded as one or more data centers where information for vehicles is hosted and preprocessed.

While this chapter does not cover issues around the human–machine interface for driver information applications in vehicles, some general guidelines hold when designing such interfaces. Driver distraction should be avoided in all cases—a holistic concept-of-use for all different interfaces in the vehicle (e.g., multifunction display, navigation screen, and speakers) should increase driver awareness and understanding of the current traffic situation. Furthermore, applications should be designed considering the driver and the vehicle need for security and protection of privacy. Finally, the already large and always increasing number of public, private, free-of-charge, and commercial sources offering information for such application creates a challenge: when two or more sources have different or even contradictory information for a road

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto178
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

segment, which version is correct? This assessment should not be left to the driver, but the vehicle should have enough meta-information available to evaluate information accuracy and timeliness.

The ultimate goal of the driver information applications is to inform—not to decide. The driver should not be overstimulated. The driver should be aware of the traffic situation on his or her journey, know about possible alternatives and consequences when driving. The well-informed driver will make well-thought-out decisions.

First, in Section 2, the technologies relevant for driver information applications are identified. Second, the different types of driver information applications are outlined: value-added services (Section 3), strategical driver information applications (Section 4), and tactical driver information applications (Section 5).

2 TECHNICAL ASPECTS

Driver information applications over recent years have experienced a surge of innovation. To understand the circumstances that led to this, it is interesting to look backward on means of driver information in the past. Until the 1990s, the only means of driver information were a map, the signs along the road, and the radio. Thanks to the breakthrough of personal computers and their increasing processing power; first route planning applications for home personal computers were released based on a digital map. These applications took the task of route planning over for the driver—of course, this was static before departure. Then, in the 1990s, three additional technology breakthroughs again changed the navigation and telecommunication world:

1. With the global positioning system (GPS) becoming available for civil use, it was then possible to identify the position of a vehicle, enabling the first in-vehicle navigation systems (see Technologies—Positioning: GNSS).
2. The Internet evolved and formed the basis for the evolution of services and online communities available nowadays.
3. Mobile communication systems [GPRS (general packet radio service), EDGE (enhanced data rate for GSM evolution), UMTS (universal mobile telecommunications system), and LTE (long term evolution)] paved the way for packet-based communication to and from vehicles (see Technologies—Communication: Mobile).

Because of the longer innovation and development cycles in the automotive domain, it took until the end of the 1990s and the beginning of the twenty-first century for new technologies to be introduced in vehicles, resulting, for example, in detailed and updated digital maps and new means of dynamic trip planning.

Some applications mentioned in the following sections, such as an emergency vehicle warning, are highly localized, in the sense that this situation affects only vehicles in the immediate vicinity. From a network design perspective, with respect to scalability, ideally vehicles should communicate directly and locally instead of relying on the infrastructure of a cellular network and a data center. This physically keeps the communication network traffic in the area where the information is relevant. As a solution, the automotive industry is working on a communication technology to enable *ad hoc* direct short-range wireless communication among vehicles, as well as between vehicles and road-side units: IEEE 802.11p (adopted in Europe according to the local frequency regulation as profile standard ETSI ITS G5) (see Technologies—Communication: Wireless LAN-based Vehicular Communication), which is a variant of the IEEE 802.11 [wireless LAN (local area network)] standard (IEEE Computer Society LAN/MAN Standards Committee, 1997). This technology enables communication with a latency of less than 10 ms (Lin *et al*., 2010) in situations where the driver needs to be aware of and/or take immediate action (tactical driver information applications). To make efficient use of the wireless channel, in Europe, it is intended that applications make use of two multipurpose messages instead of each application having its own message: the cooperative awareness message (CAM) (European Telecommunication Standards Institute, 2011), which is a heartbeat message with information such as position and speed that each vehicle sends periodically, and the decentralized environmental notification (DEN) (European Telecommunication Standards Institute, 2010), which is an event-based message.

In the past, many service providers and communities around the world have regionally deployed wireless access points based on the wireless LAN standard. These networks are intended for people to connect to a local network or the Internet, or to provide local point-of-interest information. With a corresponding communication module, such access points can also be accessed from the vehicle to retrieve information from a local server or a data center in the Internet. Figure 1 depicts all the aforementioned communication technologies in a comprehensive scenario.

3 VALUE-ADDED SERVICES

With the help of telematics, a wide range of value-added services can be offered in the vehicle. These services can be classified into map-based and map-independent services (Figure 2).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto178
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

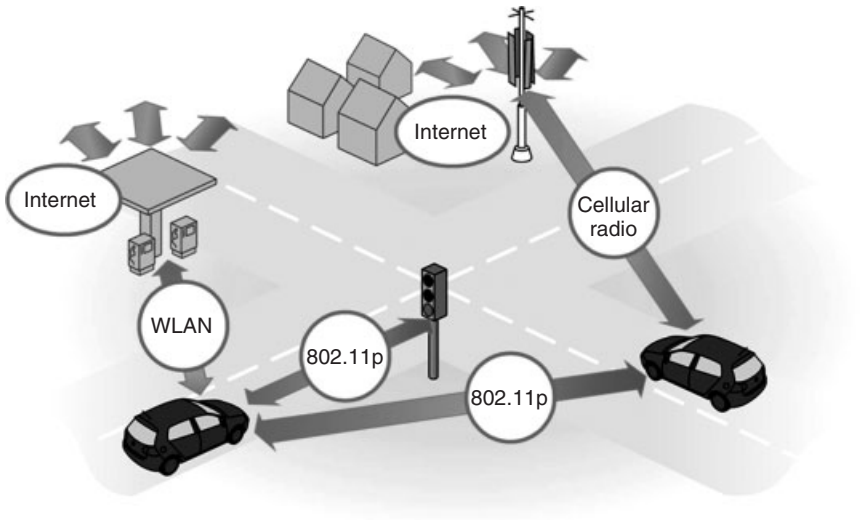


Figure 1. Illustration of communication technologies for driver information applications.



Figure 2. Community-based information in a map.

In general, all kind of information attributed with location data can provide an additional layer to a digital map (Cristani *et al*., 2008). The augmentation of a digital map with points of interest such as restaurants, landmarks, and shops is a typical example of a map-based value-added service. The map view can also be enhanced with aerial or satellite images, or with geo-tagged pictures from the area.

Another possibility is to integrate a friend-finder application (Palazzi, 2004), so that a live view of companion vehicles on the map is provided.

Many map-independent applications are available to increase the comfort of the driver and passengers. Vehicle diagnostics applications monitor the internal vehicle diagnostics status and can help in maintenance scheduling and preparation of the service at the dealership. A personal vehicle homepage on the Internet (see Figure 3 for an illustration) is a variant of a web portal (Tatnall, 2005) to display detailed vehicle status information, for example, on fuel consumption, and allow the remote upload of music as well to check and control the door lock status. Communication applications connect customers by means of voice e-mail and text messaging. An emergency crash response application transmits crash-relevant data such as location and impact speed to the service provider and/or emergency call center. Crisis assist services offer personalized assistance in providing detailed situation reports and giving directions out of dangerous situations. Stolen vehicle assistance is a security application that allows for features ranging from GPS-based tracking to possibly disable stolen vehicles.

Another value-added service is Internet radio, a very valuable addition to terrestrial audio broadcast services. It provides additional radio channels and, in particular, a multitude of special-interest programs. Further, the set of available stations does not depend on the location of the user and subscription fees do not exist. As audio entertainment is the most prominent form of in-car entertainment, it is very desirable to offer Internet radio to the driver. This holds especially true for markets where

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto178
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 3. Personal vehicle portal in the Internet.

special-interest stations are not available through broadcast services. Cellular data services are currently the only available technology to enable such a data service with reasonable coverage.

Challenges of in-car Internet radio are connectivity related. While data rates required to transport audio material are moderate, there exist precise constraints for data delay. If packet delays exceed a given threshold, playback gaps affect the user experience. Delays in packet transmission especially occur within cells under heavy load and when channel conditions are poor. Using a large-sized buffer at the receiver mitigates this problem but leads to long waiting times at startup and channel switching.

Remedies for these problems are server-based buffering methods or quality-of-service guarantees by the provider. The latter may become available by mobile networks of the fourth generation.

4 STRATEGICAL DRIVER INFORMATION: NAVIGATION AND EFFICIENT DRIVING

4.1 Navigation and traffic information

Navigation systems take over the task of finding a preferable route to the destination and providing the driver with directions while in traffic. Thus, the driver is able to focus more on the current traffic situation, reacting only to precise acoustic and visual driving instructions. Thereby, navigation systems certainly deliver increased comfort for drivers and passengers before and during the journey. Without navigation systems, getting directions on the journey and planning possible detours is often only possible with the help of a passenger, and are subject to errors because navigation and decision-making often happen as immediate reaction to changes in the traffic situation.

While navigation even without online connectivity is already an immense improvement over passenger-assisted navigation, the underlying map data is static. Thereby, it is already a common practice to include time-dependent travel times per segment. This way, recurring impacts such as rush-hour traffic are considered. Unforeseeable events (e.g., accidents, weather conditions, and road works) are not reflected in this map, which is why the computed travel time may become inaccurate on a journey. Ideally, the vehicle would be aware of all incidents and current travel times on the candidate routes (Figures 4 and 5). While travel time is the main decision criterion for most drivers, other factors such as fuel efficiency can be additional decision criteria. Map providers have recognized this, and nowadays, digital maps are available that include detailed information relevant for fuel consumption.

Contemporary navigation systems improve the lack of incident information in static maps by including information from the radio broadcast service traffic message channel (TMC) (Traffic Message Channel (RDS-TMC), 2003). As of 2012, TMC is available in most European countries, Russia, North America, China, and Australia, as well as in some countries in South America and Africa (TISA, 2012). This free-of-charge service is complemented by national commercial services, such as Navteq Traffic (formerly known as *TMCpro*) in Germany. Roadside sensors, inductive loops, and a number of vehicles provide their travel-time measurement as information sources of these services.

Besides receiving radio broadcasts, travel time and incident data can be acquired from service providers with the help of telematics. Thereby, a vehicle can play two roles: it can be information consumer (receive data) and information provider (send data from current/past journey). Nowadays, a large number of free and commercial fleet-based services are available. Waze is an example of an open-source community software for providing traffic information (Lequerica *et al*., 2010; Milo, 2011). HD traffic from TomTom (Schäfer, 2009) is a commercial example, which

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto178
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

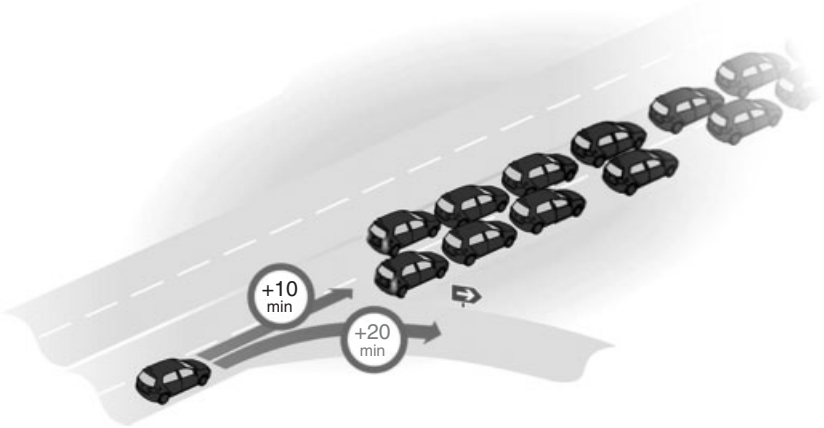


Figure 4. Travel time information.

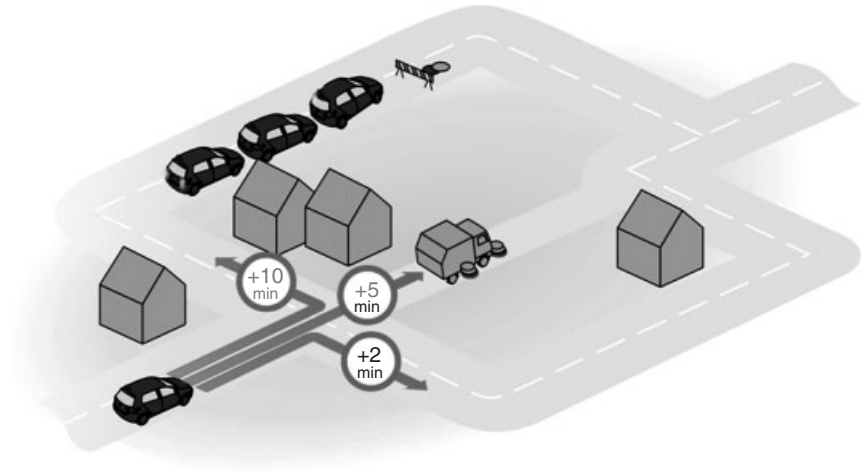


Figure 5. Incident information.

collects movements of mobile phones and GPS traces from the TomTom devices equipped with GSM modules to calculate travel times.

In research, ideas have been developed and evaluated for a wireless LAN-based self-organizing traffic information system (SOTIS) (Wischhof *et al*., 2003). Thereby, the vehicles would receive, aggregate, carry forward, and redistribute information to other nearby vehicles. Finally, telematics enable updates of all map data, for example, to include new roads or reconstructed road networks.

4.2 Parking and charging

Parking is a major issue especially when traveling into crowded regions. With the emergence of electric vehicles, it can be expected that also the availability of charging spots will play a role in destination and route choice. The associated parking search traffic is a burden for the population and other drivers. The availability of parking and charging spots affects route and destination choice in these areas; thus, the provision of accurate availability information is an important strategical driver information application. Nowadays, a number of cities already provide information on available parking spots in the Internet.

In future, it is also conceivable that vehicles distribute such information via short-range communication (Caliskan *et al*., 2006). With the vehicle sensors available nowadays, it is possible to detect free parking spots (Park *et al*., 2008). While these sensors provide information on currently available parking spots, the most convenient and reliable way of traveling would be to reserve a specific parking

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto178
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

spot for the time of arrival with the help of a booking system (Hilton, 1989).

4.3 Fleet management

Since the 1990s, shipping companies have deployed fleet management systems (FMSs). The first FMSs consisted of a trip optimization system, personnel and vehicle planning, customer databases, order processing, and invoice processing. In the beginning, trunked radio was used for communication with the driver.

With mobile communication and navigation systems, communication between the fleet and the control center is increasingly automated (Thong *et al*., 2007): vehicles are located and delivery status updates are provided. Delivery routes are adapted based on the traffic situation. Drivers can also ask for support in case of delivery problems, which minimizes search traffic and fleet downtime.

4.4 Multimodal trip planning

Even though many vehicles are equipped with navigation systems, journeys are still frequently planned well before the driver enters the vehicle. Applications for home personal computers or mobile devices such as smartphones can support the driver in planning his or her journey. Such planners can be multimodal, in the sense that they do not only include the vehicle but also other means of transport (Rehrl *et al*., 2005; Booth *et al*., 2009; Bustillos *et al*., 2011). In such planners, the user enters a destination and several means of transport (own vehicle, car-sharing, bus, train, aircraft, etc.) are then compared with regard to travel time, cost, and enviroment-friendliness.

5 TACTICAL DRIVER INFORMATION: WARNINGS AND RECOMMENDED ACTIONS

Tactical driver information applications improve traffic safety and efficiency. Drivers are given recommendations for accelerating, decelerating, or steering maneuvers, and/or they are warned of incidents right ahead to support anticipatory driving.

With the help of telematics, vehicles may immediately notify other vehicles (either directly via short-range communication or indirectly via mobile communication with a data center) when detecting dangerous situations. One example of a hazard warning is depicted in Figure 6, where a vehicle is notified of an approaching emergency vehicle. An emergency vehicle approaching an intersection is a dangerous situation, which requires the increased attention of all nearby vehicles. With communication among vehicles, the emergency vehicle is able to report its position, speed, and driving direction, so that other vehicles are aware of the situation and can react accordingly (Buchenscheit *et al*., 2009). In the end, this helps avoid accidents and supports the emergency vehicle to reach its intended destination as fast as possible.

When a lead vehicle is braking, for example, because of an obstacle ahead, the driver of the vehicle immediately behind is warned by the brake lights. Other vehicles behind often are unable to observe this braking maneuver (Figure 7). A warning message generated, in the case of a hard-braking maneuver, can help in anticipatory driving and avoiding rear-end collisions (Zang *et al*., 2008; Segata and Lo Cigno, 2011). This also avoids congestion of vehicles in dense traffic and thereby smoothens traffic flow.

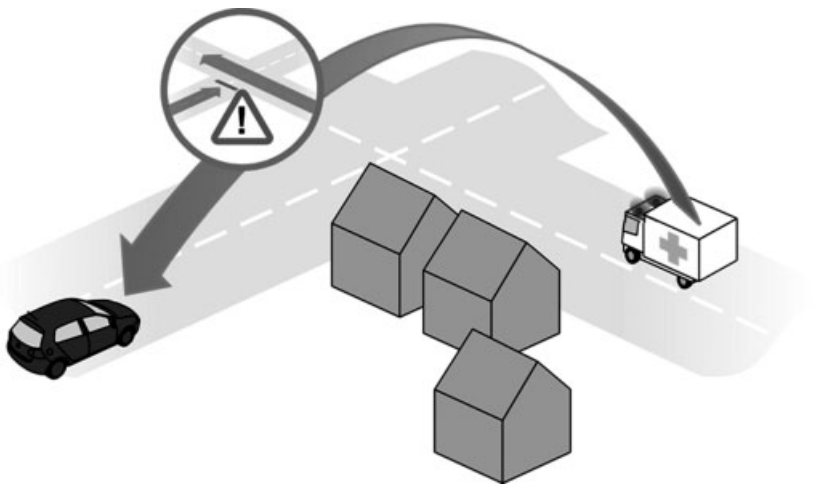


Figure 6. Emergency vehicle warning.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto178
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

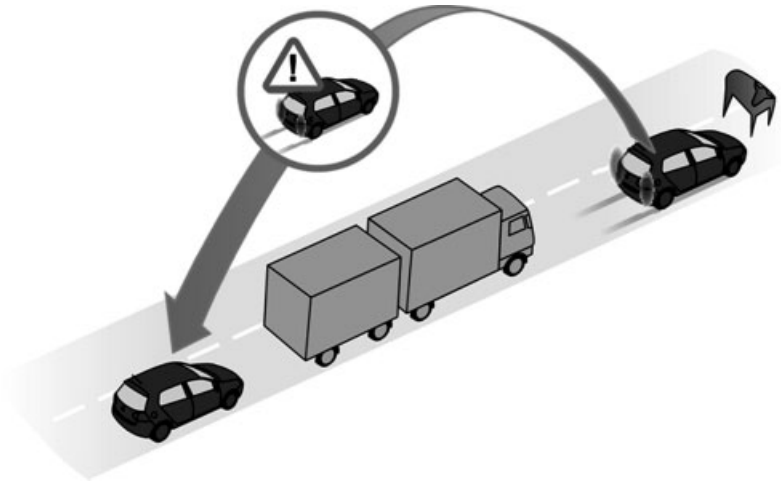


Figure 7. Electronic emergency brake lights.

Further examples of hazardous situations are slippery roads or reduced visibility because of heavy rain, snow, fog, smoke, or sand. Another kind of dangerous situation arises when a vehicle must stop. It may be easy to spot in an open-space environment, but when located behind a turn or a hill, there is always a risk that the driver of an approaching vehicle does not notice it. In this case, there is an increased accident risk. Thus, the stopped vehicle can warn approaching vehicles, for example, by sending warning messages when the vehicle warning lights are switched on or a deceleration threshold is reached. Such a warning message can also be sent by emergency vehicles securing a dangerous location. Dangerous situations may also occur at the end of a traffic jam, which are often difficult to detect by approaching vehicles, especially, when the line of sight is obstructed, for example, by a curve. However, when provided with an awareness of positions and speeds of vehicles in its vicinity via short-range communication, a vehicle can recognize it is approaching or is a part of the end of a traffic jam and report this situation to other vehicles (Kojima and Tsugawa, 2008; Padron, 2009; Bauza *et al*., 2010).

The aforementioned applications for dangerous situations in traffic are cooperative vehicle-to-vehicle applications (Caveney, 2010): they rely on vehicle-to-vehicle communication—either directly via short-range communication or indirectly via a data center. Such cooperative applications—such as telephones at the times they were introduced—suffer from network effects: the first vehicle equipped with a new communication technology will have no or only few other vehicles as communication partners. Thus, except for the feeling of being a pioneer, the use of these applications is very limited.

Therefore, the introduction of these cooperative applications should happen hand in hand with measures on the road infrastructure side; for example, the equipment of traffic lights to enable the provision of traffic light phase schedules. Figure 8 illustrates the idea of this application, which can greatly improve traffic efficiency (Wegener *et al*., 2008; Katsaros *et al*., 2011). By broadcasting its phase schedule via short-range communication, the traffic light helps vehicles waiting at the intersection to prepare for the upcoming green phase, increasing throughput at the traffic light. This can be further improved when the vehicles communicate with each other to help coordinate the driver to start and accelerate at the traffic light.

Vehicles approaching the traffic light can benefit from the phase schedule information by adapting their approaching speed accordingly. This way, vehicles can drive fuel efficiently by adapting to green waves and avoiding unnecessary braking. When traffic lights are equipped with wireless communication units, it is also conceivable that they receive information that vehicles are approaching or waiting at the traffic light. This way, wireless communication can be regarded as another sensor in addition to (or instead of) loop detectors or cameras. The traffic light can then adapt its schedule to the current traffic conditions.

Like traffic light phase schedules, information on traffic signs for the route ahead (especially, variable message signs) can be provided to vehicles. Technically, this could happen via short-range communication or via cellular networks and a data center. This way, for example, a vehicle could be able to detect when a driver is in the risk of passing a red light and issue a warning when appropriate. Vehicles may receive information about road construction, where orientation may be difficult especially when combined with

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto178
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

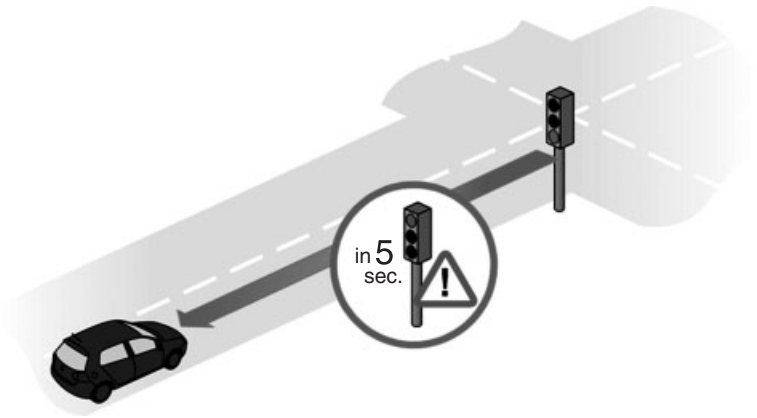


Figure 8. Traffic light phase information.

darkness or adverse weather conditions. Therein, information on lane geometry and the current traffic conditions at the construction site are important.

GLOSSARY

CAM	A cooperative awareness message (CAM) is a heartbeat message that each equipped vehicle sends out periodically delivering information such as position and speed.
DEN	A decentralized environmental notification is an event-based message delivering information about events such as obstacles on the road or the end of a traffic jam.
FMS	A fleet management system is typically used by the operators of a commercial vehicle fleet consisting of a trip optimization system, personnel and vehicle planning, customer databases, and order and invoice processing.
ITS G5	ITS G5 is a standardized protocol based on IEEE 802.11p, which defines enhancements to IEEE 802.11 necessary for vehicle-to-vehicle communication.
Multimodal	A multimodal trip includes not only the vehicle but also others means of transportation such as train or bus.
Network effect	The network effect describes the phenomenon that the first vehicle being a part of a cooperative system will have no or only few other vehicles as communication partners. Hence, the use of these applications will be limited.

REFERENCES

Bauza, R., Gozalvez, J., and Sanchez-Soriano, J. (2010) Road traffic congestion detection through cooperative vehicle-to-vehicle communications. 2010 *IEEE 35th Conference on Local Computer Networks (LCN)*, October 2010, pp. 606–612.

Booth, J., Sistla, P., Wolfson, O., and Cruz, I.F. (2009) A data model for trip planning in multimodal transportation systems. *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology,* EDBT '09, ACM, New York, NY, pp. 994–1005.

Buchenscheit, A., Schaub, F., Kargl, F., and Weber, M. (2009) A vanet-based emergency vehicle warning system. *Vehicular Networking Conference (VNC)*, 2009 *IEEE*, October 2009, pp. 1–8.

Bustillos, B.I., Chiu, Y.-C., and Papayannoulis, V. (2011) Pre-trip and en-route multi-modal travel decisions considering habitual travel times under unexpected incident scenarios. 2011 *14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, October 2011, pp. 594–599.

Caliskan, M., Graupner, D., and Mauve, M. (2006) Decentralized discovery of free parking places. Proceedings of the Third ACM International Workshop on Vehicular Ad Hoc Networks (VANET 2006), September, pp. 30–39.

Caveney, D. (2010) Cooperative vehicular safety applications. *IEEE Control Systems*, **30** (4), 38–53.

Cristani, M., Perina, A., Castellani, U., and Murino, V. (2008) *Content visualization and management of geo-located image databases*. *CHI '08 Extended Abstracts On Human Factors in Computing Systems,* CHI EA '08, ACM, New York, pp. 2823–2828.

European Telecommunication Standards Institute (2010) ETSI TS 102 637-3 V1.1.1 (2010-09). Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specifications of Decentralized Environmental Notification Basic Service, Sep 2010, ETSI, Sophia Antipolis Cedex, France.

European Telecommunication Standards Institute (2011) ETSI TS 102 637-2 V1.2.1. Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto178
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of Cooperative Awareness Basic Service, March 2011, ETSI, Sophia Antipolis Cedex, France.

Hilton, I.C. (1989) The removal of parking search traffic from the town centre. *Vehicle Navigation and Information Systems Conference, Conference Record*, September 1989, pp. 427–431.

IEEE Computer Society LAN/MAN Standards Committee (1997) IEEE Std. 802-11-8-1997. *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*. The Institute of Electrical and Electronic Engineers, New York.

Katsaros, K., Kernchen, R., Dianati, M., *et al.* (2011) Application of vehicular communications for improving the efficiency of traffic in urban areas. *Wireless Communications and Mobile Computing*, **11** (12), 1657–1667.

Kojima, M. and Tsugawa, S. (2008) An effect of the inter-vehicle communications on the traffic flow. *Proceedings of 15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting*, November.

Lequerica, I., Garcianda Longaron, M., and Ruiz, P.M. (2010) Drive and share: efficient provisioning of social networks in vehicular scenarios. *IEEE Communications Magazine*, **48** (11), 90–97.

Lin, C.-S., Chen, B.-C., and Lin, J.-C. (2010) Field test and performance improvement in ieee 802.11p v2r/r2v environments. 2010 *0IEEE International Conference on Communications Workshops (ICC)*, May pp. 1–5.

Milo, T. (2011) Crowd-based data sourcing, in Databases in Networked Information Systems, Volume 7108 of Lecture Notes in Computer Science (eds S. Kikuchi, A. Madaan, S. Sachdeva, and S. Bhalla), Springer, Berlin / Heidelberg, pp. 64–67.

Padron, F.M. (2009) *Traffic Congestion Detection Using VANET*, Florida Atlantic University, Boca Raton, Florida, USA.

Palazzi, C.E. (2004) Buddy-finder: a proposal for a novel entertainment application for gsm. *Global Telecommunications Conference Workshops, 2004. GlobeCom Workshops 2004. IEEE,* Nov-3 Dec 2004, pp. 540–543.

Park, W.-J., Kim, B.-S., Seo, D.-E., *et al.* (2008) *Parking space detection using ultrasonic sensor in parking assistance system. Intelligent Vehicles Symposium, 2008 IEEE*, June 2008, pp. 1039–1044.

Rehrl, K., Leitinger, S., Bruntsch, S., and Mentz, H.J. (2005) Assisting orientation and guidance for multimodal travelers in situations of modal change. *Proceedings of the 8th International IEEE conference on Intelligent Transportation Systems*, Vienna, Austria, September 2005.

Schäfer, R.-P. (2009) Iq routes and hd traffic: technology insights about tomtom's time-dynamic navigation concept. *Proceedings of the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, ESEC/FSE '09, ACM, New York, NY, pp. 171–172.

Segata, M. and Lo Cigno, R. (2011) Emergency braking: a study of network and application performance, in *Vehicular Ad Hoc Networks* (eds J.B. Kenney, M. Gruteser, M. Torrent-Moreno, and F. Bai), ACM, New York, NY, USA, pp. 1–10.

Tatnall, A. (2005) Web portals: the new gateways to Internet information and services, ITPro collection. Idea Group.

Thong, S.T.S., Tien Han, C., and Rahman, T.A. (2007) Intelligent fleet management system with concurrent gps gsm real-time positioning technology. *Telecommunications, 2007. ITST '07. 7th International Conference on ITS*, June 2007, pp. 1–6.

TISA (2012) Tmc world map, February 2012, http://www.tisa.org/technologies/tmc/tmc-world-map/.

Traffic Message Channel (RDS-TMC) (2003) Standard for Traffic Message Channel used with RDS.

Wegener, A., Hellbrueck, H., Wewetzer, C., and Luebke, A. (2008) VANET simulation environment with feedback loop and its application to traffic light assistance. *AutoNet 2008: Proceedings of the 3rd IEEE Workshop on Automotive Networking and Applications*, December.

Wischhof, L., Ebner, A., Rohling, H., *et al.* (2003) Sotis - a self-organizing traffic information system. *In VTC03-Spring: Proceedings of the 57th IEEE Vehicular Technology Conference*, Spring, pp. 2442–2446.

Zang, Y., Stibor, L., Reumerman, H.-J., and Chen, H. (2008) Wireless local danger warning using inter-vehicle communications in highway scenarios. Wireless Conference, 2008. *EW 2008. 14th European*, June 2008, pp. 1–7.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto178
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Driver Assistance

Dirk Wisselmann and Werner Huber
BMW Group, Munich, Germany

1 Introduction 1
2 Brief Review of Prometheus 2
3 Role of Environment Sensing 2
4 ADAS State of the Art 4
5 ADAS-Functions for Comfort 4
6 Adas Safety Functions 6
7 Future ADAS 8
8 Conclusion 10
References 10

1 INTRODUCTION

The term *driver assistance* describes, in a literal sense, the capability of a vehicle to support the driver in his or her driving task. In the context of *intelligent vehicles*, this term is used for vehicle functions that measure and interpret the near and far-field driving environments using appropriate sensor technology. A key characteristic for categorizing these functions is the so-called "degree of automation," ranging from

- informing (e.g., display of current speed limit);
- warning [e.g., lane departure warning (LDW)];
- actively intervening (e.g., distance and speed control) to
- autonomously acting systems.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

According to the RESPONSE definition (ACEA, 2009), the so-called advanced driver assistance systems (ADAS) monitor vehicle surroundings and can (autonomously) take over control of the lateral and/or longitudinal movement of a vehicle under defined circumstances, as long as the driver does not intervene. ADAS are generally characterized by the following features. They

- support the driver in his or her primary driving task;
- provide active support for lateral and/or longitudinal control;
- detect and evaluate the vehicle environment;
- use complex signal processing;
- enable direct interaction between the driver and the system.

ADAS aim to improve driving safety, efficiency, and comfort. The design task requires separate objective quality measures for each of these aims.

With the goal of achieving maximum *safety performance*, ADAS are designed for particular critical driving situations, such as unintentional lane departures or impending rear-end collisions. The criticality assessments of these situations are derived from detailed accident analyses and statistics, which are utilized to generate objective requirements on ADAS design. Objective criteria for *efficiency* (e.g., CO_2 emission per kilometer) can be derived in a similar manner.

Definition of objective ADAS design targets for *driving comfort* is complicated by the need to consider multiple aspects: increased driving comfort from the driver's point of view is achieved when driving burden is reduced and subjectively experienced driving enjoyment is increased. However, the relationship between burden and performance also needs to be considered, as shown in Figure 1. In the low workload regime, that is, in monotonous situations such as driving in traffic jams, automatic braking and acceleration

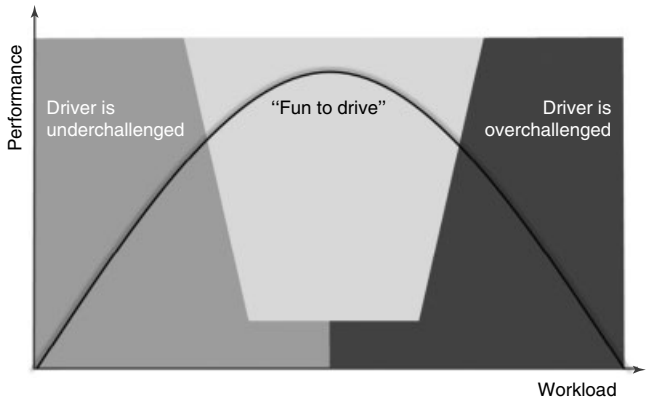


Figure 1. Driver performance versus workload (comp. Yerkes–Dodson law). (Reproduced by permission of BMW AG.)

by the vehicle reduces the driver's burden and is perceived quite positively by drivers. On a demanding, curvy stretch of road—a driving situation with typically moderate workload and requiring high performance—activation of the same function is perceived negatively by drivers, who usually enjoy driving curvy stretches of road, because system activation would actually decrease their enjoyment. In such situations, drivers are usually content with information and light/sight assistance, systems that preserve or enhance the driver's driving competence. If the driver is over-challenged, advanced driving assistance is strongly appreciated and can even play an essential role in avoiding accidents. A typical example of a conflict situation that over-challenges the driver is an impending pedestrian collision. Here, preventive safety systems can contribute strongly to accident prevention or at least mitigation of accident severity.

ADAS should be specifically designed for different driving situations, so that a "symbiotic" relationship between the driving task and the controller functionality of the ADAS is established. Consistent with this strategy, several general design principles can be formulated (Kompass and Huber, 2006) to guide development:

- ADAS operate like a virtual copilot;
- ADAS increase the competence of the driver;
- ADAS are supportive and do not dominate the driver;
- ADAS can always be overridden;
- ADAS can be operated intuitively and simply;
- ADAS can be switched on/off any time by the driver;
- ADAS exhibit transparent system behavior;
- ADAS are controllable with low effort;
- ADAS keep the driver activated;
- Individual ADAS functions interact flawlessly.

In general, driver assistance systems should always be oriented toward the driver's need for support in a particular driving and should consider the traffic situation. The degree and extent of automation appropriate for each driving task should be determined within a comprehensive, self-consistent concept including considerations such as customer acceptance and compliance, product and market philosophy, product liability, and impacts on traffic (Naab and Reichart, 1998, Bubb 2003). ADAS should support the responsible, active driver in his or her role as the controlling element in the system of driver–vehicle–environment. This requires a clear separation and recognition of the roles of the driver and the system. This recognition is particularly important at the boundaries of function activation, where controllability must be assured and over-challenging the driver is to be avoided.

2 BRIEF REVIEW OF PROMETHEUS

The essential groundwork for modern ADAS was laid in the research project "PROMETHEUS" (PROgraMme for a European Traffic of Highest Efficiency and Unprecedented Safety), which was carried out from 1987 to 1994 (Braess and Reichart, 1995). Figure 2 illustrates the project vision, which aims toward intermodal traffic with maximum safety and efficiency by means of comprehensive, mutual networking of all participants in traffic. The basic vision of the project is still valid even today. Automatic cruise control (ACC), which was the first ADAS capable of influencing vehicle dynamics that achieved production status in Europe, was mainly developed within PROMETHEUS. Key elements of image processing and night vision technologies were also advanced in this trend-setting project. Internationally, comparable projects were also carried out in Japan (ASV) and the USA (California Path), though with a smaller number of partners and a more limited scope.

3 ROLE OF ENVIRONMENT SENSING

The fundamental requirement for driver assistance systems is data acquisition to measure the vehicle environment using machine perception (Figure 3). The degree of support for tasks previously carried out by the driver is directly dependent on how comprehensively, completely, and reliably the information available to the assistance system can be supplied. Data and information can be acquired from sensors (e.g., radar, digital video processing, and lidar), telecommunication (e.g., GSM), on-board knowledge bases (e.g., navigation data), or a combination of these approaches. However, none of the currently available

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

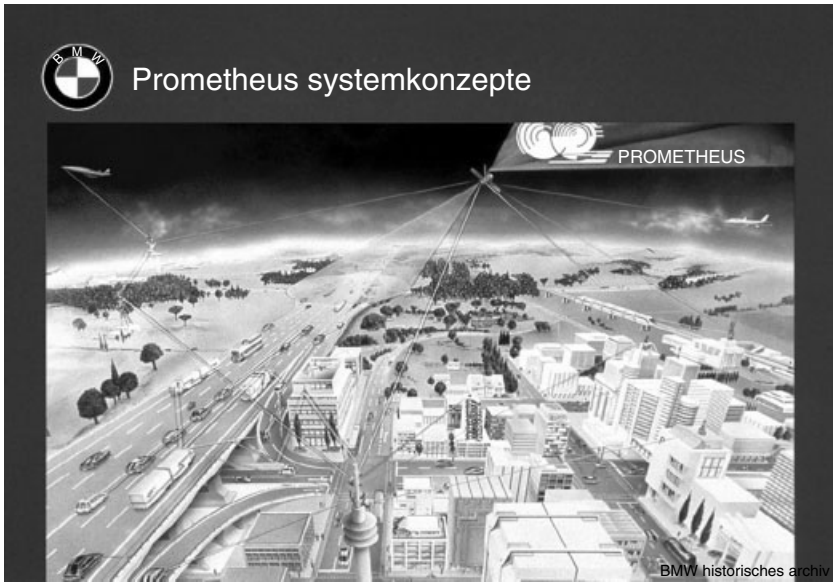


Figure 2. Vision of an interconnected intermodal traffic (PROMETHEUS). (Reproduced by permission of BMW AG.)



Figure 3. Detection and classification of objects via video camera. (Reproduced by permission of BMW AG.)

sensor systems can provide a comprehensive, 100% reliable representation of the driving environment under all relevant traffic and environmental conditions. An appropriate methodology for integrating uncertain information is thus a key aspect of development and design of driver assistance systems. Approaches for dealing with uncertainty include utilization of data redundancies or model-based plausibility testing, use of *a priori* or contextual knowledge as well as sensor data fusion. Owing to this complexity, enhancement and optimization of environmental sensing and scene

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

interpretation continue to represent one of the main challenges in ADAS development (Maurer and Stiller, 2005).

4 ADAS STATE OF THE ART

Figure 4 shows an overview of the key ADAS introduced into the European market since 1990; the number of functions has increased dramatically since 2005. This process began in the premium and upper intermediate market segments; however, currently, "basis driver assistance systems" [ACC, LDW, lane change warning (LCW), etc.] are quite prevalent even in the lower intermediate segments and are being rolled out in the economy segments. The trend toward both a broader range of vehicles offering ADAS and an increase in the number of available functions has encouraged stronger customer awareness and thus increasing demand for ADAS in new vehicles.

In the following section, important, "representative" driver assistance systems (classified according to the vehicle functions comfort and safety) are discussed in detail (Winner, Hakuil and Wolf, 2009, Eskandarian, 2012).

5 ADAS-FUNCTIONS FOR COMFORT

5.1 Automatic cruise control (ACC)

5.1.1 Basis function

As in conventional cruise control, ACC automatically regulates the vehicle speed to a target speed set by the driver. However, if there is a slower vehicle in front, ACC extends ordinary cruise control by regulating the driver's speed to consider the leading vehicle's speed and to maintain a following distance close to a target determined by the driver. Figure 5 illustrates the functionality of an ACC system. The limits of acceleration and deceleration for ACC are determined by the ISO 15622 norm. In the interests of maintaining a safety margin, these limits are set well below the corresponding values that the vehicle is technically capable of reaching. The first ACC systems were able to be activated in the speed range from 30 to180 km/h. With the advent of the so-called Stop&Go function, the current speed range of more advanced systems has been extended to include a complete stop of the vehicle, so that advanced

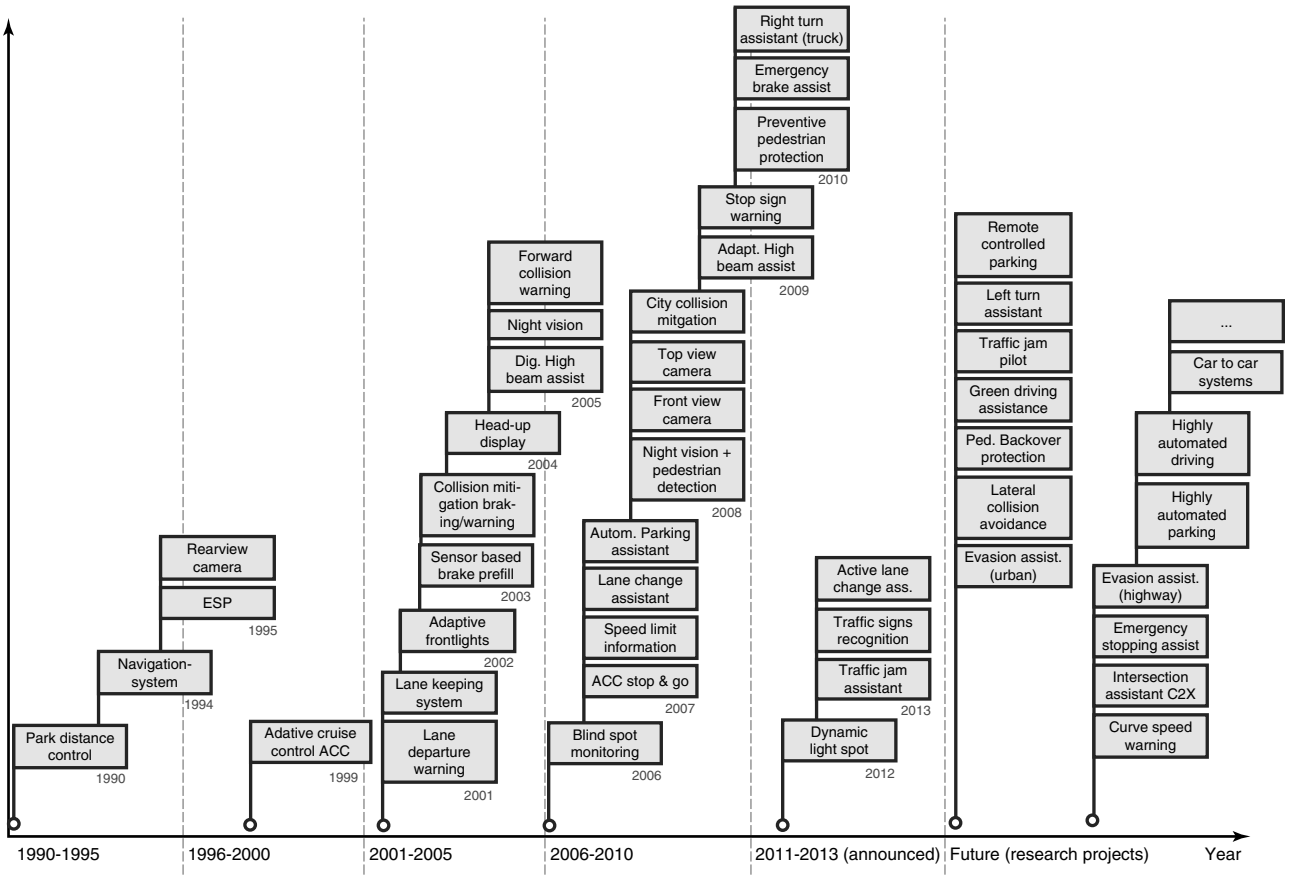


Figure 4. Series introduction of ADAS since 1990 (2013ff estimated). (Reproduced by permission of BMW AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

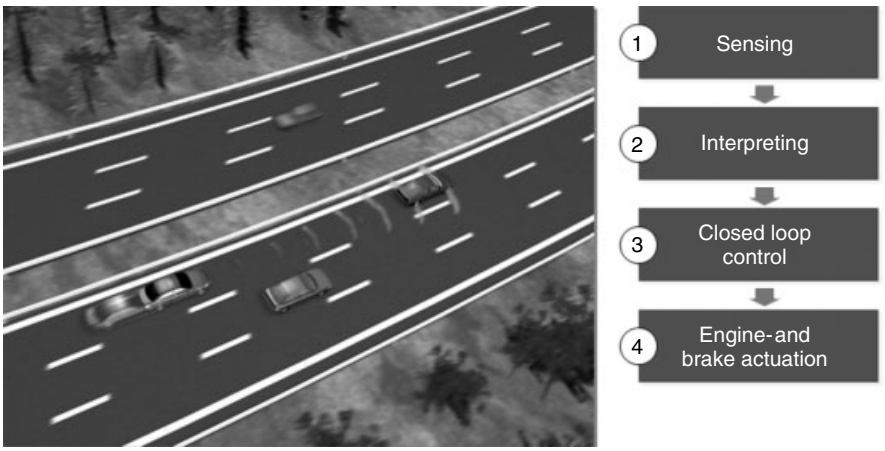


Figure 5. Principles of automatic cruise control. (Reproduced by permission of BMW AG.)

systems continue to operate even in traffic jam situations. This capability was enabled by additional sensors acting to secure the area immediately in front of the vehicle.

5.1.2 Sensors

All ACC systems currently on the market in Europe utilize radar sensors for distance and speed measurements. This choice of sensors is mainly due to the robustness of radar sensor measurements under adverse weather conditions and the ability to directly measure the leading vehicle's speed by the Doppler effect. Systems operating with lidar or stereo camera sensors are available in Japan.

5.1.3 Human–machine interface

The system is operated either by a lever located on the steering column or by buttons on the steering wheel itself. The driver can observe that a leading vehicle has been detected and that speed control is active either in the information display or in the head-up display.

5.2 Lane-keeping support (LKS)

5.2.1 Basis function and sensors

In addition to the longitudinal control (acceleration/deceleration), the driver must continually exert lateral control (steering) as a second control task: the lane-keeping support (LKS) system serves to support this driving task. Using a CMOS (complementary metal-oxide semiconductor) camera, lane separation markings of the current lane are detected, so that the lateral position of the vehicle relative to lane separation markings can be measured. Using this input signal, LKS steers the vehicle (in most cases) toward the center of the lane. The main aim of LKS is to enhance driving comfort by supporting lane keeping. LKS also leads to increased safety as a secondary benefit by preventing unintended lane departures (compare safety function *Lane departure warning*).

5.2.2 Human–machine interface

In addition to the on/off switch, the human–machine interface (HMI) of a lane-keeping system includes a readiness display indicating operating status (availability). It informs the driver whether the camera has detected lane markings and thus indicates whether the system is ready. A further key driver feedback loop utilizes a "haptic" channel, that is, by steering wheel movements informing the driver whether or not the system is properly operating. By utilizing the driver's sense of touch, the vehicle can rapidly and dependably signal the driver that he or she needs to take over lateral control if required.

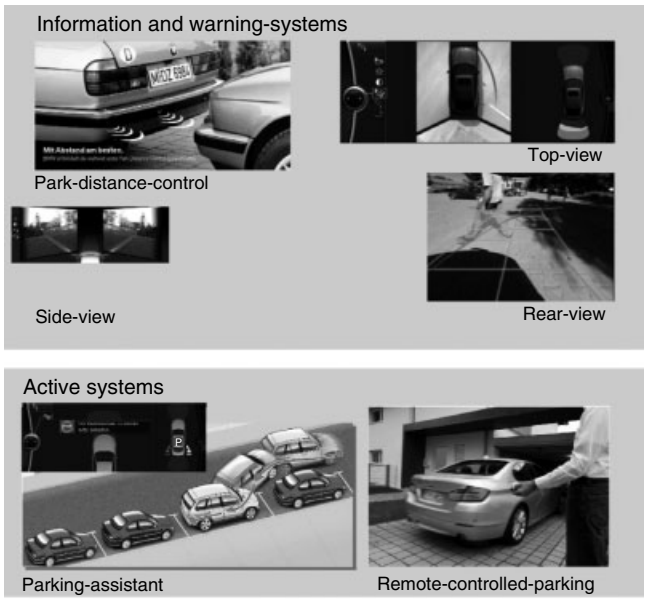
5.3 Parking maneuvers

5.3.1 Basis function and sensors

Parking (particularly parallel parking) was the first task supported by driver assistance using environmental sensors. Using ultrasound sensors, parking distance control (PDC) informs the driver of obstacles within about 3 m of the vehicle to the front and sides. Since the introduction of PDC, the number of different parking functions has increased substantially. These new capabilities and completely novel functions are partly due to installation of video cameras as additional sensors (e.g., top view). In addition, the original, purely informative PDC systems have been extended to active systems. Parking maneuver

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Classification of ADAS for parking.



(Reproduced by permission of BMW AG.)

assistants automatically measure potential parking spaces and steer automatically into the space. Table 1 shows a summary of current parking functions, classified according to whether they display information or actually steer the vehicle.

6 ADAS SAFETY FUNCTIONS

6.1 Forward collision warning (FCW) and forward collision avoidance (FCA)

Forward collision warning (FCW) and forward collision avoidance (FCA) aim to avoid rear-end collisions with a leading vehicle or mitigate their severity and are offered by practically all motor vehicle manufacturers. The term *precrash* system is often used synonymously, but in fact precrash systems usually include preventive activation of passive safety devices in the vehicle as an additional feature.

Although details vary according to design, four distinct phases of a possible accident sequence (Figure 6) can generally be distinguished for system operation. These phases are triggered according to the estimated time to collision (TTC), which is computed on the basis of sensor information and kinematics as the time available until a collision. During the first phase, the brakes are prepared for rapid emergency response by lowering the brake assistance threshold and

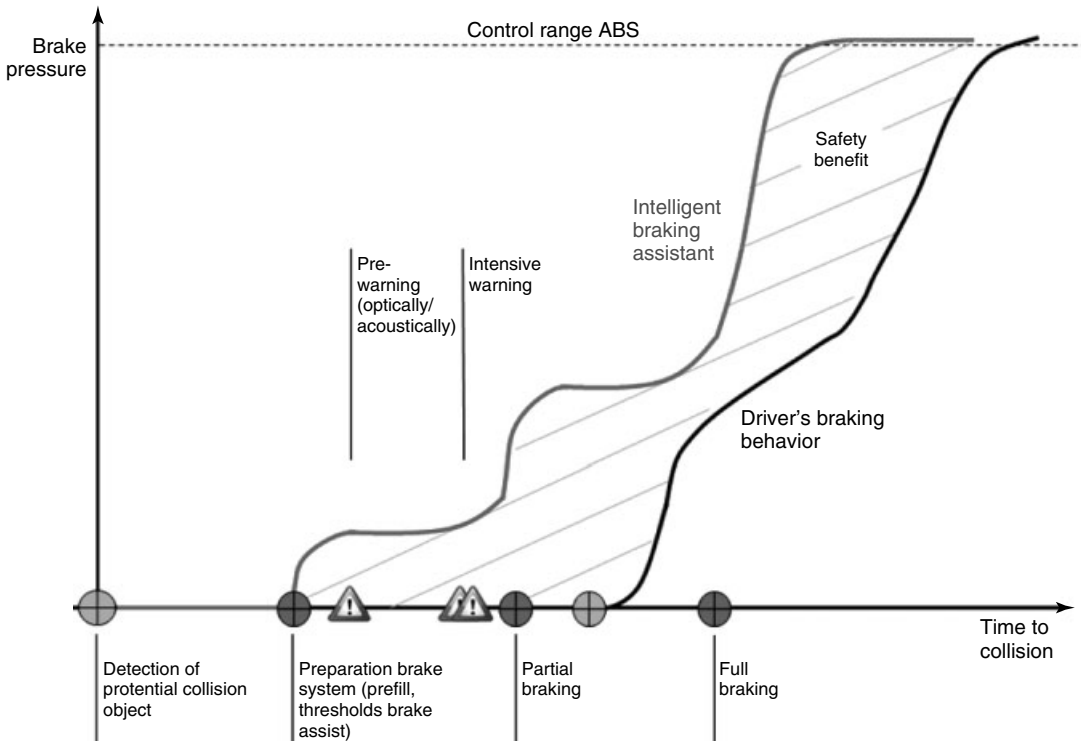


Figure 6. Phases of AEB systems. (Reproduced by permission of BMW AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

by prefilling (a slight increase of hydraulic pressure); these steps occur typically by a TTC of 2.5 s. During the second phase (generally between 2.5 and 1.5 s before collision), the driver receives an acoustic, haptic (touch), or kinesthetic warning. The third phase is characterized by active deceleration, that is, braking at about 4 m/s^2, with the aim of reducing the energy that needs to be absorbed at impact. Usually, this phase is supported by an intensive acoustical warning (acute warning). If a collision is unavoidable ("point of no return"), maximal braking deceleration is triggered for about 0.5 s.

As far as sensors are concerned, systems are mostly based on forward-looking ray sensors (radar and lidar) or cameras. The situations and speed regimes (e.g., urban or freeway traffic) that can be effectively addressed depend on the sensors or sensor combinations installed. Owing to the need to grab the driver's attention and focus on the task of rapid, effective accident avoidance actions, a key design element is the display concept. Depending on collision risk assessment, passive safety devices in the vehicle may be preconditioned, for example, by pretensioning seatbelts, closing windows, and placing seats in the optimal position. Owing to the relevance in accidents, forward front protection systems are on the way to become mandatory items, for example, in European Union (EU), from November 1, 2015 onwards for all new trucks.

6.2 Protection of vulnerable road users

In the past, the focus of vehicle-based preventive pedestrian systems was directed to passive safety, with the goal of mitigating the consequences of pedestrian collisions. More recently, the importance of preventive active pedestrian safety measures is increasing rapidly—owing in large part to European regulatory and consumer protection initiatives. If there is an impending collision, these active systems aim to avoid the collision entirely or at least decrease collision severity by countermeasures taken during the precrash phase. Effective operation of active pedestrian protection systems requires sensors capable of rapidly and reliably detecting pedestrians, a computational logic for estimating and updating collision probability and a decision logic for selecting and activating the appropriate warning and/or intervention sequence. Collision probability is estimated on the basis of analysis and prediction of the relative trajectories of the vehicle and the pedestrian. Analog to the aforementioned collision warning systems, a sequentially escalating warning and intervention strategy is also used for preventive pedestrian protection:

- timely warning of the driver;
- preparation of the braking system by prefilling (intelligent braking assistant), so that maximum braking pressure is immediately supplied if the driver initiates any braking response at all;
- active braking intervention up to maximum braking to reduce the collision speed, which has a direct impact on injury severity;
- in the future, active support of swerving to avoid the pedestrian.

Cameras, laser scanners, or photonic mixing devices (PMD)—possibly with radar sensors—are utilized for pedestrian detection. The sensors have the task of monitoring the region in front of the vehicle and proving timely and reliable pedestrian detection. These devices require a visual line of sight to the pedestrian; however, this requirement is unfortunately not always satisfied because of occlusion by objects in the way (e.g., delivery vans and parked cars). In order to detect occluded objects, considerable research is currently focused on cooperative measurement and tracking technologies that require radio contact between the vehicle and the pedestrian.

The driver is generally made aware of the hazard by means of an acoustic and/or optical warning or by a kinesthetic signal (e.g., a brief braking impulse).

Night vision assistance represents a special case of preventive pedestrian protection. Night vision extends the limits of human vision at night and thus enables foresighted driving even under low light or very low light conditions. Technologically, two current sensor approaches can be distinguished: far infrared and near infrared ("near" and "far" referring here to the closeness of the wavelength to that detectable by the human eye). The former technology is based on a heat-sensitive camera located at the front of the vehicle, which acts as a passive sensor and measures the "heat radiation" emitted by objects at a range of until 300 m in front of the vehicle. In the latter approach (a "near-infrared" system), the key sensor technologies consist of an active infrared light source together with a camera sensitive to near-infrared frequencies. A range of about 150 m can be achieved by this approach.

In the first-generation night vision systems, the processed infrared images were depicted in black and white on a vehicle display; warm objects such as animals or humans appear as lighter shapes on a darker background. The second-generation systems evaluate the camera image, highlight pedestrians or animals within the image, and warn in case of a collision hazard. In the third-generation devices, the driver's attention is focused directly onto the detected object by means of a targeted light beam, requiring an active light source (Figure 7).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 7. Far-infrared night vision systems with active light beam. (Reproduced by permission of BMW AG.)

6.3 Lane departure warning

As the name suggests, the purpose of this function is to warn the driver against unintended lane departures. Most of current systems operate by detecting the lane boundary markings by video camera image processing. The forward-looking camera generates a warning before the vehicle actually reaches the lane boundary. There are also infrared-based sensors on the market. These sensors are located in bumper and operate by scanning the roadway to detect contrast changes. The resulting warnings typically occur significantly later than in camera-based systems. There are various modes or channels used for the feedback loop to the driver, which vary somewhat according to the HMI strategy and philosophy of individual manufacturers:

- vibration of the steering wheel before lane departure;
- compensatory torque on the steering wheel;
- generation of vehicle torque by an asymmetric braking intervention;
- side-coded acoustic warning and/or seat vibrations.

6.4 Lane changing and blind spot warning

Accident statistics show that more than 5% of all accidents take place during lane changes; a majority of these accidents occur on trunk roads or motorways (Bartels, Meinecke, and Steinmeyer, 2012). These accidents can be attributed firstly to a deficit of the driver in estimating the approach speed of vehicles seen in the rear-view mirror and secondly to the existence of a blind spot, essentially an area blocked by the vehicle itself, which is not visible in the rear view mirror. Planned (intentional) lane changing maneuvers are secured using an appropriate rear-facing sensor based on radar or video image processing that can identify vehicles approaching from the rear and the side. The driver establishes his or her intention typically by signaling the lane change. If the system identifies a lane change hazard, it warns the driver by light symbols located in the side mirrors or by vibrations in the signal lever or in an active steering wheel. The driver can respond by revising the lane-changing decision or by aborting a lane change already begun. ISO standard 17387 "lane change decision aid system" differentiates between three different types of systems. The "blind spot warning" monitors the blind spot on the left and the right adjacent to the driver's own vehicle. The "closing vehicle warning" monitors the adjacent lanes to the left and the right behind the driver's own vehicle in order to detect vehicles approaching from behind. The "LCW" combines the functions of "blind spot warning" and "closing vehicle warning" (Bartels, Meinecke, and Steinmeyer, 2012).

7 FUTURE ADAS

Aside from gradual refinements in current systems, three main trends can be identified for future ADAS:

- increased degree of automation in ADAS designed for increased comfort;
- introduction of novel ADAS for traffic safety;
- introduction of ADAS for improved efficiency.

7.1 Future ADAS for driving comfort

The next generation of ADAS designed to enhance driver comfort will exhibit an increasing degree of automation. The so-called "partially-automated" ADAS will be introduced during the next few years. An example of partial automation is the traffic jam assistant, which takes over both longitudinal control and steering of the vehicle under jam situations. However, partially automated ADAS are designed to be monitored by the driver at all times (BASt, 2012). The so-called "highly automated" ADAS are designed to go a large step further: the driver may at least temporarily delegate the driving task entirely to the vehicle (Waldmann and Niehues, 2010). According to current legal interpretations, highly automated ADAS are not approved for production in Europe. In addition, key liability questions posed by introduction of highly automated ADAS

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

remain unresolved until now. However, as, on the whole, these systems are generally thought to enhance traffic safety considerably, there is an ongoing dialog between manufacturers and public authorities with the goal of defining a process for achieving approval and for dealing with liability issues. The path toward highly automated driving is evolving gradually by continued development of current ADAS for driver comfort. Examples of this evolution are the extension of ACC and LKS to the traffic jam assistant and to the highway pilot, which are initially being introduced as partially automated functions but have the potential for further development as highly automated functions (Wisselmann *et al*., 2012).

7.1.1 Traffic jam assistant and highway pilot

In the traffic jam assistant, the vehicle performs both the longitudinal and the steering control tasks. This assistance is perceived by drivers as very pleasant in traffic jam situations, as he or she avoids the need to perform tedious, repetitive operations of steering and speed/distance control and can relax while observing the traffic situation. The traffic jam assistant utilizes sensor capabilities extended from the requirements of ACC and LKS to include measurement of the area immediately adjacent to the vehicle. Series production of the traffic jam assistant has been announced by the manufacturers BMW and Mercedes for the year 2013. Increasing the upper operating limit for vehicle speed in the traffic jam assistant results in the highway pilot, which supports comfortable freeway cruising. In this function, the transition from partial to high automation can be realized by means of sensors with an extended forward detection range. A highly automated highway pilot needs to be able to predict the traffic situation independently with a sufficient forecasting horizon to allow a driver who is completely out of the driving control loop, upon receiving a take-over request, sufficient time to react, become oriented, and resume control of the vehicle. Basic research to this end is currently being performed (Damböck *et al*., 2012).

7.2 Future ADAS for safety

Further development of ADAS for safety is dependent on expansion of the area in the vehicle environment that is monitored by measurements and on improved sensor reliability. Some novel sensor concepts will also be required, such as the Car2X technology in intersection assist. The three functions evasion assistance, intersection assistance, and emergency stopping assistance illustrate these relationships.

7.2.1 Evasion assistance

Steering and evasion assistance aim to avoid an impending collision with other traffic participants by a steering (swerving) maneuver. The function identifies the need for swerving by means of object detection in the area in front of the vehicle and assessment of the available stopping distance for a braking maneuver. If the available gap is insufficient to avoid a collision by braking, there may still be time to avoid the collision by swerving, and this may be the only feasible alternative. Three different system layouts can be distinguished: driver-initiated evasion, corrective evasion, and automatic evasion.

In driver-initiated evasion (swerving), steering activity initiated by the driver is supported, and dynamic stability and traction of the vehicle are automatically controlled. The function initially facilitates an agile initial steering impulse away from the hazard, then guides the vehicle around the obstacle along a stable trajectory, and prepares the vehicle for the reverse steering impulse. This layout prevents the vehicle from losing traction and achieves a stable trajectory around the hazardous object, but operates only if the driver initiates swerving.

Corrective evasion—which is not yet an available option—takes this function to the next level by introducing "system-initiated" evasive steering (swerving). Automatic evasion carries out the complete swerving maneuver automatically and is the subject of research activities in this area (Dang *et al*., 2012).

7.2.2 Intersection assistant

Several distinct functions are subsumed in the term *intersection assistant* (Ehmanns, Hopstock, and Spannheimer, 2005):

- turning/crossing assistance (prevention of collisions when a vehicle turns into or crosses an intersection);
- left/right-turn assistance (prevention of collisions with opposing vehicles that have the right of way);
- traffic light assistance (warning if the driver is about to run a red light);
- stop sign warning (warning if the driver is about to run a stop sign).

The full spectrum of intersection assistance functions places demands on detection of the vehicle surroundings that cannot be fulfilled by current vehicle-based sensors alone. These functions require communication-based Car2X technologies, a focus of current research, which can transmit data on obscured vehicles (SimTD, 2009).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

7.2.3 *Emergency stopping assistance*

The emergency stopping assistance function can be thought of as an intelligent vehicle service feature for health-related emergencies. The vehicle switches to high automation mode and carries out a secure emergency stopping maneuver (Kämpchen *et al.*, 2010): the vehicle activates the hazard flasher, carries out a controlled, independent maneuver to the road shoulder—while considering surrounding traffic and if necessary over multiple lanes—and stops there. The unique feature is that the vehicle goes beyond supporting the driver to take over the complete driving task, as in highly automated driving. The vehicle's maneuvering strategy is based on reliable localization of the vehicle within the current lane and in particular on detection of all other vehicles and objects in the immediate surroundings of the vehicle using a comprehensive sensor system and redundancy to provide robust detection. The emergency stopping assistant thus builds on available driver assistance systems but requires technical extensions and adaptations.

7.3 ADAS for efficiency

In the future, ADAS will also be designed for optimization of driving efficiency (Liebl, 2010). This aim can be illustrated by two functions:

7.3.1 *Signal light assistance*

A vehicle approaching a traffic signal will receive the scheduled phase change times by Car2X communication. If a stop is certain to occur, then the vehicle can carry out an optimal, energy-efficient brake sequence, for example, with recuperation and/or engine switch-off.

7.3.2 *Green ACC*

In the future, the ACC function will have the capability to optimize energy efficiency of the vehicle's longitudinal control (acceleration/braking). This function will require "data fusion," that is, processing of Car2X communication data (light phases) and integration with on-board sensor and digital map data. In this way, the vehicle will be able to anticipate its speed out to an "electronic horizon" and use this information to support optimal, far-sighted driving.

8 CONCLUSION

In the next few years, ADAS will continue to represent a key focus for innovation in vehicle development. The capabilities of the vehicle to create a comprehensive operational representation of the static and dynamic driving environment using sensors and communication are continually improving. These capabilities promise to increase the range and reliability of the vehicle's electronic horizon and generate substantial benefits for driving safety, comfort, and efficiency.

REFERENCES

ACEA (2009) http://www.acea.be/news/news_detail/acea_endorses_response_code_of_practice_for_advanced_driver_assistance_syst/ (accessed 01 August 2013).

Bartels, A., Meinecke, M., and Steinmeyer, S. (2012) *Lane Change Assistance. Handbook of Intelligent Vehicles*, Chapter 28, Springer-Verlag, Berlin, pp. 730–755.

BASt (2012) Rechtsfolgen zunehmender Fahrzeugautomatisierung. Bericht der Bundesanstalt für Straßenwesen. Heft F83, Bergisch-Gladbach.

BMW Group (2010) Pressemappe Innovationstag Connected Drive, München.

BMW Group (2011) Pressemappe Vision Connected Drive, Genf.

Braess, H.H. and Reichart, G. (1995) Prometheus: Vision des "intelligenten Automobils" auf "intelligenter Straße". ATZ Automobiltechnische Zeitung 97 (1995) 4 und 6 (Teile 1&2).

Bubb, H. (2003) Fahrerassistenz - primär ein Beitrag zu Komfort oder für die Sicherheit? In Der Fahrer im 21. Jahrhundert. VDI-Berichte 1768. Düsseldorf.

Damböck, D., Bengler, K., Farid, M., and Tönert, L. (2012) Übernahmezeiten beim hochautomatisierten Fahren. 5. Tagung Fahrerassistenz. München, 15.-16. Mai 2012.

Dang T., Desens J., Franke U., *et al.* (2012) *Handbook of intelligent vehicles: 29 steering and evasion assist. Group Research and Advanced Engineering, Driver Assistance and Chassis Systems*, Daimler AG, Sindelfingen.

Ehmanns, D., Hopstock, M., and Spannheimer, H. (2005) ConnectedDrive: advanced assistance systems for intersection safety. 12th World Congress on ITS, paper 2515. San Francisco, 6–10 November 2005.

Eskandarian, A. (ed.) (2012) *Handbook of Intelligent Vehicles*, vol. 2012, Springer-Verlag, London.

Kämpchen, N., Waldmann, P. Homm, F., and Ardelt, M. (2010) Umfelderfassung für den Nothalteassistenten—ein System zum automatischen Anhalten bei plötzlich reduzierter Fahrfähigkeit des Fahrers. 11. Braunschweiger Symposium AAET. Braunschweig, 10–11 February 2010.

Kompass, K. and Huber, W. (2006) Advanced driver assistance—how far should they go? VDA Technischer Kongress, München, 22–23 März 2006.

Liebl, J. (2010) *BMW EfficientDynamics - wir haben die Segel richtig gesetzt. 30. Tagung Elektronik im Kraftfahrzeug*, Haus der Technik, Dresden.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Maurer, M. and Stiller, C. (2005) *Fahrer-Assistenzsysteme mit maschineller Wahrnehmung*, Springer-Verlag, Berlin.

Naab, K. and Reichart, G. (1998) *Grundlagen der Fahrerassistenz und Anforderungen aus Nutzersicht. Seminar "Fahrerassistenzsysteme"*, Haus der Technik, Essen.

SimTD (2009) http://www.simtd.org/ (accessed 01 August 2013).

Waldmann, P., and Niehues D. (2010) Der BMW Track Trainer—automatisiertes Fahren im Grenzbereich auf der Nürburgring Nordschleife. BMW Group Forschung und Technik, Tagung Aktive Sicherheit. Garching.

Winner, H., Hakuli, S., and Wolf, G. (2009) *Handbuch Fahrerassistenzsysteme: Grundlagen, Komponenten und Systeme für aktive Sicherheit und Komfort*, Vieweg+Teubner, Wiesbaden.

Wisselmann, D., Huber, W., Rößing, J., *et al*. (2012) Highly-automated driving—state of the art and future challenges. VDA 14. Technischer Kongress. Sindelfingen, 22–23, März 2012.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto179
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Road Traffic and Travel Information

Fritz Bolte
BASt, Federal Highway Research Institute, Bergisch Gladbach, Germany

1 Definition 1
2 Objectives of RTTI 1
3 History of RTTI, Development, and Problems Occurring 3
4 Functions 3
5 Requirements 5
6 Synergies by Cooperation 7
7 Frame Conditions and Regulations 11
8 Acquisition of and Access to Data 13
9 End User's Access to RTTI 16
10 Coexistence and Cooperation of Public and Commercial RTTI 18
11 Important CEN/ISO Standards (EasyWay (1), 2012) 19
12 Glossary (eSafety RTTI WG, 2007) 20
References 21
Further Reading 22

1 DEFINITION

Real Traffic and Travel Information (RTTI) stands for all information that drivers need to plan their journeys, carry them out, and reach their destinations safely and in time. The range of information includes transport networks, usability of their parts, limitations with respect to their categories and regulations, existing and expected operational conditions, and travel times and costs. RTTI is used both pretrip with forecasts for travel disposition and on-trip real-time while driving to adapt driver behavior and route choice to prevailing conditions.

RTTI is provided through different information channels (see Technologies—Communication: Mobile, Technologies—Communication: Wireless LAN-based Vehicular Communication, Technologies—Communication: Broadcast), for example, radio broadcasts, RDS-TMC transmission, DAB services, telephone, mobile phone, and internet (Figure 1).

RTTI services are generally based on the following functions (see In-vehicle sensors, Data Acquisition by Roadside Detection, Technologies—Data Acquisition: Data fusion):

- Data collection
- Data fusion and processing
- Service provisions
- Data distribution

RTTI services usually include activities of various actors that need to cooperate on fixed rules and standards that are described below.

In this context, the focus is on road networks although extension onto other transport networks such as rail and air could be advisable. Intermodality would be a useful complement to exploit all possible transport facilities. The limitation on road transport should help to better understand existing RTTI systems, their possibilities and problems and emerging new developments.

2 OBJECTIVES OF RTTI

Safe and efficient transport systems are a precondition to modern life in a globalized world. Mobility of people and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

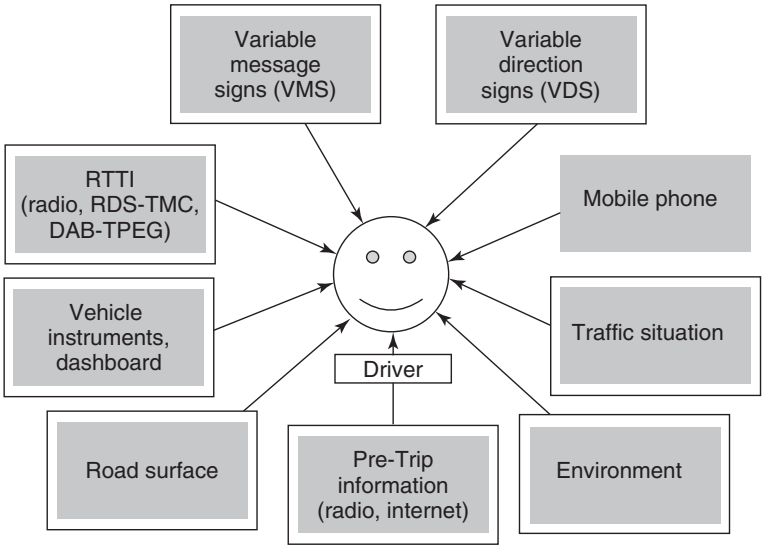


Figure 1. Driver information channels and sources. (Reproduced with permission from Bolte, 1997–2007. © Fritz Bolte.)

goods requires efficient transportation networks enabling timely accessibility of locations, production facilities, distribution and market places, recreational locations, and other places of interest. Quality of transportation is given high political priority.

In many countries, road networks carry the main load of transportation of people and goods as far as transport quantities are concerned; rail, ship, and air transport usually have different roles and importance.

2.1 Safety

Accurate, reliable, and on-time RTTI has the potential to increase the safety and efficiency of the transport system in many ways. With RTTI, road users can avoid congestion and bottlenecks, select alternative routes or transport modes, and avoid secondary accidents and incidents in urban areas and the motorway network. Informing the driver of the "operating conditions" of the traffic environment and the road conditions during her/his journey is the necessary complement to active and passive safety systems of the vehicle. Multiple services are in use in Europe, with various technologies (eSafety RTTI WG, 2007).

2.2 Exploitation of available transport capacity

The increase in the volume of road transport in the European Union associated with the growth of the European economy and mobility requirements of citizens is—together with roadworks and accidents—the main cause of increasing congestion of road infrastructure and rising energy consumption, as well as a source of environmental and social problems.

The response to those major challenges cannot be limited to traditional measures, inter alia the expansion of the existing road transport infrastructure. Innovation will have a major role to play in finding appropriate solutions for the European Union.

Intelligent Transport Systems (ITSs) are advanced applications, which, without embodying intelligence as such, aim to provide innovative services relating to different modes of transport and traffic management and enable various users to be better informed and make safer, more coordinated, and "smarter" use of transport networks (European Parliament and Council, 2010). RTTI systems are regarded as highly efficient tools to achieve these goals.

2.3 Economic aspects

Increasing congestion and accident numbers jeopardize transport economy and economic conditions generally. Transport costs are an important factor in globalized production and service provision. Vehicle and fleet operators need good and reliable transport conditions to deliver their goods and services as required and in time.

Another aspect is the negative impacts of traffic congestion on environment that require suitable and efficient measures against unfavorable traffic conditions and developments.

The creation and maintenance of favorable transport conditions is considered a political target of high priority. RTTI systems are regarded as promising tools to improve

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

traffic safety and transport efficiency and to reduce environmental impacts. Setting up and operating RTTI services is, on the one hand, challenging for engineers and organizers but, on the other hand, provides and safeguards employment.

RTTI services provide advantages for traffic flow and safety as well as for economic welfare. The introduction and operation of RTTI provides favorable potentials for high-end industrial products and various information services.

3 HISTORY OF RTTI, DEVELOPMENT, AND PROBLEMS OCCURRING

The various developments of new RTTI systems in many parts of the world reflect different geographic, economic, and technological environments and political demands in the respective countries. Design and layout of RTTI systems depend on frame conditions as set by the involved actors whose objectives may be identical, but may also be different or even conflicting. Basic information sources in this sense are drivers' experiences, their knowledge of the road network, and expected traffic and weather conditions. This store of knowledge may not be sufficient under all circumstances: additional input to the driver can be provided by RTTI systems that—due to their specific data provision—have a better overview about existing or expected road and traffic conditions.

It is difficult to identify the day traffic information services were started. In Germany, regular radio broadcasts of traffic information were set up as an answer to increasing congestion and the demand of drivers to be kept informed. Traffic broadcasting was started originally in the 1960s when increasing weekend and seasonal traffic caused heavy congestion on road networks. Drivers were warned against expected overloads of roads and especially of motorways. Information sources were experiences from the past that—in the following years—were complemented by traffic observations from police and motoring clubs with the help of patrol cars or helicopters.

With increasing motorization, more and more commuters used their cars on their way to work, and congestion became a daily phenomenon. Radio stations established daily traffic broadcastings every full hour, then every half hour, and finally by immediate program interruption for "urgent" traffic messages. Surveys revealed that traffic information was highly appreciated by most listeners and not only motorists. Consequently, all (public) broadcasters started competing against each other to provide high quality traffic information. Specific traffic information studios were set up gathering traffic information from different sources, for example, police, road maintenance units, emergency services, motoring clubs. Telephone and fax were used for communication between actors.

This caused a number of different results, positive and negative: on the one hand, more and better on-time information but, on the other, sometimes contradictory information from different sources. Another effect was an increasing number of traffic messages: drivers were "showered" with messages, most of which were irrelevant on their individual trip. Radio program directors and radio listeners were annoyed by increasing program interruptions caused by traffic messages.

In other countries, traffic information services originally were made up and operated by non-government organizations, for example, by motoring clubs or private and commercial entities, with the objective to provide good service to their members or to profit from revenues. However, despite differing developments, the problems were similar. New ways to overcome the difficulties had to be found.

The tendency to improve traffic information services through

- comprehensive information
- in-time messages
- correct messages
- understandable messages

overloaded existing radio and information technologies at the end of the 1970s. An additional challenge was the growing trans-border traffic caused by ongoing integration in Europe (European Union). International traffic means that a great number of foreign vehicle drivers are driving on European roads, who are not familiar with local road networks, traffic regulation, and signalization and do not understand local languages. They can hardly be reached by local traffic broadcast messages, although they may need them most urgently.

4 FUNCTIONS

RTTI systems include the following basic functions that had to be improved.

4.1 Data and information collection

Originally, police and road authorities had been the main providers of information from visual observations. From experience, they knew the "traffic hot spots" on the main arterials and reported incidents that occurred. With increasing traffic, the number of problem areas increased as well and it was more and more impossible to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

compile comprehensive and complete traffic reports. New information sources such as the following had to be found and exploited:

- Automatic detectors (such as inductive loops) as used for traffic management purposes to control variable message signs (VMSs) and variable direction signs (VDSs): Their data are used to automatically generate traffic messages. Usually automatic detection systems are installed in problem areas of high category roads (motorways, trunk roads). In some countries, commercial content or service providers have installed their own detection systems, especially on main roads such as motorways. As these roadside detection systems do not necessarily cover the complete network, "black spots" remain (see Applications—Intelligent Roads and Cooperative Systems: Urban Traffic Management, Advanced Highway Management Systems, Road Traffic and Travel Information (RTTI)).
- "Congestion reporters": Volunteers report, using mobile phones, their traffic observations to an information collection center operated by motoring clubs, radio stations, and the like. Reporters are registered; in order to safeguard credibility of reports, registrations are canceled if a reporter sends frequently incorrect messages. Each message must pass plausibility checks before being used for traffic information services.
- Information gained by evaluation of movements of mobile phones: Several technologies have been developed to identify the position of mobile phones and their movements (see In-vehicle sensors, Data Acquisition by Roadside Detection, Technologies—Data Acquisition: Data fusion).

4.2 Data compilation and fusion

RTTI is usually based on contents received from various sources. Automatically collected data from local detectors normally include local throughput (vehicles per time period), local speed (km/h at the measuring point), section speed (km/h on the road section), traffic density (veh/km), often per vehicle class, and statistical parameters. Reliability of these data depends on the density of measuring points and detector accuracy. Measured data can achieve a high accurateness. Incidents, for example, accidents, are usually not measured, but may be deduced from traffic model supported data interpretation or detected by visual observation (staff or video camera).

Floating car data (FCD) and floating phone data (FPD) use tracking data to determine speeds on road sections.

Information gained from human observation is less objective, but may be helpful to identify events that impact traffic flow. Video observation is increasingly used, often in connection with automatic video evaluation.

Current data from all available sources together with reported events are fused in order to get a comprehensive overview about the traffic situation and expectable trends. Interpretations are supported by historic traffic data of comparable traffic situations. Traffic models are used to get a realistic view of the current situation as well as short- and medium-term forecasts to be used for traffic management measures and for navigation devices.

4.3 Data interpretation, recommendations, and responsibilities

Road authorities, on the one hand, are mandated to manage/regulate traffic flows to reach the optimum of the whole traffic system; RTTI services, on the other, may tend to achieve the optimum for the individual driver/client. This is a critical mixture where responsibilities of road authorities and interests of RTTI service providers are concerned: Suitable agreements are needed in order to ensure that measures as indicated on VMSs do not conflict with messages of in-car navigation devices. Traffic management strategies as selected by road authorities need to be taken into account by RTTI service providers in their recommendations to their clients. This is important for route recommendations that should observe official road classifications to avoid route guidance through sensitive parts of road networks, for example, residential areas.

4.4 Distribution and presentation to end users

RTTI is processed and transmitted in coded form to be language independent, based on international standards as listed below. This includes the description of locations, events, and recommendations, which can be processed in the different stations of the message chain and finally decoded in the receiver unit to inform the end user in his desired language either in words (text-to-speech) and/or in graphic form displaying maps and symbols. Depending on the type of receiver, the encoded message can be used to update route recommendations. Message display may be according to the driving direction and area, and the road and route used. This allows for selection of relevant messages and suppression of messages that do not concern the individual trip.

A number of (independent) parties are active in the creation, processing, and transmission of RTTI and their "backstage systems." Besides the above-mentioned activities to generate content, to fuse and verify it, and finally to distribute it to the end user, the manufacturers

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of equipment—detection systems, traffic management equipment, broadcasting and telecommunication systems, and receiver units—are involved. Activities are carried out by different actors, in different ways, and with different interests that need to be elaborated on to understand the processes and possible conflicts.

5 REQUIREMENTS

The perception that former RTTI systems could not fulfill future requirements was an incentive to study the necessities of all involved actors, the potentials of new technical developments in the field of information technologies, and organizational frame conditions and barriers. Technological, operational, and organizational improvements were needed in all parts of the information chain.

5.1 Drivers as end users

It is evident that knowledge of the road network topology, as laid down on road maps on paper or electronic navigation maps, is an indispensable prerequisite for any trip. However, most driver are acquainted only with limited parts of road networks, for example, of their neighborhood or of the trunk road network. Frequent usage improves the network overview. Changes in the network—new road links or road closures—are more or less long-term events. Besides the mere network geometry, a number of specific road characteristics are important:

- Road design parameters: gradients, curvature, cross section, number of lanes, emergency lanes
- Limitations: height and width limitations, weight restrictions
- Regulations: speed limits, restrictions for specific vehicles or user groups, route guidance, environmental restrictions, etc.

Dynamic traffic management as used in heavily loaded road networks often applies dynamic regulations using VMSs and dynamic routing signs. Information such as the following is needed about permanent and temporary limitations of road usage:

- Weather dependent: road closures in winter; dangerous road surface, black-ice, prescription of snow-chains, reduced visibility, heavy rain, and snowfall
- Impairment by road maintenance operation: long-term and short-term road works, lane restrictions or blockages, and rerouting
- Expected or current traffic problems

Road users are increasingly confronted with traffic problems, when road transport demand exceeds road capacity. Road users want pre-trip and on-trip information about the following:

- Pretrip information supporting disposition of trips
- Expected traffic situations, for example, in connection with mass events such as sports events, fairs, seasonal, and weekend traffic
- Current congestion due to overload of roads, accidents, incidents, and road works
- Exceptional transports impacting road capacity and traffic flow
- Recommended alternative routes

The information demand varies among individual user groups. Commuters are usually well aware of oncoming daily road and traffic situations. Occasional road users and drivers over weekends and during seasonal traffic often lack information about the situation they have to face. Drivers of heavy goods vehicles, especially with loads to be delivered on time, urgently need information about road and traffic on their route including availability of parking lots for their prescribed pauses. The same holds for haulers and dispatchers planning and supervising transports. This enumeration is not exhaustive, but clearly indicates that specific needs of specific user groups require to "individualize," to tailor information to individual end user needs.

RTTI can reach drivers only if it is presented to them in an understandable way, that is, with symbols and graphics and/or in a language of their choice. This is an important demand to be considered in the design of RTTI systems.

5.2 Road operators and traffic managers

Motorways, or generally "A-Class roads of the trunk road network," are the most efficient category of road networks (eSafety RTTI WG, 2007). In order to improve traffic flow and to increase safety, most critical parts of the motorway network are equipped with dynamic traffic management such as stretch control (speed regulation, lane allocations, danger warnings), dynamic direction sign systems (rerouting traffic in case of incidents onto less loaded parts of the network), ramp metering systems (to avoid traffic flow breakdowns due to overload), and temporary use of hard shoulders for running traffic (especially in built-up areas).

Most of these systems are automatically controlled in closed or open loops. Comprehensive collection of traffic data and traffic-related environmental data is needed as input for possible control measures. These data are also

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

used to automatically produce traffic messages in case of incidents. These messages are often enhanced by observations from police or congestion reporters, the so-called jam busters, who are registered road users reporting their observations via the mobile phone.

Most of the above-mentioned systems use VMSs to transmit information, warnings, or regulations to users. The content of these variable message signs is limited because of the fact that drivers are only able to perceive and understand a limited amount of information when passing often at high speeds, especially when they are not capable of reading the language (e.g., if not in native language) and/or are not familiar with the geographical environment to understand exit recommendations (Figure 2).

From a traffic management point of view, it is vital to inform drivers accordingly and thus achieve the best possible rule compliance and acceptance for regulations. Therefore, in-vehicle traffic messages in national language or easily understandable icons play an important role as they complement roadside information and regulations on VMS or VDS. While VMS/VDS are only locally effective, traffic information can be transmitted to drivers anywhere on the road network, on-trip and pre-trip. RTTI is an advanced possibility to influence driver behavior, road choice, and also modal choice.

With increasing market penetration of navigation systems using digitally transmitted traffic information (RDS-TMC) to enable dynamic navigation to avoid congestion by updating route resistances according to prevailing conditions, traffic information gains even more importance in influencing road network operation. This is a fact to be considered when planning improvements and necessary enhancement of traffic information services. Navigation systems tend to reroute traffic from the motorway network to secondary roads if events or disturbances are reported on the motorway. However, traffic situations outside motorways and especially in urban areas are rarely monitored. Therefore, the situation on these secondary networks is more or less unknown to traffic managers. Available experience from the past and knowledge of normal traffic loads on these secondary networks can help overcome the situation but cannot really solve the problem. For efficient road network operation, there is an urgent need to improve traffic monitoring in all major (strategic) road networks (eSafety RTTI WG, 2007).

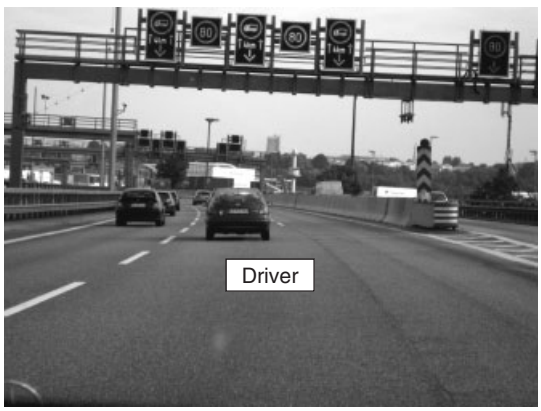


Figure 2. VMS legends with limited amount of information.

Therefore, in this context also, road operators need to be mentioned as "user group." They need information about expected and current road and traffic situation as well in order to anticipate and possibly avoid congestion by adequate traffic management measures or to react in case of incidents and accidents.

5.3 Service providers

Service providers pursue the objective of distributing/selling the best possible RTTI to their clients. Commercial companies, in competition with others, defend their market position by offering as best a service as possible. Economic results depend on acceptance by drivers and thus on the reputation of their clients and the possibility to "sell" RTTI. Establishment of own data collection is a means to improve the quality of RTTI as a commercial product.

5.4 Broadcasters and telecommunication services

Broadcasters—public as well as private companies—compete with their neighboring broadcasters covering the same transmission area. Quality of RTTI is an important factor in competition.

Public broadcasters, for example, in Germany, have been cooperating with road administrations and police since the establishment of RTTI services to distribute traffic information as a specific type of "news," which they are mandated to supply their listeners with by their concessions. Their RTTI service is given without extra pay for their listeners; it is included in the general broadcasting fee.

Private broadcasters finance themselves by broadcasting commercial advertising. There is no mandate for them to distribute RTTI free of charge. It is their own decision to deliver (free or paid) RTTI service.

Telecommunication service providers may act as content providers and RTTI distributors. Operational data of mobile phones may be used as data sources by tracking movements of their telephone clients. Mobile phones can be used to get RTTI "on demand." Economic motivations for telecommunication services are the sale of telephone airtime and the sale of content for other RTTI services, and also the provision of RTTI service for self.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 3. Early RDS-TMC receivers.

5.5 Equipment manufacturers

Equipment manufacturers need clearly defined standards for their products, enabling a variety of different products to be offered to the market to satisfy varying demands. This applies as much for encoding messages and their semantics as for transmission technologies over different channels. Global markets and trans-border traffic require language-independent RTTI messages as provided by RDS-TMC and TPEG (Transport Protocol Expert Group) standards (see appended list of CEN/ISO standards). While RDS-TMC is a transmission channel with very limited capacity, TPEG is designed for the transmission of RTTI over digital broadcasting channels with high capacity, enabling more detailed messages (CEN/TC 278/WG8 and ISO/TC 204/WG9).

On the basis of standards, manufacturers are able to design a variety of low and high end products of RTTI receivers as requested by the market (Figure 3). An increasing number of RTTI service providers operate their services, often in cooperation with other actors of the RTTI scene.

6 SYNERGIES BY COOPERATION

6.1 Actors and their interests

Based on their specific history of origin many different RTTI systems can be identified in Europe and beyond. However, similar actors are involved, although with differing degrees of involvement and commitment. Actors and their interests in participation can generally be categorized (Table 1).

6.2 Typical functions of actors

Data and information collection means the acquisition of all data describing road network topology, weather-influenced road surface conditions, current traffic situation, and events affecting traffic demand and traffic flow. It includes the collection of information on expected events and traffic situations. Table 2 lists the main actors and their main collection procedures in the case of own data sources.

Table 2 indicates that in principle all types of data/information are available, although at different levels of completeness, and in different entities. Some of them are public authorities, some are commercial companies, and others are individual drivers or motoring clubs. No entity by itself can provide a full and comprehensive overview of the current traffic situation. Cooperation and access to data pools are needed to gain a reliable basis and a full overview to provide satisfactory RTTI services. The basis for cooperation are standards, gateways to data pools, and clear and fair economic agreements between partners.

6.3 Accessibility of data sources

6.3.1 Standards

Early RTTI services were built as proprietary solutions. Their internal definitions of protocols, semantics, and data hindered exchange of information between potential partners, especially across national borders and languages. A means to overcome these obstacles is the international standardization of traffic messages and their digital coding as done in CEN/ISO 14819 Parts 1–6 of *Traffic and Travel Information (TTI)—TTI Messages via traffic message coding*. This is used as the basis for setting up the so-called Traffic Message Channel within the Radio Data System (RDS-TMC) to broadcast digitally encoded traffic messages via FM Radio.

The installation of RDS-TMC to transmit traffic messages is an important milestone for the improvement of RTTI services (see below). Meanwhile, RDS-TMC-based services are available nearly worldwide, partly as free services and partly as pay-services.

The method to code data and traffic messages (location, incident, length of queue, expected duration, recommendation, etc.) accelerates the procedure from content collection up to the distribution of the message to the end user.

Further standards enable transition to digital transmission media (DAB, GSM, DSRC, etc.) and information

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Actors and their interests.

Sector [1)]	Entity	Interests, Motivation
Public sector		
	Road authority	Political Objectives • Enabling mobility of people and goods • Safety • Exploitation of available road capacity • Transport economy • Protection of the environment (Ecology)
	Police	Political objectives • Safety • Prevention of accidents • Sanction of misbehaviour • Adequate reaction on accidents
	Emergency services	Rescue of lives
	Meteorological services	Alerts against hazardous weather conditions
	Telecom operators	Economic profit
	Map makers	Economic profit
	Broad casters	Public: public mandate to inform the public Private: economic profit
	Drivers	• Reliable trip disposition • In-time arrivals • Mobilty • Safety • Undisturbed travelling • Timely information about expectable and current obstacles and difficulties
	Motoring clubs	• Service to their members • Lead position against competitors
	RTTI Services	Economic profit
	Event managers	Easy accessibility of their events
Private (Commercial) actors		

[a] Remark: The transition between public and private sector is transient and differs from country to country.

interchange between actors in the information chain (DATEX, TPEG, etc.; see below).

6.3.2 *Data pools*

Table 2 explains that data pools needed for RTTI services exist, but are owned by different public and commercial parties.

The European Directive 2003/98/EU (Official Journal of the European Union of 31.12.2003, p. 90–96) regulates the reuse of public data, that is, the reuse of public data or public information held by public sector bodies of Member States. However, the definition of "public data or public information" and "public sector bodies" is left to the European Member States and is not done in a consistent manner. As a result, access to public road and traffic data is regulated nationally and differs across Europe (European Commission, DG TREN, 2008).

In some countries, RTTI systems originally have been established by public sector bodies (e.g., road administrations, police authorities, and public broadcasters); private services followed later with the intention to merchandise traffic information. Public basic services of satisfactory quality are available on the basis of public data pools, commercial services offer paid traffic information, and individual service, usually on the basis of public data and their own data pools.

In other countries, the private sector was to start RTTI services first on a commercial basis, establishing their data collection systems. In the latter case, public data pools are less developed. Commercial, that is, private data are protected by intellectual property rights and not accessible to other entities without specific contracts with the owner.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Typical information collection procedures of the individual entities—regardless information from other entities.

Type of Data Entity	Road Network	Traffic Statistics	Current Traffic Regulations	Traffic Management (TM) Strategies	Current Traffic Flow	Incidents–Accidents	Weather Development	Events	Traffic Forecasts
Road authorities	Design plans, maps	Census, detectors	Signalization plans	TM plans; dynamic TM	Detectors, TM systems, model-based data interpretation	Maintenance staff, model-based data interpretation	Weather sensors		TM systems, model-based data interpretation
Police forces	Observation	Experience	Planning process	Planning process	Observation	Observations, reports	Observations		Experience
Emergency service						Emergency calls			
Meteorological services							Meteo stations, network		
Telecom operators		FCD, FPD[a]			FCD, FPD[a]				
Map makers	Own map production		Static speed limit data collection						
Broadcasters					Congestion reporters	Congestion reporters			
Drivers					Observation, FCD	Observation			
Motoring Clubs					Service vehicles, congestion reporters, FCD	Service vehicles, congestion reporters			
RTTI Services		Own detection systems			Own detection systems	Model-based data interpretation			Model based Data Interpretation
Event managers								Event Schedules	

[a] FCD, floating car data; FPD, floating phone data.

Encyclopedia of Automotive Engineering. Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For the target to establish high quality RTTI services, it is recommendable to exploit all existing data pools and to avoid double investments and competing parallel and possibly contradictory information.

Cooperation between public actors and commercial companies needs to be based on clear technical, organizational, and operational regulations. European regulations, national legislation, and existing setups of RTTI services must be considered.

6.3.3 Organizational models

Access to foreign data pools was a matter of bilateral technical (proprietary data formats, coding, semantics, and exchange protocols) and commercial agreements. A variety of individual access points and gateways are needed if no coordinated access rules exist (Figure 4).

6.3.3.1 National Data Warehouse for Traffic Information. Different possibilities to organize access to separated data pools exist (Cornelissen, 2011). In the Netherlands, for instance, the so-called National Data Warehouse (NDW) for Traffic Information is operating. The reason is to improve the nationwide overview of traffic situation and the inadequate quality of data from different sources that did not allow for reliable traffic information and selection of suitable traffic management measures. This, as consequence, jeopardized optimal traffic management at the network level and resulted in more congestion than necessary.

NDW acts as a national database that pools data from all suitable sources to provide accurate and consistent traffic data and information. It acts as a single point of contact for all data suppliers on the one hand and traffic managers and service providers as users on the other.

NDW collects and distributes real-time traffic data (speed, traffic intensity, travel time, and classification), traffic situation information, and traffic management measures as well as statistics.

Data suppliers have to prove the quality of their data, which is checked regularly by NDW. Different quality levels have been defined according to road categories and importance. Loops, cameras, and floating phone data are used as data collection technologies; Bluetooth as a new technology has been tested (Cornelissen, 2011).

6.3.3.2 Mobility Data Market Place (Mobilitäts Daten Marktplatz). A different approach is the creation of a Mobility Data Market Place (Mobilitäts Daten Marktplatz, MDM) (Mobilitäts Daten Marktplatz Benutzerhandbuch, 2012). An MDM enables an internet portal, the MDM platform, to offer, search, and subscribe online traffic data or other data with relevance for traffic as well as to distribute these online data between data suppliers and data clients.

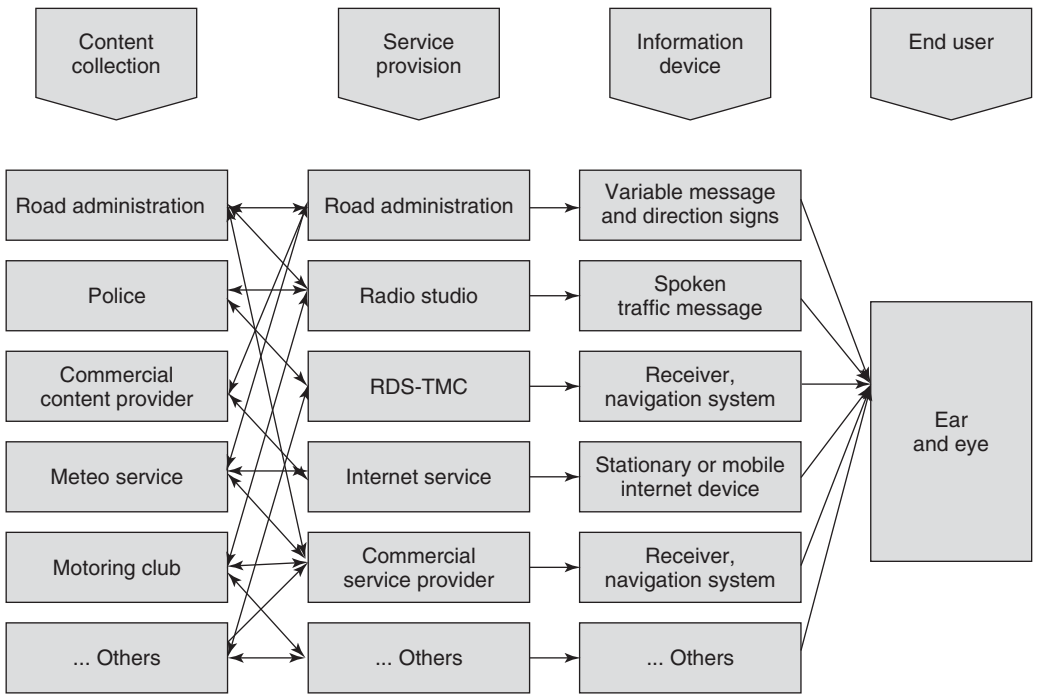


Figure 4. Multiple data sourcing and complex networking between content and service providers.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

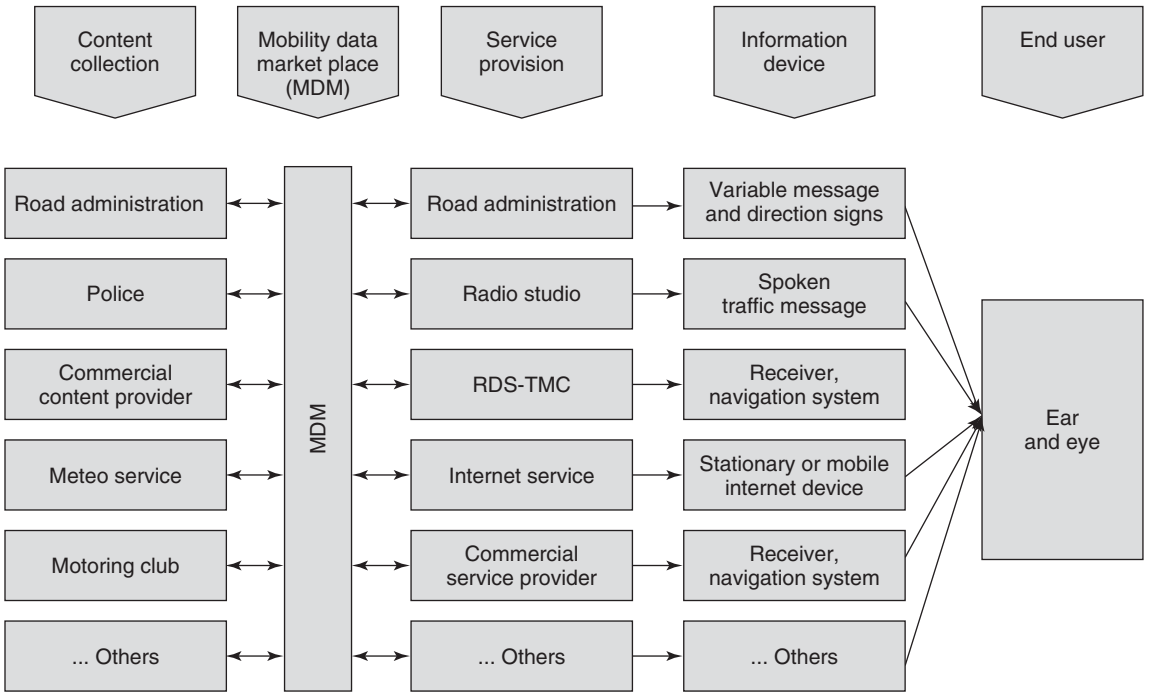


Figure 5. Multiple data sourcing and simplified networking between content providers and service providers via Mobility Data Market Place (MDM).

The function of MDM is to enable the exchange of data without checking or changing them. It is not its task to harmonize contents of data from different sources, but to enable access to different databases via one single portal (Figure 5). In contrast to the Dutch NDW, data and information are not held in a central data bank; all data handling and storage is decentralized. Users can download information about available data, their formats, interfaces, and commercial conditions to exploit data pools.

7 FRAME CONDITIONS AND REGULATIONS

The roles and task distributions between public partners and commercial actors are extremely complex questions. On the one hand, the supply of drivers with traffic information is considered as a very important contribution to traffic safety; on the other, commercial actors may consider public RTTI service as a competing service limiting their market chances.

The above-mentioned organizational schemes are technical provisions creating possibilities to collect and exchange data and information from different actors as the basis for RTTI; however, they do not touch upon the problem of which organization—public or commercial services—should deliver which type of information to drivers as end users and under what conditions: which type of information should be available free of charge and which should be paid for. Discussions differ from country to country, depending on their RTTI history.

In some countries, commercial actors that had been the first to distribute RTTI are operating pay-services exclusively for their customers. Nonsubscribers having no access to their RTTI have to restrict themselves to less developed public information sources.

Public RTTI services, for example, as operated jointly by road administrations, police, and public broadcasters in Germany, are mandated by their political assignment to take care of traffic safety, respectively, to inform their listeners as best as possible: This includes well-developed RTTI and has been done for decades. Emerging commercial RTTI services in such counties try to "limit" these public activities in order to get a broader market share.

7.1 eSafety Initiative of the European Commission

European integration with increasing cross-border traffic requires harmonized conditions for RTTI services to enable barrier-free access to traffic information. After intensive discussions, in order for "traffic information" to be

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Conclusions of the Conference on the European Commission's eSafety Initiative 5/6 June 2007.

1. *The Conference agreed that traffic information is highly relevant to road safety. Both the users and the players involved in the information chain are calling for further improvements in terms of quality and coverage in order to meet the objectives of improving road safety and to assist the driver in performing his driving task. Realistic and feasible quality standards should be developed jointly, with the involvement of all stakeholders.*
2. *The Conference notes that traffic information services should cover not only the primary networks (e.g., motorways) but also sizeable sections of the secondary networks. To this end, "strategic networks" are to be defined. Here, it is quite conceivable that different quality levels could be applied to the individual network sections.*
3. *The increasing number of traffic messages makes it necessary to transmit traffic information digitally so that it can be automatically processed in an appropriate manner, both during the generation and management of the messages and by the users' receivers. For this purpose, the "Radio Data System Traffic Message Channel (RDS-TMC)" has been developed, which is already in operation in numerous European countries, albeit with varying degrees of intensity.*
   If the secondary networks are to be included, it might become necessary to change over from analog radio channels to digital broadcast channels (e.g., DAB, DRM). To ensure universal coverage in conurbations, provision will have to be made for appropriate transmission capacity. Joint implementation strategies should be developed on a Europe-wide basis to facilitate access to this information, which is not based on a specific language. The aim is to create relatively uniform information services within the European Union.
4. *In Europe, there are both freely accessible "public" traffic information services and commercial information services, which users can access by paying a fee. The Conference notes that, from a transport policy perspective, access to safety-related traffic information should be possible worldwide without users having to pay additional costs. Some countries have already categorized safety-related information by way of example. The minimum scope of safety-related traffic information should also be defined on a Europe-wide basis. This will not rule out the possibility of individual countries going beyond this scope when providing freely accessible information.*
5. *Commercial information services have their place alongside freely available traffic information services. The services they offer may go significantly beyond those offered by public information services and cater to the individual needs of customers.*
6. *The Conference believes that Member States should, in accordance with the principle of subsidiarity, also make the necessary rules and arrangements for the free provision of safety-related traffic information within the framework of public–private partnerships (PPPs).*

(Conclusions of the Conference on the European Commission's eSafety Initiative 5/6 June 2007.)

preferably considered as "indispensable contribution to transport policy objectives" or as "commercially exploitable merchandise," a balanced proposal was made by a European working group of experts (eSafety RTTI WG, 2007) in the context of the European eSafety Initiative of the European Commission and forwarded to the EU Commission and Council.

The European eSafety Initiative was set up with the objective to cut down the number of fatalities on European roads. Experts discussed all types of new technologies and their possible contribution to this objective. Besides a number of merely vehicle-bound systems, systems with vehicle-to-vehicle and vehicle-to-infrastructure communication enabling immediate interactions between vehicles and road infrastructure RTTI services are highlighted as promising means to improve safety and traffic flow. RTTI is acknowledged as a means with the possibility to achieve results very soon because of the fact that RTTI systems are already operational in many countries, although with differing quality and with high potential for improvement. International working groups elaborated quality criteria for RTTI systems and their subsystems and proposed regulations to enable coexistence of public and commercial RTTI services.

The Conference on the European Commission's eSafety initiative on 5/6 June 2007 concluded on the key issue of Real-Time Traffic Information (Table 3) (Communication from the Government of the Federal Republic of Germany, 2007).

7.2 Free access to "safety-related traffic information"

These demands were taken up by the European Commission and Parliament, for example, in Directive 2010/40/EU of the European Parliament and the Council of 7 July 2010 on the framework for the deployment of Intelligent Transport Systems in the field of road transport and for interfaces with other modes of transport (eSafety RTTI WG, 2007).

This directive states among others as priority actions

(a) the provision of EU-wide multimodal travel information services;
(b) the provision of EU-wide real-time traffic information services;
(c) data and procedures for the provision, where possible, of road-safety-related minimum universal traffic information free of charge to users, and so on.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Besides these European regulations, national rules exist, which, if necessary, have to be adapted to these frame conditions.

Further development and deployment of RTTI systems as a prioritized ITS within the European Union need to take these frame conditions into account. It is also laid down in the directive of the European Parliament and Council (2010): "To ensure a coordinated and effective deployment of ITS within the Union as a whole, specifications, including, where appropriate, standards, defining further detailed provisions and procedures should be introduced. This latter regulation aims at providing international application and usage of Intelligent Transport Systems."

7.3 Definition of "safety-related traffic information"

For the definition of safety relevance, traffic messages had to be categorized. The idea of a classification (Table 4) as unanimously agreed by the German Traffic Information Platform in March 2006 (BMVBS/BASt, 2000–2008) was generally accepted as the basis for the following discussions at the European level.

The concept of a minimum set of important safety-related messages is still being discussed. Different sectors of the service chain, public authorities, commercial service providers, etc. all have slightly different views of which traffic messages are most important. There is a basic understanding that each meaningful message may be safety related as it reduces uncertainty and thus eliminates a source of potential stress for the driver. However, for example, the same message might be "safety related" if received by drivers immediately approaching an incident or may be just "informative" for drivers still far away. Further, the ownership of gathered data and the further processing of data, eventually combined with specific, value-adding knowledge bases, may create the need for specific arrangements of business models and responsibilities. Therefore, Member States will have to establish solutions based on their given conditions. The European Commission will have to take necessary steps to coordinate the cross-border flow of information. (eSafety RTTI WG, 2007).

Table 4. Classification of traffic messages by the German Traffic Information Platform (AG TMC-VID, March 2006) (BMVBS/BASt, 2000–2008).

Level	Hazard Prevention	Event Examples
I	+++++	High-risk situations, for example, "wrong-way drivers"
II	++++	High-risk situations, for example, "people, animals, or foreign objects on the road"
III	+++	Disruptions, for example, "road closure"
IV	+++	Disruptions, for example, "standstill"
V	+++	Disruptions, for example, "slow-moving traffic"
VI	++	Restoring normal traffic flow by cutting delays
VII	+	Maintaining smooth traffic flows with a view to lessening the economic impact

(Reproduced from BMVBS/BASt, 2000–2008.)

Discussions about safety relevance are not yet finished. In Table 5 (European Commission, DG TREN, 2008), column 1 lists traffic problems and messages as defined in the TISA list of Events (Traveller Information Service Association (2), 2012). Column 2 lists the TSIS-DATEX II definitions (CEN/TC 278/WG8 and ISO/TC 204/WG9). Column 3 tries to better describe events in a more concrete form. DATEX II incident codes can be mapped to TMC event codes.

The assessment of safety relevance differs slightly from country to country, as mentioned earlier. Some road administrations consider RTTI as an integral element and tool for traffic management. RTTI helps avoid accidents and development or prolongation of congestion by early information to drivers and supporting the efficiency of roadside VMSs and dynamic route guidance. To achieve these effects, it is of high importance that RTTI can address all drivers concerned and this can be expected to be better if RTTI is available free of charge. This exceeds the minimum requirements as laid down in the directive of the European Parliament and Council (2010) and further elaborated in the final report by the European Commission, DG TREN (2008), which must be seen as a compromise.

8 ACQUISITION OF AND ACCESS TO DATA

8.1 Payment models

Public as well as commercial RTTI content and messages present a value created by one or several actors in a value chain. Each contributor creates a part of the value of the end product. RTTI as end product is paid for. Public RTTI is financed, for example, by tax and/or by radio license fees,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 5. Classification of traffic messages by TISA and DATEX.

TISA Definition	TSIS-DATEX II Definition	Description
Ghost driver	Included in the definition for category 4	Vehicle on wrong carriageway of a dual-carriageway road
Dangerous road surface	All traffic elements of class Road conditions	All specified road conditions, for example, black ice, icy patches, oil on road, poor road surface, etc.
Danger due to reduced visibility	All traffic elements of class Poor environment conditions where minimum visibility distance is specified	Any environmental condition reducing visibility. For example, hail or fog with visibility of less than 50 m
Animal/people/debris in the road way	All traffic elements of class obstructions or disturbance activity or authority operation	All obstructions on the road (e.g., fallen trees, animals, burning vehicle). All public disorder or alerts with the potential to disrupt traffic (e.g., demonstration, assault). All authority-initiated operation or activity that could disrupt traffic (e.g., police investigation, bomb squad in action)
Blockade of road, tunnels	All operator actions of class network management	All road operator activities to manage the road network, for example, rerouting, road/lane closures, warnings (e.g., reduce speed) and recommendations (e.g., snow tires)
Unprotected accident area	All traffic elements of class accidents	All accidents
Temporary roadwork	All operator actions of class Road works where urgent road works = true	All urgent repairs on the road network
End of queue	All traffic elements of class abnormal traffic where urgency =(urgent or extremely urgent)	Any congestion that is uncommon
Comments: The definition excludes nonurgent information, that is: • Measured data, for example, real-time and predicted travel times • "Abnormal traffic," that is, incidents reporting on traffic flow • Road works • Equipment or system failures • Weather conditions not related to the road surface • Public events • Non-road-related information, for example, parking information The excluded incident reports can be defined as safety-related if labeled "urgent," for example, in case of a tailback at a dangerous location, or a blizzard warning.		

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 6. Payment models (eSafety RTTI WG, 2007).

1.	TTI "free," as public service user paid via tax (internet, radio)
2.	TTI included in broadcasting service fee (RDS/TMC via FM, DAB)
3.	TTI included in purchase of equipment (radio, navigation system)
4.	TTI included once-per-lifetime (purchase of car, radio, navigation system)
5.	TTI as subscription service (Automobile Club, service provider)
6.	TTI pay-per-use (toll number)
7.	Manufacturer pays service provider for end-user service (subscription)
8.	Service provider pays traffic information center (TIC) for TTI
9.	Traffic information center pays content provider(s) for data
10.	Public authorities co-finance broadcasting infrastructure TTI (FM, DAB)
11.	Public authorities (co-)finance private service provider(s)
12.	Public authorities (co-)finance TIC and use TTI for their own traffic management responsibilities
13.	Public authorities (co-)finance content gathering
14.	Service providers pay telco's for delivery of TTI to their customers
other	Each party has to finance their investments and operations

(Reproduced from eSafety RTTI WG, 2007. © European Commission.)

and commercial RTTI by a number of different payment methods (Table 6) (eSafety RTTI WG, 2007).

8.2 Value chain

It is reasonable that cooperation of public and commercial actors is realized in order to reach adequate quality of RTTI. Each partner can contribute his/her information or his/her content-handling procedures to increase the quality of RTTI. The value chain includes actions form different actors from content provision to message distribution. Figure 6 displays a generic model of this value chain.

The contribution of all partners has to be reimbursed according to the value added. The value of contributions is subject to economic considerations and negotiations between content providers and RTTI services. Major aspects in the commercial calculations are

- costs of content provision and handling
- (free) availability of public data
- "value" of data as seen by RTTI service provider (reliability, timeliness, etc.; see Table 7)
- willingness of end users to pay.

8.3 Quality levels of RTTI content and service

Some criteria to assess the value of data and information of RTTI are listed and defined in Table 7 (EasyWay (1), 2012).

These core criteria include factors to assess message contents as well as the service as experienced by the driver as end user.

It is economically not feasible to apply a 100% quality level of content and service to all parts of road networks. Suitable levels need to be defined according to criticality of roads, road layout and category, and traffic importance.

EasyWay (EasyWay (4), 2010) notes that the operating environments are obtained through a qualitative approach in order to make it simple and easily used in any part of the road network. The general layout starts by assigning a letter code to the different physical characteristics:

- C for critical spots (bridges, tunnels, reversible lane sections, etc.)
- T for motorways
- R for roads
- S for motorway corridors or networks
- N for road corridors or networks
- P for peri-urban road networks

With this general classification, EASYWAY proposes a more detailed schedule of roads as displayed in Table 8 (EasyWay (1), 2012).

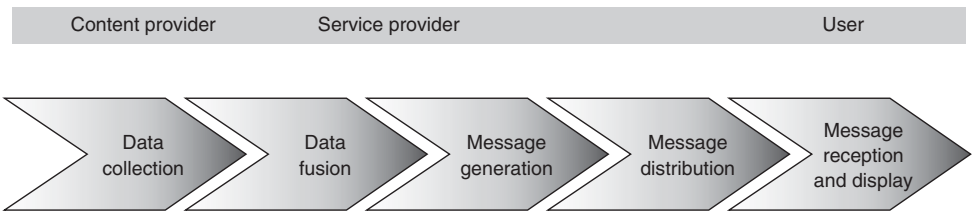


Figure 6. Generic model of RTTI value chain.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 7. Criteria to assess levels of quality (EasyWay (1), 2012).

Core Criteria	Definition
Accessibility	*Qualifies the user's exposure to the information service* For instance, the number of users reached by the information on each of the technology platforms deployed in relation to the % km of the relevant area concerned
Availability	*Qualifies the period during which the availability of the service to the specified standard is defined/guaranteed*. Degree to which (geographic) data would be available at a certain place and at a defined time. This may be by time of day/week/month or other special periods (e.g., 24/7 excluding national holiday periods)
Timeliness	*Time delay between the event detection and the provision of information to the end-user* (around 10s with new technologies)
Update frequency	*Qualifies the frequency of updating information or data update interval*. For instance, on occurrence (updates are made as and when a change occurs), periodic (regular or periodic updates) or on request update, etc
Quality assurance	*Incorporates the need to undertake partial checks* if deemed appropriate for parts of the chain (Level 1 for instance) or a full check of the service chain (Level 2) or a full check, which has the additional quality assurance through application of standards
Reliability (cross verified)	*Degree of certainty of the information*, considering whether the data value has been cross-verified from one or more additional sources, confirming the data
Accuracy	Degree of adherence of (geographic) data to the most plausible true value (can be absolute, relative, quantitative, and temporal)

(Reproduced with permission from EasyWay (1), 2012. © EasyWay ITS. EasyWay Deployment Guideline Document, Final Version December 2012 (www.easyway-its.eu).)

Table 8. EasyWay operating environments for Core European ITS Services.

C1	Critical spots, local flow-related traffic impact and/or potential safety concerns
T1	Motorway (link), no flow-related traffic impact and no major safety concerns
T2	Motorway (link), no flow-related traffic impact, potential safety concerns
T3	Motorway (link), seasonal or daily flow-related traffic impact, no major safety concerns
T4	Motorway (link), seasonal or daily flow-related traffic impact, potential safety concerns
R1	Two-lane road (link), no flow-related traffic impact, no major safety concerns
R2	Two-lane road (link), no flow-related traffic impact, potential safety concerns
R3	Two-lane road (link), seasonal or daily flow-related traffic impact, no major safety concerns
R4	Two-lane road (link), seasonal or daily flow-related traffic impact, potential safety concerns
R5	Three-/four-lane road (link), no flow related traffic impact, no major safety concerns
R6	Three-/four-lane road (link), no flow related traffic impact, potential safety concerns
R7	Three-/four-lane road (link), seasonal or daily flow related traffic impact, no major safety concerns
R8	Three-/four-lane road (link), seasonal or daily flow related traffic impact, potential safety concerns
S1	Motorway corridor or network, at most seasonal flow-related impact, possibly safety concerns
S2	Motorway corridor or network, daily flow-related traffic impact, possibly safety concerns
N1	Road corridor or network, at most seasonal flow-related traffic impact, possibly safety concerns
N2	Road corridor or network, daily flow-related traffic impact, possibly safety concerns
P1	Peri-urban motorway or road interfacing urban environment, possibly safety concerns

(Reproduced with permission from EasyWay (1), 2012. © EasyWay ITS. EasyWay Deployment Guideline Document, Final Version December 2012 (www.easyway-its.eu).)

The aim of the classification in Table 8 is to provide harmonized RTTI services allover Europe so that a road user can expect similar RTTI services in comparable road environments. EasyWay outlines a scheme of levels of quality for traffic condition and travel time information services (Table 9) (EasyWay (2), 2012).

9 END USER'S ACCESS TO RTTI

Drivers as end users can access RTTI over a number of information and communication channels, enabling them to get general as well as individually targeted information. Pre-trip information can be accessed in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 9. Levels of quality table: traffic condition and travel time information services.

Criteria	0	1	2	3
Accessibility	Only on hotspots	On main routes where problems often occur	On all routes	
Availability	Not guaranteed	Guaranteed to a minimum level	Guaranteed to a medium level	Guaranteed all the time
Timeliness	Not guaranteed 30 min	15 min	1 min around 10 s	
Update frequency	Only on an irregular basis	On a regular basis	As frequent as currently possible respectively as a significant change of traffic conditions	
Quality assurance	No regulation	Either input or output (partial check)	Service chain check (full)	Information Quality Assurance
Cross verified	Not defined	Data from one or more sources—reliability not confirmed	Data from one or more sources—reliability confirmed	Collaboration from more than one source (data fusion) reliability confirmed
Accuracy	N/A	N/A	N/A	N/A
Service grade	Not guaranteed	Guaranteed to a time interval	In real time	
Forecast horizon	Only current situation	Current situation and short time prediction	Current situation and short as well as long time prediction	

Legend:
- Accuracy: this criterion is covered by
  - LoS: level of detail and
  - LoQ: forecast horizon
- Service grade: A specific traffic condition and travel time information quality criterion, in which the service is guaranteed
- Forecast horizon: A specific traffic condition and travel time information quality criterion, which defines the prediction time for the service

(Reproduced with permission from EasyWay (1), 2012. © EasyWay ITS. EasyWay Deployment Guideline Document, Final Version December 2012 (www.easyway-its.eu).)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 10. Information and communication channels for drivers as end-users.

Time	Information/ Communication Channel	Used for
Pre-trip	Print media, radio, TV, Internet portals, telephone services	Information used for trip planning, for example, work sites, expected traffic situation, expected usability of roads, bridges, tunnels, vehicle-class-specific information, etc.
On-trip	Variable message signs (VMS)	Mandatory signs (e.g., temporary speed limits), lane allocation, temporary use of emergency lanes for flowing traffic, warning signs Limited information in order to limit mental load for drivers
On-trip	Variable direction signs (VDS)	Dynamic rerouting, management of event traffic
On-trip	Spoken traffic messages	Limited amount of general traffic information
On-trip	RDS-TMC messages	Up to 300 messages with limited local resolution, used and displayed in navigation devices
On-trip	TPEG messages	High number of detailed messages with high local resolution, precise information for navigation devices, and route calculation
On-trip	Mobile phones	Information on demand, potential to integrate internet functions yielding permanent or temporary access to data/information sources

the planning phase of trips to decide about departure time, expected arrival time, route to be used, or choice of transport mode (vehicle, public transport, train, and airplane). On-trip information helps drivers to choose an adequate way of driving, creates better awareness of oncoming traffic or road problems, and enables alternative route choices. Some main possibilities are listed in Table 10.

10 COEXISTENCE AND COOPERATION OF PUBLIC AND COMMERCIAL RTTI

In some countries, public and commercial RTTI services have agreed to cooperate, each side taking advantage from joint efforts, complementing each other, and avoiding duplication of efforts and expenses. Although some overlap

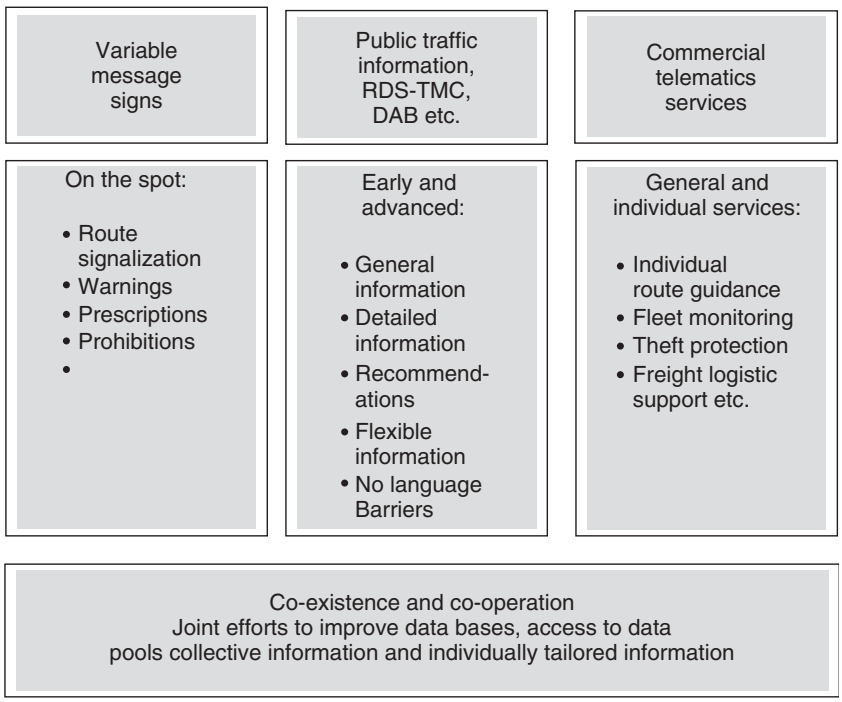


Figure 7. Synergy of public and commercial RTTI services. (Reproduced with permission from Bolte, 1997–2007. © Fritz Bolte.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

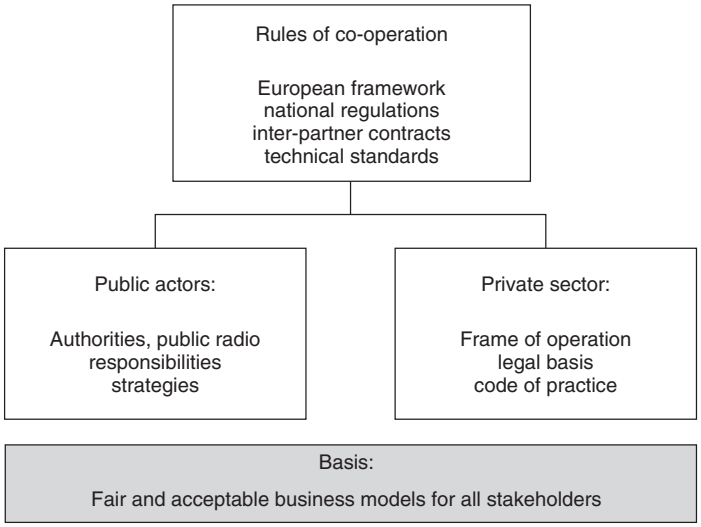


Figure 8. Hierarchy of regulations and contracts.

is possible, each side has specifically targeted services to offer. Public RTTI supplements other information media for traffic safety and management purposes, transmitting general information relevant for all drivers in the affected road or area. Commercial RTTI services can deliver more detailed and "individualized" information, tailored to the individual driver (Figure 7).

European frame conditions, completed by adequate national regulations, suitable business models, and contracts between partners, enable coexistence and fair cooperation of public and commercials RTTI services (Figure 8).

It should be noted that European regulations form the framework for RTTI services in Europe. Drivers should find consistent RTTI services in Europe. The request to deliver safety-relevant information free of charge, without extra payment, addresses the European Member States; it is their task to find suitable solutions in the context of their specific situation.

11 IMPORTANT CEN/ISO STANDARDS (EASYWAY (1), 2012)

CEN ISO/TS 14822–1:2006 Traffic and Travel Information—General specifications for medium-range preinformation via dedicated short-range communication—Part 1: Downlink (ISO/TS 14822–1:2006)-

CEN ISO/TS 14823:2008 Traffic and Travel information—Messages via media-independent stationary dissemination systems—Graphic data dictionary for pre-trip and in-trip information dissemination systems (ISO/TS 14823:2008)-

CEN ISO/TS 14907–1:2010/AC:2010Road transport and traffic telematics—Electronic fee collection—Test procedures for user and fixed equipment—Part 1: Description of test procedures (ISO/TS 14907–1:2010/Cor 1:2010)-

CEN ISO/TS 18234–1:2006 Traffic and Travel Information (TTI)—TTI via Transport Protocol Expert Group (TPEG) data streams—Part 1: Introduction, Numbering, and Versions (ISO/TS 18234–1:2006)-

CEN ISO/TS 18234–2:2006 Traffic and Traveler Information (TTI)—TTI via Transport Protocol Expert Group (TPEG) data streams—Part 2: Syntax, Semantics, and Framing Structure (SSF) (ISO/TS 18234–2:2006)-

CEN ISO/TS 18234–3:2006 Traffic and Travel Information (TTI)—TTI via Transport Protocol Expert Group (TPEG) data streams—Part 3: Service and Network Information (SNI) application (ISO/TS 18234–3:2006)-

CEN ISO/TS 18234–4:2006 Traffic and Travel Information (TTI)—TTI via Transport Protocol Expert Group (TPEG) data streams—Part 4: Road Traffic Message (RTM) application (ISO/TS 18234–4:2006)-

CEN ISO/TS 18234–5:2006 Traffic and Travel Information (TTI)—TTI via Transport Protocol Expert Group (TPEG) data streams—Part 5: Public Transport Information (PTI) application (ISO/TS 18234–5:2006)-

CEN ISO/TS 18234–6:2006 Traffic and Travel Information (TTI)—TTI via Transport Protocol Expert Group (TPEG) data streams—Part 6: Location referencing applications (ISO/TS 18234–6:2006)-

CEN ISO/TS 24530–1:2006 Traffic and Travel Information (TTI)—TTI via Transport Protocol Experts Group (TPEG) Extensible Markup Language (XML)—Part 1:

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Introduction, common data types, and tpegML (ISO/TS 24530–1:2006)-
CEN ISO/TS 24530–2:2006 Traffic and Travel Information (TTI)—TTI via Transport Protocol Experts Group (TPEG) Extensible Markup Language (XML)—Part 2: tpeg-locML (ISO/TS 24530–2:2006)-
CEN ISO/TS 24530–3:2006 Traffic and Travel Information (TTI)—TTI via Transport Protocol Experts Group (TPEG) Extensible Markup Language (XML)—Part 3: tpeg-rtmML (ISO/TS 24530–3:2006)-
CEN ISO/TS 24530–4:2006 Traffic and Travel Information (TTI)—TTI via Transport Protocol Experts Group (TPEG) Extensible Markup Language (XML)—Part 4: tpeg-ptiML (ISO/TS 24530–4:2006)-
CEN/TS 14821–1:2003 Traffic and Travel Information (TTI)—TTI messages via cellular networks—Part 1: General specifications-
CEN/TS 14821–2:2003 Traffic and Travel Information (TTI)—TTI messages via cellular networks—Part 2: Numbering and ADP message header-
CEN/TS 14821–3:2003 Traffic and Travel Information (TTI)—TTI messages via cellular networks—Part 3: Basic information elements-
CEN/TS 14821–4:2003 Traffic and Travel information (TTI)—TTI messages via cellular networks—Part 4: Service-independent protocols-
CEN/TS 14821–5:2003 Traffic and Travel Information (TTI)—TTI messages via cellular networks—Part 5: Internal services-
CEN/TS 14821–6:2003 Traffic and Travel Information (TTI)—TTI messages via cellular networks—Part 6: External services-
CEN/TS 14821–7:2003 Traffic and Travel Information (TTI)—TTI messages via cellular networks—Part 7: Performance requirements for onboard positioning-
CEN/TS 14821–8:2003 Traffic and Travel Information (TTI)—TTI messages via cellular networks—Part 8: GSM-specific parameters-
EN 12253:2004 Road transport and traffic telematics—Dedicated short-range communication—Physical layer using microwave at 5,8 GHz-
EN 12795:2003 Road transport and traffic telematics—Dedicated short-range communication (DSRC)—DSRC data link layer: medium access and logical link control-
EN 12834:2003 Road transport and traffic telematics—Dedicated short-range communication (DSRC)—DSRC application layer-
EN 12896:2006 Road transport and traffic telematics—Public transport—Reference data model-
EN 13372:2004 Road Transport and Traffic Telematics (RTTT)—Dedicated short-range communication—Profiles for RTTT applications-
EN ISO 14819–1:2003 Traffic and Travel Information (TTI)—TTI Messages via traffic message coding—Part 1: Coding protocol for Radio Data System—Traffic Message Channel (RDS-TMC) using ALERT-C (ISO 14819–1:2003)-
EN ISO 14819–2:2003 Traffic and Traveler Information (TTI)—TTI Messages via traffic message coding—Part 2: Event and information codes for Radio Data System—Traffic Message Channel (RDS-TMC) (ISO 14819–2:2003)-
EN ISO 14819–3:2004 Traffic and Travel Information (TTI)—TTI messages via traffic message coding—Part 3: Location referencing for ALERT-C (ISO 14819–3:2004)-
EN ISO 14819–6:200 Traffic and Traveler Information (TTI)—TTI messages via traffic message coding—Part 6: Encryption and conditional access for the Radio Data System—Traffic Message Channel ALERT C coding (ISO 14819–6:2006)-
EN ISO 14825:2011 Intelligent transport systems—Geographic Data Files (GDF)—GDF5.0 (ISO 14825:2011)-
ENV 12313–4:2000 Traffic and Traveler Information (TTI)—TTI Messages via Traffic Message Coding—Part 4: Coding Protocol for Radio Data System—Traffic Message Channel (RDS-TMC)—RDS-TMC using ALERT Plus with ALERT C-
ENV 12315–1:199 Traffic and Traveler Information (TTI)—TTI Messages via Dedicated short-range communication—Part 1: Data Specification—Downlink (Roadside to Vehicle)-
ENV 12315–2:1996 Traffic and Traveller Information (TTI)—TTI Messages via Dedicated short-range communication—Part 2: Data Specification—Uplink (Vehicle to Roadside)-
ENV 13998:2001 Road transport and traffic telematics—Public transport—Noninteractive dynamic passenger information on ground-

12 GLOSSARY (ESAFETY RTTI WG, 2007)

3G	"Third Generation"/Universal Mobile Telecommunications System (UMTS) mobile telephone services, characterized by high performance and high bandwidth data services
AGORA-C	Flexible on-the-fly (does not need pre-encoded locations) location referencing method

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ALERT-C	Traffic information encoding used for TMC messages—uses pre-encoded event description and location reference, decoded in the receiver using look-up tables
DAB	Digital Audio Broadcast—digital method of broadcasting, offering higher data capacity than FM radio channels
DRM	Digital Radio Mondiale—digital audio broadcasting for AM broadcast, which can fit more channels than AM, at higher quality, into a given amount of bandwidth
DVB-T	Digital Video Broadcast—Terrestrial—method of broadcasting audio, video, and other data as already used for terrestrial digital television broadcast
FPD	Floating Phone Data—as FVD but monitoring the location of mobile telephones—can be performed at GSM network level without requiring special hardware in the telephone or vehicle
FVD, FCD	Floating Vehicle (Car) Data—monitoring the locations and movements of a set of vehicles, such as through collecting locations regularly using GPS devices mounted in a fleet of vehicles, and using the data (usually anonymized) to understand more about the overall road conditions and congestion
ITS	Intelligent transport systems
LoQ	Level of quality
LoS	Level of service
RDS-TMC, TMC	Radio Data System Traffic Message Channel—a method of sending traffic information as silent messages alongside regular radio broadcasts, optimized for FM broadcast and typically received by compatible satellite navigation systems and used for driver information and dynamic navigation (rerouting). TMC is an ODA (Open Data Application) built upon RDS
RBDS	North American standard for FM radio data. Very closely related to RDS, but with some small differences that must be taken into account when implementing TMC or other data services
RTTI	Real (Real-time) Traffic and Travel Information
TISA	Traveler Information Services Association is a market-driven membership association with worldwide scope, established as a nonprofit company focused on proactive implementation of traffic and travel information services and products based on existing standards, including primarily RDS-TMC and TPEG technologies
Telematics	Telematics is a combination of the subjects "telecommunications" and "informatics." "Transport Telematics" is the application of telematics on the whole field of transport and is the basis for "Intelligent Transport Systems (ITS)"
TM	Traffic Management
TPEG	Transport Protocol Expert Group—Method of encoding and sending traffic and travel information as silent messages alongside regular radio broadcasts, optimized for DAB broadcast but also applicable to other digital bearers; requires more bandwidth than TMC but can exploit this bandwidth with more/richer services
VDS	Variable direction sign
VMS	Variable message sign

REFERENCES

BMVBS/BASt (2000–2008) Minutes of the German Traffic Information Platform, Federal Ministry of Transport, Building and Urban Development/Federal Highway Research Institute (BMVBS / BASt).

Bolte, F. (1997–2007) Real Time Traffic Information (RTTI), numerous publications and presentations at ITS World Congresses 1997–2007 and other conferences, a.o. Report given in EUROTRAVEL Conference 2008 organised by the European Broadcasting Union (EBU), Ljubljana.

Communication from the Government of the Federal Republic of Germany (2007) Communication from the Government of the Federal Republic of Germany to the European Commission of 27 June 2007 concerning eSafety Conference in Berlin on 5/6 June 2007.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

J. Cornelissen (2011) National Data Warehouse for Traffic Information 'A deliberate quality strategy'; Report Workshop Bundesanstalt für Strassenwesen (BASt), Bergisch Gladbach 23 March 2011, Published by BASt F82.

EasyWay (1) (2012) EasyWay Doc. Traveller Information Services, Reference Document; TIS Deployment Guideline Annex; TIS-DG01|Version 02-00-00|DECEMBER 2012; Available at: <www.easyway-its.eu>.

EasyWay (2) (2012) EasyWay Doc. Traveller Information Services TRAFFIC CONDITION AND TRAVEL TIME INFORMATION SERVICE, Deployment guideline, TIS-DG03-05 | Version 02-00-00 | DECEMBER 2012.

EasyWay (4) (2010) EasyWay: Operating Environments; Available at: <www.EasyWay-its.eu>.

eSafety RTTI WG (2007) Report of the eSafety Working Group on Real-Time Traffic and Travel Information (RTTI). European Commission, eSafety Initiative WG RTTI, 19 March 2007.

European Commission, DG TREN (2008) ITS ACTION PLAN; Study regarding guaranteed access to traffic and travel data and free provision of universal traffic information. FRAMEWORK SERVICE CONTRACT TREN/G4/FV-2008/475/01; D8 –FINAL REPORT EUROPEAN COMMISSION, Directorate-General Mobility & Transport; Unit B4, Rue De Mot 28, 4/37, B-1040 Brussels Belgium.

European Parliament and Council (2010) Directive 2010/40/EU of the European Parliament and of the Council of 7 July 2010 on the framework for the deployment of Intelligent Transport Systems in the field of road transport and for interfaces with other modes of transport.

Mobilitäts Daten Marktplatz Benutzerhandbuch (2012) MDM-Benutzerhandbuch -Version 1.3.2—20.01.2012; Available at: <www.mdm-portal.de)>.

Traveller Information Service Association (2) (2012) Traveller Information Service Association (TISA): TISA EVENT List. Available at: <www.tisa.org>.

FURTHER READING

Amtsblatt der Europäischen Kommission vom 2.Dezember 2011, p. 169–171 (2011) Legislative Entschließung des Europäischen Parlaments vom 6. Juli 2010 zu dem Standpunkt des Rates aus erster Lesung im Hinblick auf den Erlass der Richtlinie des Europäischen Parlaments und des Rates zum Rahmen für die Einführung intelligenter Verkehrssysteme im Straßenverkehr und für deren Schnittstellen zu anderen Verkehrsträgern (06103/4/2010—C7-0119/2010—2008/0263(COD)) (2011/C 351 E/30) ICS und Schnittstellen 461BEd01; 2011/0430 (COD) Amtsblatt der Europäischen Union C 351 E/169 vom 2.12.2011.

Best Consortium (2012) BEST Consortium presents broadcaster's vision on future traffic and travel information services; Available at: <www.tis.org/newsroom/neas/>.

Bundesanstalt für Straßenwesen (2011) Qualität von Verkehrsinformationen im Straßenverkehr (Quality of on-trip road traffic information), BASt-Kolloquium 23. und 24.03.2011; Berichte der Bundesanstalt für Straßenwesen, Fahrzeugtechnik Heft F 82.

Burgess, Lisa, Toppen, Alan, Pretorius, Pierre, 2007. Services Business Models 2007. FHWA-HOP-07-115 RTTI.

CEN/TC 278/WG8 and ISO/TC 204/WG9. CEN/TC 278/WG 8 and ISO/TC 204/WG 9: ITS Standardisation: TSIS-DATEX II. Available at: <www.itsstandards.eu)>(S38)CEN/TC 278 Road transport and traffic telematics; Available at: <www.itsstandards.eu>.

Commission of the European Communities (2008) COMMISSION OF THE EUROPEAN COMMUNITIES, COMMUNICATION FROM THE COMMISSION Action Plan for the Deployment of Intelligent Transport Systems in Europe; Brussels, 16.12.2008 COM(2008) 886 final.

Council of the European Union (2009) Council of the European Union: Council conclusions on the commission communication: "Action Plan for the Deployment of Intelligent Transport Systems in Europe"; 2935th TRANSPORT, TELECOMMUNICATIONS and ENERGY Council meeting Brussels, 30 March 2009 (Doc. 106964.pdf).

EasyWay (3) (2012) EasyWay Doc. Traveller Information Services, FORECAST AND REAL TIME EVENT INFORMATION Deployment guideline TIS-DG02, Version 02-00-00 | DECEMBER 2012.

eSafety Conference (2007) Conclusions on the key issue of Real Time Traffic Information (RTTI) and vehicle-to-Infrastructure Communication V2V) as discussed at the eSafety, Conference; European Commission eSafety Conference, Berlin 2007.

eSafety Forum (2010) eSafety Forum: Implementation Road Maps, Monitoring Report 2010, Final Draft; The Implementation Road Maps Working Group, Brussels 15 March 2010.

European Commission (2010) European Commission: Vorschlag für eine Richtlinie des Europäischen Parlaments und des Rates zur Festlegung eines Rahmens für die Einführung intelligenter Verkehrssysteme im Straßenverkehr und für deren Schnittstellen zu anderen Verkehrsträgern (erste Lesung) Brüssel, VDE C 299/2 Amtsblatt der Europäischen Union 5.11.2010 (2010/C 299/01).

European Commission (2011a) EUROPÄISCHE KOMMISSION: WEISSBUCH, Fahrplan zu einem einheitlichen europäischen Verkehrsraum—Hin zu einem wettbewerbsorientierten und ressourcenschonenden Verkehrssystem, SEK(2011) 359 endgültig SEK(2011) 358 endgültig SEK(2011) 391 endgültig, Brüssel, den 28.3.2011 KOM(2011) 144 endgültig.

European Commission (2011b) Vorschlag für eine Richtlinie des Europäischen Parlamentes und des Rates zur Änderung der Richtlinie 2003/98/EG über die Weiterverwendung von Informationen des öffentlichen Sektors {SEK(2011) 1551 endgültig} {SEK(2011) 1552 endgültig}; Brüssel, Brüssel, den 12.12.2011 KOM(2011) 877 endgültig.

European Commission Decision (2011) EUROPEAN COMMISSION: COMMISSION DECISION of 15 February 2011 concerning the adoption of the Working Programme on the implementation of Directive 2010/40/EU; Brussels, 15.2.2011 C(2011) 289 final.

European Commission, DG Mobility and Transport (2011) European Commission, Directorate-General for Mobility and Transport: Action Plan and legal Framework for the Deployment of Intelligent Transport Systems; Brussels.

European Commission, DG Mobility and Transport (2) (2011) ITS Action Plan Leaflet Luxembourg: Publications Office of the European Union, 2011, ISBN 978-92-79-18475-8, Available at: <http://ec.europa.eu/transport/its/>.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Kleine, J. (2007) Vorstudie zur Qualitätsbewertung von Verkehrsinformationen; Bundesanstalt für Straßenwesen, Referat V2—Verkehrsbeeinflussung, Telematik, December 2007.

Mobilitäts Daten Marktplatz (1) (2012) MDM-Plattform, webarchive22. Available at: <www.mdm-portal.de>.

National Data Warehouse (1) (2012) National Data Warehouse for Traffic Information, The database explained. Brochure, NDW-NL; Available at: <www.ndw.nu>.

National Data Warehouse (2) (2012) National Data Warehouse for Traffic Information: The added value of NDW. Available at: <www.ndw.nu>.

Official Journal of the European Union of 31.-12.2003, p. 90–96. Directive 2003/98/EC: On the re-use of public sector information; Brussels 2003.

Pleitgen, F. (2007) President of European Broadcasting Union (EBU): The Role of European Public Service Broadcasters in the Transmission of Information for Traffic Safety and Traffic Management; Report to eSafety Conference; European Commission; Berlin 2007.

QUANTIS Project (2009–2011) QUANTIS Project. Available at: <www.quantis-project.eu>.

Service Platform of "Mobilitäts Daten Marktplatz," (2012) Service Platform of "Mobilitäts Daten Marktplatz"; Available at: <https://service.mdm-portal.de/mdm-portal-application/>.

Traveller Information Service Association (1) (2007–2012) Traveller Information Service Association (TISA):TISA Reports; Available at: <www.tisa.org>).c/o ERTICO, Brussels.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto183
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Logistics and Fleet Management

Oliver Kunze
Hochschule Neu-Ulm, University of Applied Sciences, Neu-Ulm, Germany

1 Introduction 1
2 Process Context of Telematics in Logistics 2
3 Tier 1—Focus: Vehicle and Driver 5
4 Tier 2—Focus: Forwarding Agent and Carrier 6
5 Tier 3—Focus: Shipper and Consignee 10
6 Technology and Data 13
7 Standardization 15
8 Summary and Outlook 16
Glossary 16
Abbreviations 17
Related Articles 17
Endnotes 17
References 18

1 INTRODUCTION

Telematics solutions for logistics and fleet management (LFM telematics) are becoming an integral element of transport operations worldwide.

A recent study (Dudek and Köppel, 2011) on a panel of 116 fleet operating companies (equivalent to a total amount of 10,446 vehicles, which mainly operate in Germany and Europe[1]) has shown that 48% of the companies in the panel had their fleet completely equipped with LFM telematics, 41% of the panel had a partly equipped fleet, and 11% were not equipped with LFM telematics solutions. The study also has shown that two predominant reasons for the implementation of LFM telematics solutions were "process improvement" and "documented evidence of service toward the customer."

At the same time, the National Transport Commission of Australia (NTC) states that "the road freight sector has not, however, embraced new and innovative technologies as fast or as wholly as many other industries. As a result, the full economic, social and environmental benefits to the Australian community have not been realized." (NTC, 2011, p. 1).

However, even if the implementation rate of LFM telematics may still show differences on a global scale, its benefits to transport service providers and their customers, as well as its benefits to the public (especially because of its technical means to enforce regulations on maximal driving times and thus reduce the risk of accidents because of driver's fatigue), both promote the use of LFM telematics.

1.1 Focus of chapter

This chapter focuses on *applications* of telematics in LFM. Technical details on communication (see Cellular Mobile Networks), positioning (see Technologies—Positioning: GNSS), or sensors (see In-vehicle sensors) are not discussed in this chapter.

After introducing some selected terms and definitions, this chapter explains the process context of LFM telematics, and the three user groups of LFM telematics systems. Then, the application scopes and functionalities for the different user groups are discussed. This chapter ends with a short overview on technical components needed for an LFM telematics solution and some thoughts on standardization.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

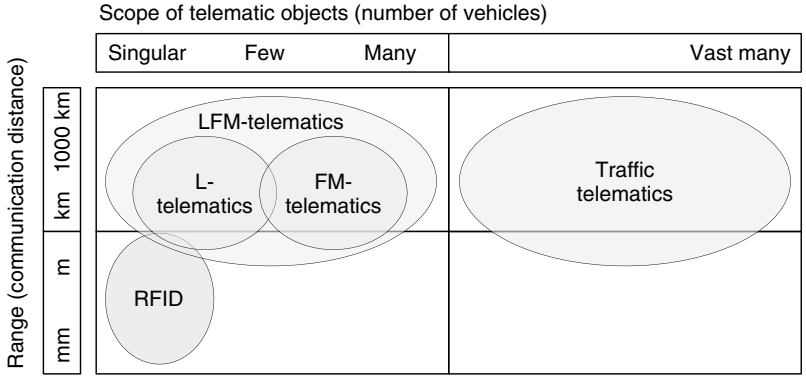


Figure 1. Delimitation of LFM telematics.

1.2 Terms and definitions

In this chapter, the term *telematics in logistics and fleet management* (*LFM telematics*) is used to describe the application of "information and telecommunication technology" (ITC) within the domain of transport logistics.

LFM telematics is different to "traffic-telematics" (see Applications—Intelligent Roads and Cooperative Systems: Urban Traffic Management, Advanced Highway Management Systems, and Road Traffic and Travel Information (RTTI)), as LFM telematics is used to support planning, execution, and monitoring of transports (L-telematics) as well as to support scheduling and monitoring of individual vehicles (FM-telematics), whereas traffic-telematics is used to monitor and control traffic flows.

Another ITC technology that is widely used in transport logistics is RFID (radio-frequency identification). It can identify objects over relatively short distances. However, even if in some cases, RFID technology may be put to use within LFM telematics applications, LFM telematics is clearly different to RFID.[2]

Using the two dimensions, "scope" (number of vehicles involved) and "range" (communication distance), the terms discussed earlier may roughly be positioned in relation to each other as shown in Figure 1. For definition of further terms, please refer to the glossary.

2 PROCESS CONTEXT OF TELEMATICS IN LOGISTICS

Key to understand the value of LFM telematics is a process-centered view of transport logistics.

2.1 Users, roles, and tasks

The transport logistics process has two main stakeholders—the customers (i.e., the shipper and the consignee of a transport) and the transport logistics operators (i.e., the service providers involved in the transport of goods—this includes forwarding agents, carriers, and in-company units, which plan and execute goods transports).

Each customer may employ several staff members who are involved in (or affected by) the execution of a transport depending on their respective roles. Allowing for some degree of simplification, these roles on the customer's side are as follows:

- the sales manager—he or she needs to know which goods are available where and when, in order to be able to make the relevant commitments toward his or her customers;
- the service manager—he or she needs to know which goods are available where and when, in order to be able to schedule service teams accordingly;
- the warehouse manager—he or she needs to provide the information on which goods are on stock and ready for shipping (outbound side) or he or she needs the information on which goods are expected when in order to control the relevant stock levels (inbound side);
- the production manager—he or she needs to provide the information on which goods are ready for shipping (outbound side—make to order goods) and he or she needs the information on which raw materials or components are expected when in order to control the relevant production processes (inbound side—especially for JIS and JIT goods).

On the transport logistics operator's side, three main roles can be defined (allowing for some simplification):

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- the driver—he or she executes the transport and, therefore, needs to know what has to be picked up/delivered where and when during his or her route;
- the dispatcher—he or she plans the transport and schedules the own vehicles and drivers or the subcontractors accordingly;
- the fleet manager—he or she is responsible for fleet maintenance, for compliance with legal regulations (especially maximum driving times and breaks of drivers) and for driver training.

The relevant tasks and obligations related to these roles are usually supported by different IT systems, which can roughly be categorized as

- "OnBoard IT" (IT installed within the vehicle either on a built-in on-board computer or on a mobile device, for example, PDA or smart phone), which is operated by the driver and often automatically fed with vehicle sensor data;
- "Fleet IT" (this includes *dispatching systems* for transport planning, *fleet monitoring systems* for monitoring of and communication with the respective vehicles and *fleet management systems* for documentation of fleet status, maintenance management, and KPI generation), which is linked to the OnBoard IT via appropriate means of data communication and operated by the staff of the transport logistics operators; and
- "Customer's IT" (especially CRM, SCM, and ERP systems), which is operated by the staff of the customers, and usually exchanges data with the Fleet IT via interfaces.

2.2 Process units

LFM telematics solutions interconnect these different IT systems by means of telecommunication in order to support the relevant processes. From a transport logistics focused point of view, these processes roughly can be categorized as *preceding processes* (any processes on customers side related to sales, service, warehousing, or production, which initiate transports), *transport planning* (route generation and vehicle and driver scheduling), *transport execution* (physical execution and monitoring of transports plus exception handling), *transport post-processing* (internal evaluation of conducted transports by means of report and KPI generation and external invoicing of transports either based on actually conducted routes or in accordance with tariffs), and *succeeding processes* (any processes on customers side related to sales, service, warehousing, production, or financial settlement, which require data about transports).

2.3 Process flow

Processes in LFM telematics are not widely standardized—they may differ significantly by vertical industry and even by company.

Still, the process flow shown in Figure 3 may be considered as an approximate general process model for telematics in logistics, which also covers the relevant aspects of fleet management telematics.

Transport planning is based on the *transport order data* and the data on available driver and vehicle resources, which stem from previously executed routes (*route execution data*). Transport planning generates the relevant *route plans*, which are needed for transport monitoring. Per route, the individual *route details* are forwarded to the relevant vehicle for *transport execution*, that is, they are transferred to the relevant OnBoard-IT system.[3] As transport execution proceeds, the relevant *status* messages (loading confirmed; route departure, break, etc.), *events* (delay, breakdown, etc.), and vehicle *positions* (GPS coordinates) are send back to *transport monitoring and exception handling*. In case of minor disturbances, the differences between plan and execution are only registered as deviations. In case of major disturbances (long delay, breakdown, etc.), the original plans need to be modified as needed (*revised route plans*).[4] In this case, the *revised route details* have to be transmitted to the effected vehicles.

While transports are executed, *transport monitoring and exception handling* needs to determine which information is relevant for the subsequent customer processes, and thus either sends *confirmations* (e.g., ETA confirmations or delivery notices) or *alerts* (e.g., excess of refrigeration temperature on a food delivery vehicle, extraordinary shock detections on a vehicle, which has loaded sensitive electronic equipment, and significant ETA delays) to the relevant customer applications as needed.

Once transport execution is completed, the *route execution data* are processed further

- by *transport KPI generation* in order to create input for controlling purposes of the transport logistics operator, for example, average waiting time at unloading location *x* or average fuel consumption of driver *y*,
- by *transport invoicing* in order to generate the data basis (e.g., mileage, loading, and unloading times) for commercial remuneration of the provided transport services with the customers,[5] and
- by *maintenance management* in order to register the vehicle usage and schedule maintenance intervals accordingly.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

As possible delays may have an impact on subsequent vehicle or driver scheduling, these data also need to be send back to *transport planning* (①).

2.4 A three-tier concept of LFM telematics

In Section 2.1, three different groups of users (Figure 2) as well as three interrelated IT systems (Figure 3) for LFM telematics systems have been described. Consequently, LFM telematics applications may be grouped into three tiers, which build on each other:

- Tier 1: ITC for vehicle operations with focus on vehicle and driver (OnBoard IT),
- Tier 2: ITC for transport logistics operators (Central IT) and
- Tier 3: ITC for staff of customer's operations (customer's IT).

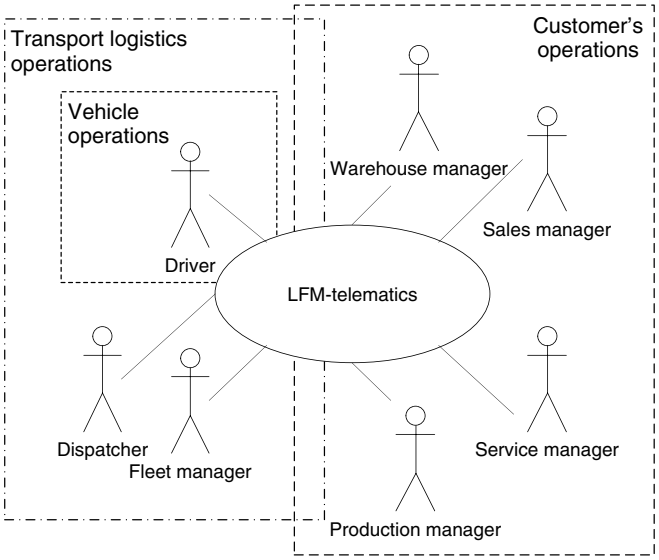


Figure 2. Users of LFM telematics.

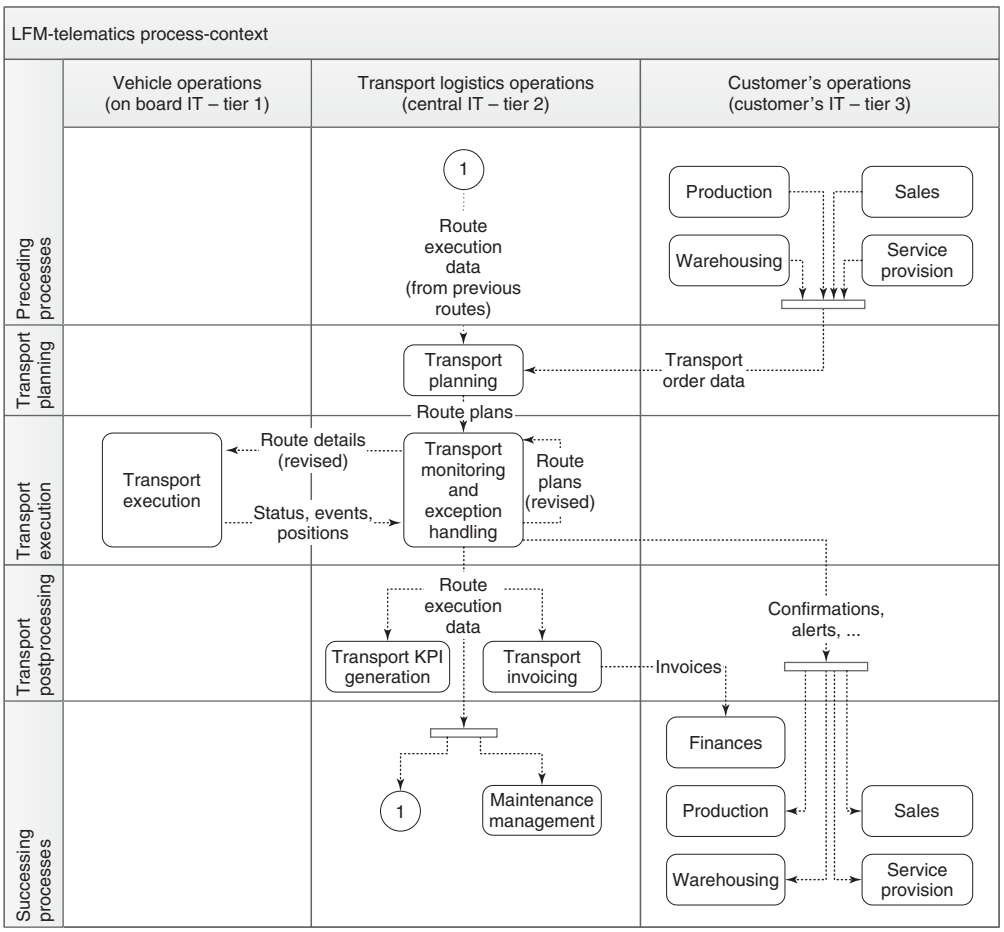


Figure 3. Process context of LFM telematics.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Tiers 1 and 2 are obvious elements of LFM telematics,[6] but tier 3 creates significant added value for customers operations, and thus requires to be considered, if one wants to assess LFM telematics solutions. Section 3 discusses the different applications by tier.

3 TIER 1—FOCUS: VEHICLE AND DRIVER

Telematics functions that support "driving as such", as for instance, "driver information systems", "driver assistance systems", or "autonomous vehicles" are discussed in the "Intelligent Vehicles" chapters of this encyclopedia (see Applications—Intelligent Vehicles: Driver Information, Driver Assistance, and Applications—Intelligent Vehicles: Autonomous Vehicles).

In addition to these intelligent vehicle functions, tier 1 of LFM telematics provides means to

- process data that stem from tiers 2 and 3 in order to support the driver (e.g., upload itineraries for navigation and receive additional pick up orders),
- capture data that are relevant for tiers 1, 2, and 3 by means of interfaces, sensors, or by means of a man-to-machine interfaces, for example, GPS positions, temperature in vehicle cargo bay, current fuel consumption rate, waiting times at customer locations or driver breaks,
- bidirectionally exchange these data with tier 2 (either online or via offline batch data exchange).

3.1 Vehicle and driver data capture

For many LFM functions, the capture of vehicle-related data is mandatory.

3.1.1 Sensor data

If no other options are available, the vehicle needs to be specially equipped with the relevant sensors (e.g., for temperature and moisture surveillance of the cargo hold, shock detection, and state of electronic seals), which require a significant amount of physical wiring and IT plugging. Therefore, one may state a *subsidiarity principle* for vehicle data capture: "if sufficient data are available via interfaces, use these instead of rigging additional sensors." Thus, standardized access to the vehicle data bus and the digital tachograph play an important role in this context.

3.1.2 Vehicle data bus

The vehicle data bus is used to exchange data between different vehicle components (a commonly used type of a vehicle data bus for trucks is the CAN-bus—for technical details, see ISO 11898, 2007; ISO 11992, 2005a). A standardized interface (see Logistics and Fleet Management, Section 7) to the vehicle data bus enables tier 1 LFM telematics applications to receive data from different vehicle components for further processing. (For further details on the vehicle bus itself, see In-Vehicle Network).

3.1.3 Digital tachograph (EU)/electronic on-board recorder (US and CAN)

The digital tachograph is a recording device for trucks, which is required by European law. Its general characteristics and functions are defined as follows.

"The purpose of the recording equipment is to record, store, display, print, and output data related to driver activities. The recording equipment includes cables, a motion sensor, and a vehicle unit. The interface between motion sensors and vehicle units shall be compliant with ISO 16844 (2005b). The vehicle unit includes a processing unit, a data memory, a real time clock, two smart card interface devices (driver and codriver), a printer, a display, a visual warning, a calibration/downloading connector, and facilities for entry of user's inputs…" (European Community, 2011, S. ANNEX I B, II. 1).

"The recording equipment shall ensure the following functions:

- speed and distance measurement, time measurement,
- monitoring driver activities, monitoring driving status,
- drivers manual entries,
- entry of places where daily work periods begin and/or end,
- manual entry of driver activities,
- entry of specific conditions,
- detection of events and/or faults,
- data downloading to external media, output data to additional external devices" (European Community, 2011, S. ANNEX I B, II. 2).

A device similar to the digital tachograph, the electronic on-board recorder for over-the-road motor carriers (EOBR), originally was introduced in the United States by legal regulations in 2010 (FMCSA, 2010). Even though this regulation was overturned by the US Court of Appeals for the 7th Circuit in August 2011 (Schulz and Berman, 2011). FMSCA indicated that they will go through a new rule-making process, and thus regulations in the United

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

States (similar to those in the European Union) are to be expected.

The Canadian Council of Motor Transport Administrators (CCMTA) indicated that they also see a "...need to consider a national EOBR standard ..." (CCMTA, 2009).

Tier 1 applications should usually have an interface to the digital tachograph (or to equivalent units) in order to avoid redundant data capture on driver activities (such as driving, work, availability, or break/rest—for details, see (European Community, 2011, S. ANNEX I B, III. 4.).

3.2 Transport and route data transfer

Transport data (transport orders) and route data (itinerary, loading list, etc.) uploaded form tier 2 to tier 1 help the driver to perform his or her tasks (e.g., checking consignments at loading, picking the right items for unloading, or navigating from stop to stop).

The download (tier 1 to tier 2) of the relevant execution status and event data per transport order (begin/end of loading, waiting times caused by customers at loading or unloading, delivery rejection by customer, etc.) or per route (traffic jam, driver break, overnight stay, etc.) is the basis for tier 2&3 applications. While some of the status and event data may be generated automatically (e.g., the arrival at a predefined GPS coordinate), usually the driver is required to key in some data manually.

3.3 Exception handling

On top of sheer transport and route data transfer, advanced LFM telematics solutions support exception handling procedures. These include timely exceptions (e.g., continuous ETA calculation and generation of proactive delay alerts), schedule exceptions (e.g., inclusion of previously unscheduled additional pickups or deliveries and additional legal driver breaks for significantly delayed routes), cargo integrity exceptions (e.g., temperature warnings for food transports; generation of additional watering stops for animal transports in case of delays), and security exceptions (e.g., automated accident alerts and panic button).

Standardizing exception handling procedures, on the one hand (tier 1), helps to reduce the communication effort for the driver (i.e., a call via mobile phone to the dispatcher in order to discuss the specific exception and the required actions). On the other hand (tiers 2 and 3), standardized exception handling rules increase the robustness and transparency within the transport process.

The relevant tier 1 in-vehicle applications, therefore, usually provide exception handling process templates by means of predefined workflows within the GUI of the application (e.g., guidance through a check-list/to-do-list).

3.4 Navigation

The list of uploaded route stop locations can be used as input for point to point navigation. This is beneficiary for the driver, as he or she can avoid searching for unknown stop locations, can circumnavigate traffic jams (increase of productivity), and, with modern navigation systems, he or she may get warnings on speed limits (reduced risk of speeding tickets).

Often specific *truck navigation* applications are required, as standard car navigation systems may not consider specifics, for example, maximum vehicle height (bridges), maximum vehicle length (hairpin bends), roads closed for dangerous goods transports (water protection areas), temporary ban of truck transports (Sunday driving bans), and others.

3.5 Geo-fencing

A virtual geo-fence is a digitally defined geographic delimiter (usually a circle or a rectangle) around the coordinates of an object of interest. The virtual geo-fence is used to detect when a monitored vehicle (e.g., by means of GPS tracking) has entered or left the vicinity of an object of interest.

Geo-fences may be used to just monitor vehicles in tier 2 applications. However, some tier 1 applications make use of this concept, too. For theft protection, for instance, the opening of a seal or valve (sensor data) outside a predefined geo-fence may either trigger an alert or be disabled altogether (Figure 4).

4 TIER 2—FOCUS: FORWARDING AGENT AND CARRIER

Tier 2 applications focus on the logistics operator as a transport service provider. Therefore, staff (excluding drivers) of external transport service providers (forwarding agents and carriers) and internal transport departments may all be seen as users of tier 2 applications. The different applications in tier 2 are described in accordance with the structure laid out in Figure 3.

4.1 Transport planning

Transport planning is one of the key elements for LFM telematics solutions, as many data needed elsewhere are a result of transport planning.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

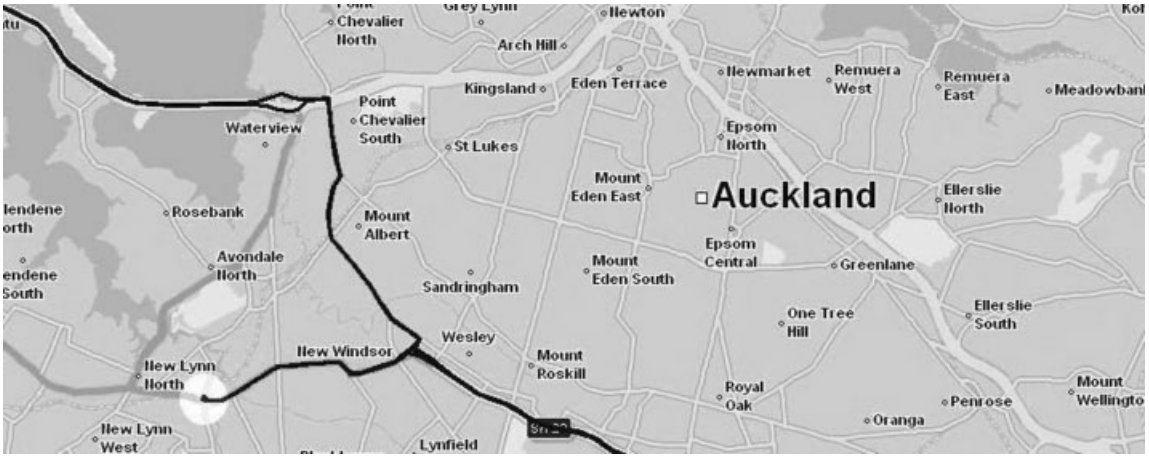


Figure 4. Vehicle track with circular geo-fence. (Reproduced with permission from Lomosoft. © Lomosoft GmbH, 2012.)

Input data for transport planning are transport order data (which stem from the customers) and the current availability of vehicles and drivers (which stem from previously generated schedules). Especially, the up-to-date information on the availability of vehicle and driver is a value adding aspect of LFM telematics, as plans based on these up-to-date availabilities can be considered more robust, and tend to require less changes than plans that have to rely on availability estimates, only.

On the basis of these input data, the dispatcher conducts the transport planning. Depending on the type of transports, different *planning algorithms* are helpful to support the dispatcher in this task. While groupage usually is planned by algorithms designed to solve vehicle routing problems (VRPs), full truck load (FTL), and less than full truck load (LTL), transports are mainly planned by algorithms designed to solve matching problems (MPs). Still, even if suitable algorithms are available for planning, some *manual dispatching*, usually, is required to compensate for gaps in data content or in data structure (i.e., data-wise nonmodeled aspects of planning).

Output data from transport planning are routes and schedules, which are then transferred to tier 1 applications (either online or offline, see Figure 7).

4.2 Transport execution

As discussed, the transport execution data are collected by tier 1 applications and sent to tier 2 applications for further processing.

Transport monitoring uses the results of transport planning to compare the current status of the transports with the plans. Most prominent data in this context are:

- Current latitude–longitude positions (unprocessed, these data usually do not have a big value, but as together with navigation systems, estimated times of arrival ETA may be computed for all subsequent stops of a route, this is a valuable input for proactive detection of delays),
- Status messages—these encompass order status (e.g., loaded, en route, delivered, and rejected), route status (e.g., loading, driving, break, waiting, and unloading), vehicle status (engine off, engine on and idle, driving), and driver status (loading, driving, unloading, waiting, and idle), and
- Event messages (e.g., break of cool chain, shock detection, crossing of virtual geo-fences, and accident).

As long as no changes to the original plans are made, one may talk about sheer transport monitoring.

As soon as significant exceptions (especially longer delays) occur, it may be necessary to revise the original plans (see "revised route plans" in Figure 3) and communicate these revised plans to the tier 1 (and tier 3) applications.

4.3 Dynamic transport planning

Dynamic planning (or continuously revised planning) could be a part of "transport monitoring and exception handling"—but dynamic planning requires all planning functionalities of "transport planning." Therefore, "transport planning" and "transport monitoring and exception handling" should be performed within one integrated system, if possible.[4]

However, dynamic planning is more than just reaction to plan deviations. Three examples illustrate this point.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Firstly, if homogeneous goods (e.g., an assortment of flowers) are transported, routes may start before precise transport orders are known. This could be the case, for instance, with flower transports, where the trucks may be heading to their destination areas without knowing how many roses and tulips to deliver where, yet. While the trucks head toward their destination areas, sales could fix the orders with the customers, and then transfer the transport order data to the tier 1 applications only after the routes have begun.

Secondly, if again homogeneous and perishable goods (e.g., a frozen food assortment) are loaded for multi-stop home delivery transports, and some regularly visited customers are not at home when the truck arrives, two things have to be cared about: (i) the customers who have not been met need to be revisited later the same day or on the following day and (ii) the subsequent customers will be visited earlier (if a previous customer is not met) or later than planned (if an additional stop to revisit a customer is inserted into the route earlier on). Here, dynamic planning is required to organize the driver's workload as well as to update ETAs [for internal tier 2 use or tier 3 use (e.g., ETA announcement via web or SMS)].

Thirdly, consider truck-meets-truck transports in FTL or LTL operations, where two trucks from different transport operators and different geographical areas need to make a delivery in the respective area. Instead of heading to their remote destinations and ending up there without a back load, they could meet half way, exchange their goods and carry out the delivery of the partner's goods in their home area. If any delays occur in such transports, this may affect the plans of two different transport operators, and thus the respective data need to be exchanged between different LFM telematics systems. In this case, process and interface standards to support this dynamic planning and execution are needed (for further details, see Kunze *et al*., 2011; Baumgärtel *et al*., 2011).

4.4 Transport postprocessing

Transport postprocessing encompasses all activities that are executed after the relevant transport has been carried out.

4.4.1 Invoicing

Invoicing usually occurs in two constellations:

- between *transport operator and shipper* (object of invoice is the performed transport service) and
- between *shipper and consignee* (object of invoice is the transported goods).

For both constellations, LFM telematics provides a significant benefit.

Invoicing of transport services is based on contractual calculation schemes. Some of these calculation schemes are based on the mileage as executed, some on the mileage as planned, and some on other schemes (e.g., zone tariffs). In case "mileage as executed" is used, the benefit of LFM telematics is obvious, as the relevant data can be extracted from data captured in tier 1.

However, also for the other calculation schemes, LFM telematics provide a twofold benefit:

- invoice initiation without delay (as soon as the transport order status "delivered" is set, the relevant invoice can be triggered) helps to improve the monetary liquidity of transport operators and
- invoice matching with transport-related costs (as transport-related calculation costs become transparent within tier-2-applications, they can be matched to the relevant invoice volumes), increases transparency on the cost-revenue-ratio for the transport operator.

The invoice initiation without delay plays a role in the *shipper–consignee relation*, too, as a "quasi-online" information to the shipper on the fact that its goods have been duly delivered is beneficiary to the invoice processes of the shipper.

4.4.2 Transport KPI generation

KPIs (key performance indicators) are needed from a transport operator's point of view as well as from a shipper's point of view.

From a transport operator's point of view, there are three main issues that require adequate KPIs on tier 2: economics, compliance with legal regulations (especially driving times), and service quality.

From a shipper's point of view, KPIs on service quality are of special interest. These KPIs can either be generated by the shipper itself based on the data provided by tier 3 applications, or those KPIs generated by the transport operator in tier 2 may be made transparent to the shipper.

4.4.2.1 Economic KPIs. As briefly discussed in Section 4.4.1, matching revenue data with related cost KPIs is important to make a profitable business. It is no trivial task to assign costs to individual transport orders (e.g., how shall the costs of a groupage route with 10 customers be distributed if each stop has a different distance to the depot, each stop receives different transport volumes, and due to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

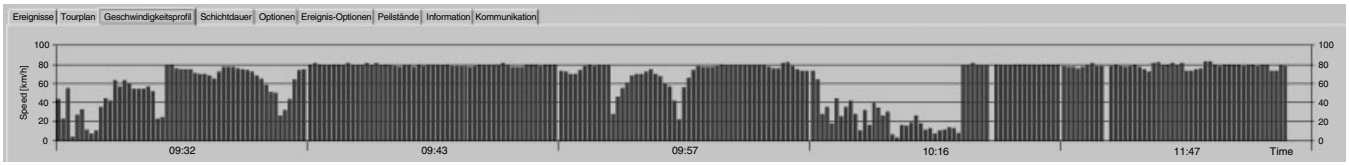


Figure 5. Speed profile. (Reproduced with permission from Lomosoft. © Lomosoft GmbH, 2012.)

different conditions at the stop premises, each stop requires a different unloading time?).

Even though these calculations tend to be difficult or imprecise, LFM telematics at least help to provide a most detailed database for the generation of these KPIs in retrospect.

However, LFM telematics also help to positively influence some of these KPIs. Two examples shall illustrate this point.

Example 1 relates to waiting times at customer locations before loading or unloading. In some vertical industries, the transport operators are kept waiting for several hours, before they can load the cargo (because of process insufficiencies of the shipper) or unload the cargo (because of process insufficiencies of the consignee). If these unproductive waiting times are monitored by loading or unloading location, they can be the basis for contractual incentive or penalty systems for shippers and/or consignees.

Example 2 relates to imprecise transport order volume data. Envision a route that is economically planned based on volume data provided by the shipper. At arrival for pickup, the real volumes differ from the ordered volumes. Then either the volumes are too small, and the route becomes noneconomic because of the unused loading space, or even worse, the volumes are too large. This may result in the need to generate additional transports on short notice, as not all originally planned transports can be handled as planned, or, if the transport operator refuses to load additional volumes, its ability to perform flexible services may be questioned. Again, transparency on this issue by means of LFM telematics combined with a contractual incentive or penalty system may help to overcome this problem.

4.4.2.2 Driver monitoring KPIs. Driver monitoring by means of LFM telematics provides a twofold benefit to the transport operator.

On the one hand, it helps (and forces) the transport operator to register the work and driving times in accordance with the legal regulations, and take adequate actions in case of violations. The relevant data are acquired by tier 1 applications (for example by the digital tachograph) and evaluated by tier 2 applications. A quasi-online connectivity is not required for this use case.

On the other hand, the driving style of a driver has an immediate impact on the fuel costs and CO_2 emissions of the transport operator. If the driver accelerates and breaks more intensely than required, the result is an unnecessary consumption of fuel. Eco[7]-drive trainings (with focus on preactive driving) in combination with tracking of individual speed profiles (Figure 5) and continuous calculation of relevant KPIs (Figure 6) may have a significant impact on the overall fleet consumption (without continuous monitoring of the drive patterns and without a related incentive or penalty system for the drivers, these positive impacts of eco-drive training tend to evaporate quickly).

In addition to these aspects, vehicle tracking and comparison of planned versus executed routes may reduce unnecessary detours. Owing to the multitude of routes, significant deviations in this respect may easily be detected by a deviation-KPI per route, which states the relevant difference. Only those routes that exceed a certain KPI limit will then need to be checked in detail.

4.4.2.3 Service quality KPIs. KPIs measuring service quality are relevant for transport operators as well as for their customers. Such KPIs are average time to delivery, average delay of delivery, average match of agreed delivery deadlines [%], and others. Basis for these KPIs are the delivery data captured by tier 1 applications. Depending on the business, these KPIs could be generated on a high level (e.g., company or region wide) or on a detail level (e.g., per consignee).

4.5 Maintenance management

Maintenance management for vehicle fleets requires scheduling maintenance works for the individual vehicles at appropriate times. This scheduling needs to consider

- the technical state of the relevant vehicle (or the state of the different vehicle components),
- the need of the vehicle as a physical resource in transport logistics operations, and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Zurücksetzen | Anzeigen

Period	Drive_style	Difficulty	Distance	Av.weight	Av.speed	Av.consumption (drive and stop)	Av.consumption (drive only)	Drivers
Zeitraum	Fahrweise [Note]	Einsatz-schwere [Note]	Fahr-strecke [km]	∅-Gew. [t]	∅-Geschw. [km/h]	∅-Ges. Verbr. [l/100 km]	∅-Fahr. Verbr. [l/100 km]	Anzahl Fahrer
12 / 2011	9.4	3.1	335	17	74	26.5	26.3	1
11 / 2011	8.8	4.3	3.549	28	67	30.0	29.7	1
10 / 2011	8.8	4.7	10.055	33	67	32.1	31.8	1
09 / 2011	8.6	4.9	4.359	38	65	33.3	32.9	1
08 / 2011	8.7	4.7	3.502	32	67	30.9	30.5	1
07 / 2011	8.6	5.0	8.835	35	65	33.8	33.4	1
06 / 2011	8.8	4.0	8.311	27	69	32.9	32.4	1
05 / 2011	8.8	4.9	11.599	34	69	34.5	34.1	1
04 / 2011	8.9	5.0	10.438	36	71	33.8	33.6	1
03 / 2011	9.1	5.0	2.464	39	77	30.8	30.6	1
02 / 2011	8.6	4.6	1.303	30	64	36.5	36.0	1
01 / 2011	8.7	4.6	6.089	31	65	35.6	35.2	1

1-50

Durchschnitts- und Summenwerte (bezogen auf alle dem Filter entsprechenden Datensätze)

∅	8.8	4.7	5.903	33	68	33.3	32.9	
Summe			70.839					

Figure 6. Eco-drive training impact on ø fuel consumption. (Reproduced with permission from Seifert Logistics. © Seifert Logistics GmbH, 2012.)

- the time slot availabilities of the different garages that are qualified to perform the required maintenance works.

Considering these three aspects, the fleet manager decides on *when which* vehicle needs to be taken *where* out of operational service, so that maintenance works can be performed as required. LFM telematics can support this task in the following way.

4.5.1 Capture of technical state

Capturing the technical state of the different vehicle components is key to determine when the next maintenance works are needed. Some of these data may be captured by tier 1 applications via the in-vehicle bus and then be transferred online or offline to the relevant tier 2 applications. Data on minor damages (i.e. repair can wait until the next scheduled maintenance break) can be captured by the driver in tier 1 man-to-machine dialogs. Other data (e.g., the remaining time until replacement of tires) can be estimated based on operational data stored in tier 2 applications.

Additional data (e.g., regular and usage-independent technical check intervals imposed by authorities) can be obtained without LFM telematics from vehicle master data records.

4.5.2 Maintenance scheduling

Scheduling maintenance breaks requires the alignment of technical maintenance needs for the different vehicle components on the one hand and schedules (vehicle schedules and garage schedules) on the other hand. Provided that all relevant data have been captured, this alignment itself might be carried out without any further use of LFM telematics. However, the results of maintenance scheduling can be transferred to the relevant drivers by means of "go to garage"-transport-orders via the "tier 2 to tier 1"-interface.

5 TIER 3—FOCUS: SHIPPER AND CONSIGNEE

One key benefit of LFM telematics is its ability to extend the reach of customer's processes (i.e., processes of the shipper or the consignee) in such a way that transports are an integral part of these processes (and are not intransparent black box activities). As soon as customer's processes are linked to LFM telematics, this establishes tier 3 applications.

As the number of potential tier 3 applications is unlimited, only a few selected examples shall be discussed in this chapter.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5.1 Automotive

Transports for the automotive industry are often JIT (just in time) or even JIS (just in sequence) transports. As safety stock levels have often been reduced to a minimum or even to zero, delays in transports of automotive parts or components may have a significant negative effect on the assembly. Therefore, tier 3 LFM telematics solutions are required which can generate precise ETAs for the consignee.

In addition to that, dock management is an issue in the automotive industry. Owing to the multitude of parts and components, and the high amount of outsourced production processes, the sheer number of transports per day poses a problem for the consignees' goods reception. To avoid queues at the unloading docks, some consignees create preassigned time windows for deliveries. These need to be considered by the logistics operator during *transport planning*, and thus this issue requires tier 3 integration toward the dock-management system of the consignee.[8]

5.2 Construction

In 2006, Günthner, Kessler, and Sanladerer designed an LFM telematics system for the construction industry (Günthner, Kessler, and Sanladerer, 2006). Since then, the application of LFM telematics in the construction industry has become more and more popular, as deliveries to construction sites often show three special characteristics: (i) certain deliveries are extremely time critical (e.g., delivery of premixed liquid concrete, which solidifies after a certain time span, or delivery of heavy goods, which require a crane for unloading), (ii) owing to possible delays on site, the delivery service may require a high level of flexibility,[9] and (iii) the goods recipient does not necessarily have permanent staff on site (staff of a contractor may only be present while certain works need to be performed, and the workers may not have a permanent office on site). Therefore, in many cases,

- a highly *precise* scheduling of delivery times,
- a *flexible* readjustment of deliveries (dynamic planning), and
- details on the *exact location* of the point of delivery on site

are required.

Precision already can be achieved by a classical tier 1 and 2 LFM telematics solution, where deliveries are preplanned and executed in accordance with the preplanned schedule, and ETA times are continuously updated and transferred to the customer's staff on site. In case of delays on the transport side, new ETA times will allow for adequate adjustments on site (e.g., if a crane was originally scheduled for unloading the now-delayed truck, it may be used for other craning activities in the meantime).

As soon as *flexibility* is required, too, an LFM telematics solution is required that allows dynamic planning (tier 2) based on input from the customer (tier 3 interfacing) plus readjustment of transports in execution (tiers 1 and 2). If, for instance, several hundred cubic meters of concrete are needed, and the concrete casting process is somehow delayed, the loaded concrete trucks cannot wait without limit. Therefore, such a delay will also affect the loading at the concrete mixing plant (and even the operative production planning there). Thus a two-face tier 3 process integration is required (dynamic planning with process integration toward construction site and concrete mixing plant).

The *exact location* of the point of delivery on site may be determined by different approaches, for example, by free text messages (e.g., "gray bureau container 50 m northwest of crane"; tier 2 to 1 communication requires exchange of free text messages), by scans of construction site maps where relevant location is indicated (tier 2 to 1 communication requires exchange of data, which can graphically be displayed by the tier 1 application), or by latitude–longitude coordinates for a navigation system (the first driver who has successfully searched for the precise delivery location sends back his or her coordinates to tier 2, and subsequently arriving drivers get these coordinates sent to their tier 1 navigation system).

5.3 Courier, express, and parcel

Courier, express, and parcel logistics (CEP logistics) sometimes is perceived as sheer physical transport services. However, information on when the relevant items were shipped and when they were delivered are an integral part of the CEP services. Nowadays, customers of CEP services want to have transparency on where the goods are even while the physical transport takes place. The related service is called *track and trace* (*T&T*), and, depending on the designed T&T granularity and the related LFM telematics system design, one can achieve gateway-based T&T (i.e., identify an item when it passes a gateway which captures its ID) or continuous T&T (i.e., identify the ships, planes, trains, or trucks the items are loaded on, track these carriers by means of locating services,[10] and thus indirectly track the loaded objects).

In order to be able to identify objects for T&T, the relevant objects need an ID, and this ID needs to be related to the relevant sender and recipient data (minimal data are names and addresses). Until recently, this identification was

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

widely done by means of a barcode sticker, which carried the ID for scanning. The related sender and recipient data were somehow entered into the IT system of the transport operator (either by data import via interfaces or via internet access frontends or by manual data capture on a scanner handheld), and then linked to the relevant ID. Nowadays, RFIDs start to be used instead of/in addition to the barcodes as identifiers (further details, see DHL, 2011).

On a regional level,[11] CEP *transport execution* performs two overlaying operations: collection of (*a priori* not necessarily identified) items from customers to consolidation centers and delivery of (previously identified) items from consolidation centers to customers. Both operations are supported by LFM telematics:

5.3.1 Pickup services

In case the items have not been identified beforehand, the driver who runs the pickup routes needs to issue an ID and capture the relevant data (this is usually done by a tier 1 handheld device). Else, he or she scans the relevant IDs (reads the relevant RFIDs) while he or she collects the items.

Regular CEP customers (customers who ship parcels every day) are visited by daily master pickup routes. As irregular customers are only served if they require a pickup, a need for LFM telematics that enables *dynamic planning* arises (see Logistics and Fleet Management, Section 4.3).

5.3.2 Delivery services

For delivery services, the routes for the relevant items need to be planned (tier 2) and then the trucks need to be loaded accordingly. To avoid wrongly loaded items or missing items on a truck, the items may be scanned at loading. A tier 1 application may then generate alerts (for missing or wrongly loaded items) or confirm the completeness of all items to be delivered on the relevant route.

To achieve continuous T&T for the loaded items, the relevant handling processes (loading, delivery,[12] intermediate storage, and return to the sender) and the relevant vehicle positions have to be traced (tiers 1 and 2), and thus a high transparence on the transport process can be achieved for (i) the logistics operator (tier 2) and (ii) the relevant shippers and consignees if relevant data are provided to them via interfaces or via T&T web frontends (tier 3). In other words, if, for instance, a web shop, a furniture store, or a reseller of white goods wants to keep its customers up to date on when to expect the delivery, they require a logistics operator who is equipped with an adequate three-tier LFM telematics system.

5.4 Mineral oil

Product availability and safety are the two main objectives in transport of mineral oil. For replenishment of fuel stations, the limited tank storage space at the fuel stations combined with fluctuating demand requires an efficient replenishment process to guarantee the product availability close to 100%. Safety during loading, transport, and unloading requires technical means as well as process means. Both aspects—product availability and safety—determine the mineral oil-specific processes, which are supported by LFM telematics. These specific processes can be matched to the generic processes shown in Figure 3 as follows:

Initially, a close link to "customer's IT" is needed to generate the relevant *transport order data* (based on current local product volumes and consumption rates per fuel station).

Transport planning needs to consider different product types (regular, premium, and gasoline) and the different compartment volumes of each truck. Transport planning also requires knowledge on which product type was loaded in which compartment of the truck during the preceding delivery run in order to determine if compartment cleansing (additional preloading time needed) is required or not. Therefore, tier 1 LFM telematics applications continuously capture loaded product volumes as well as product-compartment assignments.

During *transport execution* (loading, transport, and unloading), the workflow design of the tier 1 applications helps to increase safety by means of process standardization. Before *loading*, a check of the previously loaded product can be performed, and in case of product incompatibility, a mandatory compartment cleansing can be required by the tier 1 application. During *transport*, a specific navigation for dangerous goods transports (which avoids water reserve areas) can guide the driver, and in case of temperatures close to 0°C, a glaze alert can require the driver to reduce his or her speed.[13] Before *unloading*, a check of the current vehicle position versus a predefined virtual geo-fence can be conducted, in order to grant that no product is unloaded outside this geo-fence (theft protection). For this usage, the tier 1 application may either trigger an alert to a tier 2 application or even have access to actuators in order to block any valve operations outside the geo-fences.

Transport monitoring and exception handling are specifically required if planned delivery volumes cannot be unloaded (leftover products in truck) or if vehicle loads need to be (partially) redirected because of unexpected product shortages at a third fuel station. Such deviations from planned order volumes also need to be considered for *invoicing* between shipper and consignee, and thus

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the capture of data on real delivered volumes (via tier 1 applications connected to meters or electronic dipsticks) are vital for shippers and consignees, as well as for the transport operators (especially, if such volume deviations lead to additional mileage).

5.5 Further tier 3 applications

Tier 3 LFM telematics has a wide scope of further applications. A few more of them shall be discussed in brief.

5.5.1 Food

For food transports (fresh as well as frozen foods), the uninterrupted cold chain is of importance. Therefore, sensors in the cargo bay need to track the temperature (and sometimes the moisture too), and these tracks are recorded by the tier 1 application. In case there is a threat that the cold chain becomes broken (excess of a warning temperature level), the tier 1 application can generate a warning (see "events" in Figure 3) and the dispatcher can deal with the problem. Shipper as well as consignee can then be informed on this event via tier-2-to-3-interfaces.

5.5.2 Sensitive electronics

For transport of sensitive electronic components, a shock-free transport is of importance. Especially, if such goods are damaged by a shock somewhere between shipper and consignee, it is helpful for the transport operator that he or she can prove the shock-free transport if the shock event was caused by another party.

5.5.3 Technical services

Technical service providers (e.g., elevator repair services and kitchen installation services) can only conduct the service if the relevant items are available. Therefore, the scheduling of the items and the scheduling of the technicians often requires synchronization (tier 3 integration of item transports with technicians scheduling).

Also, the time needed for such services is difficult to assess *a priori*. If several service orders are to be carried out on the same day, and an earlier service requires more (or significantly less) time than planned, the subsequent services are affected, and, therefore, ETA information update has a high benefit for the service customers.

6 TECHNOLOGY AND DATA

The functionality of an LFM telematics system is limited by its technologic design and by the data model used.

6.1 Technical components of LFM telematics

As there is no one-size-fits-all-architecture for LFM telematics, specific LFM telematics systems show a wide range

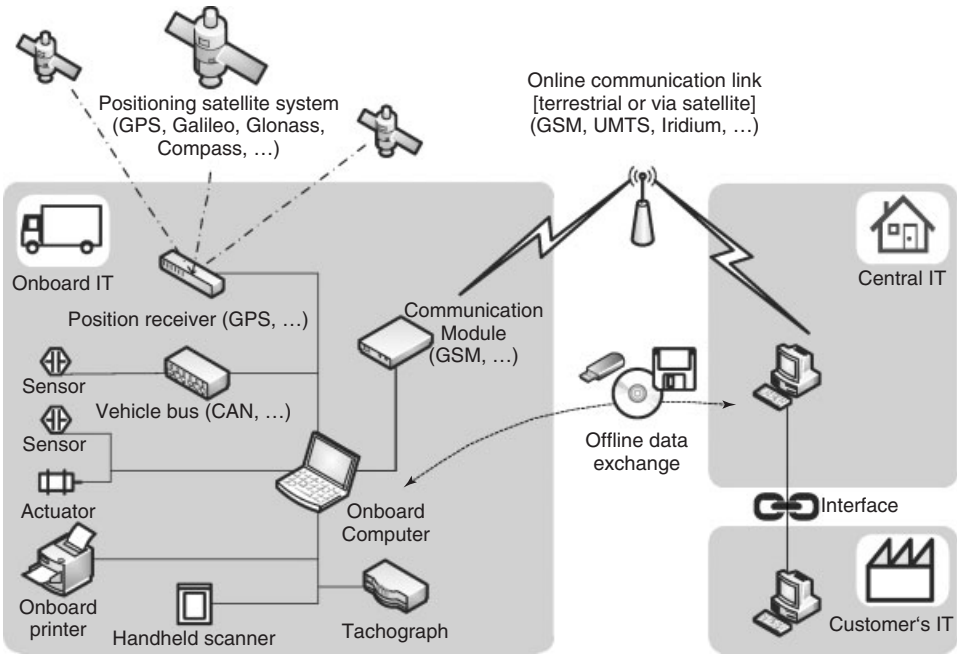


Figure 7. Generalized technical layout of an LFM-telematics system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of diversity from a physical implementation point of view. Still, Figure 7 shows a generalized layout of technical components, which establish an LFM telematics system.

The key unit of the OnBoard IT is the OnBoard Computer (OBC). It hosts the relevant LFM telematics in-vehicle tier 1 applications, provides a user interface to the driver, and ideally is linked to

- a position receiver (in order to compute the current position of the vehicle and provide navigation support to the driver)—even if today, such a receiver often is a GPS receiver, it could also be a position receiver based on Galileo (EU), Glonass (RU), or Compass (CN);
- the in-vehicle sensors directly or indirectly via the vehicle bus (e.g., to capture temperature or moisture data in the cargo hold or to log speed and fuel consumption data);
- the in-vehicle actuators (e.g., to block electronic locks of cargo bay);
- the electronic tachograph (only in case when the vehicle is equipped with an electronic tachograph in order to make use of the drive and work-time-related data, which are stored there for legal purposes anyhow);
- a handheld scanner (in order to enable the driver to scan barcode labels/read RFID tags on the cargo and manually key in data as needed outside the cab);
- an OnBoard printer (e.g., to print damage reports, proof of delivery certificates, or invoices);
- a communication module, which enables "quasi-online" data exchange with the central IT.

The term *quasi-online* is used, because a recent study (Dudek and Köppel, 2011, S. 55f) has shown that regular update cycles are as follows (Table 1).

Some mass data do not require quasi-online data transfer, and thus may be exchanged with the central IT via batch data up- and download (by means of either physical exchange of data storage media or a LAN at the fleet base).

The component layout shown in Figure 7 does not require to be implemented by means of distinct hardware units, as Figure 7 may suggest. If some features (e.g., sensor data capture, barcode scanning, or printing) are not required for a specific use case, and other features are integrated within one physical unit, the layout of technical components that establish an LFM telematics system may get as simple as shown in Figure 8.

Table 1. Update cycles.

Vehicle tracking cycle	Percentage of interviewed LFM-telematics users (%)
$\leq 1\times$/s	1
$\leq 1\times$/30 s	20
$\leq 1\times$/min	25
$>1\times$/min (ø 9 min)	54

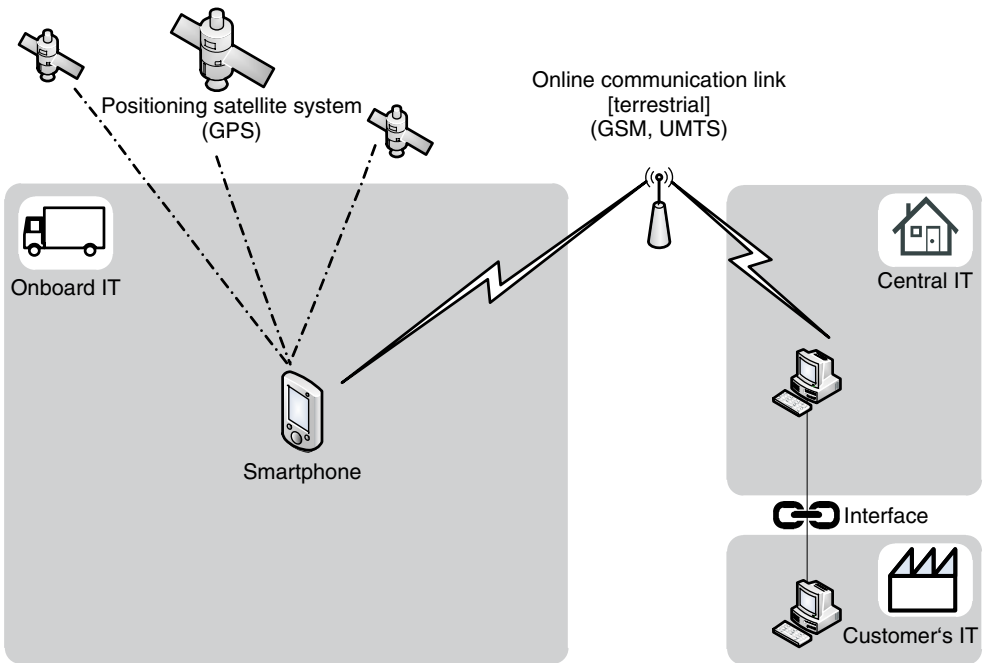


Figure 8. Simplified technical layout of an LFM-telematics system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

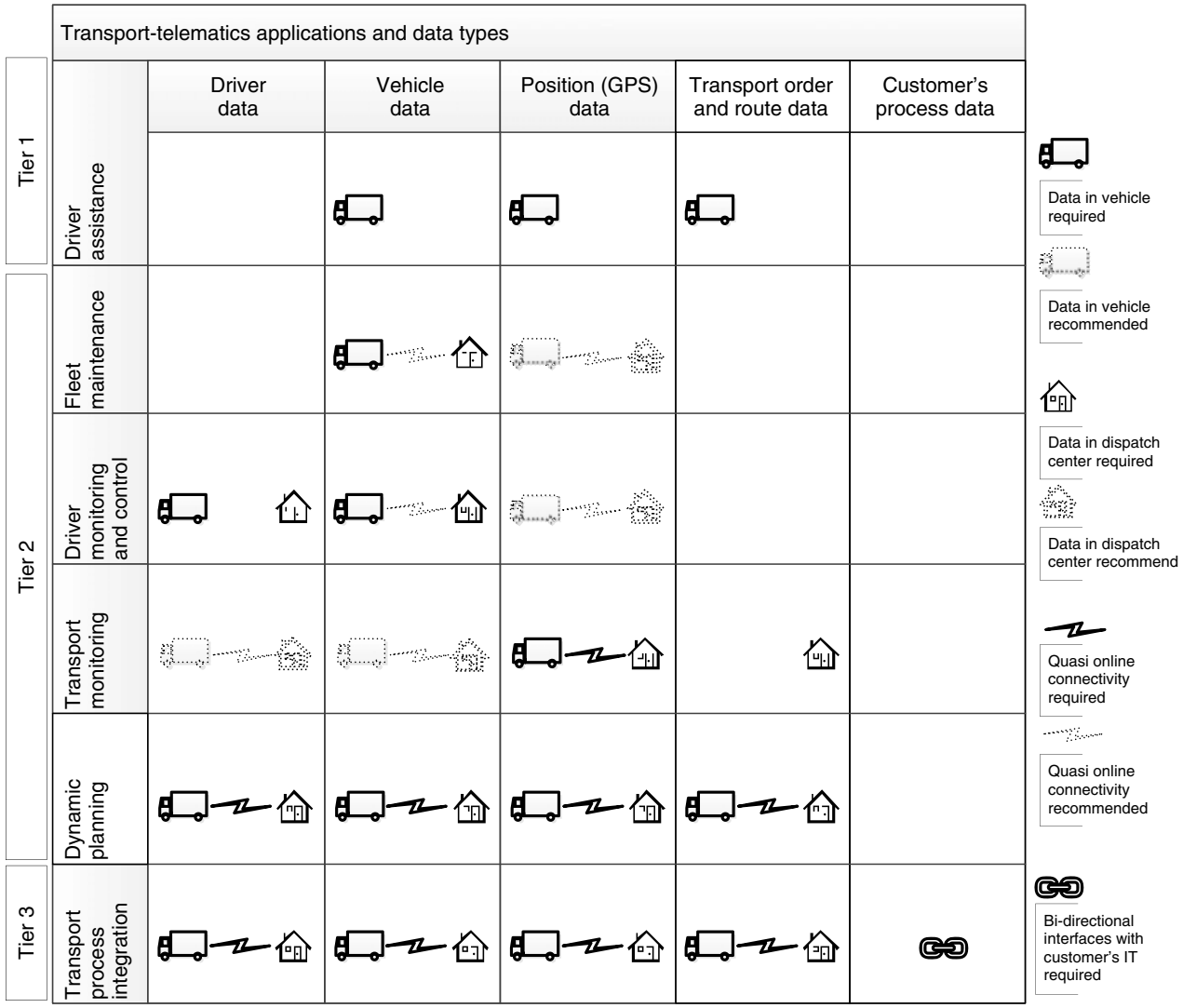


Figure 9. LFM telematics applications and data types.

6.2 Data for LFM telematics

The different applications in each tier require different data. On a high aggregation level, Figure 9 depicts which types of data are needed for which application.

Figure 9 also shows which data need to be synchronized quasi-online (i.e., in regular intervals) and which data only require offline batch data transfer (i.e., up- or download of data when the vehicle is at its home premises).

7 STANDARDIZATION

Standardization, or better the lack of sufficient standardization, today is the key hindrance factor to generate better LFM telematics systems.

If a transport operator with a heterogeneous fleet wants to implement an LFM telematics system, it is facing a number of integration challenges. These challenges concern:

- process and interface standards;
- communication standards;
- software & hardware (SW&HW) standards;
- connectivity to sensors, actuators, and/or vehicle data bus.

Again, the word "standard" is misleading, as "many standards" actually means that there is not one generally accepted standard, which is used.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

7.1 Process and interface standards

From a logistics operator's point of view, creating a three-tier LFM telematics solutions requires to integrate its own in-house processes (which are usually supported by his or her in-house software) and the processes of its customers (i.e., the shipper and/or consignee) via interfaces.

Currently, the degree of cross-company process standardization in logistics operations is not very advanced. Almost every IT system used in the industry provides own process templates, and many tailor-made adaptions of the so-called standard software can be found, as these adaptions are required to mirror specifics of the relevant transport operator. Therefore, the in-house processes of transport operators can be regarded as "individual process implementations". This issue was recently researched by Kunze and Baumgärtel, who presented a process and interface standard for cross-company dispatching of truck-meets-truck transports (Kunze *et al*., 2011).

On the customer side, the world market leading ERP systems tend to represent "quasi-industrial" process standards and interface standards.

Thus, a significant effort for process and interface integration (internal and external) is necessary to implement a three-tier LFM telematics solution.

7.2 Communication standards

The choice of communication standards (second-generation GSM, third-generation UTMS, fourth-generation LTE, satellite communication, etc.) needs to balance communication network coverage (if a vehicle is out of range of the communication net, LFM telematics is interrupted; data buffering may help to overcome small regional gaps in communication, but if coverage is too low in general, it renders LFM telematics useless), speed or data throughput (a low throughput may be compensated by applications which are designed to minimize communication data volume), and costs (flat rates have reduced the importance of this issue recently).

7.3 In-vehicle HW and SW standards

In-vehicle HW and SW for tier 1 applications are usually defined by the choice of the relevant technology provider.

7.4 Connectivity to sensors, actuators, and vehicle data bus

The issue of access to the data provided by the in-vehicle data bus has been addressed by a group of European truck manufacturers.[14] They have agreed to implement and support an open standard to give third parties access to vehicle bus data. This standard is called *FMS* (*fleet management systems interface*) and is administered by the ACEA European Automobile Manufacturers' Association (for further details, see ACEA, 2009).

In case sensors and/or actuators are needed for an LFM telematics solution, which cannot be addressed by the OBC via FMS, specific wiring and IT-integration works are required, which lead to extra costs.

8 SUMMARY AND OUTLOOK

LFM telematics solutions provide benefits not only for the drivers (driver's assistance) but also to the logistics operators (transport monitoring and dynamic planning) and to the shippers and consignees (transport process transparency). The key hindrance to generate these benefits is the low level of substantial standardization (of processes, interfaces, and technology).

Owing to outsourcing and subcontracting, today's value chains have become more and more dependent on transports. Consequently, transport logistics continues to evolve from "sheer physical box moving" to "transport, process, and information logistics". Therefore, the use of LFM telematics is likely to grow further in number of installations as well as in scope of functions.

GLOSSARY

Carrier	Company (with own vehicles) that executes transports physically.
Consignee	Recipient of goods.
Forwarding agent	Company (with or without own vehicles) that organizes transports.
Transport logistics operator	Term used to describe service providers involved in the transport of goods—this includes forwarding agents, carriers and in-company departments which plan and execute goods transports.
Shipper	Sender of goods.
Role	Term stems from process modeling, and describes a number of tasks and obligations of a person who acts within this role.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Virtual geo-fence	A digitally defined geographic delimiter (usually a circle or a rectangle) around the coordinates of an object of interest.

ABBREVIATIONS

ACEA	European Automobile Manufacturers' Association
ADAS	advanced driver assistance systems
CAN	controller area network
CEP	courier, express, and parcel
CRM	customer relationship management
DAS	driver assistance systems
DTM	dynamic truck meeting
EOBR	electronic on-board recorder
ETA	estimated time of arrival
ERP	enterprise resource planning
FMS	fleet management systems interface
FMCSA	Federal Motor Carrier Safety Administration
FTL	full truck load
GPS	global positioning system
GSM	Global System for Mobile Communications (2G Communication Standard)
HW	hardware
ID	identifier (a unique key)
IT	information technology
ITC	information and telecommunication technology
ITS	intelligent transportation systems
JIS	just in sequence
JIT	just in time
KPI	key performance indicator
LBS	location-based services
LFM	logistics and fleet management
LTE	long-term evolution (4G communication standard)
LTL	less than full truck load
MP	matching problem
OBC	OnBoard computer
PDA	personal digital assistant
RFID	radio-frequency identification
SCM	supply chain management
SMS	short message service
SW	software
T&T	track and trace
UMTS	universal mobile telecommunications system (3G communication standard)
VRP	vehicle routing problem

RELATED ARTICLES

In-Vehicle Network
Cellular Mobile Networks
Technologies—Positioning: GNSS
In-vehicle sensors
Applications—Intelligent Vehicles: Driver Information
Driver Assistance
Applications—Intelligent Vehicles: Autonomous Vehicles
Applications—Intelligent Roads and Cooperative Systems: Urban Traffic Management
Advanced Highway Management Systems
Road Traffic and Travel Information (RTTI)

ENDNOTES

1. Only 9% of the vehicles in the panel operate outside Europe, too.
2. Note that the delimitation of these terms is not coherent throughout literature. ITS (2011) for instance has defined the term *intelligent transportation systems—commercial vehicle operations* (which is similar to our definition of "LFM telematics"), but that definition includes RFID applications for cargo management systems [see (ITSA, 2011, p. 62)].
3. Without an OnBoard-IT system, the *route details* are handed over to the driver as printouts of the route itinerary and the bill of loading.
4. Note: In case "transport monitoring and exception handling" does not provide any replanning features, the relevant order-, vehicle-, and driver data have to be rerouted to "transport planning" in order to make the needed changes in the *route plans*.
5. In an increasing number of cases, "transport invoicing" is replaced by "transport crediting." This inverts the financial subprocess, but the need to match route execution data with financial transactions remains unchanged, as in the case of "transport crediting" the transport logistics operator(s) need to double check the credit notes.
6. For example, the summary of ITS applications as shown in ITSA (2011; p. 62 f) focuses on tier 2 and 1 applications, only, with very few exceptions of tier 3 applications and significant overlap of traffic-telematics applications.
7. "Eco" stands for both economic and ecological driving in this context.
8. Sometimes, the same applies to the shipper, who also wants to impose time windows for pick up.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

9. See Logistics and Fleet Management, Section 4.3 Dynamic transport planning.
10. For example, by means of GPS tracking.
11. National and international center–center transports are not considered here.
12. At delivery, the recipient confirms the reception of the items (proof of delivery) by a signature, which usually is performed on a touch pad of a tier 1 handheld.
13. And ETA calculations can be updated accordingly.
14. ACEA's Heavy Truck Electronic Interface Group: DAF trucks, Daimler, IVECO, MAN, Renault trucks, Scannia, and Volvo trucks.

REFERENCES

ACEA's Heavy Truck Electronic Interface Group (FMS Group) (2009) Information about the FMS-Standard. Hg. v. LogiCom GmbH. ACEA's Heavy Truck Electronic Interface Group (FMS Group), http://www.fms-standard.com/ (accessed 02 March 2012).

Baumgärtel, H., Kunze, O., Rosemeier, S., and Neitmann, A. (2011) Dynamic Truck Meeting (DTM) - Ein Prozess- & Schnittstellenstandard zur Realisierung von dynamischen Begegnungsverkehren mit Hilfe von Dispositions- und Telematik-Systemen. Band II—Schnittstellenstandard. KIT Scientific Report 7614, KIT Scientific Publishing, Karlsruhe.

Canadian Council of Motor Transport Administrators (2009) ELECTRONIC ON-BOARD RECORDERS PROJECT GROUP—Background and Strategic Alignment, http://www.ccmta.ca/english/committees/cra/eobrs/eobrs-tor.cfm (accessed 26 April 2012).

Deutsche Post DHL (undisclosed author) (2011) DHL SmartTruck-Dynamische Tourenplanung. Intelligent unterwegs. Deutsche Post DHL, http://www. dp-dhl.com/de/logistik_populaer/aus_den_unternehmensbereichen/dhl_smarttrucks.html (accessed 26 January 2012).

Dudek, H.-L., and Köppel, M. (2011) Telematik 2011. Ergebnisse einer Befragung von Telematiknutzern und Telematikinteressierten im Bereich Transport & Logistik, Duale Hochschule Baden-Württemberg Ravensburg. Friedrichshafen, http://www.twie.dhbw-ravensburg-studenten.de/fileadmin/downloads/wirtschaftsingenieur/PublikationTelematik2011.pdf (accessed 01 February 2012).

European Community (2011) COUNCIL REGULATION (EEC) No 3821/85 of 20 December 1985 on recording equipment in road transport, vom 11.10.2011, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1985R3821:20111001:EN:PDF (accessed 13 February 2012).

Federal Motor Carrier Safety Administration (2010) Regulation: 49 CFR Parts 350, 385, 395, and 396; Electronic On-Board Recorders for Hours-of-Service Compliance. http://www.fmcsa.dot.gov/rules-regulations/administration/rulemakings/final/On-Board-Recorders-for-HOS-Compliance.pdf (accessed 26 April 2012).

Günthner, W.A., Kessler, S., and Sanladerer, S. (2006) EDV gestützte Fahrzeugdisposition und -abrechung im Baubereich zur Optimierung der Prozesskette. In: *Logistics Journal*. Hg. v. Wissenschaftlichen Gesellschaft für Technische Logistik e.V. (WGTL), http://www. logistics-journal.de/not-reviewed/2006/10/620/sanladerer_guenthner_kessler_d.pdf (accessed 01 February 2012).

International Organization for Standardization (2005a) Standard No. 11992 Road vehicles—interchange of digital information on electrical connections between towing and towed vehicles, ISO, Geneva.

International Organization for Standardization (2005b) Standard No. 16844 Road vehicles—tachograph systems, ISO, Geneva.

International Organization for Standardization (2007) Standard No. 11898 Road vehicles—controller area network (CAN), ISO, Geneva.

Kunze, O., Baumgärtel, H., Neitmann, A., and Rosemeier, S. (2011) Dynamic Truck Meeting (DTM) - Ein Prozess- & Schnittstellenstandard zur Realisierung von dynamischen Begegnungsverkehren mit Hilfe von Dispositions- und Telematik-Systemen. Band I–Prozess-Standard. KIT Scientific Report 7613, KIT Scientific Publishing, Karlsruhe.

National Transport Commission (2011) National in-vehicle telematics strategy: The road freight sector. http://www.ntc.gov.au/filemedia/Reports/NatTelematicsStratJuly2011.pdf (accessed 26 April 2012).

Schulz, J.D., and Berman, J. (2011) Trucking news: EOBR mandate turned over by U.S. Court; Logistics Management. http://www.logisticsmgmt.com/article/trucking_news_eobr_mandate_turned_over_by_u.s._court/ (accessed 26 April 2012).

The Intelligent Transportation Society of America (2011) Sizing the U.S. and North American Intelligent Transportation Systems Market: Market Data Analysis of ITS Revenues and Employment. http://www.itsa.org/images/MDA/itsa%20mda%20report%20final.pdf (accessed 26 April 2012).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto184
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Tolling, Mobility Pricing

Martial Chevreuil
Egis, Guyancourt, France

1 Introduction 1
2 A Variety of Tolling Schemes 2
3 Electronic Toll Collection Features 6
4 Interoperability 9
5 Conclusion 11
6 Glossary 11
Endnotes 11
References 11

1 INTRODUCTION

The term *toll* is quoted in the Greek mythology where the ferryman Charon used to charge a toll to carry the souls of the dead across the River Styx separating the world of the living from the world of the dead.

Toll roads are at least 2700 years old, as tolls had to be paid by travelers using the Susa–Babylon highway under the regime of Ashurbanipal, who reigned in the seventh century BC (Gilliet, 1990).

In the ancient times, in various parts of the world, travelers were obliged to pay a toll for passing some particular points on the road (e.g., bridges and mountain passes). However, toll is different from the "octroi," which was a tax levied on some goods entering cities, invented by Romans and still existing in some countries in the world.

Tolls were in use, during fourteenth and fifteenth centuries, in the Holy Roman Empire. In fourteenth-century England, some of the most heavily used roads were repaired with the money raised from tolls.

According to FHWA (Federal Highway Administration),[1] "tolling involves the imposition of a per-use fee on motorists for a given highway facility. Historically, these fees have generally been flat tolls that may vary by number of axles and distance driven, but not by time of day.[2]" These charges were originally used for generating revenue for financing, maintaining, and operating the road, bridge, or tunnel that are used.

However, with the time, the objective of tolls has evolved, and the name itself has changed in order to better cope with the different usages of tolls. Road pricing [or road user charging (RUC)] is the general term used nowadays that encompasses the various types of tolled road infrastructures.

In another way, we can say that road pricing defines the different policies of pricing the road usage, whereas tolling is the physical way of levying the fee directly from the user. Road pricing includes the charges levied for the use of the road infrastructure; these charges may be used for repaying the infrastructure costs or compensating the social costs of road transports, that is, paying for negative externalities [noise, pollutant and GHG (greenhouse gas) emissions, visual intrusion in the landscape, and accidents]. Making the charge varying with time of day, or week, is also a way of managing the travel demand. In that case, road pricing is used as a transport demand management tool to reduce traffic intensity during peak periods and is termed *road congestion charging*. Singapore and Hong Kong have been pioneers in implementing this concept. The charges may also vary according to the tailpipe emissions of the vehicle: such measures are implemented in some cities in the world to discourage more polluting vehicles to enter the city centers.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 A VARIETY OF TOLLING SCHEMES

2.1 Toll for financing infrastructure

This is the historical reason for tolling roads: tolls are in that case a contribution that covers fully or partially the cost of road construction, maintenance, and operation. In case of partial coverage, the missing part is paid by government or local authorities' subsidies; the reason given being, for example, that the road infrastructure is also contributing to the local economy and to the society in general. The road can be publicly or privately built with a concession scheme: loans necessary to pay for the construction and maintenance of the roads are issued with the expectation that they will be paid back over time by tolls. When the loans are paid off, the road should usually revert to the authority that owns the land it was built on. In practice, a toll is often maintained in order to ensure proper maintenance and operations. In particular, during the estimated lifetime of the concession, it often appears necessary to upgrade the road because of the traffic growth.

Tolls usually vary by vehicle type, weight, height, or number of axles. We can distinguish two types of toll systems: open or closed. A closed toll system is a system where the user crosses a first tollgate when entering the road (access point to the network, where the user takes a ticket, real or virtual) and a second tollgate when exiting the network, where he or she pays after presentation of the ticket. In that case, the user pays according to the distance between the entry and the exit. Contrarily, on an open toll system, all users of the same category pay the same toll at a toll plaza, whatever their origin and destination are, and independently from the distance really traveled. Open toll systems are used for isolated stretches of road, such as tunnels and bridges, or on stretch of road networks where most of the users have similar origins and destinations of journeys, particularly in urban environment. By implementing several toll plazas on an intercity motorway, the toll paid for a long distance depends approximately of the distance traveled.

Both systems have advantages and drawbacks: the open toll system is simpler, as there is no need to calculate the toll according to the distance really traveled. Generally, the toll amounts are low. Therefore, the tollgate crossing delay is minimized, which is very important for high trafficked stretches of roads. The closed system is more appropriate for long distance road networks where the diversity of origins and destinations is large, and the toll fee may be high for certain journeys.

On a wide network, both systems can coexist. In any case, selecting the toll system type requires knowing quite precisely the travel patterns of the region, which is irrigated by the toll road network.

Until now, access to tolled roads is generally restricted by barrier arms, which open when the driver has paid or taken a ticket. Initially, tolls were collected by hand by employees at tollbooths. Owing to the fact that toll amount on open toll system is a lump sum by category, these systems have been the first to be automatized: the user deposits coins in a basket, which measures the amount and opens the barrier allowing passage if sufficient.

Nowadays, credit card terminals have replaced the coin-collecting baskets in many places, allowing at the same time automation of tolls for closed networks.

In the last decade, we have seen the rapid extension of electronic toll collection (ETC), which is defined by the use of electronic contactless communication from a transponder installed in the vehicle and the toll collection system. In case of preexisting tollgate, the communication that supports the transaction process is established at the tollgate: preexisting tollgates are generally kept for occasional users and also as a mean to control the access in order to avoid the installation of a specific enforcement system. We will see in the following sections that ETC can be implemented in different contexts where the tollgate is no more necessary. Some examples are presented on the photos hereafter (Figures 1, 2).



Figure 1. Golden Ears Bridge. (Reproduced by permission of Egis. © Egis.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 2. Free-flow tolling on A22 Portugal. (Reproduced with permission from Martial Chevreuil © Martial Chevreuil.)

2.2 Mobility pricing

Like in many domains, pricing is a tool for management. Applied to transport policy, it leads to the concept of mobility pricing. The first idea, suggested by transport economists, is to make the user (traveler or shipper for goods transport) pays the real cost of using the transport facility. However, here, we start facing a first difficulty; in theory, this real cost should integrate all aspects: amortization of the transport facility (infrastructure, rolling stock, vehicles, equipment, etc.), operation costs, and externalities (emissions, noise, etc.), some of these costs varying with the traffic load and transport occupancy! On the one hand, as some of these costs are supported by the community (public investment and externalities), they are partly paid by taxes. On the other hand, public authorities provide subsidy for public transport as an incentive measure for sustainable development. If we add to that, the different users' categories, mobility pricing is a complex policy calling for various measures and tools. Owing to the predominance of road as transport mode in developed countries, road pricing plays a key role in mobility pricing and we will elaborate on this aspect in Section 2.3.

2.3 Road pricing

Some countries have planned to implement a national road pricing scheme in order that the users pay for the "real" cost of using the infrastructure. Studies for implementing such a scheme were conducted in the Netherlands, late 1980s (Rekeningrijden project). The idea starts from the fact that increasing gas taxes is not always feasible; especially in small countries, where users can fill in their car tank in the neighboring countries where taxes are sometimes lower. Owing to political changes, the Dutch project was abandoned; however, the idea was reintroduced several times during the last 20 years, without success.

Owing to the large impact of heavy goods vehicles (HGVs) in terms of emission, energy consumption, and damage on the pavements, the idea of charging specifically the HGVs has encountered a better success, particularly in Europe, where the first step had been to introduce the "Eurovignette."[3] Different pricing schemes are existing, based on distance traveled, weight, and, more recently, on motorization (EURO classes in Europe) (Figure 3).

HGV or lorry RUC are in operation in New Zealand (called *RUC*), Switzerland [LSVA, Leistungsabhängige Schwerverkehrsabgabe (Performance-related heavy vehicle fee)], Germany (LKW-Maut, Lastkraftwagen), Austria (Go-Maut), Czech Republic, Slovakia, Poland, and four US states: Oregon, New York, Kentucky, and New Mexico.



Figure 3. Charging of heavy goods in the European Union. (Reproduced from Directive 1999/62/EC as modified by Directive 2006/38/EC and by Directive 2011/76/EU © European Commission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The enclosed map from the European Commission presents the current situation in the different European countries.

Various technologies are used for tolling. The main drivers for choosing the technology are the system implementation and operating costs, its reliability, and its ability to prevent fraud. In some cases, the costs of collecting tolls can reach 30% of the amount to be collected. Section 2.4 presents three typical examples of road pricing contexts with an introduction of the various technologies that can be used.

2.4 HGV road pricing/user charging

It is certainly New Zealand that introduced for the first time an RUC system in 1978. All the revenue from RUC goes into the National Land Transport Fund. The fund is used mainly for road construction and maintenance, along with other activities benefiting to road users. It is a distance-related charging system where rates are set according to vehicle type and weight and vary in proportion to the road usage costs. This is intended to encourage transport operators to make efficient choices when transporting freight, that is, to use vehicles that balance direct operating costs and damage to roads. The distance traveled is controlled by "hubodometers" (mechanic odometers installed on vehicles and trailers axles). In 2012,[4] some modifications have been introduced in the legislation, but the principles remain the same. The RUC scheme is applicable to vehicles over 3.5 ton manufacturer's gross laden weight and all vehicles of 3.5 ton or less, powered by a fuel not taxed at source. In New Zealand, diesel fuel is not taxed at source, as it is used for 1/3 off road. Nowadays, the mechanical distance-measuring system has been replaced by on-board units (OBUs) not only based on the odometers, now electronic, but also using satellite positioning and able to report remotely via cellular communications to their base stations the distance traveled.

Switzerland has adopted in 2000 a similar system; the toll is applicable on the whole road network. This toll scheme was adopted by a federal referendum in 1997.

Toll must be paid for all the vehicles and trailers, which

- have a total weight of more than 3.5 ton;
- are used for the carriage of goods;
- are licensed in Switzerland and abroad and drive on Switzerland's public roads network.

The rate considers

- mileage traveled within Switzerland;
- the maximum authorized weight;
- the level of pollutant emission.

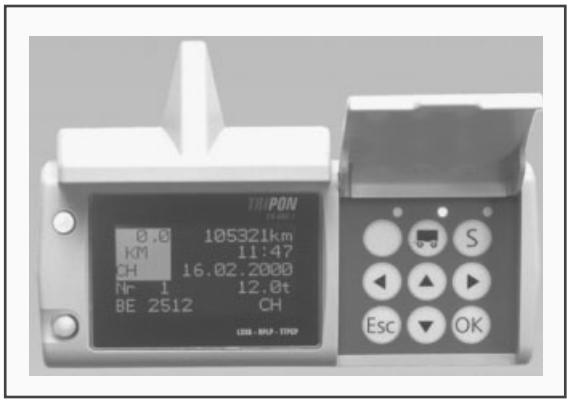


Figure 4. On-board unit for the Swiss HGV charging system.

The toll is collected on a self-declaration mode: each month, the vehicle owner must send to the customs the information registered on the OBU. Data are stored on a chip card and are transferred by internet connection, or the chip card is sent by Post (Figure 4).

The OBU is a multitechnology device that allows:

- the automatic recording of the distance traveled, thanks to a connection with the electronic tachograph;
- to start or stop this recording when crossing the border, thanks to a DSRC[5] link between the OBU and beacons;
- to interrupt the recording when abroad, thanks to GPS (global positioning system), or when the vehicle is on a train, thanks to a movement detector device;
- comparison between distances computed from the tachograph data and distances obtained by GPS tracking. This comparison may be used for odometer calibration;
- To detect if a trailer is present or not.

The OBU is mandatory for vehicles registered in Switzerland and freely available for the HGV owner. However, the owner is responsible and pays for the installation. Exempted vehicles are equipped with a DSRC tag for facilitating the control tasks.

For foreign vehicles, terminals are installed at the borders. Drivers must register the first time they enter the country and declare on the terminals the distance traveled in the country when they leave the country. For regular journeys, they can pay in advance for a defined distance.

Germany introduced a distance-based road pricing system in 2005. Only HGV with total weight over 12 ton and circulating on the federal motorway network (12 000 km) are charged. Rates vary according to the number of axles and pollutant emissions. As the toll is only applicable

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

on motorways, the system must identify precisely the position of the vehicle.

The toll collection system is based on a specific OBU equipping around 500 000 vehicles. Owing to the cost of the OBU (around €500) and the installation constraints that force to immobilize the vehicle for half a day, it appeared difficult to make it mandatory. Therefore, a manual system has been implemented for the nonequipped vehicles (around one million at the early stage).

The OBU (heavy client type) includes the following features:

- A location system based on three components: satellite positioning, in-vehicle equipment (compass and odometer), and infrared communication with roadside beacons, which is used for regular recalibration of the position, especially where the GPS signal is not reliable, and for control purpose.
- A communication module based on cellular GSM/GPRS (global system for mobile communications/general packet radio service) network.
- An embedded computer that calculates in real time the toll amount due according to the vehicle characteristics and a tariff table that is downloaded from the central office.
- A 5.8 GHz DSRC interface ready for ensuring interoperability with DSRC-based toll systems in other countries.
- A digital map of the network, regularly updated through the cellular network.

The computerized toll amount that is calculated on-board is regularly sent to the toll collection center. The digital map, the on-board software releases, and tariffs are updated by the toll center and regularly sent via the cellular radio network to each registered vehicle.

Control is ensured by fixed equipment (gantries) and mobile patrols, both equipped with infrared communication devices, which can check remotely the OBU status. In case of nonconformity, the video cameras installed on the gantries take a photo of the license plate, or the patrol intercepts the vehicle.

The manual system is based on a "reservation" by the driver of the itinerary he or she intends to follow. This can be done via Internet or on the terminals installed at the borders and on service areas. According to the vehicle parameters, the length of the itinerary, and the dates of travel, an invoice is established and the user pays online. All data are transmitted to the toll center, which can track the registered vehicle on the declared itinerary and verify that the followed itinerary is in line with the reservation.

2.5 Tolling in urban areas and congestion pricing

According to the World Road Association (PIARC, Permanent International Association of Road Congresses),[6] "congestion pricing works by shifting some less critical or more discretionary rush-hour highway travel to other transportation modes or to off-peak periods, taking advantage of the fact that the majority of rush-hour drivers on a typical urban highway are not commuters. By removing a fraction (even as small as 5%) of the vehicles from a congested roadway, pricing enables the system to flow much more efficiently, allowing more cars to move through the same physical space."

Congestion pricing or congestion charges are based on market economics principles, assuming that users have to pay for the additional negative externalities they create when they drive during peak periods. The objective is also to educate the drivers on the impact their behavior has on the environment. Application is mainly in urban areas suffering from traffic congestion during the commuting peak periods.

We can distinguish three types of systems:

- a cordon area around a city center, with charges for passing the cordon line;
- area-wide congestion pricing, which charges for being inside an area;
- targeted pricing, where access to a corridor, a lane, or a part of the street network is priced.

In 1975, Singapore was the first city in the world to introduce an area-wide congestion charging scheme system. Access to the city center was only permitted to motorists having paid a license and displayed a vignette on the windscreen. Gantries were installed in order to clearly demarcate the charging zone. These gantries were monitored by police officers who carried out visual checks and recorded any violations. At that time, the toll rate was flat ($3 a day). In 1998, the system was replaced by an electronic system: the vignette has been replaced by a tag, which communicates with beacons installed on gantries, via a DSRC link. The tag bears an electronic purse (cash card) and each time the user drives through the gantry, a certain amount is deduced. This amount varies and depends on the location and time. In case of insufficient credit, or absence of tag, a photo is taken and a fine is sent to the registered vehicle owner.

London introduced its congestion charging system in 2003 covering the London CBD (central business district) (around 21 km^2). From that date, each motorist who enters the charge zone, parks, or drives inside must pay a daily

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

charge of £10 (2012 tariff) between 7:00 and 18:00 Monday to Friday, except during bank holidays.

Payment is made in advance by registering via various means. Users can subscribe to various options (for 1 week or 1 month) and obtain discounts. Residents have a 90% discount. Registration is attached to a vehicle number plate and the control is ensured by video cameras installed at the entries and inside the charging zone, which allow with an automatic number plate recognition (ANPR) facility to check if the passing vehicles have been registered in the central database.

There are as well special measures for companies' vehicle fleet and commercial vehicles.

The Stockholm congestion charge or "Stockholm congestion tax" is a congestion pricing system implemented with the purpose of reducing the central city congestion and improves the environmental situation. It was implemented on a permanent basis on August 1, 2007, after a 7-month trial period between January 3, 2006 and July 31, 2006, and a referendum held in September 2006, which resulted in a yes from the city residents. Though the neighboring municipalities voted against, the newly elected parliament decided for a permanent implementation in June 2007. The tax is levied on weekdays between 6.30 a.m. and 6.29 p.m. No tax is charged on public holidays, the day before a public holiday, or during the month of July (Figure 5).

1. The car passes a laser detector, which activates the cameras. (When the OBU was in use, this also communicated with the transceiver aerials.)
2. The front number plates are photographed.
3. The rear number plates are photographed.

During the trial phase, the technology used was DSRC with tags equipping the vehicles. Gantries have been installed at 18 entries to the city center. It is worth to mention that Stockholm is an archipelago, which makes the access control easier than in other cities of this size.

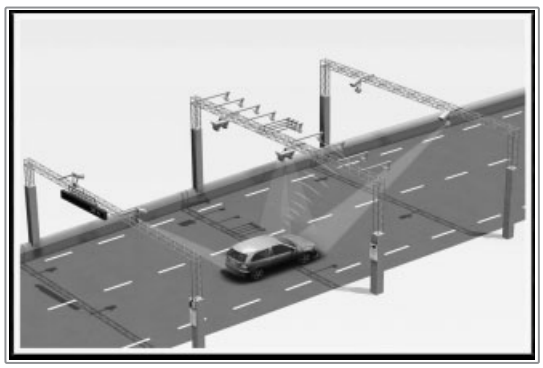


Figure 5. Stockholm toll system. (Reproduced from Swedish Transport Agency. © Swedish Transport Agency.)

When the system entered in the full operation phase, the DSRC component and the tags have been abandoned. The system relies now only on video.

Implementation of the congestion pricing has reduced car traffic in the city by 20–25% and queues by 30–50%, the impact being similar to what has been observed for London. Most of the travel demand has been transferred to public transport and a part has shift to noncharged periods of the day.

Though currently limited to a small number of cities in the world, concerns about nonrenewable energy consumption, emissions, and climate change are currently leading politicians and local authorities to think on different solutions to limit car usage in cities. Solutions can be road pricing, with charges limited to more pollutant vehicles, or even banning them from the city centers. This is the aim of the low emission zone (LEZ) policy developed by Europe.

3 ELECTRONIC TOLL COLLECTION FEATURES

ETC is a generic term that refers to all types of systems based on electronics, which allows drivers to ride on a tolled road and pay without stopping at a tollgate, the tollgate existing or not!

As introduced earlier, ETC has facilitated the development of modern infrastructures with concession to the private sector of the construction and operation, and has made possible a more equitable contribution of the transport of goods to road maintenance and operation costs. It has also made feasible the implementation of traffic demand management strategies in urban context for alleviating the congestion.

Now, a little bit of history. According to Wikipedia, in 1959, Nobel Economics Prize winner William Vickrey was the first to propose a system of electronic tolling for the Washington metropolitan area. He proposed that each car would be equipped with a transponder. "The transponder's personalized signal would be picked up when the car passed through an intersection, and then relayed to a central computer, which would calculate the charge according to the intersection and the time of day and add it to the car's bill."[7]

Several pilots were undertaken between the 1960s and the 1980s. For example, in 1986, on the eastern motorway near Paris at Coutevroult, ETC was tested for the first time with fixed transponders at the undersides of the vehicles and readers consisting of magnetic loops, which were located under the pavement.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Norway has certainly been the world's pioneer in the implementation of this technology. ETC was first introduced in Bergen, in 1986, operating together with traditional tollbooths. In the 1990s, many European countries with tolled motorways started implementing ETC systems. Boosted by industry and also by the European Commission who included in its first R&D program dedicated to transport (DRIVE I—dedicated road infrastructure for vehicle safety in Europe), various technologies were tested and implemented.

One challenge has been to introduce free-flow tolling or open road tolling (ORT) where there is no more the need for motorists to stop or slow down. The first system implemented was the E 407 near Toronto, Canada in 1997. The system is based on two technologies: motorists can choose to use a transponder (mandatory for HGV) or not. In that case, a photo of their license plate is taken and an ANPR process allows identifying them. Toll bills are sent to motorists on a monthly basis. As the operating costs of the video solution are higher than using the transponder facility, an additional fee is charged to nonregistered motorists.

Another pioneer in free-flow tolling system is that of Melbourne CityLink, opened in 2000. The innovative nature of this toll system was a deciding factor in the concession holder consortium's choice of this urban freeway. The CityLink was one of the world's largest multilane ETC systems at that time, based on DSRC technology. It is divided into eight separate toll sections. These are freeway sections where users pay a toll on entering. In order to use the CityLink, motorists must, therefore, take out a contract with the operator. Occasional users can pay the toll by phone in the days before or following their use of the toll road. The contract can be prepaid or per day. To protect against fraud, a video camera system enabling car registration numbers to be photographed and an automatic vehicle classification (AVC) system have been installed on the gantries. A major difference with E 407 is that motorists are encouraged to subscribe and get a tag. If they are not equipped with the tag and if they do not pay within the allowed period, maximum 2 days after, they are considered as defrauders and are charged with penalties.

Nowadays, in the world, a great variety of technologies for ETC are used. They are presented in Section 3.1.

3.1 Technologies

ETC scheme rely on three major components: "toll event" identification and characterization [which includes automatic vehicle identification (AVI) and classification and environment parameters], transaction (invoicing and processing), and violation enforcement, as shown in Figure 6. In addition to the pure transaction process, support functions are necessary, particularly CRM (customer relationship management) for managing the contracts with the subscribers, delivering the equipment (tags, OBU, etc.), and manage incidents, exceptions.

The three main components are somewhat independent, and can use different technologies. They can also call for different operating bodies, and this is a general trend in order to facilitate the interoperability between different operators. However, in any case, implementing an ETC scheme requires adopting a comprehensive system approach.

3.2 Toll event

The objective of this component is to collect all the information necessary to establish the toll amount due by

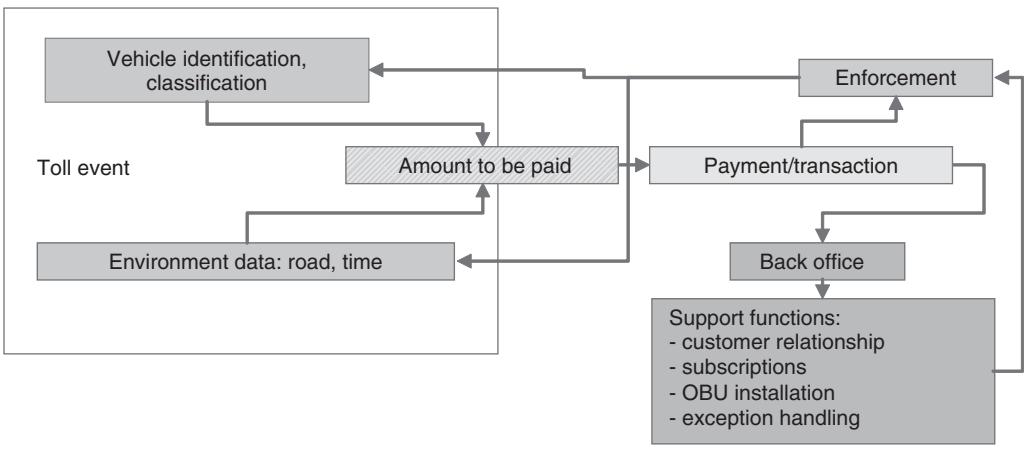


Figure 6. Tolling process. (Reproduced by permission of Egis. © Egis.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the user. It relies on different functions such as AVI and AVC.

AVI is the process of determining the identity of a vehicle subject to tolls at one point of the road network.

This point may be a point materialized on the road, for example, a classical tollgate with or without barriers, or a gantry. In that case, AVI can be performed, thanks to

- radio-frequency identification (RFID), where a beacon installed on the infrastructure communicates with a transponder on the vehicle via short-range communication. The transponder may be an RFID tag (e.g., the Salik System in Dubai) or a read/write microwave transponder (European DSRC);
- video system taking photo of the vehicle and license plates (front and/or rear plates) like in London and Stockholm.

For toll systems based on satellite positioning, the toll point is virtual, as it is not materialized on the road. The toll point is identified by the OBU within a database, or a digital map (German HGV toll system).

AVC: as most toll systems charge different rates for different types of vehicles, it is necessary to distinguish the categories of vehicle. The first and basic method is to reserve at tollgates some lanes for light vehicle equipped with a gantry at a limited height. However, this method cannot be used for meeting the current requirement of road pricing schemes that distinguish many other criteria than only the vehicle height.

The current method is to store the vehicle class in the tag or in the OBU, and use the AVI data to look up the vehicle class and parameters. However, simple tags cannot deal with vehicles that may change category, such as trucks with trailers. OBU can integrate this facility (see Switzerland).

More complex systems use a variety of sensors for counting the number of axles, as a vehicle passes over them. Light-curtain laser profilers record the shape of the vehicle, which can allow to distinguish clearly the type of vehicle but they are quite expensive.

In addition to vehicle identification and classification, establishing the toll amount requires to consider the different other charging criteria such as the time, date, and toll rate at the location. The process for calculating the toll amount is in fact between the toll event characterization and the transaction process.

3.2.1 Transaction processing

Different options exist depending on the systems used; in most case, the toll due is calculated at the toll center with the data collected from the toll point and/or the vehicle. However, in some case, the amount due is calculated by the OBU (German case) or on site (prepaid card coupled with DSRC tag, for example); the amount due is immediately deducted.

Transaction processing is performed through the back office. The processes are very similar to those of mobile phone companies. It consists of user accounts management, posting toll transactions and user payments to the accounts, and handling user inquiries. The user is in fact a consumer of road facilities and the transaction-processing component of some systems is referred to as a *customer service center*.

As we have seen in the different examples given, customer accounts may be postpaid (toll transactions are invoiced to the user on a periodic basis) or prepaid (the user funds a balance in the account that is then depleted as toll transactions occur).

3.2.2 Violation enforcement

With the development of ETC without physical barriers, the risk of fraud has increased. A violation enforcement system (VES) is necessary for limiting the number of unpaid tolls. Several methods and techniques are used to dissuade toll violation and to identify defrauders.

As ETC is based on dematerialization of processes, police visual checks are no more effective. ANPR, already used in some case as the vehicle identification method, is the basic tool in violation enforcement. Basically, the VES works in conjunction with a detection system that verifies if the vehicle is respecting the ETC rules:

- For ETC using transponders (RFID, DSRC tags, etc.), the first verification consists in detecting if the vehicle is equipped. If not, a photo of the number plate is taken and verification is made with the database of registered users: it can be occasional users who have registered for a daily pass, for example. It can be also subscribers who have obtained a tag, but have neglected to put in on the windscreen; in that case, generally, they will not be considered as violators and a simple warning will be sent to them with notification of an excess charge. If the identified vehicle does not enter in these categories, a violation notice will be issued and generally transferred to the authorities who can have access to the vehicle registration database (very few countries allow that the toll operators have access to this database).
- Concerning ETC using "smart transponders," the first verification can be more in depth: a communication is established with the roadside equipment (RSE) in order to check if the user's account is in order, if the vehicle class corresponds, and so on. The same process

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is undertaken for ETC based on satellite positioning and dedicated OBUs. It requires that the OBUs are equipped with the proper interface in order to set up this communication.

Therefore, the checking procedure necessitates that RSE are installed in various locations of the network. The equipment can be fixed, removable, or mobile. Fixed equipment is typically gantries over the road supporting the checking equipment (beacons) and the video system. Communication is ensured with a processing center via wire or wireless connection. Mobile enforcement can be also achieved with vehicle patrols equipped with similar equipment. Using patrols allows as well intercepting immediately the "suspicious vehicle."

A major issue concerning enforcement is that of the legislative context: the ANPR allows identifying the vehicle owner and not the driver, and in some countries, modification of the law has been necessary for being able to make the owner responsible of the traffic offences under some circumstances.

4 INTEROPERABILITY

With the development of toll systems, for different objectives leading to specific technical solutions within a country or a region, users and particularly professional users and commercial vehicles are facing a new difficulty, which is to be obliged to subscribe to various toll operators and be equipped with different tags or OBUs.

The problem was raised in Europe and in the United States during the early 1990s and several projects and pilot projects were undertaken to overpass the difficulty. The solution, which could have been to adopt a standardized system and choose a sole technology, was not feasible because of the investment already made and also to the different context of tolling.

In Europe, however, a first step was made by adopting a standard for DSRC-based tolling systems, with the selection of 5.8 GHz as the telecommunication bearer. This standard (CEN/ISO 14906) has been used also in different other countries such as Australia (Melbourne CityLink) and South America.

For example, in France, the seven different motorway companies that were operating in the mid-1990s, adopted for light vehicles, the first published toll standard, under the brand Liber-T and developed bilateral agreements between them. The system is still in place and the user can subscribe to any toll company provider and drive everywhere in France. All travels are recorded by each company where the user drives (toll charger). A single bill is then established by the toll service operator (the one to which the user has subscribed) after reconciliation of all travels. This system needs bilateral agreements between each of the companies, which becomes complex with the apparition of new toll operators. A further step has been to allow external service providers to offer the toll service, with specific advantages.

This evolution has led to a clear distinction of the roles between the road operator (termed *toll charger*) and the organization that provide the toll service (the toll service operator).

The toll service architecture can be represented as in Figure 7, according to the recommendations of the EU-related projects[8] and standardization bodies (CEN TC 278 and ISO TC 204).

This model has allowed to clearly identify where the standards need to be developed, that is, for ensuring proper communications at the interfaces between the four entities. A common approach has been developed between CEN and ISO in order to optimize the resources dedicated to this complex work. The WGs involved in this task are the WG1 (electronic fee collection) and the WG5 (fee and toll collection), respectively for the CEN TC 278 and the ISO TC 204.

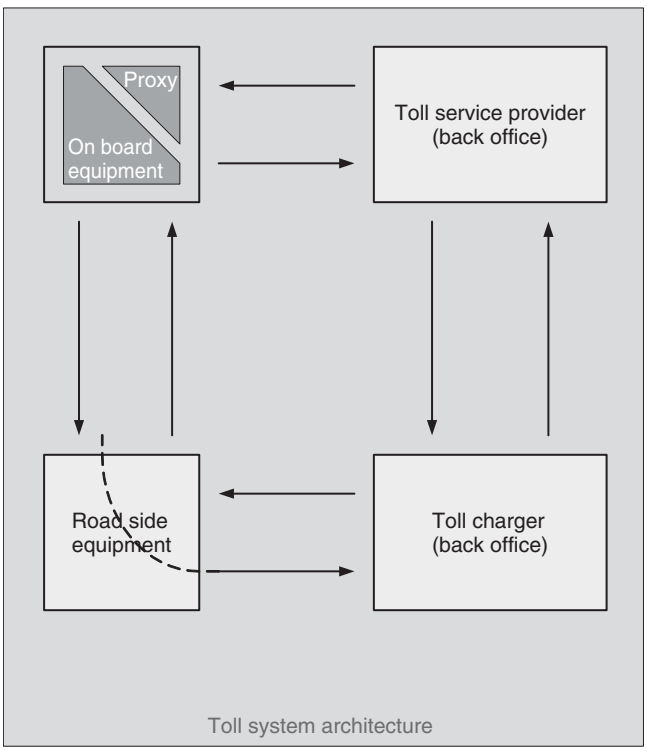


Figure 7. European EETS system architecture.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

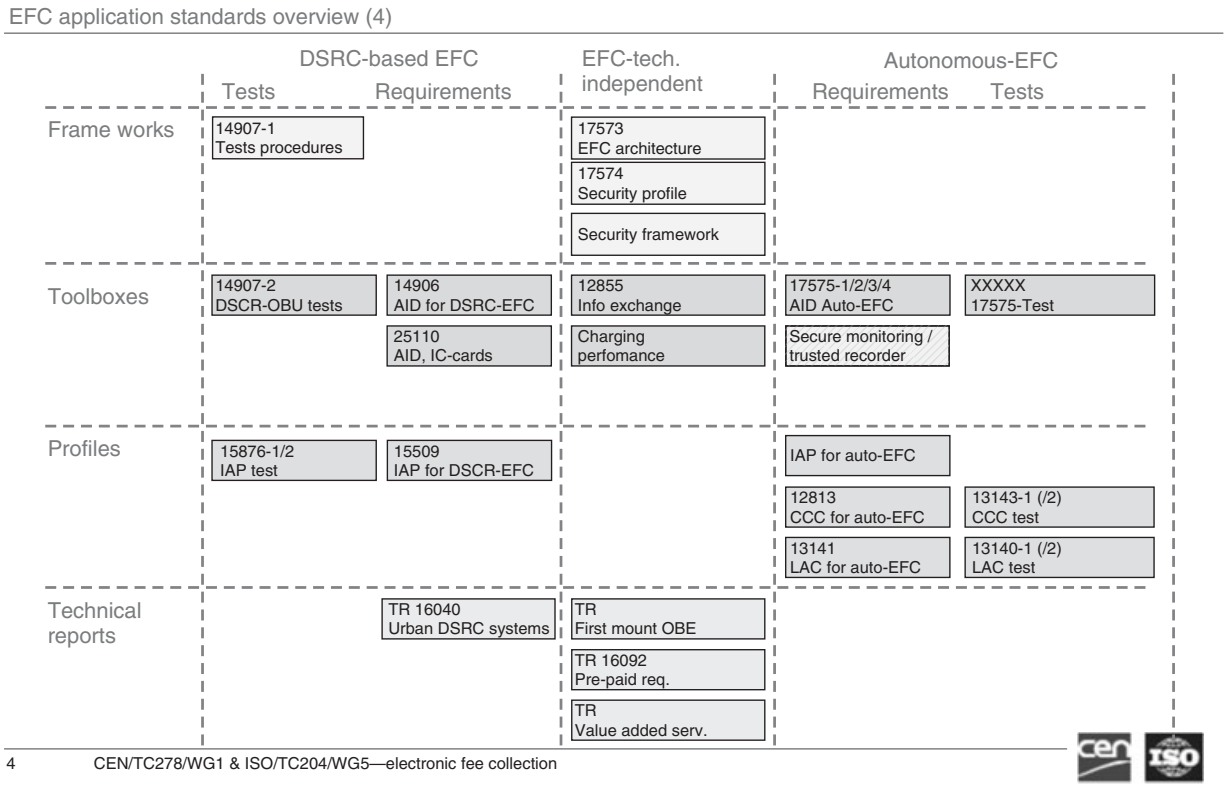


Figure 8. Overview of involved standards in ETC.

Figure 8 gives an overview of the standards that support the ETC services.

Within the RCI project,[9] 27 principal road charging stakeholders in Europe defined, implemented, demonstrated, and recommended this standard architecture for offering a European road charging service to the user based on "one box, one contract, and one invoice."

The final and ultimate task within the project was the demonstration phase in Spring 2008 when two trucks, each equipped with one interoperable OBE (on-board equipment) that seamlessly, and without user intervention, adapting functional behavior when crossing borders, drove through different countries according to the rules that apply for the different tolling schemes, that is, German (Toll Collect), Swiss (LSVA), French (TIS-PL, Télépéage Inter Sociétés pour les Poids-Lourds), Spanish (VIA-T), Italian (TELEPASS), and Austrian (ASFINAG, Autobahnen- und Schnellstraßen-Finanzierungs-Aktiengesellschaft).

Thanks to the work undertaken by the different parties and the willingness of the EU member states to go for interoperability of ETC system in Europe, which is considered as a necessity for supporting the free circulation of persons and goods; the European Union issued the Directive 2004/52/EC and the related Decision 2009/750/EC in order to set up the European Electronic Toll Service (EETS).

The directive, called *directive on interoperability*, stipulates that new toll systems using electronic solutions should use one or more of the following technologies:

1. satellite positioning;
2. mobile communications using the GSM–GPRS standards (reference GSM TS 03.60/23.060);
3. 5.8 GHz microwave technology.

The Guide for the application of Directive 2004/52/EC of the European Parliament and of the Council and of Commission Decision 2009/750/EC[10] details all aspects to be considered for setting up the EETS; the target date for starting EETS was October 2012. Some delays are expected, mainly because of contractual and commercial issues, the technical problems being theoretically solved.

In the United States, projects are also in development. The E-Zpass toll service was the first interagency service to be implemented, starting with seven toll agencies in 1991 and currently there are 25 toll agencies working with E-Zpass. Other projects are underway, with the objective of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

creating clearinghouse bodies for facilitating the interoperability of toll services.

One difference between Europe and the United States on this matter is certainly due to the difference in the road user market segments: in Europe, the first target for toll interoperability is the long distance HGV crossing several countries. Each transaction represents several tens of Euros. Toll chargers are, therefore, attached to stay in the loop and develop alliances to become EETS providers. In United States, the market is more open and the mean value of transaction is quite low. Owing to the huge amount of the total transactions, if we consider as well the parking and public transport payments, the bank sector is very much interested in being the main actor (Opiola, 2011).

5 CONCLUSION

The tolling domain is an important component of the traffic system. From the initial stage where the objective was mainly to finance new infrastructures, the toll systems have evolved in tolling services, offering new opportunities for the traffic and road network operators.

With the introduction of ICT in the transport domain and particularly in tolling, leading to the ETC concept, tolling has certainly been the pioneer of the connected car concept.

New actors have emerged for taking in charge the new opportunities of services in the domain. They will play certainly a key role in the development of the connected car services, for which the business models are still uncertain.[11]

6 GLOSSARY

ANPR	Automatic number plate recognition
AVI	Automatic vehicle identification
AVC	Automatic vehicle classification
CRM	Customer relationship management
DSRC	Dedicated short-range communication
EETS	European Electronic Toll Service
ETC	Electronic toll collection
GHG	Greenhouse gas
HGV	Heavy goods vehicle
ITS	Intelligent transport systems
OBU/OBE	On-board unit/equipment
RFID	Radio frequency identification
RSE	Roadside equipment
RUC	Road user charging
VES	Violation enforcement system

ENDNOTES

1. Federal HighWay Administration, US DoT.
2. http://www.fhwa.dot.gov/ipd/revenue/road_pricing/defined/tolls.htm
3. DIRECTIVE 2006/38/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of May 17, 2006 amending Directive 1999/62/EC on the charging of heavy goods vehicles for the use of certain infrastructures.
4. http://www.transport.govt.nz/ourwork/land/roadusercharges/
5. DSRC: Dedicated short-range communication, name given to the microwave communication bearer, operating at 5.8 GHz, which has been standardized in Europe by the CEN.
6. PIARC Technical Report, Road Network Operation Committee B2 2008–2011.
7. Quoted by the Victoria Transport Policy Institute, Canada, http://www.vtpi.org/vickrey.htm.
8. CARDME 4 (Project IST-1999-29053, Deliverable 4.1, (Final), June 1, 2002 and CESARE IV (common electronic fee collection system for a road tolling European service) project set up by ASECAP (the European Association of Toll Motorways Operators) and cofinanced by the European Commission.
9. RCI project RCI_WP8 Compendium_V1.1 See http://www.ertico.com/rci, the RCI project was cofinanced by the European Commission DG TREN.
10. http://ec.europa.eu/transport/media/publications/doc/2011-eets-european-electronic-toll-service_en.pdf
11. The connected vehicle, PIARC FISITA report *Copyright by the World Road Association and FISITA*, available the World Road Association web site http://www.piarc.org and from the FISITA web site http://www.fisita.com.

REFERENCES

Gilliet, H. (1990) Toll roads-the French experience. Transroute International, Saint-Quentin-en-Yvelines. (Transroute International became Egis Projects in 1998.)

Opiola, J. (2011) An open system architecture model for electronic payments: a new paradigm for road pricing? D'Artagnan Consulting LLP, in Routes/Roads magazine, N351 special issue on ITS.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto185
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Applications: Emergency Services and eCall

Frank Försterling

Continental Automotive GmbH, Hanover, Germany

1	Introduction	1
2	eCall Definition and Classification	2
3	eCall Technical Aspects	4
4	Pan-European eCall	8
5	ERA-GLONASS-Based eCall	8
6	Standardization	8
7	Outlook	11
Glossary		12
Further Reading		12

1 INTRODUCTION

Emergency services—which are also named *emergency management services* (*EMSs*)—are considered as one of the areas of intelligent transportation system (ITS) applications (Figure 1).

EMS comprises a wide range of incident response, public safety management, and disaster management services. EMS includes the management and coordination of emergency response resources, infrastructure protection, emergency warnings, disaster response and recovery, and evacuation and reentry management. An overview of EMS categories has been proposed by the US administration RITA (Research and Innovative Technology Administration):

- Hazardous materials management
  ITS applications associated with hazardous materials shipment can accomplish several major functions intended to provide safe and secure transport of hazardous materials by road.
- Emergency medical services
  Advanced automated collision notification and telemedicine address the detection of and response to incidents, such as vehicle collisions or other incidents requiring emergency responders.
- Response and recovery
  The variety of sensors deployed on the transportation infrastructure can help provide an early warning system to detect large-scale emergencies, including natural disasters and technological and man-made disasters.

The emergency call service—commonly named eCall (service)—is one of the most important EMSs.

The use of eCall to deploy emergency assistance will save lives and reduce the social burden of road accidents by improving the notification of such accidents; speeding up the emergency service response; and lowering the subsequent effects on fatalities, severity of injuries, and traffic flows. On the basis of impact assessment provided by the European Commission (EC), the mandatory introduction of eCall in Europe will cause the following effects:

- Reduction in traffic death rates (5%) and number of severe injuries (15%) across Europe.
- Reduction of response time 40%–50% ("golden hour"): from approximately 21 min to approximately 11 min.
- Reduction of traffic congestion by 20%.

Owing to the importance of eCall for our society, this contribution will be focused on a detailed explanation of the eCall service only.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5



Figure 1. Emergency management services as part of ITS applications.

In the following, we first present the definition of an eCall and classify the different eCall options (Section 2). Afterwards, we describe the technical aspects of an eCall service (Section 3). Section 4 describes the aspects of a standardized pan-European eCall, which most probably will be implemented from 2015 onward. In addition to the EC member states, Russia plans to introduce a mandatory eCall as well, which is called *ERA GLONASS*. On the one hand, ERA GLONASS will enable interoperability with the pan-European eCall, while on the other hand it is characterized by several differences. Section 5 considers the ERA GLONASS approach. Section 6 gives an overview of all standardization activities with relevance to the eCall (Europe, Russia). We close this chapter with an outlook on the next steps once eCall has been deployed successfully (Section 7).

2 ECALL DEFINITION AND CLASSIFICATION

2.1 Definition of eCall

The *eCall* is an emergency call either generated manually by vehicle occupants or automatically via activation of in-vehicle sensors when an accident occurs. When activated, the in-vehicle eCall system establishes a voice connection directly with the relevant service provider (SP), which is a public authority or a private eCall center. At the same time, a set of data—including key information about the accident such as time, location, and vehicle description—is sent to the SP operator receiving the voice call. The set of data may also contain the link to a potential further SP by including its IP address and phone number.

2.2 Classification of eCall

2.2.1 Private eCall

The private eCall—also called *third-party services* (*TPSs*) eCall—follows the above-mentioned definition.

- In-vehicle detection of emergency situation (manually or automatically).
- Initiation of an emergency call (any number defined by the eCall provider).
- Provision of data defined by the eCall provider.
- SP receives eCall and initiates emergency measures.

The private eCall has already been implemented by several car manufacturers and works very well.

The drawback of the private eCall solutions are as follows:

- Closed technical solution due to missing common standards;
- Limited deployment (country specific) due to missing infrastructure.

Recognizing the drawbacks of private eCall solutions on the one hand and accepting the social-economical

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

importance of the global rollout of standardized eCall solutions on the other hand, several governments decided to go for a standardized eCall solution.

2.2.2 Pan-European eCall

The pan-European eCall is based on an initiative of the EC and its member states. It is based on well-defined standards and follows the above-mentioned definition.

- In-vehicle detection of emergency situation (manually or automatically).
- Initiation of an emergency voice call (based on 112, Europe wide).
- Provision of a well-defined data set—the minimum set of data (MSD)—from the car to the SP.
- The SP is called *public safety answering point* (*PSAP*), which is a public authority or a private eCall center that operates under the regulation and/or authorization of a public body.
- The PSAP receives eCall and initiates emergency measures.
- The network operator identifies the 112 call as an eCall and marks this call via an eCall discriminator, adds location information as well as calling line identification (CLI) data.

The difference between a 112 call, an E112 call and a (pan-European) eCall is as follows:

- 112 is the single European emergency number, which was recently generalized in the entire European Union.
- E112 emergency calls support the immediate localization of the caller.
- The (pan-European) eCall is an E112 call initiated by a vehicle and which, in addition to the voice call, transfers the MSD data set.

2.2.3 ERA GLONASS eCall

The ERA GLONASS eCall is based on the decision of the Russian government to introduce a standardized eCall solution for Russia, which

- follows the definition of the pan-European eCall,
- is interoperable with the pan-European eCall,
- includes additional features for the Russian market.

The key differentiator of ERA GLONASS in comparison to the pan-European eCall is the in-vehicle identification of the position of the vehicle. The Russian government requires to apply the satellite system of Russia (which is called *GLONASS*), whereas the pan-European eCall is based on the US satellite system GPS (global positioning system).

Further differences will be explained in Section 5.

2.2.4 Personal eCall

The personal eCall utilizes for the initiation of the emergency call a Smartphone or any kind of mobile device instead of an in-vehicle system (IVS). The personal eCall is not standardized. Especially, the transfer of additional data is not part of this kind of eCall: neither the MSD nor any kind of other data.

2.2.5 eCall for PTW

The eCall for PTW (powered two-wheelers) seems to be the next important step for further improvements of road safety and life saving. Intense discussions are on at the European level for the integration of PTW eCall into the pan-European eCall standards. First PTW eCall solutions already have been implemented and are part of European interoperability tests (e.g., within the HeERO 2 project).

But the motorcycle industry faces several challenges, among them the most important once are

- insufficient crash sensors to detect an eCall situation;
- different hardware design necessary: the absence of a passenger compartment and different PTW usages require an adapted engineering of motorcycle eCalls;
- cost impact of eCall and PTW.

The amount of unresolved issues does not presently provide the necessary certainties required for making eCall available to all PTWs.

2.2.6 eCall for HGV

The rapid increase in HGVs (heavy goods vehicles) on the roads requires an extension of the pan-European eCall solution for HGV.

The key challenge for an HGV eCall is the extension of the data message as part of an eCall. A rapid assistance will benefit in early knowledge about dangerous goods involved in an accident. The current eCall MSD standard offers no designated coding scheme for information about dangerous goods.

Furthermore, owing to the time and capacity limitations of (MSD-based) in-band messaging, the provision of additional dangerous goods data might require alternative architectural approaches with implications to all parts of an eCall solution (IVS, PSAP, etc.).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Standardization activities have been started within CEN (European Committee for Standardization) in order to define solutions (Section 6).

3 eCALL TECHNICAL ASPECTS

This section describes all the technical aspects of an eCall service. The technical description is based on the pan-European eCall definition. When describing aspects of other eCall solutions, we will make an explicit note.

3.1 End-to-end consideration

For the implementation of a pan-European eCall, we have to consider three major stakeholders (Figure 2):

- The IVS, owned by the car manufacturer.
- The network infrastructure, owned by the mobile network operator (MNO).
- The PSAPs, owned by the public authorities.

Furthermore, depending on the severity of the potential vehicle crash, further stakeholders such as the road authorities, the medical emergency services, and other stakeholders might be involved in resolving the emergency situation. These further stakeholders will be informed and managed via the PSAP.

In case a serious accident occurs, in-vehicle sensors will automatically trigger an eCall. When the system is activated, the IVS establishes a 112-voice connection. As soon as the communication channel is successfully connected, an emergency data message—the MSD—is sent. The MSD includes key information about the accident, such as time, accurate location, driving direction, and vehicle description. The MSD is sent in-band via the voice channel. During the time of the MSD transfer, the voice channel is blocked (the voice part is muted) and cannot be used for direct caller–operator talks. In general, the MSD transfer takes a few seconds. Afterwards, the voice channel will be utilized for direct "PSAP operator to vehicle occupants" talks.

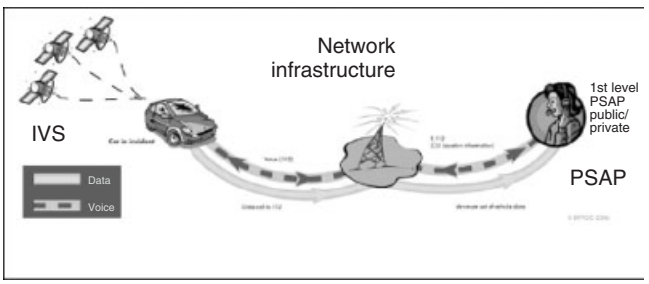


Figure 2. High level architecture of an eCall.

The eCall can also be activated manually.

The MNO identifies that the 112 call is an eCall based on the "eCall discriminator" provided by the vehicle's network access device (NAD), which is part of the IVS. The MNO handles the eCall just as any other 112 call and routes the call to the most appropriate emergency response center—the PSAP. The MNO extends the 112 call toward an E112 call by adding further information to the call such as CLI and the cell identification of the originated eCall.

The PSAP operator will receive both the MSD message and the voice call. The information provided by the MSD will be decoded and displayed on the PSAP operator screen. The location and driving direction of the vehicle can be shown using a geographic information system (GIS).

As soon as the MSD message transfer has been completed, the PSAP operator will be able to talk directly to the vehicle occupants and listen to what happened in the vehicle. This will help the operator ascertain which emergency services are needed at the accident scene (e.g., ambulance, firemen, and police) and to rapidly dispatch the alert and all relevant information to the right service organizations.

Furthermore, the PSAP operator will be able to immediately inform the road/traffic management centers about the incident. The operator provides information about, for example, the specific location and the severity of the incident. On the basis of this information, the road authorities can immediately disseminate this information to other road users and thus prevent secondary accidents. They are enabled to conduct an accelerated clearance of the carriageway, which will reduce congestion situations.

3.2 In-vehicle system (IVS)

The eCall IVS comprises different parts:

- Vehicle interface unit (VIU)
- Positioning system
- Communication system
- Human–machine interaction (HMI) unit.

3.2.1 Vehicle interface unit (VIU)

The VIU as part of the IVS provides the direct secured interface to all in-car components via the in-vehicle network (IVN). The IVN connects the necessary "crash" sensors. This includes pressure, acceleration, and roll-over sensors. In addition, seat occupancy recognition sensors indicate

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the number of potentially injured persons. The sensors measures, for example, the deceleration (loss of speed) and captures the time of air-back inflations. From these values received via the VIU, the IVS calculates the crash severity. The crash data is available after 100 ms. In the future, more sensors may be developed, in order to improve the evaluation results.

An eCall will be initiated only in case of accident recognition by at least 2 of the "crash" sensors. All necessary data will be collected and put into the MSD.

3.2.2 *Positioning system*

The positioning subsystem provides the accurate location of the vehicle as well as its direction of driving.

The pan-European eCall specifications refer to standard GPS positioning service without deeper consideration of possible vulnerabilities and limitations of the position estimation process. Later, the GNSS (global navigation satellite system) solution Galileo will be applied for the pan-European eCall. The Russian eCall is based on the GLONASS solution.

The sole utilization of a single position estimation method will leave considerable gaps in performance (e.g., in tunnels, urban areas), causing potentially dangerous effects on emergency operations and restoration of the traffic flow after the accident.

The standard GPS positioning performance can be enhanced by the utilization of, for example,

- satellite navigation augmentation systems, such as the European geostationary navigation overlay system (EGNOS) by the provision of three additional GPS-like signals;
- accelerometers: in combination with the GPS, accelerometers provide an inertial add-on solution that will continuously provide accurate position estimates with the temporal lack of GPS signals;
- map matching algorithms of modern cars (in case the vehicle is equipped with a car navigation solution);
- the network resources of the MNO, delivered by the IVS communication system.

3.2.3 *Communication system*

The communication subsystem basically comprises the NAD necessary to

- create the "eCall discriminator,"
- set up the 112 voice call,
- forward the data message (in case of the pan-European eCall the MSD)

via the public land mobile network (PLMN) through the MNOs to the most relevant PSAP.

The NAD operates in either of two modes: GSM (global system for mobile communications) or UMTS (universal mobile telecommunications system). For the pan-European eCall, the final decision of the requested network mode is still pending (status January 2013). The ERA GLONASS solution requires the support of both network types (GSM and UMTS).

The NAD has to support an embedded SIM in order to guarantee the "always available" feature. In case the communication system is configured to perform emergency calls only, the execution of mobility management procedures shall be avoided.

The eCall system uses an in-band data modem to transmit the MSD information over the voice path to the PSAP. For this purpose, the NAD (as well as the PSAP) has to be equipped additionally with a data modem. The data modem solution is based on a proposal of Qualcomm and was endorsed by the 3GPP in September 2008. This license-free approach enables the eCall solution to be quickly deployed end-to-end in-vehicle IVSs and PSAPs without modifications to the existing cellular and wireline infrastructure.

The pan-European eCall allows the transfer of data via the in-band modem solution only. Other options for the data messaging are

- SMS (short messaging service)-based data transfer (e.g., specified as a backup solution within ERA GLONASS);
- data transfer via (voice channel) out-band data channels [e.g., via USSD (unstructured supplementary service data) or GPRS (general packet radio service)].

The advantage of utilizing data messaging outside of the voice channel is the option of enlarging the amount of data to be transferred. The MSD data size is no longer a limitation for the data transfer. The disadvantage of this kind of data communication is mainly related to the potentially real-time and performance challenges. Owing to missing standardization activities, the out-band data transfer can be applied to TPS eCall solutions only.

3.2.4 *Human–machine interaction (HMI) unit*

The HMI Unit includes the necessary components to inform the vehicle occupants about the status of the eCall system, as well as the status of the eCall transaction when triggered.

Parts of the HMI unit are at least a microphone and a loudspeaker. Potentially, the vehicle's display might be used as well.

The HMI unit may be dedicated to the eCall device or part of the general vehicle HMI.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.3 Network infrastructure

MNOs have the responsibility of handling eCalls as any other 112/E112 emergency call, including the CLI and caller location information, and supporting the "eCall discriminator." The eCall has to have the same priority and reliability as any other emergency call through the MNO network.

The responsibility for processing eCalls and routing them to the correct PSAP always lies within the network serving the vehicle at the time of activation.

For the proper deployment of the pan-European eCall, the MNOs have to deploy a technical upgrade of their network infrastructure:

- MNOs need to design and implement the "eCall discriminator" (also called *eCall flag*) in their mobile switching centers (MSCs) of the MNO network infrastructure.
- 3GPP approved the "eCall discriminator" and included it in Release 8 of the technical specifications, with which the mobile telecommunications systems must comply. This discriminator will differentiate between 112 calls from mobile terminals and eCalls, and also between manual and automatically triggered eCalls.
- On the basis of this "eCall discriminator," MNOs have to identify eCalls and route the voice and the MSD to the most appropriate PSAP as defined by national authorities in accordance to the national arrangements.
- MNOs need to agree with the public authorities on the "eCall discriminator" implementation plan.
- Furthermore, the MNOs have to upgrade the dialed 112 number toward an E112 call, which includes additional location information from the network (CLI and cell originator information).

The application of TPS-based eCalls does not impact the network infrastructure. TPS eCalls are not based on the 112 numbering scheme.

Maintaining the network infrastructure, the MNOs have to ensure that the network infrastructure of the future still supports the "eCall discriminator" as well as in-band modem functionality. The MNO has to make sure that after finalizing the MSD transfer via the in-band modem, the voice channel can be reopened for voice calls.

Finally, further specification work is necessary in order to deploy the pan-European eCall in long-term evolution (LTE)-based networks as well.

3.4 Service provider

The PSAP is the public-controlled call center responsible for providing a first point of contact to a 112 call. The PSAP thus receives the emergency 112-voice call and the MSD.

On the basis of the voice connection and the MSD content, the PSAP operator decides the handover to the correct dispatcher, which will handle the remaining part of the specific emergency response. The PSAP's source of information is the voice, the MSD, and the location information provided by E112.

The PSAP's operational models vary from country to country and—in some European member states—also between the different regions. This situation caused the necessity of running several eCall implementation and interoperability trials. One of the most important interoperability trials is the HeERO project of the EC.

The transformation of a call center into a PSAP requires a few technical upgrades, as follows:

- Update the equipment of a server with in-band modem capabilities, which allow to receiving of eCalls and extracting/translating the MSD. The server should be designed according to the number of expected eCalls for this PSAP.
- Server software upgrade for the decoding of the complete MSD.
- Server software upgrade with a GIS and an incident management system in order to allow efficient handling of the emergency call.
- Ensure connectivity to all subsequent emergency organizations (ambulance, police, etc.) as well as to the appropriate road authority organizations.
- Ensure the connectivity to TPS SPs (in accordance with the standards).

The integration into the eCall chain of a third-party service provider (TPSP) and its interworking with the PSAP operator is described in the standard "EN 16102 intelligent transport systems—eSafety—TPS supported eCall—operating requirements" (Figure 3).

The TPSP involvement into an eCall identifies additional options:

- Avoidance of false calls
  Because TPS has its SP as the "bridge" to the PSAP, it can filter out "false" calls, e.g., nonemergency calls or accidental calls.
- Efficient provision of MSD
  The PSAP requires the "MSD." In case of a TPS eCall, a few of those data might be stored in the TPSP database

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

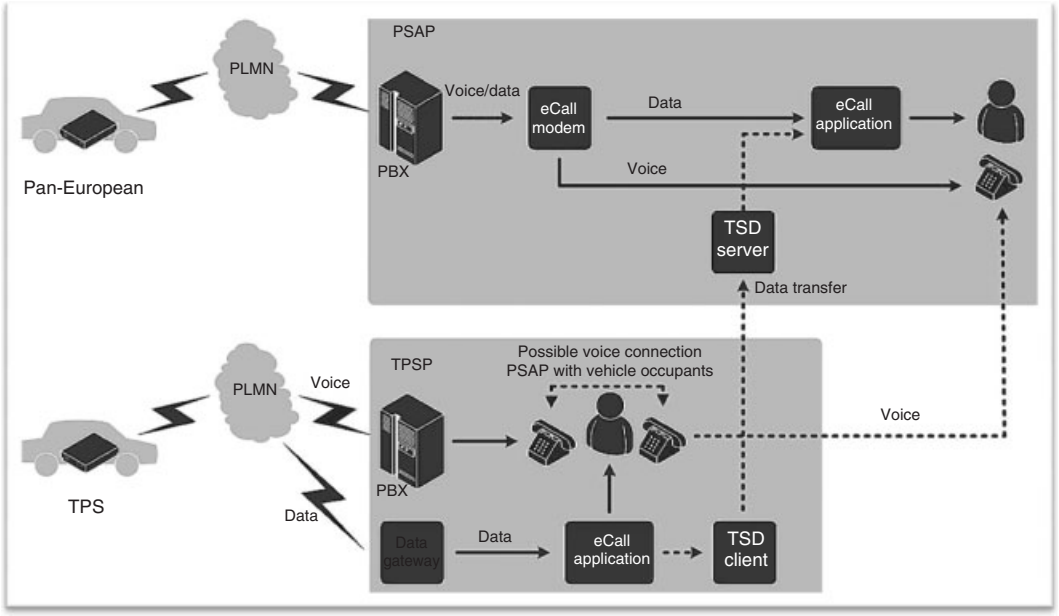


Figure 3. Integration of a third-party service provider (example).

instead of being transferred from the IVS, reducing the load on its communication with the vehicle.

- Efficient provision of additional relevant data (FSD—full set of data)
  The TPSP is also able to add additional data. For example, its customer may request that relevant medical data, or mobility information, is passed to the PSAP.
- Improved call completion rate
  In situations where the MSD transfer fails, the TPSP, although not still in contact with the vehicle, has received an alarm from the vehicle, and can forward what information it has to the PSAP.

3.5 Minimum set of data (MSD)

The MSD has been defined in the following standard: European Committee for Standardization TC 278 WG, "Road transport and traffic telematics—eSafety—eCall MSD" (Table 1).

The MSD message amounts to a maximum of 140 bytes, its main contents being:

- "When": the timestamp of the accident,
- "Where": the geographical position of the accident,
- "Who": eCall activation indicator (manual or automatic),
- "Car": vehicle type and identification number (VIN),
- number of passengers,
- optional data.

Table 1. Minimum set of data (MSD).

Name	Size (B)	Type	Description
Control	1	Integer	Bit7: Automatic activation Bit6: Manual activation Bit5: Test call Bit4: No confidence in position Bit3–Bit0: Reserved
Vehicle identification	20	String	VIN number according to ISO 3779
Time stamp	4	Integer	UTC seconds
Location	4	Integer	Latitude (WGS-84) in milliarcseconds
	4	Integer	Longitude (WGS-84) in milliarcseconds
	1	Byte	Direction in degrees. The nearest integer of 360.0 * value/255.0
Service provider	4	Byte[4]	Service provider IP address or blank field
Optional data	102	String	Further data (e.g., crash information) or blank field
Total bytes	140	—	—

The initial eCall specifications also included the attempt to define—in addition to the MSD—an FSD.

For the time being, a standardized FSD specification is not available. The provision of additional data will be

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

part of TPS eCall specifications. Technically, most of the additional data will be stored on a subscription base in an SPs database. In case of an eCall, the TPSP can retrieve the FSD locally.

Examples of an extended set of data are health data, last will data, whom to contact data, and so on. In case of the deployment of an HGV eCall, goods-related data have to be made available.

4 PAN-EUROPEAN eCALL

The use of eCall to deploy emergency assistance will save lives and reduce the social burden of road accidents by improving the notification of such accidents; speeding up the emergency service response; and lowering the subsequent effects on fatalities, severity of injuries, and traffic flows.

This led to an initiative to standardize the eCall in the European Union and to create the pan-European eCall.

The memorandum of understanding (MoU) for realization of interoperable in-vehicle eCall was submitted to the EC back in May 2004 within the context of the eSafety initiative and addressed mainly the automotive industry, MNOs, motoring associations, insurance companies, and the EU member states.

So far (as of January 2013), the MoU has been signed by

- 22 member states
- 4 further European countries
- more than 60 companies and nonprofit organizations.

The MoU lists the necessary arrangements for implementation of the eCall action plan and sets out the measures to be taken by the EC, member states, automotive industry, telecoms, and insurance industries. The MoU's key message is that eCall should work in any EU member state and that eCall should be based on the single pan-European emergency call number 112 (Figure 4).

Unfortunately, the progress of the pan-European eCall definition and deployment strategy did not show the expected results within a reasonable time frame. The EC decided to work on an eCall impact assessment.

The impact assessment on the deployment of eCall was published by the EC on 8 September 2011. The conclusion of the report, which considered a number of different options, was for the mandatory introduction of the harmonized interoperable EU-wide eCall service, based on 112 and on the pan-European standards developed by the European Standardization Organizations—in all vehicles in Europe starting by certain categories (i.e., passenger cars and light duty vehicles)—including the upgrade of MNOs and PSAPs to receive/forward and handle the eCalls. This service may coexist with the private eCall services.

The result of the impact assessment is that the EC is now actively working toward the mandated implementation of an EU-wide eCall service, with a projected introduction date of 2015 for all new type-approved vehicles.

Furthermore, the EC published an ITS action plan with 6 action areas. One of the focus areas is called *road safety and security* (area 3). This area includes as one of the important actions the "Introduction of Europe-wide eCall" (Figure 5).

The target date for the introduction of the pan-European eCall is 2015 for all new passenger and light vehicle cars.

Although the vast majority of all technical implications have been solved, a few challenges still remain open (status January 2013):

- The type approval procedure has not yet been published.
- The question of warranty is still open.
- Upgrade procedures have not yet been defined (relevant to the IVS, the communication network as well as the PSAP).
- The business model for the eCall rollout and for potential upgrade requirements is critical.

5 ERA-GLONASS-BASED eCALL

Russia has begun implementing an emergency call service, which builds upon common elements of the European eCall, extending the approach to include additional features such as GLONASS GNSS positioning and backup data transmission mechanism using SMS. The major differences between the pan-European eCall and ERA GLONASS are highlighted in Table 2.

First, ERA GLONASS deployments will be targeting dangerous cargo transportation and collective passenger transportation by October 2014. All new passenger vehicles (e.g., automobiles and light vehicles) getting new (first) type approval will be required to have the ERA GLONASS IVS installed from January 2015.

6 STANDARDIZATION

This section provides a comprehensive overview about all eCall-related standards for the purpose of introduction of the pan-European eCall as well as for the ERA GLONASS solution (Table 3).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

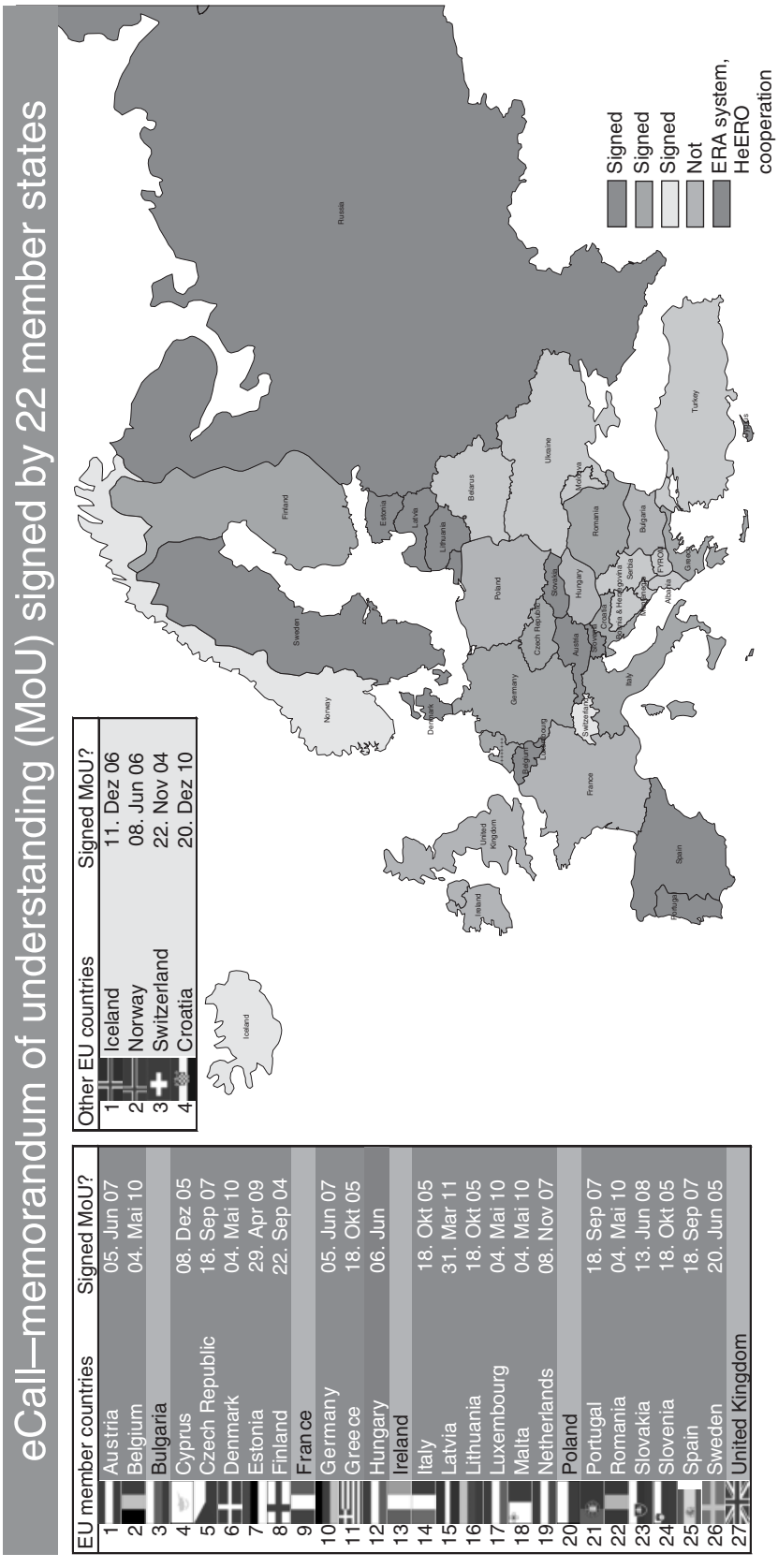


Figure 4. The eCall MoU—signature overview.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

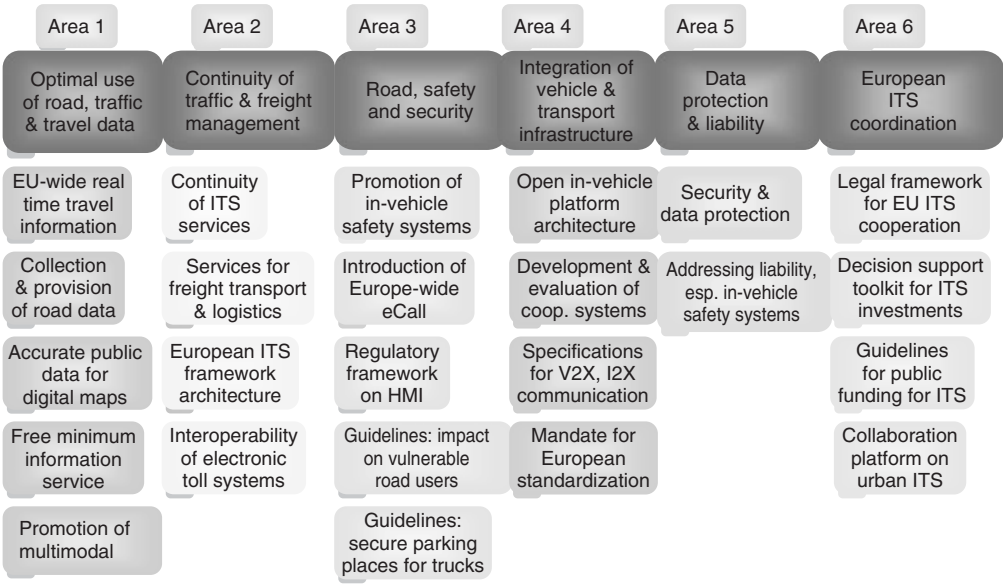


Figure 5. ITS action plan of the European Commission.

Table 2. Comparison between the ERA GLONASS and the pan-European eCalls.

Pan-European eCall	ERA GLONASS
In-vehicle unit with support of GPS (or Galileo once available)	In-vehicle unit with support of GLONASS or GLONASS/GPS
Triggered automatically by the in-vehicle device, or manually by vehicle occupants	Triggered automatically by the in-vehicle, or manually by vehicle occupants
Based on 112 emergency call (112 and E112)	Based on 112 emergency number (starting 2012)
Creates a voice link to the closest PSAP and sends a data message (minimum set of data) at the beginning of the voice call	Creates a voice link to the closest NIS Regional Center and sends a data message (minimum set of data) via voice band modem (SMS as backup)
PSAP to use local process and resources to, for example, dispatch ambulance, inform traffic centers, etc.	NIS Regional Center distributes the calls to local police PSAP, local ambulance PSAP, and local Emercom
Intention of the eCall system is to enable additional services in the vehicles and jump start telematics and ITS applications in Europe	Infrastructure of ERA GLONASS system will be served as a base for additional location-based services (e.g., fleet management, SVT)
—	Support of UMTS and GSM
—	"Blackbox" function
—	Remote IVS configuration and SW upgrade

ERA GLONASS is based on the following standards:

- GOST R 54620-2011 "GNSS. Road accident emergency response system. In-vehicle emergency call system. General technical requirements."
- GOST R 54721-2011 "GNSS. Road accident emergency response system. Base service description."
- GOST R 54618-2011 "GNSS. Road accident emergency response system. Compliance test methods for electromagnetic compatibility and environmental and mechanical resistance requirements of in-vehicle emergency call system."
- GOST R 54619-2011 "GNSS. Road accident emergency response system. Protocol of data transmission from in-vehicle emergency call system to emergency response system infrastructure."
- GOST R draft "GNSS. Road accident emergency response system. Functional test methods of in-vehicle emergency call system and data transfer protocols" (includes functional test methods and data transfer protocols).
- GOST R draft "GNSS. Road accident emergency response system. In-vehicle emergency call system. Requirements to determine the mechanism of the road

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Overview of pan-European eCall standards.

Description	References	Title
eCall requirements for data transmission	3GPP TS 22.101 ETSI TS 122 101	Third-Generation Partnership Project; technical specification group services and system aspects service aspects; service principles
eCall discriminator table 10.5.135d	3GPP TS 24.008 ETSI TS 124 008	Third-Generation Partnership Project; technical specification group core network and terminals; mobile radio interface layer 3 specification; core network protocols; stage 3
eCall data transfer—general description	3GPP TS 26.267 ETSI TS 126 267	Third-Generation Partnership Project; technical specification group services and system aspects; eCall data transfer; in-band modem solution; general description
eCall data transfer—ANSI-C reference code	3GPP TS 26.268 ETSI TS 126 268	Third-Generation Partnership Project; technical specification group services and system aspects; eCall data transfer; in-band modem solution; ANSI-C reference code
eCall data transfer—conformance testing	3GPP TS 26.269 ETSI TS 126 269	Third-Generation Partnership Project; technical specification group services and system aspects; eCall data transfer; in-band modem solution; conformance testing
eCall data transfer—characterisation report	3GPP TS 26.969 ETSI TS 126 969	Third-Generation Partnership Project; technical specification group services and system aspects; eCall data transfer; in-band modem solution; characterisation report
eCall data transfer—technical report—characterisation report	3GPP TR 26.969 ETSI TR 126 969	Third-Generation Partnership Project; technical specification group services and system aspects; eCall data transfer; in-band modem solution; characterisation report
eCall minimum set of data	CEN EN 15722	Road transport and traffic telematics—eSafety—eCall minimum set of data—draft EN 081018
Pan-European eCall operating requirements	CEN EN 16072	Intelligent transport systems—eSafety—Pan-European eCall—operating requirements
High level application protocols	CEN prEN 16062	Intelligent transport systems—eCall—high level application protocols
Data registry procedures	ISO/EN 24978:2009	Intelligent transport systems—ITS safety and emergency messages using any available wireless media—data registry procedures
Third party support for eCall	CEN EN 16102	Intelligent transport systems—eCall—operating requirements for third party support
eCall extensions for HGV	CEN TR 16405	Intelligent transport systems—eSafety—eCall additional optional data set for heavy goods vehicles eCall

accident, the algorithm determining the events and methods of testing IVS in the definition of the road accident."

- GOST R draft "GNSS. Road accident emergency response system. Test methods for wireless communication module of in-vehicle emergency call system."
- GOST R draft "GNSS. Road accident emergency response system. In-vehicle emergency call system. Compliance testing for the requirements for quality speakerphone in a vehicle."
- GOST R draft "GNSS. Road accident emergency response system. Test methods for navigation module of in-vehicle emergency call system."
- GOST R project "GNSS. Road accident emergency response system. General terms."
- GOST R project "GNSS. Road accident emergency response system. Terms and definitions."

7 OUTLOOK

The pan-European eCall service utilizes technical components (such as satellite positioning, processing, and communication capabilities) that could also provide the basis for several other in-vehicle applications and services, including those required by existing or planned regulations. Such

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

services might affect commercial or private vehicles and can be summarized as

- digital tachograph services
- electronic toll collection services
- transport of dangerous goods and live-animal-oriented services
- any kind of further TPS

that utilizes the basic infrastructure, which has been introduced via the eCall solution.

Streamlining and integration of all these services within a coherent, open-system architecture could yield better efficiency and usability. This approach might reduce costs and could enable the "plug and play" integration of future new or upgraded services.

Such a modular approach could be the basis to address further services in the areas of road safety, personal mobility, logistics support, or it could provide access to multimodal information.

The definition of an "open in-vehicle platform" concept is part of the ITS action plan, and the introduction of eCall based on this concept would positively contribute to its momentum.

The automotive industry, the telecommunications industry, the SPs and all TPSPs will benefit from new services based on the introduction of the eCall telematics platform in all vehicles.

GLOSSARY

CLI	Calling line identification
EC	European Commission
EMS	Emergency management services
ERA GLONASS (Russian)	Ekstrennoje reagirovanije pri avarijakh—the eCall specified for the Russian Market (based on GLONASS)
FSD	Full set of data
GIS	Geographic information system
GLONASS (Russian)	Globalnaja Navigazionnaja Sputnikovaja Sistema—the radio-based satellite navigation system operated for the Russian government
GNSS	Global navigation satellite system
GPRS	General packet radio service
GPS	Global positioning system
GSM	Global system for mobile communications—second-generation network
HGV	Heavy goods vehicle
HMI	Human–machine interaction
IVN	In-vehicle network
IVS	In-vehicle system
LTE	Long-term evolution—fourth-generation network
MNO	Mobile network operator
MSD	Minimum set of data
MSC	Mobile switching center
NAD	Network access device—the communication part of an IVS
PLMN	Public land mobile network
PSAP	Public safety answering point
PTW	Powered two-wheelers
SMS	Short messaging service
SP	Service provider
TPS	Third-party service
TPSP	Third-party service provider
UMTS	Universal mobile telecommunications system—third-generation network
USSD	Unstructured supplementary service data
VIU	Vehicle interface unit

FURTHER READING

CEN TC 278: http://www3.nen.nl/cen278/ (accessed 06 September 2013).

eCall Driving Group (2006) Recommendations of the DG eCall for the introduction of the pan-European eCall.

eCall Toolbox: www.ec.europa.eu/ecall (accessed 06 September 2013).

eSafety: www.ec.europa.eu/esafety (accessed 06 September 2013).

ETSI: http://www.etsi.org/ (accessed 06 September 2013).

European Committee for Standardization (2008) TC 278 WG, road transport and traffic telematics—ESafety—ECall minimum set of data.

European Commission (2009) eCall: Frequently Asked Questions, http://ec.europa.eu/information_society/activities/esafety/doc/ecall/faq.pdf (accessed 18 December 2009).

European Commission (2011) Impact assessment on the introduction of the eCall service in all new type-approved vehicles in Europe, including liability/legal issues. Final Report, Issue 2. November 2011

Filjar, R., Segarra, G., Vanneste, I., et al. (n.d.) Satellite Positioning for eCall: An Assessment of GPS Performance, http://www.heero-pilot.eu/ressource/static/files/filjar_segarra_vanneste_britvic_vidovic_rev_b.pdf (accessed 06 September 2013).

List of updated standards: http://ec.europa.eu/information_society/activities/esafety/doc/ecall/standards/annex_list_status.pdf (accessed 06 September 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto186
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

IT Security for Communicating Automotive Systems

Florian Friederici, Rafael Grote, and Ilja Radusch

Fraunhofer Institute for Open Communication Systems, Berlin, Germany

1 Introduction 1
2 ITS Communications Security 5
3 Security-Related ITS Standards around the Globe 8
Glossary 9
References 9

1 INTRODUCTION

Information technology (IT) has become an integral part of automotive systems. Therefore, these are also affected by IT security threats. Considering IT security is of vital importance in order to ensure correct system operation, to prove that the system was designed according to the state-of-the-art security, to prevent product liability trouble, and to avoid bad press from major security incidents.

Connected vehicles utilize different symmetric communications channels, such as mobile cellular-based communications and Wi-Fi networks as well as unidirectional communications such as FM broadcast. Along those, sensors allow information transmission into the vehicle as well. These include complex sensors such as positioning or even simple sensors such as temperature probes. Sensors and communications both allow transmitting information into applications, and are, therefore, subject to IT security considerations. The applications themselves may typically belong to one of the following categories: informative, assistive, autonomous, emergency, and accounting.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto188
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

This chapter is using terminologies from the following references. The definition of common terms is given in ISO 7498-2 (1989) (International Organization for Standardization), clause 3. Further terms and principles are described in Saltzer and Schroeder (1975). While both references are foundational IT security, this chapter is focused on IT security in the context of automotive telematics. According to ISO/IEC 27000 (2009) (International Electrotechnical Commission), information security is defined as the "preservation of confidentiality, integrity, and availability of information."

Independent of the concrete threats and vulnerabilities a system might face, there are always three primary security objectives. Those are confidentiality, integrity, and authenticity. *Confidentiality* refers to the ability to prevent unauthorized read access to information. That may be broadened to unauthorized information gain by any means, for example, by eavesdropping communications or observe involved devices or parties. *Integrity* refers to the converse ability of unauthorized write access to information, or more general to unauthorized information manipulation by any means. As both, confidentiality and integrity, require authorization to be applied, the last primary security objective is authenticity. *Authenticity* refers to the ability to prevent false identification, which may lead to false authorization.

Those three primary security objectives can be applied to any IT system and to different scopes within the IT system. Possible scopes are communications, data processing, and storage. Abstract roles in IT security consist of *subjects* and *objects*. *Authorization* of subjects requires a proper *identification*. Access or modification of objects is then granted based on this authorization.

Along with the primary security objectives, a couple of secondary objectives may be derived. Those are mainly related to the actions undertaken. Nonrepudiation, plausible deniability, and availability are examples for those.

Nonrepudiation refers to the ability to prove any action undertaken by the subject. *Plausible deniability* is the contrary, when it is not possible to prove that a specific subject conducted the action. As both may not be achieved at the same time, it is required to select the desired behavior of the IT system based on the purpose. An accounting system that is raising bills requires nonrepudiation of the customers, for example. An anonymous advice center, on the other hand, must offer plausible deniability to its clients.

Further examples for secondary security objectives are availability, access control, and permissions. *Availability* refers to the way the subject tries to access the object. That is typically some kind of service. Making the service unavailable is a security threat, because the system operation is halted or disabled. *Access control* and *permissions* allow a fine-grained specification of the actions allowed for a certain subject or group of subjects.

A combined security architecture that offer the former objectives are the *Triple-A* systems used in telecommunications. They utilize protocols and procedures to ensure *authentication*, *authorization*, and *accounting* (*AAA*), which allow authenticating the identity of an entity, determining the authorization for actions, and tracking of the resource consumption.

1.1 Methodology

Security engineering is often seen as black art to outsiders. It is of course no magic, but a widely unknown methodology behind. While the methodology is very deterministic and formal overall, some steps require an educated guess, which makes it to appear arbitrary to some extent. The quality of such an assessment is, however, very dependent on good assumptions. It requires some experience and imaginative power to make them.

The *ETSI TVRA* (European Telecommunications Standards Institute—threat, vulnerability, and risk analysis) method, described in TS 102 165-1 (2011) (technical specification), offers a formal way of analyzing the threats, risks, and vulnerabilities of a telecommunications system. It is commonly used in the standardization at ETSI including ETSI TC ITS (European Telecommunications Standards Institute Technical Committee—intelligent transportation systems). The method guides the security engineer through the identification of the assets within the given environment, grouping the objectives into the categories such as confidentiality, integrity, and authenticity and specification of detailed security requirements. On the basis of the knowledge of the assets and their specification, weaknesses are identified and classified into vulnerabilities and threats. The ability of an attacker to mount an attack is estimated based on the time required, the expertise necessary, the opportunity and the availability, and complexity of essentials equipment. The estimation leads to a quantified likelihood and impact of the threats. Along with that, the intensity of an attack is evaluated to determine the risk. From that evaluation, the selection of countermeasures is possible to reduce the likelihood and/or the impact of an attack. Through iteration after countermeasure selection, a comparison between the unprotected functionality to the protected functionality is possible. The TVRA for ITS is available in technical report (ETSI TR 102 893 V1.1.1, 2010).

Other evaluations criteria and requirements are TCSEC (trusted computer system evaluation criteria) (DoD 5200.28-STD, 1985), ITSEC (information technology security evaluation criteria) (ITSEC, 1991), and Common Criteria for Information Security Evaluation (ISO/IEC 15408-1, 2009; ISO/IEC 15408-2, 2008; ISO/IEC 15408-3, 2008). The *Common Criteria* are general IT security criteria and supersede the older standards TCSEC from the United States and ITSEC from Europe. While the ETSI TVRA method is focused on telecommunications, Common Criteria are applicable for general IT systems. *ETSI TVRA*, however, also refers to parts of the Common Criteria.

Security requirements for cryptographic modules are specified in *FIPS 140-2* (Federal Information Processing Standards) (FIPS PUB 140-2, 2001) for the use by US federal organizations. However, private and commercial organizations are free to adopt the specification of course. Software and/or hardware offering the functionality of a cryptographic module can be certified according to this standard.

1.2 Threats and countermeasures

A large, distributed computer system introduces a huge degree of complexity. That is what IT-equipped connected vehicles are. Such a system consists of stationary and mobile nodes, connected through various types of communication links. Dependent on the viewpoint, different scopes are possible with several emphases.

The *system perspective* is a bird's eye view on the overall system behavior and functionality. While looking from far, it is easy to understand the relationship of all participants and the way they are interacting. In addition, the bigger impact of security flaws can be imagined here. Main asset on system level is a well-done architecture with proper policies.

To better analyze potential vulnerabilities, it is useful to zoom in and have a look at the *single-node perspective*. This is focused on the functionality and interfaces of a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto188
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

single entity within the network. Besides the outside world interfaces, the secure function of the node itself is very important.

An inside look into the node allows to take the *application perspective*. Applications represent the functional logic embedded into a node, which is providing the essential use. Flaws on application level are often not by design, but by poor implementation.

Finally, the *communications perspective* highlights the information interchanges between the nodes in the network. In this perspective, nodes are considered as interconnected black boxes.

Weaknesses may be found in any of those perspectives. Separating them helps to keep the high complexity manageable. Even though, an exploited weakness in the application level may have an impact on the system level.

The single-node and application perspectives for IT-equipped vehicles are very similar to every IT system. Security here is mostly about to tamper resistant devices and to secure application design to the required extent. Nevertheless, even a single node can be consisting of several devices and sensors, which may be seen as a distributed system itself. Again, a separation of viewpoints is then possible and useful.

For this chapter, a look at the internode communications and overall system perspective is most important for two reasons. First, an attacker can easily access the air interface even over large distances, in contrast to attacking a single node, which requires physical access. Moreover, security within a single node is not specific to ITS and is widely addressed in other domains.

ITS, in general, consist of a huge number of nodes from different suppliers and applications from different vendors. The standardized secure operation of them together in the overall system of ITS is one of the biggest challenges in the telematics area.

Possible attacks on communication systems can be categorized into four groups, which are described as follows. The groups intentionally reflect the three primary security objectives that are confidentiality, integrity, and authenticity and availability from the introduction. In addition, they show prominent attacks from a communications perspective to any of them. All of the attacks mounted on communications level affect the system functionality. Pure attacks on the system perspective mainly target the policies and procedures and are not to be discussed in this chapter.

The first group is *bogus message injection and message alteration*. This refers to any attempt in manipulating information interchanged by either forging valid-looking messages with false information included or altering genuine messages in transit to convey false information. This can be done in a couple of ways, including manipulating a genuine node. This case, however, is out of scope of this chapter. In summary, the first group is all about attacking the information integrity.

The second group is *eavesdropping* the communications. This refers to unauthorized information retrieval by passively listening to the wireless communications. For example, most information exchanged in ITS communications is unencrypted by intention. The system is designed to spread information from participating nodes by broadcasting to the adjacent neighbors and by utilizing multihop communications to a certain target area. Additionally, information compilation can be done by central infrastructure elements. It is, therefore, difficult to distinguish between authorized receivers and malicious listeners. The risk at this point is that an attacker is capable of reassembling privacy infringing information from the publicly available made data and, therefore, this group is about confidentiality issues.

The third group is *identity falsification*. As explained in the introduction, it is very important to identify properly the entities in communications to assign authorization. If an attacker is able to fake his or her identity or to appear as multiple entities, he or she is able to undermine the systems policies. The impersonation of multiple entities is known as *Sybil attack*. This group is about authenticity.

The fourth and last group in this classification is *resource and network jamming*. Even though the information is neither changed nor accessed by the attacker, taking the communications out of order breaks the systems functionality. This group refers to availability.

Countermeasures in terms of security engineering are provisions taken to lower the likelihood or impact of a successful attack. In general, they can only mitigate the risks. The efforts undertaken to protect may be exceeded by the efforts undertaken to break the protection. Therefore, the selection of countermeasures is usually based not only on the technical possibility but also on a cost-benefit assessment.

Some objectives can be a priori protected, for example, the eavesdropping case. A break of integrity may only be a posteriori detected. In the latter case, a proper reaction is necessary. Such reactions range from simple “drop message” to advanced cooperative actions, involving other network participants. The exact behavior must be defined in a set of policies.

Furthermore, the system design has an impact on the vulnerabilities against certain attacks. Especially, resource and network jamming can be countered from the first by tolerant resource strategy and frequency agility.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto188
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

1.3 Applied cryptography

Implementing security measures for computers in communications is not about building walls or fences, but utilizing math to grant certain behaviors. That includes, for example, one-way functions, statistics, and probability.

The probably most famous use case for math in security is *encryption*. Even nonprofessionals do usually have an understanding of the target for encryption. Generally speaking, encryption allows transforming a given plaintext into a ciphertext using a secret key. The knowledge of the secret key is required to recreate the plaintext from the ciphertext, which means that the ciphertext is not understandable without the secret key. This technique allows an a priori protection of confidentiality and there exist many variations to encrypt data.

Attaching checksums usually does integrity protection in communications. A secure variant of a checksum is a cryptographic *signature*. Signatures offer a dual functionality; they identify the signer and validate the information. The identification, however, is dependent on additional information on trust and relationships and, therefore, not the key element of the signature itself, the integrity protection is.

Creating signatures requires the ability to make a short summary of the plaintext to be signed. Using *computational secure hash functions* does this. A hash functions simply takes an input of arbitrary length and generates a digest of a fixed length out of it. The digest is basically a number of, for example, 256 bit length. Any changes in the plaintext directly lead to changes in the digest. A good (computational secure) hash function will create digests that cannot be easily predicted and change a lot, even if only a single bit in the plaintext was changed. In this way, it is very hard to create a plaintext that matches to a given digest. As the plaintext can be a lot longer than the fixed size of a digest, collisions are still possible, but very hard to exploit for meaningful purposes.

One-way functions are foundational for doing such calculations. Those are basically mathematic functions that can be calculated efficiently in one direction, but cannot be solved easily in reverse. Examples for that are discrete exponentiation and logarithm.

Using encryption and hash functions together enables the creation of a cryptographic signature. The signer first calculates the digest of the to-be-signed plaintext. Secondly, he or she is encrypting the digest. The output of this operation is the signature, typically with the same length like the digest. A receiver can validate the signature using the same hash function on the plaintext and comparing it to the encrypted digest.

To prevent everybody that knows the secret key to generate valid signatures, a split key methodology is required. The so-called *public key cryptography* enables this. In public key cryptography, there is not a single secret key, but a key pair. One part of the key pair is private and has to be kept secret, which is the private key. The second part of the key pair is public and can be safely shared. Public key cryptography can be used for the same operations described earlier. It is possible to encrypt plaintext using the public key. The resulting ciphertext, however, can only be decrypted using the private key. This way, everybody can encrypt a plaintext and only the owner of the private key is able to decrypt the ciphertext afterwards. This procedure is widely used, for example in e-mail communications.

Using public key cryptography together with computational secure hash functions enables signatures that can only be created by the owner of the private key and can be validated by everybody who knows the public key.

The first introduced principle using a single secure key is called *symmetric cryptography*. The second introduced principle using public and private key pairs is called *asymmetric cryptography*. Asymmetric cryptography cannot replace symmetric cryptography, because both principles have different strength and weaknesses. While symmetric cryptography requires shared secret keys, implementations are faster and more secure compared to asymmetric cryptography. On the other hand, asymmetric cryptography requires longer keys and more calculations to offer the same level of security. Symmetric cryptography is, therefore, best to encrypt huge amount of plaintext. Asymmetric cryptography is better suited for signing plaintext. Both can be combined to a hybrid where the asymmetric cryptography is used for symmetric key exchange. This combination offers the best of both worlds.

1.4 Security protocols and standards

Implementations of security software and hardware that require interoperability between different vendors are based on well-established protocols and standards. Because of the nature of security functionality, it is wise to follow approved concepts to avoid rookie mistakes.

Today's most used symmetric cryptography algorithm is the *advanced encryption standard* (*AES*) specified in FIPS PUB 197 (2001). The most used asymmetric cryptography algorithms are the older *RSA algorithm*, named after Ron Rivest, Adi Shamir, and Leonard Adleman, specified in PKCS#1 (public key cryptography standards) (RSA Laboratories PKCS #1 v2.1, 2002) and the newer *elliptic curve cryptography* (*ECC*).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto188
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

From a user perspective, both RSA and ECC can be used for the same purposes. They differ in the underlying math. RSA is utilizing the difficulty of factoring large integers; in contrast, ECC is based on the algebraic structure of elliptic curves over finite fields.

A suite of algorithms that can be used for digital signatures is specified in the digital signature standard (DSS) (FIPS PUB 186-3, 2009). It includes the principles on how to create digital signatures as well as procedures to create digital signatures using the RSA algorithm and ECC. The latter is commonly known as *elliptic curve digital signature algorithm* (*ECDSA*).

Besides using hybrid asymmetric and symmetric cryptography together to exchange keys, there are also dedicated key exchange protocols for symmetric cryptography. The Diffie–Hellman key exchange protocol (IETF RFC 2631, 1999) can be used to exchange keys between communication partners to negotiate symmetric keys. It is, however, prone to man-in-the-middle attacks.

A commonly used format for certificates in the Internet is specified in X.509 (IETF RFC 5280, 2008) along with the according public key infrastructure and certificate revocation list (CRL) profile.

Revocation of certificates can be done by simple CRL or according to the online certificate status protocol (OCSP) (IETF RFC 2560, 1999). Using OCSP not only gives negative statements on revoked certificates but also allows faster revocation and additionally positive statements about the validity of certificates.

Authentication over a nonsecure network can be done using the Kerberos protocol (IETF RFC 4120, 2005). It allows the authentication of a client to a service by tickets issued from an authentication server. There are protocols for combined AAA. They are the remote authentication dial-in user service (RADIUS) specified in IETF RFC 2058 (1997) and the newer alternative diameter specified in IETF RFC 3588 (2003).

Security measures in IT are founded mainly by security policies and protocols. While policies are just a matter of having a good rule set and enforcing them, protocols have to be implemented in devices, either in hardware or in software. For protocols utilizing strong cryptography and lot of connections, the use of processing accelerators can be useful.

The underlying algorithms can be usually implemented very efficient in hardware. The key material used requires to be kept secret. Dedicated hardware can be built in a tamper evident or to some extend tamper-proof way. Key material that has to be carried around to be used as authentication token can also be stored inside a physical token. This allows easy use and better protection, because it is harder to counterfeit them.

2 ITS COMMUNICATIONS SECURITY

The *ITS communications system* consists of a novel combination of technologies involved. This leads to a slightly different set of requirements on the overall system in general, as well as the security system in particular. First, the main communications link in ITS is the 5.9 GHz wireless channel. It is used in *ad hoc* mode, that is, there is no base station required for operation. The information transmission is mainly based on broadcasting messages to all audience in the vicinity. For simplification reasons, in this chapter, a separation into message types is only made into three classes. More detailed communication patterns are classified in Schoch *et al*. (2008).

Regularly sent *beacons* contain information on the condition of the sending vehicle. Beacons are unidirectional, single-hop broadcast messages. Event-triggered *notifications* contain information on the particular event. Event notifications are unidirectional, multihop broadcast messages. They are usually restricted to a geographic area. *Unicast messages* are used for communication with backend infrastructure services or in specific V2X (vehicle to anything) applications. Unicast messages may be unidirectional or bidirectional as well as single or multihopped depending on the application or backend service.

Because of the *ad hoc* nature of the main communications link, any infrastructure-based security protocols and principles are not suitable for ITS. Even though some vehicles may offer additional cellular-based communications, the basic functionality must be available offline.

The computing devices used in automotive environment have to follow special operation requirements. That concerns the operating temperature ranges, shock resistance, energy efficiency, and budget. Such requirements lead to limited processing power, compared to desktop computers or entertainment devices. Performing calculation-intensive security processing on those devices is not possible. Therefore, it is very likely to see specialized cryptographic coprocessors to offload the calculations. As a side benefit, those coprocessors may offer a secure domain for key material.

2.1 Integrity through digital signatures

As described in the introduction, bogus message injection and message manipulation are dangers to ITS communications. An attacker might try to manipulate or fake messages with the purpose of influencing the behavior of other drivers or vehicles to his or her advantage. As many ITS use cases rely on incoming beacons and notifications, this system is particularly vulnerable to bogus message attacks.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto188
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Example attack: faked hard-brake warning lead to safety reaction. Neighboring vehicles are warned about hard brakes in hazardous situations through V2X event notifications. This leads usually to a safety reaction in following vehicles, which might be a displayed warning message or sound or, in future, maybe even an autonomous driving maneuver controlled by the vehicle itself. Through injecting faked hard-brake warnings, an attacker might exploit this use case in order to influence drivers/vehicles to stop or to trigger an emergency brake. This attack might even provoke an accident.

As a countermeasure, the integrity (and authenticity, see Section 2.2) of messages must be assured. Therefore, every ITS message must be cryptographically signed by its sender. For this purpose, an ITS node needs to own a digital certificate and a private key, which is used to calculate a signature over each outgoing message. The resulting signature is attached to the message together with the node's public certificate. This enables receiving nodes to identify the sender of incoming messages and cryptographically verifying the integrity of the data.

2.2 Authenticity through trusted certificates

Signatures ensure data integrity and identify the originator of a message, but have no implication on the validity of the originator itself. In other words, nothing prevents an attacker from generating his or her own key pair and certificate for sending signed messages, which might in fact appear perfectly valid at the first glance.

Example attack: emergency vehicle spoofing. Emergency vehicles prove themselves to be prioritized by sending specific beacons. During operations—when the light bar and sirens are activated—their special rights are indicated with the beacons, so that the neighboring vehicles are warned and may possibly bear right. An attacker might exploit that fact to stop other vehicles and have a free road on his or her own. Therefore, the attacker generates his or her own certificates to sign beacons, which claim to be from a privileged emergency vehicle. Additionally, the beacons might indicate light bar or siren usage.

Distinguishing between valid and invalid originators requires a trust anchor to legitimate certificates. Thus, certificates are issued and signed by a centralized—and trusted—certificate authority (CA). Checking authenticity means verifying the signature, not only of a message itself, but also of the originator's certificate. Illegitimate senders do not own a trusted certificate. Consequently, they can neither create valid signatures nor inject any bogus message to ITS communications.

2.3 Confidentiality through data encryption

Even though most ITS communications consist of broadcast messages, there are also use cases that require unicast messaging. This comprises, for example, communication to backend systems or specific V2V (vehicle to vehicle) use cases, where two individual vehicles communicate with each other. In these cases, messages are dedicated to a specific recipient. By eavesdropping communications, an attacker might gain confidential data.

Example attack: steal confidential data from V2I (vehicle to infrastructure) communication. The attacker records the communication between a vehicle and the infrastructure service that provides security credentials. This communication conveys personal data, which the attacker could misuse, for example, for linking identities described in the later sections.

Provided the sender knows the recipient's public encryption key, confidentiality could easily be provided through *message encryption*. Fortunately, this is most certainly the case, as certificates are usually disseminated to neighboring nodes, for example, attached to beacons. This is required for message verification on receiving nodes, but the same certificate is conversely also suitable for encryption.

However, the vast majority of messages—including regularly sent beacons and event-based sent notifications—are not dedicated to a single recipient, and remain, therefore, unencrypted. Data encryption is consequently an exclusive solution for only a few use cases.

2.4 Privacy through pseudonymity

Confidentiality of broadcast communications can, in principle, not base on secrecy, as the distributed information is intended for public consumption. The ITS system design deliberately allows tracking of neighboring vehicles and collecting information on their condition. Many use cases actually require this behavior for proper operation. At the same time, it threatens gravely the privacy of drivers and passengers in ITS-enabled vehicles.

ITS messages may include addressing information, operational conditions, and positioning data. Addressing information is required to distinguish sender and addressees. The receiver information is not required in a broadcast scenario; the sender information allows identifying the sender. Together with the included positioning data, a very accurate usage profile may be recorded by only listening to the beacons sent. The information included in ITS broadcast messages might be used to derive personal behavior and should, therefore, be considered as personal data. Consequently, confidentiality in ITS systems is primarily a matter of privacy and less of secrecy.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto188
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Example attack: illegal data mining. Marketing analysts commonly tend to gather all available pieces of information about customers. A local storeowner could, for example, be interested in data of visiting as well as potential (i.e., passing) clients. Recording ITS messages of nearby vehicles would supply him or her with a plethora of facts about his or her clients, including time and duration of visit (or pass-by) as well as detailed information on the vehicle and the client's style of driving (e.g., speed, acceleration, deceleration, steering angle, emergency stops, and other hazardous situations). Allocation between vehicles and individual clients could, for example, take place at the barrier of the store's parking site.

Anonymity is generally accepted as the most effective instrument to protect privacy. As a naive approach, one might try to separate the vehicle identity completely from the driver's identity. Indeed, it would be impossible to hide the ownership of a vehicle (or in general, the allocation to a driver) as long as it is moved in the public. Hence, a more sophisticated approach obscures the identity of the vehicle itself.

To prevent attackers from gaining information through eavesdropping, a pseudonymity concept was proposed in Dötzer (2005) and Golle, Greene, and Staddon (2004). This concept allows the required short-duration tracking, but prevents long-term tracking of vehicles. Addressing information and any static identifier is replaced by exchangeable pseudonyms. After switching a pseudonym, a vehicle is not recognizable for a listening third party any longer. Summarized, pseudonymity permits to track a vehicle during a specific traffic situation (i.e., within lifetime of a pseudonym), but not when it comes across the next day (i.e., exceeding lifetime of a pseudonym). As a result, it is not possible to map data that was collected over a long period to individual vehicles or even persons.

For the idea of changing pseudonyms, it is essential to change effectively all identifiers that are exposed including each communication layer. A single static identifier, for example, a static Wi-Fi MAC (media access control) address, would reveal the vehicle's identity. This applies naturally on identifiers on the security layer, too. In particular, the digital signer certificate is explicitly related to a single vehicle and is perfectly suitable to identify it unambiguously. Consequently, also digital certificates must be replaced on each pseudonym change.

2.5 Two-stage credential system

At this point, the security targets of authenticity and privacy are conflicting to each other. On the one hand, a message receiver wants to validate the sending vehicle to be a legitimate originator; on the other hand, the same vehicle should stay anonymous or at least pseudonymous to the receiver. In order to solve this discrepancy, a two-staged credential system has been introduced in PRE-DRIVE C2X (2009). Its rationale is splitting the single vehicle certificate up into one durable long-term certificate (enrolment credentials) and multiple short-living certificates (authorization tickets).

The *enrolment credentials* are never used in V2X communications. They generally stay secret to message receivers. Hence, pseudonym changes do not comprise replacement of the long-term certificate, which gives vehicles a consistent private identity. Enrolment credentials enable vehicles to authorize against backend systems with the purpose of equipment with new authorization tickets. To protect their confidentiality, long-term certificates are exclusively used for interaction with trusted authorities over encrypted communication links.

Authorization tickets are used in regular ITS communications for generating message signatures and encryption. They may not be resolved to any enrolment credentials or to each other. Being exposed to public V2X communications, short-term certificates must be switched along with the pseudonyms. In order to enhance privacy, it is discouraged to reuse a short-term certificate. Consequently, depending on the pseudonym change interval, a huge number of authorization tickets are required during the lifetime of a vehicle. Even though, a single certificate may consume only a few bytes of storage, it is not possible to hold enough of them for years, especially if a secure storage is used. Therefore, authorization tickets are renewed frequently, which results in the requirement of a credential enrolment infrastructure.

2.6 Certificate deployment through public key infrastructures

Resulting from the two-stage credential system, a backend infrastructure for issuing, deployment, and management of certificates is required. This infrastructure consists of several independent services. In the first place, a root *CA* builds the foundation of trust. Through a chain of trust, every valid certificate of subsidiary CAs and ITS nodes is trusted directly—or indirectly—by the root CA.

Additional CAs fulfill the task of issuing and signing enrolment credentials (enrolment authority) as well as authorization tickets (authorization authority). While issuing and deployment of enrolment credentials is a one-time operation and might be part of a vehicle's manufacturing process, the *enrolment authority* is also responsible for administration of certificates and issuing CRLs regularly. Enrolment credential revocations are

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto188
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

only relevant to the authorization authority, as enrolment credentials are secret to other ITS nodes.

The *authorization authority* issues authorization tickets. As ITS nodes frequently renew these, the deployment works over an online connection, which can be established via multihop routing or by any other communications link, such as cellular or Wi-Fi. For the process of requesting a new authorization ticket, the requesting ITS node first generates a new key pair, wraps the public key into an request envelope, signs it using its enrolment credentials, and sends the request via an encrypted channel. The authorization authority checks the validity and replies with a signed certificate. Authorization tickets may be revoked implicitly through revocation of the enrolment credentials, which prevents the affected node from requesting new authorization tickets. Alternatively, explicit revocation of authorization tickets is possible through deployment of CRLs directly to the nodes.

2.7 Remaining security risks and summary

In summary, ITS communications security comprises signing and verifying messages to ensure integrity, encryption to provide confidentiality on unicast communications, usage of frequently changing pseudonyms instead of static identifiers to protect privacy in broadcast communications, resulting from that a two-stage credential system that separates long-term enrolment credentials from short-term authorization tickets, and finally a public key backend infrastructure for credential management and enrolment of authorization tickets. Nonetheless, there are remaining security risks, which are described below.

An attacker is able to link two pseudonyms to each other by analyzing beacons shortly before and after a pseudonym change. This privacy protection vulnerability results from the fact that most data in beacons of a single vehicle varies only slightly within a short time interval. Hence, it is possible to recognize a vehicle, for example, by its position, which changes only marginally between beacons. This attack depends on the vehicle density (i.e., more vehicles in area increase privacy) and the interval between two captured beacons, that is, the attacker must "see" the pseudonym change. The risk is consequently limited to attackers that follow a vehicle or observe a large area. In both cases, tracking of vehicles does not require ITS beacons; however, it could also rely on different technologies, for example, cameras with number-plate recognition or classic observation by humans.

Another remaining risk regarding privacy consists in fixed vehicle properties, for example, vehicle width and length or manufacturer (Gerlach and Guttler, 2007). A combination of a large number of fixed properties has the same impact as a static identifier. The more the fixed properties are used, the greater is the chance for the resulting identifier to become unique within a set of vehicles. For that reason, broadcasting static vehicle properties with beacons should be avoided whenever possible.

3 SECURITY-RELATED ITS STANDARDS AROUND THE GLOBE

The principles described so far are valid all around the world. Nevertheless, systems deployed in Asia, Europe, and the United States differ in requirements and functionality. Therefore, also the security systems are slightly different.

So far, the complete US ITS stack is described in the IEEE 1609 (Institute of Electrical and Electronics Engineers) set, with IEEE 1609.2 (2006) being the security part. IEEE 1609.2-2006 was the first security standard for ITS communications published and had a major impact on the development in Europe as well. It was designed around the message set and applications used in the United States, called *wireless access in vehicular environments* (*WAVE*), and introduced a very useful set of provisions for secure ITS communications.

Main features are the format description for secure messages and certificates. Secure messages offer an envelope for signed and encrypted application message content. This and the WAVE certificate format were specially designed for the requirements in ITS communications and allow a low overhead on the communications channel compared to existing certificate and message formats.

Since the first trial-use standard from 2006, a revised version is under development, which will integrate the experience made in the first place and also requirements and results from the European development.

The organizations ETSI and CEN (Comité Européen de Normalisation) provide the European standards for ITS communications. The ITS security-related standards are assigned to ETSIs TC ITS. The WG5 (working group) is dedicated to security standardization. The process of standardization is still going on; nonetheless, first results have been already published. The foundation was made in the TVRA (ETSI TR 102 893 V1.1.1, 2010); the basic security architecture is defined in ETSI TS 102 731 V1.1.1 (2010).

Generally, the European security standards are derived from the American IEEE 1609.2. It differs particularly in the following aspects. The applications and message set to be secured fit to the European specifications. The European ITS approach features a dedicated networking protocol

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto188
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

that has to be considered for security measures as well. The two-stage credential system for privacy preserving ITS communication was considered from the beginning. Contrary to the single IEEE standard for ITS security, ETSI TC ITS WG5 is developing separate documents for the single security aspects TVRA, architecture, formats, and service access points.

The situation in Japan slightly differs from the rest of the world. There are already infrastructure-based communication systems in place that can be used for similar purposes similarly to the discussed systems in Europe and the United States. ITS communications in the same way similarly to the European and the US systems, but in a different frequency band, are specified in ARIB STD-T109 (2012) by the Association of Radio Industries and Businesses (ARIB). Security methods are, however, not specified in this standard.

GLOSSARY

AAA	Authentication, authorization, and accounting (Triple A)
AES	Advanced encryption standard
ARIB	Association of Radio Industries and Businesses
CA	Certificate authority
CEN	Comité Européen de Normalisation
CRL	Certificate revocation list
DSS	Digital signature standard
ECC	Elliptic curve cryptography
ECDSA	Elliptic curve digital signature algorithm
ETSI	European Telecommunications Standards Institute
FIPS	Federal Information Processing Standards
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
IT	Information technology
ITSEC	Information technology security evaluation criteria
ITS	Intelligent transportation systems
MAC	Media access control
OCSP	Online certificate status protocol
PKCS	Public key cryptography standards
RSA	Algorithm named after Ron Rivest, Adi Shamir, and Leonard Adleman
TC	Technical committee
TCSEC	Trusted computer system evaluation criteria
TR	Technical report
TS	Technical specification
TVRA	Threat, vulnerability, and risk analysis
V2I	Vehicle to infrastructure (communication)
V2V	Vehicle to vehicle (communication)
V2X	Vehicle to anything (communication)
WAVE	Wireless access in vehicular environments
WG	Working group

REFERENCES

ARIB STD-T109 (2012) 700 MHz band intelligent transport systems (English translation).

DoD 5200.28-STD (1985) Department of defense trusted computer system evaluation criteria.

Dötzer, F. (2005) *Privacy issues in vehicular ad hoc networks*. Workshop on Privacy Enhancing Technologies, Cavtat, Croatia, May 2005.

ETSI TR 102 893 V1.1.1 (2010) Intelligent transport systems (ITS); security; threat, vulnerability and risk analysis (TVRA).

ETSI TS 102 165-1 (2011) Telecommunications and Internet converged services and protocols for advanced networking (TISPAN); methods and protocols; part 1: method and proforma for threat, risk, vulnerability analysis.

ETSI TS 102 731 V1.1.1 (2010) Intelligent transport systems (ITS); security; security services and architecture.

FIPS PUB 140-2 (2001) Security requirements for cryptographic modules.

FIPS PUB 186-3 (2009) Digital signature standard (DSS).

FIPS PUB 197 (2001) Advanced encryption standard (AES).

Gerlach, M., and Guttler, F. (2007) Privacy in VANETs using changing pseudonyms—ideal and real. *Vehicular Technology Conference, VTC2007-Spring*. IEEE 65th, pp. 2521–2525, April 22–25 2007.

Golle, P., Greene, D., and Staddon, J. (2004) Detecting and correcting malicious data in vanets. *Proceedings of the First ACM Workshop on vehicular ad hoc Networks*.

IEEE 1609.2-2006 (2006) IEEE trial-use standard for wireless access in vehicular environments—security services for applications and management messages.

IETF RFC 2058 (1997) Remote authentication dial in user service (RADIUS).

IETF RFC 2560 (1999) X.509 Internet public key infrastructure online certificate status protocol—OCSP.

IETF RFC 2631 (1999) Diffie–Hellman key agreement method.

IETF RFC 3588 (2003) Diameter base protocol.

IETF RFC 4120 (2005) The Kerberos network authentication service (V5).

IETF RFC 5280 (2008) Internet X.509 public key infrastructure certificate and certificate revocation list (CRL) profile.

ISO 7498-2:1989 (1989) (E) Information processing systems—open systems interconnection—basic reference model—part 2: security architecture.

ISO/IEC 15408-1:2009 (2009) Information technology—security techniques—evaluation criteria for IT security—part 1: introduction and general model.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto188
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ISO/IEC 15408-2:2008 (2008) Information technology—security techniques—evaluation criteria for IT security—part 2: security functional components.

ISO/IEC 15408-3:2008 (2008) Information technology—security techniques—evaluation criteria for IT security—part 3: security assurance components.

ISO/IEC 27000:2009 (2009) Information technology—security techniques—information security management systems—overview and vocabulary.

ITSEC (1991) Harmonised criteria of France—Germany—the Netherlands—the United Kingdom.

PRE-DRIVE C2X (2009) Deliverable D1.3 security architecture.

RSA Laboratories PKCS #1 v2.1 (2002) RSA cryptography standard.

Saltzer, J.H. and Schroeder, M.D. (1975) The protection of information in computer systems.

Schoch, E., Kargl, F., Leinmüller, T., and Weber, M. (2008) Communication Patterns in VANETs. *IEEE Communications Magazine*, **46** (11), 119–125.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto188
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Driver Distraction

Klaus Bengler

Technische Universität München, Garching bei München, Germany

1 Introduction 1
2 Requirements and Solutions 2
3 Measurement and Evaluation Methods 3
4 The Need for Calibration and the Question of Validity 6
5 Relevant Process Standards 7
6 Summary 7
Glossary 7
References 7

1 INTRODUCTION

The basic human drive to be mobile is reflected in more than 125 years of automotive history. At first glance, car driving seems to be a complex but clearly defined activity. A closer look, however, reveals that driving a car is in fact a multitask activity where additional tasks, besides driving, are negotiated by the driver. Historically, drivers had to balance the operation of a vehicle and engine management along with the primary driving tasks of speed control and lane keeping. On the basis of this, the primary driving task is divided into three subtasks: stabilization, maneuvering, and navigating.

Operating a vehicle is a highly complex psychomotor activity that is accomplished by multitasking. In addition, the driver has to react to most activities within certain time frames, otherwise the vehicle would be at risk to venture off the roadway or to collide with other vehicles. It has always been of interest to investigate the mechanisms that allow the driver to solve the complex task of driving, along with the subtasks of stabilization, maneuvering, and navigation. In driving, much information is gathered via different sensory modalities and then processed continuously while leading the car (stabilization). This necessarily means that continuous input, via steering wheel and pedals, is needed in normal, nonassisted manual driving. Current research claims that most or even all multitasking phenomena can be explained by quick and successful sequencing and task switching. This would indicate that no parallel processing occurs, but rather very efficient task switching between highly trained subtasks. Therefore, any system integrated into the vehicle must not disturb driving and lead to distraction.

Following user needs and technological advances, an increasing number of additional functions and features are integrated into modern vehicles that are not directly related to driving (Freymann, 2006). They are often considered "tertiary tasks," which explains their relation to primary driving tasks (longitudinal and lateral controls) and secondary tasks (e.g., signaling, controlling the windowshield wipers, and lighting).

The number of these tertiary tasks increases with the introduction of more in-vehicle information systems (IVIS) in cars. The question regarding their distraction potential and the evaluation of the suitability of a given human–machine interaction (HMI) concept is of increasing importance. Suitability, in this case, refers to the idea that a given interaction concept is controllable, compatible (with the driving task), and efficient, in addition to being easy to use while learning the system (ISO 17287, 2003).

A differentiation should be made between distraction and diversion; whereby, distraction means that the attention of the driver is misled and not allocated toward the driving task in an adequate way. This effect can be caused by distracters

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto189
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

inside or outside the car. Lee in Regan *et al*., 2009 gives an excellent overview of different definitions of distraction, including basic explanations thereof.

Diversion also includes a shift of the driver's attention away from the driving task, induced by the driver and introduced by preparatory activities in order to compensate for the attentional cost of the diversion in advance (i.e., intense visual scanning of the environment before reading a navigation display).

Poor information and interaction design can lead to driver diversion and, ultimately, distracted driving. This could be caused by illegible displays, feedback delays, and further faults of the human–machine interface and interaction concepts.

The term *driver distraction* is primarily identified as coming from in-vehicle devices; however, it can also arise from the environment, in-vehicle applications, or from the driver (e.g., while daydreaming). Following this rationale, distraction has to be reduced and can be moderated in different ways: by suitable traffic system layout, in-vehicle design, and, additionally, driver behavior.

Numerous studies report everyday distractive driver behavior that ranges from eating, drinking, and other "all day" activities to tasks typically associated with distracted driving, such as phoning and the operation of in-vehicle functionality (Stutts *et al*., 2005; Dingus *et al*., 2006).

In general, the terms driver distraction or distracted driving are used manifoldly and describe the situation where a driver does not allocate an adequate amount of resources to manage the driving task. A reasonable question, in this context, is whether all available attentional resources of the driver have to be allocated in order to perform the driving task or the measurable driving performance can tell us whether "enough" resources were allocated to the given driving task. Several observable behavioral metrics can be used to assess online whether a driver is distracted by a certain event or operation in a given situation. Zwahlen *et al*. (1987) and many others outlined how and how much visual resources should be minimally allocated.

Another approach to investigate driver distraction can be seen when driver distraction is used as a *post hoc* explanation of a critical incident and accident analyses; specifically helping to explain why drivers failed to turn a critical situation into normal driving. A basic question is, therefore, under which conditions is distracted driving the root cause or a related, both leading to risky behavior and, finally, an accident. Here, the publications around the 100 car study give insight that glancing away from the road scene remarkably increases the accident risk (Dingus *et al*., 2006). Unfortunately, accident data seem to identify distraction as a "standard" root cause for accidents, leading to distorted statistics about this topic.

During its development, the automobile continued to become much functional and more complex. In the first half of the twentieth century, this meant the addition of a few mechanical controls and indicators. The following 50 years have been characterized by the fact that the scope of action for the driver is constantly expanding. This allows the driver to initiate multiple functions and processes, and then to let them run in parallel. A prime example of this would be extended comfort systems and related automatic climate control systems for increasing the comfort of the interior. Modern means of communication, keeping up with mobile phones, lead to a significant increase in function. In summary, it is clear that the increase in HMI in the car is characterized by a constant process of function development; existing functions to be automated, additional functionality being integrated. Many studies focus on the increasing amount of IVIS in cars that would lead to a higher distraction potential. At first glance, the increase could lead to more driving situations that include distracted driving and, therefore, an increased risk. A more differentiated analysis shows that additional information, provided that the information presentation is well designed, helps the driver to manage the driving task and thereby reduces the accident risk, even though additional glances are necessary to gather information. A typical example is the usage of navigation information and navigation systems that help drivers, especially those in unknown environments, to focus on maneuvering and stabilization. In this sense, the ongoing integration of functionalities has to increase comfort, entertain, and inform the driver in a way that emphasizes to him the importance of reducing overall driver distraction potential.

The increase of complexity in the vehicle interior can be compensated for by suitable design measures. The design of vehicle cockpits has to be an expression of the current psychological and ergonomic knowledge, and must also avoid driver distraction effects, such that all activities can be occur in harmony with the driving task.

2 REQUIREMENTS AND SOLUTIONS

Following this rationale and focusing on tertiary tasks, the questions are

- what are the guidelines that enable the optimization of HMI for in-vehicle systems and also minimize distraction;
- what are the procedures and metrics to measure the distraction potential of a given function or information presentation while being developed to avoid the *post hoc* experience.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto189
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.1 Guidelines and standards

There are clear guidelines and recommendations, specifically, the European Statement of Principles (ESOP) (Commission of the European Communities, 2006), the AAM-Guidelines (AAM, 2002), the JAMA Guidelines (JAMA, 2004), and the NHTSA Guidelines (NHTSA, 2013), which integrate existing knowledge and focus this issue while recommending suitable means of interaction design or appropriate methodologies and metrics for evaluation.

A comparison of the aforementioned guidelines reveals a great degree of overlap in their basic messages. However, significant differences are also observed in terms of specific criteria values or explicit functionalities. This is primarily due to the fact of the date of their publication and the region and stakeholders they address. A detailed overview is given by Stevens, Burns, and Akamatsu in separate chapters in Regan *et al*. (2009).

2.2 How to minimize distraction

The ergonomic quality of complex products has grown to a remarkable differentiating success factor, also addressing the property of minimized distraction. For this reason, ergonomic knowledge is applied over the whole duration of the development process. Advances in basic ergonomic research within the past decades have been very high in number. In addition, a number of ergonomic criteria, including benchmark values and evaluation methods, have been developed, empirically validated, and even some of them have been taken as in international standards.

As a selection, the above-mentioned ESOP gives recommendations on five main sections:

- installation principles;
- information presentation principles;
- principles on interaction with displays and controls;
- system behavior principles;
- principles on information about the system.

The topic of driver distraction is explicitly handled in the introducing section on Overall design goals: "the system does not distract or visually entertain the driver."

Especially, principles on installation, information presentation, interaction, and system behavior give detailed advice on potential sources for driver distraction and how to avoid them by good design or how to test for them during development [see also Stevens (2009) for more information and the development of the document.].

A slightly different structural approach is taken by the Visual-Manual NHTSA Driver Distraction Guidelines for in-vehicle electronic devices released by the National Highway Traffic Safety Administration in 2013. This document addresses clearly the given functionalities or function groups of IVIS and focuses on existing evaluation methodologies, explicitly defining criteria values (such as the AAM Guidelines). Additionally, the document provides information on relevant studies that its messages are based on.

A similar sectional structure, such as in the ESOP, can be found reading the AAM Guidelines. Although these two documents, and the shorter JAMA Guidelines, differ in their formulation and rigidity of the criteria, they agree on statements regarding the actions that shall be taken in order to minimize distraction. The most relevant formulations could only be quoted here from several ESOP principles, leaving the examples, references, and detailed recommendations, behind. Therefore, the reader should refer to the original document for more detailed information.

As several principles reference driver distraction, it is obvious that distraction may be caused by different reasons, but can also be reduced by different actions during the development process. These principles are also related to different evaluation approaches that will be further explained in the following paragraphs.

3 MEASUREMENT AND EVALUATION METHODS

Given the fact that guidelines exist, the compliance of an IVIS can be tested during development (Burns *et al*., 2010) using different methodological approaches.

Principles and different dimensions have to be considered in order to come to an ergonomic solution. Among others, these are

- geometric layout;
- interaction design;
- cognitive aspects (Table 1).

Table 1. Selected examples of evaluation dimensions for different ergonomic requirements.

Geometric	Interaction	Cognitive
Reachability of devices	Task duration	Complexity of information
Availability of space	Consistency of interaction concept	Learnability of interaction
Visibility of displays	Visual demand of interaction	
Readability of characters	Interrutibility of interaction	
Glare of displays		

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto189
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.1 Expert judgment and checklists

In many cases, human factors judgment is considered a suitable method for the qualification of a given interaction. Stevens (2009) and Green (2009) give an overview of existing checklists that are, as the TRL checklist is, highly related to the specified guideline and recommendation documents. It is important to state that especially in the early development stages, this method delivers important contributions to the reduction of distraction.

3.2 Model-based test of perceptive requirements

Meanwhile, digital human models, such as RAMSIS (Bubb *et al*., 2006), have been extended by perceptive functions that lead to a RAMSIS-kognitiv™. Considering what is known about the processing of visual information, this digital model implements features that allow the engineer to check for visual occlusions, given a specified seating position and car environment. This is possible for the analysis of interior as well as exterior visibilities. Moreover, models for age-specific acuity are integrated and allow for differentiated checks of instrumentation and infotainment display readability (Lorenz and Remlinger, 2011) (Figure 1).

Again, the engineer is able to check at an early stage of development if instruments are occluded by the steering wheel or if mobile devices fixed to the windscreen fit for different anthropometries. Additionally, the engineer can also check whether standards on the readability of letters and characters, formulated in ISO 15008, are met by the given instrumentations under specified reading distances and acuities (Figure 2).

Using RAMSIS as a model, in addition, it is possible to estimate the duration of visual diversion for a given display location. RAMSIS-kognitiv™ implements a visual behavior model that calculates the time needed for a saccade and fixation, based on the eccentricity of the display or information in relation to the driving scenery.



Figure 1. Model-based check of visibility of a secondary display using RAMSIS as a model (Lorenz and Remlinger, 2011). (Reproduced by permission of VDI Wissensforum GmbH.)



Figure 2. Model-based check of instrumentation readability (Lorenz and Remlinger, 2011). (Reproduced by permission of VDI Wissensforum GmbH.)

3.3 The occlusion test

Even if information is in the right place and readable, many distraction effects are related to the fact that the duration needed for the gathering of visual information is too long or that the driver is not able to easily interrupt the interaction (Baumann *et al*., 2004; Foley, 2010).

Poor examples include cluttered displays, too high graphical density or running texts, and banners. There are different methodologies available that assess visual demand and interruptibility, but differ in the experimental effort. It could be mentioned that the above-mentioned properties can be investigated already in early stages of development with the occlusion technique or method of interrupted vision. The principle of this method is that while the subject interacts with a given functionality, visual perception of the display is interrupted using LCD goggles. These goggles toggle between transparent intervals, where vision is permitted, and opaque intervals, where vision is not permitted, typically at a pace of 1500 ms per interval. These experiments include an occlusion session and a nonoccluded session.

The occlusion session simulates the requirement of the driving task saying that the driver should not divert from the driving scenery longer than 2000 ms (Zwahlen *et al*., 1987) and needs a given amount of time to resume the driving scenery after a diversion.

The sum of vision intervalls called *total shutter open time* (TSOT) is a very good estimate of visual demand of a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto189
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 3. Subject in interaction wearing the goggles in an intransparent interval.

display; increasing, for example, not only with the graphical complexity of the display but also with the total task time (TTT).

The ratio between TSOT and TTT in a nonoccluded condition delivers the interruptibility ratio *R*, which is a metric for the resumability of an interaction; increasing *R* values indicate bad interruptibility and can be caused by high degrees of visuomotor coordination in a bad touch concept, moving text, or animations.

Setting, procedure, and equipment are standardized in ISO 16673 (Figure 3).

The application of an occlusion test clearly classifies concepts that need improvement in the area of information presentation and/or interaction.

3.4 Simulated lane change test

Another surrogate method is the so-called simulated lane change test (LCT). This procedure (ISO 26022) is based on a simplified driving situation (Mattes and Hallen, 2009). It can be performed in a driving simulator environment or as a simple desktop setup. Three lanes are displayed and the driver has to change lanes instantly after perceiving a sign at the roadside while operating an IVIS (Figure 4).

The lane keeping performance and lane change behavior are integrated and compared to a normative (ideal) model, which delivers a deviation area (filled area in Figure 5) as a metric for the driving quality and for the distraction caused

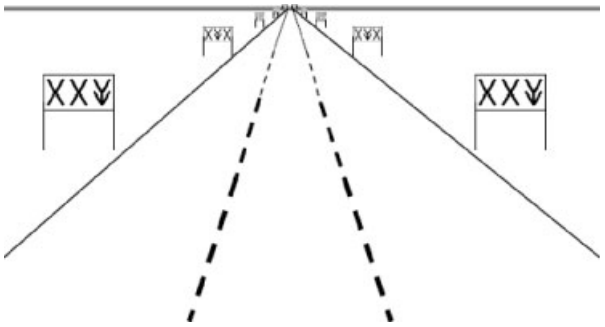


Figure 4. The driving scene. In this example, the driver should change to the right lane. (Reproduced with permission from ISO 26022, 2010. © BSI Group. Permission to reproduce extracts from ISO 26022:2010 is granted by BSI. British Standards can be obtained in PDF or hard copy formats from the BSI online shop: www.bsigroup.com/Shop or by contacting BSI Customer Services for hardcopies only: Tel: +44 (0)20 8996 9001, Email: cservices@bsigroup.com.)

by the device. The metric is called *MDEV (mean deviation)*. Bengler *et al*. (2010) show that the method—provided correctly applied—delivers highly reproducible results.

Like the occlusion test, the LCT provides information on the quality of information presentation and interaction. It is furthermore applicable to tasks that could produce distraction effects not only in the visual area but also by capturing the cognitive resources of the driver, such as the phone or listening do.

3.5 The detection response task

The increase of IVIS based less on visual–manual interaction but rather use voice input–output technologies (speech interfaces, e-books, and text to speech interfaces), question how to evaluate them considering cognitive workload and estimate their effects on driver attention. Merat and Jamson (2008) showed the potential of detection response experiments to be reliable and valid assessments.

The basic idea of this method is to present visual stimuli via LEDs in the periphery, close to the eye or tactile stimuli (i.e., vibration), while the subject operates an IVIS and measure response times and misses to assess the attentional demand of the IVIS task. The tests can be performed with or without an additional driving task. Currently, the method is under standardization to fix values for the equipment, analysis, and instructions, so that results from different assessments can be comparable (van der Horst and Martens, 2010; Engström, 2010; Bengler, Kohlmann, and Lange, 2012; Conti, Dlugosch and Bengler, 2012).

The basis for the detection respose task (DRT) stems from the multiple resource theory of attention, which states

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto189
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

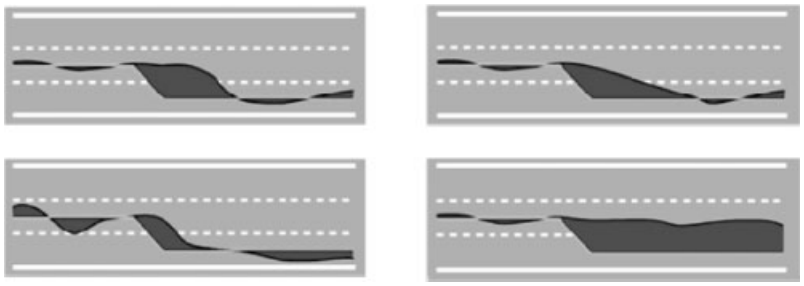


Figure 5. Different types of driver behavior in the lane change test. Late reaction, slow performance, swervy lane keeping, and signal missed. (Reproduced with permission from ISO 26022, 2010. © BSI Group. Permission to reproduce extracts from ISO 26022:2010 is granted by BSI. British Standards can be obtained in PDF or hard copy formats from the BSI online shop: www.bsigroup.com/Shop or by contacting BSI Customer Services for hardcopies only: Tel: +44 (0)20 8996 9001, Email: cservices@bsigroup.com.)

that the operation of a cognitive task should have an effect on the operation of a parallel task in a dual-task setting. The LCT and DRT settings are typical examples of the multiple task situation, realistic enough to mimic the primary driving task and the potential distraction by additional in-vehicle activities. Additionally, the DRT can be used to evaluate the effects of roadside distractors, such as information clutter, by poorly arranged traffic information, or the assessment of attentional effects related to dynamic traffic situations.

3.6 Driving simulator tests and field trials

While field testing was, and is still, an established method to assess driver distraction, the power and sophistication of driving simulation is at a level that makes it a more than appropriate method of investigating attentional behavior under standardized and safe conditions (Reed and Green, 1999). The field of simulators is very heterogeneous, ranging from static one screen to moving base simulators with 360° projection systems.

The projects ADAM (Advanced Driver Attention Metrics), adaptive integrated driver–vehicle interface (AIDE), and human–machine interface and the safety of traffic in Europe (HASTE) as well as CAMP investigated the validity of different driving simulators, which is often questioned and delivered valuable information on the metrics that can be derived from driving simulator experiments to assess distraction effects with validity. Moreover, ADAM specified a standard highway scenario, which is documented in Bengler *et al*. (2003) and incorporated in AAM (2002), and showed that valid IVIS evaluations can be performed using a static, single screen system.

As most powerful metrics proved the standard deviation of lane position (SDLP), this metric is defined as the standard deviation of all recorded n deviations from given reference point (car center) and the left or right lane boundary d, root squared, with d_{avg} as the average of all recorded distances and n the number of measurements and the number of lane exceedances (Knappe *et al*., 2007). Therefore, it is necessary to understand and document the implementation of the vehicle dynamics used for a given test, as it can highly influence these metrics (Knappe *et al*., 2007; Engström *et al*., 2005).

The usage of eye tracking equipment to record the visual behavior of a driver, while driving in the simulator or field, is highly recommended in addition to being supported by powerful analysis systems. ISO 15007/1 and 15007/2 give valuable information on how to conduct visual behavioral studies and how to treat the data.

4 THE NEED FOR CALIBRATION AND THE QUESTION OF VALIDITY

After the established methods and metrics have been presented, it has to be mentioned that they rely, as models (RAMSIS) or experiments, on human behavioral data. Especially the surrogate methods (occlusion, LCT, and DRT) have to prove that they deliver valid measurements for distraction effects that could occur in real traffic.

For the methods and models described earlier, these validation studies have been conducted and showed strong correlations between relevant visual behavior measurements (eyes of road time) or driver performance metrics in simulators and surrogate metrics: resumability metric R (occlusion), MDEV (LCT), and SDLP, and object and event detection in different driving simulators. Also, there are remarkable correlations for TSOT (occlusion) and TGT (driving simulation or field).

In this context, not only is the validity of the measurements relevant but also their reliability. This means that the repetition of an assessment, for example, in another laboratory leads to comparable results (Bengler *et al*., 2010).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto189
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5 RELEVANT PROCESS STANDARDS

On the process level, ISO 17287 provides valuable input how to avoid distraction by more sophisticated specification and helps to document the data derived from experimental studies and evaluations in a structured way. This process standard describes an approach that includes the following steps:

- intended use and usage context;
- nonintended operations (while driving);
- countermeasures to prohibit these nonintended operations;
- measures to avoid foreseeable misuse;
- ways to present system failures to the driver;
- information about experimental studies including the results.

6 SUMMARY

Given that driving is a complex continuous task and that an increasing amount of information is available for the driver inside and outside the vehicle, the assessment of distraction effects is a highly relevant topic. On the other hand, there is a well-established set of evaluation methods addressing different aspects of distraction. Also, process- and interaction-related recommendations are compiled in several documents to give different stakeholders (OEM, supplier, mobile device manufacturer, and civil engineer) guidelines for the reduction of distraction.

This should be a situation that allows the use of potential future IVISs, while simultaneously maintaining traffic safety.

GLOSSARY

In-vehicle HMI	Human–machine interaction performed while driving or in standstill using displays and controls in the interior of a vehicle.
IVIS	In-vehicle information systems (e.g., navigation, audio, route guidance, and traffic information systems).
Human factors	Scientific discipline covering the aspects of human information processing and the optimization of technology to user requirements.
Human–machine interaction	Structured system of input, output, and transformations between, and by, a user and a technical system.
Lane change test	Standardized procedure to test the visual demand of in-vehicle information systems using a standardized driving-like task.
Occlusion test	Standardized procedure to test the visual demand of in-vehicle information systems using the method of interrupted vision.

REFERENCES

AAM (2002) *Principles on Human Machine Interface (HMI) for In-Vehicle Information and Communication Systems (Draft)*, Alliance of Automobile Manufacturers, Detroit.

Baumann, M., Keinath, A., Krems, J.F. and Bengler, K. (2004) Evaluation of in-vehicle HMI using occlusion techniques: experimental results and practical implications. *Applied Ergonomics*, **35**, 197–205.

Bengler, K., Kohlmann, M. and Lange, C. (2012) Assessment of cognitive workload of in-vehicle systems using a visual peripheral and tactile detection task setting. *Work: A Journal of Prevention, Assessment and Rehabilitation*, **41** (Supplement 1/2012), 4919–4923.

Bengler, K., Mattes, S., Hamm, O. and Hensel, M. (2010) Lane change test: preliminary results of a multi-laboratory calibration study in *Performance Metrics for Assessing Driver Distraction: The Quest for Improved Road Safety*. Chapter 14 (ed. contG.L. Rupp), SAE International, Warrendale, PA, pp. 243–253.

Bengler, K., Praxenthaler, M., Theofanou D. and Eckstein, L. (2004) *Proceedings of the Investigation of Visual Demand in Different Driving Simulators within the ADAM Project. Driving Simulator Conference 2004 Europe*, Paris, pp. 91–104.

Bubb, H., Engstler, F., Fitzsche, F., *et al.* (2006) The development of RAMSIS in past and future as an example for the cooperation between industry and university. *International Journal of Human Factors Modeling and Simulation*, **1** (1), 140–157.

Burns, P.C., Bengler, K. and Weir, D.H. (2010) Driver metrics, an overview of user needs and uses in *Performance Metrics for Assessing Driver Distraction: The Quest for Improved Road Safety*, Chapter 1 (ed. G.L. Rupp), SAE International, Warrendale, PA, pp. 24–30.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto189
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Commission of the European Communities (2006) Commission Recommendation of 22 December 2006 on Safe and Efficient In-Vehicle Information and Communication Systems: Update of the European Statement of Principles on human machine interface. Brussels, 22.12.2006.

Conti, A.S., Dlugosch, C., and Bengler, K. (2012) *Detection Response Tasks: How Do Different Settings Compare? Automotive UI Conference Proceedings*, October 2012. Portsmouth, NH, USA.

Dingus, T.A., Klauer, S.G., Neale, V.L., *et al*. (2006) The 100-car naturalistic driving study, Phase II—results of the 100-car field experiment. Technical Report No. DOT HS 810593. National Highway Traffic Safety Administration, Washington, D.C., April 2006.

Engström, J. (2010) The tactile detection task as a method for assessing drivers' cognitive load in *Performance Metrics for Assessing Driver Distraction: The Quest for Improved Road Safety*, Chapter 5 (ed. G.L. Rupp), SAE International, Warrendale, PA, pp. 90–103.

Engström, J., Johansson, E. and Östlund, J. (2005) Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F*, **8**, 97–120.

Foley, J.P. (2010) Now you can see it. Now you don't. Visual occlusion as a surrogate distraction measurement technique in *Performance Metrics for Assessing Driver Distraction: The Quest for Improved Road Safety*, Chapter 9 (ed. G.L. Rupp), SAE International, Warrendale, PA, pp. 123–134.

Freymann R. (2006) HMI: a fascinating and challenging task. IEA—16th World Congress on Ergonomics. Maastricht.

Green, P. (2009) Driver interface safety and usability standards. An overview in *Driver Distraction: Theory, Effects, and Mitigation* (eds M.A. Regan, J.D. Lee and K. Young), Taylor & Francis, London, pp. 445–465.

ISO 15008. *Road vehicles—ergonomic aspects of transport information and control systems—specifications and compliance procedures for in-vehicle visual presentation*.

ISO 16673. *Road vehicles—ergonomic aspects of transport information and control systems—occlusion method to assess visual demand due to the use of invehicle systems*.

ISO 17287 (2003) *Road vehicles—ergonomic aspects of transport information and control systems—procedure for assessing suitability for use while driving*.

ISO 26022 (2010) *Road vehicles—ergonomic aspects of transport information and control systems—simulated lane change test to assess in-vehicle secondary task demand*.

ISO/DTS 14198. *Road vehicles—ergonomic aspects of transport information and control systems—calibration tasks for methods which assess demand due to the use of in-vehicle systems*. Revised Draft Version (2012).

JAMA (Japan Automobile Manufacturers Association) (2004) Guideline for In-vehicle Display Systems.

Keinath, A., Baumann, M., Gelau, C., *et al*. (2000) Occlusion as a Technique For Evaluating In-Car Displays. In Engineering Psychology and Cognitive Ergonomics Vol. V.II. *International Conference on Engineering Psychology and Cognitive Ergonomics—2000* Edinburgh, Scotland.

Knappe, G., Keinath, A. and Bengler, K. (2007) Driving Simulators as an Evaluation tool—Assessment of the Influence of the Field of View and Secondary Tasks on Lane Keeping and Steering Performance. *ESV Conference Lyon*.

Lee, J., Regan, M.A. and Young, K.L. (2009) What drives distraction? Distraction as a breakdown of multilevel control in *Driver Distraction: Theory, Effects, and Mitigation* (eds M.A. Regan, J.D. Lee and K. Young), Taylor & Francis, London, pp. 41–57.

Lorenz, D. and Remlinger, W. (2011) *Sichtauslegung eines kompakten Elektrofahrzeugs mit RAMSIS kognitiv Sicherheitsgewinn durch effiziente Sichtsimulation. 2*, Automobiltechnisches Kolloquium, Garching.

Mattes, S. and Hallen, A. (2009) Surrogate distraction measurement techniques: the lane change test in *Driver Distraction: Theory, Effects, and Mitigation* (eds M.A. Regan, J.D. Lee and K. Young), Taylor & Francis, London, pp. 107–121.

Merat, N. and Jamson, A.H. (2008) The effect of stimulus modality on signal detection: implications for assessing the safety of in-vehicle technology. *Human Factors*, **50** (1), 145–158.

National Highway Traffic Safety Administration—NHTSA (2013) Driver distraction guidelines for in-vehicle electronic devices. Docket No. NHTSA-2010-0053.

Reed, M.P. and Green, P.A. (1999) Comparison of driving performance on-road and in a low-cost simulator using a concurrent telephone dialling task. *Ergonomics* **42** (8), 1015–1037.

Stevens, A. (2009) European approaches to principles, codes, guidelines, and checklists for in-vehicle HMI in *Driver Distraction: Theory, Effects, and Mitigation* (eds M.A. Regan, J.D. Lee and K. Young), Taylor & Francis, London, pp. 90–103.

Stutts, J.C., Feaganes, J., Reinfurt, D. and Rodgman, E. (2005) Driver's exposure to distraction in their natural driving environment. *Accident Analysis and Prevention*, **37**, 1093–1101.

Van der Horst, R.A. and Martens, M. (2010) The peripheral detection task (PDT): on-line measurement of driver cognitive workload and selective attention in *Performance Metrics for Assessing Driver Distraction: The Quest for Improved Road Safety*, Chapter 4 (ed. G.L. Rupp), SAE International, Warrendale, PA, pp. 73–89.

Zwahlen, H.T., Adams, C. and DeBald, D. (1987) *Safety aspects of CRT touch panel controls in automobiles*. Paper Presented at the Second International Conference on Vision in Vehicles, University of Nottingham, England, September 14–17. Also in Gale, A.G., *et al.* (eds) Vision in Vehicles II, North-Holland, Amsterdam, 1988, pp. 335–344.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto189
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Battery Charging Standards

Peter Van den Bossche
Erasmus University College Brussels, Brussels, Belgium
Vrije Universiteit Brussels, Brussels, Belgium

1 Generalities 1
2 The Standardization Landscape 1
3 Energy Needs for Charging 4
4 Charging Modes for Conductive Charging 6
5 Wireless Charging 9
6 Battery Exchange 10
7 EMC Issues for Charging 10
8 Regulatory Issues of Charging 10
9 Communication Issues 11
10 Accessories for Conductive Charging 15
11 Conclusions 18
References 18

1 GENERALITIES

In urban traffic, owing to their beneficial effect on environment, electrically propelled vehicles are an important factor for the improvement of traffic and more particularly for a healthier living environment. One of the main benefits of the electric vehicle is its use of mains electricity as energy source, allowing, on the one hand, to displace finite fossil fuel sources with renewables and, on the other hand, a zero-emission operation at point of use, with reduced or zero indirect emissions at generation level.

On board the vehicle, the electric energy is stored in traction batteries, which are the key critical components of the vehicle. The transfer of the energy from the grid to the vehicle creates the need for charging infrastructures. The transfer can be done conductively, the most widely used system, or wirelessly by induction, eliminating the need for cables and connectors.

To allow a wide deployment of electric vehicles, they should have universal access to charging infrastructures, which highlights the need for drafting comprehensive standards on the matter. This chapter gives an overview of global developments in the field.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 THE STANDARDIZATION LANDSCAPE

2.1 Historical background

The need for the availability of standardized charging infrastructures already arose in the first golden age of the electric vehicle in the early twentieth century. The first ever standard to be developed for electric vehicles concerned in fact the charging plug, a standard sheet for which was presented in 1913 (Van den Bossche, 2003; Electric Vehicle Association of America, 1914), which would be adopted on an international level as British Standard 74 (BS74, 1917).

2.2 Global standardization

Standardization, on a global level, is mainly dealt with by three institutions: the *International Electrotechnical Commission* (IEC), founded in 1904, dealing with all things electrical, the International Organization for Standardization (ISO), founded in 1948, dealing with all other technologies, and the *International Telecommunication Union* (ITU), founded in 1865, deals with telecommunications and related matters. IEC and ISO are nongovernmental organizations, whereas ITU is an agency of the United Nations.

With the standardization of the electric road vehicle becoming a key issue, the question arises is which standardization body would have the main responsibility for electric vehicle standards? This may seem a trivial problem; however, one has to consider that the electric vehicle is a device mixing two technologies: it is, on the one hand, a "road vehicle," but, on the other hand, when it is charging connected to the grid, it becomes an "electrical appliance."

For historical reasons, the approach taken toward standardization in both these realms is fundamentally different. One can speak of radically diverging "standardization cultures."

The electrotechnical industry has a long tradition in standardization (as stated earlier, IEC was founded much earlier than ISO). In the electrical world, many things are standardized, including subjects such as the color code of wires are standardized (e.g., green and yellow for the protective or earth conductor). This standardization is based on both the strong regulatory influence, concerning electrical safety and wiring codes on the one hand, and the role of specialist component manufacturers, acting as suppliers to equipment manufacturers on the other hand.

As for electric traction equipment, there is a culture of component standardization: electric traction motors, for example, are described in elaborate international standards detailing their performances and the way to measure them. This is also due to the fact that electric traction customers are more likely to be powerful corporations (e.g., railway companies) who tend to enforce very strict specifications on the equipment they order or purchase, hence the need for more elaborate component standards to ensure the compliance of the equipment.

In the automotive manufacturing world, things are different. The major manufacturing companies have grown into vertically integrated structures that have less need for international standardization. Although there are strong government regulations regarding vehicle safety and type approval, there is less need perceived for individual component standards. Furthermore, the automobile industry is based on mass production, and routine tests as prescribed, for example, for rail traction motors would be way too expensive to perform for road vehicle motors. The automotive customer is also more likely to be a "consumer" who is less interested in compliance to specific international standards.

Technical committees specifically dealing with electric vehicle and infrastructure standardization were established in the 1970s: IEC TC69 and ISO TC22 SC21. The committees reflected their background, IEC representing electricians and component manufacturers, and ISO mostly representing vehicle manufacturers. Collaboration has not always run smooth, however, and there have been considerable discussions between the two groups as to the division of the work, in which there were a number of overlaps and issues for discussion. By the end of the 1990s, a consensus was agreed defining the specific competences of the respective committees, as shown in Table 1.

Table 1. Basic division of work in IEC/ISO.

ISO	IEC
Work related to the electric vehicle as a whole	Work related to electric components and electric supply infrastructure

The charging interface where the vehicle connects to the grid is the main contact point between the automotive and the electrical realms.

2.3 Regional standardization

2.3.1 Europe

Within Europe, CENELEC (European Committee for Electrotechnical Standardization) and CEN (European Committee for Standardization) operate as the pendants of IEC and ISO. Both have been active in electric vehicle standardization in the 1990s, through their technical committees CENELEC TC69X and CEN TC301. However, much of this work was parallel to the global standardization work, with the European standards created superseded by international standards when these were available.

Both committees went dormant around the turn of the century. TC301 was reactivated as a general CEN technical committee dealing with road vehicles. TC69X was reactivated in 2011 following the activities of the CEN–CENELEC Focus Group on electromobility, with the aim of expediting the European adoption of IEC TC69 documents. As to avoid the duplication of work, TC69X specifically avoids to do actual standardization work and prefers to follow up the work performed globally by IEC. Specific European standards in the field will only be drafted where is a clear need, for example, facing European regulations.

The new activities of the European groups were inspired by European Union initiatives, with the DG Enterprise and Industry of the European commission issuing in June 2010 its Mandate M468 concerning the charging of electric vehicles (Commission of the European Communities, 2010).

Its scope was to develop or review existing standards in order to:

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- ensure interoperability and connectivity between the electricity supply point and the charger of electric vehicles, including the charger of their removable batteries, so that this charger can be connected and be interoperable in all EU States;
- ensure interoperability and connectivity between the charger of electric vehicle–if the charger is not on board–and the electric vehicle and its removable battery, so that a charger can be connected and interoperable and recharge all types of electric vehicles and their batteries;
- appropriately consider any smart–charging issue with respect to the charging of electric vehicles;
- appropriately consider safety risks and electromagnetic compatibility (EMC) of the charger of electric vehicles in the field of LVD (low voltage directive) 2006/95/EC (Commission of the European Communities, 2006) and EMC directive 2004/108/EC (Commission of the European Communities, 2004a).

The mandate was addressed not only to CEN and CENELEC but also to the telecommunications standards body ETSI (European Telecommunications Standards Institute).

To respond to the demands of the mandate, CEN and CENELEC constituted the *Focus Group on European Electromobility—standardization for road vehicles and associated infrastructure*—as a (informal) joint working group, reporting to the CEN and CENELEC Technical Boards. The Focus Group considered, in the first instance, European requirements relating to standardization for road vehicles and associated infrastructure, and assessed ways to address them. It aimed mainly at charging infrastructure, but also covered related subject such as the standardization of batteries (in view of battery exchange systems), as well as EMC issues, where particularly the case of conducted emissions and power quality will have to be considered when introducing electric vehicle charging on a large scale.

The Focus Group is not destined to itself create standards documents nor create regulatory requirements. This remains the competence of the relevant standardization organizations and regulatory authorities.

The Focus Group had a very wide participation, including representatives of the CEN and CENELEC national members—often from local industry or Governments—and of all major European associations of stakeholders in the field. Observers fully participating have included representatives of technical standards committees in CEN, CENELEC, ISO, and IEC, from some other standards organizations, and from the European Commission services.

The final report in response to the commission mandate was presented to CEN and CENELEC technical boards in June 2011 (CEN—CENELEC, 2011), formulating a number of recommendations affecting the development of electric vehicle standardization issues in Europe. It is aimed at the relevant standards committees, also involving other committees that are relevant to the subject such as IEC TC64 (safety of electrical installations), IEC TC13 (metering), IEC SC17D (low voltage switchgear and control gear assemblies), IEC SC23H (plugs and sockets), and IEC TC57 (smart grid), as well as the corresponding European technical committees in the framework of CENELEC.

For the sake of communication, interaction will also be sought with ETSI.

The work of the Focus Group will be pursued by the CEN–CENELEC Electromobility Coordination Group, which first met in March 2011 and which has the aim to support coordination of standardization activities during the critical phase of writing new standards or updating existing standards on electromobility, and make recommendations accordingly.

2.3.2 *USA*

In the United States, although global IEC and ISO standards are followed and collaborated to, there is a strong tradition of sectoral standardization. The Society of Automotive Engineers (SAE) thus also performs in the field and publishes standards and recommended practices concerning various aspects of electric vehicle technology, including charging infrastructure.

2.3.3 *Japan*

In Japan, electric vehicle standardization is coordinated by the Japan Automobile Research Institute (JARI) that performs activities on national level as well as acting as national committee for both IEC and ISO.

2.3.4 *China*

In China, electric vehicle standardization is conducted by the Electric Vehicle Subcommittee SC27 of the Technical Committee TC114 of the Standardization Administration of the People's Republic of China (SAC).

It was established in 1988 with its secretariat located at the China Automotive Technology and Research Centre (CATARC) in Tianjin City, China. It is supervised by both the Technical Committee TC114 of the SAC and the National Technical Committee of Auto Standardization (NTCAS) under the Ministry of Industry and Information Technology (MIIT). The electric vehicle standardization in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

China has an interface with international standards ISO TC22 SC21 and IEC TC69, and the World Forum for Harmonization of Vehicle Regulations, which is a working party (WP 29) of the United Nations Economic Commission for Europe (UNECE). It is tasked with creating a uniform set of regulations for vehicle design to facilitate international trade. The forum works on regulations covering vehicle safety, environmental protection, energy efficiency, and theft resistance.

Until the writing of this chapter, there were over 50 of Chinese EV standards issued, 12 EV standards are pending for approval, 14 new EV standards are under development, and 45 EV standards are under preliminary research. The issued standards discussed earlier include electric vehicle standards, key components standards, and infrastructure standards.

In China, there are two types of standards. First is National Standard (GB, Guobiao) supervised by the SAC. Second is Industrial Standard (QC, quality control) supervised by the MIIT.

2.4 Regulatory aspects

One should always consider the fundamental difference between standards and regulations: while standards are technical documents issued by independent bodies and in principle voluntary, regulations come out of governmental sources and are mandatory. There are, however, common points as in the European "New Approach" directive where reference is made to harmonized standards, the adherence to which presumes conformity with the regulations.

For vehicle type approval reference should be made of the UNECE regulations that contain specific requirements for the electrical safety of electric vehicles (UNECE–R100, 2011). On the other hand, electrical installations have to conform to electric wiring regulations, which may differ strongly between different countries. Electrical safety is a vital requirement and protection against electric shocks remains the key driver of electric equipment standardization.

3 ENERGY NEEDS FOR CHARGING

3.1 Energy consumption

Although the actual energy consumption of electric road vehicles is strongly dependent on the vehicle's mission and usage profile such as the number of stops, a good approximation of the grid consumption of a battery-electric vehicle with current technology in mixed city traffic can be given by the empirical formula (Van den Bossche, 1992):

$$E_s = 80 + \frac{80}{m} \tag{1}$$

where

- E_s is the specific energy consumption in $\frac{\text{Wh}}{\text{T•km}}$
- m is the mass of the vehicle in tonnes.

A typical medium-sized vehicle, for example, weighing 1500 kg would thus have an energy consumption of

$$E = 1.5 \times \left(80 + \frac{80}{1.5}\right) = 200\frac{\text{Wh}}{\text{km}} \tag{2}$$

To drive this vehicle over a distance of 50 km, a typical urban range for battery-electrics or plug-in hybrids, the following amount of energy would be needed from the grid:

$$E = 50\ \text{km} \times 200\frac{\text{Wh}}{\text{km}} = 10,000\ \text{Wh} = 10\ \text{kWh} \tag{3}$$

The time needed for charging will depend on the power available and on the rating of the charger.

In this context, the notion of *charging speed* could be introduced, corresponding to the distance that can be covered by the amount of electrical energy charged during 1 h.

3.2 Charging power levels

Several power levels can be defined according to the power taken from the grid and the associated charging speed possible.

3.2.1 "Standard" charging

Standard or *Normal* charging uses a power level typical for a standard residential power outlet. This corresponds to the concept *level 1* charging as defined in the United States (Toepfer, 1994).

This level 1 charging only offers a limited performance, because of the low 120 V supply voltage in North America. A standard 15 A outlet will yield up to 1.8 kW, giving a "charge speed" of 9 km/h.

European countries use the 230 V voltage rating, with current up to 16 A (United Kingdom and Switzerland have 13 and 10 A, respectively), giving a "charge speed" of around 15 km/h. The power offered by a residential outlet is fully adequate for overnight charging (typically most of the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

charging of both private and commercial electric vehicles will be done overnight); for opportunity charging, however, the power is rather limited.

3.2.2 "Semifast" charging

Semifast or *accelerated* charging uses higher current levels beyond those of a standard domestic outlet, but which are readily available in a typical residential or commercial installation. This corresponds to *level 2* charging in the United States.

In this context, one could use either a higher current single-phase connection (e.g., 32 A, which corresponds to the rating of a typical electric kitchen range) or a three-phase connection. Three-phase system, the most common system, is a four-wire distribution 3×400 V, with a phase voltage of 230 V and a line voltage of 400 V. In some countries such as Norway, Italy, and parts of Belgium, three-wire 3×230 V networks with a phase voltage of 230 V, without neutral wire, can also be found.

This three-phase connection allows considerably higher power for a modest current, with just 16 A per phase, one can get:

$$P = \sqrt{3} \times 400 \times 16 \times \cos\varphi = 11.1 \text{ kW} \quad (\cos\varphi = 1) \quad (4)$$

This allows charging, for example, the vehicle in just within 1 h for a range of 50 km, or a "charge speed" of 55 km/h.

With a current of 32 A, put to 22 kW are available. This allows considerably fast charging for most small- and medium-sized vehicles (for heavy vehicles such as buses, this power level will correspond to a "normal" charge/dots), using current levels that can be implemented more easily than the high single-phase current of nearly 100 A, which would be required to deliver this power.

The use of three-phase connection has the further advantage to be more beneficial for the load spreading of the electric network.

The on-board battery charger on the vehicle shall of course be configured to accept a three-phase connection.

An overview of the available power for normal and semifast charging is given in Table 2 (Van den Bossche, 2010a). The last two columns of this table give the section of copper needed for the cable (based on the standard ratings of 16 A for 2.5 mm^2 wire and 32 A for 6mm^2 wire, and not including neutral or earth conductors), and the relationship between power and copper section. It can clearly be seen that the use of three phases, allows a much better utilization of the conductors, and hence the use of lighter cables and accessories.

Table 2. Power levels for charging ($\cos\varphi = 1$).

	Voltage	Phases	Current	Power (kW)	Copper (mm^2)	kW/mm^2
Standard	230	1	16	3.7	5	0.74
Semifast	230	1	32	7.4	12	0.62
Semifast	230	3	16	6.4	7.5	0.85
Semifast	400	3	16	11.1	7.5	1.48
Semifast	400	3	32	22.2	18	1.23

3.3 "Fast" charging

For "fast" charging (called *level 3* charging in the United States), higher power levels are used, which create the need for specific infrastructure beyond standard domestic or industrial socket outlets. The charging can be either performed with a DC or an AC connection between the vehicle and the charging post.

In the DC case, a fixed battery charger (rectifier) is connected to the battery, and more heavy and expensive fixed infrastructure is thus needed, whereas, for AC fast charging, the rectifying is done on board the vehicle, most commonly using the traction inverter that is able to recharge the battery at a high current (for regenerative braking) and can also be fed by the grid (Figure 1).

Fast charging infrastructure is being proposed for power levels up to 250 kW (Szczepanek and Botsford, 2009), claiming to be able to charge an electric vehicle in < 10 min, comparable with the refueling time of a gasoline-powered legacy vehicle.

The high flexibility and apparent user-friendliness of fast charging comes, however, with a number of drawbacks:

- the high cost of the fixed infrastructure involved compared with semifast charging;
- the need for heavy cables, which for practical purposes are usually fixed to the charging station (case "C" in

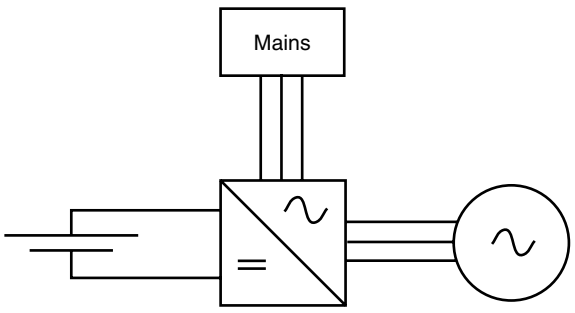


Figure 1. Charging with inverter.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

IEC61851), exposing them to copper theft, which is a growing problem in most countries;
- the high burden on the distribution network with high level point loads, and the high cost of electricity at peak times in particular. This can be alleviated by providing an energy buffer such as a stationary battery at the charging station (which would be charged overnight at low rates); however, such solution increases the investment cost even more while introducing additional losses because of battery and charger efficiencies;
- the fast charging cycle provides only opportunity charging and does not allow final charge, which takes a certain time at a low current, periodical (overnight) charging will thus remain necessary;
- the high strain on the battery through fast charging may have an influence on battery cycle life and hence cost.

Fast charging will thus be more expensive for the user and most charging will make use of normal or semifast charging points located both at private locations (residential garages or corporate parking places) and at publicly accessible charging points.

The presence and availability of fast charging stations, however, provide a psychological advantage to electric vehicle drivers, allowing them to fully exploit the range envelope of their vehicle and to overcome "range anxiety." A few well-located fast charging stations in an urban area can fulfill this need. They will be accessed mostly for "emergency" uses or in case of an unexpected change in mission, with the bulk of the electric energy delivered to the vehicles by cheap overnight charging.

The psychological advantage of the fast charging stations is comparable in fact to the plug-in hybrid vehicle, where the presence of an auxiliary power unit gives users the confidence to use the full range of their battery, overcoming the range anxiety typical for many battery-electric drivers who do not have specialist knowledge of the behavior of their battery (Botsford and Szczepanek, 2009).

The high power connection of the "fast" charging station makes it furthermore particularly interesting for "vehicle-to-grid" applications.

4 CHARGING MODES FOR CONDUCTIVE CHARGING

While Section 3 dealt with the power levels to be used, one should also consider the way the infrastructure is used for connecting the vehicle. This leads to the definition of the so-called *charging modes*, introduced in the international standard IEC61851-1 (IEC61851-1, 2010).

4.1 Mode 1 charging

Mode 1 charging refers to the connection of the EV to the AC supply network (mains) utilizing standardized socket outlets (i.e., meeting the requirements of any national or international standard), with currents up to 16 A. This corresponds to nondedicated infrastructures, such as domestic socket outlets, to which electric vehicles are connected for charging.

These socket outlets can easily and cheaply deliver the desired power, and owing to their availability, mode 1 charging is the most common option for charging electric vehicles, particularly when existing infrastructures are to be used.

There are, however, a number of safety concerns to be considered. The safe operation of a mode 1 charging point depends on the presence of suitable protections on the supply side: a fuse or circuit breaker to protect against overcurrent, a proper earthing connection, and a residual current device (RCD) switching off the supply if a leakage current greater than a certain value (typically 30 mA) is detected.

In most countries, RCDs are now prescribed for all new electric installations. There are, however, still a lot of older installations without RCD, and it is often difficult for the EV user to know, when plugging in the vehicle, whether or not an RCD is present. While some countries leave this responsibility to the user, mode 1 has, therefore, been outlawed in a number of countries such as the United States.

Some countries such as Italy do not allow mode 1 charging for charging places accessible to the public and limit its use to private premises, out of concern that live standard socket outlets in public places may be exposed to the elements, vandalism, or unauthorized access.

In countries where the use of mode 1 charging is allowed, its simplicity and low investment cost make it a preferred mode for private premises (including residential garages and corporate parking lots). Mode 1 allows charging in full safety when used in a reliable electrical installation.

However, the uncertainty faced by the user about the presence of an RCD when plugging in the electric vehicle in an arbitrary standard outlet makes that a potential hazard may be present. For this reason, vehicle manufacturers steer away from mode 1 charging, preferring mode 3, with mode 2 as a transitory solution.

Mode 1 charging will, in the future, thus mostly be relevant as the main mode for small vehicles such as scooters and quadricycles (CEN–CENELEC, 2011).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.2 Mode 2 charging

Mode 2 charging connection of the EV to the AC supply network (mains) also makes use of standardized socket outlets. It provides, however, additional protection by adding an *in-cable control box* (ICCB) with a control pilot (CP) function (Section 4.3.2) between the EV and the control box.

The introduction of mode 2 charging was initially mainly aimed at the United States, as a transitional solution in the first edition of the IEC61851-1 standard. It is now, however, receiving new interest to replace mode 1 for charging at nondedicated outlets and is used by most car manufacturers.

The main disadvantage of mode 2 is that the control box protects the downstream cable and the vehicle, but not the plug itself, whereas the plug is one of the components more liable to be damaged in use. Furthermore, the use of the ICCB is not elegant and not always very practical, for example, in public environments.

The CEN–CENELEC Focus Group recommended that occasional charging on private premises should preferably be done using mode 2 to ensure RCD protection (CEN–CENELEC, 2011).

4.3 Mode 3 charging

4.3.1 Definition

Mode 3 charging involves the direct connection of the EV to the AC supply network utilizing dedicated electric vehicle supply equipment (EVSE). This refers to private or public charging stations. The standard IEC61851-1 (IEC61851-1, 2010) mandates CP protection between equipment permanently connected to the AC supply network and the electric vehicle.

4.3.2 Control pilot

For mode 3 charging, the IEC61851-1 standard foresees additional protection measures to be provided by the so-called *CP*, a device that has the following functions mandated by the standard:

- verification that the vehicle is properly connected;
- continuous verification of the protective earth conductor integrity;
- energization and de-energization of the system;
- selection of the charging rate (ampacity).

This function is typically performed through an extra conductor in the charging cable assembly, in addition to the phase(s), neutral, and earth conductor. Annex A of IEC61851-1 specifies the CP circuit given in Figure 2, showing the operation of the system. A control signal (1 kHz PWM, pulse-width modulation) is sent through the CP conductor. The switch on the vehicle allows to control the charging, whereas the duty cycle of the PWM signal controls the current absorbed by the charger, thus allowing dynamic ampacity control (Section 9).

When no vehicle is connected to the socket outlet, the socket is dead. This provides a key safety advantage particularly for publicly accessible charging points. Power is delivered only when the plug is correctly inserted and the earth circuit is proved to be sound.

The new version of IEC61851-1, now under development, further refines the use of the CP for the interaction

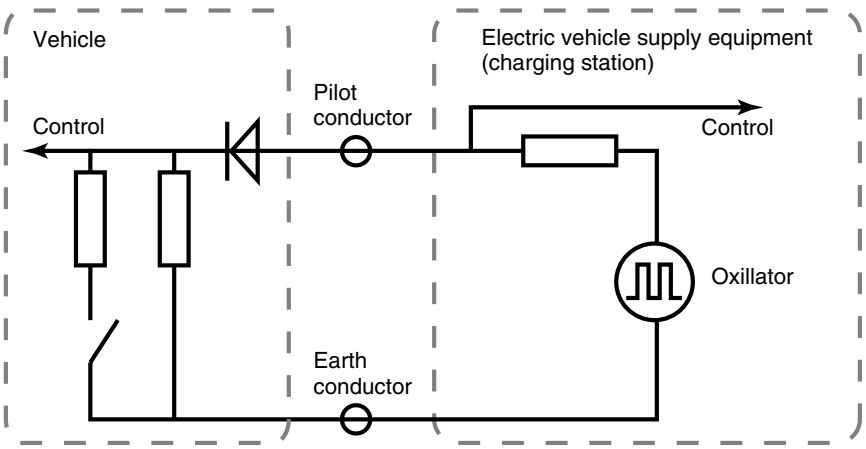


Figure 2. Control pilot conductor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

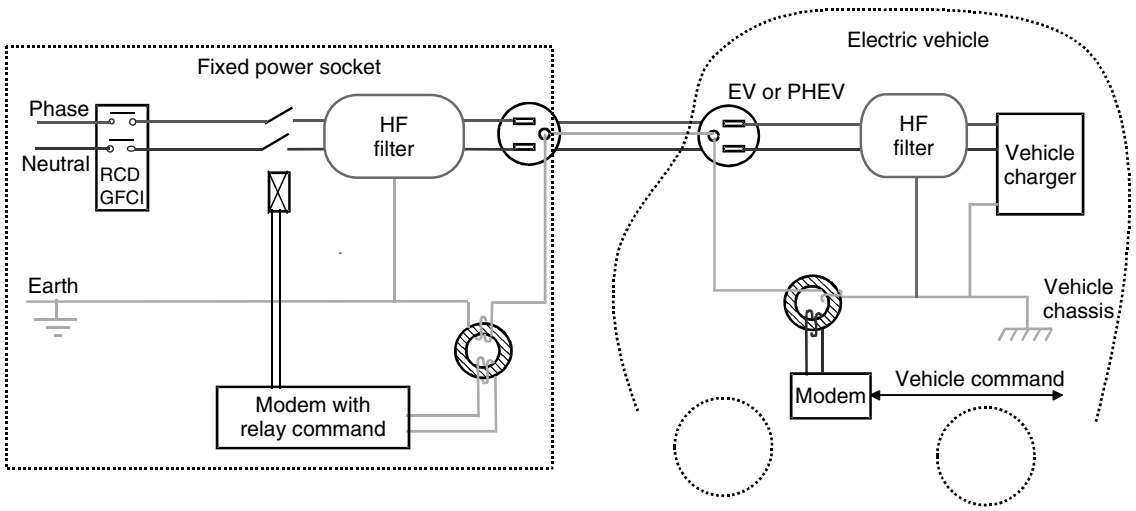


Figure 3. Control pilot function with power line communication.

between the electric vehicle and the EVSE, defining several states of operation:

Table 3. Charging states.

A	Vehicle not connected
B	Vehicle connected, not ready for charging
C	Vehicle charging, ventilation not required
D	Vehicle charging, ventilation required
E and F	Vehicle connected, EVSE not available for charging

The use of a CP conductor in charging equipment was first proposed around 1990 for charging stations for electric boats deployed in the Norfolk Broads in England (Benning Ltd., 1990).

The use of a CP function with fourth wire is also included in the SAE standard J1772 (SAE J1772, 2009).

4.3.3 Control pilot conductor alternatives

The use of a dedicated conductor for the CP necessitates an extra conductor and thus the use of special cables and accessories. The functionality of the CP can, however, also be achieved in alternative ways; the standard 61851-1 (IEC61851-1, 2010) defines the functional requirements described in Section 9 of the *CP function*, mandatory for mode 3 charging, which can be achieved by a pilot conductor or by other means. These include various wireless data transfer systems and power line communication. An interesting implementation of the latter has been developed by Electricité de France (Bleijs, 2009). The principle as described in an informative annex to 61851-1 is illustrated in Figure 3.

In this implementation, the CP signal is a common-mode signal between the phase wires and the earth conductor, using a 110 kHz carrier frequency. This signal is generated by the vehicle electronics and transmitted to the earth wire through a transformer (ferrite torus). Filter circuits are present as to avoid the unwanted transmission of data signals from the charging system to the mains, and to be compliant with the relevant standards concerning electrical equipment to be connected to the grid, which proscribe any earth-line communication upstream of the EVSE (EN50065-1, 2002).

The system is able to perform all CP functionalities over a three-wire connection. This basic protection can be implemented using a cheap and light set of electrical components and would thus also be suited for light vehicles such as electric two wheelers, where the extra cost coming with the use of special accessories should be avoided.

The proposed system presents, however, several other interesting opportunities, as it is able not only to carry the CP signal but also to perform data exchange functions to be used in smart charging and billing (Section 9).

Unfortunately, this system has not been adopted yet by vehicle manufacturers.

4.3.4 Implementation of mode 3

The inherent safety features, as well as the potential for smart grid integration, make mode 3 a preferred solution. It was thus recommended by the Focus Group for public charging stations as well as for home charging using dedicated outlet (CEN–CENELEC, 2011).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Most electric vehicles appearing on the market nowadays are mode 3-ready.

4.4 Mode 4 charging

Mode 4 charging is defined as the indirect connection of the EV to the AC supply network (mains) utilizing an off-board charger, where the CP conductor extends to equipment permanently connected to the AC supply (IEC61851-1, 2010).

This pertains to DC charging stations, which are mostly used for fast charging. As the charger is located off board, a communication link is necessary to allow the charger to be informed about the type and the state of charge of the battery to provide it with the right voltage and current.

5 WIRELESS CHARGING

5.1 Introduction

All charging systems described earlier rely on a conductive coupling and thus have the need for cables, plugs, and sockets. The use of wireless systems has been proposed to allow an improvement of electric safety on the one hand and an enhanced user friendliness on the other hand, removing the need of handling cables and all the associated hazards.

Most wireless charging systems are based on inductive power transfer, using a two-part transformer with the primary connected to the network and the secondary installed on the vehicle. Charging can be performed after juxtaposition of the two parts.

One type of inductive charging has been introduced and extensively promoted by General Motors in the 1990s, making use of a paddle-shaped primary inductor coil to be inserted in a slot on board the vehicle. This lightweight, high frequency system could be used over a wide power range, but it still necessitated the use of a cable and had a much higher cost than conductive solutions. Further development of/and interest in this system was halted by the end of the twentieth century. International standardization work had already been started, but did not proceed beyond CD (committee draft) level (IEC61980-1/CD, 2000; IEC61980-2/CD, 2000).

5.2 Automatic connection

The real advantage of wireless systems is to do away with cables. Cable-free charging can be made possible by appropriate design and location of the inductors: it is sufficient to park the vehicle adjacent to the primary inductor, which can be done in one of the following ways (Figure 4):

1. special charging post design;
2. fixed devices, flush with road surface;
3. fixed devices, not flush with road surface;
4. devices on road surface, which include moving parts;
5. special kerb designs.

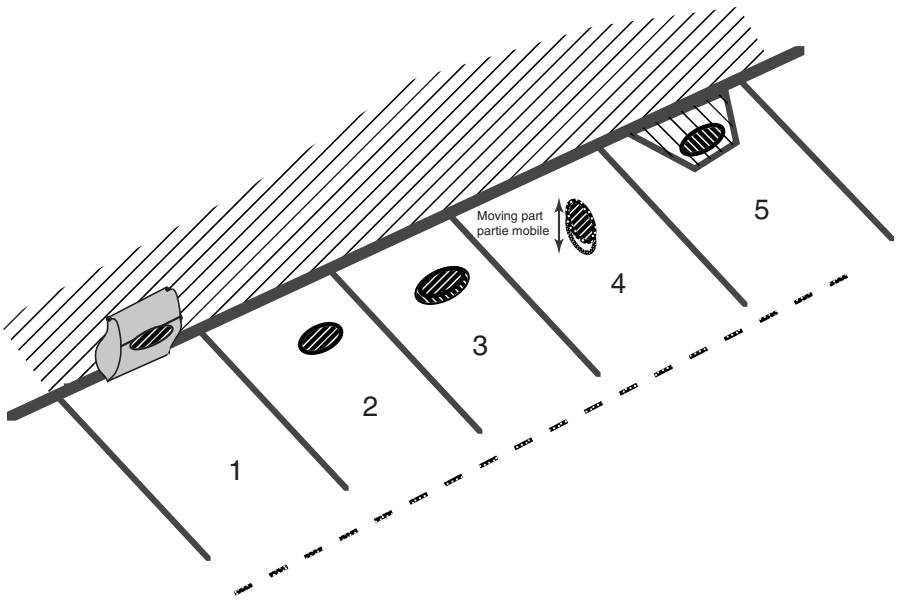


Figure 4. Automatic charger connections.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

As such, systems require a specific approach to both the vehicle and the charging point architecture. Their use will be mostly limited to niche applications such as captive fleets for special use and automatic rent-a-car systems. This should also be seen in the framework of the inductive energy transfer systems now being implemented for powering trams and electric buses without overhead cables.

5.3 New developments for standardization

Wireless charging systems have recently known a renewed interest and the standardization work on the subject has been revived. The project team on 61980 started again in the spring of 2011 and will at first focus on the general requirements of the inductive charging systems, particularly highlighting the safety aspects involved with exposure to magnetic fields in the vicinity of the inductive coils. Other wireless transfer systems, for example, based on capacitive energy transfer, will also be considered.

6 BATTERY EXCHANGE

A fast replenishment of the energy on board of the vehicle can be performed by replacing the whole battery pack with a freshly charged one. This technology, which has been used in the past for niche applications such as industrial electric vehicles, has now gained new interest for general use. Its implementation, however, will entail specific standardization problems.

On the one hand, exchangeable batteries will need a standardization of pack dimensions and interfaces. Battery technology, particularly in the field of lithium batteries, is rapidly evolving, and standardization of battery module sizes can only be reasonably achieved when the technology is sufficiently mature. First steps are now, however, being taken to standardize battery cell sizes, supported by both the electrical community (through IEC TC21) and the automotive community (through ISO TC22 SC21).

On the other hand, the exchange of batteries between vehicles and the deployment of "second-life" vehicle batteries for other applications such as stationary energy storage for grid support necessitate means to accurately estimate the *state of health* of a battery. The state of health of a battery reflects the general conditions of a battery and its ability to store electric energy and deliver specified performance, compared with the rated performances of a new battery. Determination of the state of health is not straightforward, as it is highly dependent on the understanding of a battery's chemistry and environment and the evolution of aging processes. The CEN–CENELEC Focus Group recommends that parameters for state of health should be defined in standards to allow for second-life use of batteries (CEN–CENELEC, 2011).

7 EMC ISSUES FOR CHARGING

The influence of the extended use of power electronic converters as used in battery chargers will have to be closely followed up in order to avoid potential problems regarding electromagnetic interference either in the form of radiated electromagnetic waves or as conducted interference on the interconnecting cables. EMC is defined as the ability of an equipment or a system to function satisfactorily in its electromagnetic environment without introducing intolerable electromagnetic disturbances to anything in that environment.

EMC is subject to heavy regulation; in the European Union, there is, on the one hand, the EMC directive 2004/108/EC (Commission of the European Communities, 2004a) that pertains to electric and electronic equipment and, on the other hand, the vehicle EMC directive 2004/104/EC (Commission of the European Communities, 2004b) that pertains to road vehicles.

Furthermore, there are numerous international standards published by IEC, ISO, and CISPR (International Special Committee on Radio Interference) dealing with the matter.

The current set of standards covers most of the EMC needs for low frequency phenomena, except in the frequency range between 2 and 150 kHz, a frequency band that contains the typical operating frequencies of power electronic converters as used in electric vehicle traction systems and battery chargers. It will thus be necessary that this frequency range is addressed by standardization as soon as possible.

The low frequency range below 2 kHz is mostly relevant for conductive emissions (power quality and harmonics), for which limits are given by international standards such as IEC61000-3-2 (IEC61000-3-2, 2005) and 61000-3-12 (IEC61000-3-12, 2011).

The specific EMC problems related with electric vehicle charging are being treated in the new edition of the IEC61851-21 standard, work on which has started in 2011.

8 REGULATORY ISSUES OF CHARGING

The difference in national electric regulations and wiring codes constitutes a major challenge for the specification of a single-charging system usable in various countries.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Some nations, for example, require shutters on socket outlets in domestic environments, whereas others do not.

A typical discussion point is also to define the extension of the scope of the European Union LVD (Commission of the European Communities, 2006). The vehicle itself is explicitly excluded from this scope, but the charging equipment and the cable are clearly covered. The vehicle inlet (where the connector is coupled) is still a matter of discussion, however.

Standardization work is now being performed by IEC TC64 on particular requirements for low voltage installations intended to supply energy for electric vehicles (IEC60364-7-722/CD, 2012).

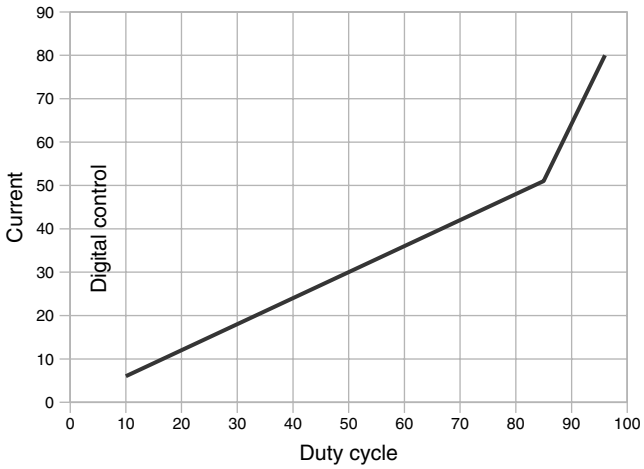


Figure 5. PWM signal. (Taken from data in IEC61851-1 (2010).)

9 COMMUNICATION ISSUES

9.1 Generalities

The vehicle needs to establish a communication with the charging post, which can be developed in several ways, with increasing sophistication.

In mode 1 or mode 2 charging, where standard nondedicated socket outlets are used, there is no communication at all.

9.2 Control pilot PWM communication

Mode 3 introduces communication through the CP function. The principle is defined in IEC61851.

In its most basic way of operation (as developed in the first edition of 61851), the CP only fulfills its essential safety function. The pilot signal is just a small current sent through the CP loop to ensure that the vehicle is properly connected and the earth connection is sound.

More functionality can be added using a PWM signal in the CP circuit. The PWM signal can convey information about the maximum ampacity of the charging point through the variation of its duty cycle.

This feature presents several operational benefits:

- The charger can adjust itself to the maximum allowable current that can be delivered by various charging points, for example, a standard value of 16 A, reduced value of 10 A, and higher value of 32 A for semifast charging points.
- The charging point can control the amount of current absorbed by the charger, in the framework of a smart grid load management or to optimize the tarification of the electric energy.

The relationship is illustrated in Figure 5. A duty cycle $< 5\%$ conveys the message that charging current is controlled through advanced serial communication; for duty cycles $> 97\%$, charging is impeded.

9.3 Control pilot baseband communication

As to allow further functionalities to be included in the CP signal, the new version of IEC61851-1 currently under development also proposes to use the CP for baseband serial communication, which is a bidirectional, half-duplex, point-to-point serial communication link on the CP. It provides an extended CP function that is backwards compatible with either vehicles of charging stations fitted with the standard PWM CP (albeit without the extra communication options of course). The actual protocol used for the communication is covered in other standards; using the CAN (controller area network) protocol (ISO11898-1, 2003) is an option.

9.4 Communication for DC charging (mode 4)

Off-board chargers, which supply a direct current to the vehicle battery, must communicate with the vehicle battery in order to supply it with the correct voltage and current. This is particularly the case with nondedicated chargers as used in public charging stations, which should be able to supply vehicles with varying battery voltages and chemistries.

Early development in this field was described in the CENELEC Technical Specification 50457-2 (CENELEC TS50457-2, 2006), with a protocol largely based on the ISO road vehicle diagnostic standards as defined in ISO14229

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and ISO14230 (ISO14230-1, 1999; ISO14230-2, 1999; ISO14230-3, 1999; ISO14229-1, 2006). These concern requirements for diagnostic systems are implemented on a serial data link layer, which allows a tester to control diagnostic functions in and on vehicle electronic control unit.

The protocols were specifically adapted for the selected application: after the initialization phase by the off-board charger, the vehicle's charge control unit controls the charging process of the off-board charger. Contrary to the standard communication according to ISO14230 where the server and the client are fixed during all the session, their roles are definitively reversed after the initialization phase.

This document is now, however, to be considered obsolete and will be superseded by the new standards now under development.

The new developments in the field of DC charging have led to the drafting of IEC61851-24 on a global level. The work has started in 2010 and is now at CD stage (IEC61851-24/CD, 2011).

This standard will define the messages of digital/data communication to be used during charging control between off-board DC charging system and electric road vehicle. It states the basic specifications of the communication circuit, which is based on "CAN" principles. The document focuses on the upper two layers (6 and 7) of the OSI (open system interconnection) concept (ISO/IEC7498-1, 1994) (Figure 6); the lower two layers being described in the CAN standard ISO11898 (ISO11898-1, 2003; ISO11898-2, 2003).

Annexes to IEC61851-24 will describe several practical protocols reflecting the solutions described in the annexes IEC61851-23 (IEC61851-23/CD, 2011), among which the "CHAdeMO" concept.

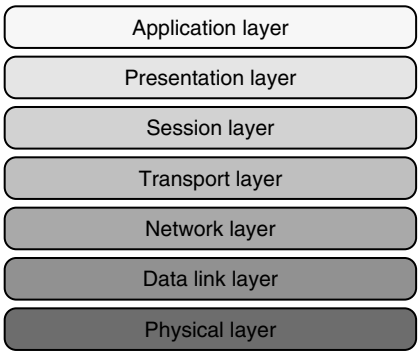


Figure 6. Open system interconnection layers.

9.5 High level communication and grid management

The development of new concepts such as "smart grid" or "vehicle to grid" has created the need for an appropriate communication protocol for electric vehicle charging, allowing functionalities such as:

- vehicle identification and billing, allowing payment for charging at public charging stations;
- individual billing of used energy to the user's account when the vehicle is charged at any outlet connected to a smart meter outside of the vehicle's common operating area (roaming);
- charge cost optimization by choosing the most appropriate time window where electricity rates are the lowest;
- grid load optimization by controlling charger ampacity in function of grid demand;
- peak-shaving functionality using electric vehicles connected to the grid as a spinning reserve (vehicle-to-grid);
- appropriate billing and user compensation functions for vehicle-to-grid operation.

Communication protocols for this issue will be typically based on the well-known 7-layer OSI reference model defined in international standards (ISO/IEC7498-1, 1994) (Figure 6).

The development of a communication protocol involves several actors, including physical devices such as the charging post or the vehicle controller, entities such as electricity suppliers or grid operators, and last but not the least the vehicle user. An overview is shown in Figure 7.

The local or remote communication system may have the function of a "clearing house" for the authentication, collection, and consolidation of grid and billing parameters from the actors as well as transmitting charging process information to the respective actors. Not all such functions are necessarily required for the basic charging functions, and some may be performed locally or remotely. The system can thus become either simple or rather complex, and several issues are still to be resolved.

Owing to the complex and multidisciplinary nature of the subject, its standardization of this issue is being addressed by a joint working group uniting ISO TC22 SC3 (electric equipment on road vehicles, including on-board communication systems) and IEC TC69 (electric road vehicles).

This joint working group is currently drafting a family of standards called *ISO/IEC15118* "road vehicles—vehicle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

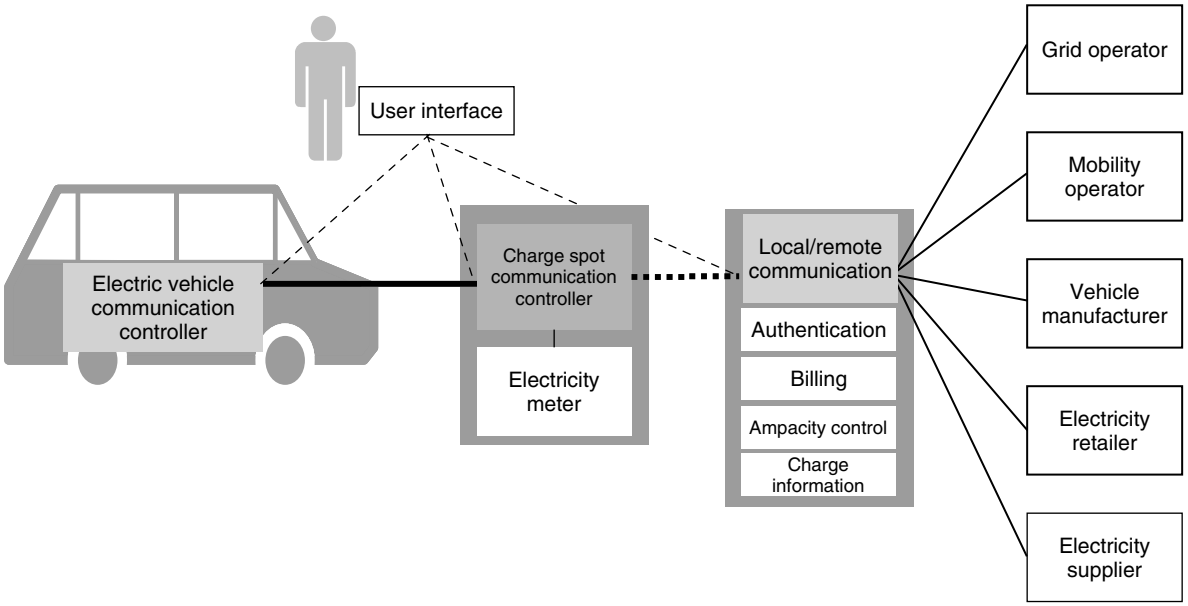


Figure 7. Factors involved in the charging process. (Reproduced with permission from Van den Bossche, 2010b. © World Electric Vehicle Association.)

to grid communication interface," to describe the communication, in terms of data format and message content, between the electric vehicle (including battery-electric vehicles and plug-in hybrid electric vehicles) and the EVSE (charging post), as well as message content and data structure to enable billing communication and grid management. Provisions for additional communication aspects (such as vehicle charge status information and configuration) are being considered to allow for interoperability of all vehicles with all charging stations. As the communication parts of those generic equipments are the electric vehicle communication controller (EVCC) and the supply equipment communication controller (SECC), this standard describes the communication between these components. All connections beyond the EVCC and how the messages will be exchanged are considered to be out of the scope as specific use cases.

The basic document of the ISO/IEC15118 family cases is part 1 "General Information and Use-case Definition" (ISO/IEC15118-1/CDV, 2012), which circulated as CDV (committee draft for voting) in the Spring of 2012. ISO/IEC15118-1 provides a general overview and a common understanding of aspects influencing the charge process, payment, and load leveling. It specifies furthermore the initial startup process and security issues for charging.

Communication between vehicle and charging post is defined by two different concepts called *basic signaling* and *high level communication*.

- Basic signaling shall be used to define items such as states and duty cycle for safety and initialization of the charging (CP function, cf. Section 9.2).
- High level communication shall be used to define issues such as identification, load leveling, and value-added services. In all cases, charging itself shall be controlled by the EV and the EVSE does not need to have a charging load controller. Information exchange with high level communication only occurs if both EV and EVSE are equipped with a high level communication device.

The specific actors involved with the charging process can be divided into primary actors, which directly involved with the charging process, and secondary actors, which may be involved because of the supply of information needed for the charging process.

All envisageable charging processes have to be contextualized in the so-called "use cases" in order to define communication needs. To this effect, the charging process is separated into eight functional groups where elementary use cases can be defined:

- beginning of charging process;
- communication setup;
- identification;
- authorization;
- target setting and charge scheduling;
- charge controlling and rescheduling;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

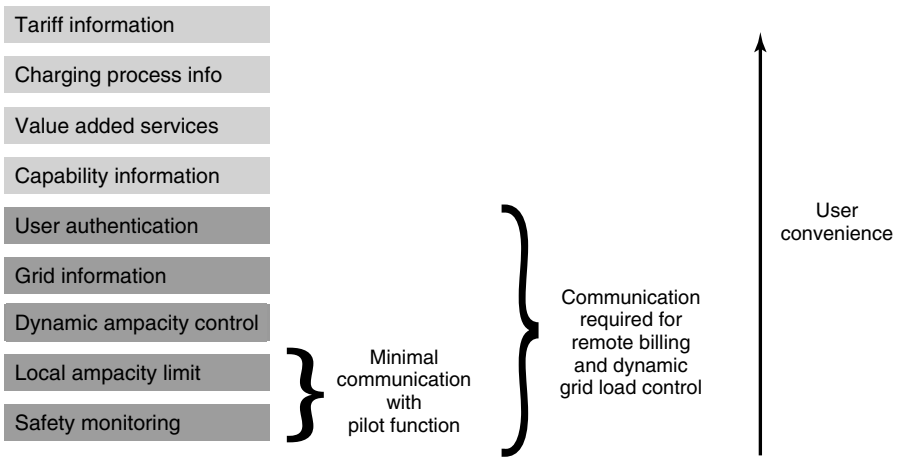


Figure 8. Communication and use cases. (Reproduced with permission from Van den Bossche, 2010b. © World Electric Vehicle Association.)

- value-added services;
- end of charging process.

Each use case will be constituted of a combination of elementary use cases, which can be used to construct specific charging scenarios.

These scenarios allow functionalities to be implemented with increased complexity, as illustrated in Figure 8; the increase of user convenience necessitates an extension of data structures to be exchanged.

Part 2 of ISO/IEC15118, which was also circulated as CD in the Spring of 2011 (ISO/IEC15118-2/CD, 2011), is called *technical protocol description and OSI layer requirements*. This document focuses on the top two layers of the OSI: the physical layer and the data link layer, for the communication between vehicle and EVSE.

Part 3 describes the physical and data link layers for a high level communication. It covers the overall information exchange between all actors involved in the electrical energy exchange. This document was circulated as CD in the Fall of 2011 (ISO/IEC15118-3/CD, 2011).

The interaction between the parts of the standard is shown in Figure 9.

For each use case, different scenarios are to be defined, relating to the desired control of charging by the grid operator, the used billing scheme, and the communication system for the user. Several systems are now under consideration and/or used in experimental fleets:

- use of an RFID (radio-frequency identification) tag;
- communication over the CP conductor;
- communication (at low or high data rate) through power line communication;
- communication with the vehicle's CAN-Bus system;
- wireless communication through Bluetooth or ZigBee devices;
- communication via mobile phone.

In order to define the communication messages, every such scenario shall be translated into a sequence diagram in unified modeling language (ISO/IEC19501, 2005).

It is clear that considerable standardization work remains to be performed in this field. The drafting of ISO/IEC15118 will form a good base for a generally applicable family of standards.

To be a true global solution, however, there should not be reliance on proprietary protocols allowing global use by all concerned parties.

9.6 Billing

The practice of charging of electric vehicles at public charging stations also raises the problem of billing the user for the energy consumed. Payment systems can make use of coins (vulnerable to vandalism), credit cards (creating the necessity of communication systems and involving transaction costs), or dedicated access devices (cards or RFID).

As the value of the electricity typically charged in one opportunity charging session is quite low compared to the parking cost in city centre environments, one can consider to charge the user according to time rather than to energy used, which dispenses the need for (more expensive) electricity counters. Some legal issues have also been considered here, as in some countries, the sale of electricity as such is heavily regulated. Furthermore, the real cost of electricity is not constant but varies with the peak load of the grid.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

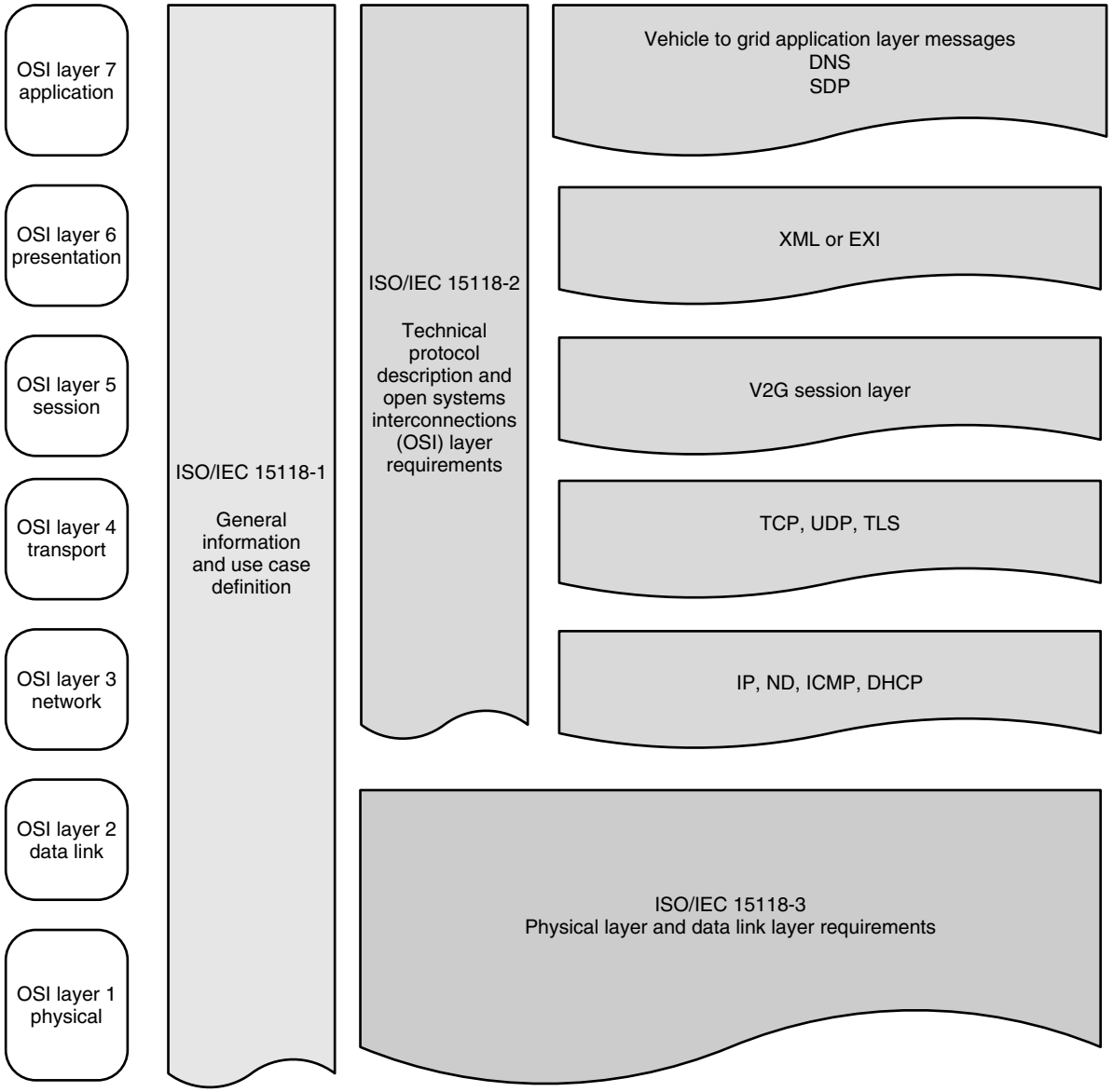


Figure 9. Parts of ISO/IEC15118.

The new developments in communication will allow a more sophisticated approach of this issue, with user identification and communication using wireless devices or mobile phones, differential tariffication according to time of the day and grid load, and compensation for energy returned to the grid under a "vehicle to grid" scheme. Furthermore, vehicles being charged at varying locations in a "smart grid" environment will charge the user in a transparent "roaming" way: wherever the user charges his or her vehicle, it will be charged on his or her own bill.

One should consider the complexity and cost of the proposed system facing the actual energy cost: most charging taking place in privately owned premises can be billed on existing meters without the need for complex business models.

10 ACCESSORIES FOR CONDUCTIVE CHARGING

10.1 Connection cases

The connection of the cable between the vehicle and the charging outlet can be carried out in three ways as defined in IEC61851-1 (IEC61851-1, 2010):

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- Case "A": where the cable and the plug are permanently attached to the vehicle. This case is generally found only in very light vehicles.
- Case "B": where the cable assembly is detachable, and connected to the vehicle with a connector. This is the most common case for normal and semifast charging.
- Case "C": where the cable and the vehicle connector are permanently attached to the supply equipment. This arrangement is typically used for fast charging (mode 4), so that drivers do not have to carry heavy cables around. Public charging stations using this case are, however, at a higher risk of copper theft.

10.2 Connection accessories

10.2.1 Standard domestic plugs

For mode 1 and mode 2 charging (also for mode 3 charging with power-line communication), standard plugs and sockets can be used encompassing only phase, neutral, and earth contacts. In most areas, this will usually be the standard domestic plugs as described in various national standards, and typically rated 10–16 A (IEC/TR60083, 2009).

One has to recognize, however, that these domestic plugs, particularly not the low cost versions mostly used on consumer grade equipment, are not really suited for the heavy-duty operation of electric vehicle charging, characterized by

- long-time operation at near rated current;
- frequent operation, including disconnection under rated load;
- exposure to outdoor conditions.

This leads to a shorter lifetime of the accessories and to contact problems, which may cause hazardous situations. It is thus recommended to limit the rating of the charging equipment using such plugs to a lower value, up to 10 A; their use being confined to small vehicles such as scooters (for which this current level is largely sufficient), as well as for occasional charging of larger vehicles (the "grandma" solution).

10.2.2 Standard industrial plugs

A better alternative is to use industrial plugs and sockets as defined by the international standard IEC60309-2 (IEC60309-2, 2005). These plugs (in standard, blue color for 230 V and red for 400 V) are widely used, particularly in Europe, for industrial equipment but also for outdoor use by the general public on venues such as camping sites and marinas, where they function in an operation mode comparable to an electric vehicle charging station. Their "industrial" nature and their relatively high insertion force, particularly for higher current versions, has been cited as an issue affecting user-friendliness for electric vehicle deployment.

These accessories are widely spread on the market and are relatively inexpensive, making them the preferable solution for mode 1 or mode 2 charging. The Focus Group (CEN–CENELEC, 2011) proposed them as interim solution pending development of dedicated accessories.

10.2.3 Dedicated electric vehicle accessories

The use of a physical CP conductor necessitates the introduction of specific accessories for electric vehicle use. Such plugs and sockets are described in the international standard IEC62196 "Plugs, Socket-Outlets, Vehicle Couplers and Vehicle Inlets–Conductive Charging of Electric Vehicles."

Part 1 of this standard (IEC62196-1, 2011) gives general functional requirements; it integrates general requirements from the industrial plug standard IEC60309-1 (IEC60309-1, 2005) with the electric vehicle requirements of IEC61851-1 (IEC61851-1, 2010).

Physical dimensions are treated in part 2. IEC62196-2 (2011), published in 2012. It does present standard sheets for several types of accessories:

- Type 1 (Figure 10)
  Type 1 is a single-phase coupler rated for 250 V and 32 A (30 A in the United States and Japan). It is fitted with two extra contacts: one for the CP and the other for an auxiliary coupler contact (CS), which can be used to indicate the presence of the connector to the vehicle and to signal the correct insertion of the vehicle connector into the vehicle inlet. With a diameter of 44



Figure 10. Type 1 coupler. (Reproduced with permission from Peter Van den Bossche. © Peter Van den Bossche.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 11. Type 2 plug. (Reproduced with permission from Peter Van den Bossche. © Peter Van den Bossche.)

mm, this connector is made in a compact way. Type 1, corresponding to the SAE J1172 coupler (SAE J1772, 2009), is intended as a vehicle connector only to be used in a case "C" configuration in mode 3, or with an ICCB in mode 2.

- Type 2 (Figure 11)
  European car manufacturers and utilities, however, recognizing the potential benefits of three-phase charging and the availability of three-phase supply in most European countries (cf. Section 3.2.2), expressed the desire for three-phase accessories.
  Type 2 is a three-phase coupler rated for currents up to 63 A, and has two auxiliary contacts. It is illustrated in Figure 11 based on a realization by the German company Mennekes. It comes in both a "plug–socket outlet" and a "connector–vehicle inlet" combination, which are of similar build but not intermateable. For the European market, the automobile industry is presently considering mounting both type 1 and type 2 inlets on cars and light trucks.
- Type 3 (Figure 12)
  Type 3 is also a three-phase type, based on a design by the Italian company SCAME further adopted by the "EV Plug Alliance." Its design is derived from a single-phase plug adopted as national standard in Italy (CEI-69-6, 2001) where it is in widespread use particularly for light electric vehicles such as two wheelers.
  One main difference between type 2 and type 3 accessories is the presence of "shutters" on the latter, providing IPXXD protection.

The CEN–CENELEC Focus Group recommended to define a unique footprint for the AC plug and socket outlet, encompassing five power contacts (three phases, neutral, and earth) and two auxiliary contacts to allow mode 3 charging. Both type 2 and type 3 fit this definition, the final choice between both types should be made rapidly. At this moment, type 2 plugs seem to be preferred in most



Figure 12. CHAdeMO DC (left) and type 1 AC (right) inlets on a Nissan Leaf vehicle. (Reproduced with permission from Peter Van den Bossche. © Peter Van den Bossche.)

European countries except where shutters are required by law such as in France and Italy.

10.3 Connectors for DC charging

Accessories for DC charging are being treated in IEC62196-3, which is presently under development. This will only pertain to connectors and vehicle inlets, as DC charging stations will typically use a case "C" connection.

Solutions considered include the DC connector proposed by the CHAdeMO association (Figure 13).

Furthermore, solutions are proposed encompassing both AC and DC connections in one unit, the so-called "combo" connectors.

The distinct parts of IEC62196-2, which are being prepared, will likely have the character of "catalogue standards," presenting standard sheets for various solutions without, however, making a distinct choice for one particular solution.

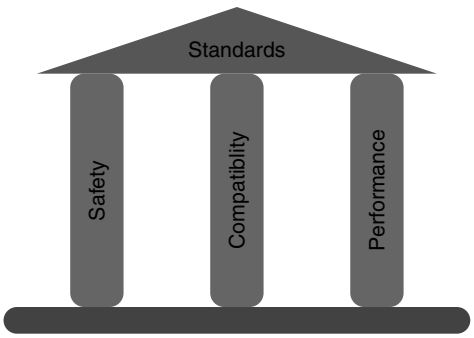


Figure 13. The house of standardization (Van den Bossche, 2010b). (Reproduced with permission from Van den Bossche, 2010b. © World Electric Vehicle Association.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

10.4 Battery connectors

Battery connectors are widely standardized for use in industrial electric vehicles. Functional requirements are given in the European standard EN1175-1 (EN1175-1, 1998). Several families of connectors in use conform to this standard, such as the so-called Euroconnectors (FEM 4.007b, 1993; DIN43589-1, 1994) and the "Anderson" connectors. These connectors are available with or without auxiliary contacts, for DC currents up to 350 A.

Although such connectors are sometimes found in electric road vehicles for internal connections, their use for (mode 4) charging of road vehicles is not advisable, particularly by the general public, as they are not designed for the higher battery voltage levels now in use or for connecting to cable assemblies. Furthermore, they lack earth conductors and are not designed to break a load.

11 CONCLUSIONS

The charging of electrically propelled vehicles remains a key issue for future standardization work. As with all standardization matters, charging standards pertain to the three main pillars of the house of standardization: safety, compatibility, and performance.

Safety standards ensure protection against electric shock and other related hazards, as well as controlling EMC issues, allowing the charging infrastructure to be used safely in all its potential environments.

Compatibility standards obviously refer to the definition of suitable plugs and sockets for electric vehicle charging. They also cover the communication needs of charging and allow the electric vehicle to be deployed in an extended area and the infrastructure to be universally usable.

Performance measurement standards, in the framework of this study, pertain to the management of energy measurement for billing as well as battery state of charge and state of health.

Intensive work is now being performed by international standardization committees in order to realize unified solutions, which will be a key factor in allowing the deployment of electrically propelled vehicles on a global level, highlighting the technical and societal relevance of standardization.

REFERENCES

Benning Ltd. (1990) Safe power for electric boats, brochure.

Bleijs, C. (2009) Low-cost charging systems with full communication capability. In *EVS-24*, Stavanger.

Botsford, C. and Szczepanek, A. (2009) Fast charging vs. slow charging: pros and cons for the new age of electric vehicles. In *EVS-24*, Stavanger.

BS74 (1917) British standard specification for charging plug and socket for vehicles propelled by electric secondarybatteries, BSI.

CEI-69-6 (2001) Foglio di unificazione di prese a spina per la connessione alla rete elettrica di veicoli elettrici stradali. CEI - Italian Electrotechnical Committee.

CEN–CENELEC (2011) CEN–CENELEC Focus Group on European Electro-Mobility Final Report to CEN and CENELEC Technical Boards in Response to Commission Mandate M/468 concerning the charging of electric vehicles, CEN–CENELEC, 6.

CENELEC TS50457-2 (2006) *Conductive Charging for Electric Vehicles – Part 2: Communication Protocol Between Off-board Charger and Electric Vehicle*, CENELEC.

Commission of the European Communities (2004a) EU OJ L 390, 31.12.2004 Directive 2004/108/ec of the European parliament and of the council on the approximation of the laws of the member states relating to electromagnetic compatibility and repealing directive 89/336/eec.

Commission of the European Communities (2004b) Directive 2004/104/ec of 14 October 2004 adapting to technical progress council directive 72/245/eec relating to the radio interference (electromagnetic compatibility) of vehicles and amending directive 70/156/eec on the approximation of the laws of the member states relating to the type-approval of motor vehicles and their trailers, EU OJ L337, 13.11.2004.

Commission of the European Communities (2006) EU OJ L 374, 27.12.2006 Directive 2006/95/ec of the European parliament and of the council on the harmonisation of the laws of member states relating to electrical equipment designed for use within certain voltage limits.

Commission of the European Communities (2010) Standardisation mandate m/468 to cen, cenelec and etsi concerning the charging of electric vehicles, 29.6.2010.

DIN43589-1 (1994) Geräte-Steckvorrichtungen 80, 160, 320 A, 150 V für Elektro-Flurförderzeuge—Teil 1: Anschlußmaße, Werkstoff, Kennzeichnung, DIN.

Electric Vehicle Association of America (1914) Standard charging plugs and receptacles. *The Central Station*, **13** (7), 304.

EN1175-1 (1998) *Safety of Industrial Trucks. Electrical Requirements. General Requirements for Battery Powered Trucks*, CEN.

EN50065-1 (2002) *Signalling on Low-voltage Electrical Installations in the Frequency Range 3 kHz to 148, 5 kHz - Part 1: General Requirements, Frequency Bands and Electromagnetic Disturbances*, CENELEC.

FEM 4.007b (1993) *Industrial Trucks—Connectors fo Traction Batteries Up to and Including 96V—Dimensions*, Fedération Européenne de la Manutention.

IEC/TR60083 (2009) *Plugs and Socket-outlets for Domestic and Similar General Use Standardized in Member Countries of IEC*, 6th edn, IEC.

IEC60309-1 (2005) *Plugs, Socket-outlets and Plugs for Industrial Purposes – Part 1: General Requirements*, IEC.

IEC60309-2 (2005) *Plugs, Socket-outlets and Plugs for Industrial Purposes – Part 2: Dimensional Interchangeability Requirements for Pin and Contact-tube Accessories*, 4.1 edn, IEC.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

IEC61851-24/CD (2011) Number 69/208/CD. IEC TC69 PT61851-24. *Electric vehicle conductive charging system - digital/data communication of d.c. charging control between off-board d.c. charger and electric vehicle*.

IEC61980-1/CD (2000) *Electric vehicle inductive charging systems – Part 1: General requirements*, Number 69/125/CD. IEC TC69 WG4.

IEC61980-2/CD (2000) *Electric vehicle inductive charging systems – Part 2: Manual connection system using a paddle*, Number 69/126/CD. IEC TC69 WG4.

IEC60364-7-722/CD (2012) *Low-voltage Electrical Installations - Part 7-722: Requirements for Special Installations or Locations - Supply of Electric Vehicles*, 1st edn, IEC TC64.

IEC61000-3-12 (2011) *Electromagnetic Compatibility (EMC) – Part 3-12: Limits - Limits for Harmonic Currents Produced by Equipment Connected to Public Low-voltage Systems with Input Current >16 A and ≤ 75 A Per Phase*, 2nd edn, IEC.

IEC61000-3-2 (2005) *Electromagnetic Compatibility (EMC) – Part 3-2: Limits - Limits for Harmonic Current Emissions (Equipment Input Current ≤ 16 A Per Phase)*, 3rd edn, IEC.

IEC61851-1 (2010) *Electric Vehicle Conductive Charging System – Part 1: General Requirements*, 2nd edn, IEC.

IEC61851-23/CD (2011) Number 69/206/CD. IEC TC69 PT61851-23. *Electric Vehicle Conductive Charging System – Part 23: d.c. Electric Vehicle Charging Station*, 1st edn.

IEC62196-1 (2011) *Plugs, Socket-outlet and Vehicle Couplers – Conductive Charging of Electric Vehicles - Part 1: Charging of Electric Vehicles up to 250 A a.c. and 400 A d.c.*, 2nd edn, IEC.

IEC62196-2 (2011) *Plugs, Socket-outlet and Vehicle Couplers – Conductive Charging of Electric Vehicles - Part 2: Dimensional Interchangeability Requirements for Pin and Contact-tube Accessories with Rated Operating Voltage up to 250V a.c. Single Phase and Rated Current up to 32A*, 1st edn, IEC SC23H.

ISO/IEC15118-1/CDV (2012) *Road Vehicles—Vehicle to Grid Communication Interface – Part 1: General Information and Use-case Definition*, 1st edn, ISO/IEC.

ISO/IEC15118-2/CD (2011) *Road Vehicles—Vehicle-to-Grid Communication Interface – Part 2: Technical Protocol Description and Open Systems Interconnections (OSI) Layer Requirements*, 1st edn, ISO/IEC.

ISO/IEC15118-3/CD (2011) *Road Vehicles – Vehicle to Grid Communication Interface – Part 3: Physical Layer and Data Link Layer Requirements*, volume 69/204/CD, ISO/IEC.

ISO/IEC19501 (2005) *Information Technology – Open Distributed Processing – Unified Modeling Language (UML) Version 1.4.2*, ISO/IEC.

ISO/IEC7498-1 (1994) *Information Technology – Open Systems Interconnection – Basic Reference Model: The Basic Model*, ISO/IEC.

ISO11898-1 (2003) *Road Vehicles – Controller Area Network (CAN) – Part 1: Data Link Layer and Physical Signalling*, 1st edn, ISO.

ISO11898-2 (2003) *Road Vehicles – Controller Area Network (CAN) – Part 2: High Speed Medium Access Unit*, 1st edn, ISO.

ISO14229-1 (2006) *Road Vehicles – Unified Diagnostic Services (UDS) – Part 1: Specification and Requirements*, ISO.

ISO14230-1 (1999) *Road Vehicles – Diagnostic Systems – Keyword Protocol 2000 – Part 1: Physical Layer*, ISO.

ISO14230-2 (1999) *Road Vehicles – Diagnostic Systems – Keyword Protocol 2000 – Part 2: Data Link Layer*, ISO.

ISO14230-3 (1999) *Road Vehicles – Diagnostic Systems – Keyword Protocol 2000 – Part 3: Application Layer*, ISO.

SAE J1772 (2009) *Electric Vehicle Conductive Charge Coupler*, SAE.

Szczepanek, A. and Botsford, C. (2009) *Electric vehicle infrastructure development: an enabler for electric vehicle adoption*. In *EVS-24*, Stavanger.

Toepfer, C. (1994) *EV Charging Systems, Report of the Connector and Connecting Station Committee*. Number TR-104623-V2, Project 2882, Electric Power Research Institute.

UNECE-R100 (2011) *Uniform provisions concerning the approval of vehicles with regard to specific requirements for the electric power train*. United Nations Economic Commission for Europe, rev. 1 edition.

Van den Bossche, P. (1992) *The "Twelve Electric Hours" competition: a good way to evaluate electric vehicles in city traffic*. In *EVS-11*, Firenze.

Van den Bossche, P. (2003) *The electric vehicle: raising the standards*. PhD thesis, Vrije Universiteit Brussel.

Van den Bossche, P. (2010a) in *Electric and Hybrid Vehicles*, 1st edn (ed. G. Pistoia), Chapter 20, Elsevier, pp. 517–544.

Van den Bossche, P. (2010b) *Matching accessories: standardization developments in electric vehicle infrastructure*. In *EVS-25*.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto192
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Hardware-in-the-Loop Simulation

Susanne Köhl and Markus Plöger
dSPACE GmbH, Paderborn, Germany

1 Introduction to and Objectives of Hardware-in-the-Loop Simulation 1
2 Roles in ECU Software Testing 2
3 Components of a HIL Simulator 5
4 Integration into the Electronics Development Process 10
5 Outlook 12
6 Conclusion 13
Acknowledgments 14
Related Articles 14
References 14
Further Reading 14

1 INTRODUCTION TO AND OBJECTIVES OF HARDWARE-IN-THE-LOOP SIMULATION

Hardware-in-the-loop (HIL) simulation enables electronic control unit (ECU) functionality to be tested in real time and in a realistic test environment during ECU software development, without the real environment (vehicle, other ECUs, and connected actuators/sensors) actually being available. The static and dynamic processes in the ECU's environment are simulated in real time. The unit under test and the simulation environment form a closed loop (Figure 1).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Tests on a HIL simulator range from early tests of single functions to release tests for the entire ECU network, and have become an integral part in the ECU development processes at original equipment manufacturers (OEMs) and suppliers. This has given the electronics development process another milestone for enhancing vehicle quality by eliminating errors. ISO 26262, the ISO (International Organization for Standardization) standard for safety-relevant in-vehicle electric/electronic systems, also acknowledges HIL to be an important element in developing and testing safety-relevant functions.

ECUs for all vehicle domains—powertrain and vehicle dynamics, chassis, comfort and safety, and so on—are tested on HIL systems. In recent years, topics such as electrification and driver assistance systems have been added to this list (Plöger and Köhl, 2006).

The ECU's output signals and bus signals are measured in real time. These signals are input values for the real-time simulation model of the controlled system. The actuators, driven by the output signals of the ECU, are either electrically simulated or included in the simulation model. The sensor signals that the ECU expects for its operations are generated by the HIL simulator according to the state of the simulated controlled system. The bus signals [CAN (controller area network), LIN (local interconnect network), FlexRay, Ethernet, MOST, proprietary serial buses, and so on], which in a real vehicle would be sent by other ECUs or other sensors, are simulated and generated in parallel to this. Simulating the communication network in this way is also called *restbus simulation*. In addition to an ECU's good behavior being simulated, its behavior in response to a fault can also be tested by means of electrical failure simulation at the input, output, or bus channels.

HIL tests differ from other functionality test methods in the following ways: ECU benches and test drives are expensive and labor-intensive, tests on these systems can be

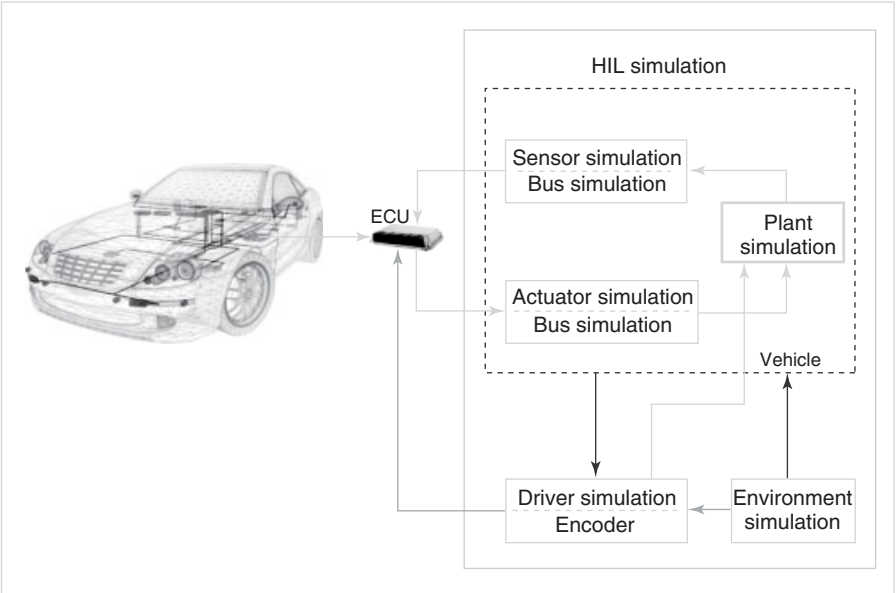


Figure 1. Signal flow in HIL simulation.

automated only partially or not at all, and can, therefore, not be reproduced. Such tests cannot be performed until the required components are actually available. With HIL, individual components can be tested before all the vehicle electronics are available. In test drives, it is dangerous to test at the extremes of ranges. Unlike tests on a test bench, HIL system testing allows operating points to be set in any way desired, for example, to cover the entire engine speed range. Automation can be used on a HIL system to make reproducibility considerably more precise. The cost of testing is reduced because fewer test vehicles are needed. Up to 90% of all faults occurring in test drives can be reproduced by means of HIL simulation (Lamberg, 2006).

It is automation that makes regression tests possible at all. Automation also means that testing can be done round the clock and on seven days a week (24/7 tests) (Plöger and Köhl, 2006).

2 ROLES IN ECU SOFTWARE TESTING

One task performed by car manufacturers is to integrate the electrics/electronics components into the overall system. They, therefore, also have the task of testing how ECUs from different suppliers interact in the overall system (ECU network testing). At this point in time, the suppliers should already have verified that the individual ECUs (or smaller networks provided by them) fulfill the specifications.

Until now, car manufacturers have also often performed functional tests on the ECUs. To save time and money, however, it must not be allowed to happen that a manufacturer discovers errors at the component level while performing network tests (Plöger and Köhl, 2006).

2.1 Task assignment between OEM and suppliers

Before delivering a new ECU version for integration into the overall vehicle, the supplier (which can also be a department within the OEM) must first run component tests to verify the version. Subsequently, integrating into a (sub)system the various components supplied by external suppliers is a critical phase in the development process. Misunderstandings and imperfect specifications can result in errors not being found until the car manufacturer performs tests on a subsystem (e.g., within one vehicle domain) or the overall system (Honisch, Hutter, and Schmid, 2004). However, the real objective of the manufacturer's system tests is to verify the interaction between the components and to verify functions that are implemented across several ECUs. These distributed functions (the central algorithm, sensors and actuators, and operating and display elements are located on different ECUs) can be verified only during interaction between all the ECUs involved.

If the suppliers and the OEM use compatible test systems, they can easily exchange and compare tests and test results. Shared concepts and clearly apportioned responsibilities avoid unnecessary redundancies and increase test case coverage.

In some circumstances, the OEM has to provide a restbus simulation of the other ECUs, so that the supplier can

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

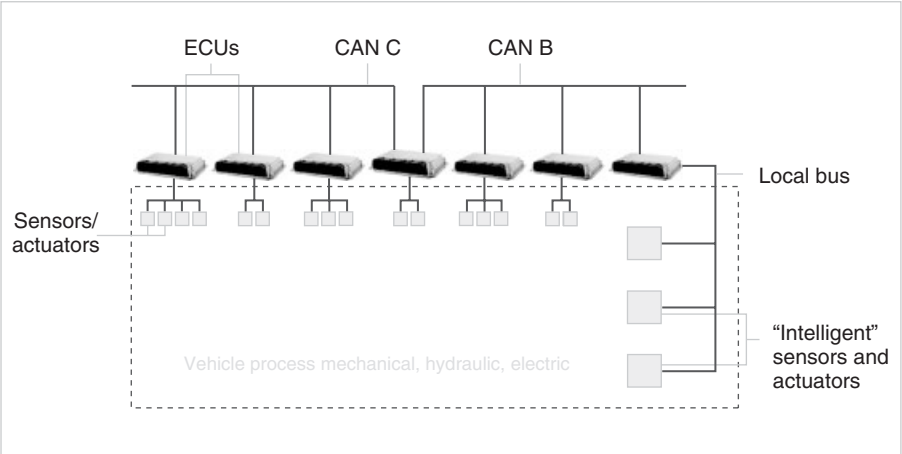


Figure 2. ECU interfaces (Plöger and Köhl, 2006). (Reproduced by permission of Springer Vieweg.)

run tests at component level (Figure 2). In the other direction, the supplier often provides the OEM with parameterized models representing the environment of the delivered ECU.

The roles involved in the different development phases and the resulting requirements for the HIL system are explained later (Köhl and Jegminat, 2005). Figure 3 provides an overview.

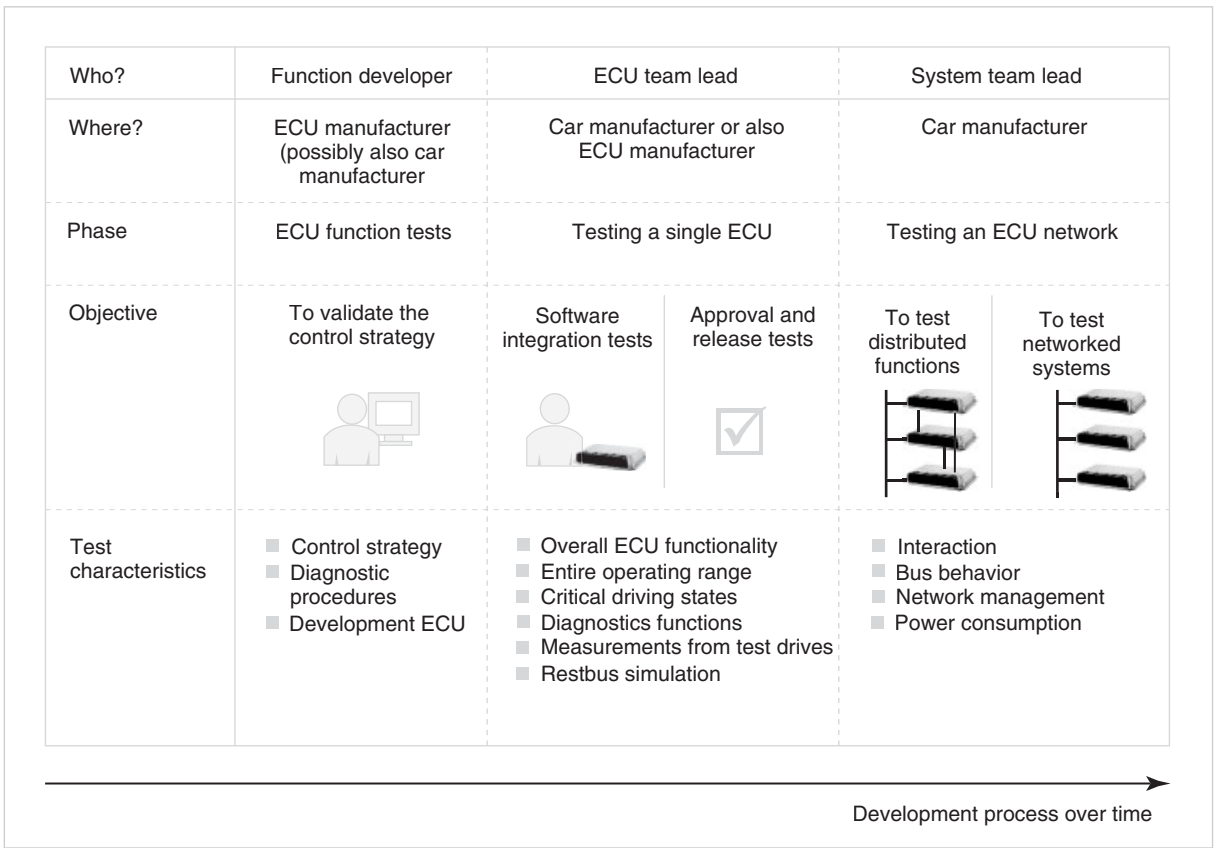


Figure 3. Roles in the development process when using HIL simulation (Plöger and Köhl, 2006). (Reproduced by permission of Springer Vieweg.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.2 Function developer at supplier (or at OEM)

The function developer tests single functions and the control strategies. The typical objective is function approval.

The following requirements for the HIL system are typical:

- It is a common practice for suppliers to reuse a control algorithm in different ECU variants. This requires flexible simulator hardware that is easy to adapt to different ECU variants. The HIL software components (SWCs), such as submodels, parameter sets, and test scripts, have to be suitably managed.
- Simulator experiment software that can be used flexibly and interactively is also desirable.
- Usually, only a small amount of automation is necessary (test scripts are often created only in parallel to ECU/function development).

Because ECU diagnostics depend on signal values that are not (or cannot be) calibrated until later, function testing in this phase is frequently performed on an ECU with its diagnostic functionality deactivated.

To save time and effort, tests that were performed during function development should not have to be repeated in the integration phase (Plöger, 2006).

2.3 ECU team lead at OEM (or supplier)

As soon as the functions have been integrated on the target hardware together with the lower software layers (operating system and I/O drivers, e.g.), the ECU has to undergo macroscopic tests. These include tests on overlapping administration layers, for example, in the handling of diagnostic memory.

Either the manufacturer or the supplier performs the ECU release test. The objective is to release the complete ECU, including diagnostics, as error-free.

Automated tests are indispensable for ECU integration tests. Interactive testing is necessary only for test development or if the source of a detected error has to be found.

Typically, there are several integration stages with different quality gates. If new software versions are supplied, or if OEMs use ECUs from different suppliers but with identical functionalities (second source principle), regression tests are the best form of testing. In ECU (integration) tests, managing the simulator's SWCs (submodels, test scripts, and so on) is even more important than during function tests (Plöger, 2006).

2.4 System team lead at OEM

As already mentioned, for efficiency reasons, tests that were already finished at the component level should not be repeated for (sub)system testing. When system-release tests are performed on a HIL system, the focus lies explicitly on testing distributed functions and bus communication. Network management is also tested in this context. The test objective for HIL release tests is an error-free overall system, including diagnostics.

Frequently, different vehicle configurations have to be tested. Country-specific variants, different equipment variants, and special vehicle models make it necessary to handle different configurations. Combinatorial tests are, therefore, needed. The HIL system must support numerous different controlled system models, I/O channels, and bus communications.

Automated tests are again indispensable. Tests designed for the system must be easy to repeat for all ECU/vehicle variants. The higher the degree of automation, the higher the test coverage is.

Another important aspect is that tests verifying the functionality during vehicle development can also be used to examine warranty claims after vehicle production has begun (Plöger, 2006).

At this point, it should be mentioned that simulation always carries with it a residual risk because of imprecisions in the simulation models and situations that were not anticipated. With in-vehicle system tests, there is no longer any risk from any issues neglected during model creation. In addition, functionalities that require driver interaction can better be tested in a driving simulator or in a test drive. Additive test drives are indispensable for these very reasons. The final release of the entire system must, therefore, still be carried out in the vehicle (Schäuffele and Zurawka, 2003).

2.5 Roles within one company

To cope with the complexity involved in ECU development and HIL simulation, there is often a clearly defined task separation among system team leads, function team leads, test system experts, test developers, and test users.

The test system expert is often already involved in project planning for the test system and putting it into operation. He or she is frequently someone from the HIL test system supplier. Two increasingly important roles are those of test developer and test user. The test developer creates the test specification or test plan together with the system team lead or function team lead. The test developer then bases the actual test implementation on this

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and makes it available for test execution. The test user is responsible for test execution. He or she puts together the tests required for a specific test task, parameterizes them, and executes them. With major projects, an overall test manager coordinates the activities (Wallentowitz and Reif, 2006; Lamberg, 2006).

3 COMPONENTS OF A HIL SIMULATOR

Although requirement details vary in different application fields (Section 2), the essential components of the HIL simulator are always similar. The individual components of a HIL simulator and the requirements that apply to them are described in more detail later (Schütte *et al*., 2001). Figure 4 shows all the essential components for a HIL simulator for a single ECU.

3.1 Hardware

3.1.1 Real-time processors

With its increasingly complex and precise models (Section 3.2), HIL simulation makes particularly tough demands on the processing power of the real-time processors. The models are usually computed with a step size of 100 μs to 1 ms to meet the requirements of real-time behavior. Motor sports are one example of short processing times. Highly dynamic electric motors, such as those installed in modern hybrid or electrical vehicles, need to be simulated even faster than 100 μs. As a rule, the simulation models for these are simulated synchronously (pulse-center-aligned) to the frequency of the pulse width modulation (PWM) control on the real-time processor. Sample times of <100 μs are usual.

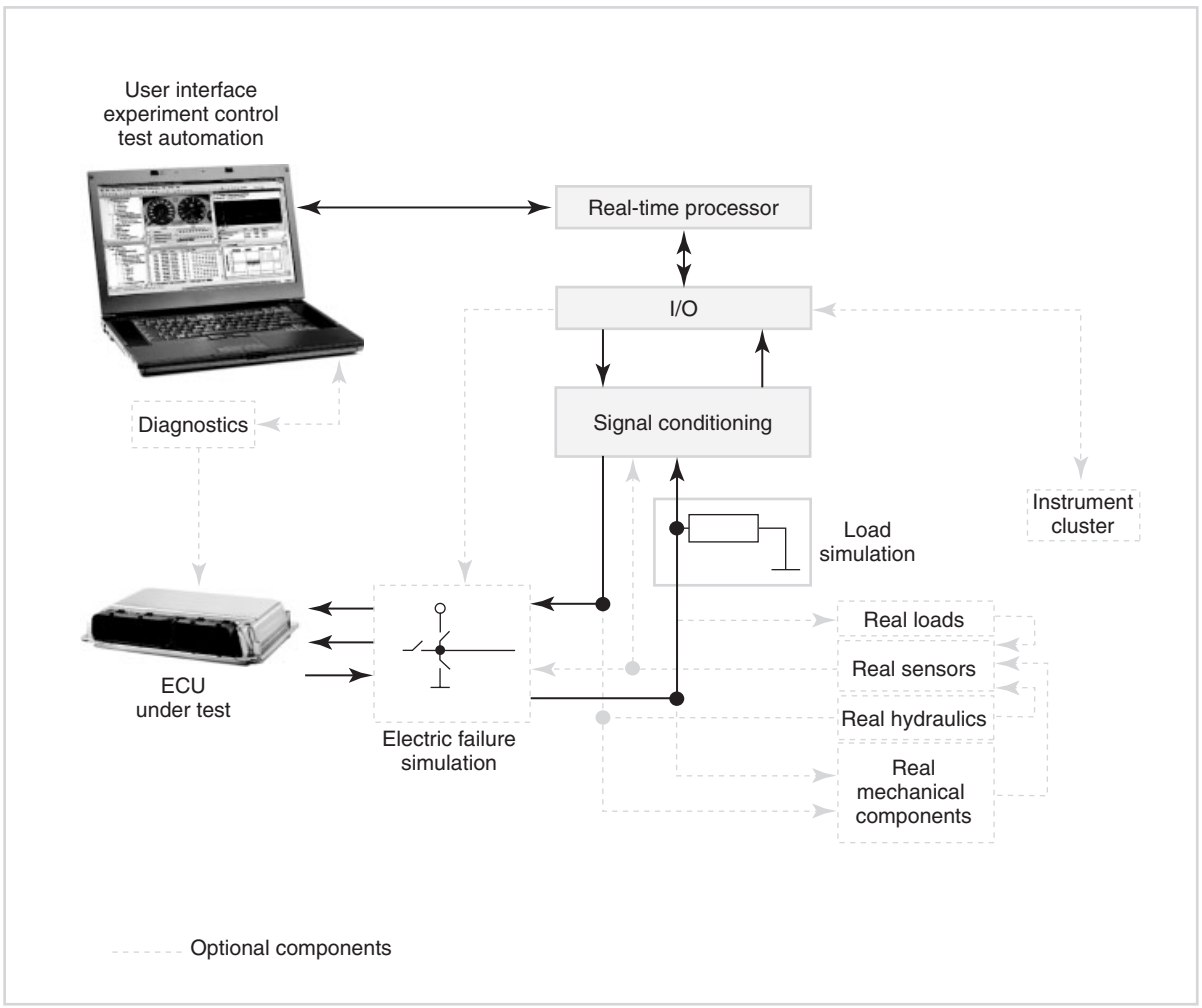


Figure 4. Components of a HIL simulator (Plöger and Köhl, 2006). (Reproduced by permission of Springer Vieweg.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For even higher requirements, parts of the model such as the electric winding model are executed on an additional FPGA (field-programmable gate array). This makes it possible to achieve sample times of 1 μs and far less. It also provides far more precise representation of currents because of motor inductivities compared with processor-based electric motor simulation. In addition, the effect of discontinuities (e.g., switching events of diodes and transistors) can be simulated more precisely and easily by means of fast-running models of the motor and the power electronics on FPGAs (Schulte, Kiffe, and Puschmann, 2011).

The complexity of the required models has grown continuously over the last few years, and the required processing power has risen with it. PC processors have become more important in HIL technology because they now outperform the PowerPC processors that were previously used. More and more multicore processors are also being used. These perform distributed and, therefore, parallel computation of complex models.

To couple several single simulators to create a networked system, processors can be connected to form a distributed multiprocessor system. Nowadays, with distributed multiprocessor systems, HIL simulators for ECU networks are implemented to test all the ECUs installed in a vehicle. Note that it is perfectly possible for a top-of-the-range vehicle to have over 60 ECUs (Gehring and Schütte, 2002; Köhl, Lemp, and Plöger, 2003).

Next to pure processing power, fast connection of I/O simulation is also extremely important. Standard PC interfaces are not usually designed for this, so high performance HIL systems use proprietary protocols to ensure real-time performance. With flexible network topologies and high bandwidths, spatially separated I/O boards can also be connected without the need for a local processor node (dSPACE GmbH, 2011; Plöger and Köhl, 2006).

3.1.2 Generating and capturing signals

The I/O components used in general measurement and control engineering are often unsuitable for representing sensor signals and capturing actuator control signals from an ECU. For example, special requirements apply to generating and capturing powertrain signals in real time. Here, it is the combination of angle- and time-based signals that requires special preparation:

- Ignition and injection signals must be captured precisely with regard to crankshaft position and duration.
- It must be possible to specify any arbitrary signal forms for position encoders and generate them angle-synchronously with high precision.
- Knock signals and ionization currents or cylinder in-pressure sensor signals must be triggered crank angle-synchronously, and also generated on a time basis or partly on an angle basis. Particular characteristics of the signals (damping, dependence on current engine load, and so on) need to be changed online in the simulation raster.

These requirements can be met by special hardware architectures (e.g., in FPGA technology). Resolutions of 0.01° crankshaft angle over the entire speed range 0–30,000 rpm are no longer any problem.

The HIL simulation of fast electric motor controllers also brings with it new requirements for generating and capturing I/O signals. For example, the typical six PWM signals of a synchronous motor ECU have to be measured with high precision. The duty cycle, period times, and power-up times of individual signals, as well as the dead times between neighboring signals, have to be measured synchronously to the PWM frequency at defined sample time points. Interrupts for triggering the associated simulation model also have to be issued at the same point in time (see earlier). Because the electric motor control requires the exact angle position of the rotor, this must also be computed and output with a high time resolution (25 ns is common). I/O modules with FPGA technology are also used for these tasks (Plöger and Köhl, 2006).

3.1.3 In-vehicle communication

Another fundamental requirement is for the real-time system to support all the automotive bus systems. The CAN bus, with all its variations such as high speed, low speed, single-wire fault tolerant, and 24-V trailer CAN, has been the most important of these for a long time. An example of CAN is, therefore, given here.

For HIL operation of one or a few ECUs, the first step is to read in the CAN messages sent by the ECU(s) in real time, typically in a 1-ms task. However, to simulate the environment of the ECUs under test, it is also important to simulate the CAN bus traffic of ECUs that are not actually present and couple it with the test object (restbus simulation). Where several, or even all, of the ECUs on a CAN bus have to be tested in the network, it may be necessary to run a series of tests where the messages exchanged between the real ECUs are manipulated. This can be done by means of a CAN error gateway.

The automotive industry is also increasingly using other bus systems apart from CAN. The HIL system must of course support these, too. They include the LIN bus (especially used in interior/comfort functions) and increasingly

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

also time-triggered buses such as FlexRay. The MOST bus, which has been used for infotainment until now, is being replaced more and more by Ethernet.

3.1.4 Signal conditioning

The main task of signal conditioning is to adjust the signal levels between the ECUs and the I/O boards. The amplitudes of incoming analog signals usually have to be reduced. Some real sensors output only very low signal voltages; therefore, for the HIL system to simulate these sensors, the output voltage of the I/O channel must be reduced without losing resolution. With analog signals especially, it is essential to ensure that the signals are passed to the ECU differentially. This avoids errors because of mass displacement. Moreover, for digital signals, signal conditioning must represent the characteristics of the real signals. For example, the digital outputs of the HIL system must be configured among open collector, pull-down, and pull-up. Signal conditioning also filters analog signals and protects I/O boards from overvoltage and overcurrent.

Systems that were specially developed for use in HIL systems already have signal conditioning integrated. It is then most important that the signal conditioning is as flexible and configurable as possible. This can be achieved by systems with software-configurable signal conditioning for typical signal levels (such as SCALEXIO from dSPACE) (Plöger and Köhl, 2006).

3.1.5 Load simulation

Each ECU monitors its own outputs for electrical failures such as short circuits or broken wires. To prevent the ECU assuming a broken wire when tested during normal operation, a real load or suitable load simulation must be connected to each ECU output. There are different approaches to load simulation, depending on how sensitive the diagnostics are.

In the simplest case, it is enough to connect a replacement load with higher impedance compared to the real load. This can often be plugged directly into the system or the cable harness to the ECU, as in normal cases, it does not require cooling.

If the ECU also monitors the current flowing at a critical output, at least an electrically equivalent load must be installed and connected in the HIL simulator. This could be a load resistance mounted on a heat sink, for example.

In particular cases, the real load needs to be integrated, for example, if it is a current-regulated hydraulic valve whose inductance is difficult to simulate.

The HIL simulator usually measures the control signals directly at the load and uses them as input values for the real-time simulation model. With current-regulated loads, the load current first has to be converted into a voltage by suitable converters (Hall sensors or shunt resistors) and filtered in the HIL simulator before it can be read in.

Sometimes, it is necessary to simulate a load with a complex I/O behavior, for example, an inductive load with high dynamics. Electronic load modules are used for these purposes. They provide current sink and source capability with high speed current regulation. Combining them with an FPGA board running a model of the load to be simulated (e.g., an E-motor) yields an emulation of the electric motor on electric power level (Plöger and Köhl, 2006).

3.1.6 Test benches

Sometimes, it is necessary to test the ECU together with real loads, which need physical stimulation. For example, the developer of an electronic power steering (EPS) system wants to test the controller, power stages, and steering motor. In these cases, a test bench is built in which a second load motor is mechanically coupled to the steering motor. The load motor continuously receives its set points for torque and speed from the real-time model. As a result, the overall steering system can be tested in a realistic HIL set up.

3.1.7 ECUs with integrated sensors

Some ECUs have integrated sensors inside their enclosures, and these sensors have to be physically stimulated in order to operate the ECU. One example is transmission ECUs that are installed directly at the transmission. They detect the input or output speed by means of integrated Hall sensors. Another example is inclination sensors that are integrated into ECUs for electric park brake functionality. The HIL system cannot stimulate the sensors electrically in these cases. The same applies to sensors that communicate with the ECU via a secret protocol. This is the case with crash sensors, the speed sensor of a tachograph, and so on. There are two ways to treat these sensors in HIL simulation.

One way is to use the real physical value to stimulate the sensor. For example, an integrated speed sensor can be stimulated by an alternating magnetic field generated by electric magnets, and the electric park brake ECU can be inclined by mounting it on a tilting platform.

An alternative way is to open ("crack") the ECU or the sensors to access the electrical interface. This is a suitable method for the acceleration sensors of an airbag ECU, for example (Plöger and Köhl, 2006).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.1.8 Electric failure simulation

HIL simulation is used not only to test ECU functions in normal operation; another important use case is to test diagnostic functions. This begins by testing whether a fault such as a detached sensor is detected correctly. The next tests check whether the ECU activates any associated limp home function, for example, by calculating a substitute value for a missing sensor. Powerful, automatable electric failure simulation is a necessary prerequisite for these tests and should cover at least the following fault cases: broken wire, short circuit to ground, short circuit to battery, and short circuits between ECU pins. To simulate poor contacts, it might be necessary to switch modifiable line resistances between two ECU pins, or between an ECU pin and a reference (mass or battery voltage). Loose contacts at switch inputs can be simulated by high frequency pulse patterns (Plöger and Köhl, 2006).

3.2 Software

The software for a HIL system can be separated into real-time software running on the simulator hardware on the one hand and experiment and test automation software running on the host PC on the other (Figure 5).

3.2.1 Real-time behavior models

To test ECUs in virtual environments, the controlled system and its environment need to be simulated. This is variously called *the simulation model*, *the plant model*, *the behavior model*, or *the real-time model* of the controlled system.

HIL testing requires the model and the simulator to perform in real time. The requirements regarding mathematical plant models vary greatly according to application area. The models should be as precise as possible to provide plausible signals to the ECU in all operating states, that is, to close all the control loops, such that the ECU's diagnostics do not detect a fault.

The modeling requirements vary according to vehicle domain.

For engine simulation including diagnostics, complex models are common. Often, these models have crankshaft-angle-synchronous tasks. In modern engines, it is also necessary to simulate the in-cylinder behavior: this again is even more complex: Pressure behaviors are no longer simulated in mean-value models. On-board diagnostic functions in modern engine ECUs require relatively precise dynamic modeling of the exhaust gas system including temperatures and pressures or more precise catalytic converter models than was necessary just a few years ago. The more sensitive the diagnostic functions, the greater the necessary model complexity.

Typically, planetary transmission models are no longer signal flow-based but acausal, which allows object-oriented, physical modeling.

When driver assistance systems have to be tested, traffic simulation is important. Besides sensors for object detection, it is also important to simulate

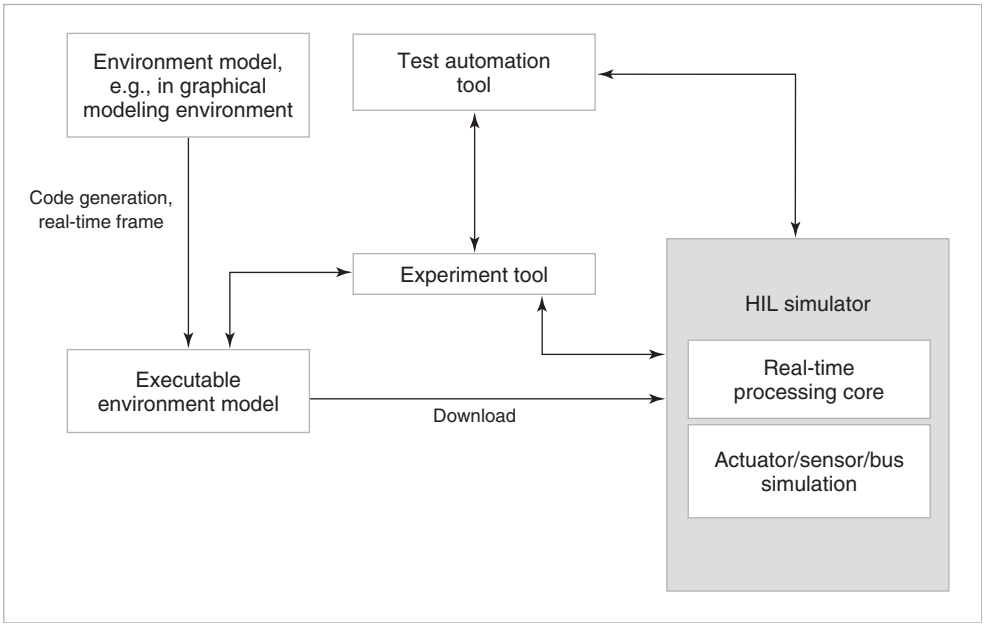


Figure 5. Software components in HIL simulation (Plöger and Köhl, 2006). (Reproduced by permission of Springer Vieweg.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

other traffic participants, such as oncoming traffic or pedestrians.

To test comfort functions, often all that is needed is quite simple functional, nonphysical behavior models for simulating the controlled system.

More precise models typically need higher processing power. Obviously, there is also more effort involved in parameterizing the models in these cases.

The controlled system models are usually available in a block-oriented graphical simulation environment. The currently most widely used tool for this is MATLAB®/Simulink®. Tools such as SimScape, SimulationX, and Dymola are used to design acausal models.

If very fast models are required, model calculation is transferred to FPGA. These are implemented in hardware description languages (HDLs). To avoid manual HDL programming, higher level tools are available to generate the HDL code directly from MATLAB, Simulink, and so on.

Many models are already available from offline simulation. As it is now increasingly possible to give the HIL systems enough processing power for complex modeling, models from offline simulation can often be reused in HIL simulation. Alternatively, parameters are exchanged via standard interfaces between the complex and the more simple models to allow maximum continuity within the virtual development process (Plöger and Köhl, 2006).

3.2.2 *Model implementation and hardware configuration, including I/O and restbus simulations, and interprocessor communication*

In addition to the actual controlled system models, the real-time processor also computes all the I/O functions including restbus simulation. All communications involved in multiprocessor operation can of course be implemented transparently for the user in high quality graphical form.

From the behavior model, which is graphically modeled, C code is typically generated automatically. The hardware configuration and the connection to the interfaces of the behavior model are performed in a simulator-specific configuration and implementation tool (such as ConfigurationDesk). This tool allows the signal path between the ECU or load pins and the behavior model interfaces to be configured and displayed. Besides simulator configuration, the tool also implements the behavior model code (from MATLAB/Simulink/Simulink Coder™) and I/O function code (from ConfigurationDesk) on the HIL hardware. The entire build process for a real-time application is handled by the configuration and implementation tool.

In addition to the I/O functionality, the allowed failure cases per channel are also configured at this stage.

If necessary, settings for partitioning the model to several cores or processors are also made in this context.

Comprehensive documentation options and graphical displays provide project transparency, especially in large-scale HIL projects. In addition to the model and the I/O configuration, external devices such as ECUs and loads, including their signal properties (descriptions, electrical properties, failure simulation settings, and load settings) can be defined and documented. Dedicated tools such as ConfigurationDesk allow the documentation to be generated in Microsoft® Excel® files (as required when testing according to ISO 26262).

To support distributed teamwork, it is essential that the required HIL hardware for a specific project can be assembled and configured offline as a "virtual system," in other words, as a purely software-based configuration. A real-time application needs to be executed for test runs even if parts of the necessary (and configured) I/O hardware are not physically available.

3.2.3 *Real-time software in general*

As mentioned earlier, the software running on the real-time processor is usually C code generated from the block-based graphical simulation environment or the physical acausal graphical simulation environment by means of automatic code generation. The currently most widely used tool for graphical simulation is MATLAB®/Simulink®. However, physical models are also becoming more important.

As a rule, the programs run in a real-time frame that should enable real multitasking as well as supporting timer, software, and hardware interrupts (Plöger and Köhl, 2006).

3.2.4 *Human–machine interface*

The manual operation of the test system, visualization and storage of real-time variables, and adjustment of model parameters are usually done on a host PC with the help of graphical front-end tools as the simulator's interactive experiment and visualization interface. These graphical user interfaces (such as ControlDesk Next Generation from dSPACE) are used for managing experiments, projects, and platforms. Virtual instruments allow visualization of a virtual instrument panel and access to the parameters of the real-time simulation model for manual or automatic modifications. The important issues include synchronous measurement and/or recording on all data sources (i.e., HIL simulator and ECU), and integrated measurement data analysis.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Typically, the same tool is used for layouting, logging, monitoring, and replaying the bus communication. CAN messages, LIN frames, and FlexRay PDUs (protocol data units) can be handled with modules such as the Bus Navigator from dSPACE. Instruments are available to view RX messages/frames/PDUs and to configure TX messages/frames/PDUs. For detailed tests on the control unit's bus communication, the messages, frames, and PDUs can be manipulated before transmission, or excluded from being transmitted.

The failure insertion unit of the HIL simulator is also remote-controlled via the experiment software. Failures in the wiring of an ECU such as an ECU pin short-circuited to the ground or the battery voltage are enabled and disabled.

Purely numeric displays and simple plotter instruments are no longer sufficient, particularly for testing vehicle dynamic functions or adaptive cruise control (ACC) functionality. A 3-D real-time animation tool (such as MotionDesk from dSPACE) is a far better way to visualize a vehicle's movements and assess them at a glance.

3.2.5 Software support for automated tests

The potential of HIL simulation cannot be fully harnessed without systematic, complete automation of all components. Development of the necessary test specifications begins with ECU or function specifications and system specifications (Sax and Schmerler, 2004). Thus, the HIL testing of ECUs is usually requirements-based testing and uses functional or black box tests.

The tasks involved in HIL simulation are extremely varied and require flexible and versatile means of test description. The past few years have, therefore, seen the arrival of numerous tools and solutions that address the issue of test description in different ways.

Script languages are very widely used in connection with automated testing on HIL simulators. For example, there are numerous libraries (APIs, application programming interfaces) for automating HIL simulators with the Python script language (www.python.org). Test scripts can be implemented in Python to access the different test system components such as the real-time model, failure simulation hardware, diagnostics, and calibration systems, via the APIs.

In 2009, the ASAM (Association for Standardization of Automation and Measurement Systems) Board released version 1.0 of the HIL API as a common standard to access HIL systems from test automation systems (http://www.asam.net/).

Script languages are extremely flexible to use, but a comparatively long time is required to learn them. Thus, they are usually suitable for test developers but not for testers. For this reason, several solutions have been developed to provide easier user access to specific test types. This means that testers do not have to bother with the script language itself, but can parameterize their tests conveniently with the input options provided by the front end.

Nowadays, however, even test development does not have to be performed exclusively by scripting. Modern tools such as AutomationDesk (dSPACE GmbH, 2004) provide graphical test description, like that from the Unified Modeling Language (UML) (www.uml.org), as an additional option.

The exact, automatable reproduction of test scenarios is the basic reason why HIL simulation boosts efficiency (Wallentowitz and Reif, 2006).

4 INTEGRATION INTO THE ELECTRONICS DEVELOPMENT PROCESS

For numerous car manufacturers, successful tests on the HIL simulator are one of the preconditions for a vehicle's release approval.

4.1 Goals

The later in the development process that an error is found, the more it costs (Figure 6). Problems can be detected and solved at an early stage by systematically using HIL test systems as early as possible. In the development process. In this way, the greatest possible vehicle maturity can be achieved quickly (Honisch, Hutter, and Schmid, 2004).

As a precondition for speedy error remedies, the development, commissioning, and verification of the ECUs must be closely interlinked. For example, when new software versions come along, information on the implemented changes has to be passed on, so that tests can be restricted to essential features, and also be extended if required (Plöger and Köhl, 2006).

4.2 HIL process

For efficient integration into the overall development process, HIL simulation must be treated as a subprocess. This subprocess begins by specifying the required HIL system based on the information that is available on the vehicle and the ECUs installed in it. Following preparatory assembly of all the ECU-independent components, the simulator can be adapted to the ECUs without losing time as soon as the ECU specifications are available. This includes activities such as implementing load simulations

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

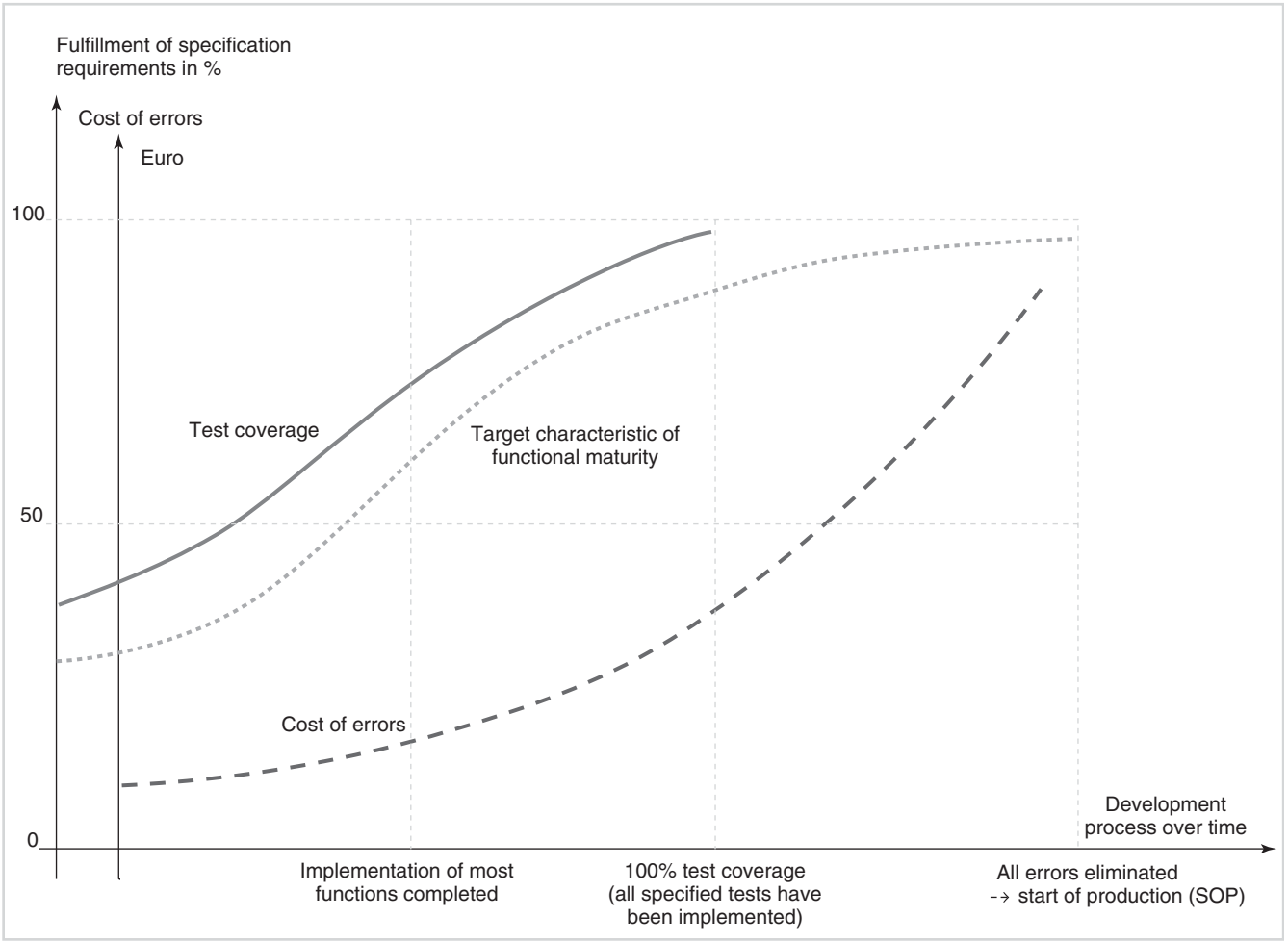


Figure 6. Functional Maturity in the electronics development process (Plöger and Köhl, 2006). (Reproduced by permission of Springer Vieweg.)

and preparing ECU mounts. It is still possible to perform iterations when ECU specifications change. After being put into operation with the defined number of ECUs, the simulator is ready and the test objects can be tested. The tests to be executed later can be developed in parallel to construction of the test system, using the ECU specifications as a basis.

When an ECU is updated, the simulator is once more put into operation (after any necessary adaptation has been performed) and regression tests are executed.

Test schedules and test cases are developed on the basis of the ECU/function specifications. In addition to the requirements documents, constant extension of the test suite also comes from test cases that were already implemented for predecessor ECU or vehicle models and from the developers' experience. Feedback from trials or service dealers is also used. Test coverage is not 100% in relation to the defined requirements until the second half of the development cycle (Figure 6) (Plöger and Köhl, 2006).

4.3 HIL testing in the context of ISO 26262

ISO 26262 will become the generally and globally accepted, automotive-specific standard and define what is "state of the art." Thus, it is highly relevant for the HIL testing of software for automotive ECUs, namely functionality, functional safety, diagnostics, and communication. ISO 26262-6 recommends HIL testing as a test method but does not state how to verify that a specific HIL system is suitable for testing safety-related functions.

Appropriate measures must be taken to ensure the proper operation of a HIL system. These measures can be based on detailed knowledge in HIL systems and procedures agreed

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

upon with long-term experienced users of HIL systems. The suitability of a HIL system for use in a specific safety-related project is verified by testing the entire system. This is required for the first time after system assembly and recurrently for a system in use.

The suitability of a tool and its automotive safety integrity level (ASIL) classification can be verified in consultation between customer and tool vendor. This is always carried out for a specific safety-related project.

It is not necessary to verify the suitability of software tools for a HIL system according to ISO 26262.

4.4 Integrating HIL simulation into the development and test process

Vehicle development generally begins 30 months before the planned SOP (start of production) date or earlier. Software development often begins at the same time as function specification, so it is a frontloaded design and preparation phase within the vehicle development process. HIL testing in software development is a firm component in ECU integration and release tests. Car manufacturers who have incorporated HIL simulation into their processes in this way have numerous HIL simulators constantly in action, testing subsystems and systems for the product range in active development. These systems typically work 24 h a day, 7 days a week (Plöger and Köhl, 2006).

Figure 6 shows the targeted functional maturity over the period of a development cycle. Functional maturity means the percentage of specification requirements that were tested "OK." When production begins, there should be no more errors present (100% maturity). If all the specified tests have been implemented for the defined requirements (100% test coverage), a high level of automation can be achieved. Now, all errors are eliminated in succession until there are no more known errors at the time of production start-up.

4.5 Connecting to requirements management

Tools such as DOORS or TestDirector are frequently used to collect and manage requirements and specifications. They are also very good for managing test specifications and results. The same applies to test plans and test execution. Direct connection to the HIL test environment (as is available with AutomationDesk) is, therefore, a logical step in process integration (Plöger and Köhl, 2006).

5 OUTLOOK

As electronics and software continue to grow, advanced, new systems will be developed. Also, complexity is constantly growing. One well-known way to prevent complexity leading to chaos is to use methods and tools, particularly HIL tests and model-based designs. These fields have witnessed an enormous amount of change over the past 10 years (Hanselmann, 2012).

Nowadays, whole vehicles are tested thoroughly using network simulators, and HIL tests for single ECUs and subsystems have been standard procedures for a long time. Entire ECUs are now developed using model-based design, and the ECU code is largely generated automatically (Hanselmann, 2012).

It is no longer individuals who perform these tasks, but whole departments and separate teams both within one company and, partly, in more than one company. Those in charge report having to handle up to 100,000 HIL test cases, 40,000 calibratable ECU parameters, or 400 vehicle variants. The volume of data generally created during ECU development, especially when methods and tools are used, is immense and still increasing. The ability to reuse this data in other projects, by other people or for other variants of the functions or systems under development, is a major factor in efficient development and also helps ensure quality (Hanselmann, 2012).

The precondition for reliable reuse is the central management of models, tests, parameters, and architectures, and of development results and statuses in general, including the underlying requirements. The numerous interdependencies and connections between all these data also have to be considered. It is no longer possible to master this management task solely using central file storage or configuration management systems: file-based management is not only too coarse-grained, it also does not provide sufficient traceability for observing dependencies (Hanselmann, 2012).

In practice, at dSPACE, we see tasks such as test management, model management, central parameter and signal management, and HIL hardware configuration management, in the context of the numerous variants that are created in the automotive industry, and in conjunction with the associated requirements (Hanselmann, 2012).

5.1 Test management

We understand test management not only as the general management of tests, test scripts, and test results but also test planning and the representation and analysis of test results. It also includes connecting to test execution, for example, by directly coupling test automation tools in the HIL environment (Hanselmann, 2012).

Test management must also create links to models, parameter sets, and variants.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5.2 Model management

In the development of complex systems, constructing giant, monolithic models quickly proves to be a dead end. That is why modularity is particularly important. It is a precondition for teamwork and the reuse of (sub)models in other projects, by other people, or for new variants. However, all this inevitably produces an even larger number of models that have to be handled together with their interrelationships.

Clean descriptions of the interfaces are also necessary to achieve modular structures and reuse individual submodels. The descriptions must, therefore, not be created from a local viewpoint, but coordinated on a global level (Beine, 2012).

Model management involves the central management of function models and plant models for real-time and offline simulations alike. It is closely connected with central parameter and signal management. The managed parameters are used to provide values to functions, models, and ECU code. The interfaces of systems, functions, models, and ECUs can be described by means of signals, laying the foundation for integrating individual submodels in an overall model (Hanselmann, 2012).

5.3 Data management

However, viewing these topics separately, mastering the individual stages in the development process and increasing the quality and productivity of each stage, is not enough. The overall goal is efficient, integrated management of the enormous volumes of complex and interdependent data (Hanselmann, 2012).

This management must cover all the process steps and all the individuals and teams who are users and producers of the data. In practice, however, it is often the case that only solutions for individual issues and for small teams are needed initially. In such cases, it is important to select a solution that is scalable and usable for other tasks. It must also be capable of being extended to cover the entire development process, and in connection with this, entire departments and companies (Hanselmann, 2012).

Using our years of experience and knowledge from countless ECU development projects, we (dSPACE) aim to give teams distributed worldwide the support they need throughout the entire E/E development cycle, based on an established data management platform. We will support the relevant standards and build a bridge to established development tools from dSPACE, and from other vendors as well (Hanselmann, 2012).

5.4 Virtual ECU testing

Another issue that we are intensively focusing on is frontloading, in other words, testing virtual ECUs.

Today's ECU software comprises numerous SWCs with intensive interactions. In the large ECU networks frequently installed in current vehicles, the number of SWCs can easily reach the thousands (http://www.dspace.com/en/inc/home/products/systems/virtual_ecu_testing.cfm).

In addition, because the task of developing ECU components is usually shared by several departments or even different companies, not only the SWCs themselves have to be tested and validated but also the interactions between them. The earlier in the development process that errors and inconsistencies are found, the quicker and cheaper it is to correct them (http://www.dspace.com/en/inc/home/products/systems/virtual_ecu_testing.cfm).

Virtual ECU testing means implementing early verification strategies in the ECU development process by means of more intensive virtualization—the current trend. At the heart of these verification strategies are the so-called virtual electronic control units (V-ECUs), which despite being pure software prototypes can realistically simulate the functional behavior of real ECUs including the operating system, task scheduling, and hardware-independent basic software parts. These V-ECUs can be used for initial (and automatable) integration testing by offline simulation on a PC very early in the development process, for high quality restbus simulation on a HIL simulator, as substitutes for ECU hardware prototypes that are not yet available during commissioning, and even in integration testing itself (Krügel, Geburzi, and Krisp, 2011; Krügel and Schulze, 2012; Krügel *et al*., 2012).

6 CONCLUSION

Vehicle electronic systems are becoming more advanced and complex. HIL simulation for testing electric/electronics has played an important role in the development process for many years and will continue to do so. In the same way as software and model complexity is growing, so is users' need for well-engineered procedural methods for handling large models and comprehensive software structures. HIL tests and model-based design are meanwhile being supplemented by the frontloading of tests, which means performing HIL tests earlier in the development process and also virtualizing HIL tests. Both these aspects enable users to reduce development time in spite of increasing complexity, while at the same time maintaining or even enhancing the quality of the electronics.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ACKNOWLEDGMENTS

This chapter is based in part on the article "Hardware-in-the-loop-Software" in Wallentowitz and Reif (2006). Text and figures were reproduced with permission.

RELATED ARTICLES

Active Safety, Pre-collision Safety and Other Safety Products (millimeter wave, image recognition, laser)
AT Control—Actuation Methods & System Integration, Gear Choice, Gear Shift Strategy & Process, Adaptive Features
Body and Lighting ECU (Key-less Entry, Sonar, HID, LED Usage for Lamps)
Body ECU (airbag)
Body ECU Cluster
Car air conditioning and electronics: analog, digital control and zone management
Chassis Control Systems—A Look into the Future
Chassis ECU (ACC and sensor)
Chassis ECU (Vehicle dynamics, ABS)
Control Systems and Strategies for Automated Manual and Double Clutch Transmissions
CVT Control—system integration, ratio choice, shift dynamics & strategy, adaptive features, engine calibration, em assist
Diversification of electronics and electrical systems and the technologies for the integrated systems
ECU Chassis (Steering)
Engine ECU systems
Engine Management Systems
Hybrid Systems and High Voltage Components
In-Vehicle Network
Interfaces between Sensors and ECUs
Semiconductor Sensors (3): Optical Sensors
Various types of sensors
Vehicle Safety, Functional Safety, OBD Diagnosis

REFERENCES

Beine, M. (2012) Die Modellvielfalt Verwalten, AUTOMOBIL ELEKTRONIK.

dSPACE GmbH (2004) Produktinformationen zu AutomationDesk, TargetLink, MTest.

dSPACE GmbH (2011) Produktinformationen zu SCALEXIO.

Gehring, J., Schütte, H. (2002) Automated Test of ECUs in a Hardware-in-the-Loop Test Bench for the validation of Complex ECU Networks. *Proceedings of the SAE World Congress*, Detroit, USA.

Hanselmann (2012) Elektronik Automotive 2012, E/E Data Management Instead of Chaos.

Honisch, A., Hutter, A., and Schmid, H. (2004) Vollautomatisierter Test von Steuergeräte-Netzwerken, ATZ-MTZ extra Mercedes-Benz A-Klasse.

ISO 26262 *Road Vehicles—Functional Safety*, www.iso.org.

Köhl, S., Lemp, D., and Plöger, M. (2003) *Steuergeräte-Verbundtests mittels Hardware-in-the-Loop Simulation*, ATZ, Wiesbaden, Deutschland, pp. 948–955.

Köhl, S., and Jegminat, D. (2005) How to do hardware-in-the-loop-simulation right. SAE-Paper No. 05AE-279, Detroit, USA.

Krügel, K., Stockmann, L, Holler, D., and Lamberg, K. (2012) IAV—Simulation und Test für die Automobilelektronik IV. Simulation-based development and testing environment for electric Vehicles.

Krügel, K., Geburzi, A., and Krisp, H. (2011) *The Next Generation of Validation*, Hanser Automotive, Munich.

Krügel, K., and Schulze, T. (2012) *Modellbasierte Funktionsentwicklung von Steuergeräte-Funktionen für Motoren mit vollvariablem Ventiltrieb*, MTZ, Wiesbaden.

Lamberg, K. (2006) Software-Testen in *Handbuch Kraftfahrzeugelektronik* (eds H. Wallentowitz and K. Reif), Vieweg und Teubner Verlag, Wiesbaden.

Plöger, M., and Köhl, S. (2006) Hardware-in-the-Loop-Software in *Handbuch Kraftfahrzeugelektronik* (eds H. Wallentowitz and K. Reif), Vieweg und Teubner Verlag, Wiesbaden.

Sax, E., and Schmerler, S. (2004) Qualitätssteigerung in der KFZ-Steuergeräte-Entwicklung. Test Guide Design and Verification.

Schäuffele, J., and Zurawka, T. (2003) *Automotive Software Engineering*, ATZ-MTZ-Fachbuch, Wiesbaden.

Schulte, T., Kiffe, A., and Puschmann, F. (2011) HIL simulation of power electronics and electric drives for automotive applications. 16th International Symposium on Power Electronics Ee, Novi Sad, Republic of Serbia.

Schütte, H., Plöger, M., Diekstall, K., *et al*. (2001) *Testsysteme im Steuergeräte-Entwicklungsprozess, Automotive Electronics*, pp. 16–21.

Wallentowitz, H. and Reif, H. (eds) (2006) *Handbuch Kraftfahrzeugelektronik*, Vieweg, Wiesbaden.

FURTHER READING

Lamberg, K (2001) Methodik zur systematischen Bereitstellung von HIL-Testsystemen für Kfz-Steuergeräte. Dissertation. Herbert Utz Verlag.

Wältermann, P., Schütte, H., and Diekstall, K. (2004) *Hardware-in-the-Loop Test verteilter Kfz-Elektroniksysteme*, ATZ, Wiesbaden, Deutschland, pp. 416–425.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto193
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Vehicle Safety, Functional Safety, OBD Diagnosis

Stefan Kriso, Matthias Klauda, and Reinhold Hamann
Robert Bosch GmbH, Abstatt, Germany

1 Introduction 1
2 Definitions 1
3 Safety of Road Vehicles 1
4 Functional Safety 3
5 Side Issues of Safety 12
6 Summary and Conclusion 15
Acknowledgments 15
Endnotes 15
References 16
Further Reading 16

1 INTRODUCTION

With a significantly increasing number of electrical and electronic systems within road vehicles, the complexity of such systems dramatically increases. To prevent these increasingly complex systems from becoming more and more fault-prone is one of the most important challenges of the future. Systems have to be intrinsically safe whereby owing to the cost pressure within the automotive industry, it is not possible just to double or triple the systems, as it is usual in the avionics industry. This chapter describes the general concept of vehicle safety—especially with regard to functional safety that came into focus within recent years with the advent of a new standard for functional safety of road vehicles—the ISO 26262.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 DEFINITIONS

2.1 Risk

Combination of the probability of occurrence of harm and its severity.

2.2 Safety

Absence of unreasonable risk, which is judged to be unacceptable in a certain context according to valid societal moral concepts.

2.3 Functional safety

Absence of unreasonable risk due to hazards (potential source of harm) caused by malfunctioning behavior of electrical/electronic systems.

3 SAFETY OF ROAD VEHICLES

3.1 Safety relevant systems in road vehicles

The number of software intensive systems in road vehicles is increasing and becomes more and more important for future innovations. Today, they can be found everywhere in a modern car, for example, powertrain, entertainment, driver assistance, stability control, body applications, and passive safety. Many electrical/electronic systems within road vehicle can be assumed as safety relevant, that is, a malfunction of such a system can lead to harm of an involved person. Some examples for such malfunctions are

- stability control: unintended brake force up to almost locking pressure of one front wheel;

- engine management: unintended vehicle acceleration;
- airbag: faulty activation during driving;
- electric window lift: unwanted closing of a window; and so on.

A systematic determination leads to more than 1000 safety-relevant functions of electrical/electronic automotive systems.

3.2 Trends

3.2.1 E-mobility

The most upcoming trend today is e-mobility, that is, the electrification of the powertrain. This leads to new and partly unknown aspects of the safety of a vehicle:

- An electronic control unit of a conventional vehicle has about 10,000 operating hours, which corresponds to the driving hours of the vehicle. In an e-vehicle, the operating time is dominated by charging, which takes about 30,000–50,000 h. So new risks come up from the charging procedure and energy storage.
- High voltage within a car brings new risks for people maintaining and repairing the vehicle. This will lead to requirements for occupational health and safety procedures as well as for operating, service, and decommissioning of vehicles.
- Not only risks from electrical or electronic systems may arise, but also chemical risks from the battery ("thermal runaway") or risks for the environment (e.g., acid disposal) arise.

3.2.2 Miniaturization

Owing to the miniaturization of electronics devices, technology might reach its limits regarding capability to produce fault-free devices in a cost-effective way under given automotive quality requirements:

- Reliability may decrease with decreasing structure size;
- Amount of embedded logic and memory per device will increase, which may lead to an increase of the failure rate of the device (as the failure rate per bit is—in the best case—constant);
- Operating hours will increase (e-mobility).

In the end, the faults of semiconductors might become more and more common. But, as the safety on system level has to be reached anyway, the system design has to deal with this situation to ensure the overall level of system safety.

3.2.3 Variants

The number of variants of modern vehicles is more and more increasing. Software intensive systems allow vehicles to be individually customized, for example, just by activating/deactivating functionality (e.g., driver drowsiness detection is pure software function that can be located on different electronic control units in the vehicle). Vehicles provide more and more open interfaces to "external" systems, so that functions can be downloaded by the user (like a Smartphone app). This leads to huge number of variants with a lot of possible combinations of features. It must be ensured that all these possible feature combinations are safe and reliable. In this area, we see a correlation to the area of security that needs to provide measures to ensure the integrity of vehicle systems against malicious IT attacks and thereby safeguard the vehicle.

3.3 Special constraints of the automotive industry

To ensure safety of road vehicles, the special boundary conditions of the automotive industry have to be considered.

3.3.1 Driving cycle

A typical driving cycle of a road vehicle consists of three phases with different possibilities to ensure the safety of the system:

- Start-up: enhancing testing possible, increased test coverage. Can include "destructive" tests like memory checks.[1] The start-up time is strongly limited, for example, 50 ms for communication tests. Typically, all safety relevant measures are to be checked during start-up.
- Driving: often only limited nondestructive tests are feasible. The functionality of the system must be maintained and monitored, functional restrictions like interrupt blocking times must be considered.
- Shut-down: similarly to start-up, some safety functionality tests may be shared between start-up and shut-down. The shut-down cycle may be interrupted at any time.

3.3.2 Cost structure

The automotive industry is very cost sensitive because of the high production volume with small margins. Low unit cost for electronic control units translates into restricted memory size and limited computing capacity (Schäuffele and Zurawka, 2005). Consequently, strong reliance on

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

redundancy on system level as a safety concept (e.g., doubling or trebling of electronic control units) found elsewhere for example, in airborne systems, is not feasible within automotive systems.

3.3.3 Maintain and repair, field observation

Cars are sold to end customers, where it cannot be foreseen which kind of maintenance and repair strategy they have. So it is very difficult (or impossible) for a vehicle manufacturer to influence the maintenance and repair behavior of the user. Systems must be safe even if the end customer is not willing to maintain and repair their car regularly. Preventive maintenance as can be found in airborne systems is normally not very successful, as it cannot be ensured that all vehicles in the field will be included.

In addition, it is nearly impossible to get statistically significant data from field observation, even during the warranty time of the car. So, an evaluation of the safety of a vehicle system based on commonly recognized industry data basis would lead to overdesigned systems, as the data basis contains safety margins so that the data are normally too pessimistic.

3.3.4 Education and training

Vehicle users (drivers) are mostly educated and trained once—at the beginning of their "driver's career" when they earn their driver's license. Training with malfunctioning equipment such as in airborne industry does not take place. In most cases (except, perhaps, drivers of buses or similar), there is no requalification foreseen. The safety concept of a vehicle must consider this and different capabilities of drivers. In ISO 26262, this is directly mapped to the controllability of malfunctions (Section 4.4.3). This is not easy to determine as the capabilities of the "average driver" especially in highly critical situations are not well known and may differ regionally.

4 FUNCTIONAL SAFETY

4.1 ISO 26262: new standard for functional safety of road vehicles

ISO 26262 was published in November 2011 and addresses the functional safety of road vehicles. As it is the automotive specific adaptation of the IEC 61508 (functional safety of electrical/electronic/programmable electronic systems) it is going to become the most important safety standard within the automotive industry. Its introduction [*] states:

> *ISO 26262 is the adaptation of IEC 61508 to comply with the needs specific to the application sector of electrical and/or electronic (E/E) systems within road vehicles. This adaptation applies to all activities during the safety life cycle of safety-related systems comprised of electrical, electronic, and software components.*
> *[Functional] safety is one of the key issues of future automobile development. With the trend of increasing technological complexity, software content, and mechatronic implementation, there are increasing risks from systematic failures and random hardware failures. ISO 26262 includes guidance to avoid these risks by providing appropriate requirements and processes. ISO 26262*
>
> 1. *provides an automotive safety lifecycle and supports tailoring the necessary activities during these life cycle phases;*
> 2. *provides an automotive-specific risk-based approach to determine integrity levels [automotive safety integrity levels (ASILs)];*
> 3. *uses ASILs to specify applicable requirements of ISO 26262 so as to avoid unreasonable residual risk;*
> 4. *provides requirements for validation and confirmation measures to ensure a sufficient and acceptable level of safety being achieved; and*
> 5. *provides requirements for relations with suppliers.*
>
> *Figure* 1 *shows the overall structure of ISO 26262. It is based on a V-model as a reference process model for the different phases of product development.*

As ISO 26262 is very complex and addresses a lot of different issues, we focus only on some key issues of this standard in the following sections.

4.2 Legal point of view of ISO 26262

4.2.1 Introduction

The summary given here refers to the specific legal situation in Germany. Nevertheless, it may help to gain a better understanding of the legal background of ISO 26262, as the basic principles on product liability are very similar worldwide.

Technical standards are recommendations drawn up by acknowledged experts to be used as technical, best-practice solutions for sector- or product-specific problems. These recommendations also establish minimum safety requirements for the products in question. The fact that technical standards are to be regarded as pure recommendations without binding legal force does not in any way mean that they are of no legal consequence. On the contrary, compliance with a legally nonbinding recommendation is a crucial yardstick for all four areas of law dealing with liability for defective products:

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

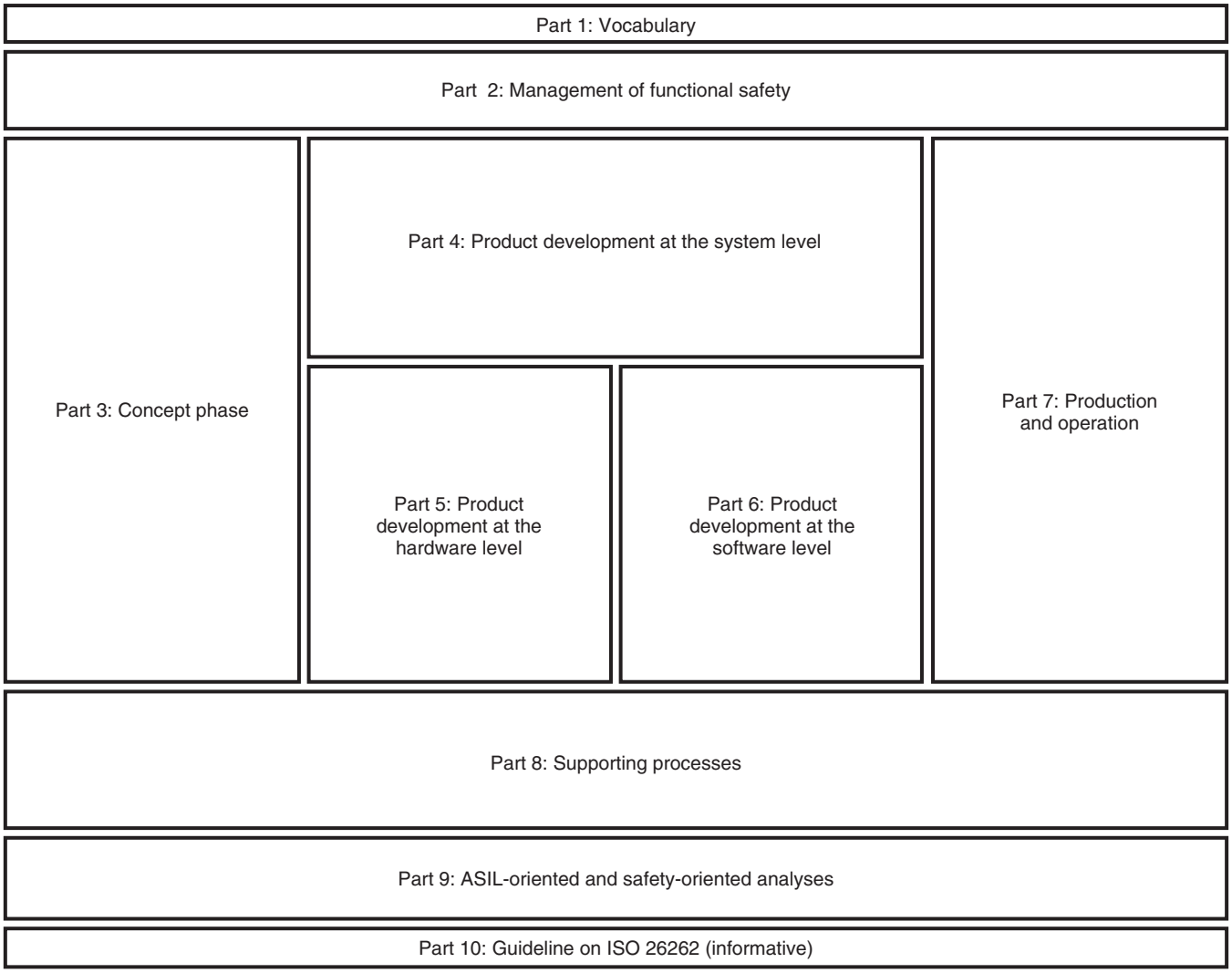


Figure 1. Overview of ISO 26262 [*]. (Reproduced by permission of DIN Deutsches Institut für Normung e.V.)

4.2.2 Contract law—warranty

Whether the obligations of the seller under a sales contract have been properly met initially depends, under Section 434 I of the German Civil Code (BGB, Bürgerliches Gesetzbuch), on whether the product sold has the properties as agreed between the seller and the purchaser. The terms of tender and/or the general conditions of purchase within the automotive industry regularly contain an obligation on behalf of the supplier to comply with all the technical standards relevant to the delivered item. This makes ISO 26262 one of the agreed properties once the contract has entered into force—even if this standard was not expressly mentioned in the tender. Even if compliance with the applicable standards has not been explicitly agreed as a requirement, ISO 26262, in its capacity as the key standard on functional safety in the automotive industry, becomes an integral part of the supply contract via Section 434 I No. 1 or 2 of the German Civil Code. If the supplier is unable to prove that he or she has met the requirements arising from ISO 26262, the product concerned is deemed to be nonconforming. As it is virtually impossible to meet the requirements of ISO 26262 after the fact, the OEM (original equipment manufacturer) customer may reject corresponding deliveries as noncompliant. The supplier is then liable without limitation for all subsequent costs caused by the deficient delivery.

4.2.3 Product liability

The legal duty to maintain safety, as far as product liability is concerned, requires (in accordance with established

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

case law) that the manufacturer employs due diligence in keeping with the state of technology at the time the products are brought into circulation, so as to ensure that the products are as safe as might legitimately be expected. In the event of an accident involving personal injury, the manufacturer of the product that caused the accident must prove that he or she has exercised due diligence in the normal course of business. If the manufacturer cannot even prove that he or she has at least fulfilled the minimum requirements of applicable technical standards, then the manufacturer has no realistic chance of defending himself or herself.

4.2.4 Public safety

The authorities are responsible for ensuring that public safety is not endangered by unsuitable products. The German Equipment and Product Safety Act give the authorities wide-ranging powers to intervene in order to fulfill this obligation, up to and including the closure of companies. An important criterion for the authorities in determining whether to intervene or not is whether the manufacturer has complied with the applicable standards. If he or she is unable to prove that he or she is in compliance when problems emerge in the market, then this would be grounds for action on the part of the authorities.

4.2.5 Criminal proceedings

If a product defect has led to an accident that caused severe injuries or even death, the public prosecutor's office must check whether a criminal offense has been committed. If the product that caused the accident does not meet the minimum requirements of the applicable standards, then this can be the trigger for criminal proceedings against the manufacturer's employees responsible.

In summary, it therefore can be said that technical standards indeed might not be legally binding in that they are nothing more than recommendations. However, they develop enormous legal significance as a yardstick for the care due in the normal course of business.

4.3 Key concept of ISO 26262

Goal of the activities that are addressed in ISO 26262 is the prevention of malfunctioning behavior of electrical and electronic systems. After defining the "item" (system or function under consideration), its malfunctions have to be determined and their impact on the safety of the involved persons (driver, passenger ...) has to be evaluated. This is done during the hazard analysis and risk assessment (Section 4.4) and leads for each malfunction to an "ASIL".

In a mathematical notation, the safety relevance of a function or a system can be expressed by the following term:

$$\text{Safety relevance} \in \{\text{QM}, \text{ASIL} | \text{ASIL} \in \{\text{A}, \text{B}, \text{C}, \text{D}\}\}$$

The higher the ASIL is, the higher will be the necessary risk mitigation, that is, for an ASIL D malfunction more risk mitigation measures are necessary than for an ASIL A malfunction. Not safety-related malfunctions are classified as "QM," that is, a standard quality management for example, according to ISO/TS 16949 is sufficient for the development of this function. In this case, there are no further requirements by ISO 26262.

In case of a given safety relevance of the item (safety relevance $\neq$ QM), the risk has to be reduced to a reasonable level. The measures to reach this are described in ISO 26262; the requirements in ISO 26262 and the proposed methods are dependent on the ASIL and therefore dependent on the initial risk outgoing from the item.

There are two kinds of sources of risks: systematic failures and random hardware failures. Both have to be considered in the development of the item. As systematic failures are mainly introduced during the development of the item, ISO 26262 defines a safety life cycle with requirements for the management of functional safety (for the project independent organization as well as for the project developing the item and for the phases after release for production), for the development process of the item, for the production, operation, service, and decommissioning phases, and for the supporting processes (e.g., interfaces within distributed development, requirements management, configuration management, change management, verification, and documentation).

Not all potential systematic failures can be avoided with the mentioned requirements for the safety life cycle. So, the product must be able to control the remaining systematic failures (e.g., errors in a piece of software) as well as random hardware failures. So, ISO 26262 gives requirements for development and test on system level, on hardware level (e.g., introduction of redundancies), and on software level (e.g., diagnosis).

In the end of the development phase of the item, an overall safety validation ensures the achieved functional safety of the system at vehicle level. This includes for example, the controllability of the malfunctions of the item and the effectiveness of safety measures for controlling random and systematic failures. In addition, measures to evaluate the implementation of the process requirements are introduced, see Section 5.1.

Figure 2 summarizes the key concept of ISO 26262.

In addition to the requirements, ISO 26262 gives a lot of recommendations for the development of functionally

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

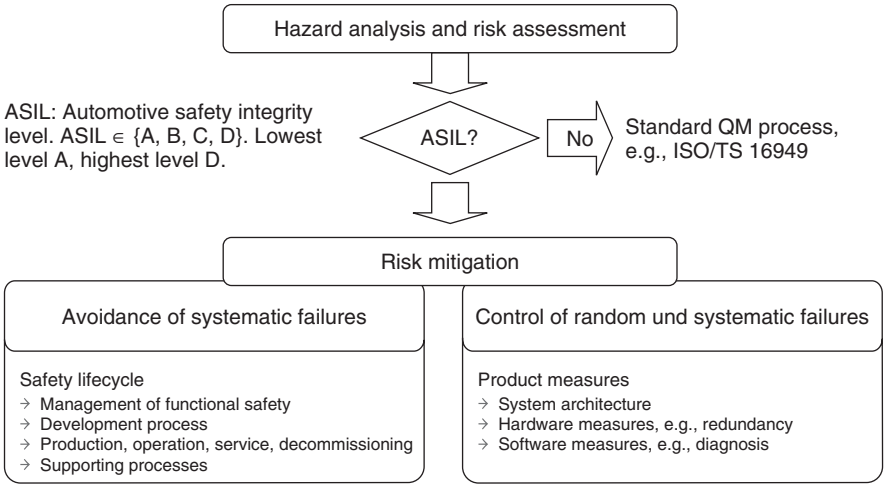


Figure 2. Key concept of ISO 26262.

safe products. These helpful recommendations may be implemented, but they are not necessary to proclaim conformance with the standard. Also, the requirements itself are not fully obligatory, as it is allowed to omit them based on a documented justification: "*When claiming compliance with ISO 26262, each requirement shall be complied with, unless one of the following applies: [. . .] b) a rationale is available that the noncompliance is acceptable and the rationale has been assessed in accordance with ISO 26262-2*." [*]

As a result of this, a one-to-one implementation of ISO 26262 is not required, and different organizations will find different ways of implementation of the standard.

4.4 Hazard analysis and risk assessment, ASIL classification

4.4.1 Introduction

Hazard analysis, risk assessment, and ASIL determination are used to determine the safety goals of the item such that an unreasonable risk is avoided. The scheme of ISO 26262 for hazard classification recognizes that a hazard in an automotive system does not necessarily lead to an accident. The outcome will depend on whether the persons at risk are actually exposed to the hazard in the situation in which it occurs, and whether they are able to control the outcome of the hazard. Figure 3 gives an example of this concept applied to a failure that affects the controllability of a moving vehicle.

The item is evaluated with regard to its potential hazardous events. Safety goals and their assigned ASIL are determined by a systematic evaluation of hazardous events. The ASIL is determined by considering the estimate of the

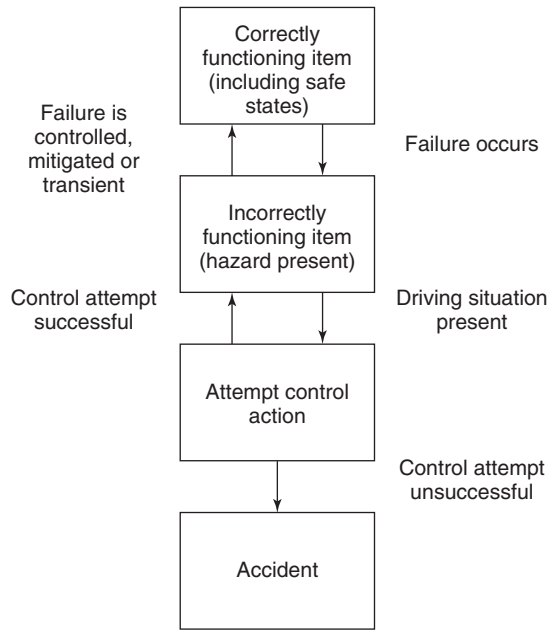


Figure 3. State machine of automotive risk [*]. (Reproduced by permission of DIN Deutsches Institut für Normung e.V.)

impact factors, that is, severity, probability of exposure, and controllability. It is based on the item's (i.e., system's) functional behavior; therefore, the detailed design of the item does not necessarily need to be known.

The detailed process to determine the ASIL is described in ISO 26262, part 3, Chapter 7. Figure 4 summarizes this procedure.

For each item (system under consideration), the potential malfunctions have to be determined. In a situation analysis,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

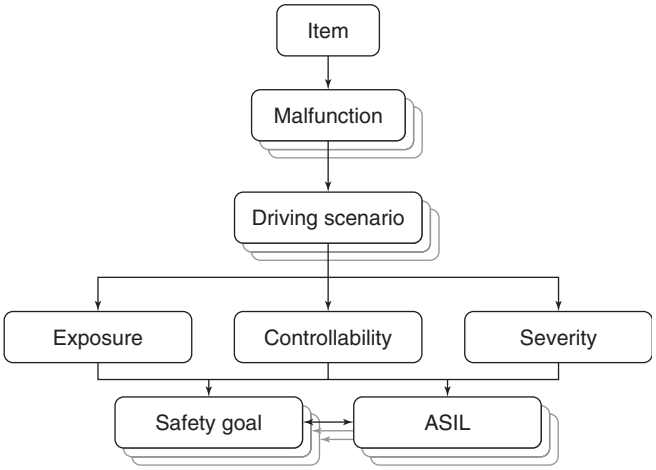


Figure 4. Hazard analysis and risk assessment according to ISO 26262.

the operational situations and operating modes in which the item's malfunctioning behavior will result in a hazardous event shall be described. These driving scenarios shall consider the correct usage of the vehicle and its incorrect usage in a foreseeable way.

In the subsequent hazard identification, the hazards are described and classified in a systematic way. For the classification of hazards, the following parameters are used:

4.4.2 Exposure (E)

To answer the question "How often does the driving situation occur?" the probability of hazardous driving scenario has to be estimated and classified into one of five probability classes according to Table 1.

It is important to know that only the probability of the situation is of interest, independent from the probability of the malfunction itself. For instance, the malfunction "unwanted deployment of an airbag" has a very low probability, but the driving scenario "driving on a highway," where the malfunction could become hazardous, has—according to ISO 26262, part 3, table B.2—a very high probability and therefore the exposure E4.

4.4.3 Controllability (C)

The next question is how the malfunction within the driving scenario can be controlled by the driver or by any other involved person. This controllability parameter shall be classified using Table 2.

It is assumed that the driver of the vehicle is in an appropriate condition to drive (e.g., he or she is not tired), has the appropriate driver training (he or she has a driver's license), and is complying with all applicable legal regulations, including due care requirements to avoid risks to other participants. Reasonably foreseeable misuse has to be considered, for example, driving at 60 km/h when the speed limit is 50 km/h.

4.4.4 Severity (S)

If a malfunction occurs and it cannot be controlled, then it is the question about the severity of the resulting hazardous event. The severity parameter is classified according to Table 3.

The focus is on the harm to each person potentially at risk, including the driver or the passengers of the vehicle causing the hazardous event, and other persons potentially at risk such as cyclists, pedestrians, or occupants of other vehicles.

4.4.5 ASIL determination

After determining the parameters E, C, and S, the ASIL can be simply determined using Table 4.

Table 1. Classes of probability of exposure regarding operational situations (ISO 26262-3, Table 2).

	Class				
	E0	E1	E2	E3	E4
Description	Incredible	Very low probability	Low probability	Medium probability	High probability

Reproduced by permission of DIN Deutsches Institut für Normung e.V.

Table 2. Classes of controllability (ISO 26262-3, Table 3).

	Class			
	C0	C1	C2	C3
Description	Controllable in general	Simply controllable	Normally controllable	Difficult to control or uncontrollable

Reproduced by permission of DIN Deutsches Institut für Normung e.V.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Classes of severity (ISO 26262-3, Table 1).

	Class			
	S0	S1	S2	S3
Description	No injuries	Light and moderate injuries	Severe and life-threatening injuries (survival probable)	Life-threatening injuries (survival uncertain), fatal injuries

Reproduced by permission of DIN Deutsches Institut für Normung e.V.

Table 4. ASIL determination (ISO 26262-3, Table 4).

Severity Class	Probability Class	Controllability Class		
		C1	C2	C3
S1	E1	QM	QM	QM
	E2	QM	QM	QM
	E3	QM	QM	A
	E4	QM	A	B
S2	E1	QM	QM	QM
	E2	QM	QM	A
	E3	QM	A	B
	E4	A	B	C
S3	E1	QM	QM	A
	E2	QM	A	B
	E3	A	B	C
	E4	B	C	D

Reproduced by permission of DIN Deutsches Institut für Normung e.V.

The ASIL is defined within the range from ASIL A to ASIL D, where ASIL A is the lowest safety integrity level and ASIL D the highest. In addition, the class QM denotes that there are no requirements to comply with ISO 26262.

QM classified malfunctions are not safety relevant in the means of ISO 26262, what does not mean, that the system is not safety relevant at all. The classification QM assumes that there is a functioning QM system implemented, for example, according to ISO/TS 16949, and that this is sufficient to develop such a QM classified system.

Table 5 gives some examples for the ASILs of some malfunctions in road vehicles.

As the determination of the parameters E, C, and S are very subjective, the resulting ASIL cannot be objective. Currently, there are activities within the automotive industry to harmonize the method to determine the ASIL. A detailed harmonization of the ASIL itself is not possible as it depends on the concrete vehicle configuration (e.g., number of airbags that influence the severity of a malfunction of another system).

4.5 Safety element out of context [*]

The automotive industry develops generic elements for different applications and for different customers. These generic elements can be developed independently by different organizations. In such cases, assumptions are made about the requirements and the design, including the safety requirements that are allocated to the element by higher design levels and on the design external to the elements. Such a safety element out of context (SEooC) is not developed in the context of a particular vehicle.

An SEooC can be a system, an array of systems, a subsystem, a software component, or a hardware component or part. Examples are system controllers, ECUs, microcontrollers, software implementing a communication protocol or an AUTOSAR software component, and so on. Thus, the development of the SEooC is based on assumptions; on an intended functionality and use context that includes external interfaces. These assumptions are set up in a way that addresses a superset of items, so that the SEooC can be used later in multiple different, but similar, items.

Figure 5 shows the relationship between assumptions and SEooC development. The development of a SEooC can start at a certain hierarchy level of requirements and design. The correct implementation of the requirements for the SEooC will be verified during the SEooC development. The validation of these requirements and assumptions are then established during the development of the item.

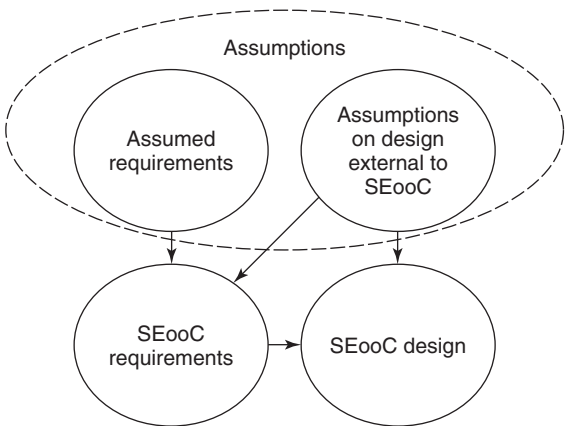


Figure 5. Relationship between assumptions and SEooC development [*]. (Reproduced by permission of DIN Deutsches Institut für Normung e.V.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 5. Exemplary ASIL classifications of some possible malfunctions of electrical/electronic vehicle systems[a].

System	Malfunction	Driving Scenario	Exposure	Controllability	Severity	ASIL
Electronic stability program (ESP ®)	Unintended brake force up to almost locking pressure of one front wheel	Country road, oncoming traffic, or obstacles on the side	E4 (according to ISO 26262-3, Table B.2)	C3 The majority of the drivers will not be able to keep the vehicle in the lane	S3 Because of asymmetric braking the vehicle will be destabilized, accidents with road users/objects may lead to life- threatening or fatal injuries	D
Airbag	Unintended airbag deployment during driving	Country road, oncoming traffic, or obstacles on the side	E4 (according to ISO 26262-3, Table B.2)	C3 (according to ISO 26262-3, Table B.4)	S3 Life-threatening injuries (survival uncertain) or fatal injuries cannot be excluded.	D
Engine management	Unintended acceleration	Country road, oncoming traffic, or obstacles on the side, driving in a curve at high speed, vechicle close to destabilization	E2 Significant amount of time for the critical scenario where an acceleration can result in leaving the lane is <1%	C3 Only a few drivers can handle this, as exceptional reaction is required	S3 Leaving the lane may lead to life-threatening or fatl injuries	B
	Sudden unwanted loss of vehicle acceleration	Country road, passing with oncoming traffic, "vehicle slows down to zero"	E1 Critical distance to oncoming traffic during passing very seldom	C2 Normally controllable, also by other traffic participants	S3 Passing can not be finalized, collision with oncoming traffic leading to life-threatening or fatal injuries	QM
Electric window lift	Unintended closing of window	Parking not belted passengers (children) leaning out of the window	E2 Parking with open window and children leaning out	C2 Adequate reaction possible in most cases, also by other traffic participants	S3 Can cause pinching of extremities or other parts of the body (head, cervix, etc.)	A
Turn indicator	Side-inverted activation	Country road, vehicle 1 wants to turn left (without a turn off lane in a small crossway), actuates indicator to turn left, but flasher indicates the wrong direction (i.e., indicates turning right) The following vehicle 2 wants to drive past left to the vehicle he is expecting to turn right	E2 Country road, crossroad without a traffic light, no turn off lane, intended overtaking in the area of turn off	C2 Following road traffic sees braking vehicle with wrong direction indication. This calls increased attention to most drivers, therefore they will abandon from driving past left	S3 Side collision, Delta-v > 35 km/h. Life-threatening injuries (survival uncertain) or fatal injuries cannot be excluded	A

[a] The current classifications for engine management lead to an ASIL B for the malfunction "unwanted vehicle acceleration" and QM for "sudden unwanted loss of vehicle acceleration." However, in the automotive community, we recognize the beginning of a trend toward ASIL C/ASIL A for these malfunctions; therefore, in future, the common understanding of these ASIL classifications may remain on this higher level.

Encyclopedia of Automotive Engineering. Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For further information and detailed examples of the concept of SEooC refer to ISO 26262 part 10.

4.6 Confidence in the use of software tools

4.6.1 Introduction

The goal of the qualification of software tools is to get confidence that the use of a software tools does not lead to the violation of a safety goal (i.e., that an erroneous output could not lead to safety critical behavior of a product) by maintaining quality assurance for certain tools and their utilization. To this end, the ISO 26262 provides for the following measures:

- appropriate measures that can be introduced during the product development phases and that keep a tool from generating faults in and of itself;
- faults generated by a tool can be identified and eliminated in a subsequent process step or by means of organizational measures.

The term *software tool* used in the ISO 26262 context not only refers to the tools that are used in software development but also describes the application programs that are used in hardware development and ECU calibration or deployed at the system level (e.g., for requirements management). It comprises the huge range from commercial tools to small scripting enhancements.

Thus, development organizations may often have a huge number (typically more than a thousand) of tools. It is obvious that the qualifying tools according to ISO 26262 may lead to a high effort and in an extreme case may prohibit tool use. The challenge is to handle this immense number of tools adequately without degrading safety conditions. For that, we provide a methodology that is based on a fine granular tool classification that is based on a detailed use cases analysis. In addition, this section introduces a two-stage tool qualification that is based on generic uses cases and sample processes.

4.6.2 Tool classification

The first step for providing confidence in the use of software tool addresses the *classification of all use cases* for which a given tool is deployed during the development process. Basis for all further activities is a detailed use case analysis. In the case that a use case is not safety relevant, no further qualification measures are required and the qualification process is finished for this case (tool impact, TI). The other case calls for an investigation into the degree of probability with which faulty contents issued by the tool in a subsequent development step may be discovered, for example, through reviews or testing (tool error detection, TD). Quality control measures can be introduced through the deployment of an additional tool, or by superimposing a process step (e.g., plausibility checks). If it is not possible to insert suitable verification steps into the development process to exclude tool-induced errors with a high degree of probability (i.e., to increase the TD), measures suited to tool qualification are called for. The result of the classification (Figure 6) essentially depends on the context within which a software tool is deployed in the product development process.

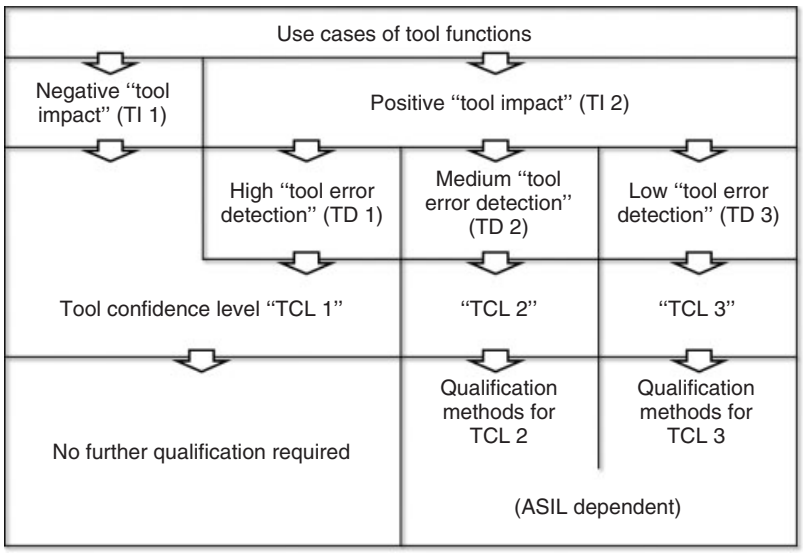


Figure 6. Two-staged process for qualification of software tools in accordance with ISO 26262-8.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 6. Methods for qualification of software tools classified TCL3 in accordance with ISO 26262.

Methods		ASIL			
		A	B	C	D
1a	Increased confidence from use in accordance with 11. 4.7	++	++	+	+
1b	Evaluation of the tool development process in accordance with 11.4.8	++	++	+	+
1c	Validation of the Software Tool in accordance with 11.4.9	+	+	++	++
1d	Development in accordance with a safety standard[a]	+	+	++	++

[a] No safety standard is fully applicable to the development of software tools. Instead, a relevant subset of requirements of the safety standard can be selected.
Example: Development of the software tool in accordance with ISO 26262, IEC 61508, or RTCA DO-178.
Reproduced by permission of DIN Deutsches Institut für Normung e.V.

4.6.3 Tool qualification

Tool qualification can be carried out in accordance with one of the four available methods. According to ISO 26262, they are

- increased confidence from use,
- evaluation of the development process,
- validation of the software tool, and
- development in compliance with a safety standard.

The recommendation for the usage of these alternative measures is to be determined considering the ASIL classification of the product to be developed, and from the result of the tool classification. Table 6 shows an overview of the qualification methods recommended for tools classified with "TCL (tool confidence level) 3" for the different ASILs.

4.6.4 Use case analysis based fine granular tool classification

Efforts during the tool qualification phase may reach an enormous level because of the typical use of an immense number of tools. For example, the Automotive Technology Business Sector of Robert Bosch GmbH covers approximately 1500 tools within 56 organizational units. The question is how to handle tool qualification adequately with this amount of tools.

In attaching importance to a fine granular tool classification, it is possible to contain the efforts that may be caused during the tool qualification phase.

At first sight, this seems to be contradictory as a use case's refinement and an increase of detail level are accompanied by an increase in the number of use cases, and thus it could be expected to result in an effort increase and just not an effort reduction. Therefore, the gain of a fine granular tool classification is not directly obvious.

But, this perspective is too simplistic. The increase of number of use cases is undisputed. As a reward for this, a detail level will be reached that is already located in software documentations, technical specifications, user guides, development manuals, former use cases analysis, and—what matters most is—descriptions for development processes.

Today's development processes currently reflect quality control measures. The intrinsic feature is that a tool chain will be described as a network of cooperating tools and not as a set of stand-alone tools whereby output from a predecessor tool may serve as input for a successor tool. This allows us to group relevant tools in terms of their data flow and not in terms of logical or physical borders. On the other hand, successor tools often provide safety as a precondition for further processing. It is to exploit this for ensuring tool correctness and improving functional safety.

Using this methodology, the Robert Bosch GmbH could demonstrate that the majority of tools embedded in a tool chain are protected within a development process and thus contain mostly use cases with TCL1. Hereby, only 2% of all use cases are classified as TCL2 or TCL3. In absolute numbers, that means, only 31 tools (from a total amount of 1500 tools) have to be considered in the subsequent qualification phase where the real efforts are estimated (e.g., in a validation). In this way, the increased effort in the first classification phase is more than compensated by a very slender second qualification phase.

As a spin-off of this methodology, it becomes clearly evident that critical tools in terms of ISO 26262 tool qualification are mainly located at the development process' periphery (concept phase and transition to production and operation), because further quality control measures with the scope of development process are very difficult to implement there.

This conspicuity has not been necessarily expected and can be approved by our experience over time—in the early beginning of tool qualification; there have been compilers and code generators that seemed to be the most critical tools. Thus, it is necessary to turn one's attention to tools that are located at the immediate beginning or at the end

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of a development process, the top section of the V-shape process flow.

4.7 Random hardware faults and hardware metrics

4.7.1 Fault classes [*]

In general, the combination of faults that are considered are limited to combinations of two independent hardware faults, unless analysis based on the functional or technical safety concept has shown that n point faults with $n > 2$ are relevant. Therefore, for a given safety goal and a given hardware element, a fault can be classified in most cases as either one of the following fault classes:

- Single-point fault: this fault can lead directly to the violation of a safety goal and is a fault of a hardware element for which no (not one) safety mechanism prevents some of the faults of the hardware element from violating the safety goal.
- Residual fault: this fault can lead directly to the violation of the safety goal and is a fault of a hardware element for which at least one safety mechanism prevents some of the faults of the hardware element from violating the safety goal.
- Detected dual-point fault: this fault contributes to the violation of the safety goal, can only lead to a safety goal violation in combination with one other independent hardware fault that is related to the dual-point fault and is detected by a safety mechanism that prevents it from being latent.
- Perceived dual-point fault: this fault contributes to the violation of the safety goal but will only lead to a safety goal violation in combination with one other independent hardware fault that is related to the dual-point fault and is perceived by the driver with or without detection by a safety mechanism within a prescribed time.
- Latent dual-point fault: this fault contributes to the violation of the safety goal but will only lead to the violation of the safety goal in combination with one other independent fault and is neither detected by a safety mechanism nor perceived by the driver. Until the occurrence of the second independent fault, the system is still operable and the driver is not informed about the fault.
- Safe fault: safe faults can be faults of one of the two categories:
  - all n point faults with $n > 2$, unless the safety concept shows them to be a relevant contributor to a safety goal violation or
  - faults that will not contribute to the violation of a safety goal.

For further information about the hardware faults and a flow diagram for fault classification and fault class contribution calculation, please refer to ISO 26262 part 5 and part 10.

4.7.2 Hardware metrics

In the development of safety-related systems according ISO 26262, quantitative requirements for the above-mentioned fault classes have to be calculated. These quantitative requirements are ASIL dependent. For example, target values for the probabilistic metric for random hardware failures are $<10^{-7}\,\mathrm{h}^{-1}$ for ASIL C and $<10^{-8}\,\mathrm{h}^{-1}$ for ASIL D.

It is important to know that (other than in other safety standards) these quantitative requirements do not have any absolute significance. This means that these values do not correlate with real failure rates in the field. The quantitative requirements can be calculated based on a commonly recognized industry source, for example, IEC/TR 62380, EN 50129, and SN 29500. The quantitative requirements are only useful to compare a new design with an existing one or to compare two new alternative solutions. So, there is no necessity to align these results with real field data.

5 SIDE ISSUES OF SAFETY

5.1 Quality assurance

5.1.1 Quality gates

To assure the quality of a product, during the development of the product, there are several milestones defined where the product quality is checked using checklists. These so-called quality gates are typically defined as follows:

- QG0: project started
- QG1: release of technical specification
- QG2: completion of production planning
- QG3: scope of delivery is available from serial equipment
- QG4: product and process approval
- QG5: project completion.

Figure 7 shows a simplified process model with the mentioned quality gates.

The achieved level of product safety (together with other product and process quality criteria) is evaluated within quality gate meetings using checklists.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

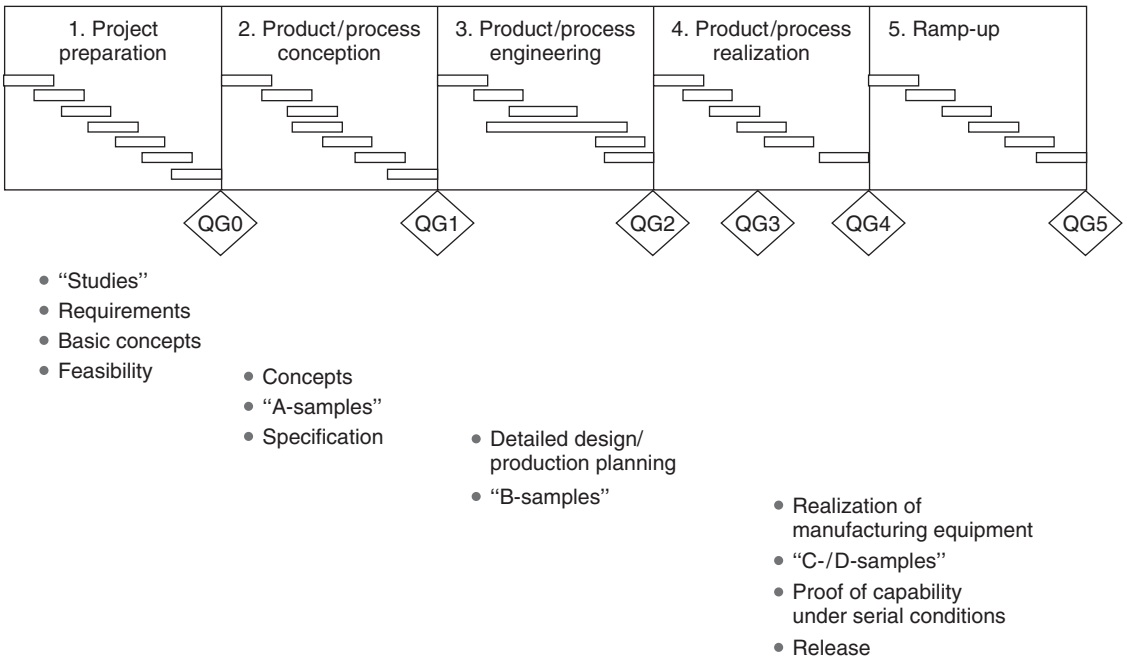


Figure 7. Quality gates of automotive development process.

5.1.2 Confirmation measures according to ISO 26262

ISO 26262 requires some further confirmation measures that are independent from the quality gates, but that are normally linked to them:

- Functional safety audit: evaluation of the process required for functional safety.
- Functional safety assessment: evaluation of the achieved functional safety of the item.
- Confirmation review: evaluation of the compliance of work products with the corresponding requirements of ISO 26262.

An overview of the subjects for evaluation, expected results, responsibilities of persons performing the confirmation measures, timing during the safety life cycle, scope, and depth is shown in Table 7.

In typically large organizations within the automotive industry, the process landscape is defined as follows (see also Figure 8).

On corporate level, directives and process definition define boundaries for the implementation on organizational level. These processes are evaluated within organizational safety audits if the definition of the processes on this level is compliant to ISO 26262.

On product engineering level in a functional safety audit, it is checked whether the product has been developed according to the defined processes. In parallel functional safety product assessments evaluate the achieved level of product safety. Questionnaires in mandatory quality gate meetings ensure that functional safety audits and functional safety assessments are done in each project.

Normally functional safety audits and functional safety assessments can be done internally, that is, it is possible to ensure the required independency of functional safety auditors or functional safety assessors without certification by an external party.

5.2 On-board diagnosis

5.2.1 Introduction

In case that a safety-related system is not able to guarantee its safety performance adequately and this could lead to a hazard of the system, it is necessary to mitigate this increased risk by safety measures. For this, it is necessary to have a detection mechanism for the system's faults, failures, and malfunctions.

5.2.2 Monitoring and fault recognition

Malfunctions or failures of a system can lead to deviations from the "normal" system status. So, the system status is subject of monitoring to recognize such deviations as early as possible as a trigger for mitigation activities, for example,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 7. Procedural requirements for confirmation measures according to ISO 26262.

Topic	Confirmation Review	Functional Safety Audit	Functional Safety Assessment
Subject for evaluation	Work product	Implementation of the processes required for functional safety	Item as described in the item definition in accordance with ISO 26262-3:2011, Clause 5
Result	Confirmation review report[a]	Functional safety audit report[a] in accordance with 6.4.8	Functional safety assessment report in accordance with 6.4.9
Responsibility of the persons that perform the confirmation measure	Evaluation of the compliance of the work product with the corresponding requirements of ISO 26262	Evaluation of the implementation of the required process	Evaluation of the achieved functional safety Provision of a recommendation for acceptance, a conditional acceptance or a rejection, in accordance with 6.4.9.6
Timing during the safety lifecycle	After completion of the corresponding safety activity Completion before the release for production	During the implementation of the required process	Progressively during development, or in a single block Completion before the release for production
Scope and depth	In accordance with the safety plan	Implementation of the processes against the definitions of the activities referenced or specified in the safety plan	The work products required by the safety plan, the implementation of the required processes and a review of the implemented safety measures that can be assessed during the item development

[a] This report can be included in a functional safety assessment report.
Reproduced by permission of DIN Deutsches Institut für Normung e.V.

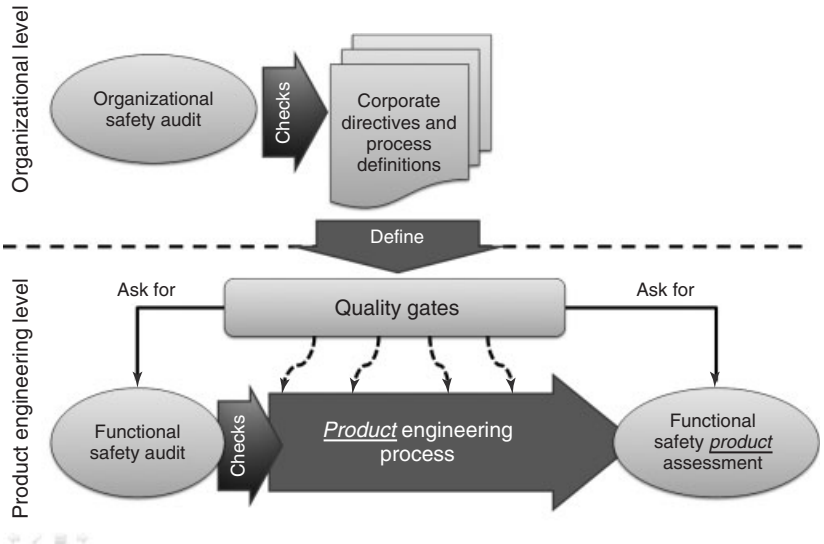


Figure 8. Example for an implementation of functional safety audits and assessment within a huge organization.

reboot or switch off of the system. Goal is to prevent that a fault lead to a safety critical system status by bringing the system into its safe state, for example, switching off the engine to prevent the vehicle to accelerate unintentionally.

Examples of fault recognition methods, which can be used for monitoring purposes, are:

- reference value check
- redundant value check
- monitoring communication links
- handshake
- monitoring physical properties
- monitoring program execution.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For a deeper discussion of these examples, see Schäuffele and Zurawka (2005).

5.2.3 System classes

5.2.3.1 Fail-safe system (FS). On systems capable of assuming this type of safe state, such as the defined emergency shutdown (also frequently called *emergency stop*), a safety response may consist of the initiation of this state. It must be ensured that a safe state of this kind may be exited only in a controlled manner, that is, not as a consequence of additional faults, malfunctions, or failures. A system featuring this type of safety response is also known as a *fail-safe system* (*FS system*).

5.2.3.2 Fail-reduced system (FR). In some cases, the adoption of a safe state is naturally followed by the transition into a degraded operating mode. Whenever this result is continued—albeit restricted—system serviceability, such as the limp-home operating mode, this is termed *fail-reduced system* (*FR system*). There may be situations in which the degraded operating mode is deliberately introduced because of a scarcity of resources or with a view to risk minimization.

5.2.3.3 Fail operational system (FO). Sometimes introducing a safe state is not possible because of technical obstacles. This applies, for example, to many vehicle functions that are mandatory while driving. If a system happens to fail, the adverse effect of the system failure on the behavior of the vehicle must be neutralized, and a fallback switchover to a suitable backup system will be required. In principle, the implementation of the referred backup system may be similar to, or different from, the type of system suffering in failure. A system featuring this type of safety response is known as a *fail-operational system* (*FO system*).

For example, system requirements in terms of safety logic are frequently specified in the form of FO/FO/FS or FO/FO/FR. This means that a system meeting these specifications must remain fully operational, regardless of two successive internal failures. It is the occurrence of a third that would permit the system to transition into a safe state and/or enter the limp-home operating mode.

6 SUMMARY AND CONCLUSION

Safety of road vehicles becomes more and more important, as the vehicle systems get more complex in future. Managing this complexity leads directly to the challenge of managing the safety of such a complex system. The most important standard in this field is the newly released ISO 26262, a standard for functional safety of road vehicles. To reduce the risk in case of a product liability issue, it is necessary to implement the ISO 26262, as it contributes to the definition of the existing state of the art. Nevertheless, its implementation is not sufficient to reach the state of the art, as this could include much more than the contents of the standard. Key concept of ISO 26262 is a risk-based assessment and implementation of processes and product measures to reduce the residual risk to a reasonable level. The ISO 26262 requires several issues that seem to be new for the automotive industry, for example, the qualification of software tools and the application of hardware metrics. To evaluate the achieved functional safety of the product and the implemented processes, functional safety assessments and audits are performed. The results of these confirmation measures are typically checked at certain quality gates. To enable systems to tolerate faults and to reach and to stay within safe states, monitoring and diagnostics functions are used. This leads normally to FS systems, whereas in future, FO systems may come more into focus.

ACKNOWLEDGMENTS

Citations from ISO 26262 (marked with [*] within the text above) are reproduced by permission of DIN Deutsches Institut für Normung e.V. The definitive version for the implementation of this ISO-Standard is the edition bearing the most recent date of issue, obtainable from Beuth Verlag GmbH, Burggrafenstraße 6, 10787 Berlin, Germany.

In addition, we would like to thank our following colleagues for their inputs, fruitful discussions, and review of the content of this document:

Dr. Bernd Müller, Chief Expert for "Safety, Reliability, Availability," Corporate Research and Advance Engineering,

RA Andreas Reuter, Attorney-at-law, Vice President Corporate Legal Services,

Dipl.-Phys. Carsten Gebauer, ISO 26262 expert and member of ISO TC22/SG3/WG16, Corporate Research and Advance Engineering,

Dr. Jürgen Klarmann, Expert for tool qualification, Cross Divisional Group—Software, Methods and Tools.

ENDNOTES

1. "Destructive" in the context of memory tests during the start-up phase means that the test overwrites a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

high amount of memory cells and therefore the original content of the memory is lost.

REFERENCES

Hamann, R., Kriso, S., Williams, K., *et al*. (2011) ISO 26262 release just ahead: remaining problems and proposals for solutions. SAE World Congress 2011 (SAE 2011-01-1000).

ISO 26262-X (2011) (E) Road vehicles—functional safety—part X: .... International Organization for Standardization, Geneva (X = 1...9).

ISO 26262-10 (2012) (E) Road vehicles—functional safety—part 10: guideline on ISO 2662. International Organization for Standardization, Geneva.

FURTHER READING

Löw, P., Papst, R., and Petry, E. (2010) *Funktionale Sicherheit in der Praxis—The Application of DIN EN 61508 and ISO/DIS 26262 in the Development of Series Products*, 1st edn, dpunkt. Verlag, Heidelberg. ISBN: 978-3898645706 (currently only available in German).

Schäuffele, J. and Zurawka, T. (2005) *Automotive Software Engineering*, 1st edn, SAE International, Warrendale, PA. ISBN: 0-7680-1490-5, SAE Order No. R-361.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto194
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Interfaces between Sensors and ECUs

Hiroaki Hoshika and Shinya Igarashi

Hitachi Automotive Systems, Ltd., Hitachinaka, Japan

1 Introduction 1
2 Requirements Necessary for Interfaces Between Sensors and ECUs 1
3 Classification of Interfaces 1
4 Analog Interfaces 1
5 Pulse Interfaces 4
6 Serial Communication Interfaces (Pulse Code Modulation) 5
7 Network Nodes 6
8 Conclusion 9
References 10
Further Reading 10

1 INTRODUCTION

Nowadays, automotive electronics systems use a lot of Electronic Control Units (ECUs) and sensors. In general, multiple sensors are connected to one ECU. They are connected with various interfaces. We are going to describe several types of interfaces that connect sensors to ECUs as well as their characteristics and synopsis.

2 REQUIREMENTS NECESSARY FOR INTERFACES BETWEEN SENSORS AND ECUS

Sensors convert some physical quantity, displacement, and/or the like into electric signals. The signals are used as an input to control an ECU. Therefore, various types of interfaces are used that are able to suppress deterioration and time lag to surely transmit sensor-collected information to an ECU. In automotive environments, various types of external electric disturbances can happen. This is why various measures are taken, for example, to minimize the influence of such disturbances, surely let an ECU know a failure of the sensor for a prompt failsafe process if a sensor fails, and transfer information to an ECU with some low cost communication means even from a sensor located at a distance from the ECU. Conventional interfaces have only such functions that transmit sensor-collected information to ECUs. In recent years, however, the needs are high for bidirectional information exchanges involving direct connections to vehicle networks such as LIN (local interconnect network) and CAN (controller area network).

3 CLASSIFICATION OF INTERFACES

Table 1 lists some of the typical interfaces that connect sensors and ECUs.

4 ANALOG INTERFACES

This is one of the interfaces that have been used for the longest time, and is still widely in use nowadays. The physical quantity detected by a sensor is replaced by a voltage or amperage for transfer. In automotive systems, the voltage type is frequently in use.

Analog interfaces convert the voltage acquired on the side of ECU into another physical quantity. Therefore, the relationship between voltage and physical quantity needs to be interrelated as the output characteristics of a sensor in advance. These characteristics do not necessarily need

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto195
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Table 1. Typical interfaces

Classification by hardware	Classification by information transfer methods	Characteristics
Analog interface	Voltage	Physical quantity is transferred through voltage signal. This is one of the most common methods and is still applied in wide fields. This method, however, is under influence of voltage drops, external noises, grounding voltage, ECU's reference voltage, and so forth and, therefore, tends to cause errors. This is why precautions are necessary at the stage of designing wiring and signal filters
Analog interface	Current loop	Physical quantity is transferred through current in the signal lines. A current loop is formed with a sensor and an ECU. Consequently, this method is not susceptible to signal deterioration or external noises. However, electric circuits are somewhat more complicated than those of voltage methods
	Differential voltage	Physical quantity is transferred through voltage difference of two signal wires. This method is not susceptible to external noises that are not often in phase with two signal wires simultaneously. This method, however, has a complex electric structure and, therefore, is not usually used for automobiles
Pulse interface	Frequency modulation	Physical quantity is transferred in linkage with the frequency of pulse signals. This method is not susceptible to the fluctuation in signal voltage if some appropriate threshold is selected. This method has been in use with sensor signals for a long time but is not free from errors in modulation and demodulation caused by errors in frequency sources
	PWM modulation	Physical quantity is transferred in linkage with the duty ratio of a pulse signal. Duty ratio is acquired by dividing a high or low level time by the periods of pulse signals. This method, therefore, is not susceptible to errors in frequency sources, but obscure waveforms processed by a noise removal filter can be an error factor
	Frequency and PWM modulation	Frequency modulation and PWM modulation are processed simultaneously to link two types of physical quantities with one signal wire. Two types of signals are transferred with a single wire. Therefore, it is possible to reduce the wiring cost to connect an ECU
Serial communication interface	SENT (SAE-J2716)	This is the unidirectional communication with the pulse time modulation of 4 bits per pulse and 2 bytes per message. This method is seen as a replacement of analog signal communications. This method, however, supports IDs and CRCs in transfer data; therefore, multiple channels and error detections are compatible. Circuit configurations are simple enough to use LSIs. This method is used for angle sensors and so forth
	DSI/PSI5	Only two wires are used to enable power supply and signal transfer. Multiple sensors can be connected in parallel. Amperage readings are processed in Manchester encoding for data transfer. This method is suitable for an environment where you need to position a large number of sensors with relatively low power consumption. This method is used in, for example, airbag systems
Network nodes	CAN(ISO11519-2) (ISO11898)	This is a two-line differential type and enables multimaster bidirectional communications of 1 Mbps at the maximum. This method is widely used for inter-ECU communications. This method is one of the international standards but is rarely applied to sensors because of its cost of circuits. For more information, see Section 7.2
	LIN (LIN 2.x) (SAE-J2602)	This is a single-line type and enables single-master bidirectional communications (semiduplex communications) of 20 Kbps at the maximum. As the circuits are not expensive, this method is used for sensors that do not require very high response performance. For more information, see Section 7.1

(Reproduced by permission of Hitachi Automotive Systems.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto195
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

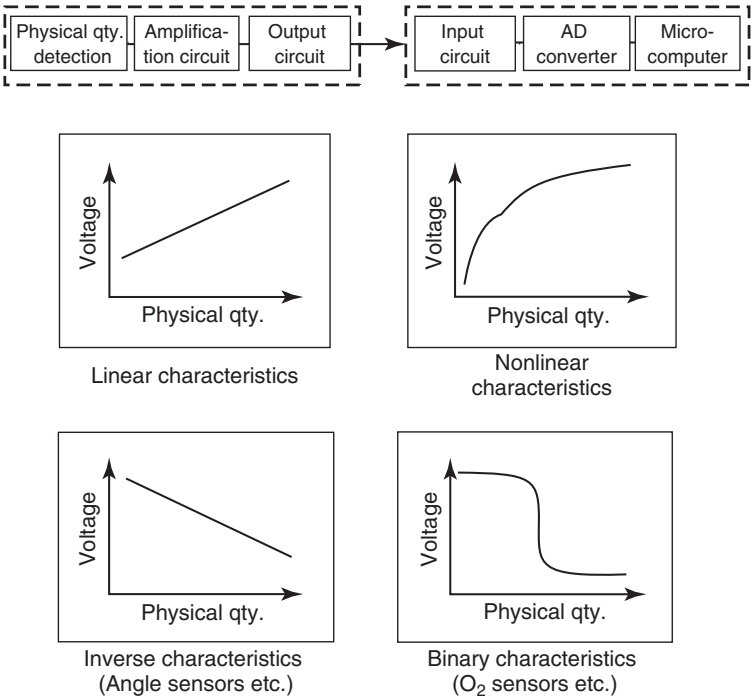


Figure 1. An example of the relationship between voltage and physical quantity. (Reproduced by permission of Hitachi Automotive Systems.)

to take a linear form. Contrarily, these characteristics can take various forms depending on the detection principle of such a physical quantity and/or the range of significant measurements (Figure 1).

When an analog interface is in use, signals are transferred from a sensor by way of a wire harness. On the side of an ECU, there is a filter to remove any intrusion of signal noises that may be caused by external electrical disturbance. Generally speaking, a low pass filter (LPF) is provided to cut off high frequency elements equal to or higher than the cut-off frequency (f_c). Figure 2 shows an example of such LPFs.

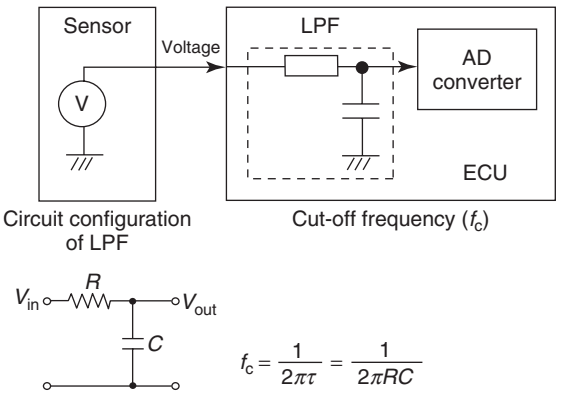


Figure 2. An example of low pass filters. (Reproduced by permission of Hitachi Automotive Systems.)

4.1 Precautions for using analog interfaces

If the voltage at the input terminal is different from the output of the sensor signal, the ECU causes an error in the physical quantity recognized on its side. Therefore, some measures should be taken to prevent any difference from being generated at the sensor output signal and the terminal voltage on the side of ECU. Such measures can differ depending on the situations. For example, a wire harness generates resistance. Accordingly, the current flowing down a signal wire or a grounding wire causes a voltage drop. Further, an overshoot can be generated in a grounding wire. As countermeasures, for example, the amperage in signal wires is reduced, and/or the grounding current is separated using a dedicated grounding wire. If, moreover, the reference sources are different between the side of a sensor and its corresponding ECU, such a configuration can cause errors. For this reason, when some high precision is required with respect to voltage, a ratiometric-type output circuit may be used where a reference voltage is provided from the side of the ECU to the sensor. Figure 3 shows some elements that tend to cause problems while a voltage-type analog interface is used.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto195
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

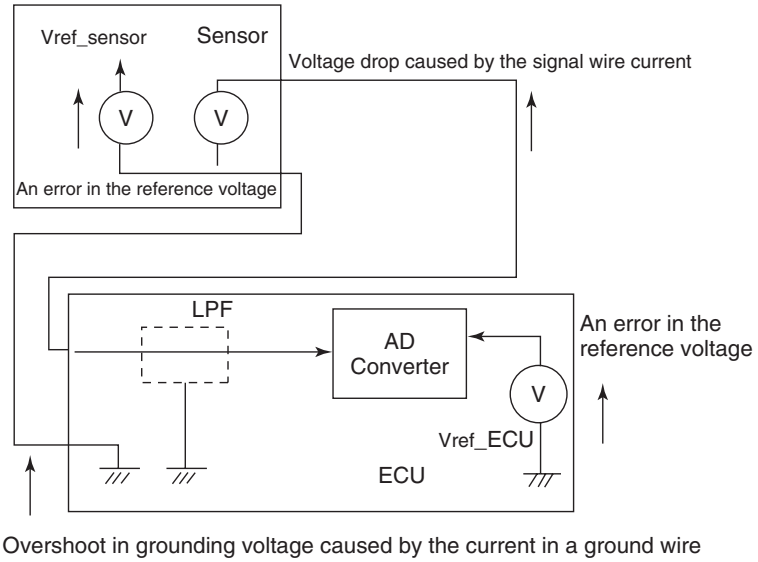


Figure 3. Elements that tend to cause problems when a voltage-type analog interface is used. (Reproduced by permission of Hitachi Automotive Systems.)

5 PULSE INTERFACES

When analog interfaces are used, absolute voltage and absolute amperage are susceptible to various types of external disturbances. When pulse interfaces are used, on the other hand, external disturbances are not so influential because only two levels of voltage, a high level and a low level, need to be transferred as signals. This means that it is only necessary to reach and exceed the threshold of each level. On top of that, in this method, it is possible to replace the AD converter in an ECU with a timer circuit. Furthermore, it is also possible to simplify the signal output circuit on the side of the sensor. It is, more often than not, analog signals that are detected by a sensor. Consequently, these signals are converted into pulse signals in a modulation circuit.

The pulse frequency modulation is the so-called FM modulation; that is, the signals detected by a sensor are replaced with the frequency of pulse signals. For example, the analog voltage signal acquired on the side of a sensor is converted into the frequency with a VCO (voltage-controlled oscillator). On the side of an ECU, a counter circuit counts the number of pulses in pulse signals with regular intervals to acquire the counts in proportion to the average frequency in one period. VCO circuits are not too complicated. This is why they are widely in use as automotive sensors, which are mostly analog circuits. They, for example, are used to acquire output signals from an airflow sensor.

The PWM (pulse-width modulation) modulation is a type of information transfer where the ratio (the duty ratio) between the high and the low level times is used. The pulse frequency from one rising pulse to the next rising pulse (called a *carrier frequency*) is always constant, whereas the ratio between the high and the low level times of signals is varied. The ECU includes a timer circuit to detect each of the high and the low level times. If you divide the high or the low level time by the period of signals (the sum of the high and low level times), you can obtain the duty ratio. The PWM modulation does not change the duty ratio even if the period of pulses fluctuates. This is an advantage that enables the signal transfer with little errors, not requiring a stable oscillator on the side of the sensor. Figure 4 shows

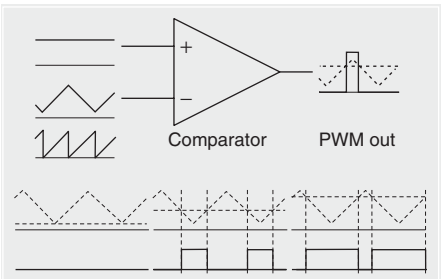


Figure 4. A schematic view of a PWM modulation circuit based on an analog circuit. (Reproduced by permission of Hitachi Automotive Systems.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto195
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

a schematic view of a PWM modulation circuit based on an analog circuit (Kumagai, 2011).

PWM modulation pulses are also easily generated by a digital circuit. Therefore, microcomputer-equipped sensors, in general, use PWM output units without any alteration. As for ECUs, they usually include a built-in microcomputer equipped with a timer/counter unit that provides an operation mode suitable for acquiring a duty ratio from these PWM signals.

The ratio between the high and the low level times is significant for PWM modulation pulse signals. The time duration necessary for a level transition generates an error. Therefore, only a minimum number of processing units such as an LPF, necessary for noise reduction, should be inserted in the signal I/O circuit, depending on the carrier frequency. We ought to pay attention in order to avoid increasing transition time.

Both the frequency and the PWM modulations require one signal wire for one piece of analog information. On the other hand, complex PWM modulation is also used, which varies the carrier frequency of the PWM modulation to transfer two pieces of analog information via one signal wire. For example, analog signal A varies the frequency of pulse signals, whereas analog signal B varies its duty ratio. This modulation reduces the number of wires, which means a cost reduction.

6 SERIAL COMMUNICATION INTERFACES (PULSE CODE MODULATION)

Serial communication interfaces use AD converters or similar to encode analog signals from sensors. After this, they transfer such signals one by one by the unit bit. A clock synchronization type processes the transferring timing of each of such bits with a dedicated clock line, whereas a start–stop synchronization type has a fixed duration of transmission time for each one bit and controls only starting points. Of these two types, the start–stop synchronization type is more commonly used in automotive systems. In the broad meaning of the word, network communications such as CAN and LIN belong to serial communications. In this section, however, we classify as serial communication interfaces those that do not establish mutual communications between sensors and ECUs and, besides this, that mainly use signal transfer from sensors to ECUs.

SENT (Single Edge Nibble Transmission) and DSI/PSI5 (distributed system interface–peripheral sensor interface), for example, are among the ones that are standardized and already spread.

6.1 SENT

SENT is one of the communication standards based on SAE-J2716. The widths of 5-V amplitude pulses are defined and classified into 16 levels. Every one pulse transfers 4-bit information (called a *nibble*). Every one message transfers 8-nibble information (32 bits). Among the information, two nibbles convey a status and CRC (cyclic redundancy check). To sum up, a sensor practically transfers 24-bit data. The smallest unit of pulse widths is a tick. One tick is equivalent to 3 μs. The pulse width varies depending on data transferred. One message takes 152–272 ticks (456–816 μs) to complete a transfer. In the field of automobiles, analog signal sampling periods are approximately 1 ms per signal at the shortest. Therefore, the performance of this interface, able to transfer two-channel signals (24 bits in total) with the accuracy of 12 bits in one wire in 816 μs or shorter, is sufficient. As the pulse widths are controlled, a reference clock is required on the side of a sensor. The fluctuation of the reference clock on the side of a sensor is allowed within the range of 20%, as the pulse widths are classified into no more than 16 levels and synchronized pulses are measured on the side of an ECU to correct pulse-width information. Accordingly, even a simple oscillator can work as a sensor. In addition, you can use a timer with a small number of bits as a pulse-generating timer. No communication-receiving functions are available from an ECU. The signals are something like what you would obtain by replacing conventional analog signals with serial communication signals to improve the reliability in information transfer. This interface has already been applied to the interface of angle sensors (Figure 5).

6.2 DSI/PSI5

As of now, these two are not international but noninternational standards under management of consortiums joined by parts suppliers, semiconductor manufacturers, and so forth. These two are different communication standards. However, they, besides both being a two-wire-type sensor interface, share several common features as described later. Figures 6 (DSI Consortium, 2011) and 7 (PSI5 Consortium, 2011) show their connection topologies.

(1) Operation requires only two wires for the power supply and for signals.
(2) Communications are based on master–slave interactions. The master sends a command and the slave returns a response.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto195
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

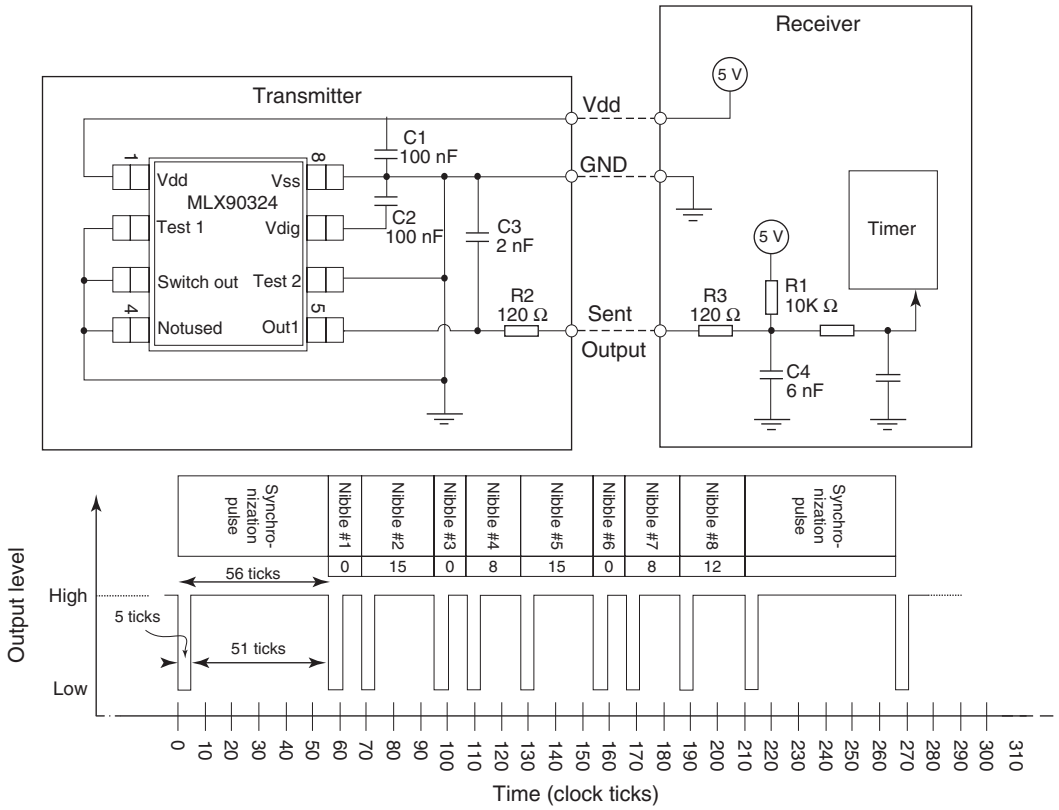


Figure 5. An example of interfaces based on SENT. (Reproduced by permission of Hitachi Automotive Systems.)

(3) Supported connections include topologies such as separate connections, bus connections, and daisy-chain connections.
(4) The master (ECU) controls the voltage of the power supply line. The slave (sensor) manipulates electricity for its own consumption to process communications.

These two look alike as types of interfaces. They, however, do not have any compatibility. DSI, for example, supports the commands of Manchester encoding based on voltage (it is not a level but a direction of transition that has logical significance). In its logic, responses are in the unit of nibble (4 bits) with which the electric current of a slave is controlled to one of three levels. PSI5, on the other hand, supports the commands based on the logic of voltage, whereas the responses are generated with Manchester encoding based on current C amperage. Their communication rates are also different from each other. DSI has a function that supports the power function class, an operation of an actuator. It is possible to operate an actuator by temporarily raising the voltage and amperage. These two standards are already in practical use in applications that require multiple sensors arranged in a wide range on one vehicle such as airbag systems.

7 NETWORK NODES

In recent years, there has been growth in the needs for sensors that support connections with networks. Automobiles use networks of various standards that correspond to their purposes. In this section, we are going to describe LIN and CAN, two typical communication standards generally used for sensors.

7.1 LIN (local interconnect network)

This was, at its origin, a network standardized by automobile manufacturers and parts suppliers in Europe, as a network they were able to configure with low cost hardware that did not require very fast communication speed such as small actuators and ECUs. Nowadays, this network is under management of a consortium joined by automobile manufacturers and parts suppliers from all over the world. This network has the characteristics described later. In general, this network is used as a subnetwork of a CAN.

The physical layer consists of single wires, which is the same as the standard of ISO9141 (International Organization for Standardization). The network has a bus structure

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto195
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

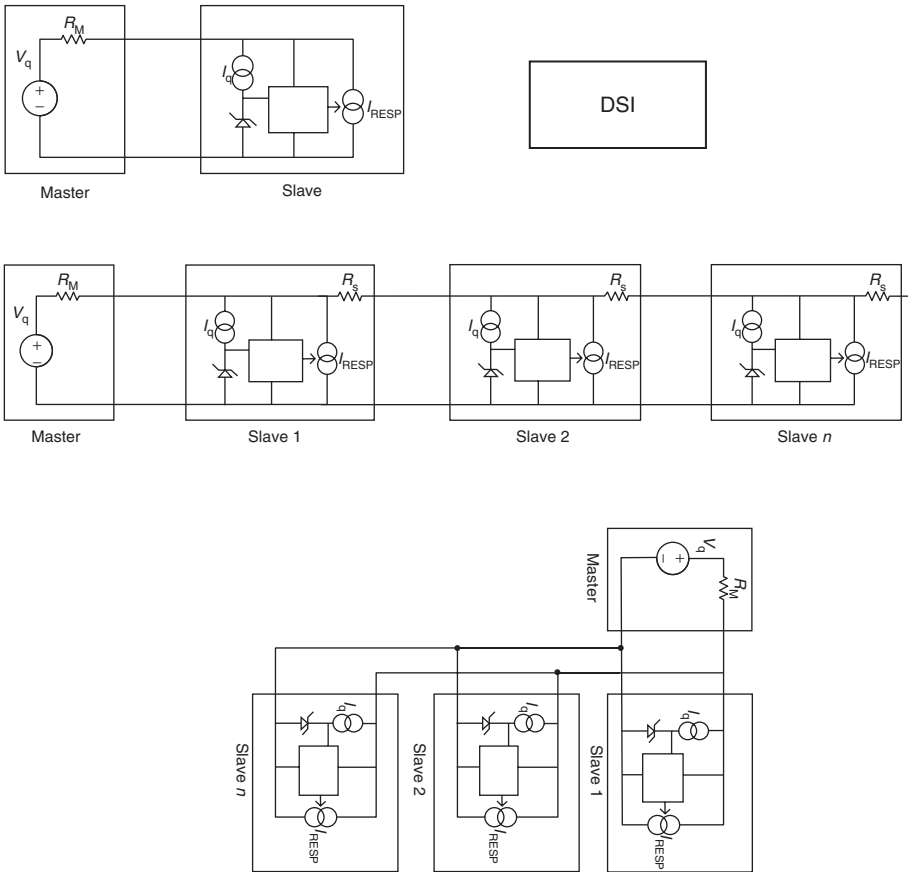


Figure 6. DSI connection topologies. (Reproduced by permission of Hitachi Automotive Systems.)

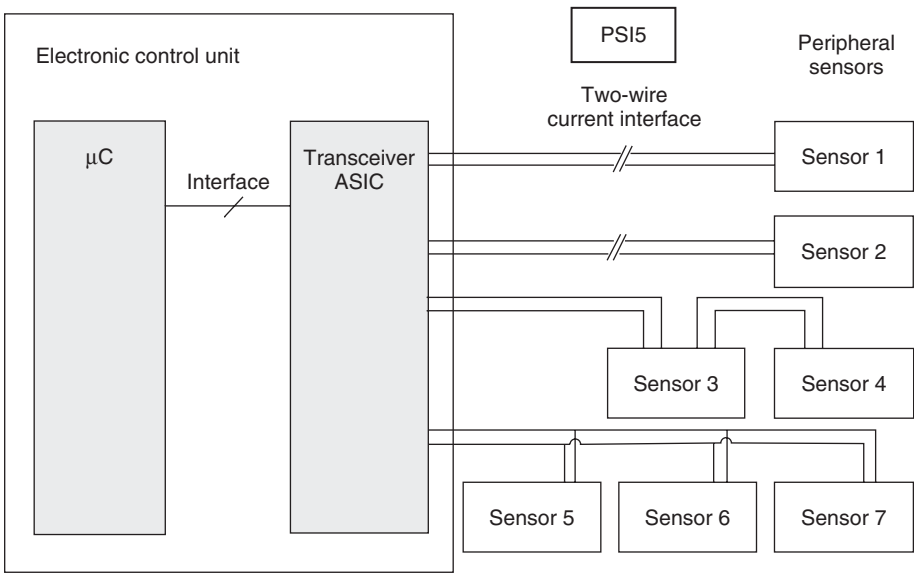


Figure 7. A connection topology of PSI5. (Reproduced by permission of Hitachi Automotive Systems.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto195
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

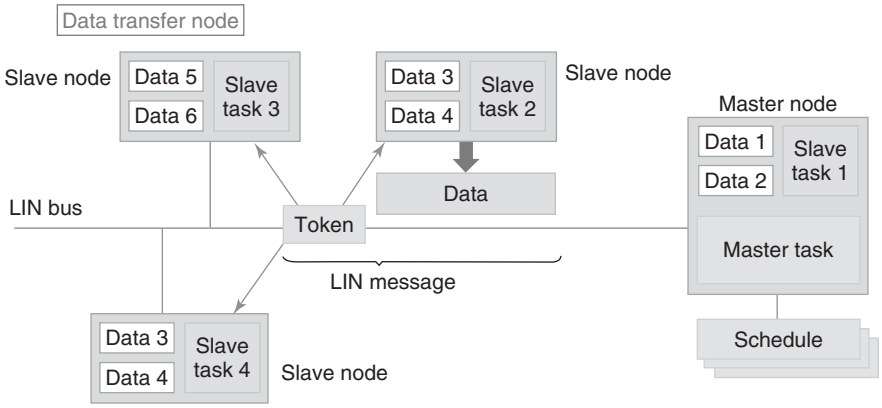


Figure 8. Schematic view of a token in LIN. (Reproduced by permission of Hitachi Automotive Systems.)

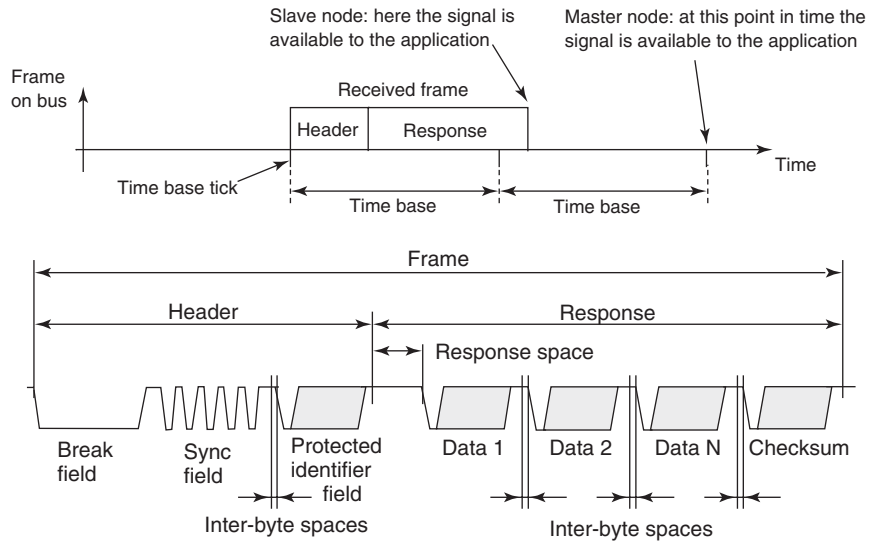


Figure 9. Configuration of the LIN communication frame. (Reproduced by permission of Hitachi Automotive Systems.)

of a linear type. There is only one master that sends tokens, which are received by a plurality of slaves at the same time. As message addressing is used, only the slave with the corresponding ID sends back a response to the master. In networks with sensors, the ECU works as the master, while the sensors work as the slaves. During communications, a time trigger protocol is used. Communication timings are all controlled by the master. The data transfer rate is 20 kbps at the maximum (10.4 kbps according to SAE-J2602), which is not suitable for transferring the signals in high speed responses. Figures 8 and 9 show a schematic view of tokens and the configuration of a communication frame (ITmedia Inc., 2010; Lin Consortium, 2006).

These serial communications make use of the UART (universal asynchronous receiver/transmitter) interface (start–stop synchronization). Slaves are synchronized based on the synchronization signal included in the message sent from the master. Therefore, slaves do not need any high accuracy oscillator. Messages are as short as 8 bytes at the maximum, including parity bits and checksums for protecting transferred data.

The transport protocol is based on ISO15765. Diagnoses and standard transport mechanisms are available. Therefore, various types of bidirectional communications are possible.

7.2 CAN (controller area network)

The specifications of CAN were published by Robert Bosch in Germany in 1986. Afterward, the specifications were standardized (ISO11898/ISO11519). As of now, this bus-type network is applied to almost all vehicles. Figure 10 shows an example of CAN connection and its data frame

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto195
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

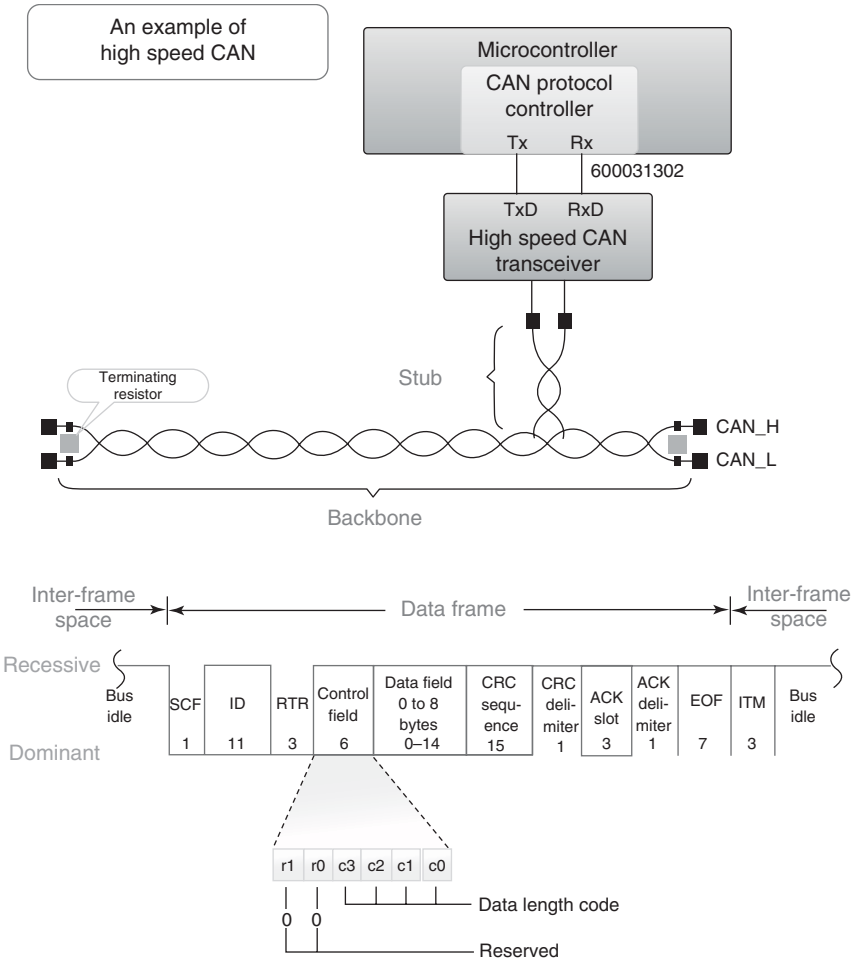


Figure 10. An example of CAN connection and its data frame configuration. (Reproduced by permission of Hitachi Automotive Systems.)

configuration (ITmedia Inc., 2008). This network is roughly classified into two types; that is, low and high speeds. The low speed CAN is mainly used for vehicle body systems, whereas the high speed CAN is mainly used for power train systems. Besides these, the single-wire CAN is also standardized. The high speed CAN is based on the two-wire voltage operation with twisted pair wires. This network is highly reliable. The standard maximum communication speed is 1 Mbps. However, the communication speed of 500 kbps is used for high speed CAN. Any node can start communications when the CAN bus is open. If, on the other hand, two or more nodes are trying to start communications at the same time, the node with the smallest ID number is given the right to start communications. This system makes it possible to control communication priorities.

This system requires a microcomputer. In most cases, a microcomputer equipped with a CAN controller is used. In addition, a CAN transceiver is used for the physical layer of CAN. As circuits are necessary for the network communication processing as well as a high performance microcomputer, this network is not suitable for small-scale sensors. This network is used for sensors and/or sensing systems with high reliability and complex functions. On top of that, this is a multimaster network, and it is easy to add CAN nodes. For example, if you want to add a sensor to a system for the purpose of control, and if it is difficult to modify the ECU, you only have to modify the software to add a sensor to a CAN bus.

8 CONCLUSION

As a trend of electrical interface between sensor and ECU, signals are generally shifting to digital transmission type such as SENT, LIN, DSI, and PSI5. It brings less cost of total vehicle electrical systems, higher safety redundancy,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto195
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and shorter development time of vehicle. In the future, it is expected that some of simple sensors will be managed by multifunction sensor-integrated communication unit to link vehicle network system. This maybe makes optimized sensor information topology.

REFERENCES

DSI Consortium (2011) *DSI3 Bus Standard*, Revision 1.00, February 16, http://www.dsiconsortium.org/downloads/DSI3_%20Bus_Standard_r1.00.pdf.

ITmedia Inc. (2008) *The CAN System from the Viewpoint of the Physical Layer Helpful for Implementations and Tests: IT Mono ist*, http://monoist.atmarkit.co.jp/mn/articles/0809/10/news140.html. (Translated by Hitachi Automotive Systems).

ITmedia Inc. (2010) *The Basic Knowledge of LIN You Should Know, Part 1: IT Mono ist*, http://monoist.atmarkit.co.jp/mn/articles/1009/10/news104_3.html. (Translated by Hitachi Automotive Systems).

Kumagai, M. (2011) *Power Circuits and Actuators*, http://www.tohoku-gakuin.ac.jp/en/.

LIN Consortium (2006) LIN Specification Package Rev.2.1.

PSI5 Consortium (2011) Specification.

FURTHER READING

Melexis Microelectronic Systems (2008) MLX90324 Data Sheet.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto195
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Various Types of Sensors

Shinya Igarashi, Junji Onozuka, Hiroaki Hoshika, and Masahide Hayashi
Hitachi Automotive Systems, Ltd., Hitachinaka, Japan

1 Introduction 1
2 Airflow Sensor 4
3 Rotation Sensor 9
4 Knock Sensor 12
5 Temperature Sensor 13
6 Humidity Sensor 13
7 Combined Inertia Sensor (Acceleration and Angular Rate Sensor) 16
8 Conclusion 17
References 18

1 INTRODUCTION

Automotive electronics systems make the best possible use of multiple types of sensors and sensing technologies for each and every function. The required amount of detected physical quantity depends on the purposes and means for control and information processing. The selection of sensors depends on how best to materialize the purposes and means with an excellent cost performance. Therefore, when developing automotive sensors, the purposes and effectiveness of the system should be considered in relation to the interrelationship with other systems and the influence of the technology.

Automotive electronic systems were started to reduce the toxic gases in the engine exhaust and to try to achieve high engine efficiency and a reduced fuel consumption rate. As next stage, the system development went on to improve ride comfort and operability. These purposes have been spreading to safe driving performance, preventive safety, information communications, traffic system linkage, and autonomous and/or automatic driving.

The details of specific sensors are described in the following sections. Sensing systems can take various types and forms and their technologies have evolved to apply the advanced technologies. The discussions in the following sections cover only a part of such systems and technologies. Automotive electronic systems have been expanding and requiring more physical quantity detection or information collection. When the form of the system is determined, the sensor specification will be clear such as what physical quantity detection is needed with what level of accuracy and the environment conditions. However, the sensors suitable for the system are varied. There are many principles of detection and different materials are used. There is some difficulty in selecting the best sensor for the system.

The next point of consideration should be how to materialize the system with the lowest possible cost. From an opposite viewpoint, it is necessary to find a way to materialize the system with a selected sensor to propose and/or create the system with high performance and high functionality, which all automobiles should follow thereafter. Developing sensors and sensing technologies with low costs requires consideration from various standpoints, such as the following:

1. pursuit of new sensors and sensing technologies based on new materials and new theories;
2. sensor types and structures oriented toward size reduction and high mass productivity;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Table 1. Needs expected from sensors in automotive systems.

System	Sensor	Information collected
Engine exhaust/fuel consumption control	Air mass flow rate	Intake air (mass), engine load
	Manifold absolute pressure and flow rate	Intake air (volume) and engine load
	Atmospheric pressure	Altitude and density compensation
	Suction temperature	Temperature and density compensation
	Aspirated gas humidity	Ignition timing and combustion temperature compensation
	NO_X	NO_X emission
	Soot	Soot (particulate carbon) emission
	Fuel composition	Combustion generated heat Target air–fuel ratio
	Combustion pressure	Combustion process, status Combustion speed Engine output Knocking Cylinder stroke Combustion temperature
	Ion current	Combustion process, status
	Throttle opening angle	Torque requirement Intake air Idle engine speed
	EGR valve opening angle	EGR
	EGR pressure	EGR
	EGR flow	EGR
	Knocking (sound and vibration)	Knocking
	Fuel pressure	Fuel supply correction Mixture generation in combustion chamber
	Speed	Load
	Engine speed	Load
	Exhaust temperature	Combustion status
	Exhaust pressure	Combustion status and catalytic efficiency
	Fuel vapor pressure	Fuel vapor leak
	Fuel vapor purge	Fuel supply compensation
	Crank angle	Engine cycle time and engine start timing
	Cam angle	Intake valve opening angle
	Starter rotor angle	Engine start timing
	Torque	Engine/motor output
	Oil quality	Loss compensation, diagnosis
	Engine noise	Trouble diagnosis
Transmission control	Throttle position	Required vehicle speed
	Shift lever position	Transmission mode
	Transmission oil temperature	Transmission efficiency
	Transmission oil pressure	Transmission efficiency
	Drive speed	Transmission torque
Traction control	Engine speed	Transmission torque
	Steering angle	Transmission torque
	Steering changing rate	Transmission torque
ABS	Wheel speed	Vehicle speed and slip rate
	Brake switch	Brake mode
	Brake oil pressure	ABS pattern
	Acceleration/deceleration speed	ABS efficiency
	Brake oil level	ABS efficiency, diagnosis
	Yaw rate	ABS balance
Cruise	Vehicle distance	Cruise mode
	Vehicle speed	Cruise efficiency
	Brake switch	Cruise cancellation
	Front video camera	Front vehicle, oncoming vehicle, pedestrian, obstacle distance, and traffic lane
	Selection switch	Cruise pattern
Idle engine speed	Air conditioner clutch	Load
	Power steering oil pressure	Load
	Shift lever position	Idle engine speed mode
	Engine speed	Idle engine speed

(*continued overleaf*)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. (*Continued.*)

System	Sensor	Information collected
Suspension control	Vehicle height	Suspension mode
	Steering angle	Suspension mode
	Steering direction	Suspension mode
	Acceleration	Suspension mode
	Vehicle weight	Suspension mode
	Road surface information	Suspension mode
	Skid angle	Suspension efficiency
	Yaw rate	Suspension efficiency
	Horizontal displacement	Suspension efficiency
	Vehicle speed	Suspension mode
	Absorber pressure	Suspension efficiency
	Absorber load	Suspension efficiency
	Selector switch	Suspension mode
Airbag	Deceleration speed	Collision
	Bag pressure	Occupant protection efficiency
	Seat load Vehicle speed	Airbag deployment criteria
Protective safety	Front obstacle detection radar	Obstacle, front vehicle
	Front detection monitor	Lane, traffic condition information
	Lateral detection monitor	Lateral condition
	Rear detection monitor	Rear condition
	Vehicle speed	Warning level
Navigation	Gyro	Rotation angular velocity
	Steering rotation speed	Rotation direction
	Terrestrial magnetism	Direction
	Vehicle speed	Mileage
	Camera	Rear/bottom/side check
	GPS	Absolute position
Air conditioner	Cabin temperature	Air conditioner operation condition
	Cabin humidity	Air conditioner operation condition
	Cabin airflow	Fan operation mode
	Outdoor temperature	Air conditioner operation condition
	Air purity	Filter switching condition
Driver information	Vehicle speed	Indicator information
	Engine speed	Indicator information
	Oil pressure	Maintenance information
	Fuel level	Indicator information
	Oil level	Maintenance information
	Oil quality and condition	Maintenance information
	Cooling water pressure	Maintenance information
	Cooling water condition	Maintenance information
	Cooling water temperature	Indicator information
	Outdoor temperature	Indicator information
	Cooling water level	Indicator information
	Window washer level	Indicator information
	Transmission oil level	Maintenance information
	Tire air pressure	Maintenance information
	Tire surface temperature	Maintenance information
	Battery liquid level	Maintenance information
	Raindrops	Wiper mode
	Sunlight	Air conditioner mode
Automatic cruise	Road guide tracking	Road information
	Leak cable TR/RC	Road/vehicle distance guide
Traffic information and control	Light/radio beacon	Traffic information
	ETC	Road toll
Communication	Digital high speed data	Data communication service
	Digital audio	Car telephone
	Satellite communication	Satellite communication service
	Infrastructure information receiver	Environment and navigation

ETC, electric toll collection.
Reproduced by permission of Hitachi Automotive Systems.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3. sophisticated information processing and communication technology such as sensor fusion based on digital conversion and intelligence deployment;
4. improvement in electronics systems oriented toward fewer sensor allocations or no sensor allocations.

These standpoints may include some elements that are incompatible with each other, yet designers and developers are constantly trying—from different viewpoint—to achieve the lowest cost system possible and the necessities of such a system for controlling and information processing, which ought to be common objectives among them. As a result, only the best combination at the time can be found.

Table 1 lists the needs that should be fulfilled by sensors in automotive systems. These are selected as combinations and/or means for achieving the purposes of control and information processing systems. In recent years, the automobile driving systems involving electric motors in place of engines are applied to electric vehicles (EVs) and hybrid electric vehicles (HEVs), for example. For such applications, different sensing technologies are required. These are very different from conventional engine control and the needs expected from the sensors in such systems will become clear in the future. Furthermore, the viewpoints of driver information, safety control, automatic driving, and communication-related processes vary reflecting the times. Table 1 shows general trends.

2 AIRFLOW SENSOR

Airflow sensors are major sensors for electronics control fuel injection systems used for measuring the rate of airflow into cylinders. Electronics control fuel injection systems enable optimal fuel supply and ignition by operating fuel injection valves and ignition systems on the basis of intake of air signals. The air–fuel ratio, the weight ratio between air intake and fuel, is one of the most important factors that determine the exhaust and fuel-efficiency characteristics of engines. In recent years, high accuracy and long-term reliability are required from the air–fuel ratio control for emission improvement. Therefore, high accuracy is required of airflow sensors. Furthermore, the selection of airflow measurement methods causes a great to input configurations of computer systems. To satisfy the new functional requirements of future engine control systems, the manufacturers of parts and automobiles have been developing a large number of airflow detection methods.

2.1 Specifications of airflow sensors

Table 2 lists some of the specifications needed from airflow sensors in electronics control fuel injection systems. To materialize highly accurate air–fuel ratio control, airflow rate ought to be expressed in terms of mass flow rate. This requires a wide range of measurement coverage as well as high accuracy. Airflow, while operating with an almost fully opened throttle valve, becomes large-amplitude pulsating flow with the frequency of 20–200 Hz, which results from the reciprocating motion of a piston. In some extreme cases, such pulsating flow can be accompanied by reverse flow. Sensors are supposed to take highly responsive measurements of such pulsating flow in order to acquire an average flow rate. On top of that, the above-mentioned requirement needs to be satisfied in such a severe environment as is characteristic of automobile engines, where there may be backfires, dust in intake airflow, oil drops (blow-by gas), and a wide range of air temperature changes.

Table 2. Specifications of automobile airflow sensors.

Item	Specification
Measurement accuracy in mass airflow rate	±2%
Measurement range	1 : 100 or larger
Response time	10 ms or less
Pressure loss	5 kPa or less
Measurement environment	Air temperature: −40 to 130°C
Durability	Mileage: 150,000 km
Vibration resistance	20 G 20–200 Hz

Reproduced by permission of Hitachi Automotive Systems.

In addition, long-term reliability is required as well as compactness, lightness, and low cost. Currently, there are no such airflow sensors that satisfy all these specifications. In practical use, lacking aspects are supplemented with control systems.

2.2 Types of airflow sensors

Airflow sensors have been applied to practical use in the form of the moving vane type, the Karman vortex type, and the thermal type, named in chronological order.

The moving vane type represents the flow rate sensors applied in the earliest days (Figure 1). The kinetic pressure from airflow is received with a moving vane (a rotatable plate). The moving vane opens, opposing the force from the return spring. This opening angle is measured to obtain the volume flow rate in this type.

The opening angle of the moving vane is measured on the basis of its sliding resistor position. The value of the sliding resistor's measurements is matched with the area of air passage formed by the moving vane. Thus, the signals from the flow rate instrument are interrelated with

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

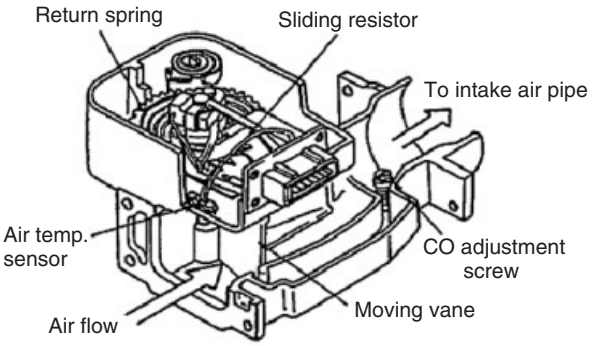


Figure 1. Moving vane-type flow rate instrument. (Reproduced by permission of Hitachi Automotive Systems.)

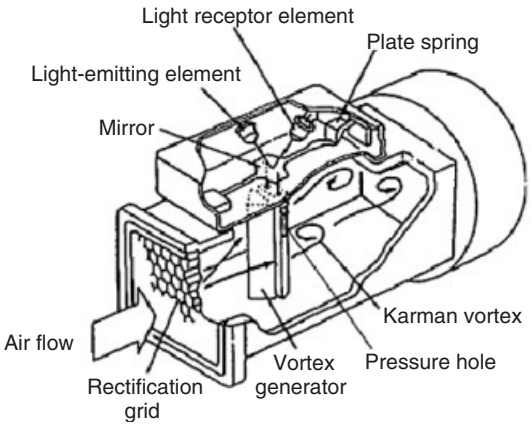


Figure 2. Karman vortex-type flow rate instrument. (Reproduced by permission of Hitachi Automotive Systems.)

the airflow rate values by means of logarithmic functions. There are, however, some boundaries with respect to durability and reliability that are brought by moving parts and sliding parts. Specifically, moving vanes are used with air damper chambers that suppress vibration. They make moving vanes larger in external dimensions and somewhat slower in response. In addition, there are some more shortcomings. For example, air pressure and temperature need to be compensated or corrected to obtain the mass flow rate.

Afterward, the Karman vortex type was applied in practical use. This type makes use of the fact that Karman vortices pass the downstream side of a vortex generator with regular intervals in proportion to the airflow rate. The frequency of the Karman vortices is measured to obtain the volume flow rate. As the means for detecting Karman vortex frequencies, the following methods have been adopted:

1. In the first method, an ultrasonic wave oscillator and a receiver oppose each other in the downstream of a vortex generator. Vortices are detected with synchronized modulation of ultrasonic frequencies generated by such vortices.
2. In the second method, the pressure variance resulting from vortex generation is led to a mirror by way of pressure holes on both sides of the vortex generator to vibrate the mirror. The vibration cycle of the mirror is detected with a light coupler (Figure 2).
3. In the third method, the pressure variance is measured with a micromanometer in the same manner as the above-mentioned method.
4. In the fourth method, vortices are measured with thermal wires.

This generation of Karman vortices, however, tends to be unstable at the time when they are with pulsating flow or low speed airflow. Therefore, it is not easy to measure large-amplitude pulsating flow or ultra-low flow rates. In addition, atmospheric pressure and temperature need to be used for compensation and correction to obtain mass flow rate, to name but a few disadvantages.

Following this method, thermal wire-type airflow sensors were applied. It is possible to directly detect the mass flow rate of air. If compared with the two methods mentioned earlier, the measurable range is wider than those of the above-mentioned methods. This type of sensor does not involve any moving parts; they are small and are manufactured at a low cost. This is why most of the airflow sensors are of this type nowadays.

Besides the above-mentioned types, there are corona discharge types, ultrasonic wave types, swirl vortex types, and so forth, which were not applied to engine control but were once considered for such applications. Other than the airflow sensors mentioned earlier, the speed density type, which acquires airflow rate indirectly from intake pipe pressure and engine revolution speed, and the throttle speed type, which acquires airflow rate indirectly from throttle valve opening angle and engine revolution speed, are applied to practical use. The former has a disadvantage; that is, the measurement accuracy of airflow rate deteriorates when exhaust air returns. Thanks to its low cost and high reliability; however, this type has been widely applied to such fuel injection systems that have a small rate of exhaust return and have no influence over measurement accuracy. The latter, on the other hand, are not so widely applied so far firstly because high accuracy is required of the throttle valve opening angle and secondly because the measurement accuracy of airflow rate deteriorates if atmospheric pressure varies. However, it is possible to compensate by Electronic Control Unit (ECU), so that the trend is variable.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.3 Thermal wire-type airflow sensor

2.3.1 Principle of thermal wire-type airflow sensors

Thermal wire sensors supply electric current to a heat resistor made of platinum wire or platinum thin film in order to heat the resistor. This type makes use of the fact that the amount of heat transmitted to the air depends on airflow speed. As shown in Figure 3, a thermal wire probe, which detects airflow rate, and a temperature probe, which detects air temperature, are on the sides of a bridge circuit. The electric current supplied to the thermal wire probe is increased or decreased to regulate the temperature difference between the thermal wire probe and the air temperature probe regardless of airflow rate. At this time, the electric current supplied is converted into an airflow rate signal in this method. The interrelationship between the airflow rate and a signal is expressed by the following equation:

$$I^2R = A + B\sqrt{G_a} \tag{1}$$

where I is the electric current supplied, G_a is the airflow rate, and A and B are constants. In this method, the electric current is measured, and the above-mentioned equation is used to acquire airflow speed, G_a. When flow rate is low, electric current I, the measurement to be converted into a flow rate signal, varies a great deal, getting closer to showing logarithmic characteristics.

In Figure 4, the rectifying cylinder in the intake air passage contains the platinum wire of 70 μm in diameter spanned across the rectifying cylinder as the thermal wire probe. Further, the platinum thin film resistor formed on the ceramic plate is used as the air temperature probe. After the engine key is turned off, electric current flows to the thermal wire probe to heat it up to approximately 1000°C, thus burning and removing any adhering dust particles. This burn-off circuit is provided to prevent any adhering dust particles from degrading accuracy.

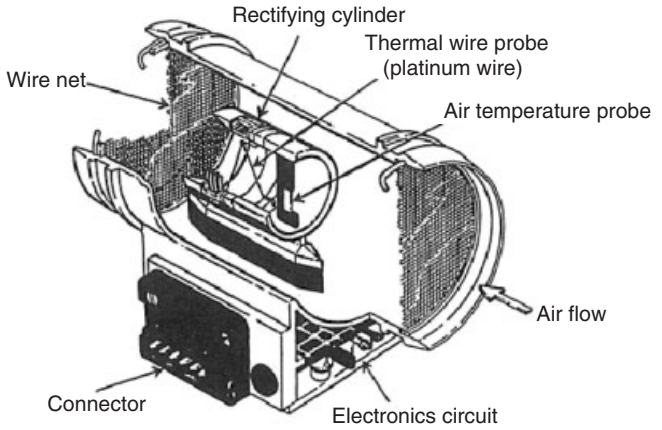


Figure 4. Thermal wire-type flow rate instrument. (Reproduced by permission of Hitachi Automotive Systems.)

Figure 5 shows a bypass type, in which a thermal wire probe and an air temperature probe are disposed in a bypass passage branching from the intake air passage. The purposes of this bypass type are to take highly accurate measurement of pulsating flow and protect the thermal wire probe against the backfire at the time of engine disorder. Both the thermal wire and the air temperature probes are of a wire wound type, which have a platinum wire wound in

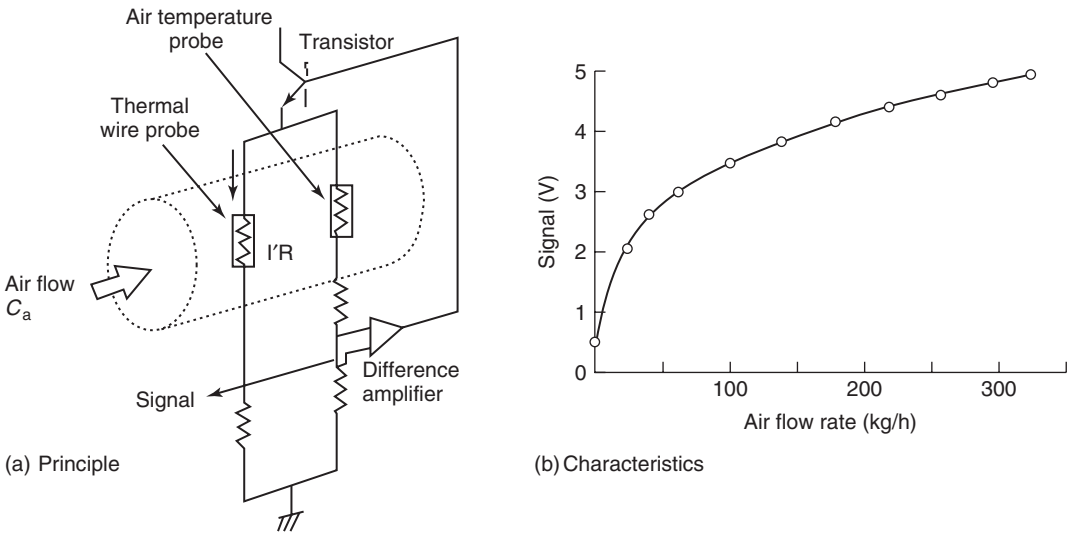


Figure 3. (a,b) Principle of the thermal wire-type flow rate instrument. (Reproduced by permission of Hitachi Automotive Systems.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

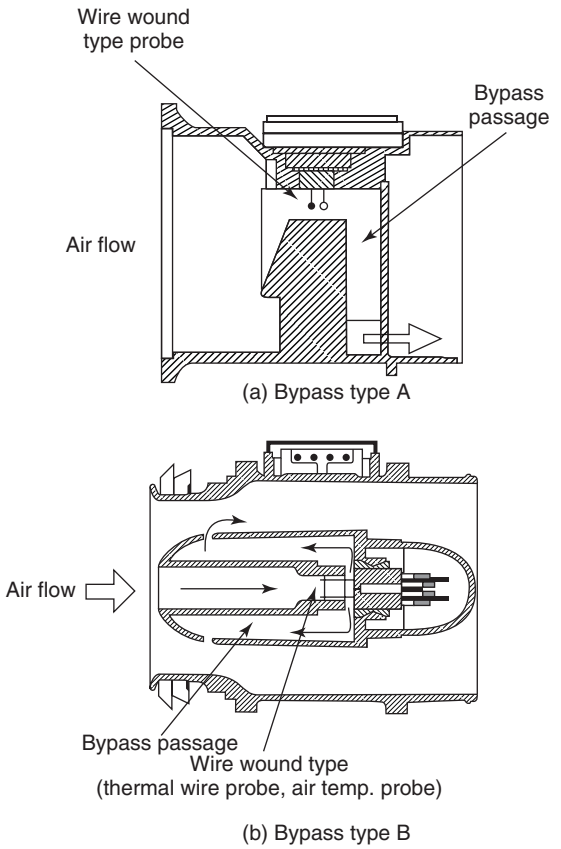


Figure 5. (a,b) Bypass-type thermal wire flow rate instrument. (Reproduced by permission of Hitachi Automotive Systems.)

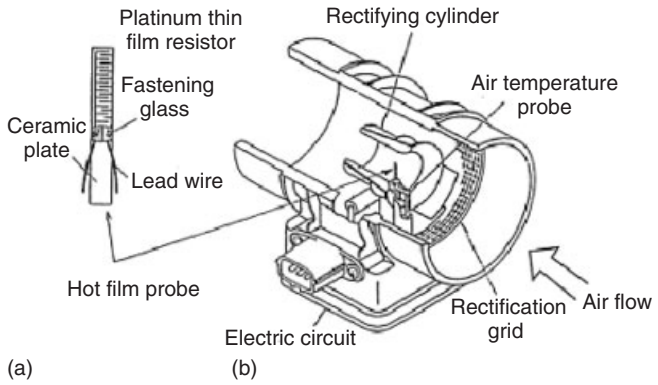


Figure 6. (a,b) Hot film-type airflow instrument. (Reproduced by permission of Hitachi Automotive Systems.)

the shape of a coil around the cylindrical ceramic bobbin that is covered with glass.

Figure 6 shows a hot film type, which has a rectifying cylinder in the intake air passage containing a hot film plate filmed with a platinum thin film resistor on a thin ceramic plate. The hot film plate is inclined by approximately 30° with respect to airflow, letting airflow go along the plate. This is an attempt to reduce the noises emitted from flow rate signals triggered by the disturbance in airflow.

Figure 7 shows an insert type, which can be attached in any position in an intake air passage. The thermal wire probe, the bypass passage, and the electronics circuit are configured all in one and inserted into the sidewall of the intake air passage. The purpose of this type is to make the intake air pipe compact, including the airflow sensor. The differences in the cross-sectional areas of intake air pipes are automatically compensated by the air–fuel ratio closed-loop control with an air–fuel ratio sensor and the memory of the compensation values of this closed-loop control.

2.3.2 Structure of thermal wire-type airflow sensor

2.3.2.1 External shapes of sensors. Among those that have been applied to practical use, there are, from the structural viewpoint, the hot wire type, which directly spans a platinum wire in airflow; the wound wire type, which has a ceramic bobbin wound with a platinum wire; and the hot film type, which contains a platinum thin film resistor on a ceramic substrate. The hot wire type uses a platinum thin wire. Therefore, its thermal capacity is small and its response is excellent. However, the dust particles accumulating on a thin wire tend to result in some deterioration in the accuracy of flow rate measurement. On the other hand, the wire wound type and the hot film type have a large thermal capacity and, therefore, their responsiveness is inferior. Yet, the advantage of these types is that the influence of dust particle accumulation is much less and, therefore, the accuracy is less deteriorated.

The accumulation of dust particles on a thermal wire probe is affected—besides the shape of the probe—by the temperature of the thermal wire probe. While dust particles accumulate on the surface of a probe, it is mainly the moisture and the oil in the dust particles that attract the dust particles to one another. With this problem considered, the temperature of the thermal wire probe is raised to 200–250°C to vaporize the moisture and the oil on the surfaces of the dust particles, reducing the amount of accumulating dust particles. The wire wound type has a wire wound-type probe with some smaller surface area. Accordingly, it is practically easy to raise the temperature as mentioned earlier.

2.3.2.2 Electronics circuit. The temperature difference between a thermal wire probe and an air temperature probe is constantly controlled to a prescribed level, thus compensating any variance in air temperature. Besides the conventional electronics circuit shown in Figure 3, another

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

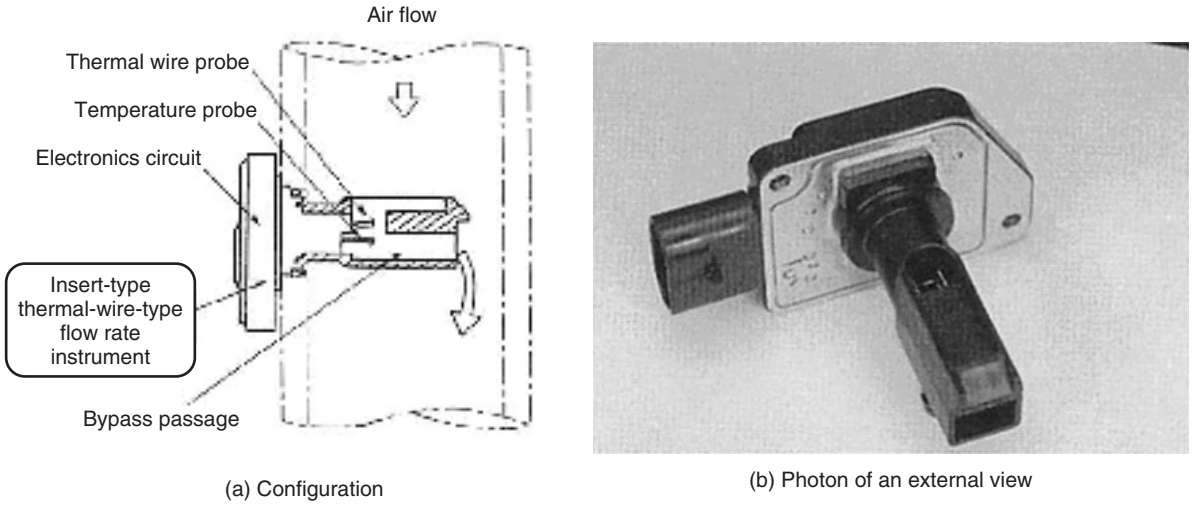


Figure 7. Insert-type airflow rate instrument. (Reproduced by permission of Hitachi Automotive Systems.)

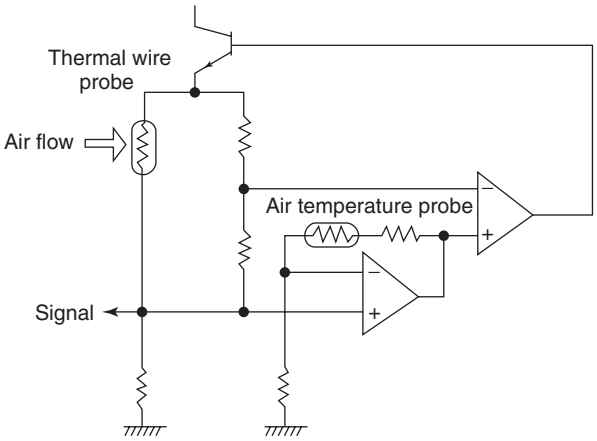


Figure 8. An example of constant temperature difference circuits. (Reproduced by permission of Hitachi Automotive Systems.)

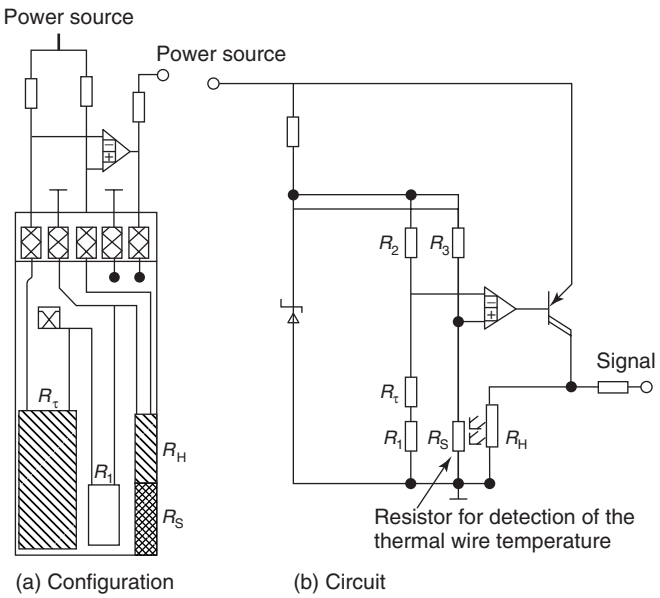


Figure 9. (a,b) Electronics circuit of a hot film-type flow rate instrument. (Reproduced by permission of Hitachi Automotive Systems.)

new electronics circuit has been developed as described later.

Figure 8 shows a circuit where another amplifier is added to a conventional circuit to make it possible to use probes having the same resistance for the air temperature probe and the thermal wire probe. As opposed to conventional ones, this air temperature probe causes less resistance and is more compact. Consequently, even if the air temperature changes suddenly, it is detected with no delay, and, therefore, it is possible to compensate the temperature change. Moreover, when it is assumed that the heat conduction from the probe-fastening end deteriorates the accuracy, identical shapes can be given to both the air temperature probe and the thermal wire probe, offsetting the influence from the heat conduction and hence reducing such influence.

Figure 9 shows a hot film-type circuit, in which a resistor is added to detect the temperature of the heat element to omit a fixed resistor having been supposed to be connected to the heat element in series. The sensor module is somewhat complicated, as the heat elements and the temperature-detecting resistors are stacked and disposed

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

there. Nevertheless, there is an advantage. As no fixed resistors that require cooling means are built-in, the electronics circuit is simplified compared to the circuits that come with a heat element emitting a great amount of thermal energy.

2.3.2.3 Bypass passage. The bypass type measures the flow rate in a bypass passage that branches off from an air passage. This type reduces the error in average flow rate measurement of pulsating airflow. When pulsating flow is measured with a thermal wire-type flow rate instrument, the following two problems could be faced.

2.3.2.3.1 Averaging the average flow rates of pulsating flow. As far as pulsating flow is concerned, the average of flow rate signal is smaller than the actual average airflow rate. One cause of this is, as shown in Figure 3, that the signals from a thermal wire-type flow rate sensor show nonlinear characteristics with respect to airflow rate. The larger the pulsating flow amplitude is, the slower the response is and the smaller the measurement will be than the actual value.

Another cause is that the amplitude is large enough to be accompanied by small reverse flows. When this happens, and if the air heated up by the thermal wire probe flows back in the opposite direction, then the air covers the thermal wire probe once again. Thus, the power supply to the thermal wire probe becomes higher, which means that the flow rate signals become higher. The bypass type uses the effect of the increase in the average flow rate, brought about by the inertia of the airflow in the bypass passage, in order to correct the error mentioned earlier. Airflow rate u in a bypass passage is acquired as follows:

$$\Delta P \sim \frac{u^2}{2} + \ell \frac{\mathrm{d}u}{\mathrm{d}t} \tag{2}$$

Here, ΔP is the pressure difference added to the inlet/outlet of the bypass passage that is generated by the airflow in the intake air passage, u the flow rate in the bypass passage, and ℓ the length of the bypass passage. The length of the bypass passage, ℓ, is determined such that the increase in the average flow rate in the bypass passage offsets the error mentioned earlier.

2.3.2.3.2 Reverse flow compensation. Normally, thermal wire-type airflow sensors cannot detect the direction of the flow. These sensors, even if the airflow travels in the reverse direction, will output the same signals as those at the time of the airflow traveling in the forward direction. Therefore, when an average airflow rate is acquired on the basis of actual measurement, reverse flow is measured in the same manner as forward flow. Consequently, the measurement can be larger than the actual average.

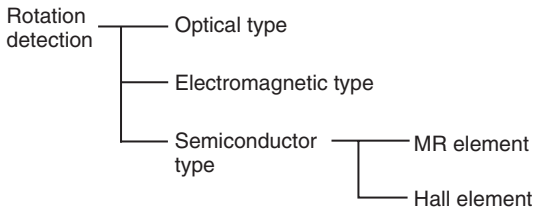
The bypass sensor type makes use of a strongpoint of fluid diodes to prevent airflow from traveling in a bypass passage in the reverse direction. As a result, the error is reduced to a half of the conventional one. To avoid the measurement error resulting from such reverse flow, you must subtract the flow rate of reverse airflow from that of the forward airflow, thus obtaining an actual flow rate. In order to take highly accurate measurements of average flow rate of the pulsating flow, including reverse direction flow, a thermal-type flow rate sensor is applied to a practical use, which acquires the direction and the flow rate of the flow at the same time. This type of sensor is described in a separate section about microelectromechanical systems (MEMS) airflow sensors (Section 6.4.7), together with semiconductor sensors.

3 ROTATION SENSOR

The purposes of rotation sensors range over various things such as detection of engine rotational angles, rotational speed of transmissions, and rotational speed of tires. In this section, examples such as crank angle sensors and cam angle rotation sensors used for the ignition timing of gasoline engines are described.

3.1 Types and principles of detection

Described below is a familiar detection method of rotation sensors.



3.2 Optical type

This is a type of rotation sensor used in the earlier days of electronic sensors. The optical type has a basic structure in which a distributor and a base component supply an ignition plug with power and the crank angles are detected with a light-emitting diode and a photodiode across a slit plate. Figure 10 shows a schematic view of its structure.

3.3 Electromagnetic type

Electromagnetic induction is used based on the detection principle. A coil and a magnet work in combination with

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

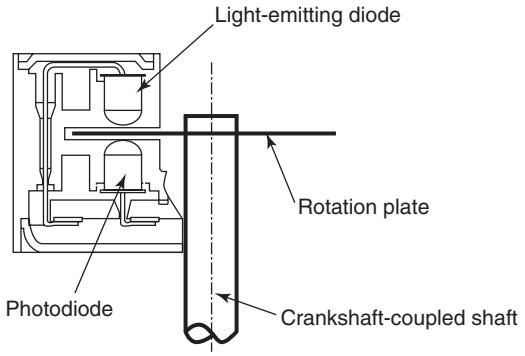


Figure 10. Optical-type rotation sensor. (Reproduced by permission of Hitachi Automotive Systems.)

each other. As a magnetism detector (gear or the like) moves, the magnetic flux varies, which enables the coil to generate electromotive force. The following equation expresses the relationship between the electromotive force and the rotational speed.

$$E = \frac{\Delta B}{L^2} \times N \times v$$

Here, E is the electromotive force, ΔB the change in the magnetic flux generated with a detection tooth, L the distance of the gap from the detection tooth, N the number of coils, and v the rotation speed of the detection tooth.

The electromagnetic type has several advantages in that the structure is simple and no power supply is necessary. There are, on the other hand, several disadvantages, for example, detection is not easy when the rotative object is standing still, the phase with respect to the rotative object varies depending on the rotational speed, and the temperature characteristics vary in a wide range.

Figures 11 and 12 show the structure and output characteristics of the electromagnetic type.

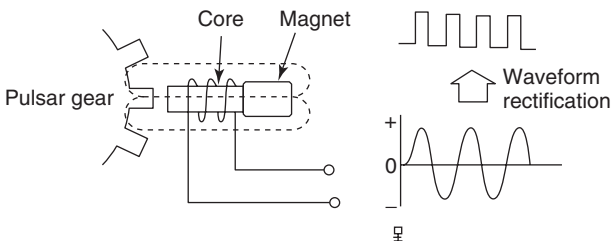


Figure 11. (a,b) Electromagnetic-type rotation sensor. (Reproduced by permission of Hitachi Automotive Systems.)

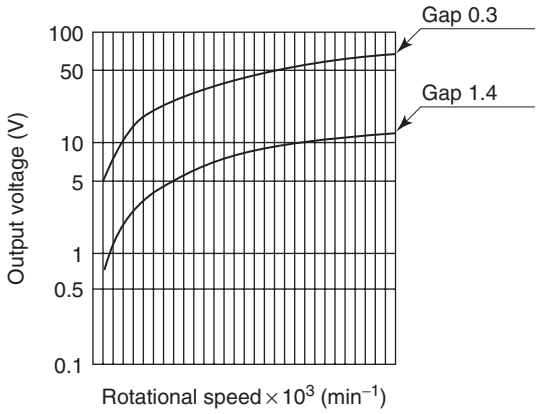


Figure 12. Output characteristics of the electromagnetic-type rotation sensor. (Reproduced by permission of Hitachi Automotive Systems.)

3.4 Semiconductor type

The advantage lies with its sensing element. MR and Hall elements are used for sensing. The mainstream applications contain one chip, which includes signal-processing circuits based on LSI (large scale integrated circuits) for compensation of temperature characteristics of elements, correction of rotation phases, and so forth.

Figure 13 shows the detection principle of a Hall element.

Figures 14 and 15 show an example of IC circuit structures (Sanken Electric Co, Ltd, 2007) and a block diagram (Sanken Electric Co, Ltd, 2007).

3.5 Structure

Rotation sensors are used in an environment where it is hot and oil is scattering. Therefore, housings are made with resin materials, which are resistant to high temperatures.

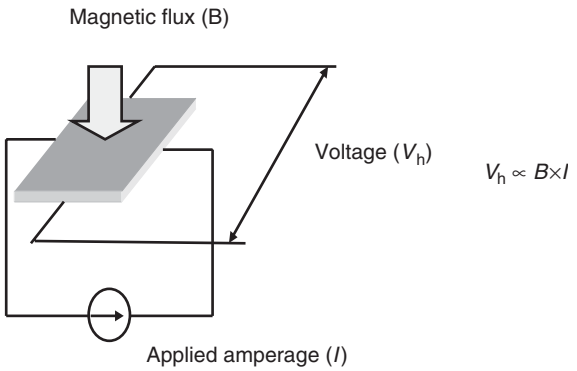


Figure 13. Detection principle of the Hall element. (Reproduced by permission of Hitachi Automotive Systems.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

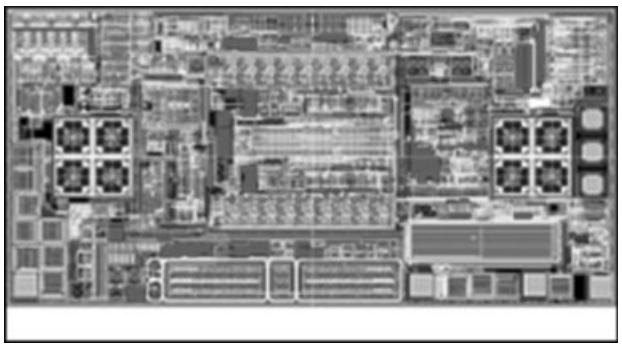


Figure 14. An example of the circuit structure of a semiconductor-type rotation sensor chip. (Reproduced by permission of Hitachi Automotive Systems.)



Figure 16. Examples of rotation sensors. (Reproduced by permission of Hitachi Automotive Systems.)

In addition, to be fit for the environment with scattering oil drops, housings are often equipped with an O-ring. Furthermore, to prevent erodent substances from infiltrating into the electronic circuit, duplicated housing structures may be formed, for example, with laser welding on interface surfaces to ensure reliability (Kato, 2010). Figure 16 shows examples of the external shapes of rotation sensors (Kato, 2010).

3.6 Characteristics of rotation sensors

Rotation sensors do not make contact with a rotating body to detect rotation. Therefore, the larger the distance (gap) from a rotating body, the more freedom is available for installing those sensors. As the distance (gap) depends on the magnetic flux density of the magnet, the sensitivity of the Hall IC, and the tooth shape of the rotating body,

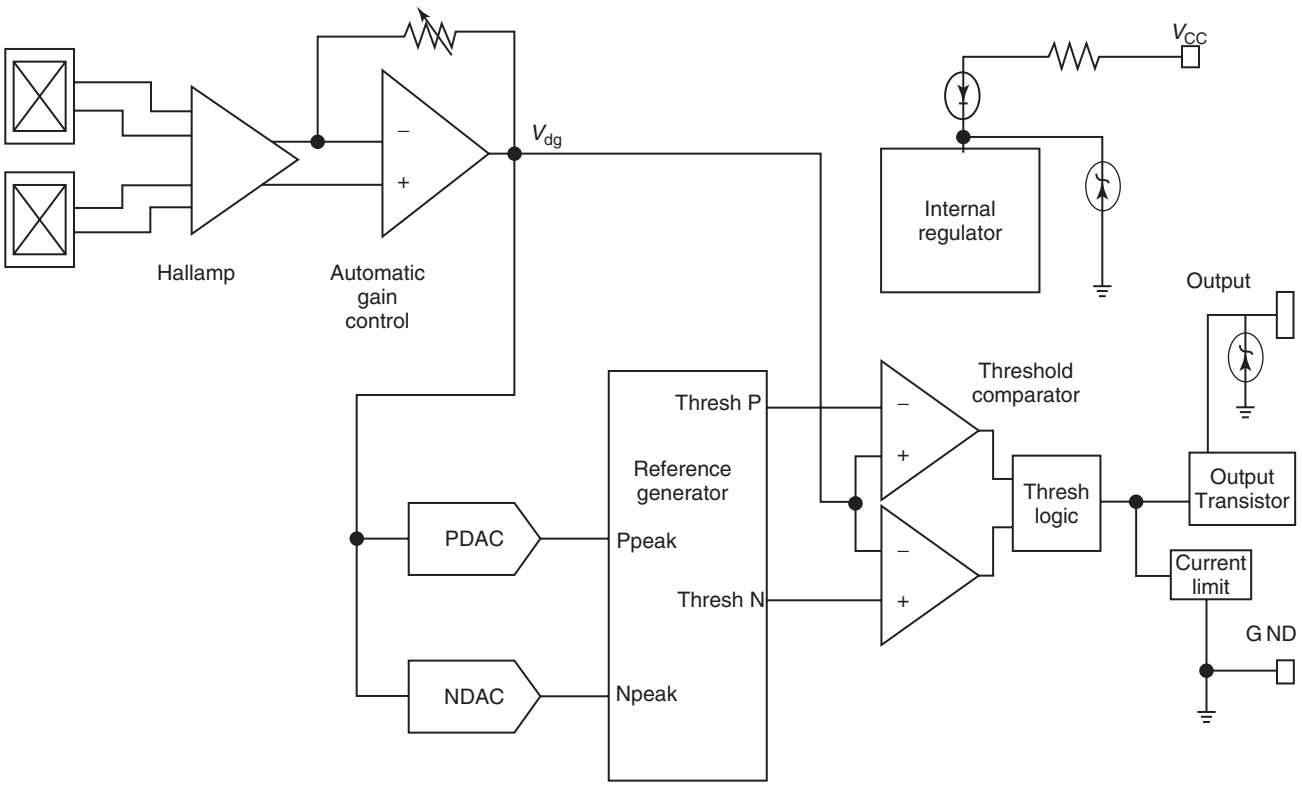


Figure 15. An example of the sensor block diagram of a semiconductor-type rotation sensor. (Reproduced by permission of Hitachi Automotive Systems.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

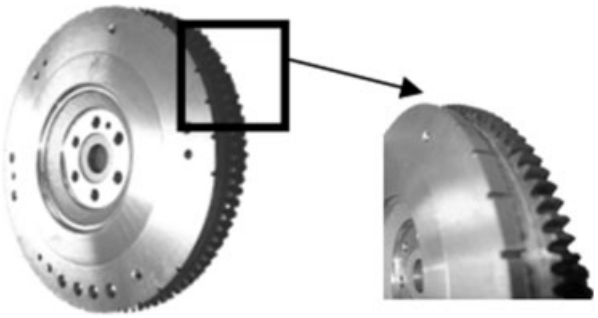


Figure 17. An example of pectinate teeth for detecting crank angles. (Reproduced by permission of Hitachi Automotive Systems.)

appropriate designs need to be made. In general, the distance (gap) is approximately 1 mm.

Figure 17 shows an example of the tooth shapes for the detection of crank angles.

4 KNOCK SENSOR

Knock sensors detect the vibration generated by abnormal combustion in a cylinder when knocking occurs in the engine. These sensors are used for engine control (ignition advance control). Knocking occurs in engines, varying depending on factors such as compression ratio, spark advance angles, fuel octane numbers, intake air temperature, humidity, or engine load. Therefore, an appropriate advance angle control needs to be performed to suppress knock generation.

4.1 Detection principle

Knock sensors use piezoelectric elements [lead zirconate titanate ($Pb(Zr, Ti)O_3$)] to detect vibration, which is subject to the piezoelectric phenomena. Knock sensors can be one of the two types: the resonance type, which has the same vibration frequency as that of the engine, and the nonresonance type, which does not have such vibration frequency. The resonance type has different resonance frequencies depending on engine types. Accordingly, the resonance of sensors needs to be adjusted for specific engine types. Nowadays, nonresonance type is in the mainstream.

4.2 Structure

To convert the vibration energy into pressure (physical force) to transmit it to piezoelectric elements, the piezoelectric elements have a structure in which appropriate weights are stacked. Its electromotive force is generated basically according to the following equation:

$$E = \alpha \times G \times W$$

Here, E is the electromotive force, α the vibration acceleration, G the piezoelectric constant, and W the mass of a weight.

When knocking occurs, the frequency is as high as 10 kHz or even higher. Consequently, the elasticity of the

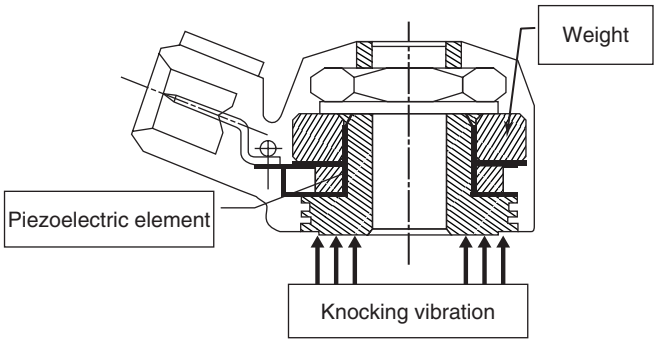


Figure 18. An example of cross-sectional structures of a knocking sensor. (Reproduced by permission of Hitachi Automotive Systems.)



Figure 19. Internal structure of a knock sensor. (Reproduced by permission of Hitachi Automotive Systems.)



Figure 20. An example of an external view of a knock sensor. (Reproduced by permission of Hitachi Automotive Systems.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

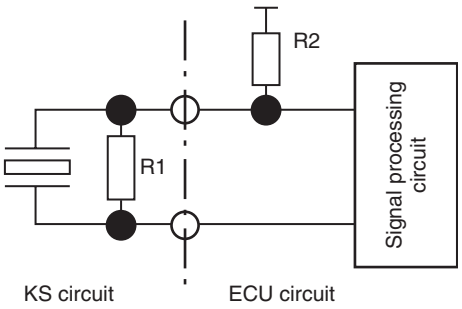


Figure 21. Electric circuit of a knocking sensor. (Reproduced by permission of Hitachi Automotive Systems.)

materials constituting the sensor tends to have an influence on that sensor. The resin material of the housing, in particular, is selected such that it does not have an influence on the frequency characteristics. Figures 18–21 show the structure and the electric circuit of a knock sensor.

5 TEMPERATURE SENSOR

The purposes of temperature sensors cover a wide range of applications such as the measurement of water temperature, engine-oil temperature, intake-air temperature, and EGR (exhaust gas recirculation) recirculating gas temperature. Temperature sensors for exhaust gas are under influence of high temperature. Therefore, platinum thermometers and thermocouples are applied.

5.1 Detection principle

To detect temperature, NTC (negative temperature coefficient)-type thermistors are mainly used. The characteristics of the thermistors depend on temperature ranges and detection sensitivity.

Thermistors vary their electric resistance exponentially with respect to temperature. Accordingly, it is necessary to make conversion into such signals that are easy for the ECU to process. In general, partial pressure circuits with series resistors (Figure 22) are widely applied. The relationship between a thermistor and the resistance is expressed by the following equation:

$$R_T = R_0 \times \exp\left\{B \times \left(\frac{1}{T} - \frac{1}{T_0}\right)\right\}$$

Here, R_T is the thermistor resistance at temperature T, R_0 the thermistor resistance at reference temperature T_0, and B a constant.

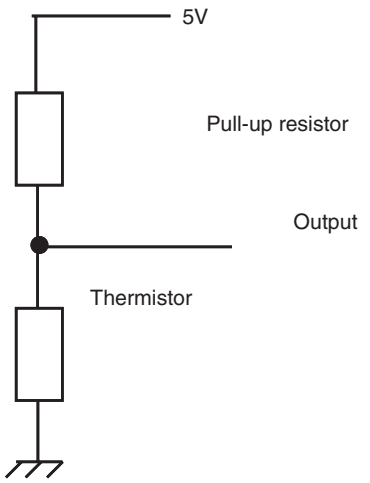


Figure 22. Partial pressure circuit with resistors. (Reproduced by permission of Hitachi Automotive Systems.)

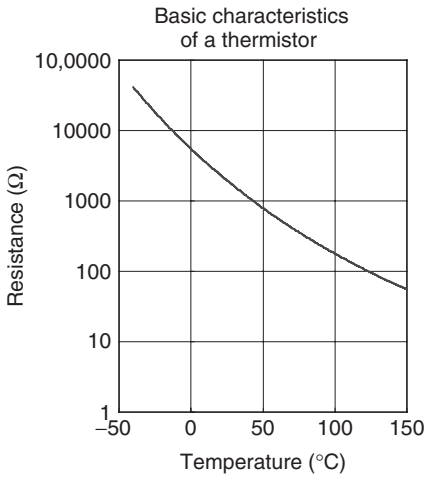


Figure 23. An example of thermistor characteristics. (Reproduced by permission of Hitachi Automotive Systems.)

Figures 23–25 show examples of the characteristics of intake-air thermistors (Mitsubishi Materials Corporation, 2008), the characteristics of a partial pressure circuit with resistors, and its structure.

6 HUMIDITY SENSOR

6.1 Introduction

There is a growing awareness of global warming and resource saving. The endeavors for the improvement of car fuel mileage seem to be a never-ending challenge. Besides powertrain management, the energy management

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

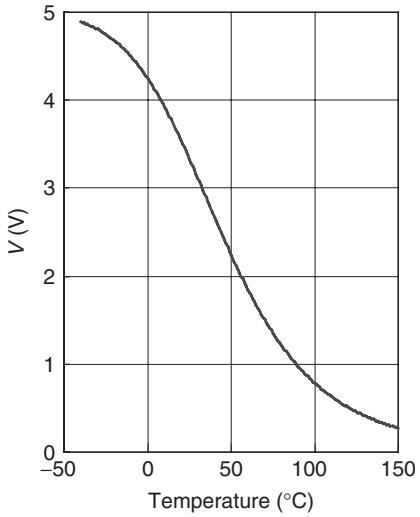


Figure 24. An example of temperature sensor characteristics. (Reproduced by permission of Hitachi Automotive Systems.)

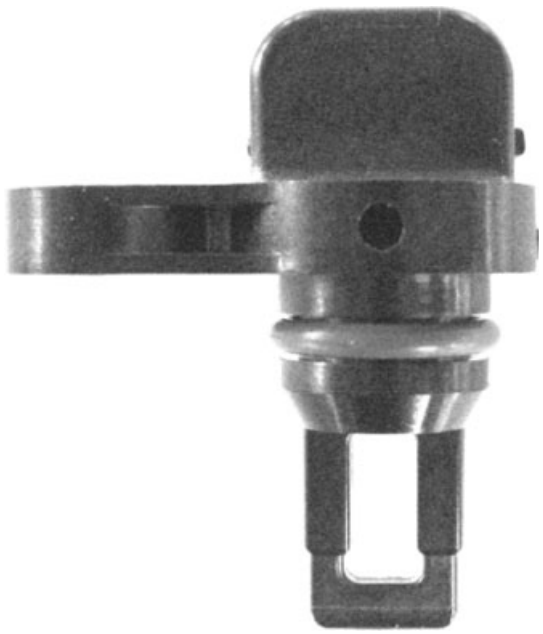


Figure 25. An example of an external view of a temperature sensor. (Reproduced by permission of Hitachi Automotive Systems.)

of air-conditioning systems and controls for minimizing front/rear glass frosting are also among the purposes of the improvement of the balance between the safety and the comfort. In recent years, more and more specific control is required for vehicles and powertrain systems. The relative humidity of the atmosphere is measured to improve the fuel mileage and reduce the cost of vehicles as well as to improve the safety and the comfort. All these efforts are picking up. In this section, we discuss the humidity sensors installed on vehicles for such purposes as mentioned earlier.

6.2 Overview of humidity sensors

Among the humidity sensors in practical use as of today, most are for the purpose of measuring the relative humidity of the atmosphere. They can be independent products or integrated with some other sensors for an application of practical use. Several major purposes are listed as follows.

6.2.1 Air-conditioning/cabin systems

Relative humidity is an important application. This is the detection of the humidity of the atmosphere let in from the outside, the humidity control in a cabin, the prevention of the front glass from being frosted (the temperature of the glass and the humidity in the cabin are detected to control dehumidification and/or rear-glass heaters before glass frost is formed), and so forth.

6.2.2 Power units and relevant components

Absolute moisture content is determined, usually to compensate the mixture ratio for moisture in the intake air to an engine. This is for ignition advance compensation/control, EGR rate control (the EGR rates that make combustion unstable will vary depending on humidity), estimation of combustion temperature, estimation of emissions of harmful substances in the exhaust (the combustion temperature varies depending on humidity, generating a different amount of harmful substances in different compositions, which are to be estimated), compensation of engine-generated torque, and so forth. Besides these, there are various purposes of use.

6.2.3 EV and so forth

A humidity sensor is disposed in a battery unit. This is for the safety management, for minimizing the energy consumed by the air conditioner system to improve the cruising range, and/or for some other purposes.

6.3 Classification of humidity sensors

In the household sector, various types of hygrometers and humidity sensors are used. In the automotive field, the humidity sensors should satisfy the requirements such as

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Principles, use purposes, and characteristics of typical humidity sensors.

Classification	Principle of detection	Use purpose and characteristics
Macromolecule capacitance type	Polyimide system polymer materials have dielectric constant ε varying due to moisture absorption, forming a condenser structure. Its capacity variance is measured to detect the relative humidity	Engine control, air-conditioning systems, and so on This type is widely used in the household sector and the automotive field, providing the accuracy sufficient for the practical use in a wide temperature range. With some ingenuity in the electrode structure, some sensors are resistant to dew condensation. Others have a sensing module, temperature sensor, correction circuit, and so forth integrated on a conductor chip and are already in practical use. This type reacts to some types of volatile gases. An adequate check is necessary
Macromolecule resistor type	The wet and dry polymers with additives are coated on a pair of electrodes. The free ions from the additives vary the AC resistance, which is measured to detect the relative humidity	Air-conditioning systems This type is less costly and has made achievements in a wide range of fields in the household sector. Its accuracy in the low temperature and low humidity range, however, is not very good in general. As the characteristics vary due to dew condensation, an antimoisture coating needs to be applied, or this type should be used in an environment where dew condensation does not occur

Reproduced by permission of Hitachi Automotive Systems.

the operation in a wide temperature range and the long-term reliability, so much so that the sensors with inherent durability are used. In addition, it is necessary for humidity sensors to be used in consideration of a variety of factors such as a waterproof structure, the availability of signals for external components, and the integration with some other functions. Table 3 lists typical humidity sensors in practical use for automobiles.

There are various kinds of forms to implement these sensors. Small ones are widely used. When absolute moisture is used in particular, it is necessary to calculate saturated vapor pressure. Therefore, you need to dispose a temperature sensor very close to a relative-humidity detecting component. Figure 26 shows several examples of automotive humidity sensors in practical use.

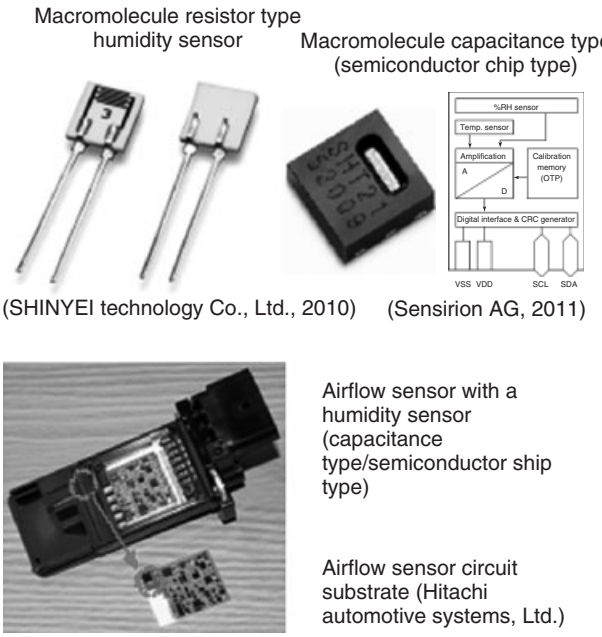


Figure 26. Examples of humidity sensors. (Reproduced by permission of Hitachi Automotive Systems.)

6.4 Characteristics of humidity sensors

For the evaluation of the performance of automotive humidity sensors, the following items, in general, can be included among the criteria.

6.4.1 Detection range of relative humidity

Normally, it is necessary to guarantee operation, in general, in the temperature range from −40 to 125°C for powertrain systems and from −40 to 85°C in a cabin. The functions should work at any temperature within these temperature ranges.

6.4.2 Accuracy of relative humidity

In the field of household products, the tolerance in the measurement of relative humidity is determined, usually, within the temperature range 20–25°C. In the field of automobile-related products, on the other hand, one needs to know the real-world tolerance across the entire temperature

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

range of the guaranteed operation. You should make use of a humidity generator or the like at low and high temperature ranges to take highly accurate measurement.

6.4.3 Responsiveness

An automobile can be under influence of radical temperature/humidity changes while driving and the responsiveness varies depending on the humidity. Therefore, a sufficient preparatory study of the application of a sensor is required before using it. Not only must the humidity sensor per se be considered but also the response lags caused by an antimoisture coating. Moreover, if absolute moisture is going to be acquired, the balance with the temperature sensor is important, too.

6.4.4 Hysteresis

Some minor amount of hysteresis is observed as a result of changes from a low humidity to a high humidity and vice versa.

6.4.5 Over-time changes

As macromolecules are used, the characteristics change over time, although such change is small. Automotive sensors ought to be subject to such changes only within a sufficiently limited area from a practical viewpoint.

6.4.6 Characteristics of recovery from dew condensation or freezing

Some sensors have an antimoisture coating, and others have detection macromolecules directly exposed. They are different in characteristics. An adequate examination of whether the characteristics change after the sensing module is under influence of dew condensation, or whether it has been frozen if the sensor has detection macromolecules exposed directly, is needed.

6.4.7 Durability

It is necessary to adequately assess whether the sensor is resistant to various types of gasses and chemicals that could be generated in the automotive environment such as high temperature and high humidity, thermal shock, vibration, temperature cycles, dew condensation, and exhaust gas durability.

6.4.8 Miscellaneous

When it is necessary to acquire the amount of moisture, for example, for the calculation of the dew point or the mixture ratio, the temperature around a humidity sensor is extremely important. Therefore, a thermometer is normally disposed in the vicinity of the humidity sensor. Particularly, when the temperature is high, attention should be paid to the accuracy of the temperature sensor and where the temperature sensor should be located as well as the accuracy of the humidity sensor.

7 COMBINED INERTIA SENSOR (ACCELERATION AND ANGULAR RATE SENSOR)

7.1 Outline and application of sensors

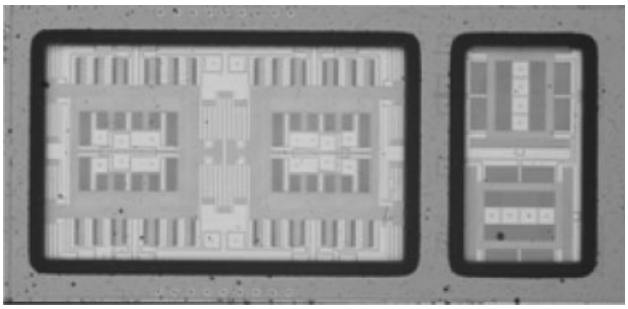
Combined inertia sensors take measurement of angular rate (rotation) and acceleration. These sensors have been already applied in some considerable amount to those systems that prevent vehicle skids. In the future, it is expected that the field of application will be expanding to such fields as suspension and rollover controls.

The MEMS semiconductor micromachining technology is used to integrate angular rate sensors and acceleration sensors on one SOI (silicon on insulator) substrate in two or more parts (Figure 27).

7.2 Principles of operation

7.2.1 Angular rate sensor

Figure 28 shows the operation principle of the angular rate sensor. Drive force F_d is generated by the electrostatic



(a) Angular rate sensor (b) Acceleration sensor module

Figure 27. (a,b) MEMS-combined sensor elements. (Reproduced by permission of Hitachi Automotive Systems.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

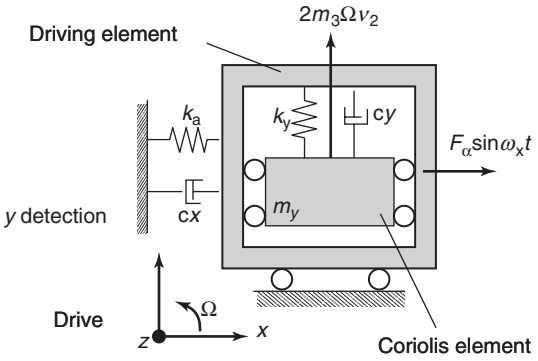


Figure 28. Operation principle of the angular rate sensor. (Reproduced by permission of Hitachi Automotive Systems.)

force. It excites the driving element in the drive direction (in the *x* direction). At this time, as the driving element is excited, the Coriolis element, which is connected with the driving element by the spring, starts to vibrate, too. While this is happening, an angular velocity around the *z*-axis is added. The Coriolis force is generated in the driving element being vibrated and the Coriolis element in the detecting direction (in the *y*-axis direction) orthogonal to the drive direction (in the *x*-axis direction) and the axis with the angular rate impressed (to the *z*-axis). As a result, the Coriolis element shifts itself in the detecting direction (in the *y*-axis direction) in proportion to the degree of the impressed angular velocity. The distance of such shift is detected as the variance in capacitance. As this variance is output as a signal, the impressed angular rate is detected (Hayashi, 2009).

7.2.2 Acceleration sensor

Figure 29 shows the operation principle of the acceleration sensor. As the acceleration is added, moving part m_y shifts. Two differential capacitance detection electrodes C1 and C2, disposed in the detecting direction, change their capacitances. Such change is output as an electric signal representing the acceleration.

7.3 Signal processing circuit

Figure 30 shows the constitution of a signal processing circuit. The MEMS element, connected with the analog end of the signal processing circuit, sends a signal, which is converted into a digital signal by the $\Delta\Sigma$ converter. This signal is processed by the Digital Signal Processor (DSP) for synchronous detection and feedback (Jeong, 2011).

7.4 Structure

Combined inertia sensors are installed in various places such as the inside of an engine compartment or the inside of a passenger compartment. Therefore, depending on the location in the vehicle, their structures are selected for implementation, for example, from ceramic packages, plastic packages, and transfer mold packages (Figure 31).

8 CONCLUSION

Sensors and sensing technologies, which have been developed and improved over long period of time, are now essential elements for improving the performance, safety, and

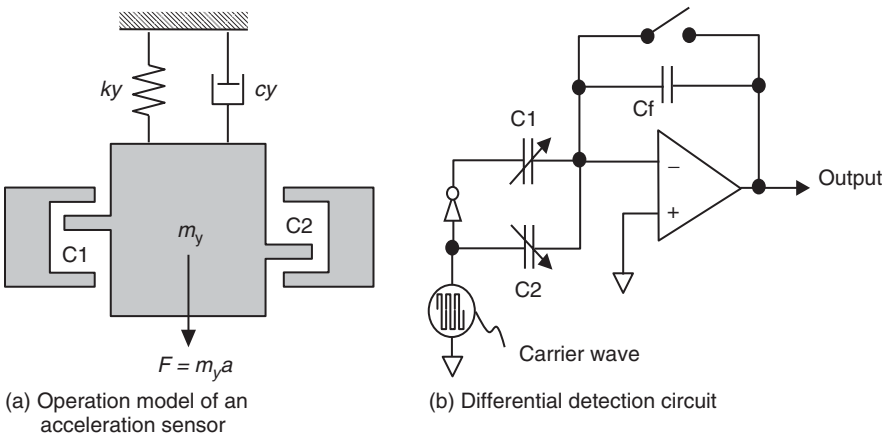


Figure 29. (a,b) Operation principle of the acceleration sensor. (Reproduced by permission of Hitachi Automotive Systems.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

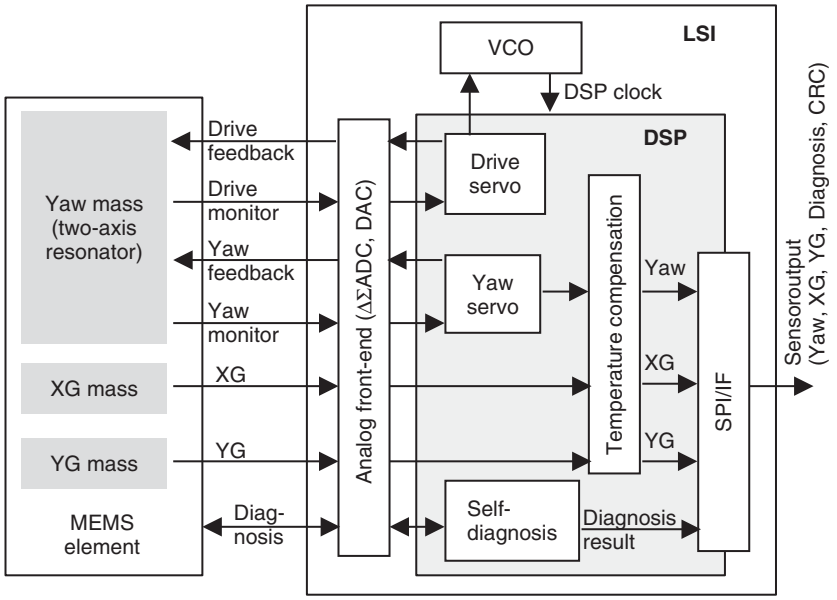


Figure 30. An example of the constitution of a processing circuit. (Reproduced by permission of Hitachi Automotive Systems.)

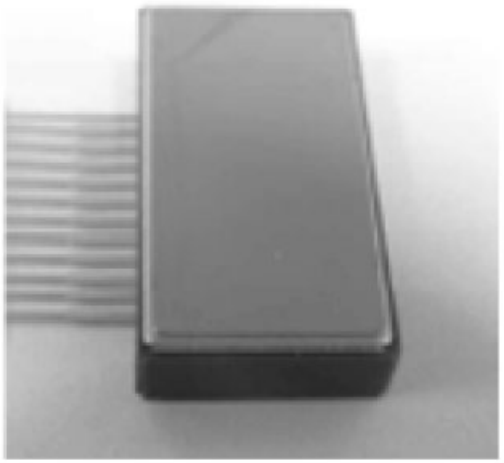


Figure 31. An example of structures (ceramic package). (Reproduced by permission of Hitachi Automotive Systems.)

environmental compatibility of automobiles. New sensing technologies that add new functions to automobiles, new sensor structures that realize sensor cost reductions, and new sensor usages that improve the performance of automobiles will be developed continuously in the future.

REFERENCES

Hayashi, M. (2009) *Trends of Vehicle MEMS Sensors in the Hitachi Group: Hitachi Review* **91** (10), 38–41.

Jeong, H.W. (2011) Three axes MEMS combined sensor for electronic stability control system. *Transactions E*, **131** (8), 279–285. The Institute of Electrical Engineers of Japan

Kato, M. (ed.) (2010) *Illustrated Car Electronics*, Nikkei Business Publications, Inc., Tokyo.

Mitsubishi Materials Corporation (2008) *DTN Thermistor Catalog*.

Sanken Electric Co., Ltd. (2007) *Allegro: ATS626 Catalog*.

Sasayama, T. (ed.) (1997) *Automotive Electronics*, Sankaido Publishing, Tokyo.

Sensirion A.G. (2011) SHT2x Data Sheet.

SHINYEI Technology Co., Ltd. (2010) http://www.shinyei.co.jp/stc/eng/.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto196
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Semiconductor Sensor (1)—Sensors for Power Trains

Keiji Hanzawa, Kenji Nakabayashi, Akio Yasukawa, Shinya Igarashi, and Hiroaki Hoshika

Hitachi Automotive Systems, Ltd, Hitachinaka, Japan

1 MEMS Airflow Sensor 1
2 Semiconductor Pressure Sensors 4
3 Conclusion 10
References 10

1 MEMS AIRFLOW SENSOR

1.1 Introduction of MEMS airflow sensor

From the point of view of global environmental conservation, the regulations on automotive emission reduction are getting stricter year by year. To comply with such strict regulations, it is essential to make the air fuel ratio control highly accurate. It is required to take highly accurate and reliable measurement of intake air rate in a wide measurement range (Sasayama, 1997; Ishikawa, 2006).

In recent years, the emission regulations have been enhanced, and improved fuel mileage is wanted. Under these circumstances, DI (direct injection) systems, in which the fuel is directly injected into cylinders, are being applied. So are diesel engines, variable displacement functions, and so forth. Accordingly, a larger number of engines are coming up with larger pulsations in the intake air, thus letting the intake air flow back into the intake port. Such pulsating flow can be extreme. The flow rate of the air coming into a cylinder needs to be measured accurately. Thermal wire type airflow sensors, however, cannot adequately detect the direction of airflow. It was, therefore, necessary to develop the MEMS (microelectromechanical system) airflow sensor to make possible accurate measurement of intake airflow rate with such reverse flows and extreme pulsating flows (Figure 1) (Hayashi, 2009; Matsumoto, 2010).

1.2 Principles of the MEMS airflow sensor

The MEMS airflow sensor has a diaphragm structure formed by etching the rear surface of a silicon substrate. On this diaphragm, a heater resistor and, at both sides of the heater resistor, two temperature sensor resistors are disposed. The power is supplied to the heater to heat and control it to a certain temperature. The amount of heat emission and distribution to the air, and to the sensors on both sides, depends on the airflow rate. This is the principle utilized. The heater resistor and the temperature sensor resistors are treated as extremely thin diaphragms of several micrometers. Their heat capacities are so small that high speed response, with the response time of several milliseconds, is made possible.

Figure 2 shows the principle of the airflow rate detection. When no air current is coming, the temperature distribution on the silicon diaphragm makes a line symmetry image with the heater as the axis. When an air current occurs, the temperature at the downstream temperature sensor increases as much as ΔT, depending on the direction and the flow rate of the air current.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto197
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

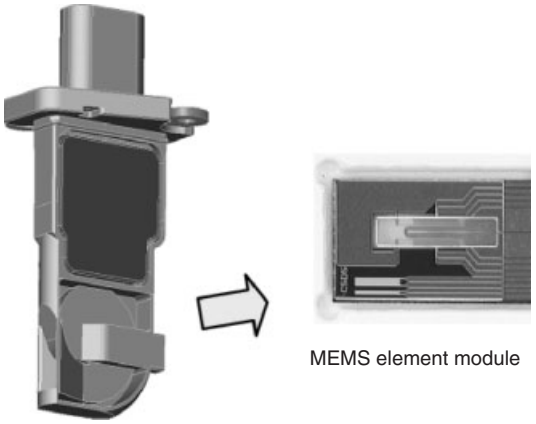


Figure 1. External view of a MEMS airflow sensor.

If a reverse flow occurs, the temperature at the upstream temperature sensor increases by ΔT. Temperature difference ΔT is determined based on the difference in the resistances between the two temperature sensors disposed on both sides of the heater, thus making it possible to acquire the flow rate and the direction. This temperature difference ΔT is processed with the constant temperature difference bridge circuit and the temperature difference detecting circuit and, thereby, is converted into a signal corresponding to the flow rate.

The heater and the temperature detection resistors are made with metal films such as platinum or polysilicon membrane. Figure 3 shows an example of sensing elements made with polysilicon membrane. The upper and lower sides are covered with the polysilicon membrane and the oxide films with excellent adhesive properties. Furthermore, to prevent moisture intrusion, the silicon nitride membrane is deposited in this structure.

1.3 Manufacturing processes of the MEMS airflow sensor

Figure 4 illustrates several manufacturing processes of a sensing element. To manufacture a sensing element, a monocrystal silicon substrate is covered with the silicon nitride membrane, the oxide film, and the polysilicon membrane in a prescribed order. The polysilicon membrane is treated in the patterning process to form the heater and the temperature sensors. On top of that, the oxide film, the silicon nitride membrane, and the polyimide membrane

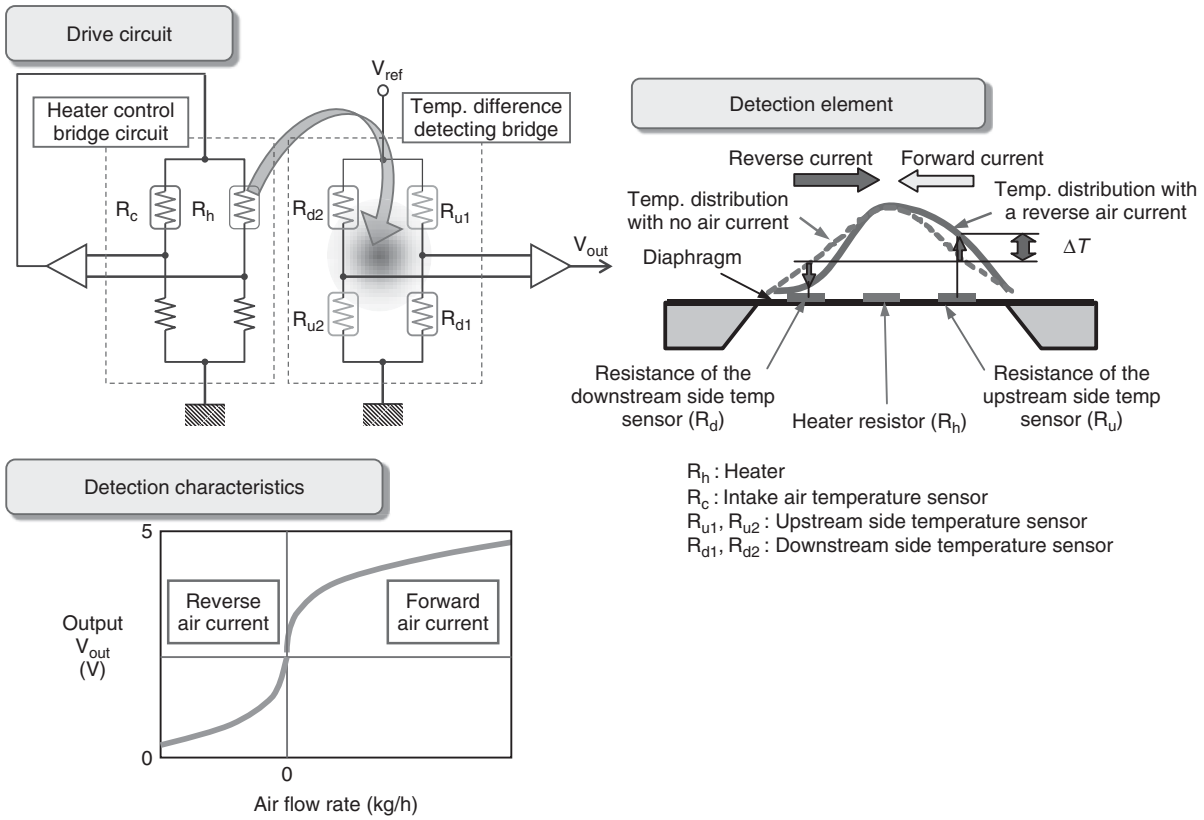


Figure 2. Principle of the detection by the MEMS airflow sensor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto197
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

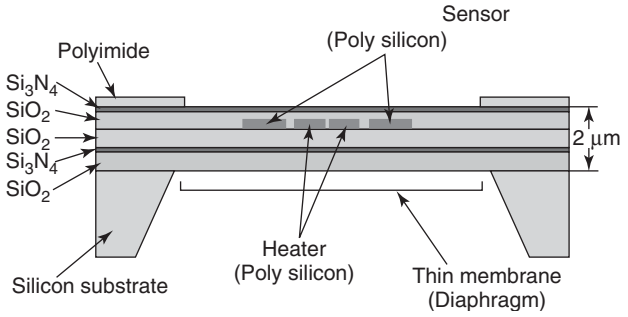


Figure 3. The cross-sectional view of a sensing element.

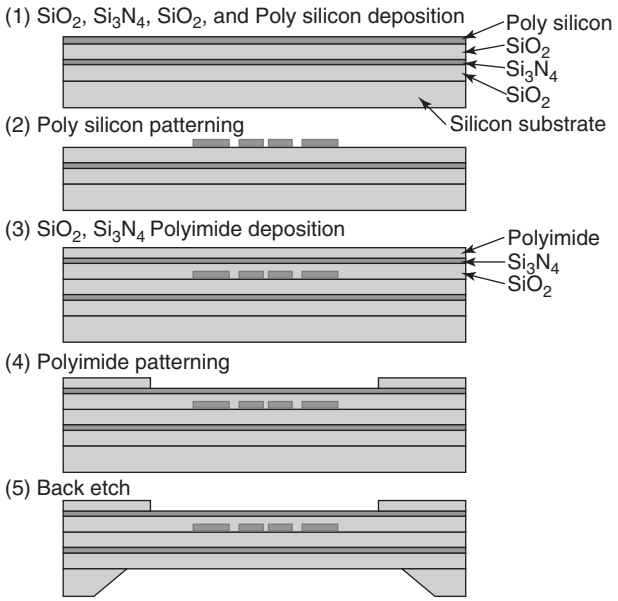


Figure 4. Manufacturing processes of a sensing element.

are formed. The polyimide membrane on the heater and the temperature sensors is removed. Then, the rear surface of the monocrystal silicon substrate is removed by alkali etching or the like. Thus, a diaphragm is formed with the thin films that provide excellent thermal insulation.

1.4 Possible problems and preventive actions for automotive application

1.4.1 Protection against floating foreign substances

The sensing elements of MEMS airflow sensors have a thin film structure. It is important to protect the thin films when a MEMS airflow sensor is applied to an automotive engine.

The intake air flowing into an automotive engine contains dust, rain drops if it is raining, oil evaporating from the engine, and the like, which are all accelerated with

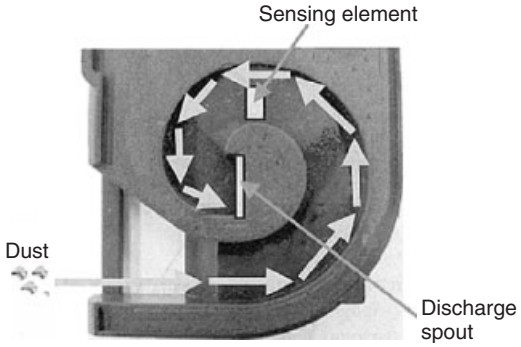


Figure 5. Structure of a cyclone type bypass passage.

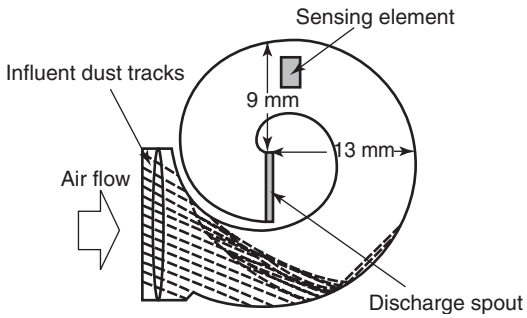


Figure 6. An example of particle tracking simulations of the dust.

the air current. The sensing element should be protected against these substances. To protect the sensing element, in particular, from the dust accelerated with the airflow, a bypass passage structure such as the one shown in Figure 5 is adopted.

This bypass passage is in the shape of a spiral. The dust accelerated with the airflow will be separated by the centrifugal force. Consequently, the dust flows along the wall of the bypass passage. As disposed in a position distant from the wall, the sensing element is free from collision with the dust.

Figure 6 shows an example of the result of particle tracking simulation of the dust. The dust is 10 μm in diameter and is accelerated to 50 m/s. In this simulation, the airflow of 50 m/s was generated around the bypass passage shown in Figure 5. Then, the speed distribution of the airflow inside and outside the bypass was acquired with fluid simulation. Finally, a plurality of tracks of the dust currents in the airflow was obtained.

The result of this tracking simulation shows that the dust, having flowed into the bypass passage, traveled to the wall of the bypass passage. The dust, when getting closer to the sensing element, traveled only along the wall. After this, the dust flowed out of the vent to the outside.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto197
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

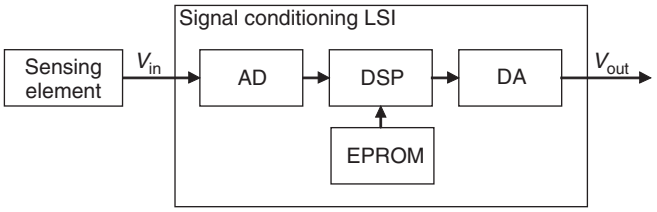


Figure 7. Signal processing LSI.

1.4.2 Catching up with high precision

The MEMS element makes fast responses and has a function to detect reverse flows. Furthermore, the element shows high affinity with semiconductor processes. Besides the conventional technologies for analog circuits, mixed signal technologies have been advancing in combination with digital technologies.

Figure 7 shows an example of implementation of a signal processing LSI. This LSI converts externally input signals provided from the sensing module into digital signals using the AD (analog-to-digital) converter. The individual adjustment information of each sensor is stored in the EPROM (erasable programmable read-only memory). On the basis of this information, the DSP (digital signal processor) makes digital calculation. Finally, the DA (digital-to-analog) converter outputs analog voltage out of this structure. In addition, the DSP has a built-in function that can compensate complex error curves provided from the sensing module.

Moreover, the MEMS type element, which has a fast response and is capable of detecting reverse currents, and the digital signal processing technology can be combined to make possible the compensation for the reverse current errors generated with pulsating flow.

Figure 8 shows the characteristics of the MEMS airflow sensor in the flow range of reverse currents. It is understood that the actual air pulsations are detected more accurately than those detected by the conventional thermal wire type under the conditions that generate reverse currents.

MEMS airflow sensors are manufactured with the semiconductor-manufacturing processes. This can be an advantage firstly because it is possible to make the sensors compact and secondly because these processes are highly suitable for mass production. At this moment, an accurate air fuel ratio controlling technology is under development, in which, besides airflow rate, EGR (exhaust gas recirculation) flow rate is measured and a plurality of information sources are utilized such as pressure sensors and open throttle angle sensors.

It is expected that the sensors will be made more compact and more intelligent with digital signal processing in the future. Besides these, it is also expected that they be going to be more accurate and have more functions.

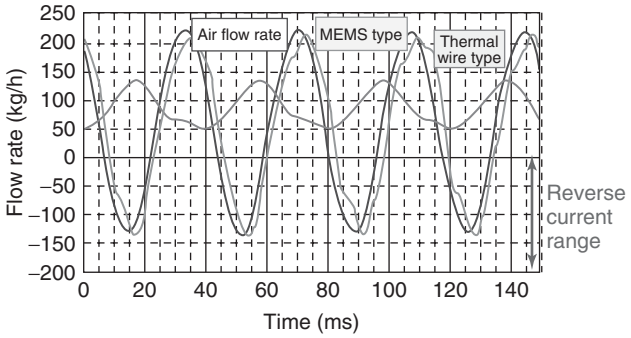


Figure 8. Flow rate characteristics of the thermal wire type sensor and the MEMS type sensor.

2 SEMICONDUCTOR PRESSURE SENSORS

2.1 Introduction of semiconductor pressure sensors

When electronic control began to be used in respect to automobiles, semiconductor pressure sensors were used as absolute pressure sensors at an intake manifold to control combustion in engines. Since then, a wide range of use has been seen in the middle pressure area such as hydraulic pressure applications including transmissions and suspensions, in the low pressure area that detects vaporized gasoline, and in the high pressure area in which fuel pressures are measured from gasoline DI engines and diesel common rail fuel injection systems (Sasayama, 1997; Kato, 2010).

This section describes the purposes of semiconductor pressure sensors as well as their technologies.

2.2 Purposes of semiconductor pressure sensors

Figure 9 shows some purposes of semiconductor pressure sensors.

The negative pressure of an intake manifold is measured to calculate the airflow, and thus control an appropriate mixture ratio with the gasoline. In these processes, the intake pressure is used. Furthermore, the intake pressure is also used for the measurement of boost pressure of turbo engines. Moreover, there are also such pressure sensors that can be integrated with temperature sensors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto197
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

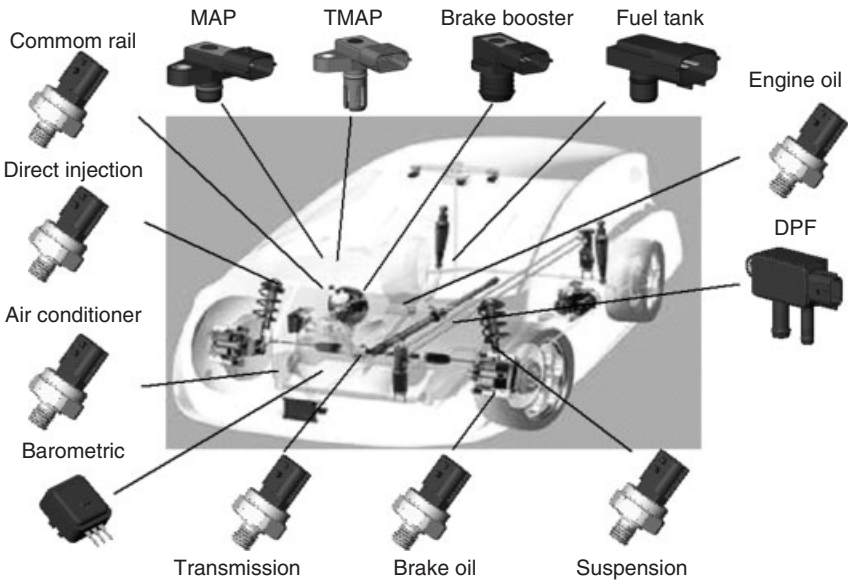


Figure 9. Purposes of semiconductor pressure sensors.

Atmospheric pressure sensors correct the air pressure during driving at high altitudes to control engine combustion to an optimal level.

The negative pressure in a brake booster is measured to acquire a brake booster pressure. In recent years, in particular, the brake booster pressure is used to control restarting an engine in an engine with stop/start capability.

The tank pressure was introduced with OBDII (on-board diagnostic II) to monitor the leakage of vaporized gasoline.

The oil pressure is used to take measurement of transmissions, engine oil, suspensions, and brake oil.

The fuel pressure is used to measure the pressure on the fuel for injecting gasoline directly into an engine as well as the pressure of the fuel for a common rail fuel injection system of a diesel engine.

Besides these, sensors are used for exhaust pressure, clog detection of diesel particulate filters on diesel engines, coolant pressure of air conditioners, and so forth.

Figure 10 shows the interrelations between some purposes and usable pressure ranges of semiconductor pressure sensors.

The pressure covers as wide a range as 5 kPa to 200 MPa.

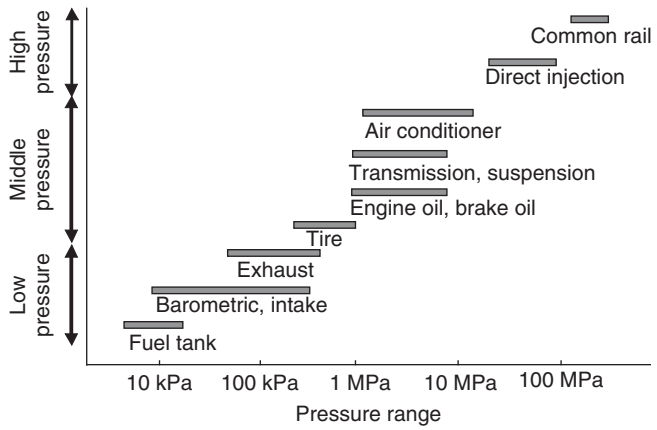


Figure 10. Purposes and usable pressure range.

2.3 Overview of pressure sensors

Semiconductor pressure sensors make use of the excellent characteristics of silicone for the elastic body. They, with micromachining technologies, form a thin film called a *diaphragm*, enabling the elements to convert pressure changes into electric signals.

Semiconductor pressure sensors are based on detection principles that can be generally classified into a piezo-resistor type and a capacitance type.

Figure 11 compares the detection principles of pressure sensors.

2.3.1 Detection principles

2.3.1.1 Piezo-resistor pressure sensors. Figure 12 shows the principle of piezo-resistor type pressure sensors. There is a diaphragm at the center of a silicon monocrystal. The diaphragm is formed thin in this structure, in a part of which a piezo-resistor gauge is spread and formed.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto197
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Principle	Resistance variance $V=(\Delta R/R)\cdot Vcc$			Capacitance variance $C=e\,A/d$
Diaphragm	Semiconductor silicon	Metal	Ceramic	Ceramic
Formation of pressure sensing part	Diffusion	Thin film	Thick film	Evaporation
Output	Analog	Analog	Analog	Frequency
Responsiveness	○	○	○	△
Temperature characteristics	△	○	○	△
Accuracy	○	○	△	○
Linearity	○	○	○	△
Pressure resistance	Low to medium pressure	High pressure	High pressure	High pressure

○ = Good (better), △ = Middle (even)

Figure 11. Comparison of principles of pressure sensors (Saito, 2009). (Reproduced with permission from CMC Publishing, Ltd.)

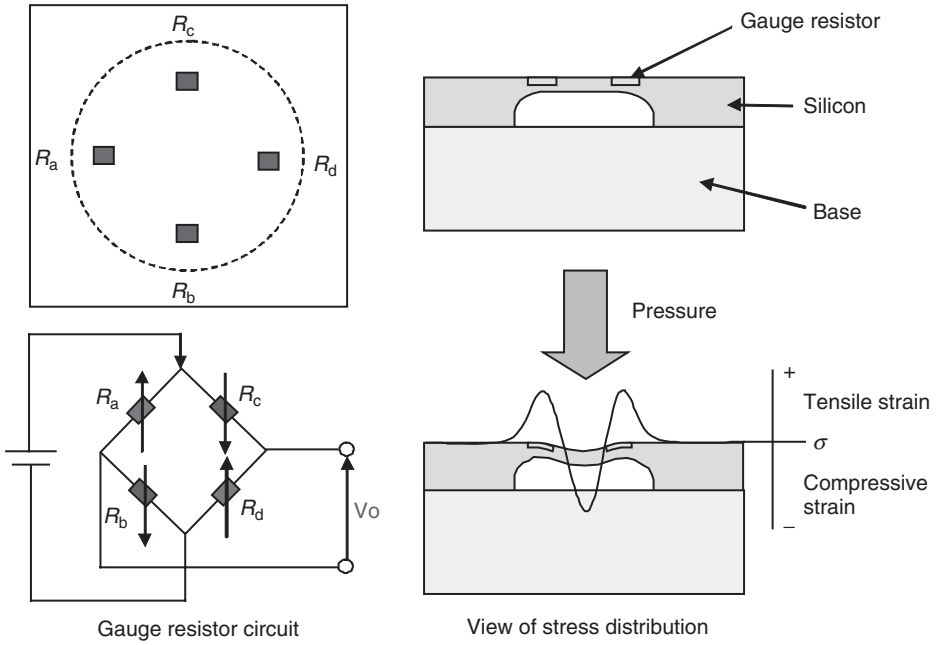


Figure 12. Detection principles of piezo-resistor types (Saito, 2009). (Reproduced with permission from CMC Publishing, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto197
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

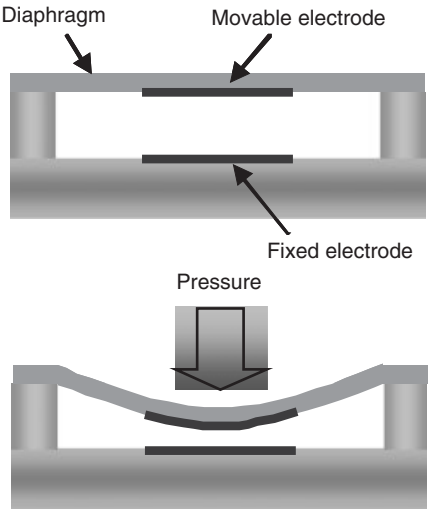


Figure 13. Detection principle of a capacitance type (Saito, 2009). (Reproduced with permission from CMC Publishing, Ltd.)

When some pressure is applied on the diaphragm, the strain generates stress (σ), causing resistance change (ΔR) to the gauge resistor.

$$\Delta R = \frac{1}{2} \bullet R \bullet \Delta\sigma \bullet \pi$$

ΔR resistance change to the gauge resistor
R initial gauge resistance
$\Delta\sigma$ stress generated in the gauge
π piezo-resistor coefficient.

A Wheatstone bridge output is expressed by the following equation:

$$\mathrm{Vo} = \frac{(Ra \bullet Rd) - (Rb \bullet Rc)}{Ra + Rb + Rc + Rd} \bullet I$$

Accordingly, the voltage signals are output in proportion to voltage.

2.3.1.2 Capacitance type pressure sensors. Figure 13 shows the detection principle of the pressure sensors of capacitance types. A silicon diaphragm is used as a movable electrode. On an insulated substrate, two opposing electrodes are formed in this structure.

When pressure is applied on the diaphragm, the shape and position of the diaphragm are transformed.

$$w = k \bullet P$$

w transformation of the diaphragm
k pressure/transformation coefficient of the diaphragm
P pressure.

The relationship between capacity variance ΔC and transformation w is expressed by the following equation:

$$\frac{\Delta C}{2C} = \frac{w}{d}$$

d electrode aperture

$$P = \frac{d}{k} \bullet \frac{\Delta C}{2C}$$

From the two equations above, the relationship between pressure P and capacitance variance ΔC is expressed by the following equation:

If this variance in the capacitance is converted into a change of voltage, voltage signals are output in accordance with pressure.

2.3.2 Processing circuit

Piezo-resistor type signal processing circuits consist of a temperature compensation circuit, an adjustment circuit, and an electric noise protection circuit.

In this section, we discuss temperature compensation circuits and adjustment circuits. A bridge is constituted with a P type Si piezo-resistor element, which is as impure as $10^{20}/\mathrm{cm}^3$. Here is the reason. The P type Si resistor of this density has a positive temperature coefficient. On the other hand, the piezo-resistor coefficient (π), which determines the sensitivity, has the negative temperature characteristics; that is, the coefficient decreases as the temperature rises. These characteristics are used to form a self-temperature compensation circuit. The principle figure of a self-temperature compensation circuit and adjustment circuit is shown in Figure 14.

The error characteristics of pressure sensors include the primary element errors such as offset voltage, temperature characteristics of the offset voltage, sensitivity, and temperature characteristics of the sensitivity as well as the secondary element errors such as nonlinearity of the temperature characteristics offset voltage, nonlinearity of the temperature characteristics of sensitivity, and nonlinearity of pressure characteristics. There are several methods to correct and reduce these errors in circuits; for example, (i) a method of using a laser to trim a thin film resistor and (ii) an electric trim method that uses an EPROM.

In the early days of commercialization, the gauge and the circuits, which formed an amplifier circuit and an adjustment circuit, were separately constituted. After that, the gauge was integrated with the circuits when those sensors were productionized later. Currently, the technology for integration of semiconductor circuits is in full use

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto197
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

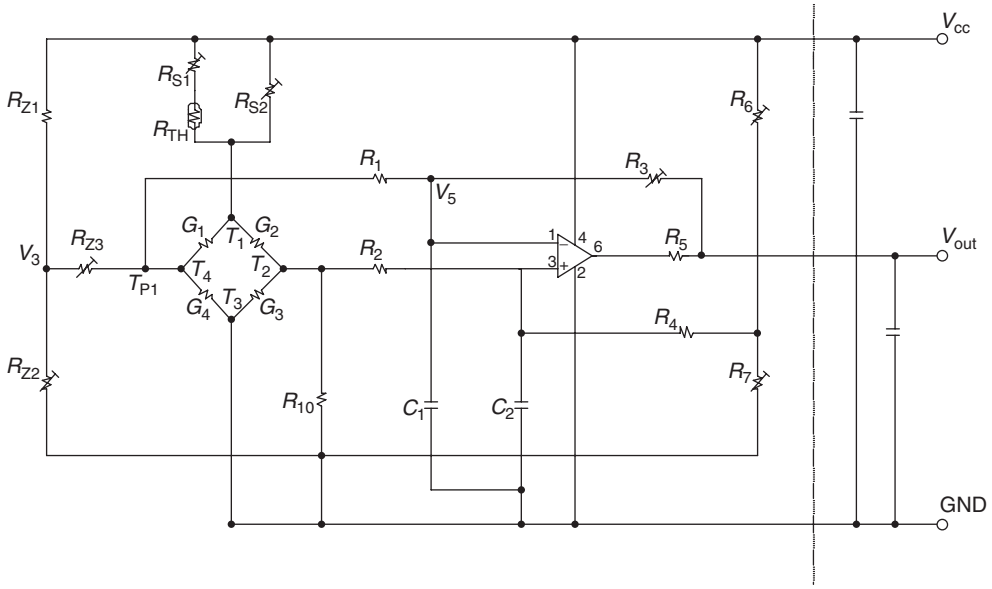


Figure 14. The principle of self-temperature compensation circuits.

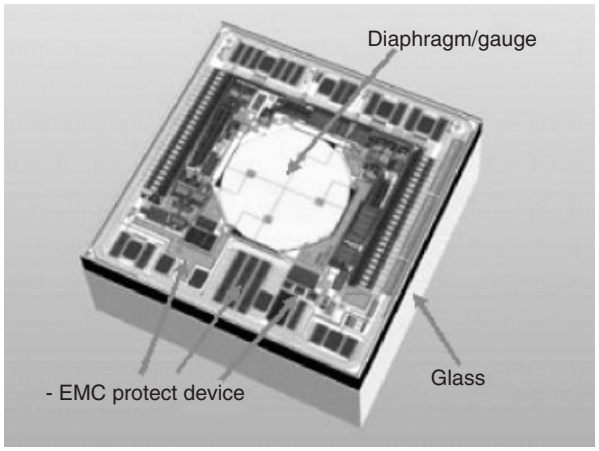


Figure 15. An example of integrated pressure sensors (Saito, 2009). (Reproduced with permission from CMC Publishing, Ltd.)

for small-sizing, cost cutting, and high reliability. Consequently, an EMC (electromagnetic compatibility) protection circuit is also integrated on one chip.

Figure 15 shows an example of integrated pressure sensors.

2.3.3 Production processes

Figure 16 shows an example of the production process flow of integrated pressure sensors (one-chip type).

2.3.3.1 Circuit formation process. The gauge resistor and the circuit elements are spread and formed. The electrodes are formed by evaporation. Furthermore, a passivation film is formed as a protection film.

2.3.3.2 Rear side polishing process. The rear side of the wafer is mirror-finished in the back grind process and the polish process.

2.3.3.3 Diaphragm formation process. Wet etching and/or plasma etching are used to process a concave cross section and to form a diaphragm.

2.3.3.4 Vacuum base chamber formation process. The electrostatic bonding is used to bond the silicon wafer and the glass wafer and thus to form a vacuum base chamber.

Figure 17 shows the principle of electrostatic bonding.

In the electrostatic bonding, a voltage is applied on the silicon and the glass. As the Na ion in the glass transfers, the electric charge moves. The glass in the vicinity of the silicon interface loses Na from its layer. Thus, the glass and the silicon bond to each other.

2.3.4 On-board installation (packaging) and types

In this section, we discuss typical intake pressure sensors and common rail pressure sensors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto197
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

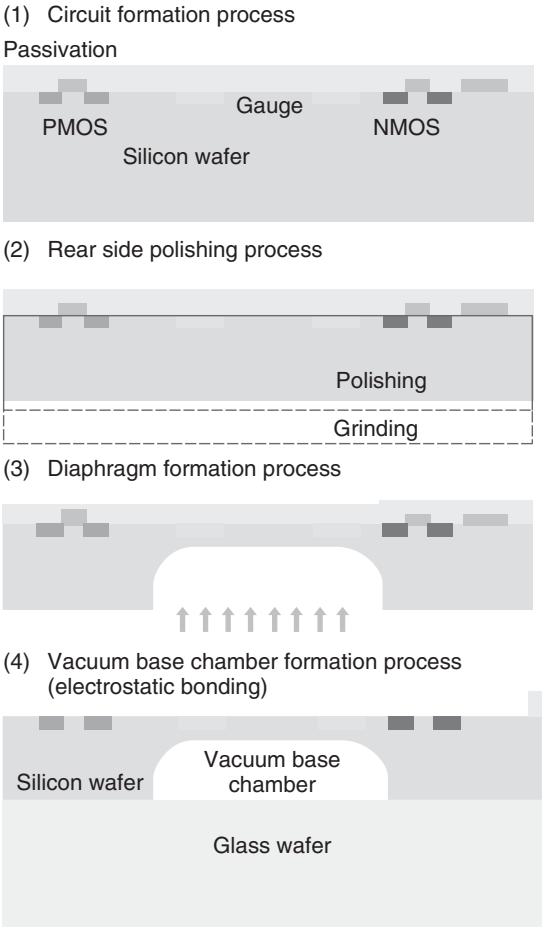


Figure 16. Production process flow (Saito, 2009). (Reproduced with permission from CMC Publishing, Ltd.)

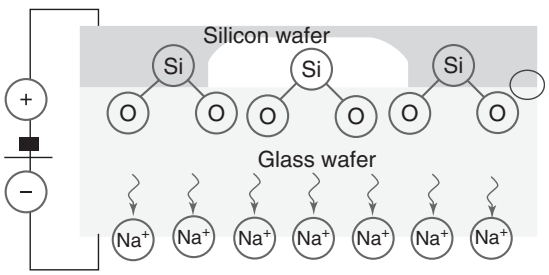


Figure 17. Principle of electrostatic bonding (Saito, 2009). (Reproduced with permission from CMC Publishing, Ltd.)

2.3.4.1 Intake pressure sensors. Figure 18 shows the structure and a picture of an intake pressure sensor of a surface pressure receiving type.

Intake pressure sensors form a sensor module for sensing pressure, a pressure guide module, and a connector. The sensor module is held by the housing and is fastened and air-sealed by the encapsulant. In addition, an O-ring ensures the air-tightness with the intake manifold in this structure.

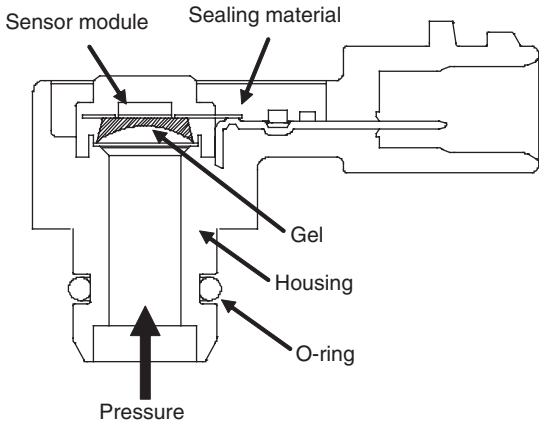


Figure 18. Structure of an intake pressure sensor.

As this is a surface pressure receiving type, the surface of the sensor chip may be directly in contact with foreign substances. This is why the sensor chip is covered with a gel material in this structure. The gel material protects the sensor chip against foreign substances included in a pressure medium. The gel material is made of such raw materials that have high elasticity (or low elastic modules) in order to minimize the stress to the diaphragm.

Furthermore, these materials have excellent chemical-resistant characteristics.

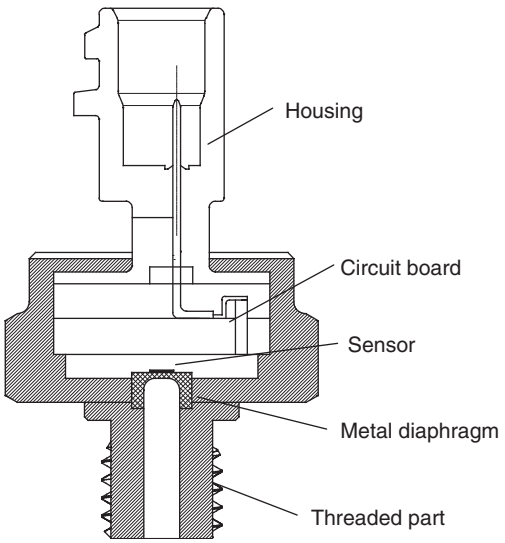


Figure 19. Structure of a common rail pressure sensor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto197
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.3.4.2 Common rail pressure sensor. Figure 19 shows the structure of a common rail pressure sensor. The pressure medium (fuel) for the super high pressure of 200 MPa receives the pressure directly with the thin wall metal diaphragm. The metal diaphragm and the threaded part are fixed by welding. To detect the pressure, the stress generated to the thin wall metal diaphragm is sensed by the Si piezo-resistor element connected on the diaphragm.

3 CONCLUSION

MEMS airflow sensors and semiconductor pressure sensors are already widely used in automotive applications. Moreover, the increase in demands for MEMS airflow sensors with high performances is expected because of the increase in needs for the accurate measurement of the intake airflow volume for increasing performance of automotive engines. On the other hand, the semiconductor pressure sensors will be used in a wider range of applications because of their high applicability to various systems for improving fuel mileage, emission reduction, and the safety in automobiles. Moreover, multifunctional sensors integrating a plurality of sensing functions, such as airflow, pressure, and humidity sensing functions, will be in greater demand in the future.

REFERENCES

Hayashi, H. (2009) Trends of vehicle MEMS sensors in the Hitachi Group *Hitachi Review*, **91** (10), 38–41. Hitachi, Ltd, Japan.

Ishikawa, H. (2006) Technology for measurement of high precision engine air flow rate *Transactions E*, **126** (8), 381–386. The Institute of Electrical Engineers of Japan, Japan.

Kato, M. (ed.) (2010) *Illustrated Car Electronics*, Nikkei Business Publications, Inc., Japan.

Matsumoto, M. (2010) High reliability, high precision automotive micro air flow sensor *Transactions E*, **130** (3), 80–85. The Institute of Electrical Engineers of Japan, Japan.

Saito, K. (2009) Automotive pressure sensors in *The Latest Trend in Sensors for Automobiles* (ed. M. Kimata), CMC Publishing, Japan. ISBN: 978-4-7813-0097-9

Sasayama, T. (ed.) (1997) Automotive Electronics, Sankaido Publishing.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto197
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Chassis ECU (Vehicle Dynamics, ABS)

Yasuhiro Abe, Toshihisa Kato, Shinya Takemoto, and Takahiro Okano

Advics Co., Ltd, Kariya, Japan

1 Introduction 1
2 Antilock Brake System 1
3 Electronic Stability Control 4
4 Regenerative-Friction Brake Coordination 9
5 Future Control Brakes 12
References 13
Further Reading 13

1 INTRODUCTION

Brake control systems experienced a rapid evolution in the 1970s with antilock brake system (ABS) as a technology to improve automotive active safety. ABS technology prevents wheels from locking and maintains vehicle steerability and stability under braking conditions. ABS was followed by traction control system (TCS), which prevents wheel spin at start-up or acceleration, and then electronic stability control (ESC) system, which prevents vehicle instability because of abrupt steering. Additional features, such as adaptive cruise control (ACC), use the ESC pressure control and have been widely adopted. In addition, brake pressure control plays an important role not only in active safety but also in regenerative brake coordination. This is an important technology for gas-mileage improvement of ecologically friendly cars such as hybrid and pure electric vehicles (EVs). This chapter describes the configuration of the sensors; electronic control unit (ECU); actuator; and control of ABS, ESC, and regenerative braking systems as a new technology.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 ANTILOCK BRAKE SYSTEM

ABS prevents wheels from locking to maintain optimum brake force during abrupt braking or on slippery road surfaces. It also aids vehicle stability and steerability by maintaining cornering force.

Equation 1 shows the relationship among the slip rate (S), coefficient of friction (μ) between the tire and road surface, and cornering force, which are ABS control standards. The maximum coefficient of friction is obtained when the slip rate is approximately 20%. When the slip rate is 100%, wheels are in the locked state and cornering force is close to zero. The less cornering force there is, the more stability and steerability become deteriorated. In order to maintain sufficient brake force, control, and stability, it is necessary to control the brake hydraulic pressure of each wheel so that the slip rate is approximately 20%.

$$\text{ABS slip rate } (S) = \frac{\text{Vehicle speed} - \text{Wheel speed}}{\text{Vehicle speed}} \times 100\% \quad (1)$$

2.1 System

The basic system configuration of ABS is shown in Figure 1. Generally, ABS is comprised of four-wheel speed sensors provided for each wheel, ECU, and brake actuators. The integrated configuration of the ECU and actuators is largely industry-wide. In the case of four-wheel drive vehicles, they may include a longitudinal acceleration sensor to make the vehicle speed more accurate. The incorporation of a longitudinal acceleration sensor into the ECU has recently

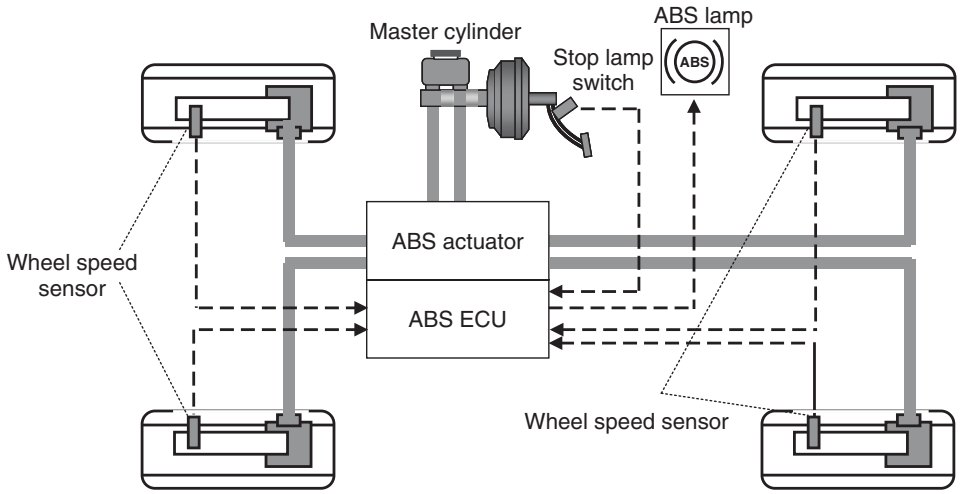


Figure 1. Example of ABS system configuration.

increased. The ECU judges the slip condition of each wheel from the input wheel speed sensor signal and uses the actuators to control the brake hydraulic pressure of each wheel.

2.2 Electronic control unit

A typical ECU configuration in the ABS is shown in Figure 2. It is mainly comprised of a CPU (central processing unit), which performs operations; an ABS interface IC, which performs inputs and outputs; a motor relay, which drives the motor; a solenoid relay, which supplies or cuts off power to the solenoid; a CAN (controller area network) driver, which drives the lamp; and an EEPROM (electrically erasable programmable read-only memory), nonvolatile memory. The mainstream CPU is usually a 16- or 32-bit microcomputer. The ABS interface IC controls the basic system inputs and outputs. It monitors wheel speed sensor signals and controls the solenoid outputs to the actuator. The motor relay is a semiconductor relay to drive the motor, and the solenoid relay is a fail-safe relay to cut off the power to the solenoid. The driver is notified of a system error by the illumination of a warning lamp. The lamp is

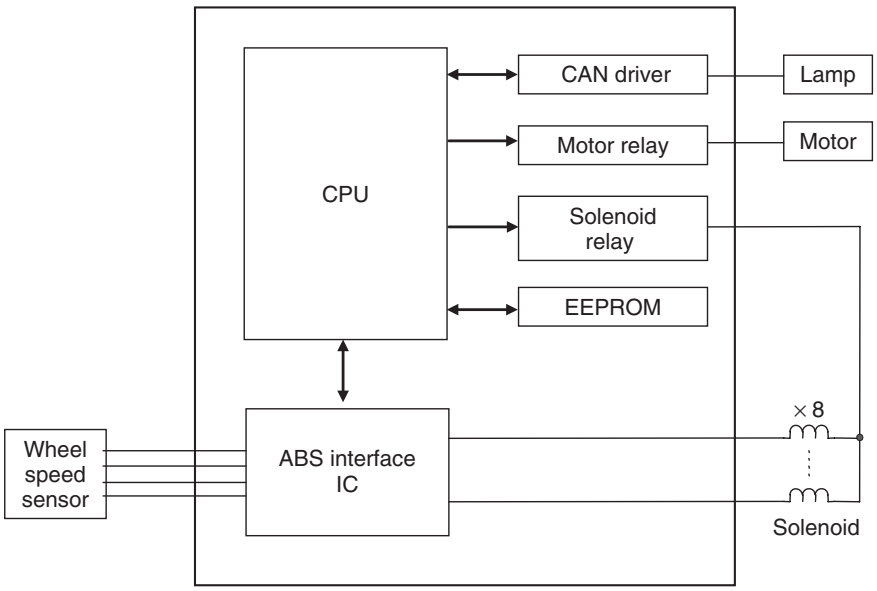


Figure 2. Example of ABS–ECU circuit configuration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

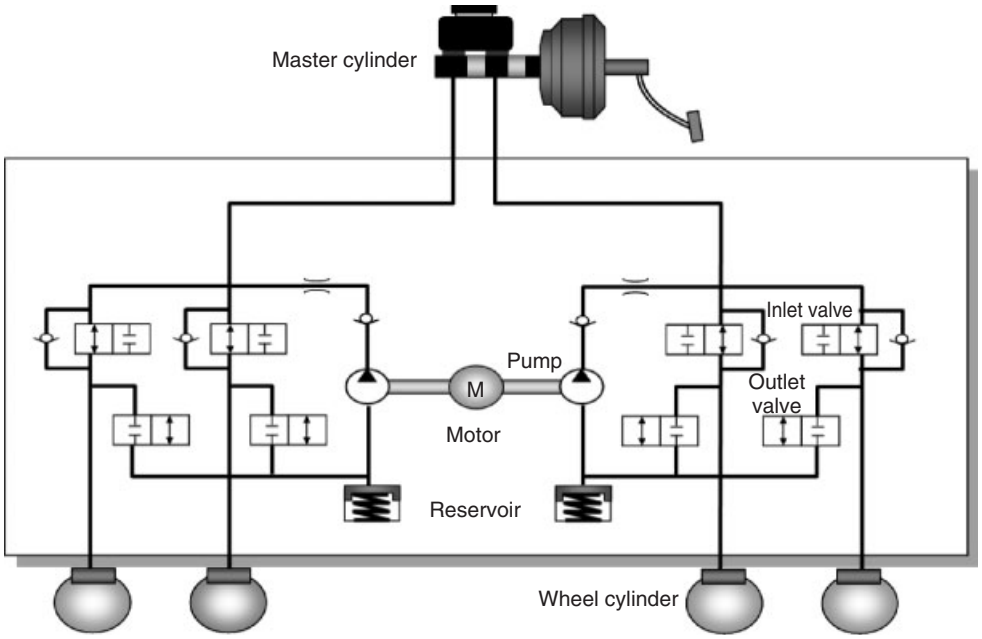


Figure 3. Example of ABS hydraulic circuit—ADVICS "ADS-A2."

controlled electronically by the instrument cluster or meter ECU. The ABS–ECU and meter ECU use CAN communication to report the system status. Failure information is stored in the EEPROM to avoid its loss even if the ignition switch is turned off.

2.3 Sensors

The wheel speed sensor is a sensor that detects the wheel rotational speed. It detects the speed of the rotor rotating in synchronization with the wheel by the electromagnetic pickup system and converts it to a voltage as an input to the ECU. Use of a hall element in the semiconductor system is now industry-wide.

The acceleration sensor is a sensor that detects the vehicle longitudinal acceleration. In the early years of ABS, a pendulum optical switch, mercury switch, and other types of sensors were used. A semiconductor system that can linearly detect vehicle acceleration is now the industry standard.

2.4 Actuator

The ABS actuator is an equipment that adjusts the brake hydraulic pressure of wheel cylinders according to the control command from ABS–ECU. The actuator has various configurations, and one example of the hydraulic circuit is shown in Figure 3. Each wheel has two solenoid valves (inlet valve and outlet valve). Three hydraulic patterns (apply, hold, and release) are created through on/off driving of those solenoid valves to control the brake hydraulic pressure of the wheel cylinder on each wheel. In the case of release, brake fluid is stored in the reservoir and pumped up from the reservoir to the master cylinder by running the motor. Recently, the common system has reduced noise, vibration, and harshness during ABS operation by performing motor speed control or solenoid valve current control.

2.5 Basic control of ABS

ABS control is performed by using individual wheel speed and wheel acceleration. This is obtained using the signal from the wheel speed sensors. These wheel speed sensors are also used to estimate the vehicle speed. An overview of ABS control is shown in Figure 4.

When the brake pedal is pressed, the hydraulic pressure in the wheel cylinder is increased, thus decreasing the wheel speed and increasing the slip rate. The wheels tend to lock abruptly when the slip rate exceeds the maximum coefficient of friction. The slip rate causing the wheel lock is continuously monitored using the wheel speed or wheel acceleration, and then the hydraulic pressure is released accordingly. The slip rate of the wheel speed or wheel acceleration is further monitored in order to avoid excessive release and the process is optimized by switching between hold and release judgment.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

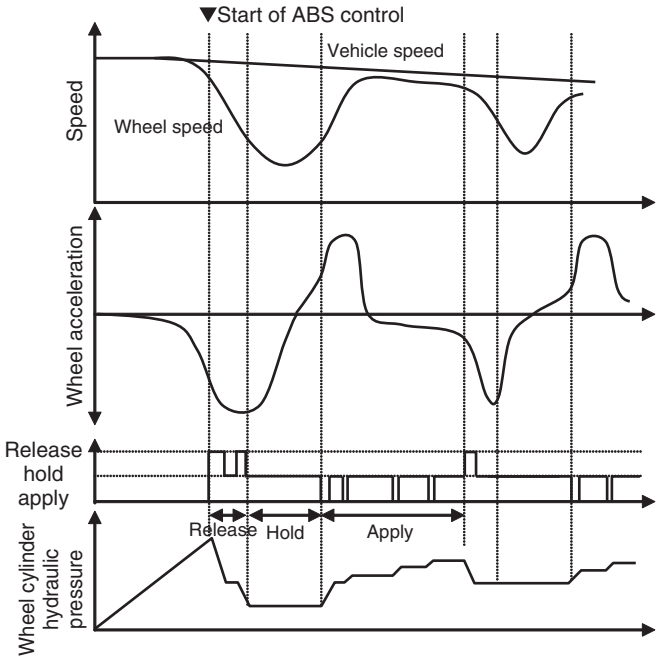


Figure 4. Overview of ABS control.

As the wheel speed increases and the slip rate decreases, a command is given to immediately apply brake hydraulic pressure to the wheel cylinders. When a sufficient hydraulic brake pressure is restored, the pressure pattern is switched to a relatively slow increase. Thus, high brake hydraulic pressure is maintained on the wheel. The above process is repeated to provide sufficient brake force and maintain stability and steerability of vehicle during braking.

The tendency of a wheel to lock differs on rough road surfaces such as gravel, dirt roads, potholes, manholes, and snowpacked roads. To avoid ABS malfunction on these road types, appropriate measures are taken for the calculation of slip rates of wheel speeds and wheel accelerations used in the judgment of wheel lock.

2.6 ABS application control

Electronic brake force distribution (EBD) is an ABS application function that controls the hydraulic pressure on rear wheels using the ABS actuator. Brake force distribution varies in cases of different loading conditions and because of a longitudinal load shift while braking. This causes an increased wheel slip on the rear wheels compared to the front wheels and results in vehicle instability. Brake hydraulic pressure on rear wheels used to be limited to the ideal brake force distribution by using a proportioning valve. These were replaced by EBD with the installation of ABS units.

EBD starts on the rear wheels with smaller amount of wheel slip than ABS. Stability of the vehicle is maintained by temporarily holding the brake pressure of the rear wheels. The hydraulic pressure on the rear wheels is thus controlled to limit the rear-wheel slip. Sufficient brake pressure is reapplied while continuously monitoring the rear-wheel slip. Intermediate pressure increase is performed during holds, considering the pedal feeling. The hydraulic pressure on the rear wheel can be reduced if the amount of slip increases excessively (Figure 5).

3 ELECTRONIC STABILITY CONTROL

Vehicles corner according to the driver's steering operation under normal conditions. However, a rear-wheel skid (oversteer) or front-wheel skid (understeer) may occur because of a sudden change in the road surface condition or sudden operations such as emergency avoidance (Figure 6).

The ESC detects vehicle behavior by various sensors such as the yaw rate sensor, steering angle sensor, longitudinal acceleration sensor, and lateral acceleration sensor. Skids are controlled to maintain vehicle stability by controlling the hydraulic pressure of each wheel and engine torque.

Analysis of data, both domestically and abroad, has shown a significant reduction in vehicle accidents among vehicles that have ESC systems installed. The benefits of ESC have become clear and it is a safety system that should be promoted for a safe automotive society. The installation of ESC has become mandatory in various parts of the world and its use has progressed rapidly.

Each vehicle manufacturer refers to ESC in its own way [ESP (electronic stability program), VDC (vehicle dynamic control), VSC (vehicle stability control), VSA (vehicle stability assist), etc.], but this chapter refers to it only as ESC.

3.1 System

An example of the ESC system configuration is shown in Figure 7.

The ESC system includes many sensors such as those which detect vehicle behavior (yaw rate sensor, longitudinal and lateral acceleration sensors) and those which detect a driver's operation (steering angle sensor, brake pressure sensor), in addition to each wheel speed sensor equipped with the ABS system. The ECU controls the brake actuator, which can automatically distribute brake pressure for four-wheel independent control. The ESC–ECU communicates with engine and meter ECUs on a vehicle

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

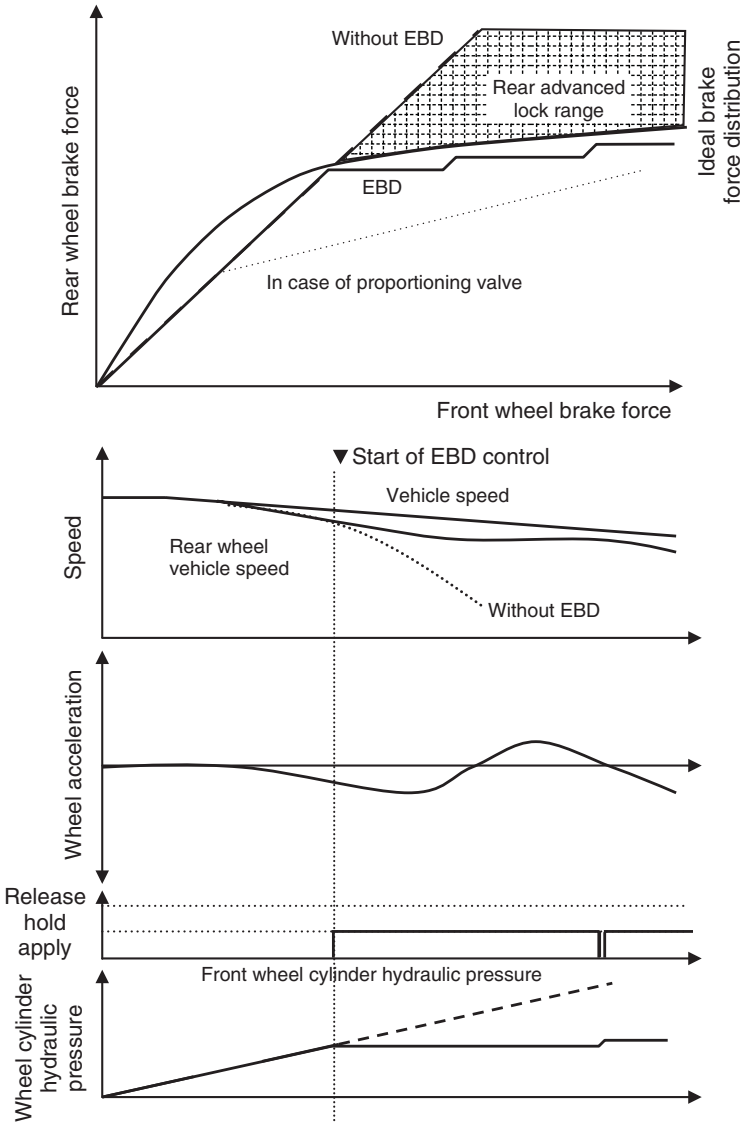


Figure 5. Overview of EBD control.

network, such as CAN, to control torque and lamp status.

3.2 Electronic control unit

The typical ECU configuration in the ESC system is shown in Figure 8. The main differences with ABS–ECU are as follows. The mainstream CPU is a 32-bit microcomputer. The linear solenoid driver for automatic brake pressure distribution for ESC is added. It is connected to other systems such as the engine and sensors over the CAN communication. In some cases, the yaw rate sensor and G sensor are incorporated into the ECU. Failure information and the sensor zero-point learning values are stored in EEPROM nonvolatile memory.

3.3 Sensors

As stated earlier, the ESC system includes many sensors such as those which detect vehicle behavior (yaw rate sensor, longitudinal and lateral acceleration sensors) and those which detect a driver's operation (steering angle sensor, brake pressure sensor), in addition to the wheel speed sensor used in ABS.

The yaw rate sensor detects the vehicle rotational speed by a gyro with the rotary inertia used, or a tuning fork

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

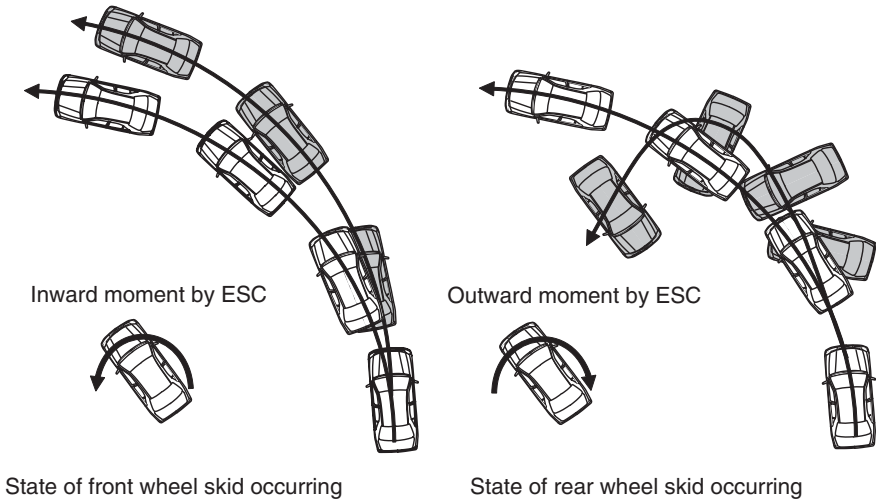


Figure 6. Effects of ESC.

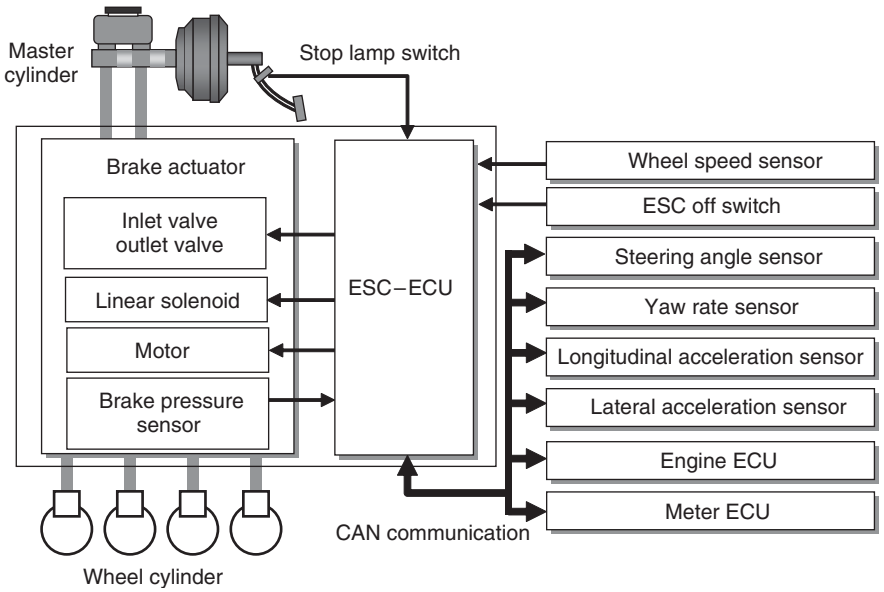


Figure 7. Example of ESC system configuration.

using a coriolis force. It converts that speed to a voltage and outputs it to the ECU.

There are two acceleration sensors mounted in total (longitudinal, lateral) to detect the acceleration in vehicle longitudinal and lateral directions. The acceleration sensors detect the acceleration by a semiconductor element, convert it to a voltage, and output it to the ECU. As a detection system, the piezo element system and capacitance system are mainly used.

The steering angle sensor is a detection element mounted on the steering column. It detects the steering rotational pulse and converts the steering angle calculation result to a voltage to output it to the ECU. It is used to detect the driver's steering speed and steering direction. As the detection system, there is an optical system as well as systems with an magnetic resistance (MR) element or hall element used.

The brake pressure sensor is a piezo element mounted on the master cylinder pressure input of the unit. It converts the master cylinder pressure to a voltage and outputs it to the ECU. It is used to detect the driver's demanded braking condition. The piezo elements for detection include the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

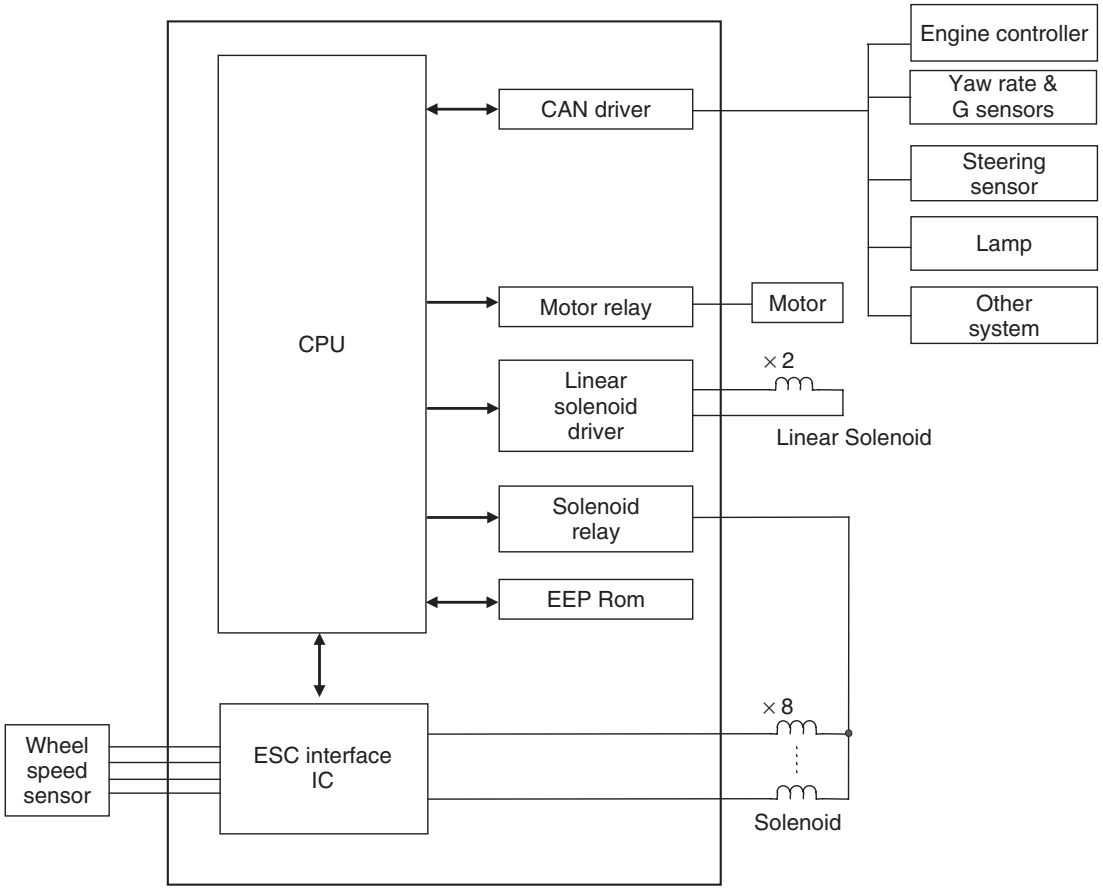


Figure 8. Example of ESC–ECU circuit configuration.

alloy wire resistance strain gauge system, semiconductor strain gauge system, and capacitance system.

3.4 Actuator

An ESC brake actuator must have the ability to apply pressure very quickly to control vehicle behavior. An ESC actuator requires additional hydraulic channels, solenoid valves, and a brake pressure sensor, as well as improvements in the capacity of the pump and motor to achieve the high performance that is required.

An example of the hydraulic circuit of the ESC brake actuator is shown in Figure 9. A linear solenoid is added upstream to the inlet valve for pressurization to make the pump-generated pressure linearly adjustable according to the current command from the ECU. In addition, the reservoir is equipped with the function of forming an inflow path from the master cylinder to the pump when the driver does not press the brake pedal. The ESC–ECU is generally integrated into the ESC actuator and installed in the engine compartment similar to the ABS–ECU and actuator.

3.5 Basic control of ESC

ESC detects the skids of front (understeer) or rear wheels (oversteer) to control the brake hydraulic pressure of each wheel and engine torque to stabilize the vehicle. The ESC–ECU compares the driver's intended direction (control target) detected from the steering angle sensor, accelerator position, and brake pressure with the vehicle condition detected from the yaw rate sensor, acceleration sensor, and wheel speed sensor (Figure 10).

When there is a tendency for a front-wheel skid (understeer) to occur, brake pressurization is commanded mainly to the turning inside rear wheel to give an inward moment to the vehicle. When there is a tendency for a rear-wheel skid (oversteer) to occur, brake pressurization is commanded to the turning outside front wheel to give an outward moment to the vehicle. The vehicle is controlled in a stable

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

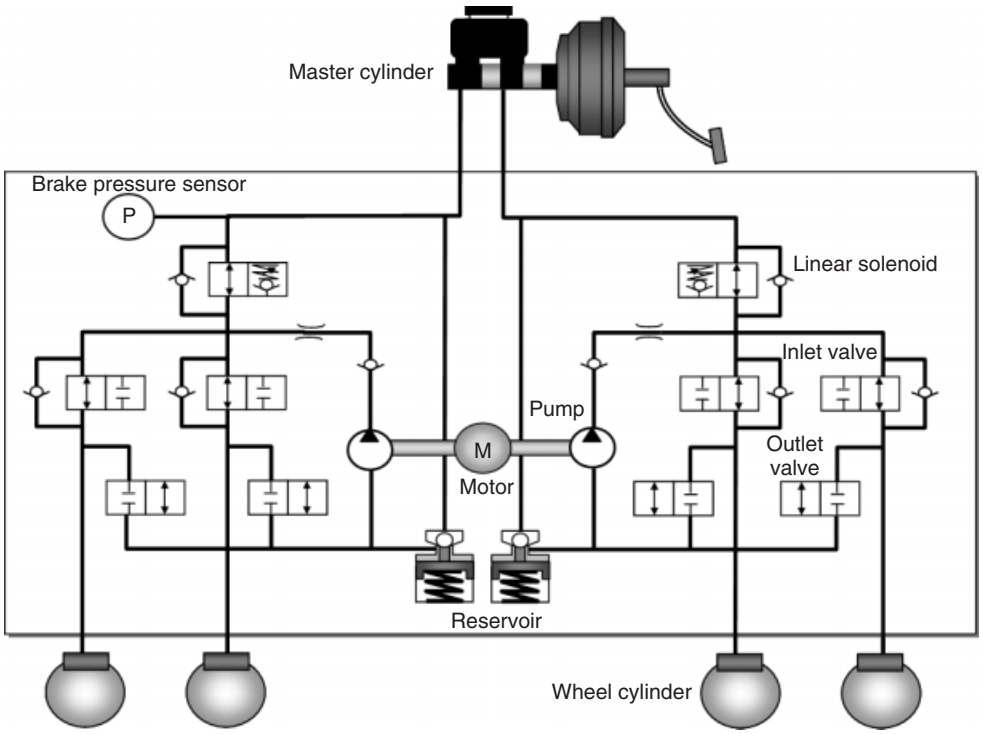


Figure 9. Example of ESC brake actuator—ADVICS "ADS-V2."

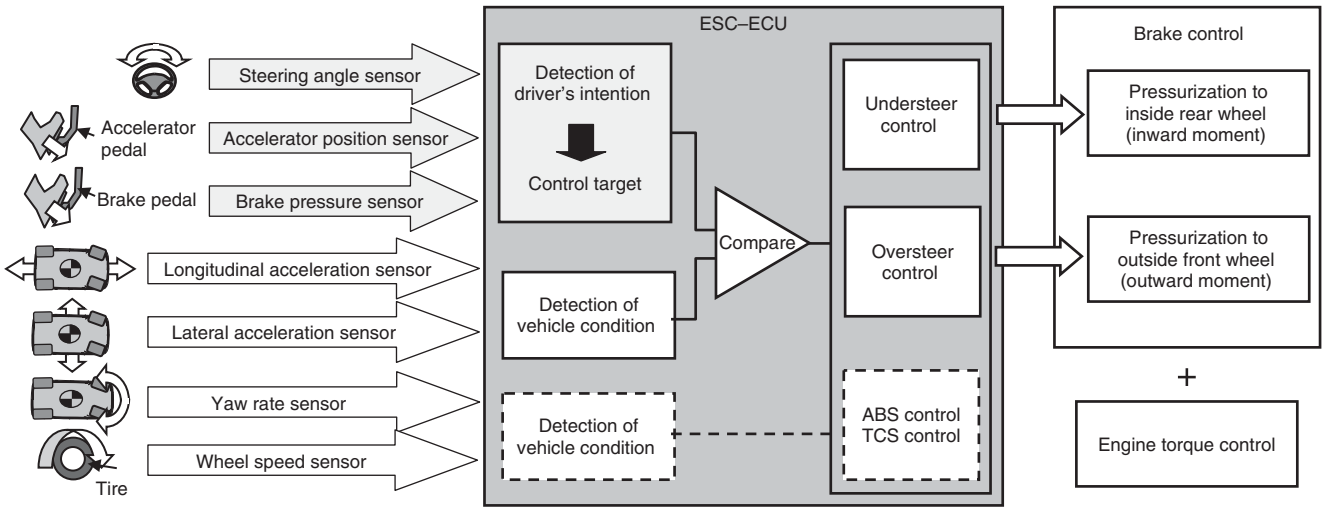


Figure 10. Schematic diagram of ESC control.

direction by combining these brake pressurization controls and engine torque control.

3.6 ESC application control

Additional functions such as TCS, brake assist system (BAS), and ACC take advantage of the ESC systems automatic pressurization function and engine torque control function. The TCS and BAS are described here.

3.6.1 Application control (1): TCS

TCS is a brake control system that maintains vehicle stability and acceleration by preventing driven wheel

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

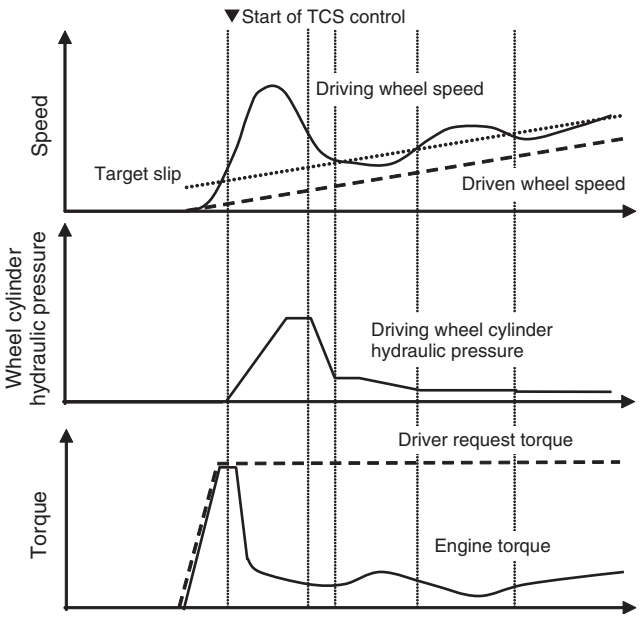


Figure 11. Overview of TCS control.

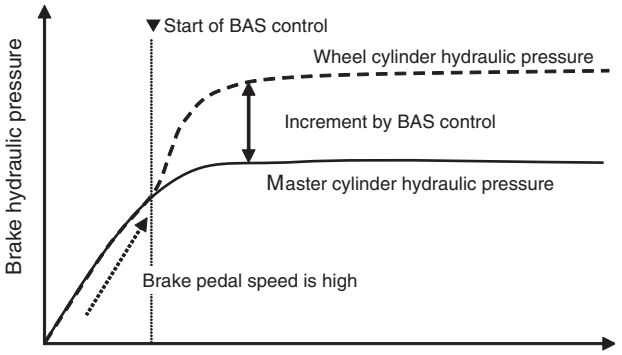


Figure 12. Overview of BAS control.

spin-up generated by slippery road surfaces or excessive driving force. As is the case with the $\mu-S$ characteristic in ABS, there is a relationship among the wheel slip rate, coefficient of road-surface friction, and tractive force. TCS calculates the wheel slip rate from the wheel speed of driven wheels and vehicle speed. The drive torque is controlled so that a proper slip rate, which can ensure a sufficient driving force and cornering force, can be maintained.

TCS performs control to limit the torque of driven wheels based on wheel slip. It judges slip by comparing the driven wheel speeds to the estimated vehicle speed. Estimated vehicle speed is calculated by using driven and non-driven wheel speeds as well as wheel acceleration. Control is performed by combining requests for engine torque reduction and controlling brake pressure of driven wheels (Figure 11).

When the driver presses the accelerator pedal, the force of the driven wheels increases with the increase in the engine output torque. When the driving force is higher than the friction force generated between the tires and the road surface, the driven wheel speed suddenly increases and a wheel slip occurs. When the wheel speed of driven wheels exceeds the control start threshold level, TCS control starts.

During TCS, brake pressure control is continued so that the wheel slip rate will come closer to the target slip threshold level. When the wheel speed of driven wheels decreases and falls below the TCS control completion threshold level, brake pressure control and engine control are gradually completed. The TCS system is designed to maintain vehicle stability and acceleration by controlling the slip rate of driven wheels. If the road has a different coefficient of friction between the right and left driving wheels, the wheel cylinder hydraulic pressure of driving wheels is controlled independently between the right and left.

3.6.2 Application control (2): BAS

In case of emergency, the pressure applied by many drivers on the brake pedal may not be sufficient for braking. In this situation, the pressurization function of ESC is used to increase the brake force in addition to the pressure from the driver. Over the last few years, the installation of BAS is mandatory in a number of regions to improve pedestrian safety.

Brake pedal speed and brake stroke are used by BAS to determine an emergency situation. The brake pedal speed and stroke are obtained from the pressure change gradient. Brake pressure is obtained from the brake pressure sensor equipped with the ECS brake actuator. BAS increases the hydraulic pressure of each wheel using the pressurization function of ESC when an emergency situation is determined (Figure 12).

Pressurization can be performed by driving the motor and linear solenoid within the ESC brake actuator (Figure 9).

4 REGENERATIVE-FRICTION BRAKE COORDINATION

Brake systems in hybrid vehicles and pure EVs are expected to have a function that contributes to a reduction of exhaust gas emission in addition to safety and comfort as in the case of conventional vehicle brake system (hereinafter called the *conventional braking system*). The coordination

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

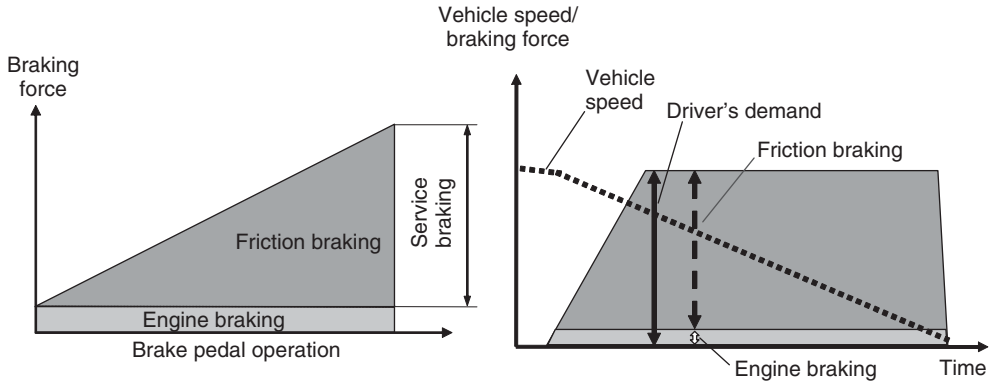


Figure 13. Sharing of braking force in conventional brake system.

of regenerative and friction brake should feel as natural as a conventional system while recovering the maximum kinetic energy during deceleration. This energy is known as the *regenerative energy* and is an effective technique for gas-mileage improvement.

4.1 Regenerative brake method

In a conventional braking system, vehicle deceleration is obtained by engine braking and service brakes (Figure 13). In this process, the kinetic energy of the vehicle is converted to heat energy in the engine and the wheel cylinders. This heat energy is thus released in the atmosphere.

Regenerative brake systems are broadly divided into two types; the first is with the simple regenerative-friction brake coordination method and the other with the more advanced regenerative-friction brake coordination method. In the first method with simple regenerative-friction brake coordination, the regenerative brake force is produced by the engine brake to recover some of the kinetic energy of the vehicle. As there is no coordination with the fiction brake system, the amount of kinetic energy recovered is limited to engine braking only. In this method, once the regenerative braking is no longer available from the kinetic energy corresponding to engine brake, vehicle deceleration is acquired by actual engine braking. Thus, the regenerative braking in this method does not require compensation from the service brake. Therefore, braking with the service brake delivers a natural brake feeling similar to conventional braking system, while still recovering some kinetic energy (Figure 14).

The second method recovers even more kinetic energy by coordinating regenerative-friction braking (Figure 15). This method recovers kinetic energy not only from the engine braking but also from the service brake by using regenerative brake. As the kinetic energy recovery amount is more than the simple regenerative-friction brake coordination method, lower fuel consumption is expected from this method. To allow maximum recovery of kinetic energy, maximum regenerative brake torque is developed within the range of brake force required by the driver. The remaining

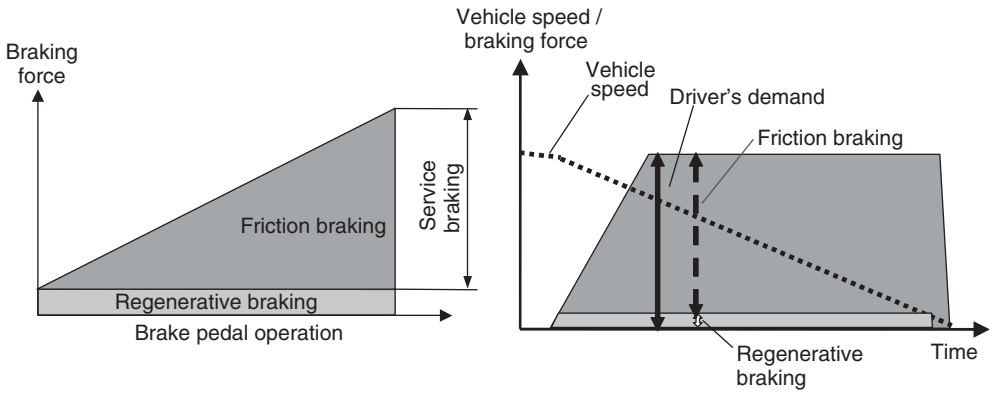


Figure 14. Simple regenerative-friction brake coordination method.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

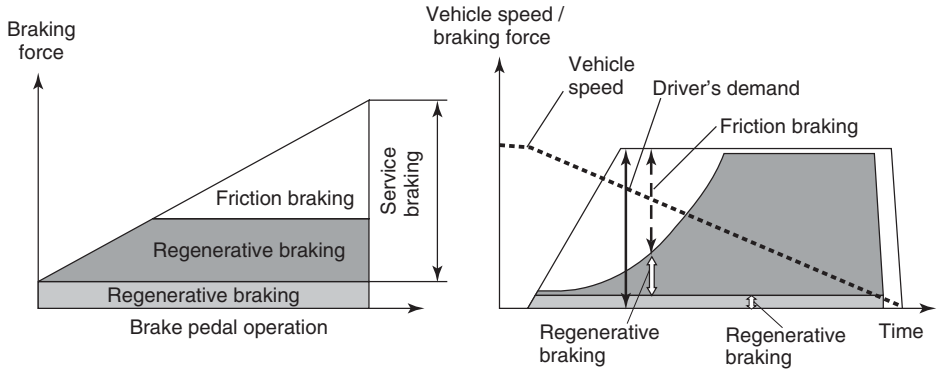


Figure 15. Advanced regenerative-friction brake coordination method.

brake force required by the driver is compensated by the friction brake torque. Advance regenerative-friction brake coordination method requires the friction brake torque to be controlled independently from the brake pedal operation amount by the driver.

4.2 Technology of friction brake torque control required for regenerative-friction brake coordination

With regard to friction brake torque adjustment methods, various techniques have been proposed. The following are typical methods:

4.2.1 Control of wheel cylinder pressure by active vacuum booster

In this method, an active vacuum booster is used to adjust the required friction brake torque during regenerative-friction coordination. This helps adjust the hydraulic pressure in the master cylinder, which helps achieve desired hydraulic pressure into the wheel cylinder (von Albrichsfeld and Karner, 2009) (Figure 16).

4.2.2 Control of wheel cylinder pressure by electric motor

In this method, the required friction brake torque is adjusted by driving an electric motor. The electric motor adjusts the hydraulic pressure in the master cylinder to introduce the desired hydraulic pressure into the wheel cylinder (Hano and Hakiai, 2011; Obata *et al*., 2011) (Figure 17).

4.2.3 Control of wheel cylinder pressure by linear solenoid valve

This method uses the linear solenoid valves to adjust the required friction brake torque during regenerative-friction coordination. This uses the high pressure brake fluid stored in the accumulator to directly introduce the desired pressure into the wheel cylinder (Nakata *et al*., 2009) (Figure 18).

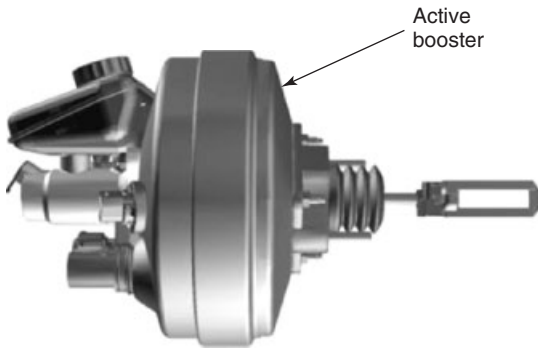


Figure 16. Adjustment of friction brake torque by control of servo-assisted vacuum. (Reproduced from von Albrichsfeld and Karner, 2009. © SAE International.)

4.2.4 Direct control of friction brake torque by electric motor

This method uses an electric motor directly mounted on the wheel cylinder to adjust the friction brake torque and achieve the required brake force.

4.3 Isolation technology of brake pedal stroke and brake pedal force from the friction brake torque

A natural brake feeling is required during the coordination of regenerative brake torque and conventional friction brake torque. To achieve natural brake feeling, the regenerative brake coordination with friction brake torque should not have any effect on the brake pedal stroke and brake pedal

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

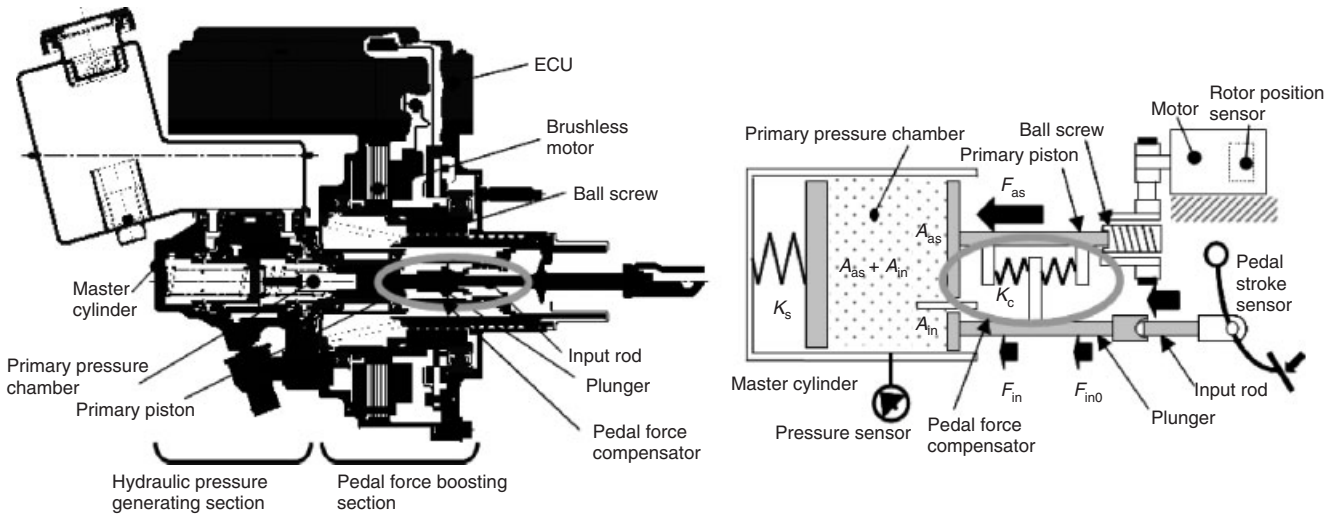


Figure 17. Adjustment of friction brake torque by electric motor control. (Reproduced from Hano and Hakiai, 2011 © SAE International.)

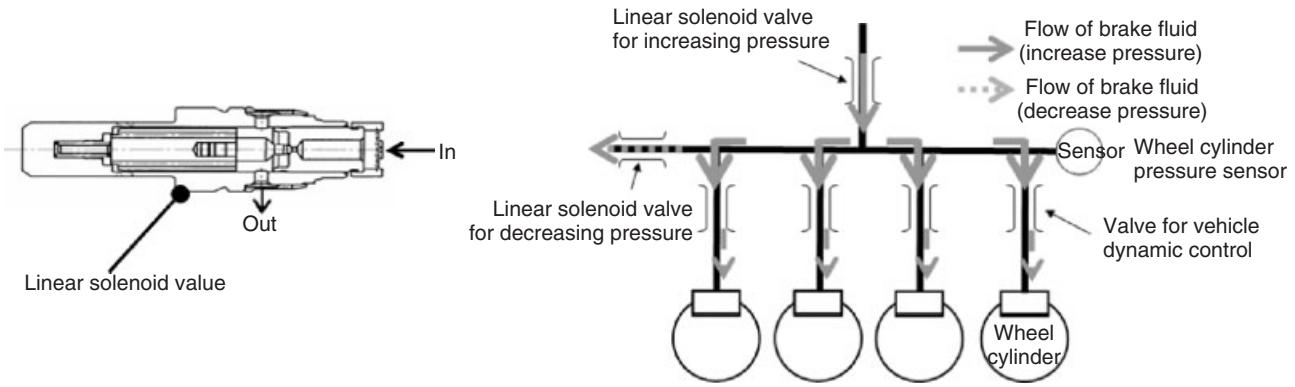


Figure 18. Adjustment of wheel cylinder pressure by control of linear solenoid valve. (Reproduced with permission from Nakata *et al*., 2009. © Society of Automotive Engineers Japan.)

force. This should simultaneously perform coordination of the regenerative-friction brake torque according to the required driver's intended vehicle deceleration.

In order to avoid brake pedal feedback due to the friction brake torque adjustment, use of pedal stroke simulator is common. This creates a natural brake pedal feeling (Nakamura *et al*., 2002; Nakata *et al*., 2009). This adjustment in the friction brake torque can be separated from the brake pedal feeling by controlling the variation in pedal stroke and pedal force.

The pedal stroke simulator uses the brake pedal stroke and reaction force generated by the consumption of brake fluid to recreate a natural brake pedal feeling that feels similar to a conventional braking system. The consumption of brake fluid in the wheel cylinder reacts to the coordination of regenerative-friction brake, and the pedal feeling does not depend on this pressure. This method can be explained by visualizing a spring and piston assembly to recreate the reaction force characteristics corresponding to wheel cylinder (Figure 19).

5 FUTURE CONTROL BRAKES

The benefits of ESC in vehicles have been socially recognized. The effect on accident reduction has prompted various regions around the world to change laws and regulation to make ESC mandatory on all new vehicles. These changes mean the installation of ESC is no longer limited to only luxury vehicles but has also increased to compact

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

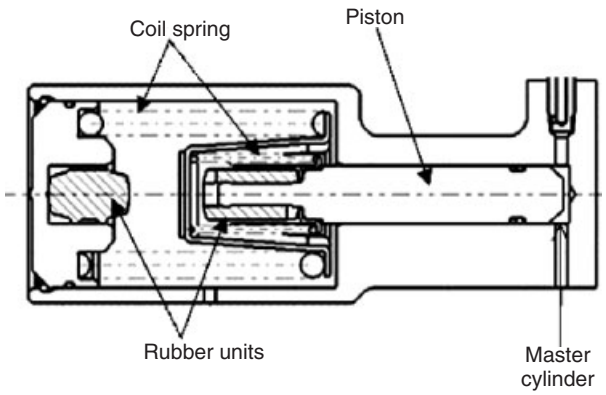


Figure 19. Example of pedal stroke simulator using springs and piston. (Reproduced from Nakamura *et al*., 2002. © SAE International.)

cars. Technology has advanced with demands for more lightweight, compact, and low cost ESC. In addition, there is a focus on collision mitigation, which can be realized by using the automatic brake function of ESC. This functionality can be realized by combining a sensor that detects an obstacle ahead (millimeter wave sensor, camera, etc.) and ESC. With the advances in this technology, there is a lot more attention being brought to the area of collision mitigation. It is expected that this will expand laws in many areas to make ESC standard on vehicles. The evolution into safer brake control technology is expected by linking with information over vehicle-to-vehicle communication or road-to-vehicle communication as well as perimeter monitoring sensors mounted on vehicles.

The role of the brake control has been increasing in the field of environmentally friendly vehicles. Advancements in engine technology to improve fuel efficiency have caused a reduction in engine vacuum levels. The utilization of brake control to reduce brake pedal force for low vacuum situations or for regenerative braking in the case of hybrid electric vehicle (HEV) and EV has been pursued. This means the brake control would not operate only during emergency situations but under normal braking conditions as well. Consequently, this drives the demand more than ever to advance the technology to produce ESC units that operate quieter than ever before.

REFERENCES

von Albrichsfeld, C. and Karner, J. (2009) Brake System for Hybrid and Electric Vehicles. *Proceedings of the SAE World Congress*, Detroit, USA.

Hano, S. and Hakiai, M. (2011) New challenges for brake and modulation systems in hybrid electric vehicles (HEVs) and electric vehicles (EVs). SAE Technical Paper 2011-39-7211.

Nakamura, E., Soga, M., Sakai, A., *et al*. (2002) *Development of Electronically Controlled Brake System for Hybrid Vehicle*, SAE International, Warrendale, USA.

Nakata, D., Nakamura E., Fukasawa, T., and Ohya, K. (2009) Development of the spread type electronically controlled brake system for hybrid vehicle. JSAE Technical Paper 20095631.

FURTHER READING

Obata, T., Ohtani, Y., Shirakawa, N., *et al*. (2011) Development of electronically-driven intelligent brake actuator with regenerative braking system. JSAE Technical Paper 20115172.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto200
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Body ECU Cluster

Isamu Sakurai
Yazaki Parts Co., Ltd, Shizuoka, Japan

1 Introduction 1
2 General Meters and Gauges in the Cluster 1
3 History and Development of Clusters 3
4 Types of Gauge Readings 4
5 Components that form a General Cluster 5
6 Conclusion 10

1 INTRODUCTION

A body electronic control unit (ECU) cluster (hereafter called *cluster*) is a device that is located at an easy-to-see position within a driver's line of sight and displays the conditions of a vehicle including information required by laws and regulations such as the speed of a vehicle. The information that the cluster displays allows the driver to not only monitor the conditions of the vehicle but also recognize that the vehicle may be in trouble. Consequently, the driver can take preventive measures before an accident happens. The most common location of the cluster is such that the meters or gauges can be viewed through the steering wheel. However, in some vehicles, the cluster is located at the center of the dashboard or at the position where the driver's line of sight for viewing the meters or gauges is over the steering wheel. How the cluster displays the information varies from one vehicle model to another. Therefore, this chapter focuses on the most popular cluster configuration that is introduced in vehicles released until 2011. The subsequent sections provide a detailed explanation of the history and development of the cluster, types of gauge readings, and components that form the general cluster.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto201
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 GENERAL METERS AND GAUGES IN THE CLUSTER

This section describes the key meters and gauges included in the modern cluster and the information that they display.

2.1 Speedometer

A speedometer is a meter that displays the speed of a vehicle (hereafter called *speed*). Speeds are expressed in either of the two units: kilometers per hour (km/h) or miles per hour (MPH). In countries where miles are used to measure distance, the speedometer displays speeds in both MPH and km/h, though MPH is used as the primary unit and km/h as the secondary unit. On the other hand, in most of the countries where the metric system is used to measure distance, speeds are expressed only in km/h. In Canada, speeds are expressed in both the units (km/h is used as the primary unit and MPH as the secondary unit). Many of the regions including Europe, China, Japan, the Middle East, and South Africa comply with Regulation No. 39 of the Economic Commission for Europe of the United Nation (ECE39) in terms of tolerances of the measuring mechanism of the analog cluster. In these regions, the speedometer has a plus tolerance so it is designed to read a little faster than the actual speed.

2.2 Tachometer

A tachometer is a meter that displays the number of revolutions of the engine (hereafter called *engine RPM*).

The engine RPM is expressed in the unit r/min, 1/min, or revolution per minute (RPM). Particularly, the analog cluster gives readouts in the unit $\times 1000$ r/min, 1/min $\times$ 1000, or RPM $\times$ 1000. As a result, the number of digits on the dial is minimized to present readouts as simply as possible. In addition, many of the tachometers have redlining to indicate that the engine is being revved up to the maximum safe limit.

2.3 Fuel gauge

A fuel gauge is a gauge to display the amount of fuel remaining in the tank at present. Many of the fuel gauges show the ratio of the full amount of fuel to the remaining amount. The pointer points to "F" or "1" indicating that the tank is filled up, or to "E," "0," or "R" indicating that the tank is empty. The typical fuel gauge has tolerances of the measuring mechanism. To be more specific, when the tank is filled up, the fuel gauge has a plus tolerance so it reads past "F" or "1." On the other hand, when the tank is empty, the fuel gauge has a minus tolerance so it reads past "E," "0," or "R." In addition, an increasing number of modern clusters have a fuel gauge symbol with an arrow indicating on which side the filler is located.

2.4 Coolant temperature gauge

A coolant temperature gauge is a gauge to indicate that the engine coolant temperature is at an appropriate level. Overheating or overcooling of the engine coolant may adversely affect the function of the engine of a vehicle. Therefore, the coolant temperature gauge helps the driver check to see that the engine coolant temperature falls within a safe range while driving.

2.5 Odometer/tripmeter

An odometer is a gauge that displays the total distance traveled by the vehicle over its life. A tripmeter is a gauge that displays the distance traveled by the vehicle up until the present moment since it was last reset to zero. As is the case with the speedometer, distance traveled is expressed in the unit km in many countries, but miles are also used in countries such as the United States or Britain. In addition, a switch is used to switch the display between odometer mode and tripmeter mode or to reset the tripmeter to zero. (Note that some tripmeters contain multiple pieces of data.)

2.6 Additional driving information

An increasing number of modern vehicles are equipped with a cluster that displays additional driving information besides odometer and tripmeter distance, such as instantaneous or average fuel economy, average speed of the vehicle, traveling hours, and distance that a vehicle can cruise until the next tank fill.

2.7 Information displayed by the cluster on a hybrid electric vehicle (HEV), electric vehicle (EV), or plug-in hybrid electric vehicle (PHEV)

Many of the hybrid electric vehicles (HEVs), plug-in hybrid electric vehicles (PHEVs), or electric vehicles (EVs) are equipped with a cluster that displays the amount of power consumed or the state of the battery charged by the regenerative braking system. Moreover, the indicators in the cluster assist the driver in monitoring his or her contribution to environmental conservation while driving in an eco-friendly manner. Such contribution is presented in visual form using a bar graph indicator or eco-driving symbols. In addition, EVs and PHEVs have indicators to show the remaining battery power available or the charger cable being attached to the charger.

2.8 Indicators

The driver can be informed about the operating state or failure of the systems in the vehicle by turning on or blinking the symbols arranged in the cluster. Some of the symbols illuminate with a beep sounded as a warning. In most countries, laws and regulations concerning vehicles define required indicators, symbols, and their colors. For instance, vehicles are mandated by law to have a brake warning indicator indicating that the brake is in trouble or a seatbelt warning indicator indicating that the seatbelt is not fastened.

2.9 Other information

Today, a wide variety of ingenious ideas are incorporated into clusters equipped with monochrome dot-matrix displays or color thin-film transistor (TFT) liquid displays, which may also differ from region to region or from vehicle to vehicle. For instance, such clusters are capable of

- prompting the driver to take a break based on the collected data about his or her driving behavior or continuous driving for longer hours than specified;
- assisting the driver in identifying the orientation of the wheels when starting the vehicle;
- informing the driver about preset calendar dates such as anniversaries as soon as the engine starts.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto201
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3 HISTORY AND DEVELOPMENT OF CLUSTERS

The first automotive meter was a governor-type speedometer introduced in Oldsmobile in 1901 (Figure 1). In recent years, a multiplex communication system has been introduced so a large amount of data required to display the information (e.g., speeds, engine RPM, and warnings) is transmitted as an input to the cluster from ECUs. Such data is processed in the microcontroller unit (MCU) built into the cluster, and consequently presented in either analog or digital form.

Now we consider the history and development of analog speedometers in the last 50 years as an example. An approach to signal conveyance and gauge reading is being developed to ensure higher accuracy. In early speedometers, a speedometer cable was used to mechanically convey the rotational speed of the axles to the speedometer. As the cable rotated, the magnet spun. The spinning magnet produced eddy currents, and the pointer turned in response to the strength of the produced eddy currents. Later on, an approach of converting the rotational speed of the axles into electric pulse signals was invented. The produced electric pulse signals were input to the speedometer. A current was passed to cross coils according to the input signals, and then the pointer turned in the direction of the generated magnetic field. Today, a multiplex communication system is used to convey speeds to the speedometer in accordance with defined protocols. A stepper motor moves the pointer at an angle so as to read speeds according to the input signals (Figure 2).

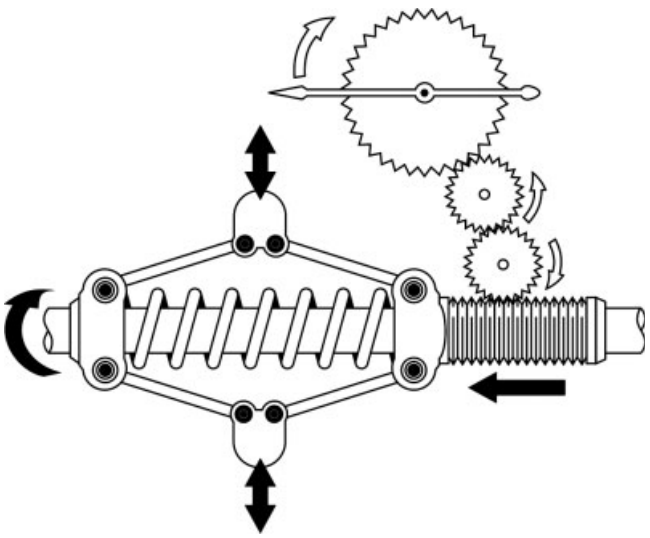


Figure 1. Simplified diagram of a governor-type speedometer.

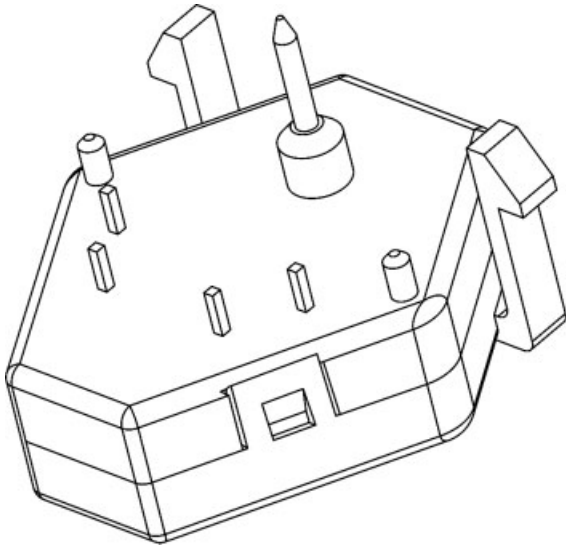


Figure 2. Simplified diagram of a stepper motor.

The cluster design has also been changing. Almost all the meters or gauges were separately enclosed in round housings, but now they are combined into in a single enclosure, which is called a *cluster*.

Next, the material that the meter, gauge, or cluster housings are made out of has changed from metal to general-purpose resin. A transparent material that the cluster front glass is made out of has changed from inorganic glass to transmissive resin such as polymethylmethacrylate (PMMA) or polycarbonate (PC).

An approach to illuminating the cluster has also undergone many changes for higher visibility especially at nighttime. An indirect lighting system was popular, where the dial and the pointer were irradiated with light from the front or side of the cluster. However, such system was replaced by a transillumination system to illuminate the dial, or by a self-luminous system to illuminate the pointer. The transillumination system is designed to illuminate the dial by allowing the transmission of light from the back. In the self-luminous system, the pointer seems to illuminate itself by light emitted from the light emitting diode (LED).

Moreover, the light source has changed from a bulb to a LED. Since the emergence of the digital cluster in the late 1970s, a dot-matrix display or a high definition display has taken the place of a segment display. In addition, a monochrome display has been replaced by a color display.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto201
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4 TYPES OF GAUGE READINGS

4.1 Use of a movement

Movement-equipped clusters are capable of indicating continuous and variable readings (e.g., speeds) by moving the pointer to point to an appropriate mark or number printed on the dial. Many of the clusters include movement-equipped meters or gauges such as a speedometer, a tachometer, a fuel gauge, and a coolant temperature gauge. This type of cluster usually comes with a small display unit that indicates odometer and tripmeter distance in digital form. This type of cluster is hereafter called *analog cluster*. Figure 3 illustrates an external view of a typical analog cluster.

4.2 Use of a display unit

In general, a cluster that employs a display unit is designed to convert measured values (e.g., speeds) into numbers and display discrete values in numeric or segment form on a liquid crystal display (LCD) panel or a vacuum fluorescent display (VFD). This type of cluster is hereafter called *digital cluster*. In this chapter, the term *digital cluster* refers to the cluster that shows key driving information (e.g., speeds) on the display, though some of the digital clusters use indicators rather than the display unit to indicate particular information. Figure 4 illustrates an external view of a typical digital cluster.



Figure 4. External view of a typical digital cluster.

As shown in Figure 5, an increasing number of digital clusters present graphic images on the color display unit that look exactly like the analog cluster.

4.3 HUD (head-up display)

Recent years have seen more vehicles equipped with a head-up display (HUD). The HUD system consists of a

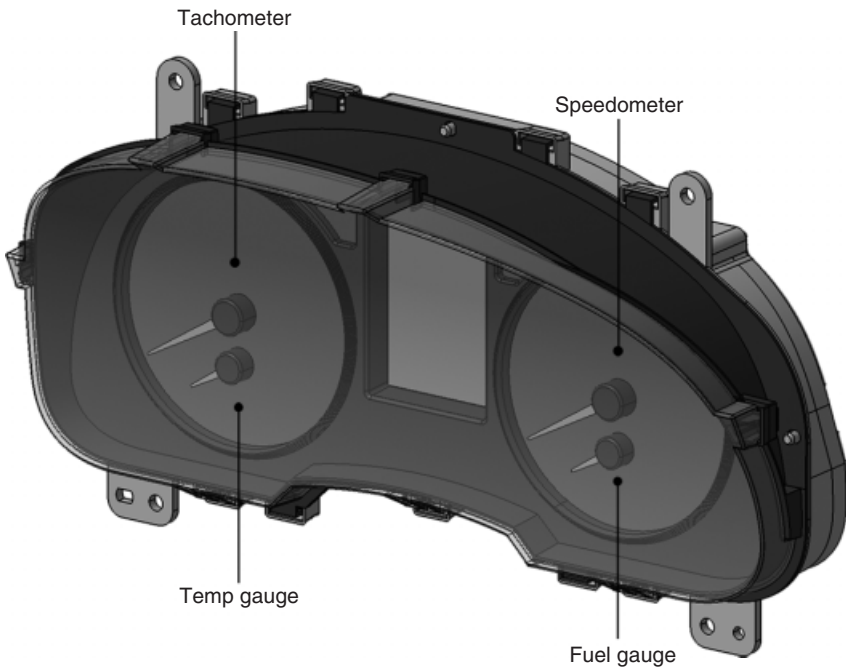


Figure 3. External view of a typical analog cluster.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto201
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

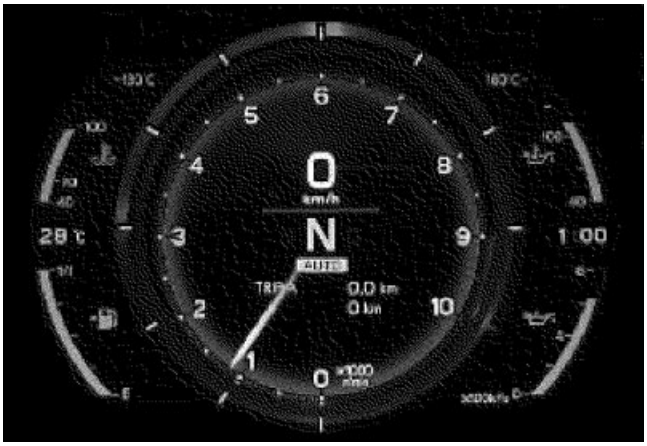


Figure 5. "Faux-analog" digital clusters.

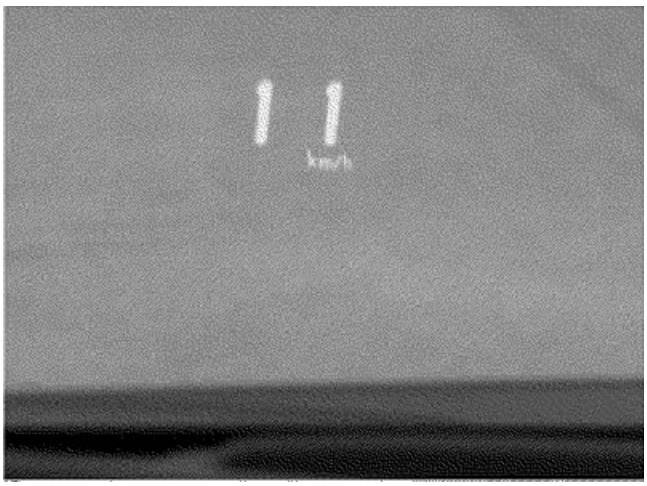


Figure 6. HUD that projects information on the windshield.

display projector, which is installed inside the dashboard and hidden from the view of the driver. Driving information on the display projector is projected on a transmissive and reflective panel either on the windshield (Figure 6) or on a specific reflector panel (Figure 7) pulled up from the dashboard. The information is projected in such a way that it appears a little farther away from the driver than it actually is. This allows the driver to get the projected information into his or her sight in less time. The HUD system also reduces the number of times the driver moves his or her eyes up and down. Therefore, the driver can recognize the projected driving information at a glance with his or her eyes kept on the road. These features of the HUD system are of great benefit to safe driving. Figure 8 illustrates the mechanism of the HUD system.

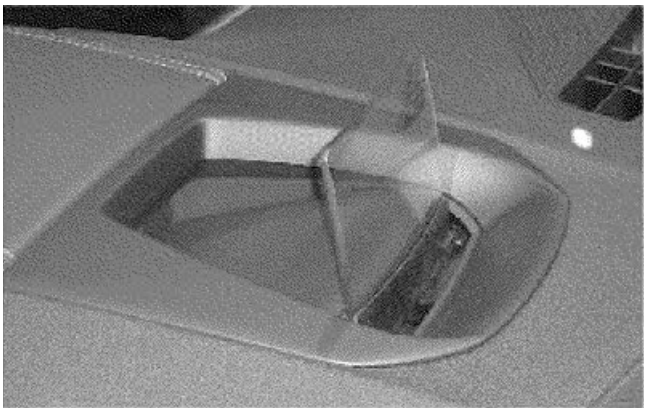


Figure 7. HUD that projects information on a specific reflective panel.

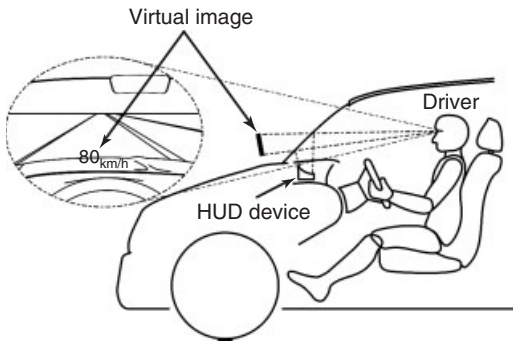


Figure 8. Mechanism of a HUD.

5 COMPONENTS THAT FORM A GENERAL CLUSTER

5.1 Parts breakdown

Figure 3 shows the external view of the general analog cluster, and Figure 9 its parts breakdown. This section, taking the speedometer as an example, describes the basic structure of the meters and gauges that gives the driver readouts using the pointer. A range of numbers and marks are printed on the dial plate (d in Figure 9). The pointer (c in Figure 9) is located at the center of the dial, surrounded with the printed numbers or marks. Once the movement (i in Figure 9) has been driven, it moves the pointer to read speeds. The light sources (k in Figure 9) are located inside the speedometer so as to illuminate both the dial and the pointer at nighttime (or at anytime of the day or night). The

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto201
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

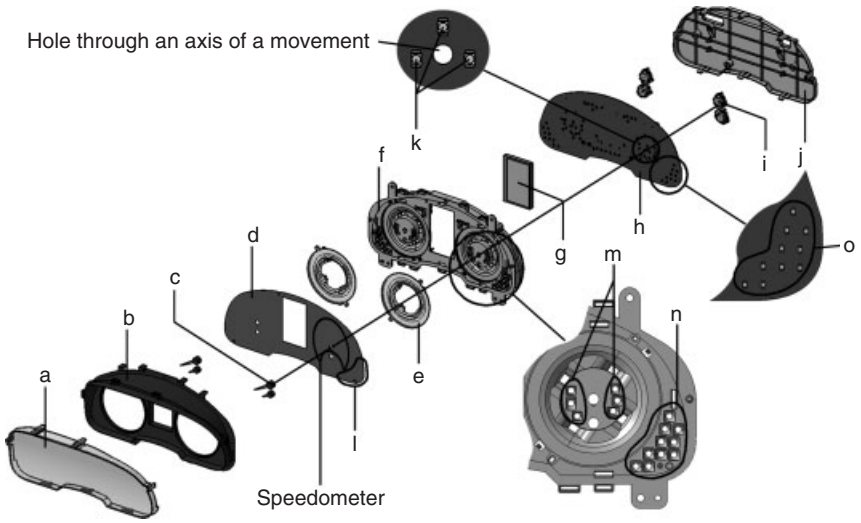


Figure 9. Parts breakdown of a typical analog cluster.

tachometer, fuel gauge, and coolant temperature gauge are structured in the same manner as described above.

There is space reserved for mounting the indicators (l in Figure 9). Below such space are individual lamp cases (n in Figure 9), in which the light sources (o in Figure 9) to illuminate the indicators are housed. As an increasing number of indicators are included in the cluster in recent years, more space needs to be reserved for them. Therefore, some of the indicators may be placed within available open areas located inside the individual gauges (m in Figure 9). In addition, the small display unit (g in Figure 9) is provided to show text or numeric information such as odometer or tripmeter distance.

Besides the above components, the cluster is composed of the following basic components: the transmissive front glass (a in Figure 9) to cover the cluster, the housing (b in Figure 9) to hold the front glass, the case (f in Figure 9) to secure the printed circuit board (PCB) (h in Figure 9) to the dial plate, and the rear cover (j in Figure 9) to protect the PCB. Some of the clusters contain prisms (e in Figure 9) immediately below the dial plate in order to maintain high brightness uniformity on the dial.

Next, Figure 4 shows the external view of the general digital cluster, and Figure 10 its parts breakdown. A wide variety of display units are introduced in the digital cluster as described in Section 5.2. Figure 10 shows the

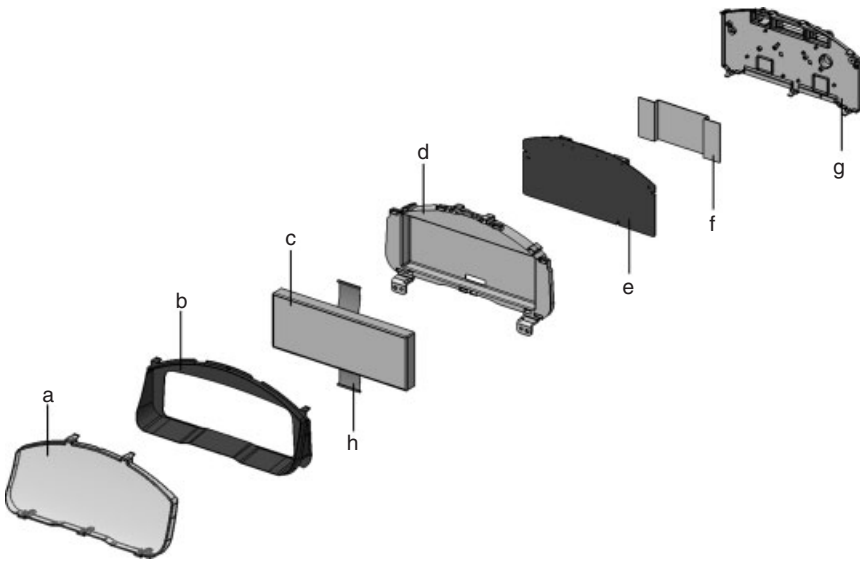


Figure 10. Parts breakdown of a typical digital cluster.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto201
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

components of the cluster equipped with a color TFT-LCD module (hereafter called *display module*). The display module (c in Figure 10) is secured in front of the case (d in Figure 10). The PCB (e in Figure 10) is mounted behind the case. The flexible printed circuits (FPCs) (h in Figure 10) connector, which is mounted on the display module, is electrically connected to the PCB. As is the case with the analog cluster, the front of the display module is covered with a transmissive front glass (a in Figure 10). The front glass is secured with the housing (b in Figure 10). At the back of the cluster is the rear cover (g in Figure 10), which is used to protect the PCB. If a controller (graphic display controller or GDC) that operates on a high clock frequency of 100 MHz or over is used to display graphic images, the cluster may be equipped with a noise-suppressing shield case (f in Figure 10).

5.2 Display types

Display types included in the cluster range widely, from a small display in the analog cluster to indicate odometer or tripmeter distance to a large full-color display in the digital cluster that looks exactly like the analog cluster. Display types vary in panel type or size and driving system depending on the cluster type in use. Therefore, the subsequent sections list some of the typical display types available in the clusters today.

5.2.1 Monochrome segment display

A monochrome segment display, which typically uses an LCD, consists of seven character-forming segments for displaying numeric information in odometer or tripmeter mode, a unit of distance (e.g., km), fixed icon patterns, and a bar graph indicator to present a fuel level in visual form. There are two types of segment displays: positive display and negative display. The image on the positive display becomes opaque when the segment is turned on, but becomes transparent when the segment is turned off. Conversely, the image on the negative display becomes transparent when the segment is turned on, but becomes opaque when the segment is turned off. A twisted nematic (TN) display is commonly used for the positive display. On the other hand, a vertical alignment (VA) display is commonly used for the negative display, featuring a high contrast and a wide viewing angle. In addition, in some of the display units, the entire display area or only the specific segment patterns are colored for better looks. They are usually colored by affixing a printed sheet to the LCD or by color printing on the inner surface of the glass plate of the LCD.

5.2.2 Monochrome dot-matrix display

A monochrome dot-matrix display, which typically uses an LCD, is designed to form text and numeric characters or graphics from a matrix of small dots. The dot-matrix display is useful in displaying text or numeric information including speeds in the font that cannot be formed by the seven-segment display. In addition, the monochrome dot-matrix display is capable of visually representing an increase or a decrease in a level using a bar graph indicator or messages in much greater details than the seven-segment display. A dual scan super twisted nematic (DSTN) display or a temperature compensation film super twisted nematic (TC-FSTN) display is a commonly used type of monochrome dot-matrix display. In the DSTN display, a compensated cell is affixed to a driving cell to compensate for optical property variations caused by changes in temperature. In the TC-FSTN display, an inexpensive compensation film is used as an alternative to a compensated cell.

5.2.3 Full-color display

A full-color display, which typically uses a TFT, allows more flexible graphical expression, and is capable of presenting high resolution images or animations. One pixel is composed of three subpixels: red, green, and blue. Detailed color representation is provided pixel by pixel as a result of controlling a gradation of levels of each subpixel. Compared with the monochrome dot-matrix display, the full-color TFT display features a higher resolution, a higher contrast, and a higher responsivity at low temperatures.

5.2.4 Other types of display

In addition to the display types described in the preceding sections, a VFD or an organic electroluminescence (EL) display is often introduced in the cluster. Both of the display types show speeds using segments and the engine RPM or fuel level using the bar graph indicator. In addition, a matrix of lights are arranged in such a way to express text and numeric characters or graphic images by turning selected lights on or off. Both of the display types feature a high contrast, a wide viewing angle, and high responsivity, thanks to their self-luminous system.

5.3 Hardware

Many of the modern clusters are designed to cover all of the basic circuit functions on a single PCB. For instance, the analog cluster has a circuit configuration as shown in Figure 11. It is the MCU (a in Figure 11) that is responsible

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto201
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

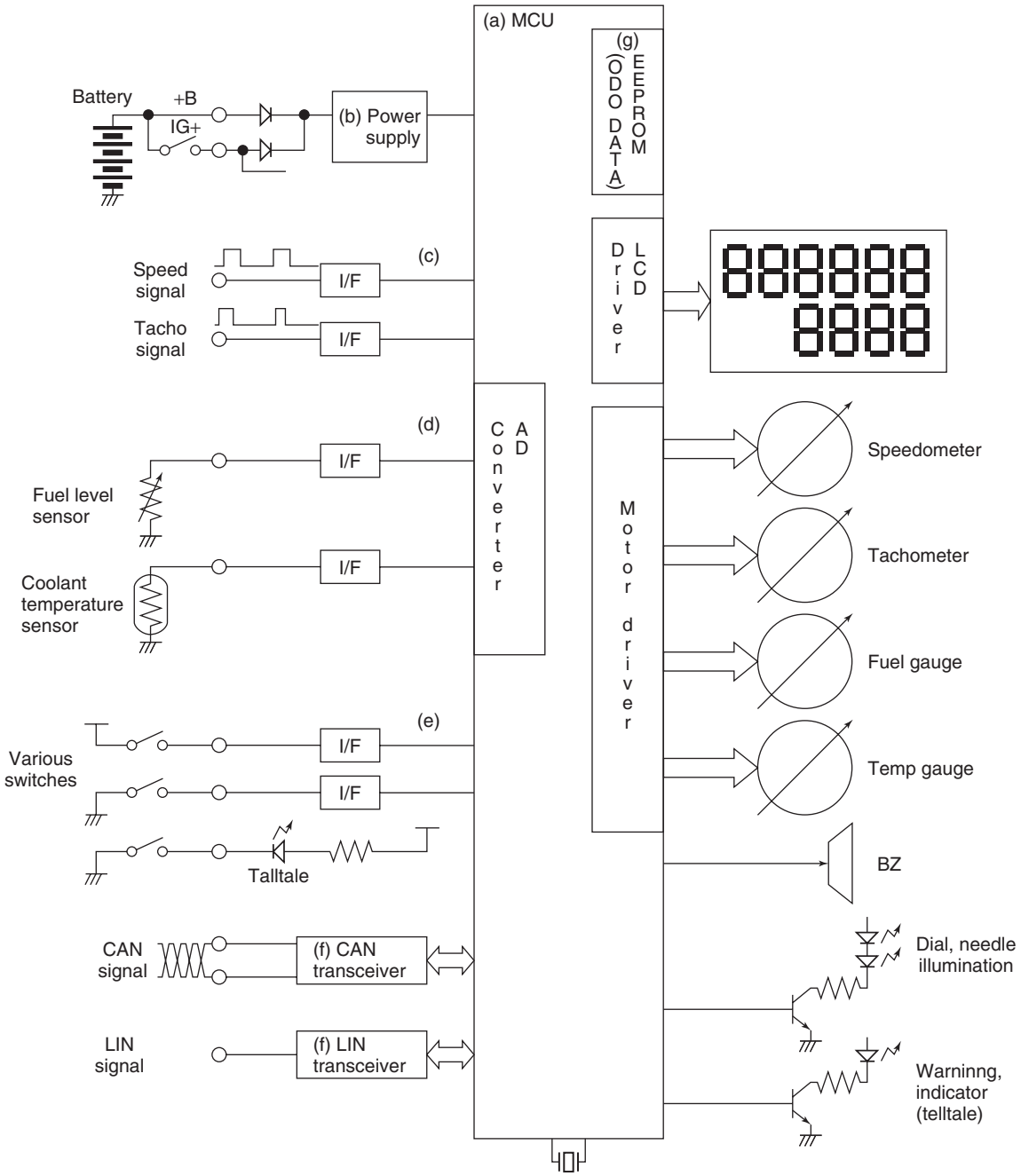


Figure 11. Example of a circuit diagram.

for almost all cluster functions. Power, specifically for the MCU, which is generated on the power integrated circuit (IC) (b in Figure 11), is derived from 12 V (or 24 V) supplied from the battery. Power is supplied to the MCU and the peripheral circuits. Such signals, as the speed signal and the engine RPM signal (c in Figure 11), which are transmitted to the cluster, need to be measured periodically. The voltage signals (d in Figure 11) from the fuel gauge sensor (the amount of fuel remaining) or the coolant temperature gauge sensor (coolant temperature) are captured into the analog to digital (A/D) converter. The switch signals (e in Figure 11) detect the on/off level of the brake, lighting switch, or A/T (automatic transmission) shift position. These signals are captured into the MCU via the specific interface (I/F) circuit, and undergo necessary processing within the MCU.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto201
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The produced output signals drive the stepper motor for gauge reading, the LCD to display odometer or tripmeter distance, the LEDs for the indicators, the buzzers to give warnings, or the LEDs to illuminate the dial or pointer. Some drivers are included in the MCU and capable of driving the stepper motor, LCD, or LEDs directly, and others are provided external to the MCU.

Odometer distance must be retained even when the battery is removed. Therefore, the data is stored in the nonvolatile electrically erasable and programmable read-only memory (EEPROM) (g in Figure 11).

Many signals are transmitted using a multiplex communication system such as controller area network (CAN) or local interconnect network (LIN). Such signals are captured into the MCU via the specific transceiver (f in Figure 11), and input/output processing is performed. As a communication system is more multiplexed, multiple signals are included in a CAN or LIN signal and transmitted together at once. This leads to a reduction in the number of circuit components required for the input I/F as well as simplification of the system. In recent years, an increasing number of clusters utilize a multiplex communication system to transmit not only the speed signal and the engine RPM signal but also the fuel level, the coolant temperature, and the switch signals to detect the on/off level.

5.4 Software

In modern clusters equipped with the MCU, the data input into the cluster is processed by software (or a program) contained in the MCU at any time. The meter or gauge readings (e.g., speedometer, tachometer, and indicators) are controlled in real time and kept updated. The C language or assembler is a commonly used program language. In digital clusters with a large full-color display unit, an authoring tool is also used because it allows a screen layout to be visually designed without in-depth knowledge of software.

As an example of processing performed by software in the analog cluster, the stepper motor is software controlled for smoother moving of the pointer instead of just being controlled according to the input signals from the vehicle. In addition, in the fuel gauge, where the input signals largely vary, filtering is performed. Filtering allows the gauge to display the current fuel level based on not only the instantaneous input signal but also averaging of the multiple input signals sampled. Such software control increases accuracy and provides the driver with an easy-to-view display. It applies to all software-processed gauge reading in both analog and digital clusters.

5.5 Illumination

In the analog cluster, both the dial and the pointer are illuminated in order to ensure high visibility of the cluster even at nighttime. Many clusters use LEDs as light sources. Printing on the dial plate divides the dial plate into two areas: transparent areas, which allow the transmission of light (e.g., printed marks or numbers), and opaque areas, which do not allow the transmission of light (areas other than the printed marks or numbers). High visibility is ensured by a contrast between these two areas when the dial plate is irradiated with light from the back. Light emitted from the LED reaches the hot-stamped section of the pointer. The light is reflected diffusely there and causes the pointer to be illuminated.

In some clusters with a dial plate made with authentic metal, an indirect lighting system is applied to illuminate it because it does not allow the transmission of light. Moreover, using an indirect lighting system in combination with such a dial plate can create a more luxurious appearance. Some clusters are illuminated at all times. In this case, they have a mechanism to dim the light at nighttime compared to daytime. Others allow the driver to adjust the brightness on his or her own. The digital cluster is typically illuminated at all times, and dimmed at nighttime.

5.6 Movement

In many analog clusters, the pointer is moved by the stepper motor. The stepper motor is capable of turning the pointer accurately in accordance with the output signals from the MCU because the rotational angle of the motor shaft is in proportion to the number of pulses input to the motor. Pulse signals can be used as input to the motor. In addition, the circuit configuration between the MCU and the motor can be simplified. For these reasons, use of the stepper motor has become the mainstream way of moving the pointer in the modern MCU-controlled analog cluster.

5.7 Decorations

The cluster is located in the driver's line of sight when he or she takes the driver's seat. Therefore, the cluster plays an important role as not only a display device but also a part of the interior design. Modern clusters are more stylized than ever before, including a stylish dial, pointer, or illumination. The dial is decorated with a ring. Moreover, finishing touches are added on the housing or ring with ornamental patterns, coating, plating, or printing for a more attractive appearance.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto201
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6 CONCLUSION

Since the beginning of the twentieth century, when a cluster basically had only a speedometer, modern clusters have made such great strides that they can now accurately provide all the information (as described in Section 2) to the driver in a manner that is easy to see and understand. It is likely that more and more information will be provided to drivers by various means in the vehicle. In addition, it is likely that the amount of information that drivers demand will grow steadily. In conclusion, clusters are sure to play even more important role in the future in presenting accurate, legible, and intelligible information to drivers.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto201
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Car Navigation

Shouji Yokoyama
Aisin AW Co. Ltd, Okazaki, Japan

1 Introduction 1
2 Sytem Overview 2
3 Car Navigation-Based System Integration with Other System 10
4 Future of the Car Navigation System 11
Related Articles 12
References 12
Further Reading 12

1 INTRODUCTION

Car navigation systems help drivers determine current location, search for destinations, and provide pinpoint guidance. They gained popularity in the 1990s and current usage has spread around the world. Before car navigation systems achieved mass market production, many earlier systems helped evolve the technologies behind the current systems. This section describes a brief history of these evolutions.

In 1909, one year after the launch of the first mass market automobile, the Ford Model T, J. W. Jones invented "Jones Live Map" in the United States. His equipment showed the real-time direction and distance to the destination using a turntable connected to a vehicle wheel. Chadwick's "Automatic Route Guide" and Rodes' "Route Indicator" are also known as *primitive route guidance systems* with predefined route data. These systems disappeared soon after printed road maps became widespread.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

In the late 1960s, Rosen *et al*. proposed the electronic route guidance system (ERGS), which realized route guidance according to the information from terrestrial beacons. Similar systems were investigated in Japan and West Germany in the 1970s. Although these systems achieved good results, none of them were realized due to the huge infrastructure costs required to deploy the systems.

The world's first inertial car navigation system is said to be the Honda Electro Gylocator launched in 1981 for the Japanese market. This system mainly consisted of a Gyroscope, CRT display, and replaceable printed maps on transparent films. The system calculated the current location from the gyroscope data and vehicle speed data. Then, it superimposed a bright dot indicating current location on the CRT overlaid by the map film. In 1985, Etak in the United States launched the world's first digital-map-based system, "Navigator" with map matching function. The car navigation system did not achieve mass market production until the 1990s when the Navstar GPS constellation became fully operational.

GPS was developed by US Department of Defense and one of the signals is freely accessible to anyone. The first GPS satellite was launched in 1978. Since then, many GPS-based car navigation systems were studied. For example, Mitsubishi Electronic showed a GPS and inertial car navigation system prototype at the 1983 Tokyo Motor Show and GM announced a GPS, inertial, and Loran-C hybrid system in 1984.

In March 1990, Mazda and Mitsubishi Electronics launched a GPS and inertial car navigation system for the Japanese market. Soon after, Pioneer launched an aftermarket system in June 1990. Car navigation systems gained popularity in Japan where the population is crowded into densely populated plains and the road structure is extremely complicated. Although the first generation systems only showed current location on the map, many companies

joined the market and technical innovations such as route guidance functions accelerated. In 1992, Toyota and Aisin AW launched a Voice Navigation System with fully automatic route planning and voice guidance. This system realized all the basic functions of current navigation systems.

In Europe, BMW and Philips launched a car navigation system in 1994. Mercedes and Bosch followed with "Auto Pilot" in 1995. Turn by Turn navigation systems without map display function gained initial popularity in Europe where road structure and road signs are well organized and driving speeds are higher than that in other regions. In 1997, the RDS-TMC traffic information service was launched and has grabbed an almost de facto standard position in the current traffic information service worldwide.

In the United States, Visteon launched "Avis Satellite Guidance" for rental cars in 1994. Although many car-navigation-related technologies were developed in the United States, safety- and security-oriented communication services such as GM's "OnStar" spread earlier than car navigation systems.

Location accuracy was advanced by methods such as using an accelerometer to detect slope, image recognition for precise positioning, and so on. In 1998, Toyota and Aisin AW realized an automatic transmission control system based on the function of the car navigation system.

Paralleling the mobile phone technology evolution, wireless-communication-based car navigation system technologies have been actively developed. In 1996, Denso launched its in-car communication information terminal "MUIT". Since then, various car-navigation-related communication services such as internet-based destination search, traffic information delivery and sharing, map data update delivery, route delivery, and so on have been realized by many service providers.

In 2000, "Selective Availability" was discontinued and GPS positioning accuracy was greatly improved. Simple location systems with only a GPS receiver could now obtain sufficient accuracy, so portable navigation device (PND) development accelerated. In 2004, TomTom in the Netherlands launched "TomTom Go". These types of PNDs exploded in popularity due to the high performance to cost ratio and excellent usability. Recently, mobile phone technologies have reached a level sufficient to realize navigation functionality. In 2009, Google launched a free navigation service with advanced voice recognition function for its Android smartphone.

The following sections mainly describe basic technologies of in-car navigation systems.

2 SYTEM OVERVIEW

Figure 1 shows the overall configuration of the car navigation system. A car navigation system consists of a navigation unit installed on the vehicle, a satellite positioning system, and a traffic information service.

The navigation unit consists of input sensors, processing, map database, and output devices.

2.1 Input sensors

The input sensors consist of a satellite signals receiver, distance travelled sensor, travel heading sensor, traffic information receiver, and user input device. Positioning is the most important function of the car navigation system. Position is calculated using a satellite positioning system as an absolute positioning sensor on the earth and using information from a distance travelled sensor and travel heading sensor as relative positioning sensors where no satellite signal reception is available. Relative positioning sensors are also used to adjust for positioning error. Owing to cost, simple car navigation systems in many cases do not use relative positioning sensors and instead only use the satellite positioning system.

2.1.1 Satellite positioning system (Absolute positioning sensor)

The US NAVSTAR GPS system and the Russian GLONASS are current satellite positioning systems, with others planned such as the European Union Galileo. This section explains the principle of widely used GPS positioning.

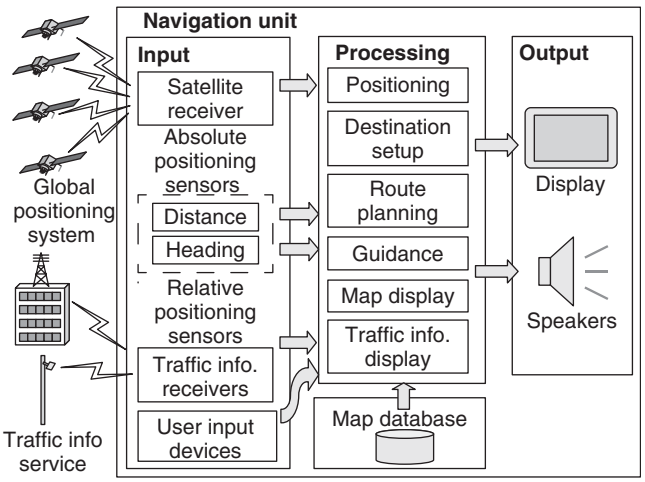


Figure 1. Car navigation system overview. (Reproduced with permission from Aisin AW Co. Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

A GPS receiver calculates the time of signal transmission from GPS satellite "" by checking the difference between the time the satellite transmits the GPS signal and the time the GPS receiver receives it. The distance from the satellite r_i is calculated by multiplying message transit time by the speed of the light c, where the subscript i is the satellite number.

Since the GPS signals contain their orbital (absolute position) information, as shown in Figure 2, it is theoretically possible to determine a receiver position (x, y, z) on the earth if the receiver receives the signals from at least three satellites. In reality, the GPS receiver's clock information contains a significant error δ. So, messages from at least four satellites are needed to be received in order to solve the equations with unknown values x, y, z, and δ. By assuming the true distance from satellite i is the calculated distance r compensated by the clock error δ, the equations for each satellite are as follows;

$$(x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2 = (r_1 + \delta c)^2$$
$$(x - x_2)^2 + (y - y_2)^2 + (z - z_2)^2 = (r_2 + \delta c)^2$$
$$(x - x_3)^2 + (y - y_3)^2 + (z - z_3)^2 = (r_3 + \delta c)^2$$
$$(x - x_4)^2 + (y - y_4)^2 + (z - z_4)^2 = (r_4 + \delta c)^2$$

The absolute receiver position (x, y, z) and the clock error δ can be derived by solving these equations. When the vehicle is in motion, it is also possible to calculate the absolute vehicle speed and the absolute vehicle heading by detecting and synthesizing the carrier frequency deviations of the GPS signals caused by the Doppler effect.

The positioning accuracy of current GPS receivers is around 10 m under open sky conditions. The accuracy is degraded by multipath reflection from the building, terrain, and road surfaces. The largest cause of GPS positioning error in an open sky condition is the unpredictable delay through the ionosphere.

Although technology that can compensate for ionosphere delay by using correction data from fixed stations (D-GPS or RTK-GPS) has been realized, current car navigation systems do not use them because of the additional cost. It is also possible to correct for ionosphere effects by checking the difference between at least two signals from a satellite. The US government is planning to add new civil signals in different frequencies. Because of this, the expectation is for positioning accuracy to be improved in the future.

The road surface reflection can be suppressed by installing the GPS antenna on the metallic roof, or putting a metallic plate (or specially designed rings known as a *choke ring*) under the antenna. It is also possible to filter the deteriorated reflection signals while the vehicle in motion.

2.1.2 Distance travelled sensors

The vehicle wheel speed sensor is often used as a distance travelled sensor. Owing to slip occurring at the drive wheels, the non-driven wheel speed is preferable. Some car navigation systems use integration of the accelerometer output for easier installation, but care is needed about significant accumulated errors.

2.1.3 Travel heading sensors

There are two ways to detect travel heading: absolute direction detection and integration of relative direction movement. Geomagnetic field detection is the typical method to get absolute direction, but the magnetic field has a tendency to be affected by the surrounding environment such as bridges or railroads, magnetization of the vehicle body, and the magnetic field inside the vehicle system. While

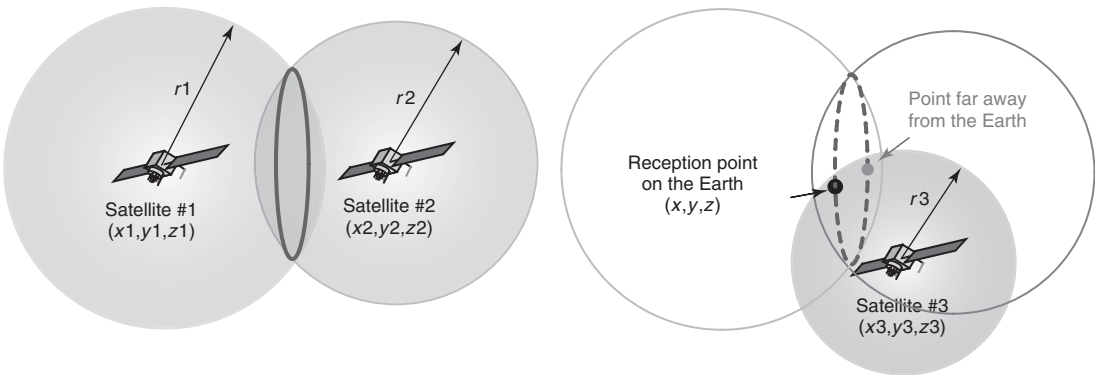


Figure 2. GPS positioning. (Reproduced with permission from Aisin AW Co. Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the vehicle is in motion, the above-mentioned GPS-based absolute heading method is more accurate and stable.

A gyroscope is generally used for measuring heading. Car navigation systems use low cost and compact Coriolis Vibratory Gyroscopes instead of fiber optic gyros that are typically for aviation use. It is also possible to detect vehicle heading from the wheel speed difference between the left and right wheels or steering angle information, but few systems use these methods.

The car navigation system determines the heading of the vehicle by a combination of absolute and relative headings. In many cases, systems mainly use the GPS-based absolute heading and compensate using the gyroscope during low speeds or when there is no GPS signal reception.

2.1.4 Traffic information service

By utilizing traffic information on the car navigation system, the driver can understand the real-time road situation, alternate routes to avoid congestion, and a more accurate estimated time of arrival.

Traffic information is generally generated at public or private traffic information centers and is based on roadside sensor data, communication terminal data in vehicles (Probe Car Data), local police department information, and so on.

There are several traffic information transmission protocols. For example, the widely used TMC (Traffic Message Channel) protocol is standardized as ISO-14819. Also used in car navigation systems in some regions are the Japanese VICS, VICS-based Chinese RTIC, and TMC's next generation protocol, TPEG (Transport Protocol Expert Group) that can also deliver other information such as fuel price and public transport information.

The car navigation system acquires the traffic information via broadcast or wireless communication such as FM subcarrier (radio data system (RDS), data radio channel (DARC)) broadcast, digital broadcast, roadside beacon, or mobile phone. The system then displays the traffic situation on a map, provides congestion guidance by either voice or display, recalculates the route considering the congestion, and shows the updated estimated time of arrival.

Some car navigation systems or traffic information services store historical traffic patterns and utilize it for future traffic estimation according to the time of day, day of the week, and so on.

2.1.5 User operation input

Touch screen displays and control methods such as a rotary switch or joystick are widely used user operation input methods. Numerous systems display software keys on a touch screen where a user can enter information by touching the screen, operating a rotary switch, and so on.

Resistive and capacitive touch screens are widely used in car navigation systems. A resistive touch screen consists of two thin transparent electrically resistive layers facing each other with a thin gap. One layer applies electrical voltage, and the system detects the touch location by analyzing the associated voltage on the other side. Capacitive touch screens detect the position by measuring the change in capacitance between the transparent conductor-coated screen and the finger.

Since the user only needs to directly touch the software key on the screen, the touch screen is intuitive and easier to use. On the other hand, the screen has to be installed close to the driver seat so that the driver can reach it.

Remote control methods such as a rotary switch or joystick are less intuitive than the touch screen method and the user has to become acclimated to operating them. Since the screen and the operation control are separated, it is possible to install the screen far from the driver so that the driver's eye movement is reduced, and the operation control is close to the driver.

Some systems add a microphone for speech input. In many cases, speech input is integrated with a touch screen or remote control. Speech input is usually used for complicated inputs such as address or POI (point of interest) input.

2.2 Map database

A map database is the digital data representation of required information for the car navigation system. The data consist of background data, road data, index data, and other data. The background data are used for map display. The index data are used for destination search. The road data are used for positioning, route planning, and map display. Reference tables in the other data are used for traffic information display.

2.2.1 Background data

The background data represent areas (ocean, lake, pond, river, building footprint, wooded area, forest, farm, etc.) in polygons, lines (railroad, narrow stream, etc.) in polylines, and the location of texts or symbols in map lettering information (Figure 3).

The polygon or the polyline data contain the category of the shape and its drawing rules (conditions to show or not show the shape, etc.). The map lettering information also contains the category of the text or the symbol and its drawing rules (drawing the text along the road shape, etc.).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

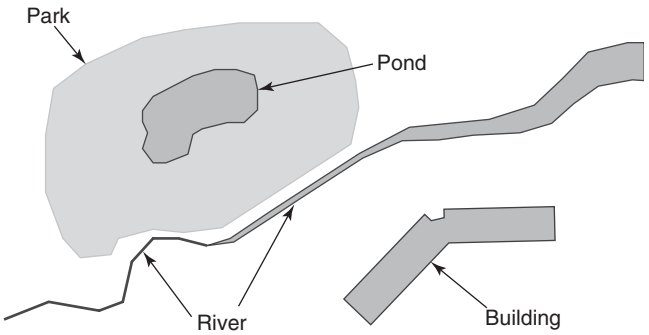


Figure 3. Background data. (Reproduced with permission from Aisin AW Co. Ltd.)

2.2.2 *Road data*

The road data are polyline representations of the centerlines of the corresponding roads. In some cases, both directions of a single road are represented by separated polylines or some portions of the lanes are represented by individual polylines. Crossroads are represented by points, which are connected to the polylines of the corresponding roads. The above-mentioned polylines and points are defined as links and nodes. A node represents an intersection point of the road network such as a crossroad, and a link represents the road portion between the nodes (Figure 4).

The road representative polyline has attributes such as the road administrator information, the function of the road (main line, ramps, service road, etc.), structure (highway, bridge, tunnel, etc.), width, number of lanes, one-way or two-way, speed limit, and so on. The node has attributes such as traffic signals existence, stop sign locations, traffic regulations, and simplified maps for the route guidance function. The road data also often contain a main road network data set in order to reduce the calculation time in long distance route planning.

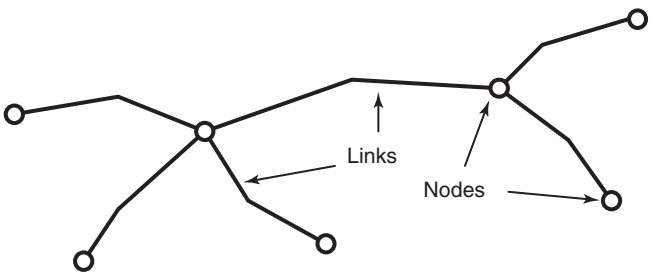


Figure 4. Links and nodes. (Reproduced with permission from Aisin AW Co. Ltd.)

2.2.3 *Index data*

The index data correspond to the location coordinates of a polygon of an area (or a building, or a point when there is a representative point). The index data can consist of a name of the area, an address, a postal code, a name of the intersection, a name or category of the POI, more detailed information such as business hours, and so on.

2.2.4 *Other data*

Other data mainly consist of a data for linking external information to the map database of the car navigation system. One example is a reference table for linking the external traffic information to the roads in the map database. Voice data for the route guidance function are also categorized as other data.

2.2.5 *Storage format*

The storage format of the database should follow a standard in order to reduce the development cost and time for designing or updating the map database for the car navigation system.

As an example standard, this section includes an abstract of the logical data storage format of the Japanese JIS-D0810 format proposal for ISO/TC4/WG3 international map database standardization working group (Hamada, 2002). The logical format portion was published as TS20452 in 2007.

Data structure describes the real world by shape, attribute, and connection-related information. They are represented by background data, road data, network data, and index data (Figure 5). The area is divided by the rectangular-shaped "parcel" data for map display, positioning, route guidance, and so on, and by the arbitrary-shaped "region" data for route planning. Both the parcel data and the region data have a layered structure in order to realize quick indexing in certain situations.

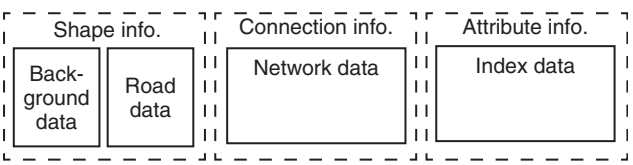


Figure 5. Data structure. (Reproduced with permission from Aisin AW Co. Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.3 Processing

The processing consists of a positioning module that determines the vehicle position, a destination setup module that supports the user to set the destination, a route planning module that calculates the route to reach the destination considering various conditions, a guidance module that generates the route guidance via both voice and image at the appropriate timing, a map display module, and a traffic information display module that superimposes the traffic information on the map.

2.3.1 Positioning

Positioning can be realized by a hybrid navigation method, which is a combination of inertial navigation, map matching methods, and GPS positioning.

An inertial navigation is a simple method to estimate the current location by integrating the above-mentioned distance travelled sensor and travel heading sensor information. It has a drawback of accumulated error due to integration of the detection errors.

A map matching method can compensate the problems of inertial navigation. Map matching is a method to compare the inertial navigation trace and the corresponding road shape on the map and determine the most probable current location. The combination of these methods is useful as long as an initial location is determinable, the map data are accurate, and the vehicle is traveling on a mapped road (Figure 6).

GPS positioning is a method to obtain the vehicle's latitude, longitude, altitude, and traveling direction. Although it is a discrete value, the system can estimate the absolute position as a circle (sphere) with a specific probability. Many simple car navigation systems only use GPS positioning as it can derive an absolute position and direction very simply. On the other hand, they cannot identify the location when there is no GPS satellite signal reception such as in tunnels, in urban canyons, on roads under highway overpasses, in underground car parking lots, on sides of steep mountains, and so on.

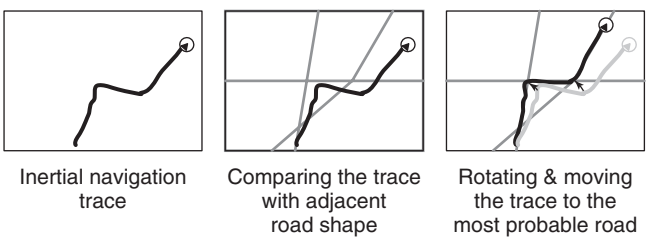


Figure 6. Map matching. (Reproduced with permission from Aisin AW Co. Ltd.)

Many car navigation systems use a hybrid navigation method that combines the inertial navigation, map matching, and GPS positioning. The system modifies the current location based on the GPS location when there is no inertial navigation trace data at the system initial setup, or when the estimated position is not located inside the GPS's probability circle. The inertial navigation and map matching are used when there is no GPS satellite signal reception such as in a tunnel.

2.3.2 Destination setup

The destination setup module supports the user to set the destination (latitude, longitude) interactively using the user input devices. The car navigation system provides various ways to support destination inputs: pointing to the destination directly from the map, searching the map database by inputting the name of the destination or address, selecting from the category list, and combinations of the above-mentioned methods. Some of the methods differ from country to country, region to region, and language to language.

2.3.3 Route planning

The route planning module derives the appropriate route to reach the destination from the road network data consisting of nodes and links.

As nearly all roads lead to the destination indirectly, there are almost infinite paths in the real road network. It is time-consuming and not practical to check all the possible paths. Several algorithms have been proposed to simplify the calculation. A Dijkstra's algorithm is one of the classic methods and often used in car navigation systems.

Furthermore, the road network should be layered in order to reduce calculation time for effective route planning. In the general case, people select the main network roads rather than the minor streets when the travel distance is long. So, the database structure is layered so that the upper layer, which only contains main network road data, is used in longer distance route planning (Figure 7).

In order to calculate the appropriate route, the road data contain various additional information called *costs*. In addition to the length, the link data contain various costs according to category, number of lanes, toll road status, speed limit, and so on. The node data contain traffic signal cost, left-turn cost, right-turn cost, going-straight cost, and so on. When the user defines the route planning preference (e.g., allowing or avoiding toll roads), the system selects the costs to be considered in order to calculate the route that takes the user's preference into account.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

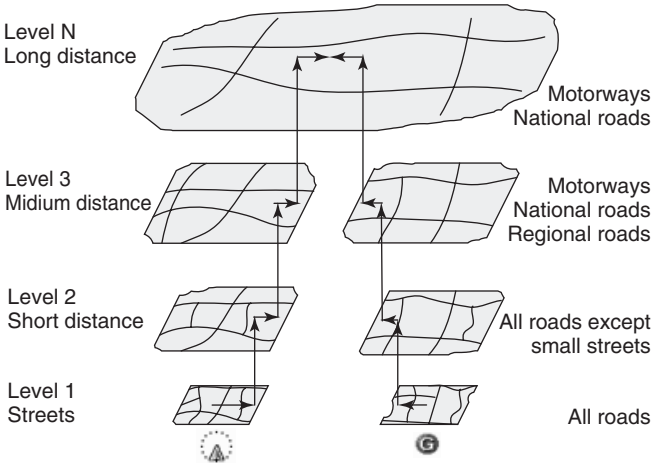


Figure 7. The layered road network database for route planning (Region data). (Reproduced with permission from Aisin AW Co. Ltd.)

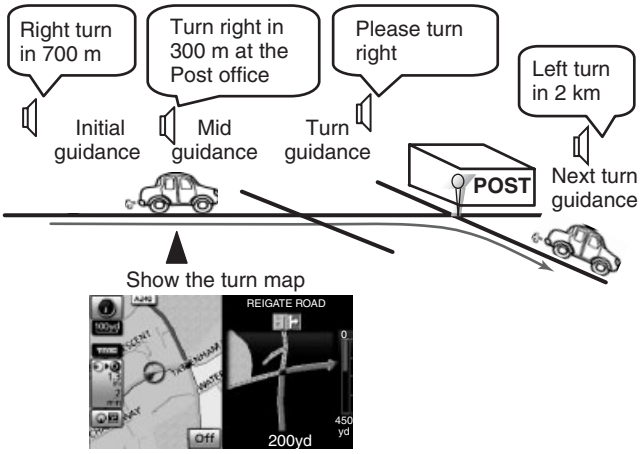


Figure 8. Route guidance example. (Reproduced with permission from Aisin AW Co. Ltd.)

The real-time traffic information can be considered in order to derive the "quick" route that considers congestion, road closures, road work, and so on. In this case, the system increases the cost of the links with traffic problems according to the situation.

Some systems have begun to utilize wireless communication to provide "quicker" routes that consider precise and wide area coverage traffic information, region-specific routes, fuel-efficient routes, scenic routes, and so on.

2.3.4 Route guidance (simplified turn map, voice guidance)

The route guidance module shows the necessary information for reaching the destination such as turn directions, landmarks, and the like to the user by voices and images at appropriate timing while the vehicle is traveling along the planned route.

The system determines at which turns or crossroads along the route to provide guidance from the relationship among the entrance link, exit link, and other links; the connection angle among the links; the road category; the road name; the road width; and so on. In some very difficult cases, additional field surveys alone can determine the necessity of guidance.

Turn guidance timing is determined according to the road category, road width, road shape, and so on. Some systems adjust the timing according to the vehicle speed in order to make guidance at appropriate timings.

A basic route guidance example is shown in Figure 8.

There are other kinds of guidance functions along with the basic route guidance. For example, a "turn by turn list" shows the list of turns with distances and directions, a "simplified highway map" shows the list of highway facilities and infrastructures with their names, distances, and estimated times of arrival (Figure 9).

Some systems use a secondary display inside the instrument panel or the heads up display (HUD) that can project



Figure 9. Turn by turn list and Simplified highway map. (Reproduced with permission from Aisin AW Co. Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

various types of information such as direction, distance to the next turn, lane information, and the like on the windshield (Figure 10).

The route guidance differs from region to region. For example, a crossroad is defined by the combination of the street names in Europe and the United States, whereas a crossroad is directly defined by the crossroad name in Japan. There are also regionally specific turn types such as Michigan lefts in the United States and P-turns in South Korea (Figure 11). Consideration is also needed due to ideal route guidance timing and display also differing from region to region.

Some systems capture the front view image by a windshield mounted camera and superimpose the route guidance on a live image. In the future, guidance images will be integrated with above-mentioned HUD or displayed inside the instrument panel.

The ultimate route guidance is something like a passenger very familiar with the area who recognizes the situation with their eyes (cameras) and brain (ECU), then shows the direction by voiced instructions (speakers) and gestures (displays) at the appropriate timing, and adjusts according to the driver's skills and surrounding environment.



Figure 10. Route guidance example using a secondary display. (Reproduced with permission from Aisin AW Co. Ltd.)

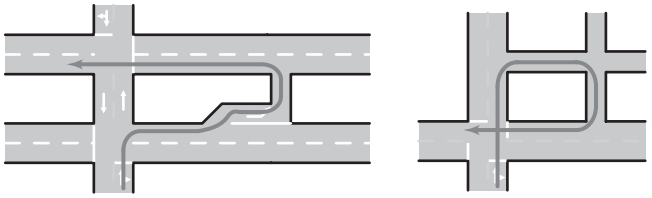


Figure 11. Michigan lefts and P-turns. (Reproduced with permission from Aisin AW Co. Ltd.)

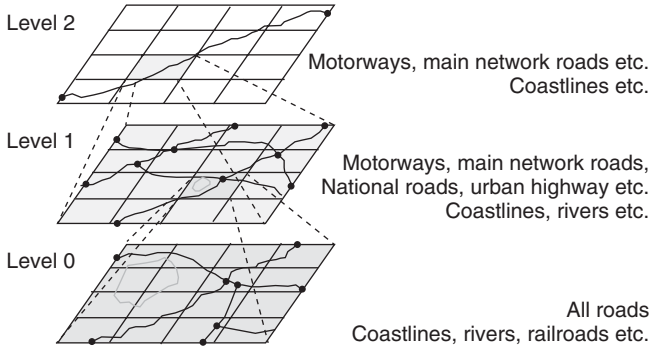


Figure 12. Layered map database structures (parcel data). (Reproduced with permission from Aisin AW Co. Ltd.)

2.3.5 Map display

The map display module provides location information to the user by showing the adjacent background, road data, and so on, stored in the map database (Section 2.2), then superimposing the current location, route information, destination, and so on.

The map database has a layered structure. Each layer corresponds to a certain scale of the map and stores background and road data according to the level of layer as a "parcel" of information (Figure 12).

First, the system determines the area to be displayed from the desired map scale, position, display size, and drawing orientation of the map. After reading the corresponding parcel data, it transforms the coordinates and draws the transformed data.

Since the map database has a layered structure, the system can realize quick zooms in and out, and wide area drawing while keeping the data size small.

The user can select the drawing orientation from north-up, heading-up that changes orientation according to the travel heading, or bird's-eye that can show more nearby detail, and so on. Each orientation can be realized by transforming the coordinates. For example, the coordinate transformation equation for heading-up is as follows:

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}$$

where $[X,Y]$ is the value on the original coordinates, $[X', Y']$ is the value after the transformation, and θ is the transformation angle (Figure 13).

Some systems show an intuitive and attractive 3D terrain map display using altitude information and 3D image transformation (Figure 14).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 13. North-up map and heading-up map. (Reproduced with permission from Aisin AW Co. Ltd.)



Figure 14. Bird's-eye map and 3-D terrain map. (Reproduced with permission from Aisin AW Co. Ltd.)

Future map displays will obtain more up-to-date information from external resources and show relevant and desired information in more intuitive ways.

2.3.6 Traffic information display

The traffic information display module provides traffic information such as congestions, hazards, and so on to the user by displaying the information by text, simplified image, or superimposition on the map.

The navigation system translates the location format of traffic information from the traffic information centers into the road data format of the navigation system's map database in order to superimpose the information on the map. The location format of the traffic information is defined by each traffic provider (VICS link numbers are used for VICS, location codes are used for RDS-TMC). The system stores location reference tables in the map database (Section 2.2.4), and refers the location of the traffic information to the corresponding road information, then superimposes the traffic information on the map. An example of the VICS traffic information display is shown in Figure 15

The Japanese VICS has three types of methods to display the traffic information: text, simplified image, and superimposition on the map (Figure 16).

For RDS-TMC, the traffic information is superimposed as icons and arrows that show the traffic flow situation

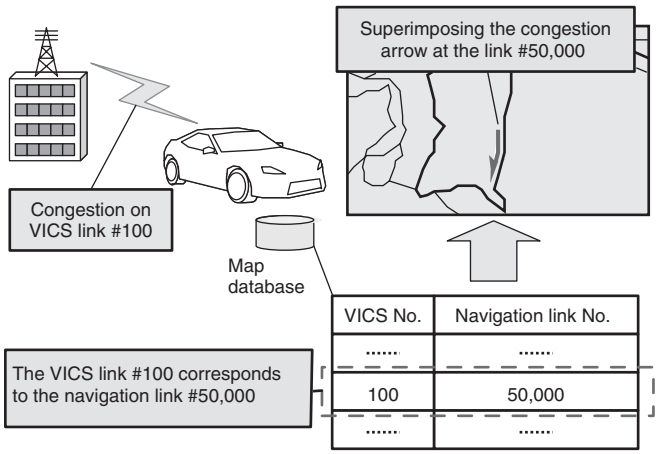


Figure 15. Traffic information display. (Reproduced with permission from Aisin AW Co. Ltd.)

on the map. The RDS-TMC also expresses the traffic situation in text list format. The driver can understand the traffic situation details by selecting items from the text list (Figure 17).

2.4 Output

The output devices consist of displays to show various types of information and a speaker to provide information by sound and voice.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 16. VICS simplified image and superimposition on the map. (Reproduced with permission from Aisin AW Co. Ltd.)

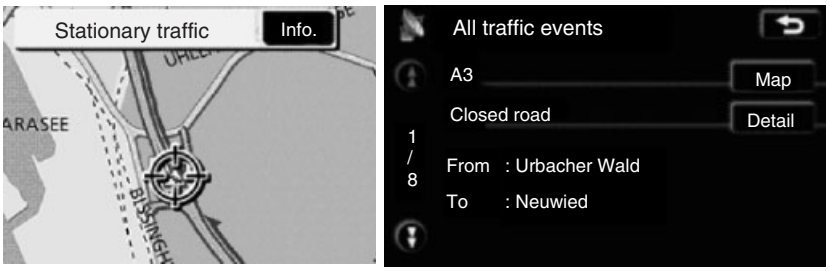


Figure 17. RDS-TMC superimposition on the map and text list. (Reproduced with permission from Aisin AW Co. Ltd.)

2.4.1 Display

A display provides the map and information such as various menus on high resolution screens installed around the cockpit. Some systems have a twin display (main display and secondary display) configuration, and others have an ultrawide display.

2.4.2 Speaker

The car navigation system normally shares the speaker for voice output with the car audio system. Only a subset of speakers is used for the voice guidance or the guidance voice is superimposed over the audio sound so that the user can notice the guidance voice of the car navigation system while continuing to listen to the sound from the car audio system.

3 CAR NAVIGATION-BASED SYSTEM INTEGRATION WITH OTHER SYSTEM

Car navigation systems can support realizing innovative systems by connecting to the cellular communication network or the vehicle system network.

3.1 Telematics

The car navigation system can get external information and send the information from the vehicle to the external world by incorporating a wireless communication function such as a cellular phone module. Various functions such as phone, emergency call, operator assistance, remote vehicle diagnosis, and destination search from the internet have been realized in this area. This section describes some parts of such functions.

3.1.1 Probe car data

Some navigation systems integrate a probe car traffic information technology, in which the car sends the measured position or speed data to the traffic information center and the center sends back traffic information on the roads beyond the coverage areas of current traffic information services such as VICS or RDS-TMC.

There are various methods to implement probe car traffic information: upload current position as determined in the positioning module, upload speed, or upload congestion information determined by the car navigation system. Uploading the congestion information can reduce the communication cost, while a simple method of uploading the current position has an advantage in expanding the usefulness because of its flexibility. Some systems also

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

implement uploading the fuel consumption data for fuel-efficient route planning.

The probe car system is also effective for supporting the recovery from disaster. Honda, Toyota, Nissan, Pioneer, and ITS Japan integrated their probe car data after the massive earthquake on March 11, 2011, and provided a "passable road map" of the disaster area.

3.1.2 Map update

The car navigation system needs to use the latest map database in order to maintain good performance. Off-board navigation systems and onboard navigation systems with incremental map update functions have been realized using wireless communication.

The off-board navigation system has the map database on the server side, and only the necessary portion of information is transmitted via wireless communication. Off-board solutions are widely used in cellular-phone-based navigation systems. They become useless outside the coverage area of the wireless communication, although some systems solve this issue by downloading the necessary information in advance.

An onboard navigation system with an incremental map update function can keep the latest map database with low communication cost by downloading only the updated portion of the map database and synthesizing inside the car navigation system. The synthesizing process is a background task, so performance reduction needs to be taken into account when the incremental map update function is designed.

3.2 Collaboration with the vehicle system

By connecting the car navigation system to the in-vehicle network, the car navigation system can utilize various vehicle sensor data and support the vehicle control system.

3.2.1 Car-navigation-system-based driver support system

Car navigation systems can predict the oncoming road conditions from road shape data in the database, current position, speed, and the planned route. They can support improving vehicle system control by providing feed-forward information.

For example, an automatic transmission control system can detect the driver's intention to decelerate from the upcoming corner shape information, driver's accelerator pedal operation, and brake pedal operation, and then the system can prevent unnecessary upshifting. It can also downshift based on the driver's accelerator or brake pedal operation before a downhill corner so that the vehicle can accelerate smoothly after passing the corner.

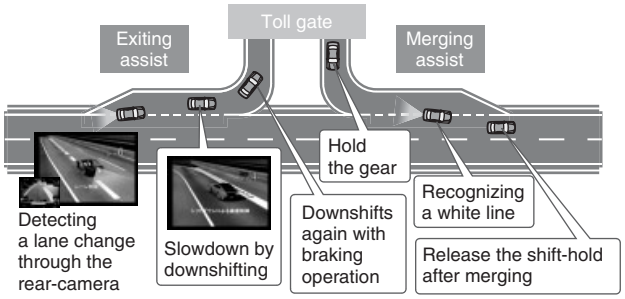


Figure 18. Camera-based automatic transmission control on a highway. (Reproduced with permission from Aisin AW Co. Ltd.)

3.2.2 Collaboration with the in-vehicle camera

Recently, rearview cameras and forward view cameras are beginning to be installed on vehicles to improve safety. Some car navigation systems can recognize road paint patterns (lane marker, stop line, pedestrian crossing, etc.) from the captured images and compare with map data in order to improve positioning. This precise vehicle position is used for advanced navigation system functions such as lane change guidance or for driver assistance systems such as automatic transmission downshifting to support deceleration while the vehicle is exiting a highway (Figure 18).

There is also image recognition technology to acquire the speed limit information from traffic signs and then display the speed limit information on the car navigation system screen.

4 FUTURE OF THE CAR NAVIGATION SYSTEM

Thanks to the automobile, people can freely move to distant places with relatively low cost. On the other hand, people have to concentrate on their location and finding the directions to their destination when driving in unfamiliar places. Some people may enjoy exploring places. However, sometimes it may lead to unsafe distracted driving.

Car navigation systems can support the driver by identifying the current location on the map and guidance to the destination. So, they have realized safer and more comfortable driving, meeting driver's demands. Since they can enhance the freedom of movement, car navigation systems have gained worldwide popularity.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In the future, the car navigation system will become a standard function of the rapidly evolving mobile phone and will incorporate public transport and pedestrian navigation functions for supporting the total movement of people. In-vehicle terminals will also incorporate wireless communication, sharing various types of information between vehicles, and so on. The driver will be able to have safer and more satisfying driving experiences on demand.

On the other hand, the driver will be provided with tremendous amounts of information, including the entertainment functions, while driving the vehicle. So the in-vehicle system will need to allow through filtering only the necessary information, just like an internet search engine selects the most useful information from the oceans of internet data, and then provides the information with easy-to-use interactive interfaces. Car-navigation-system-based autonomous driving functions may also support safer driving so that the driver can manage a large load of information in some situations.

RELATED ARTICLES

Intelligent Transport Systems: Overview and Structure (History, Applications, and Architectures)
Technologies—Positioning: GNSS
Technologies—Positioning: Optical positioning
Tracking and Navigation for Goods and People
Road Traffic and Travel Information (RTTI)

REFERENCES

Hamada, T. (2002) *Kiwi Format and Telematics –Kiwi-W Consortium*, http://www.kiwi-w.org/documents/kiwi_and_telematics_eng.ppt (accessed 15 January 2014)

FURTHER READING

Kato, M. (2010) *Zukai Car Electronics,* Book 1, Nikkei Business Publications, Inc., Tokyo.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto202
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Series Hybrid Electric Vehicles (SHEVs)

Joeri Van Mierlo and Omar Hegazy

Vrije Universiteit Brussel (VUB), Brussels, Belgium

1 Introduction 1
2 Concept of Series Hybrid Electric Vehicles 2
3 Series Hybrid Electric Drivetrains 3
4 Hybridization Rate of Series Hybrid Electric Vehicles 6
5 Operating Modes 7
6 Energy/Power Management Control 8
7 Conclusion 11
List of Abbreviations 11
References 12
Further Reading 12

1 INTRODUCTION

Electric vehicles (EVs) can be an alternative to the internal combustion engine vehicles (ICEVs). EVs are powered by electric batteries, which need to be recharged with electricity from the grid. Furthermore, the EVs can provide an ideal solution to reduce the environmental impact of transports and reduce the energy dependency because they have low energy consumption and zero local emissions. In other words, battery electric vehicles (BEVs) are zero-emission vehicles (ZEVs) (Chan, Bouscayrol, and Chen, 2010; Van Mierlo and Maggetto, 2001). However, the BEVs still have some challenges, which need to be solved. These challenges are limited driving range, long charging time, battery lifetime, and high initial cost. Therefore, the concept of the series hybrid electric vehicles (SHEVs) can be a viable solution to overcome the disadvantages of the BEVs.

SHEVs comprise two or more energy sources to drive the vehicle, normally one is the main source [such as internal combustion engine (ICE) or fuel cell (FC)], and another is an auxiliary source [such as battery, supercapacitor (SC), or flywheel]. These energy sources can be arranged in different topologies and can cooperate in different ways. Recently, most hybrid vehicles in automotive market are hybrid electric vehicles (HEVs) with an ICE or FC and at least one electrical machine. The commercialization of HEVs has mainly been possible due to the advances in the battery packs, power electronics interfaces (PEIs), and control strategies (Emadi, Lee, and Rajashekara, 2008; Emadi, Williamson, and Khaligh, 2006; Hegazy, Van Mierlo, and Lataire, 2011c).

In the recent years, many research studies in the control strategy (CS) of HEVs have been proposed (Salmasi, 2007; Gao *et al*., 2009; Bayindir, Gozukucuk, and Teke, 2011; Shen, Shan, and Gao, 2011; Hegazy, Van Mierlo, and Lataire, 2011c; Chan, Bouscayrol, and Chen, 2010; Van Mierlo and Maggetto, 2001). As reported in these studies, the main objectives of the power CS are to improve fuel economy, reduce the exhaust emissions, and maintain the battery packs in their desired state of charge (SoC) while ensuring seamless operation of the powertrain. Therefore, the major challenge for the development of SHEVs is the CS of multiple energy sources and PEIs. This necessitates the utilization of an appropriate CS to achieve high performance SHEVs and relatively many references can be found (Salmasi, 2007; Gao *et al*., 2009; Bayindir, Gozukucuk, and Teke, 2011; Shen, Shan, and Gao, 2011). Furthermore, the development of the PEIs and the electrical machines has been investigated in the literature (Hegazy, Van Mierlo, and Lataire, 2011a; Van Mierlo and Maggetto, 2001; Emadi, Williamson, and Khaligh, 2006; Emadi, Lee,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

and Rajashekara, 2008). In other words, the proper selection of the power electronics technologies (such as DC/DC converters, inverters, and chargers) is essential for the successful development of efficient and high performance SHEVs. In addition, the electric machines are another key component required for the emergence and acceptance of high performance SHEVs.

This chapter presents a comprehensive review of the SHEVs to give a clear picture on their powertrain and power management strategy for the next-generation SHEVs. It should be pointed out that the SHEVs have significant potential to not only improve fuel economy and the performance of the vehicle but also reduce the emissions.

2 CONCEPT OF SERIES HYBRID ELECTRIC VEHICLES

Compared to BEVs, SHEVs are potentially capable of extending the driving range and battery lifetime and reducing the battery size. The SHEVs comprise an ICE/generator, electric motor (EM), single or multiple energy storage systems (ESS), power electronic converters, and controllers. In a SHEV configuration, the tractive effort (or traction power) comes only via an electric drive, which obtains electricity from either the ICE/generator (called *GENSET*) directly or from the battery packs with/without power electronics converter. In other powertrains, an auxiliary power unit (APU) is also used as an alternative to GENSET.

In addition, the SHEV has the ability to operate the GENSET on high power when the power demand from the driver is low, which implies a high efficiency for the powertrain. The extra power generated can be stored in the ESS (such as batteries) for future use. Furthermore, the overall efficiency can be significantly improved by operating the electrical machine as a generator during deceleration of the vehicle (called *regenerative braking mode*). Therefore, the SHEVs have the possibility to recover this energy during regenerative braking mode instead of losing the kinetic energy as heat in mechanical brakes as in conventional vehicles (CVs). On the other hand, if the power demand of the EM is higher than the output power of the GENSET, the required power is supplied from the battery packs (called *hybrid traction mode*). Figure 1 illustrates the schematic diagram of the concept of the series hybrid configuration.

The SHEVs have many advantages compared to the CVs, which are represented as follows (Chan, 2007; Chan, Bouscayrol, and Chen, 2010; Bayindir, Gozukucuk, and Teke, 2011):

1. There is a mechanical decoupling of the ICE and the wheels; thus, the GENSET can be located anywhere. Moreover, it is possible to operate the ICE very close to maximum efficiency.
2. SHEVs are relatively the most efficient during stop-and-go city driving.
3. The ICE can operate in a narrow revolutions per minute range (its most efficient range), even when the vehicle changes the speed.
4. SHEVs are easy to control; thus, ICE can operate at maximum efficient point by controlling the output power of the battery to satisfy the required power of vehicle.

However, the SHEVs have some drawbacks, which are given as follows (Chan, 2007; Chan, Bouscayrol, and Chen, 2010; Bayindir, Gozukucuk, and Teke, 2011):

1. The GENSET and the EM must be large enough to satisfy the performance of the vehicle. Consequently, the total weight, cost, and volume of the powertrain can be excessive compared to BEV powertrain.
2. During the highway driving cycle, the overall efficiency is less optimal because of the many energy conversions

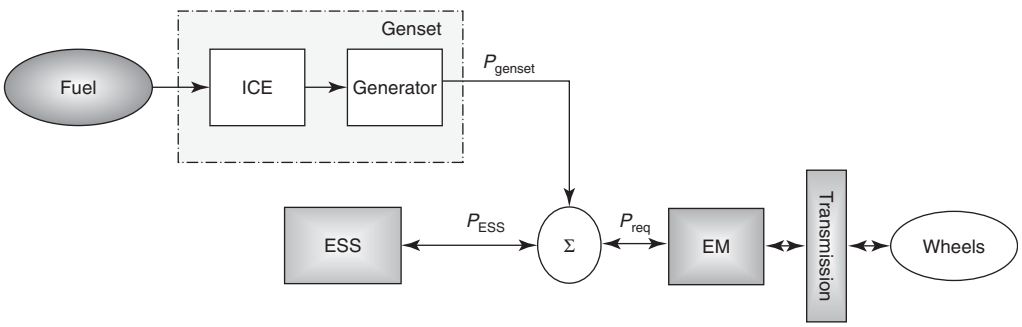


Figure 1. The schematic diagram of SHEV.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

occurring in the different series-connected components of the propulsion system.

3 SERIES HYBRID ELECTRIC DRIVETRAINS

3.1 Powertrain based on ICE

There are several possible configurations of SHEVs. For example, as shown in Figure 2a, the ICE drives a three-phase electric generator (called EM_1) and a three-phase rectifier (named PEI_1) instead of directly driving the wheels, whereas a high-voltage battery system (HV-BS) is directly connected to the DC-bus. The EM (called EM_2) is the only means to provide the power to the wheels. This configuration is the simplest and is of lowest cost. However, this configuration comprises a heavy ESS (i.e., battery) because a large number of battery cells are connected in series to reach the voltage of the high voltage DC-bus (Figure 2a). It is important to point out that this configuration has a low regenerative braking capability, especially at low speed. At low speed, the voltage generated by the EM is less than that generated by the DC-bus, where the DC-bus generates battery voltage. For this reason, the regenerative braking capability is limited at low speed. To improve the performance of the powertrain during regenerative braking, the low-voltage battery system (LV-BS) is connected to the DC-bus through bidirectional DC/DC converter (*buck/boost*), as shown in Figure 2b. In this configuration, the low-voltage battery has the ability to recover more energy from the DC-bus during regenerative braking. To achieve this operating mode, the bidirectional DC/DC converter works as buck converter to transfer the power from DC-bus to battery system.

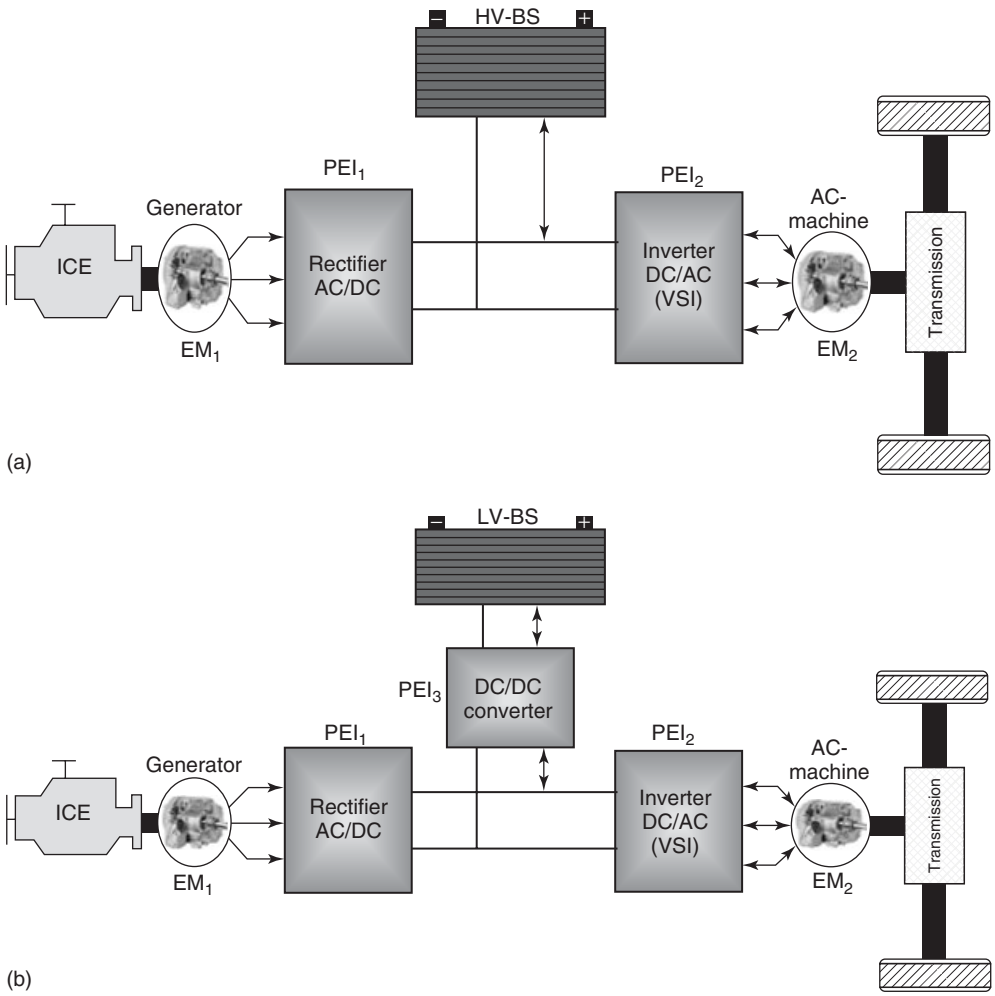


Figure 2. Different powertrain configurations based on ICE. (a) High-voltage battery (configuration I) and (b) low-voltage battery with bidirectional DC/DC converter (configuration II).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

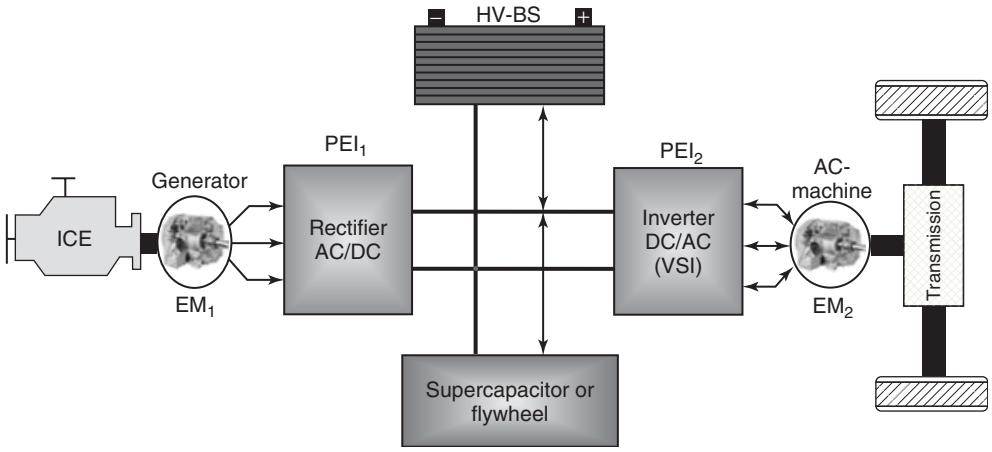


Figure 3. SHEVs with peak power source.

In addition, SHEVs can be assisted by SCs (or flywheels), which can improve the efficiency by minimizing the losses in the battery. They deliver peak power during acceleration and recuperate the regenerative energy during braking (Van Mierlo and Maggetto, 2001; Bayindir, Gozukucuk, and Teke, 2011). Therefore, the SCs are kept charged at low speed and almost empty at maximum speed. In this configuration, the SCs not only improve the performance of the powertrain but also extend the lifetime of the battery. Figure 3 shows the SHEVs with peak power source. In addition, for low-voltage ESS (batteries and SCs), the DC/DC converter should be placed between the ESS and the DC-bus, which may lead to features such as light ESS, less cost, and high regenerative braking capability, especially at low speed.

3.2 Powertrain based on FC

A special SHEV has an FC hybrid structure. In this structure, the engine/generator group is replaced by an FC system producing electric energy starting from stored hydrogen or from a fuel tank feeding a reformer to produce hydrogen. The FCs have emerged as one of the most promising candidates for fuel-efficient and emission-free vehicle power generation. In particular, the proton exchange membrane fuel cells (PEMFCs) have received much attention for automotive applications because of their solid electrolyte, low operating temperature, and high power density. However, the high cost and its slow dynamics are the major challenges for the commercialization of fuel cell electric hybrid vehicles (FCHEVs) (Feroldi, Serra, and Riera, 2009; Van Mierlo, Van den Bossche, and Maggetto, 2004; Hegazy, Van Mierlo, and Lataire, 2011b; Van Mierlo and Maggetto, 2001; Chan, 2007; Chan, Bouscayrol, and Chen, 2010; Feroldi, Serra, and Riera, 2009). To improve the transient performance of FCs and recover energy through regenerative braking, as mentioned in Section 1, FC systems should be typically hybridized with ESS, such as batteries or SCs to form the FCHEV powertrain. In other words, it is important to hybridize an FCHEV with an ESS for the following reasons:

1. To provide a fast dynamic response for the powertrain.
2. To enable recovery of regenerative braking energy, and to improve the overall efficiency of the vehicle.
3. To reduce the size of the FC, where an FC of lower rated power can be significantly lighter and cheaper than an FC of higher rated power.
4. Improvement in the CS allows the FC to operate more often in the high efficiency region.

As illustrated in Figure 4, one or more high-power DC/DC converters are commonly required within an FCHEV powertrain. The major challenge of multiple DC/DC converters is the CS, which would be complicated. In other words, for these applications, a high-power DC/DC converter is a key component that interfaces the FC or ESS to the DC-bus in the powertrain of the SHEVs (Emadi, Williamson, and Khaligh, 2006; Emadi, Lee, and Rajashekara, 2008; Hegazy, Van Mierlo, and Lataire, 2011a). Therefore, the design of high-power DC/DC converters and their controller would play an important role in the future powertrain system to control the power regulation, particularly for a common DC-bus. Several topologies of DC/DC converters based on the count of their components, advantages, and disadvantages are discussed and compared in the literature (Solero *et al*., 2005; Hegazy, Van Mierlo, and Lataire, 2011a, 2011c). Moreover, for high

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

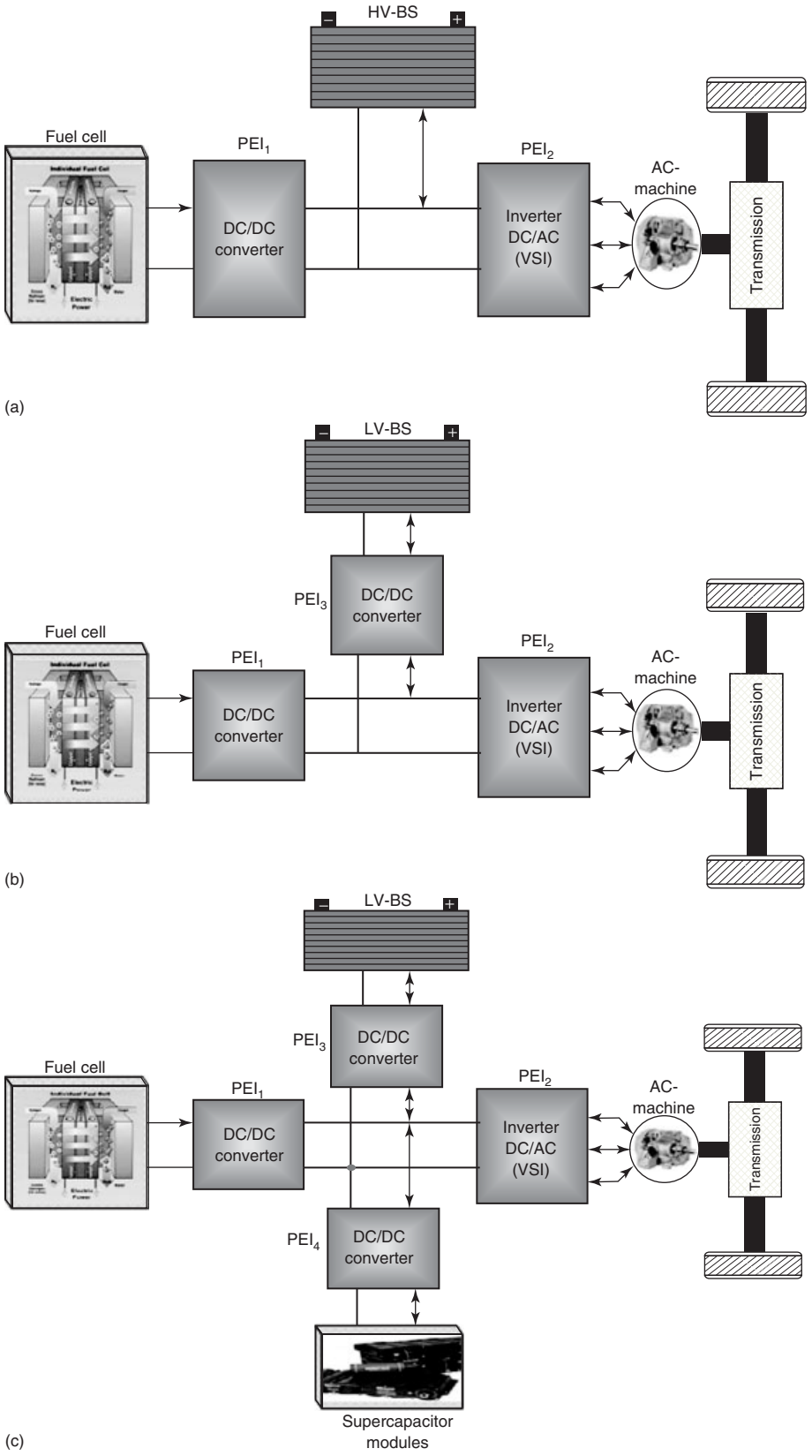


Figure 4. Different powertrain configurations based on FC.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Major characteristics and features of the SHEVs.

Item	SHEV-based ICE	SHEV-based FC
Propulsion	• Electric motor drives • ICE/generator	• Electric motor drives • FC
Energy storage system	• Battery • Supercapacitor • Flywheel	• Battery • Supercapacitor • Flywheel
Energy sources	• Gasoline stations	• Hydrogen
Characteristics	• Low emissions • Extended driving range • High fuel economy compared to ICEVs • Commercially available	• Zero emission • High efficiency • Satisfied driving range • High cost • Under development
Major issues	• Managing multiple energy sources (control strategy) • Powertrain efficiency • Dependent on the driving cycle • Battery sizing and management	• Fuel cell cost • Hydrogen production • Control strategy • Powertrain sizing

power applications, multiphase interleaved converters have been proposed for use in HEV applications to improve their efficiency and reliability. As is clear from previous studies, the selection of PEI is based on some significant factors such as lower cost, higher efficiency, electrical isolation, free ripple, and reliable operation.

In addition, the fuel consumption, driving range, emissions, and powertrain efficiency are the major issues when comparing SHEV based on ICE and SHEV based on FC. The SHEV based on FC can provide zero emissions, high fuel economy, high efficiency, and satisfied driving range compared to SHEV based on ICE. However, SHEV based on FC has a high cost, and it is still under development. Table 1 summarizes the major characteristics of the SHEV based on ICE and the SHEV based on FC (Chan, Bouscayrol, and Chen, 2010; Bayindir, Gozukucuk, and Teke, 2011).

4 HYBRIDIZATION RATE OF SERIES HYBRID ELECTRIC VEHICLES

There exist three definitions of the hybridization rate for the SHEVs. These definitions are (i) electric hybridization rate (EHR), (ii) combustion hybridization rate (CHR), and (iii) rate of hybridization (RoH) (Maggetto and Van Mierlo, 2000; Van Mierlo and Maggetto, 2001).

4.1 Electric hybridization rate (EHR)

The EHR gives an indication of the SHEV performance. The EHR is the ratio between the electric power and the total traction power as seen in Equations 1 and 2. Moreover, the EHR is expressed in percentage. The high percentage illustrates that the vehicle tends to a pure BEV.

$$\mathrm{EHR} = \frac{\text{electric power}}{\text{traction power}} \tag{1}$$

In SHEVs, the traction power corresponds with the nominal EM power as the EM is the only means to provide the power to the wheels. Consequently, the EHR for SHEV is given by

$$\mathrm{EHR} = \frac{\text{battery power}\,(P_{\mathrm{bat}})}{\text{electric motor power}\,(P_{\mathrm{mot}})} = \frac{P_{\mathrm{mot}} - P_{\mathrm{gen}}}{P_{\mathrm{bat}} + P_{\mathrm{gen}}} \tag{2}$$

where P_{gen} is the GENSET power.

4.2 Combustion hybridization rate (CHR)

To obtain an idea of the relative contribution of the ICE, the CHR is to be defined as the ratio between the thermal power and the total traction power, as indicated in Equations 3 and 4. Furthermore, CHR is expressed in percentage, and a high percentage indicates that the vehicle tends to be a pure thermal vehicle.

$$\mathrm{CHR} = \frac{\text{thermal power}}{\text{traction power}} \tag{3}$$

For SHEV, the thermal power is the rated GENSET power (P_{gen}). Therefore, the CHR is defined by

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

$$\mathrm{CHR} = \frac{P_{\mathrm{gen}}}{P_{\mathrm{mot}}} = \frac{P_{\mathrm{mot}} - P_{\mathrm{bat}}}{P_{\mathrm{bat}} + P_{\mathrm{gen}}} \quad (4)$$

It should be pointed out that the sum of EHR and CHR is always equal to 1 and each concept is just an indication of whether the vehicle tends toward a pure BEV or a pure thermal vehicle.

4.3 Rate of hybridization (RoH)

To investigate the relative contribution of each energy source, the rate of hybridization (RoH) has to be defined. If both energy sources have an equal contribution to the traction power, the RoH is defined as equal to 1 (for maximum value). If the contribution of the ICE system is higher than the electric battery system, the RoH ($\mathrm{RoH_{th}}$) is the ratio of the battery power (P_{bat}) to the thermal power (P_{gen}) and vice versa of the $\mathrm{RoH_{el}}$ (Equations 5 and 6).

$$\text{If } P_{\mathrm{bat}} < P_{\mathrm{gen}}, \quad \text{then} \quad \mathrm{RoH_{th}} = \frac{P_{\mathrm{bat}}}{P_{\mathrm{gen}}} \quad (5)$$

$$\text{If } P_{\mathrm{gen}} < P_{\mathrm{bat}}, \quad \text{then} \quad \mathrm{RoH_{el}} = \frac{P_{\mathrm{gen}}}{P_{\mathrm{bat}}} \quad (6)$$

Figure 5 illustrates the hybridization rate of SHEVs.

5 OPERATING MODES

In SHEVs, the power flow control can be illustrated by seven operating modes, as shown in Figure 6. Seven possible operating modes exist for the SHEVs:

1. *Mode 1* (called *pure GENSET drive*): the GENSET only supplies the required power.
2. Mode 2 (named *zero-emission driving*): in this mode, the GENSET is turned off and the vehicle is only supplied by the batteries. This mode can be applied in urban driving.
3. *Mode 3* (called *hybrid drive*): the traction power is supplied by the GENSET and the batteries. This mode is used when the power requirement is high.
4. *Mode 4* (named *engine traction and battery charging*): the GENSET provides the energy needed for charging the batteries and the propulsion of the vehicle.
5. *Mode 5* (called *battery charging and no drive*): in this mode, the GENSET charges the batteries, whereas the traction motor is not used.
6. *Mode 6* (named *regenerative braking*): during braking or deceleration, the GENSET is turned off and the traction motor is operated as generator to charge the battery.
7. *Mode 7* (*hybrid battery charging*): both the GENSET and the traction motor operate as generator to charge the batteries.

Furthermore, these operating modes of SHEVs are summarized in Table 2.

As a result, these various power flow paths can provide a tremendous flexibility of vehicle operation compared to CVs. With proper CS, applying a specific mode for a special operating condition can significantly improve the overall performance, efficiency, and emissions. In the following section, a general overview of the control strategies is presented.

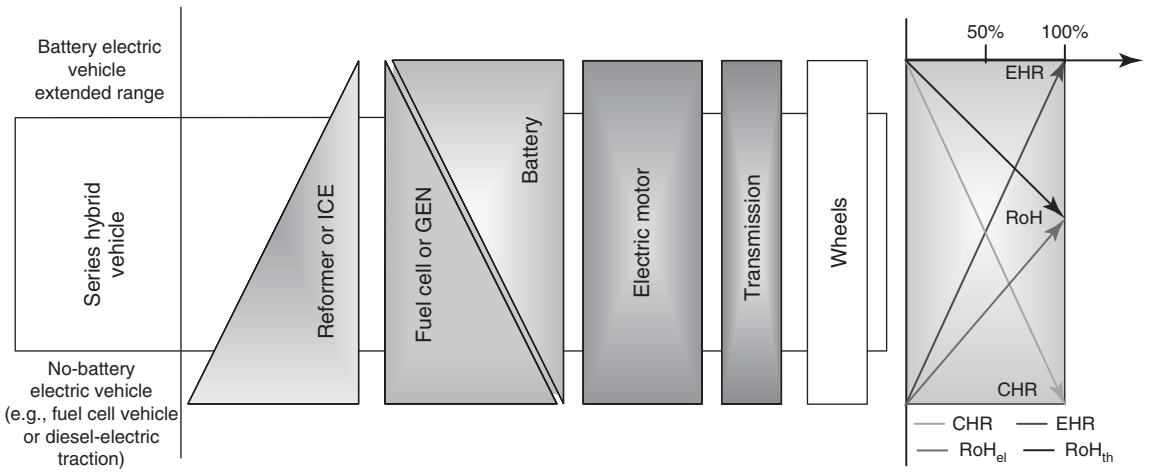


Figure 5. Hybridization rate of SHEV.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

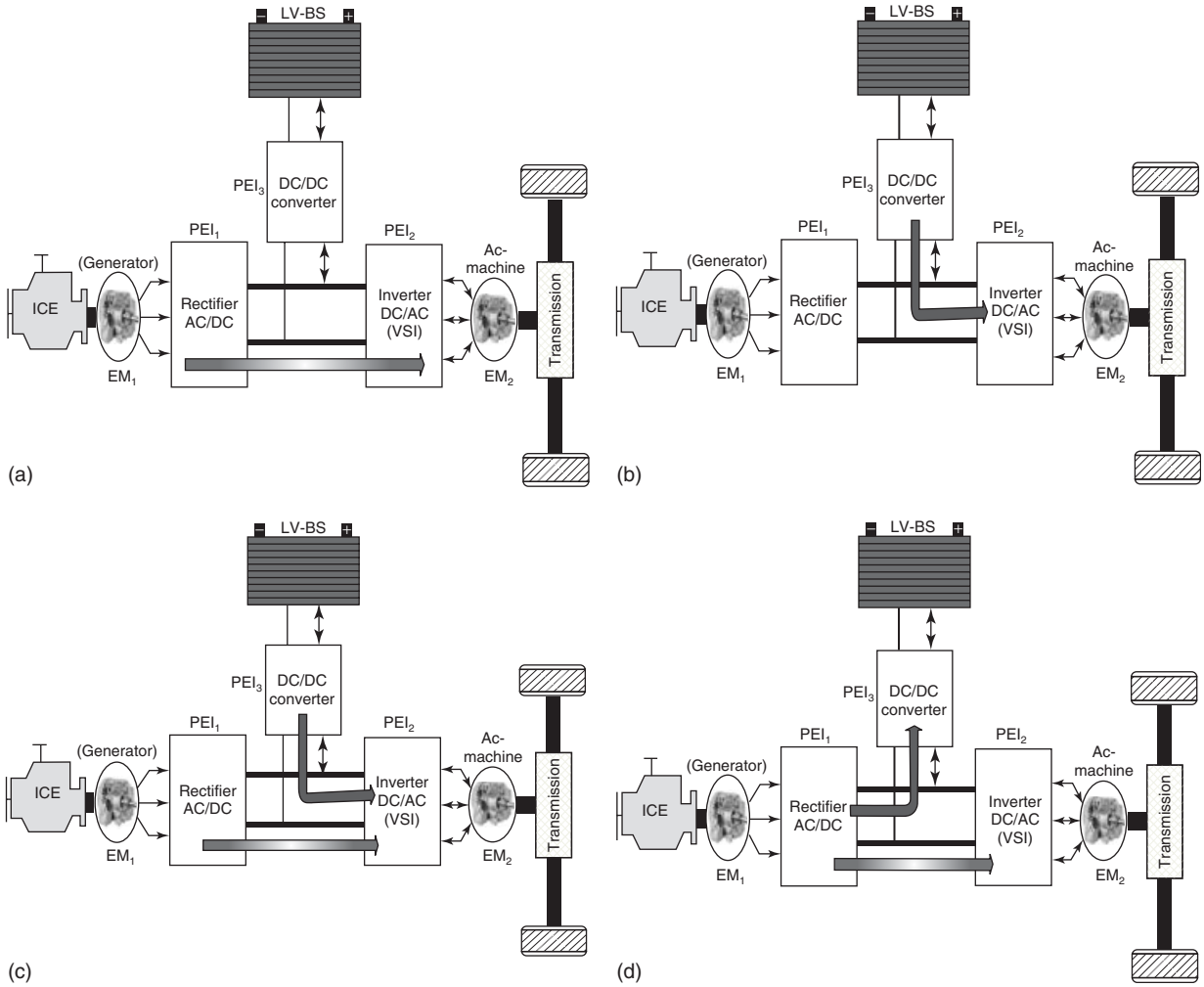


Figure 6. SHEV operating modes. (a) Mode 1: ICE-EM_1-PEI_1-PEI_2-EM_2. (b) Mode 2: BS-PEI_3-PEI_2-EM_2. (c) Mode 3: ICE-EM_1-PEI_1-PEI_2-EM_2 and BS-PEI_3-PEI_2-EM_2. (d) Mode 4: ICE-EM_1-PEI_1-PEI_2-EM_2 and ICE-EM_1-PEI_1-PEI_3-BS. (e) Mode 5: ICE-EM_1-PEI_1-PEI_3-BS. (f) Mode 6: EM_2-PEI_2-PEI_3-BS. (g) Mode 7: ICE-EM_1-PEI_1-PEI_3-BS and EM_2-PEI_2-PEI_3-BS.

Table 2. Operating modes of SHEVs.

	Mode	Function/operation
1	Mode 1	Pure GENSET drive
2	Mode 2	Pure electric drive or zero-emission driving
3	Mode 3	Hybrid drive
4	Mode 4	Engine traction and battery charging
5	Mode 5	Battery charging and no drive
6	Mode 6	Regenerative braking
7	Mode 7	Hybrid battery charging

6 ENERGY/POWER MANAGEMENT CONTROL

Owing to the multiple energy sources and PEIs, there exist several powertrain configurations and different control strategies to control the energy source, resulting in different operating modes. The aim of the energy/power management strategies is to satisfy a number of objectives for HEV. The main objectives are represented as follows:

1. Low fuel consumption
2. Low emissions
3. Low noise
4. Good drivability
5. Minimum propulsion system cost
6. Acceptable performance.

The concept of sharing the requested power between the ICE and ESS (i.e., battery packs) for traction mode during vehicle operation is referred to as *vehicle supervisory control* or *energy/power management control*. Therefore, there are a number of energy/power management strategies

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

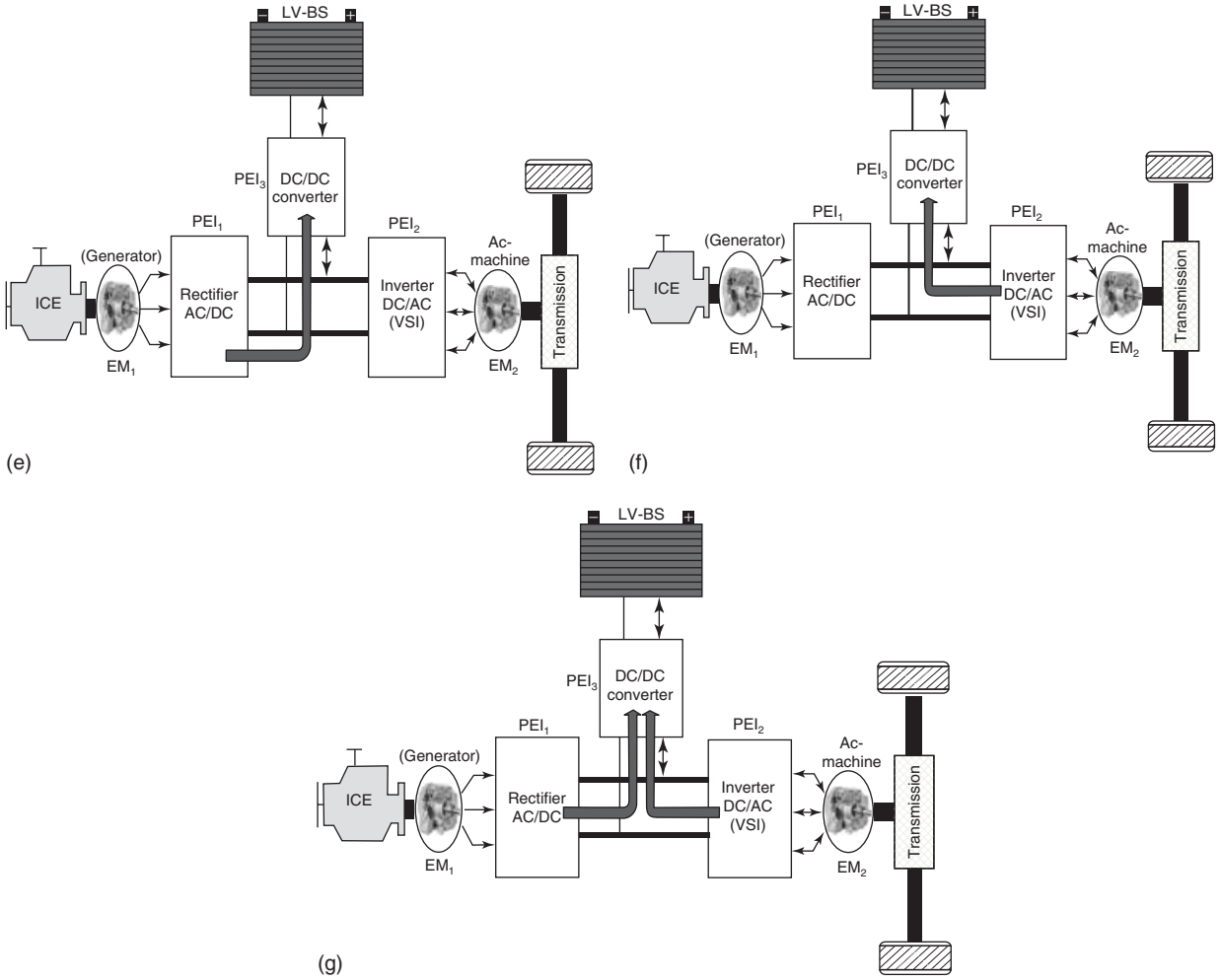


Figure 6. (*Continued*)

proposed in the literature of hybrid vehicles to minimize the fuel consumption and reduce the emissions (Salmasi, 2007; Gao *et al*., 2009; Bayindir, Gozukucuk, and Teke, 2011; Shen, Shan, and Gao, 2011; Barrero *et al*. 2009). As illustrated in Figure 7, the energy/power management control strategies can be classified into two main categories: rule-based control strategy (RBCS) and optimization-based control strategy (OBCS) (Salmasi, 2007; Hegazy, Van Mierlo, and Lataire, 2011b). These control strategies will be presented in more detail in Energy Management Systems of HEVs.

It is clear that different control strategies are possible to control the power flow in a hybrid powertrain, as shown in Figure 7. Majority of the proposed solutions for the power CS can be classified into two types: (i) RBCS and (ii) OBCS. RBCSs consist of deterministic and fuzzy logic-rule-based methods, while OBCSs are typically utilized global optimization and real-time optimization when determining the CS. The following subsections provide an overview of the commonly used control strategies.

6.1 Rule-based control strategies (RBCSs)

The main aspect of the RBCSs is their effectiveness in real-time supervisory control of power flow control in a hybrid powertrain. The rules are determined based on heuristics, human expertise, and mathematical models without a prior knowledge of a predefined driving cycle. These strategies can be classified into deterministic and fuzzy rule-based methods, as shown in Figure 7 (Salmasi, 2007; Gao *et al*., 2009; Bayindir, Gozukucuk, and Teke, 2011; Shen, Shan, and Gao, 2011; Barrero *et al*. 2009). The main concept of RBCSs is commonly based on "IF-THEN" type of control rules, which determine, for example,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

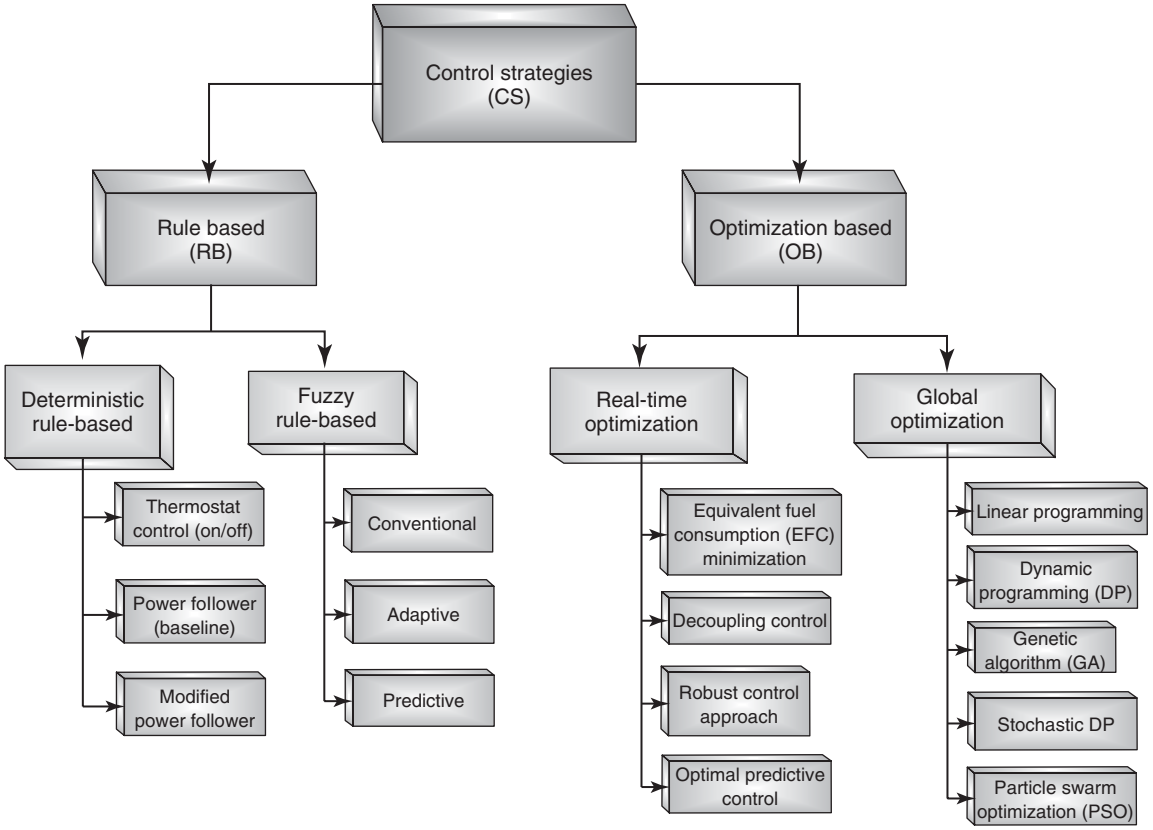


Figure 7. Classification of the hybrid electric vehicle control strategies.

when to shut down the ICE or the amount of electric charging/discharging powers. These control strategies perform load leveling or balancing within vehicle operation. The idea behind load leveling is to move the ICE operation point as close as possible to optimal region of fuel economy, efficiency, and emissions at particular ICE speed. However, for this type of system, a high fuel economy can be found at lower ICE torques, while the efficiency may not be the best value. Thus, small acceleration demand can result in higher fuel economy. The difference between the power demand and ICE-generated power will be compensated by the EES or utilized to charge the EES.

6.1.1 Deterministic rule-based control strategies (DRBCSs)

To realize these control strategies, the rules are designed with the aid of power flow in a hybrid powertrain, efficiency/fuel maps of ICE, and driving experience. Furthermore, these rules are generally implemented via lookup tables in order to split the requested power between the ICE and ESS.

6.1.1.1 Thermostat (ON/OFF) control strategy. Thermostat (ON/OFF) CS is a popular strategy for energy management in current hybrid powertrains (Barrero *et al*., 2009). This strategy attempts to obtain the maximum efficiency of the ICE. For this purpose, the ICE runs at the most efficient point of the efficiency map, providing a constant power. The difference between this set point and the power needed is absorbed by the ESS. In this simple CS, the battery SoC is always maintained between predefined high and low levels by simply turning on or off the ICE. In addition, the benefits of this strategy are an optimal ICE efficiency and almost no inertial losses. The drawbacks are that a larger ESS is needed so that the engine's ON-and-OFF cycles are not too short. In addition, high losses are produced on the ESS as the power flow is very high. The operation of the thermostat (ON/OFF) CS is illustrated in Figure 8.

6.2 Optimization-based control strategies (OBCSs)

These control strategies are used to maximize the fuel economy and the efficiency of the powertrain while

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

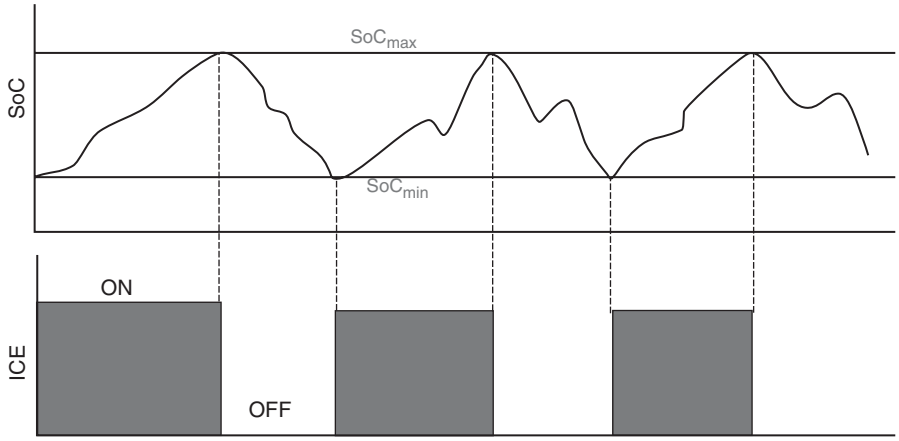


Figure 8. Operation of thermostat (ON/OFF) control strategy.

minimizing the losses in powertrain. The optimal reference torques/currents for the power converters and the optimal hybridization degree are calculated by minimization of a cost function generally representing the fuel consumption or emissions. If this optimization is implemented over a given driving cycle, a global optimum solution can be found. Indeed, the global optimal solution is inherently noncausal, in that it provides a minimum fuel consumption using knowledge of future and past power demands. However, these approaches cannot be directly utilized for real-time energy management and they are not easily implementable for practical applications. The classification of the OBCS is depicted in Figure 7.

6.2.1 Global optimization

There exist several presented solutions to achieve performance targets by optimization of a cost function representing efficiency, fuel consumption, and emissions over a drive cycle, yielding to global optimal operating points. These techniques are not applicable because of their preview nature and computational complexity (Salmasi, 2007; Bayindir, Gozukucuk, and Teke, 2011; Shen, Shan, and Gao, 2011). Some control algorithms for global optimization, based on a priori knowledge of a scheduled driving cycle, have been proposed to achieve fairly fuel economy with minimum cost. A commonly used technique for determining the globally optimal energy management in hybrid powertrain is dynamic programming (DP) technique. The DP technique yields the minimum but has the disadvantage of its high computational load and the need for high memory storage capacity. It is a memory- and computational-intensive algorithm, especially when dealing with large-scale optimization problems. Recently, the particle swarm optimization (PSO) algorithm is used to achieve the global optimization for hybrid powertrain. PSO algorithm is a member of the wide category of swarm intelligence (SI) methods (Hegazy, Van Mierlo, and Lataire, 2011b). Compared to genetic algorithms (GAs) and DP, the PSO is simpler and requires less memory and less computational time. Moreover, the PSO algorithm has no evolution operators such as crossover and mutation. In addition, the PSO can reliably solve this complex constrained optimization problem easily and quickly.

7 CONCLUSION

This chapter has presented a comprehensive review of the SHEVs to give a clear overview on their powertrain and energy management control strategies for the next-generation SHEVs. Moreover, the benefits and the drawbacks of the SHEVs are provided. In this chapter, the different operating modes of the SHEVs are investigated. Furthermore, a comparative study between SHEV based on ICE and SHEV based on FC has been presented in this study. It is noted that the SHEV based on FC is a very promising topology for improving the fuel economy and achieving zero emissions. Furthermore, it is important to point out that the SHEVs can have predominance in many applications, such as city vehicles and urban buses.

LIST OF ABBREVIATIONS

APU	auxiliary power unit
BEVs	battery electric vehicles
CHR	combustion hybridization rate
CS	control strategy
CVs	conventional vehicles

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

EHR	electric hybridization rate
EM	electric motor
ESS	energy storage systems
EVs	electric vehicles
FC	fuel cell
HEVs	hybrid electric vehicles
HV-BS	high voltage battery system
ICE	internal combustion engine
ICEVs	internal combustion engine vehicles
LV-BS	low voltage battery system
OB	optimization based
PEIs	power electronics interfaces
RB	rule based
RoH	rate of hybridization
SCs	supercapacitors
SHEVs	series hybrid electric vehicles
SoC	state of charge
ZEVs	zero-emission vehicles

REFERENCES

Barrero, R., Tackoen, X., Van Mierlo, J., and Coosemans, T. (2009) Hybrid buses: defining the power flow management strategy and energy storage system needs. *WEVA Journal*, **3**, 1–12.

Bayindir, K.C., Gozukucuk, M.A., and Teke, A. (2011) A comprehensive overview of hybrid electric vehicle: powertrain configurations, powertrain control techniques and electronic control units. *Energy Conversion and Management*, **52**, 1305–1313.

Chan, C.C. (2007) The state of the art of electric and hybrid, and fuel cell vehicles. *Proceedings of the IEEE*, Special issue on Electric, Hybrid and Fuel Cell Vehicles, **95** (4)

Chan, C.C., Bouscayrol, A., and Chen, K. (2010) Electric, hybrid, and fuel-cell vehicles: architectures and modeling. *IEEE Transactions on Vehicular Technology*, **59** (2), 589–598.

Emadi, A., Lee, Y.J., and Rajashekara, K. (2008) Power electronics and motor drives in electric, hybrid electric, and plug-in hybrid electric vehicles. *IEEE Transactions on Industrial Electronics*, **55** (6), 2237–2245.

Emadi, A., Williamson, S.S., and Khaligh, A. (2006) Power electronics intensive solutions for advanced electric, hybrid electric, and fuel cell vehicular power systems. *IEEE Transactions on Power Electronics*, **21** (3), 567–577.

Feroldi, D., Serra, M., and Riera, J. (2009) Design and analysis of fuel-cell hybrid systems oriented to automotive applications. *IEEE Transactions on Vehicular Technology*, **58** (9), 4720–4729.

Gao, J., Sun, F., He, H., *et al*. (2009) A Comparative Study of Supervisory Control Strategies for a Series Hybrid Electric Vehicle. *Power and Energy Engineering Conference, APPEEC 2009*, Asia-Pacific, 27–31 March 2009.

Hegazy, O., Van Mierlo, J., and Lataire, P. (2011a) Analysis, control and implementation of a high-power interleaved boost converter for fuel cell hybrid electric vehicle. *International Review of Electrical Engineering*, **6** (4), 1739–1747.

Hegazy, O., Van Mierlo, J., and Lataire, P. (2011b) Design optimization and optimal power control of fuel cell hybrid electric vehicles based on swarm intelligence. *International Review of Electrical Engineering*, **6** (4), 1727–1738.

Hegazy, O., Van Mierlo, J., and Lataire P. (2011c) Analysis, Control and Comparison of DC/DC Boost Converter Topologies for Fuel Cell Hybrid Electric Vehicle Applications. *14th European Conference on Power Electronics and Applications, IEEE EPE 2011*, 30 August–1 September 2011, Birmingham, UK.

Maggetto, G. and Van Mierlo, J. (2000) Electric and Electric Hybrid Vehicle Technology: A Survey. *Proceedings of IEE Seminar on Electric, Hybrid and Fuel Cell Vehicles*, pp. 1/1–1/11.

Salmasi, F.R. (2007) Control strategies for hybrid electric vehicles: evolution, classification, comparison, and future trends. *IEEE Transactions on Vehicular Technology*, **56** (5), 2393–2403.

Shen, C., Shan, P., and Gao, T. (2011) A comprehensive overview of hybrid electric vehicles. *International Journal of Vehicular Technology*, Article ID 571683, 7 pp

Solero, L., Lidozzi, A., and Pomilio, J.A. (2005) Design of multiple-input power converter for hybrid vehicles. *IEEE Transactions on Power Electronics*, **20** (5), 1007–1015.

Van Mierlo, J., Van den Bossche, P., and Maggetto, G. (2004) Models of energy sources for EV and HEV: fuel cells, batteries, ultracapacitors, flywheels and engine-generators. *Journal of Power Sources*, **128** (1), 76–89.

Van Mierlo, J. and Maggetto, G. (2001) Vehicle simulation programme: a tool to evaluate hybrid power management strategies based on an innovative algorithm. *Journal of Automobile Engineering*, **215**, 1043–1052.

FURTHER READING

Chau, K.T. and Wong, Y.S. (2002) Overview of power management in hybrid electric vehicles. *Energy Conversion and Management*, **43** (2002), 1953–1968.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto203
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Parallel Hybrid Electric Vehicles (Parallel HEVs)

Joeri Van Mierlo and Omar Hegazy

Vrije Universiteit Brussel (VUB), Brussels, Belgium

1 Introduction 1
2 Concept of Parallel Hybrid Electric Vehicles 2
3 Parallel Hybrid Electric Drivetrains 3
4 Hybridization Rate of the Parallel HEV 5
5 The Operating Modes of the Parallel HEV 6
6 Mechanical Coupling 8
7 Power Flow Control Strategies 8
8 Conclusions 9
List of Abbreviations 9
References 9
Further Reading 10

1 INTRODUCTION

Owing to the concerns about the environmental and energy issues, many research studies have been carried out to enhance the performance of the internal combustion engine vehicles (ICEVs) and launch new-generation vehicles. Furthermore, it is widely accepted that the fuel consumption improvements of ICEV are reaching their limits. To improve the performance of the ICEVs and recover energy through regenerative braking, the vehicle should be hybridized with energy storage systems (ESSs), such as batteries, supercapacitors, or flywheels (Van Mierlo and Maggetto, 2001; Van Mierlo, Van den Bossche, and Maggetto, 2004; Chan, 2007; Chan, Bouscayrol, and Chen, 2010). Nowadays, hybrid electric vehicles (HEVs) have become an interesting alternative to ICEVs because of their ability to reduce the fuel consumption and achieve zero-exhaust emissions. HEVs can significantly provide high performance and flexibility over conventional vehicles.

In general, HEVs consist of an internal combustion engine (ICE) and one or more electric motors (EMs). In other words, HEVs comprise two or more energy sources to drive the vehicle, normally one is the main source (called *ICE*) and another is an ESS (such as battery and supercapacitor). In HEV configuration, the ESS can be connected to DC power bus through a bidirectional power electronics interface (PEI) (called *DC/DC converter*), whereas the EM is connected to DC power bus by means of a motor controller, which is a bidirectional DC/AC inverter. Conceptually, it is inherently an EM that assists the ICE for achieving lower emissions and fuel consumption. Furthermore, the ESS can improve the dynamic performance of the vehicle and recuperate the energy during regenerative braking.

On the basis of different combinations of connecting EM, transmission system, and ICE, HEVs are generally classified into two basic configurations: series hybrid electric vehicles (SHEVs) and parallel hybrid electric vehicles (parallel HEVs). Recently, with the combination of the features of both SHEVs and parallel HEVs, the classification has been extended to three arrangements: series, parallel, and series–parallel or power-split hybrid vehicles (Van Mierlo and Maggetto, 2001; Van Mierlo, Van den Bossche, and Maggetto 2004). As described in the previous chapter (see Series Hybrid Electric Vehicles (SHEVs)), SHEVs can provide a viable solution to extend the range of the battery electric vehicles (BEVs).

This chapter provides an overview of common parallel HEVs and gives a comprehensive review on their powertrains and power management strategies. In this chapter, different transmission systems are presented. Furthermore,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto204
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

the hybridization rate of parallel HEVs is explained in this chapter.

2 CONCEPT OF PARALLEL HYBRID ELECTRIC VEHICLES

Parallel HEVs comprise an ICE and at least one EM, which are both mechanically connected in parallel to drive the wheels. Figure 1 shows the schematic diagram of the concept of the parallel HEV configuration, where the output torques of the ICE and the EM are mechanically connected through a torque coupler or speed coupler. As illustrated in Figure 1, the power relationship can be expressed as follows:

$$P_{\text{req}} = P_{\text{ICE}} + P_{\text{EM}} = P_{\text{ICE}} + \begin{cases} \eta_{\text{EM}} P_{\text{ESS}}, & \text{if } P_{\text{EM}} \geq 0 \text{ (motor mode)} \\ \frac{P_{\text{ESS}}}{\eta_{\text{EM}}}, & \text{if } P_{\text{EM}} < 0 \text{ (generator mode)} \end{cases} \quad (1)$$

where

P_{req} the required power
P_{ICE} the ICE power
P_{EM} the electric motor power
P_{ESS} the ESS power
η_{EM} the electric motor efficiency.

Thus, the overall powertrain efficiency is increased because of low power transmission losses compared to the SHEV powertrain. The parallel HEV powertrain is compact because the generator is not needed and the EM size is small compared to the EM of the SHEV. In addition, the parallel HEV offers more freedom to select the combination of the traction source. For example, the EM can be connected to the powertrain in different ways and can be located at different positions along the transmission (see the subsequent text).

In parallel HEV configuration, both ICE and EM can propel the vehicle, either independently or simultaneously, thanks to the use of clutches. The main advantage of parallel HEV is that the vehicle can recuperate the energy from the regenerative braking instead of losing the kinetic energy as heat in mechanical brakes as in ICEVs. Moreover, the parallel HEV allows to utilize the EM as starter motor, which allows to switch off the ICE when the vehicle is standing still at a red traffic light.

The parallel HEVs have some benefits compared to the ICEVs, which are summarized as follows (Van Mierlo and Maggetto, 2001, 2004):

1. The ICE can be operated in a region (in the torque–speed engine map) where it has a good efficiency. In conventional vehicles, the ICE can sometimes be operated at working conditions where the efficiency drops below 20% (see the dark gray line in Figure 2).
2. Both the ICE and EM are directly connected to the wheels. Consequently, the energy loss may be low, which increases the overall powertrain efficiency. In other words, the parallel HEV has fewer energy conversion stages than the SHEV.
3. The EM allows stop–start functionality: switching off the ICE when the vehicle is standstill.
4. Braking energy can be recuperated into the battery because of the fact that the EM works as a generator during vehicle deceleration.
5. Owing to the absence of the electric generator, the parallel HEV is more compact than the SHEV.

However, the parallel HEVs have some disadvantages, which are represented as follows:

1. The mechanical transmission system is complex.
2. The control of parallel HEV is more complicated than SHEV because of the mechanical coupling.
3. The mechanical coupling of the ICE to the wheels leads to a limited choice of operating points of the ICE in a narrow speed and torque range.

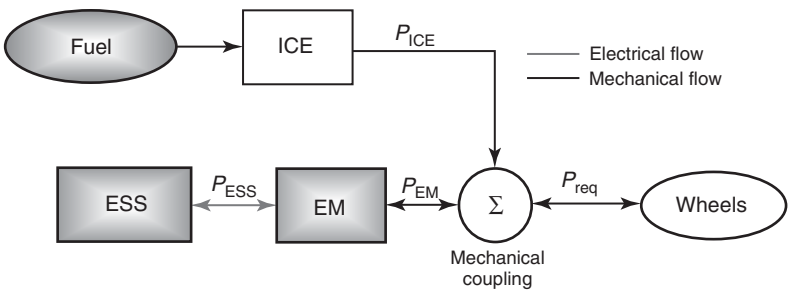


Figure 1. Schematic diagram of parallel HEV.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto204
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

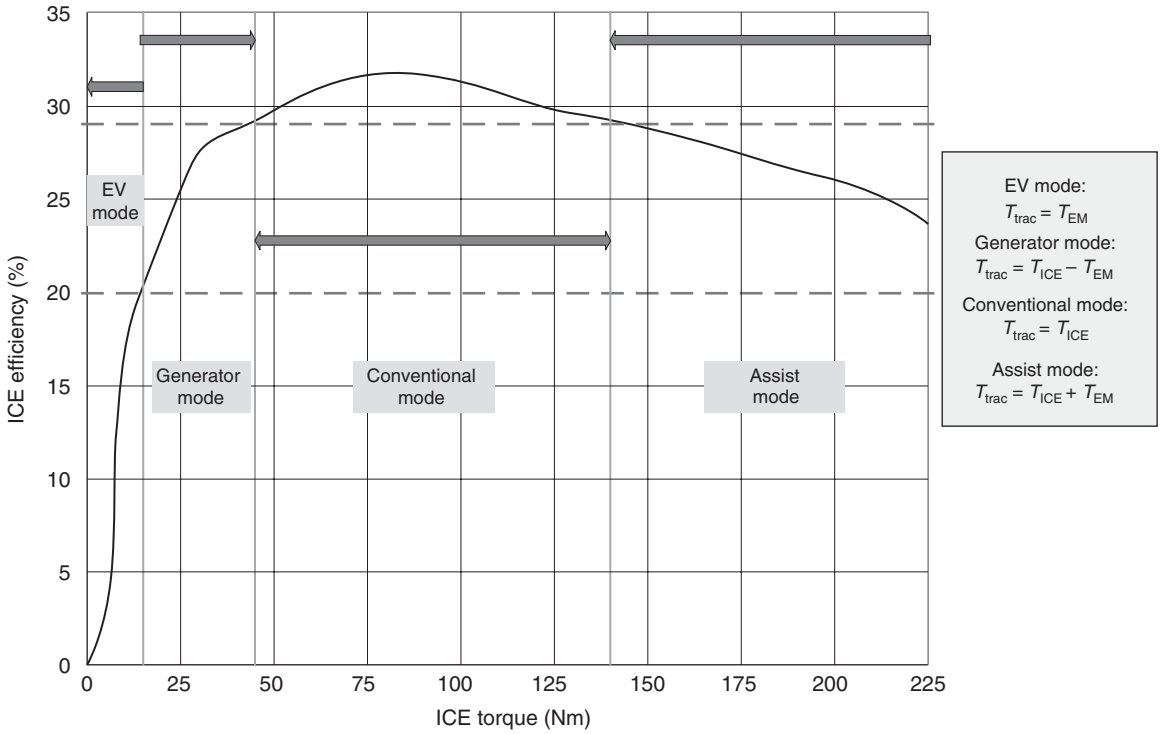


Figure 2. Different operating modes based on ICE torque and ICE efficiency at a given speed. T_{trac}, the traction torque; T_{EM}, the EM torque; T_{ICE}, the ICE torque.

Figure 2 presents the operating modes depending on the ICE torque and ICE efficiency at a given speed (Debal *et al*., 2009). This figure clearly demonstrates the ability of a parallel HEV to optimize the fuel consumption to operate the ICE in a region where it has a high efficiency. This is possible only during the presence of the EM.

Indeed, at low requested torque, the parallel HEV will operate in electric vehicle mode (EV mode). Only the EM will drive the wheels.

At a higher requested driving torque (example shown in Figure 2 between 17 and 40 Nm), the parallel HEV will be operated in generator mode. This means that the EM will function as a generator, demanding a higher power to the ICE than needed for driving. Hence, the ICE is operated at efficiency higher than 20% (in this example). In this generator mode, the battery is charged.

Above 40 Nm, the vehicle will work as a conventional vehicle (conventional mode). Only the ICE drives the wheels.

Above 140 Nm, the parallel HEV will be operated in power assist mode (assist mode). This means that the ICE and EM are driving the wheels. This mode not only achieves higher fuel economy, but also keeps the vehicular emissions low. The gray dashed lines indicate the ICE efficiency at beginning of each operating mode.

3 PARALLEL HYBRID ELECTRIC DRIVETRAINS

There exist different parallel HEV propulsion topologies. The parallel HEV drivetrain configurations comprise a conventional drivetrain (ICEV) with an electrical drivetrain (ED) in parallel. On the basis of the location of the transmission system relative to the ICE and EM, the parallel HEV drivetrains can be further classified as pretransmission (Figure 3) or posttransmission parallel HEV (Figure 4). The stop–start functionality is easier to implement in a pretransmission parallel HEV. Furthermore, the requested EM maximum torque is smaller than in the case of a posttransmission parallel HEV. On the other hand, the posttransmission parallel HEV has the advantage that the power does not need to pass through the transmission (while driving as well as braking), which reduces the losses. The challenge of the posttransmission parallel HEV is the required high starting torque of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto204
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

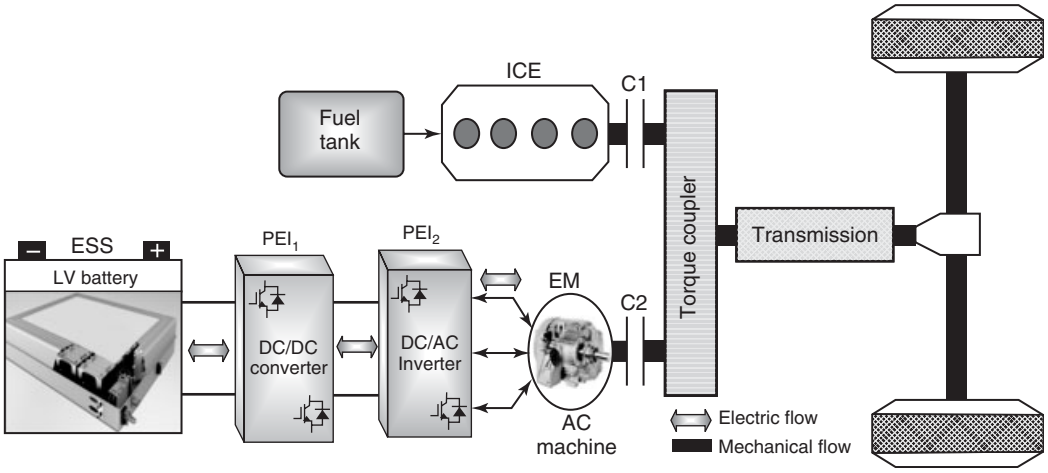


Figure 3. Pretransmission parallel HEV configuration.

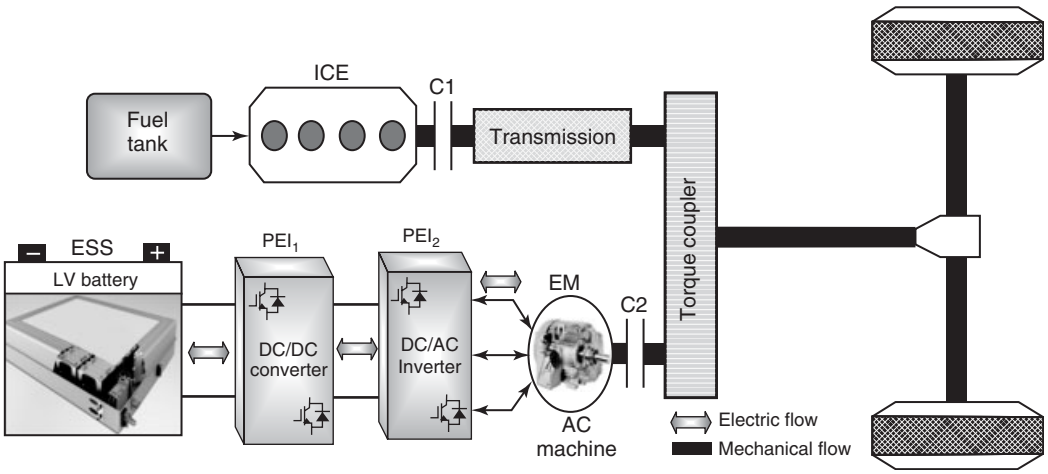


Figure 4. Posttransmission parallel HEV configuration.

the EM. A third solution is the split-transmission parallel HEV, which is also known as *separated axle configuration* (Figure 5).

In a pretransmission parallel HEV, the transmission system is located on the main drive shaft after the torque coupler, as shown in Figure 3. Hence, the gear speed ratios apply on both the ICE and the EM. The mechanical power flows are summed at the transmission system. In other words, the generated torque from the EM is added to the generated torque from the ICE at the input shaft of the transmission system. In addition, there are two clutches (C1 and C2), which are used to disconnect the ICE or the EM from the drivetrain, as illustrated in Figure 3. Therefore, this configuration can provide an independent use of the ICE and the EM, thanks to these clutches.

On the other hand, in a posttransmission parallel HEV, the transmission system is located on the ICE shaft before the torque coupler, as depicted in Figure 4. Hence, the gear speed ratios only apply on the ICE. In a posttransmission parallel HEV, the generated torque from the EM is added to the generated torque from the ICE delivered on the output shaft of the transmission system (Gao and Ehsani, 2006). The transmission system is used to adapt the operating point of the ICE in order to improve the vehicle performance. Furthermore, the posttransmission parallel HEVs can be used in HEVs with a higher degree of hybridization. To further improve the ICE efficiency, a continuous variable transmission (CVT) can be used.

As can be seen in Figures 3 and 4, a low-voltage (LV) battery system can be used to recover more energy from the DC link during regenerative braking. To achieve this operating mode, a bidirectional DC/DC converter is needed to work as a buck converter in order to transfer the power from DC link to LV battery system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto204
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

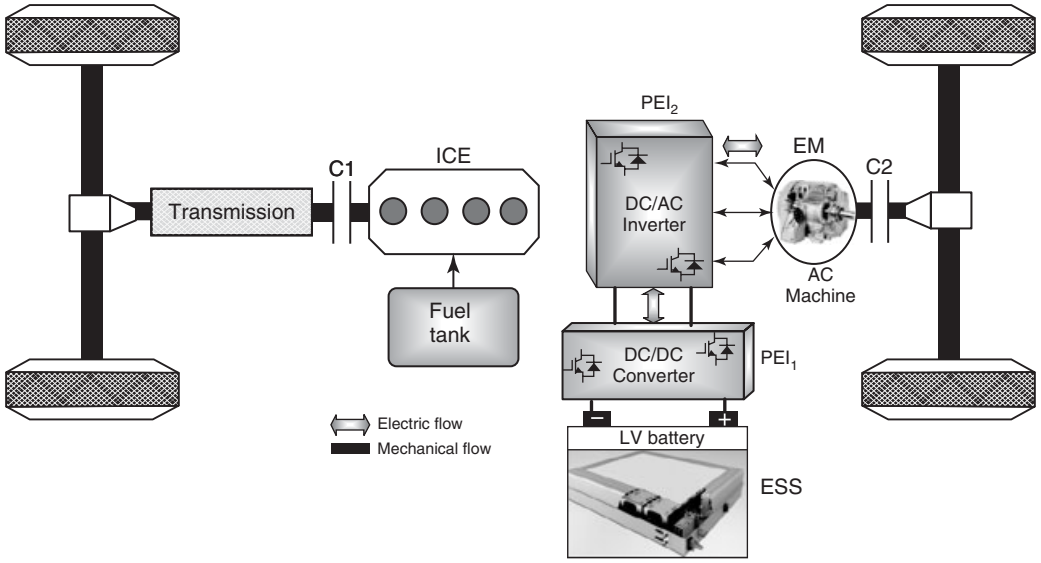


Figure 5. Through-the-road parallel HEV configuration.

As demonstrated in Figure 5, the ICE and the EM can drive on an all-wheel drive vehicle, which is known as *through-the-road configuration* or *separated axle configuration* (Ehsani, Gao, and Emadi, 2010). In this configuration, the ICE and the EM can separately provide the propulsion force to the wheels. The implementation of this configuration is to connect the ICE to the rear wheels, while the EM is connected to the front wheels. In this way, more power can be recuperated during the regenerative braking mode because the major weight is located on the front wheels during braking. Furthermore, this configuration is effective on slippery surfaces by providing vehicle longitudinal stability control that is not as easy with other types of parallel HEV configurations. However, in this configuration, the battery packs cannot be charged from the ICE when the vehicle is in a standstill mode. Furthermore, in the driving mode, the battery is charged by the ICE through an inefficient power path, where the power is transferred from the ICE via the road wheels to the battery system.

Most operating modes described in the SHEVs are still effective for parallel HEVs (see Series Hybrid Electric Vehicles (SHEVs)). It should be pointed out that the SHEVs may be considered as an extension of a purely BEVs with a small ICE, while the parallel HEVs can be considered as an improvement of the ICEVs with small ESS and EM. However, the control strategy of the parallel HEVs is more complicated than the control of the SHEVs because of the mechanical coupling between the ICE and the EM. As mentioned in SHEVs section, one or more PEIs (such as DC/DC converter and DC/AC inverter) are commonly used in parallel HEV powertrain (Emadi, Lee, and Rajashekara, 2008; Van Mierlo and Maggetto, 2001). Furthermore, one or more EMs are required in parallel HEV powertrain. Therefore, the major challenges of the development of the parallel HEV are PEIs, EMs, mechanical coupling systems, and control strategies.

4 HYBRIDIZATION RATE OF THE PARALLEL HEV

The concept of the hybridization rate of the parallel HEV is mainly classified into three basic definitions: (i) electric hybridization rate (EHR), (ii) combustion hybridization rate (CHR), and (iii) rate of hybridization (RoH) (Maggetto and Van Mierlo, 2000; Van Mierlo and Maggetto, 2001). As explained before in the previous chapter (see Series Hybrid Electric Vehicles (SHEVs)), the EHR indicates as to which extent the propulsion system tends toward a BEV. In general, the EHR is the ratio between the electric power and the total traction power. In the case of a parallel HEV, the EHR is equal to the ratio of the EM power to the total traction power, as illustrated in Equation 2.

$$\text{EHR} = \frac{\text{power of the electric motor}}{\text{power of the ICE} + \text{power of the electric motor}} = \frac{P_{\text{EM}}}{P_{\text{ICE}} + P_{\text{EM}}} \quad (2)$$

On the other hand, the CHR is the ratio between the ICE power and the total traction power, as shown in Equation 3.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto204
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Furthermore, the CHR can be expressed in percentage, and a high percentage indicates that the vehicle tends to be an ICEV. The CHR can be written as follows:

$$\mathrm{CHR} = \frac{\text{power of the ICE}}{\text{power of the ICE} + \text{power of the electric motor}} = \frac{P_{\mathrm{ICE}}}{P_{\mathrm{ICE}} + P_{\mathrm{EM}}} \quad (3)$$

To investigate the relative contribution of each energy source to another, the RoH can be defined. Hence, if both energy sources have an equal contribution to the traction power, the RoH will be 1 (or 100%). As was described for SHEVs, if the contribution of the ICE power is higher than the EM power, the RoH ($\mathrm{RoH_{ICE}}$) is the ratio of the EM power (P_{EM}) to the ICE power (P_{ICE}) and vice versa of the $\mathrm{RoH_{EM}}$, as demonstrated in Equations 4 and 5.

$$\text{If } P_{\mathrm{EM}} < P_{\mathrm{ICE}} \Rightarrow \text{Then} \quad \mathrm{RoH_{ICE}} = \frac{P_{\mathrm{EM}}}{P_{\mathrm{ICE}}} \quad (4)$$

$$\text{If } P_{\mathrm{ICE}} < P_{\mathrm{EM}} \Rightarrow \text{Then} \quad \mathrm{RoH_{EM}} = \frac{P_{\mathrm{ICE}}}{P_{\mathrm{EM}}} \quad (5)$$

Figure 6 presents the hybridization rate of the parallel HEVs. In the following section, the operating modes of the parallel HEV are described.

5 THE OPERATING MODES OF THE PARALLEL HEV

Owing to the multiple sources and coupling systems, the parallel HEV has different operating modes in general. These operating modes are illustrated in Figure 7. In addition, these operating modes are described in detail as follows:

1. In *mode 1* (pure electric drive mode), as shown in Figure 7a, the ICE is switched off and the parallel HEV only runs on its EM, which receives the energy from the battery pack. This mode occurs when the parallel HEV speed is low.
2. In *mode 2* (hybrid drive mode), as illustrated in Figure 7b, when the traction power is higher than the maximum power of the ICE, both ICE and the EM are operated to drive the parallel HEV.
3. In *mode 3* (hybrid battery-charging mode), as shown in Figure 7c, the battery pack can be charged by the energy that delivered from the ICE and regenerative braking, when the battery has a very low state of charge (SoC), and the regenerative braking energy is not enough to recharge the battery.
4. In *mode 4* (regenerative braking mode), as depicted in Figure 7d, the braking energy can be recuperated to charge the battery pack by using the EM as a generator.
5. In *mode 5* (pure ICE drive mode), as shown in Figure 7e, the ICE is switched on to drive the parallel HEV, when the battery is fully charged, and the required power is less than the ICE power that can deliver in its optimal operating line.
6. In *mode 6* (engine traction and battery-charging mode), as demonstrated in Figure 7f, the ICE is turned on to propel the parallel HEV and charge the battery, where the battery SoC is below the upper limit, and the required power is less than the ICE power. In this mode, the ICE can operate in

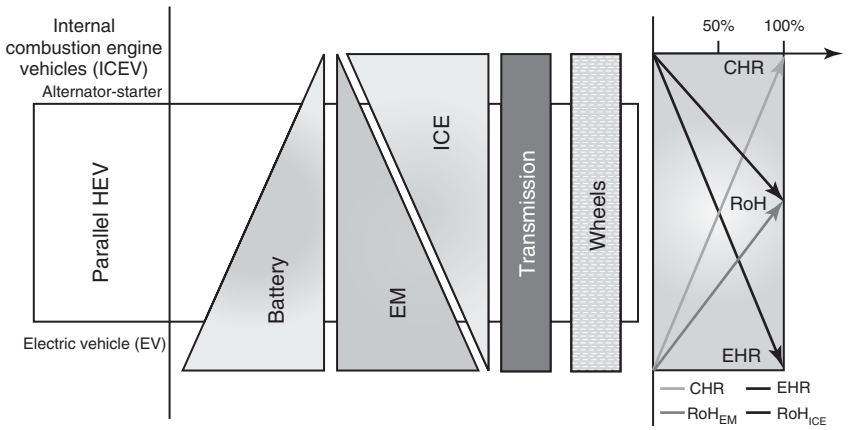


Figure 6. The hybridization rate of the parallel HEV.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto204
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

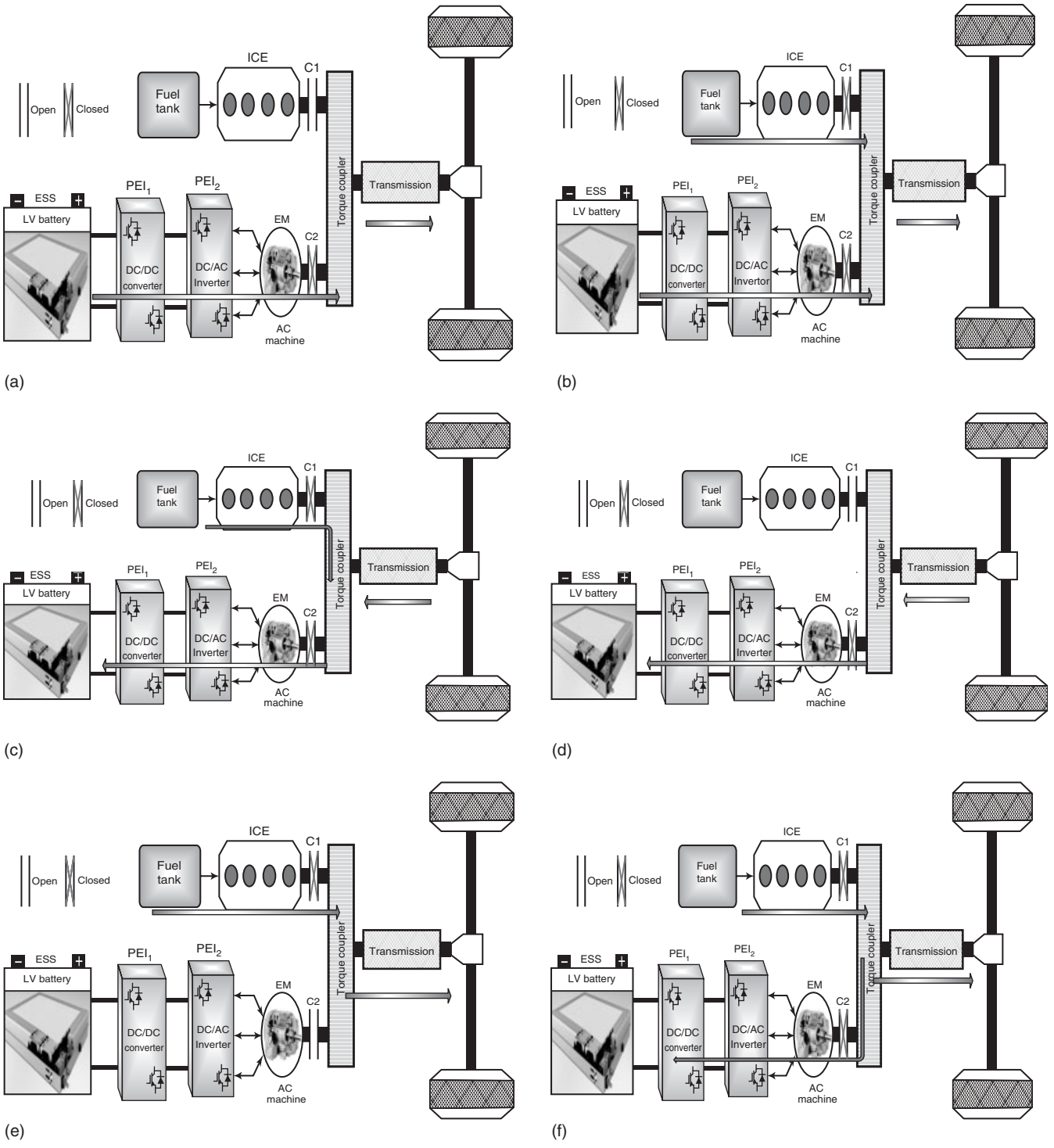


Figure 7. The operating modes of the parallel HEV. (a) Mode 1: pure electric drive. (b) Mode 2: hybrid drive. (c) Mode 3: hybrid battery charging. (d) Mode 4: regenerative braking. (e) Mode 5: pure ICE drive. (f) Mode 6: engine traction and battery charging.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto204
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

its optimal operating line to improve the powertrain efficiency.

6 MECHANICAL COUPLING

In parallel HEV powertrain, the main goal of the mechanical coupling, irrespective of the type of transmission, is to keep the ICE speed and torque within its operating point regardless of the speed and torque requirements at the wheels, as explained Section 2 (Van Mierlo and Maggetto, 2000, 2001). This adaptation of speed can be achieved in different ways for manual and automatic transmission systems. In other words, the two energy sources may be coupled together by either torque or speed coupling.

In the torque-coupling system, the speed values of the ICE and EM have always the same relation to each other. Furthermore, the torque coupling sums the generated torques of the ICE and EM and delivers the total torque to the wheels. As shown in Figure 8, the typical torque coupling is a shaft-fixed gear set. The speed relation between the inputs and outputs can be expressed as follows (Gao and Ehsani, 2006; Van Mierlo and Maggetto, 2000, 2001, 2004):

$$\omega_3 = \frac{1}{i_{g1}}\omega_1, \quad \omega_3 = \frac{1}{i_{g2}}\omega_2, \quad \text{and} \quad \omega_2 = \frac{i_{g2}}{i_{g1}}\omega_1 \tag{6}$$

In the traction mode, the torque relation between the inputs and outputs can be given by

$$T_3 = T_1 \cdot i_{g1} \cdot \eta_1 + T_2 \cdot i_{g2} \cdot \eta_2 \tag{7}$$

However, in the regenerative braking mode, the torque relation between the inputs and outputs can be calculated as follows:

$$T_3 = \frac{T_1 \cdot i_{g1}}{\eta_1} + \frac{T_2 \cdot i_{g2}}{\eta_2} \tag{8}$$

where

η_1	the efficiency of the ICE shaft to the output shaft
η_2	the efficiency of EM shaft to the output shaft
T_1	the ICE torque
T_2	the EM torque
T_3	the output torque
i_{g1}	the gear ratio of the ICE shaft to the output shaft
i_{g2}	the gear ratio of the EM shaft to the output shaft
ω_1	the speed of the ICE shaft
ω_2	the speed of the EM shaft
ω_3	the speed of the output shaft.

7 POWER FLOW CONTROL STRATEGIES

Owing to the mechanical coupling, the control structure of parallel HEV is more complicated than SHEV. The main objectives of the control strategies are to satisfy the requirements of the vehicles (Van Mierlo and Maggetto, 2000, 2001) and improve the performance of the parallel HEVs. In the parallel HEV configuration, the supervisory control strategy comprises energy management rules to split power requirements between the ICE and the EM as well as to determine the transmission gear ratio. Consequently, the

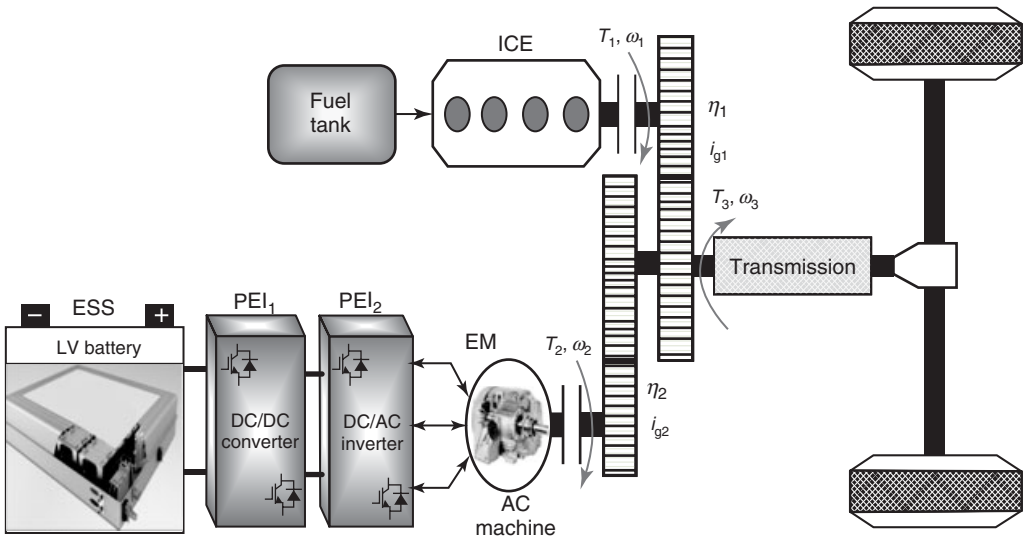


Figure 8. Gear set as the torque coupling for parallel HEV.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto204
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

main objective of the power control strategy is to operate the ICE along its optimal operating line, that is, where the highest efficiency is achieved for every given power level (see the dark gray line in Figure 2). In order to achieve this target, it is important that the correct number of gears and the correct individual gear ratios would be chosen.

Recently, different control strategies in the literature have been reported for parallel HEVs to improve the fuel economy, reduce exhaust emissions, and maintain the battery packs in their desired SoC (Gao and Porandla, 2005; Salmasi, 2007; Gao *et al*., 2009; Bayindir, Gozukucuk, and Teke, 2011; Shen, Shan, and Gao, 2011; Sciarretta, Back, and Guzzella, 2004; Salmanx, Schouten, and Kheir, 2000). These energy management strategies can be generally classified into two kinds and they are as follows:

1. rule-based control strategies
2. optimization-based control strategies.

The rule-based control strategies consist of deterministic and fuzzy logic rule-based methods, while optimization-based control strategies are typically utilized global optimization and real-time optimization when determining the control strategy.

Furthermore, the main concept of rule-based control strategies is commonly based on "IF-THEN" type of control rules. These control strategies only operate the main energy source (i.e., ICE) when it can operate at a high-efficiency region because of high power demands or low SoC of the ESS (such as batteries). The rule-based control strategies can be used to operate the parallel HEV in purely electrical mode from standstill to a low vehicle speed. When the vehicle speeds up from this speed, the ICE will turn on and operate based on optimal operating line. Thereby, the fuel consumption and the exhaust emissions can be reduced, and the total efficiency of the parallel HEV powertrain can be improved. However, as in all HEV propulsion systems, the SoC of the ESS should be controlled as well.

On the other hand, the optimization-based control strategies are used to minimize the fuel consumption and maximize the efficiency of the powertrain. In fact, the global optimal solution is inherently noncausal, in that it provides a minimum fuel consumption using knowledge of future and past power demands. However, these control strategies cannot be directly used for real-time energy management and they are not easily implementable for practical applications (Gao and Porandla, 2005; Salmasi, 2007; Gao *et al*., 2009; Bayindir, Gozukucuk, and Teke, 2011; Shen, Shan, and Gao, 2011; Sciarretta, Back, and Guzzella, 2004; Salmanx, Schouten, and Kheir, 2000).

Nowadays, the Honda Insight, Honda Civic, Ford Escape hybrid SUV, and Honda Accord are some examples of parallel HEVs, which are commercially available in the market.

8 CONCLUSIONS

This chapter presents an overview of the parallel HEV powertrains. Moreover, the advantages and the disadvantages of the parallel HEV are presented. In this chapter, the different operating modes of the parallel HEVs are provided. Furthermore, different mechanical coupling systems that used in parallel HEV powertrains are described. It should be pointed out that parallel HEVs can have predominance in many applications, such as city vehicles and trucks.

LIST OF ABBREVIATIONS

Parallel HEVs	parallel hybrid electric vehicles
ESS	energy storage systems
ICE	internal combustion engine
ICEVs	internal combustion engine vehicles
HEVs	hybrid electric vehicles
PEIs	power electronics interfaces
EM	electric motor
CVT	continuous variable transmission
BEVs	battery electric vehicles
RoH	rate of hybridization
EHR	electric hybridization rate
CHR	combustion hybridization rate
i_g	gear ratio
C_1 and C_2	clutches
SoC	state of charge

REFERENCES

Bayindir, K.C., Gozukucuk, M.A., and Teke, A. (2011) A comprehensive overview of hybrid electric vehicle: powertrain configurations, powertrain control techniques and electronic control units. *Energy Conversion and Management*, **52**, 1305–1313.

Chan, C.C. (2007) The state of the art of electric and hybrid, and fuel cell vehicles. *Proceedings of the IEEE*. *Special issue on Electric, Hybrid and Fuel Cell Vehicles*, **95** (4).

Chan, C.C., Bouscayrol, A., and Chen, K. (2010) Electric, hybrid, and fuel-cell vehicles: architectures and modeling. *IEEE Transactions on Vehicular Technology*, **59** (2), 589–598.

Debal, P., Faid, S., Bervoets, S., *et al*. (2009) *Development of a Post-Transmission Hybrid Powertrain*. EVS24 International Battery, Hybrid and Fuel Cell Electric Vehicle Symposium, May 13–16, 2009.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto204
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Emadi, A., Lee, Y.J., and Rajashekara, K. (2008) Power electronics and motor drives in electric, hybrid electric, and plug-in hybrid electric vehicles. *IEEE Transactions on Industrial Electronics*, **55** (6), 2237–2245.

Ehsani, M., Gao, Y., and Emadi, A. (2010) *Modern Electric, Hybrid Electric and Fuel Cell Vehicles*. ISBN: 978-1-4200-5398-2

Gao, W. and Porandla, S. (2005) Design Optimization of a Parallel Hybrid Electric Powertrain. *Proceedings of the IEEE Vehicle Power and Propulsion Conference*, Chicago, pp. 530–535.

Gao, Y. and Ehsani, M. (2006) A torque and speed coupling hybrid drivetrain architecture, control, and simulation. *IEEE Transactions on Power Electronics*, **21** (3), 741–748.

Gao, J., Sun F., He, H., *et al*. (2009) A Comparative Study of Supervisory Control Strategies for a Series Hybrid Electric Vehicle. *Power and Energy Engineering Conference, APPEEC 2009*, Asia-Pacific, 27–31 March 2009.

Maggetto, G. and Van Mierlo, J. (2000) Electric and Electric Hybrid Vehicle Technology: A Survey. *Proceedings of IEE Seminar on Electric, Hybrid and Fuel Cell Vehicles*, pp. 1/1–1/11, 2000.

Salmanx, M., Schouten, N.J., and Kheir, N.A. (2000) Control Strategies for Parallel Hybrid Vehicles. *Proceedings of the American Control Conference Chicago*, Illinois, June 2000.

Salmasi, F.R. (2007) Control strategies for hybrid electric vehicles: evolution, classification, comparison, and future trends. *IEEE Transactions on Vehicular Technology*, **56** (5), 2393–2403.

Sciarretta, A., Back, M., and Guzzella, L. (2004) Optimal control of parallel hybrid electric vehicles. *IEEE Transactions on Control Systems Technology*, **12** (3), 352–362.

Shen, C., Shan, P., and Gao, T. (2011) A comprehensive overview of hybrid electric vehicles. *International Journal of Vehicular Technology* Article ID 571683, 7 pp.

Van Mierlo J. and Maggetto G. (2000) Views on hybrid drivetrain power management strategies. The 17th World Battery, Hybrid and Fuel Cell Electric Vehicle Symposium EVS-17, Montréal, Canada.

Van Mierlo, J. and Maggetto, G. (2001) Vehicle simulation programme: a tool to evaluate hybrid power management strategies based on an innovative algorithm. *Journal of Automobile Engineering*, **215**, 1043–1052.

Van Mierlo, J. and Maggetto, G. (2004) Innovative iteration algorithm for a vehicle simulation program. *IEEE Transactions on Vehicular Technology*, **53** (2), 401–412.

Van Mierlo, J., Van den Bossche, P., and Maggetto, G. (2004) Models of energy sources for EV and HEV: fuel cells, batteries, ultracapacitors, flywheels and engine-generators. *Journal of Power Sources*, **128** (1), 76–89.

FURTHER READING

Chau, K.T. and Wong, Y.S. (2002) Overview of power management in hybrid electric vehicles. *Energy Conversion and Management*, **43**, 1953–1968.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto204
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Series–Parallel Hybrid Electric Vehicles

Joeri Van Mierlo, Omar Hegazy, Jelle Smekens, and Cedric De Cauwer

Vrije Universiteit Brussel (VUB), Brussels, Belgium

1 Introduction 1
2 Concept of Series–Parallel Hybrid Electric Vehicles 1
3 Transmission Systems 3
4 Series–Parallel Hybrid Electric Vehicle Configurations 9
5 Power Flow Control Strategies 12
6 Conclusions 15
List of Abbreviations 15
Endnotes 17
References 17
Further Reading 17

1 INTRODUCTION

In general, hybrid electric vehicles (HEVs) can be classified into series hybrids, parallel hybrids, and series–parallel hybrid vehicles. The series–parallel hybrid electric vehicles (series–parallel HEVs) incorporate the features of both series hybrid electric vehicles (SHEVs) and parallel hybrid electric vehicles (parallel HEVs). This powertrain comprises an additional mechanical link as compared with the SHEVs and also an additional generator as compared with the parallel HEVs. Because of the fact that series–parallel HEVs can operate in both parallel and series modes, the fuel economy and drivability can be optimized based on the operating condition of the vehicle. However, due to increased components and complexity, the series–parallel HEVs can be more expensive than a series or parallel HEV.

The series–parallel HEV is also known as a *combined hybrid electric vehicle* (*CHEV*) or *power-split HEV* (Liu and Peng, 2008; Chan, 2007; Chan, Bouscayrol, and Chen, 2010; Bayindir, Gozukucuk, and Teke, 2011; Van Mierlo and Maggetto, 2000, 2004). The Toyota Hybrid System (THS) is a series–parallel drivetrain, which was developed and used on Toyota Prius in 1997 in Japan (Liu and Peng, 2008). As is reported in the literature, the series–parallel HEVs are investigated and developed to overcome the drawbacks of series and parallel architectures (Grammatico, Balluchi, and Cosoli, 2010).

This chapter reviews the state of the art of the series–parallel HEVs, and gives an overview on their powertrain and power management strategy. In this chapter, different transmission systems [such as continuous variable transmission (CVT) and electric variable transmission (EVT)] are described. In addition, the operating modes of the series–parallel HEVs are explained in detail.

2 CONCEPT OF SERIES–PARALLEL HYBRID ELECTRIC VEHICLES

There are a number of variations of series–parallel HEV configurations. The series–parallel HEV contains an internal combustion engine (ICE) and one or more energy storage systems (ESSs), as illustrated in Figure 1. Configurations 1 and 2 also consist of two electric motors (EMs) and a power-split device. The power coupling contains a torque and/or speed coupling and possibly one or more single- or multigear transmissions. With the invention of EVTs, a third configuration has been made possible in which the EVT fulfills coupling, motor, and generator

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

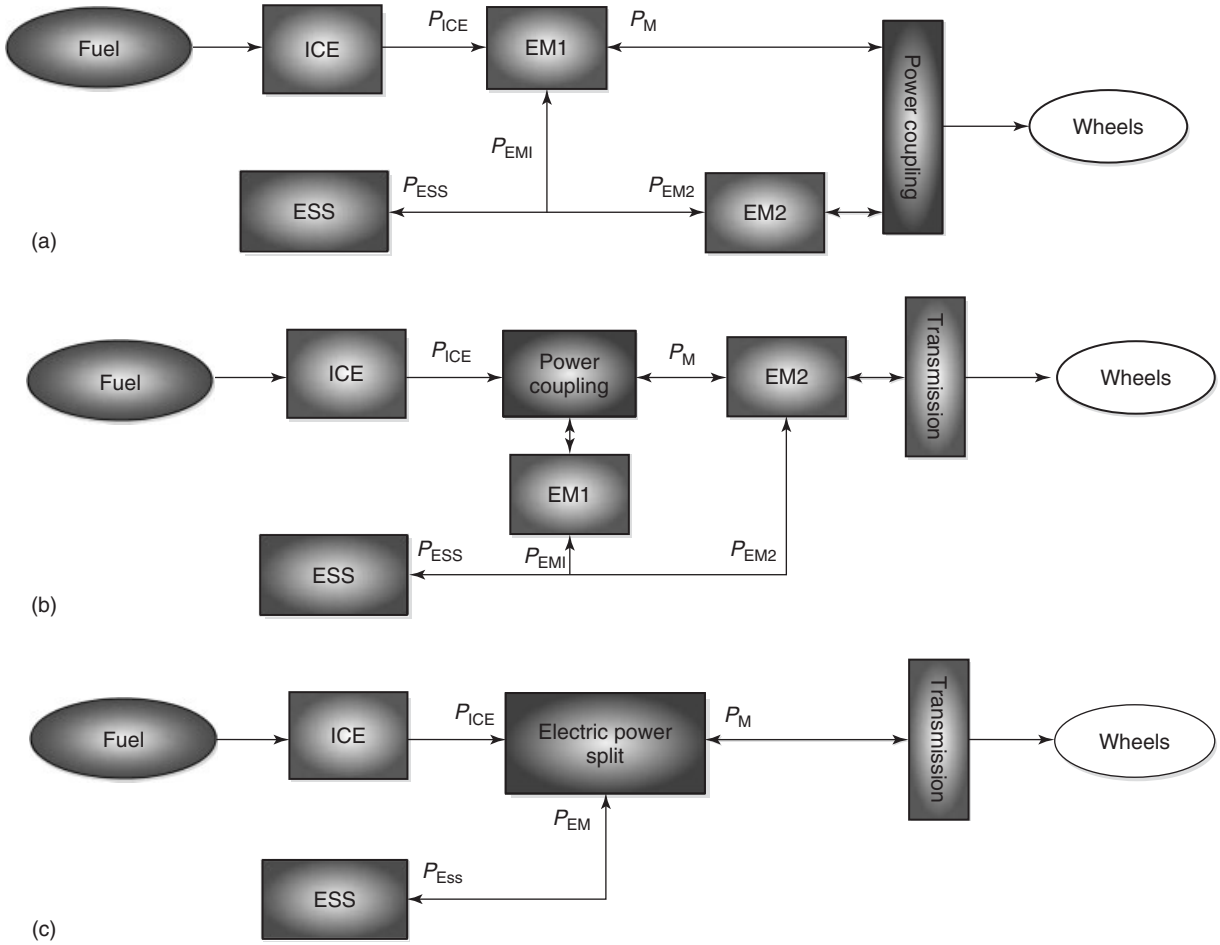


Figure 1. Schematic diagram of series–parallel HEVs. (a) Configuration 1; (b) configuration 2; (c) configuration 3.

functions. In configuration 3, the EVT is contained in the electric power split together with a single- or multigear transmission. Because the EVT fulfills multiple functions, it is convenient to introduce the concept of a power split to compare with the other configurations. In Section 2, the power split is introduced, and we will see that the combination of power coupling device and two electric machines can realize the same power conversions as a power split. In the schematic diagrams in Figure 1, no electronic power converters or controllers are depicted, but it is presumed that they are implicitly present.

The advantage of a series–parallel HEV is that it has several more operation modes than those of an SHEV or parallel HEV separately and thus can benefit from the advantages of both, as described in the previous chapters (see Series Hybrid Electric Vehicles (SHEVs) and Parallel Hybrid Electric Vehicles (Parallel HEVs)). Furthermore, one can observe the differences between the three topologies from the point of the view of the design of a drivetrain.

- In an SHEV, the rated power of the traction motor is directly determined by the design requirements of the vehicle. So, there is no degree of freedom (DoF) in downsizing the EM. However, the ICE can be downsized compared to the conventional ICE-powered vehicle. In addition, the ICE can run in its high-efficiency operating point when charging the battery.
- In a parallel HEV, there is more liberty in downsizing the ICE or the EM because their power is added to deliver the power to the wheels. However, the mechanical coupling of the ICE to the wheels makes it impossible to choose the operation of the ICE independent of the speed. The total power of the ICE and the EM is equal to the traction power. So, depending on the control strategy of this drivetrain, the ICE can be operated in its optimal region.
- In a series–parallel HEV, the power distribution among the ICE, EM1, and EM2 can be controlled to give more DoF compared to SHEV and parallel HEV powertrains.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Moreover, the power split allows the rotational speed of the motor to be decoupled from the speed of the vehicle.

With a series–parallel HEV, the carmaker has more DoF to look for an optimal powertrain relative to SHEVs and parallel HEVs. However, the disadvantage of the series–parallel HEV is that control strategies are more complex because of the increase in subsystems.

In the following section, the concept of a planetary gearbox, which functions as speed coupling, is explained in detail.

3 TRANSMISSION SYSTEMS

To realize the concept of a series–parallel HEV, a particular power conversion is required: the mechanical power (P_{ICE}) from the ICE and the electrical power (P_E) from the battery are considered to be power inputs and have to be summed up to deliver the traction power (P_T) transmitted to the wheels. As shown in Figure 2, this apparatus should have a bidirectional electric power input (P_E), an unidirectional mechanical power input for the ICE (P_{ICE}), and a bidirectional mechanical power output to connect the wheels of the vehicle (P_T). The output should be bidirectional for regenerative-braking purposes. The operating modes this device should deal with are the same, as specified in Section 3. It is convenient to call such a transmission system a power split. As mentioned earlier, a power split can be realized by the combination of a power coupling (i.e., a speed or torque coupler) and two electric machines or by an EVT. Below a speed coupler is introduced as a particular power coupling.

Analogous with the torque coupler that is introduced in the chapter on parallel HEVs (see Parallel Hybrid Electric Vehicles (Parallel HEVs)), we can define a speed coupler. A torque coupler signifies a three-port power transmission of which the output torque is a linear combination of the input torques, whereas a speed coupler signifies a three-port power transmission of which the output speed is a linear combination of the input speeds as written in Equation 1.

$$\omega_T = S_{ICE}.\omega_{ICE} + S_E.\omega_E \tag{1}$$

where S_{ICE} and S_E are constants associated with the design of the speed coupler. ω_{ICE} and ω_E are, respectively, the rotational speed of the ICE and the rotational speed that can be associated with the electrical input. ω_T is the rotational speed of the output shaft. By introducing energy conservation and neglecting the losses, we see that this power converter offers two degrees of freedom: ω_{ICE} and ω_E. The torques are linked together by the following equation of which the lowest torque determines the other two (Ehsani, Gao, and Emadi, 2010).

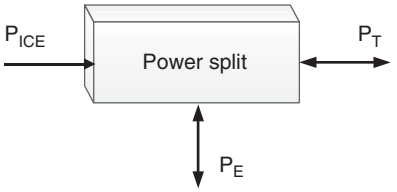


Figure 2. The block diagram of the power split.

$$T_T = \frac{T_{ICE}}{S_{ICE}} = \frac{T_E}{S_E} \tag{2}$$

where T_T, T_{ICE}, and T_E are, respectively, the torques related to the input and two output shafts. In the Toyota Prius, the speed coupling is realized using a planetary gearbox.

3.1 Planetary gearbox

A speed coupler can be realized using a planetary gear in which the torque values of all the shafts have a constant relation to each other. Planetary gearing is widely used in many HEVs as a speed coupling (Gao and Ehsani, 2006; Van Mierlo and Maggetto, 2000, 2001). It features a compact structure, high load ability, and a high efficiency. Figure 3 illustrates the section drawing of the planetary gear. It includes three parts, a sun gear, a ring gear, and several planet gears connected by a gear carrier. The sun gear, the ring gear, and the carrier gear are the three shafts a planetary gear set connects to the outside. The planetary gear set is often utilized in an automatic transmission. Such a gear set can produce one or two gear reductions and reverse based on simultaneously engaging or locking various elements of the planetary systems. For example, the gear reduction can be produced in the planetary gear set by turning the sun gear and holding the ring gear. The speed relation between the inputs and outputs can be expressed as follows (Van Mierlo and Maggetto, 2000, 2001):

$$\omega_c \cdot (1 + \rho) = \rho \cdot \omega_s + \omega_r \tag{3}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

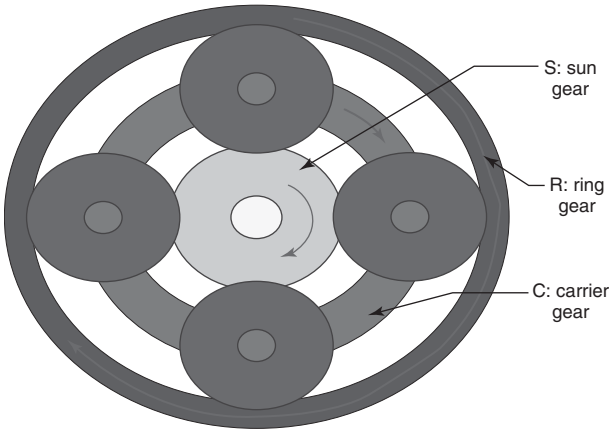


Figure 3. Planetary gearbox for speed coupling.

The torque relation between the inputs and outputs can be expressed by the following equation (Van Mierlo and Maggetto, 2001):

$$T_C = \frac{(1+\rho)}{\eta_S \cdot \rho} \cdot T_S = \frac{(1+\rho)}{\eta_R} \cdot T_R \qquad (4)$$

where

T torque
η_S efficiency from sun to carrier
η_R efficiency from ring to carrier
ω_s sun gear angular speed
ω_r ring gear angular speed
ω_c carrier angular speed
ρ planetary gear ratio (number of the sun gear teeth/number of ring gear teeth)

As an example, we will explain the operational modes for a configuration 2-type series–parallel HEV. In the Toyota Prius II powertrain, the generator is mechanically connected to the sun gear, the motor to the ring gear, and the ICE to the carrier. The relation between the angular speeds of the three gears is described in Equation 5.

$$\omega_s = \left(1 + \frac{n_r}{n_s}\right) \times \omega_c - \frac{n_r}{n_s} \times \omega_r \qquad (5)$$

where in this particular case,

n_s number of teeth of the sun gear
n_r number of teeth of the ring gear
ω_s sun gear angular speed equal to the generator angular speed
ω_r ring gear angular speed equal to the motor angular speed
ω_c carrier angular speed equal to the ICE angular speed

Additionally, the vehicle speed is linked to the motor speed by means of a gearbox (Figure 9, Section 3). Equation 5 is illustrated by a nomogram in Figure 4. In this figure, one can see three vertical lines representing all possible values of, respectively, the generator speed, the ICE speed, and the motor speed (or sun, carrier, and ring gears). The relative spacing between these vertical lines is related to the planetary gear ratio. The lines given in different colors represent the speed relation for different operational modes, which are discussed later. For example, with a generator speed of 2000 rpm and the ICE turned off (i.e., zero angular speed), the motor speed is determined by the intersection of the third vertical line with the straight line through these two points (generator speed of 2000 rpm, ICE of zero speed).

Operational modes as represented in Figure 4 are listed in the following:

- *Stand still* is represented by a red line, where all three speeds are zero.
- *Zero-emission driving* is represented by a yellow line. At low speeds, the ICE is turned off. Therefore, the ICE speed should be zero.
- *Reverse driving* is represented by a skin-colored line. Reverse driving is a particular form of zero-emission driving where the motor speed is negative.
- *Cranking aid* is represented by a purple line, where the ICE is driven by the generator, while the motor speed is zero. When the vehicle speed increases to a nonzero speed, the ICE's speed increases and it provides additional power, illustrated by the purple line migrating toward the position of the upper green line, which represents hybrid driving.
- *Hybrid driving* is represented by the green lines. In the upper line, the generator has a nonzero speed, while in the lower line, the generator is standing still. If the vehicle speed is set, the generator speed fixes the ICE speed. To preserve the freedom in choosing the operating point of the ICE, the generator must turn accordingly. This means that the generator must work on idle (no torque) to obtain the hybrid drive mode, that is, no battery charging with hybrid driving when ICE speed requirements fix the generator speed to a nonzero value.

3.2 Electric variable transmission

As described in the beginning of Section 3, the EVTs are contained in the electric power split in configuration 3 (Figures 1c and 10). Here, we will go in more detail on this device and give a review on the latest developments. An EVT equipped with the proper electric converters is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

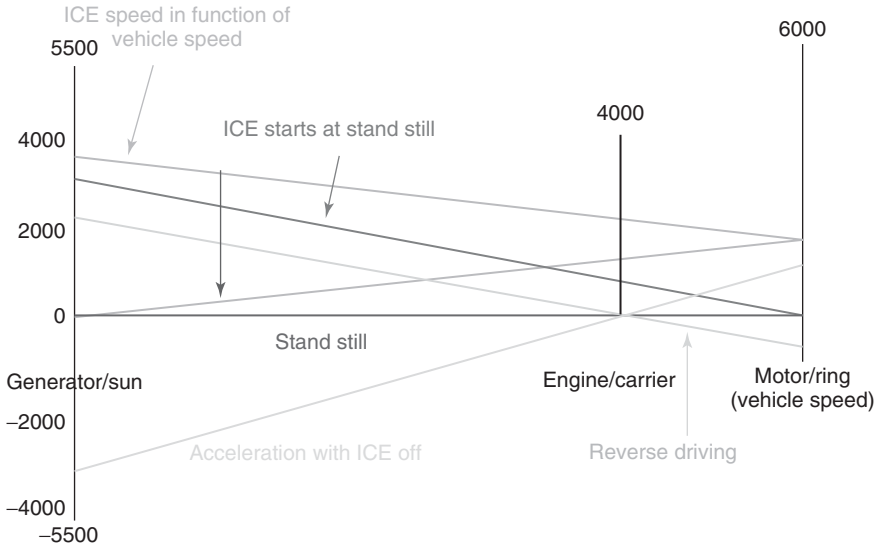


Figure 4. Nomogram of the Toyota Prius.

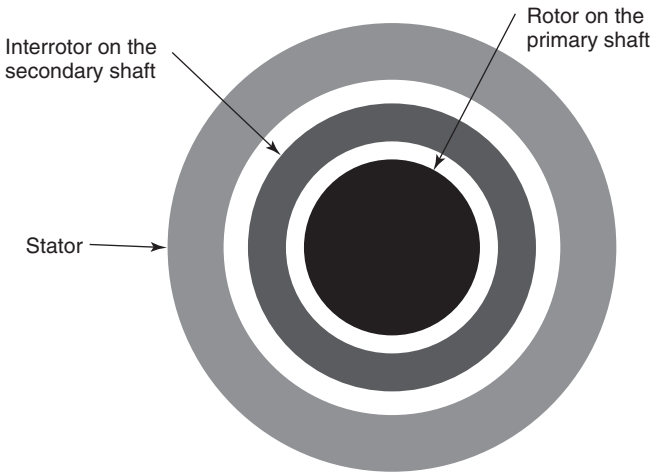


Figure 5. Two concentrically arranged electric machines.

capable of the same power conversions as a power split, hence the name electric power split.

The basic concept is to combine two electric machines concentrically based on magnetic coupling (Figure 5). As indicated in Figure 5, this concept is realized using two rotors and one stator. The rotor, which is fixed with the secondary shaft and is rotating between the rotor and the stator, is called *interrotor*. The electric power is exchanged via two power electronics converters with the DC-bus. The result is a power split with two mechanical ports and one electrical port (Miller, 2006).

One mechanical connection, the primary shaft, is connected to the ICE and the secondary shaft is connected to the driveline. For the inner machine (rotor and interrotor), we can write the following energy balance (Figure 6) (Kessels and van den Bosch, 2009).

$$P_{ICE} = T_{ICE}.\omega_{ICE} = P_{IM} + P_D \tag{6}$$

where P_D is the direct mechanical power, which is transferred to the secondary axis, and P_{IM} is the electrical power involved in this power transfer. For the outer machine, one can write the mechanical power as follows:

$$P_D = P_{EM} + P_M \tag{7}$$

where P_M $(= T_M \cdot \omega_M)$ is the mechanical output power and P_{EM} is the electrical power involved in this power transfer. The net electrical power is the sum of the electrical power from the power converters. This power is given by Equation 7:

$$P_E = P_{E1} + P_{E2} \tag{8}$$

where

- P_E includes the sum of the power from the electric converters and this power is accumulated in the battery.
- P_{E1} and P_{E2} are the electrical powers that are transferred to the inner and outer rotors, respectively. They are given by the following two equations.

$$P_{E1} = \begin{cases} P_{IM}\eta_1, & \text{if } P_{IM} \geq 0 \\ \frac{P_{IM}}{\eta_1}, & \text{if } P_{IM} < 0 \end{cases} \tag{9}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

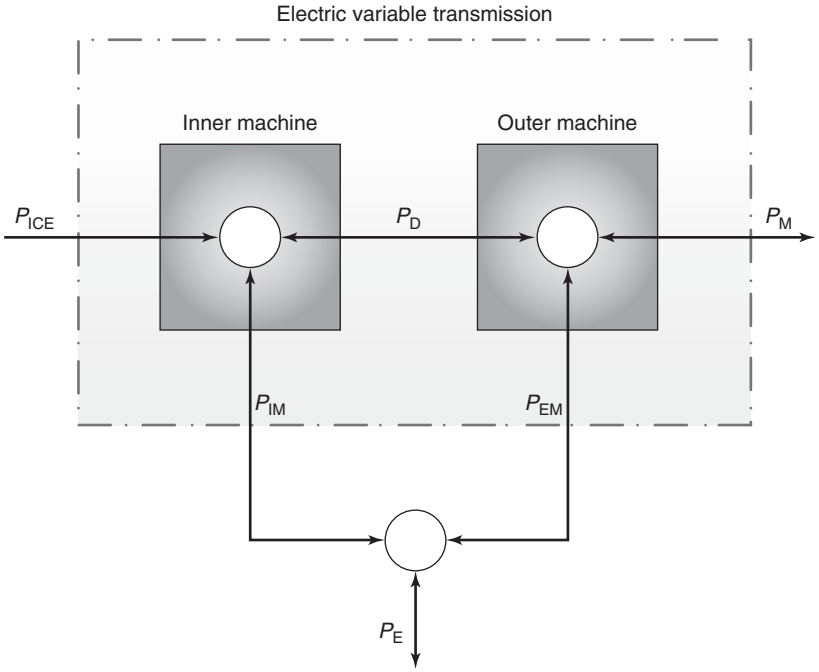


Figure 6. Power flows in an EVT.

$$P_{E2} = \begin{cases} P_{EM}\eta_2, & \text{if } P_{EM} \geq 0 \\ \frac{P_{EM}}{\eta_2}, & \text{if } P_{EM} < 0 \end{cases} \tag{10}$$

where η_1 and η_2 are the efficiencies of the electrical machines and the electronic power converters.

Although there is a direct power exchange, there is no mechanical link between the primary and secondary shafts. Furthermore, the two rotational speeds of the inner and outer rotors are independent. Figure 6 illustrates the power flows in an EVT. To understand how the ICE and battery are coupled to this apparatus, we refer to Figure 10 (Section 3).

Hoeijmakers and Rondel (2004) and Hoeijmakers (2003) proposed an EVT for HEVs. It was composed of two concentrically arranged induction machines (IMs): the outer machine is a common squirrel-cage induction machine (SCIM) with three-phase windings on the stator and the squirrel-cage on the interrotor. The inner machine is also an SCIM but with the three-phase windings on its inner part, the rotor, and its squirrel cage is on the outer part, the interrotor. The interrotor is made thin to make the machine light and compact, the fields of both machines turn with the same speed and the slip frequency is the same. The behavior of the two SCIMs cannot be separated and thus is not the same as two individual IMs. This is explained as follows (Figure 7): when the current in the rotor increases, the flux of the inner machine may saturate the interrotor material. Consequently, a part of the flux passes through the air gap of the external machine. In case a current passes in another direction of the stator windings, an electromechanical torque will be produced between the rotor and stator.

Another interesting way of accomplishing the variable transmission and the power split is the combination of two permanent magnet (PM) machines. Many configurations are possible depending on where the PMs are placed: on the rotor, interrotor, or stator (Xizheng *et al*., 2009; Chen, Quan, and Zhu, 2011). PM machines offer the advantage of high power density and good efficiency, but it is difficult to weaken the magnetic field and they are relatively expensive because of the scarce materials used for PMs. Especially for

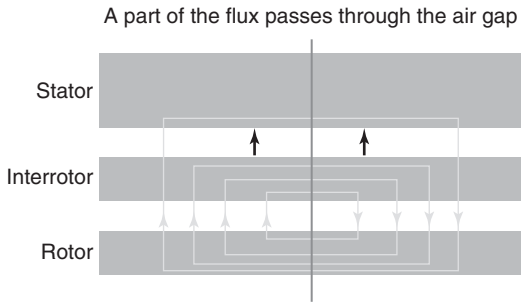


Figure 7. Magnetic coupling between the inner and the outer machines.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

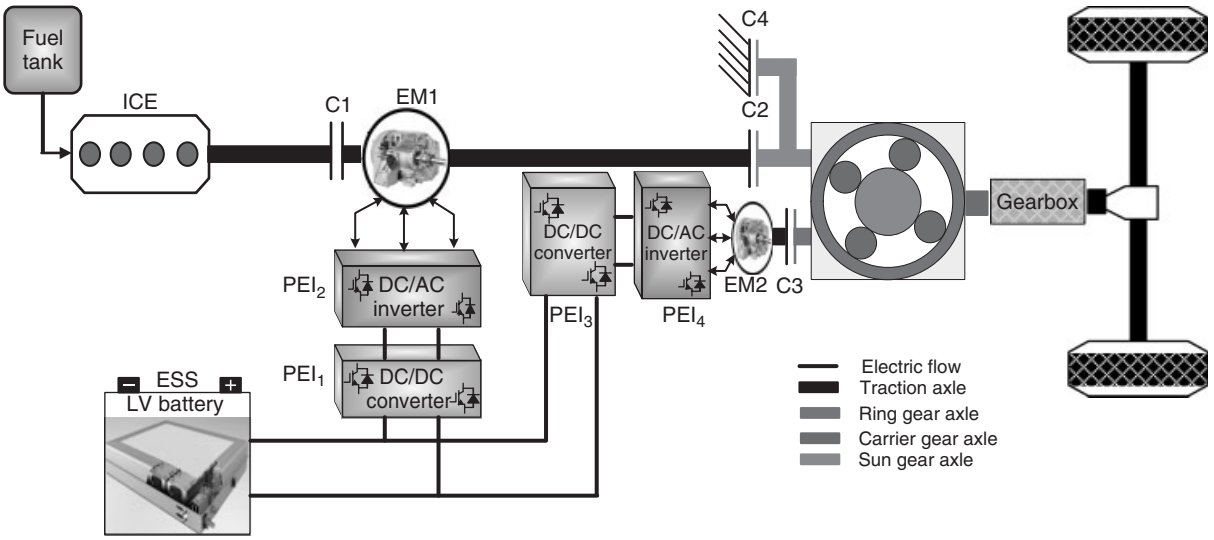


Figure 8. Configuration 1.

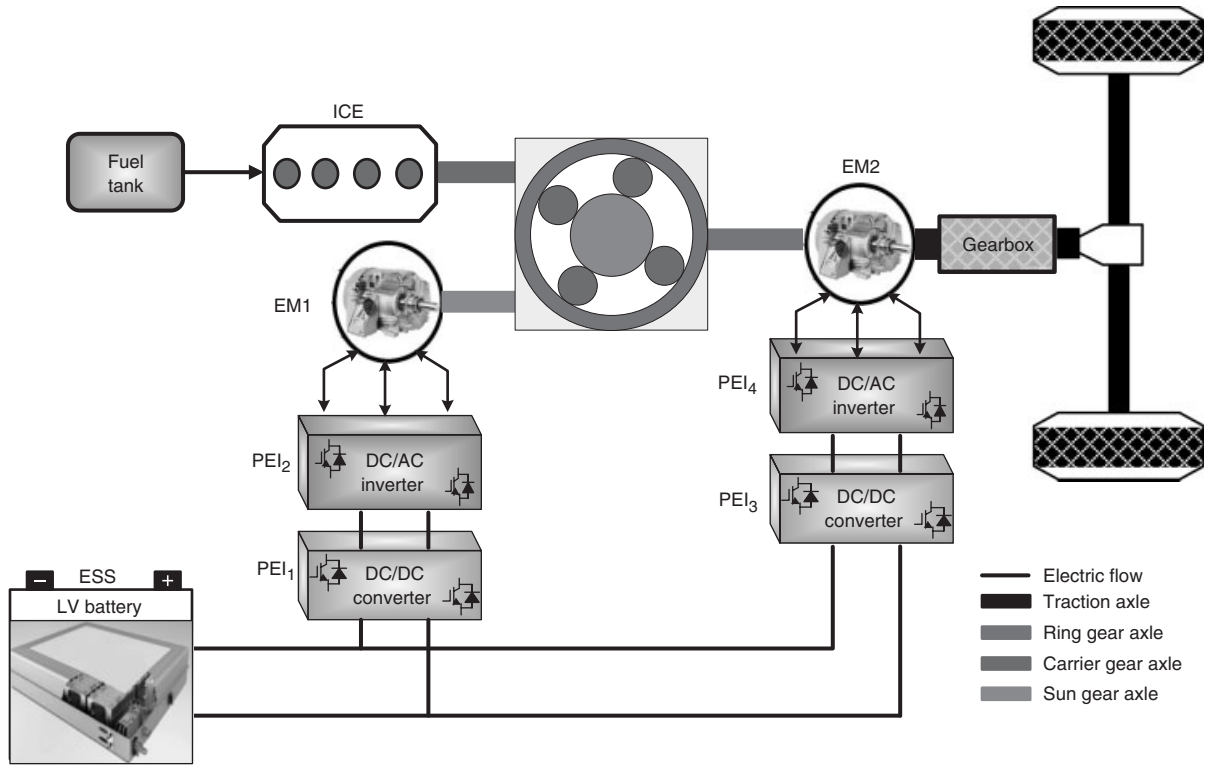


Figure 9. Configuration 2 based on CVT.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

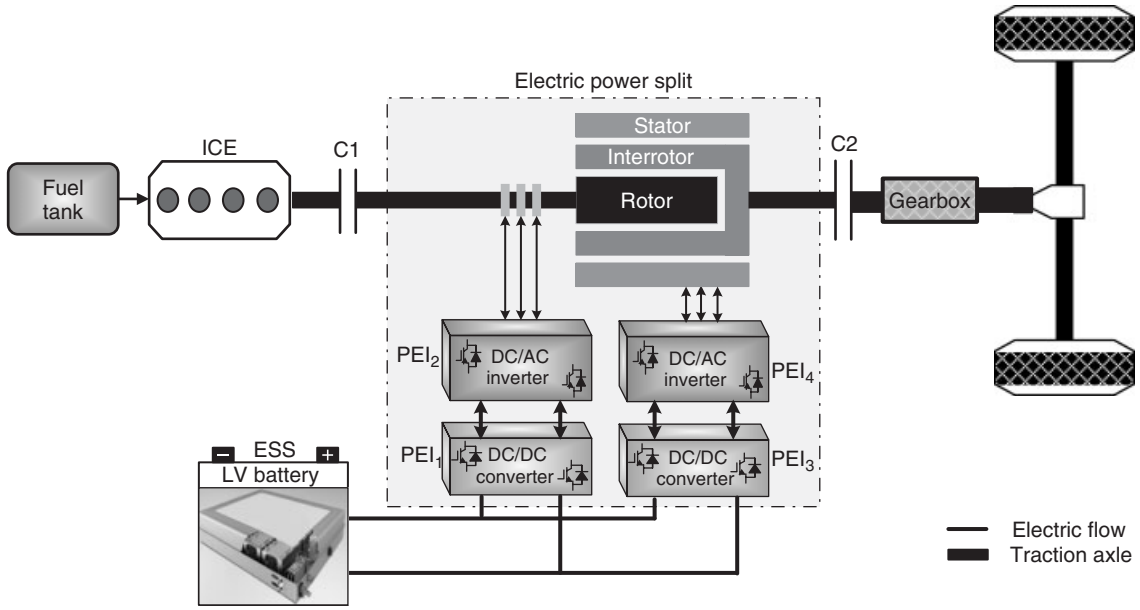


Figure 10. Configuration 3 based on the EVT.

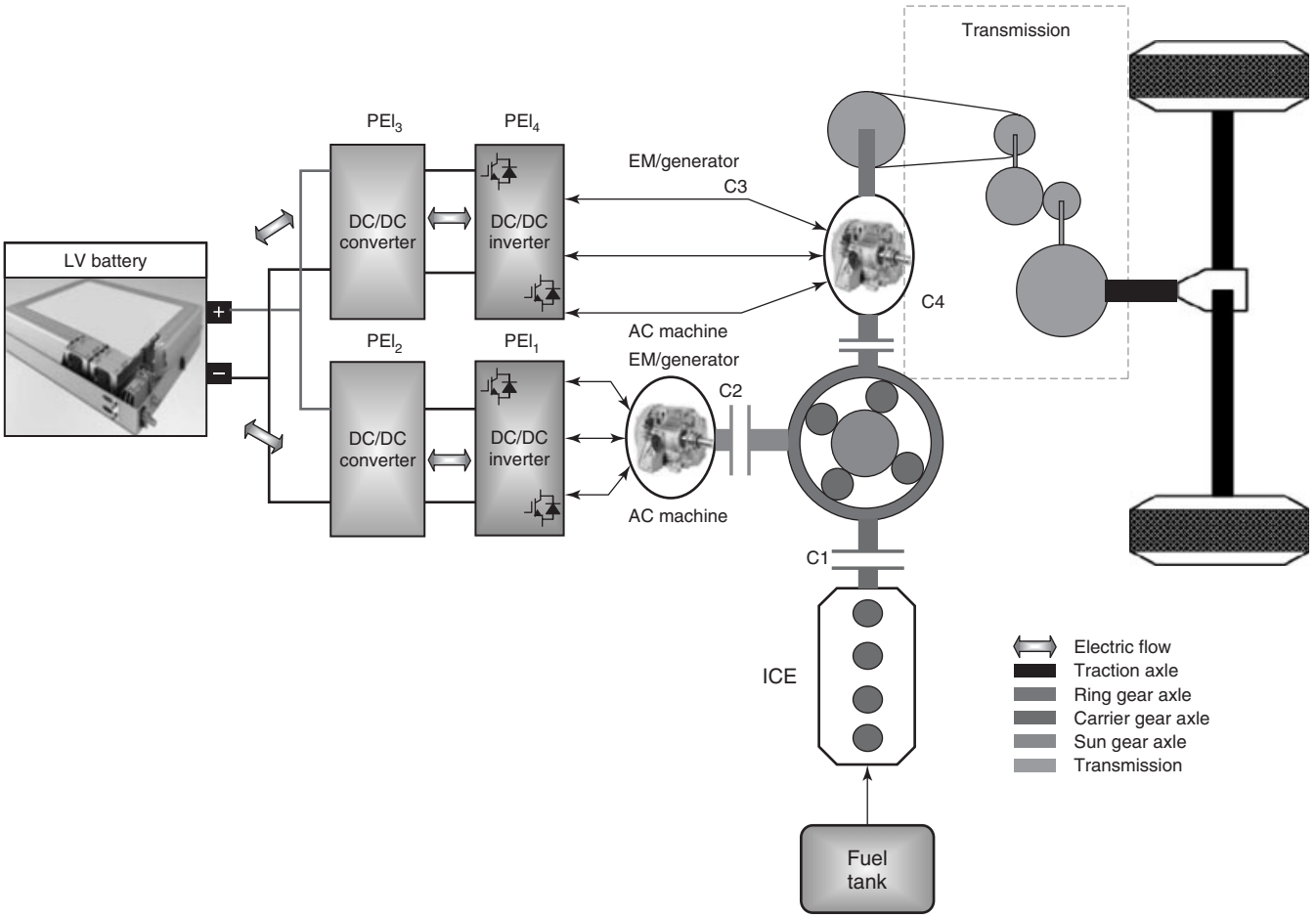


Figure 11. The physical implementation of an example of a configuration 2-type combined hybrid architecture.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

higher speeds, their inability for field weakening control is a drawback.

A third way that is being researched to make an EVT is the implementation of two switched reluctance machines (SRMs). SRMs are simple, reliable, have a lower manufacturing cost than IMs, and show superior thermal and mechanical robustnesses. They can accept high rotation speeds and high temperatures but they are noisy and produce a large torque ripple (Van Mierlo and Maggetto, 2001; Chen, Quan, and Zhu, 2011).

With the PM machines and SRM solution, the magnetic coupling between the inner and outer machines can be ignored. With SCIM, this is not the case, presenting an extra difficulty in the controllability of this machine.

As reported in the literature, the EVT can achieve all operating modes, which are required for the series–parallel HEV (Hoeijmakers and Ferreira, 2006; Hoeijmakers and Rondel, 2004).

4 SERIES–PARALLEL HYBRID ELECTRIC VEHICLE CONFIGURATIONS

In general, power can flow from the ICE completely and directly to the wheels (traction power), completely to the battery (charging power), or can be split in traction power and charging power. Electric power can flow from (traction power) and to the battery (braking or charging power). Combinations of these power flows result in eight practical operating modes, which are discussed in Section 4.1. As seen in Section 2, three configurations are possible to realize all the power flows. They are discussed later. In configuration 1 (Figure 8), the power from EM1 ($\eta_{EM1} \cdot P_{EM1}$) is added with the power from the ICE (P_{ICE})—they are on the same axis to deliver the mechanical power P_M. P_M is added to the mechanical power from EM2 ($\eta_{EM2} \cdot P_{EM2}$)

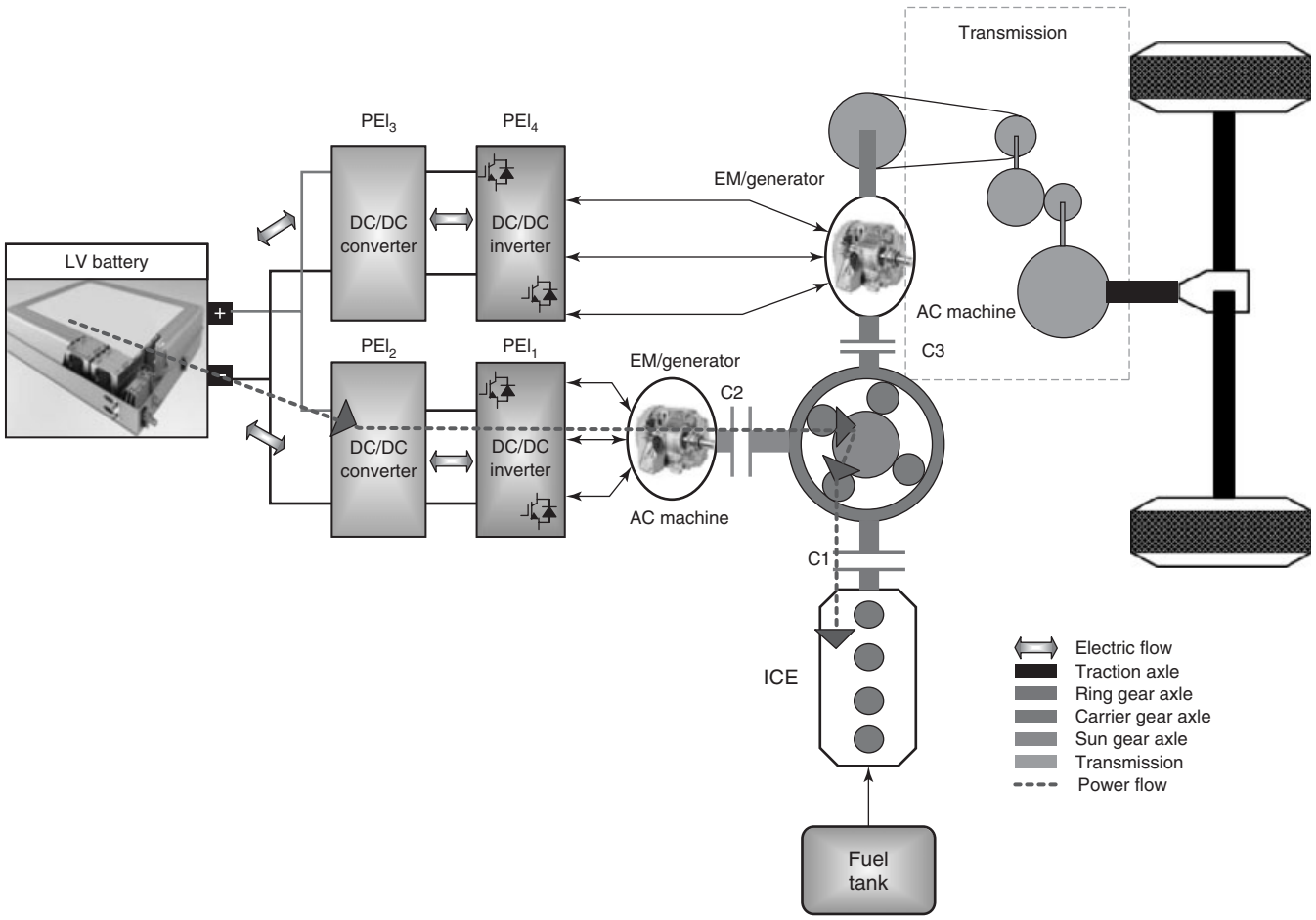


Figure 12. Operational mode 1: cranking aid.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

by a mechanic coupling. An example of this configuration is the Opel Ampera/Chevrolet Volt. When the vehicle is driving at low speeds, the pure electric mode is most efficient. In this case, only EM2 is generating traction power, and clutch 3 (C3) and clutch 4 (C4) are closed, while clutch 2 (C2) is opened to decouple EM1. At higher speeds, more traction power is required; C4 is opened, while C2 is closed. Consequently, EM1 can be operated to deliver the additional traction power. C1 is opened to limit the inertia and friction losses from the ICE. When the battery is almost depleted, the ICE is turned on to extend the range of the vehicle. EM1 is used as a starter-motor (cranking aid) to start the ICE. C1 is closed to realize this effect. Then, EM1 can be used as a generator. At low speeds, all the power from the ICE is converted into electric power (C2 open and C4 closed; cf. mode 6 in Section 4.1), and at high speeds a part of this power is transferred to the wheels (C2 closed and C4 opened; cf. mode 5 in Section 4.1).

In configuration 2 (Figure 9), the mechanical power from the ICE (P_{ICE}) is added with the output of EM1 ($\eta_{EM1} \cdot P_{EM1}$) by a power split device. The output axis of the power split is the same as the rotor of EM2. An example of this configuration is the Toyota Prius II, where the power split is also a planetary gearbox. Note that in both configurations, the electric machines EM1 and EM2 can be operated as a generator or motor to make all modes possible (Section 4.1). Note the absence of clutches in this powertrain.

As demonstrated in Figures 8 and 9, configurations 1 and 2 can both be accomplished with the same number of components: an ICE, an ESS, two EMs, a planetary gearbox, and transmission and electronic power converters for the EMs. Consequently, these configurations are comparable in complexity, size, and weight. In the case of configuration 3, the EM1, EM2, and the mechanic coupling are replaced by one electric machine, an EVT (Figure10). This configuration is lighter and likely to be more efficient.

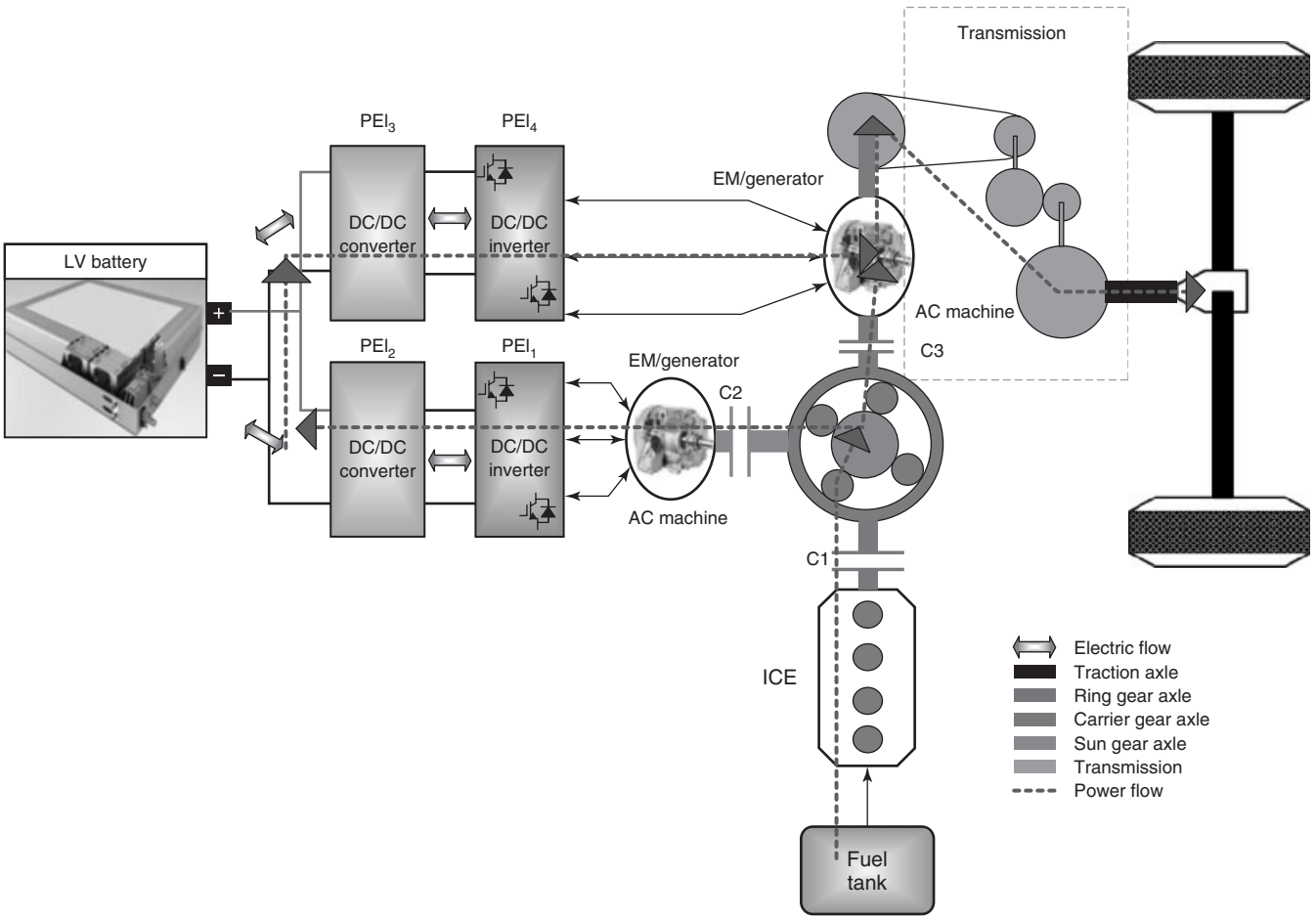


Figure 13. Operational mode 2: pure ICE drive.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.1 The operating modes of the series–parallel HEV

All operating modes described for the SHEV and parallel HEV (see Series Hybrid Electric Vehicles (SHEVs)and Parallel Hybrid Electric Vehicles (Parallel HEVs)) are still valid for the series–parallel HEVs, but the realization of the operating modes can be different from the SHEV because of the specific architecture of the series–parallel HEV, that is, a mechanical connection between the EM and the ICE. Moreover, the mechanical connection between the EM and the ICE allows another operating mode: it allows the ICE to be driven by the EM when starting up the car. Therefore, eight possible operating modes exist for the series–parallel HEVs.

The different operational modes are illustrated by means of the power flow in an example of a configuration 2-type series–parallel HEV. The power split device is planetary gearbox, with the ICE connected to the carrier, the generator to the sun wheel, and the motor to the ring wheel. The implementation of the series–parallel HEV architecture for this example is depicted in Figure 11.

The different operating modes of the series–parallel HEVs are described as follows:

1. *Mode 1* (cranking aid): the motor/generator is used to drive the ICE for start-up of the ICE, as illustrated in Figure 12.
2. *Mode 2* (*pure ICE drive*): only the ICE supplies the required power. This mode can either be achieved through the mechanical connection between the ICE (see Parallel Hybrid Electric Vehicles (Parallel HEVs)) and the wheels or through the numerous electrical conversion steps (see Series Hybrid Electric Vehicles (SHEVs)). Both power flows are shown in Figure 13. The power flow strategy will determine which of the both can be used according to the circumstances.

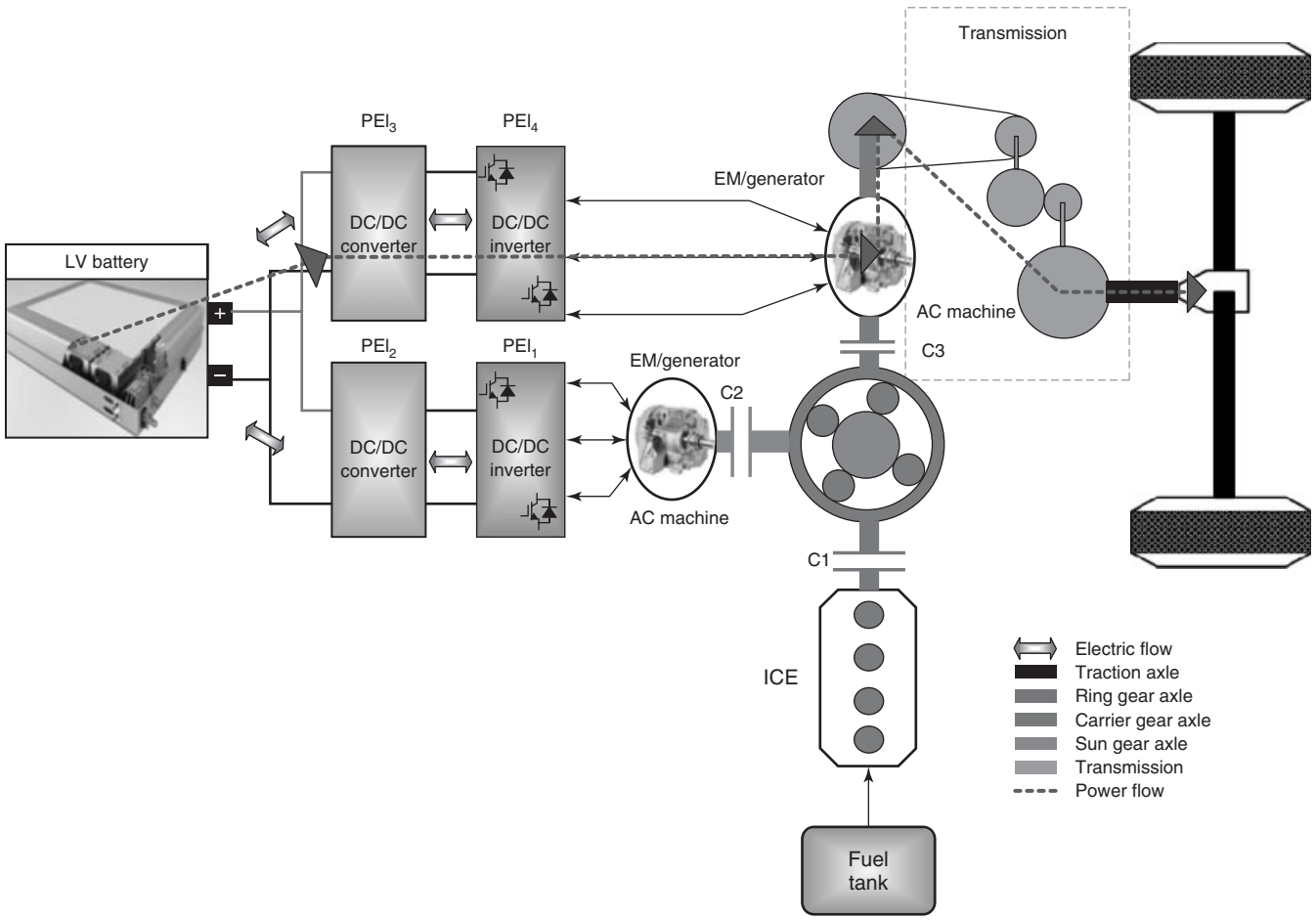


Figure 14. Operational mode 3: electric mode.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

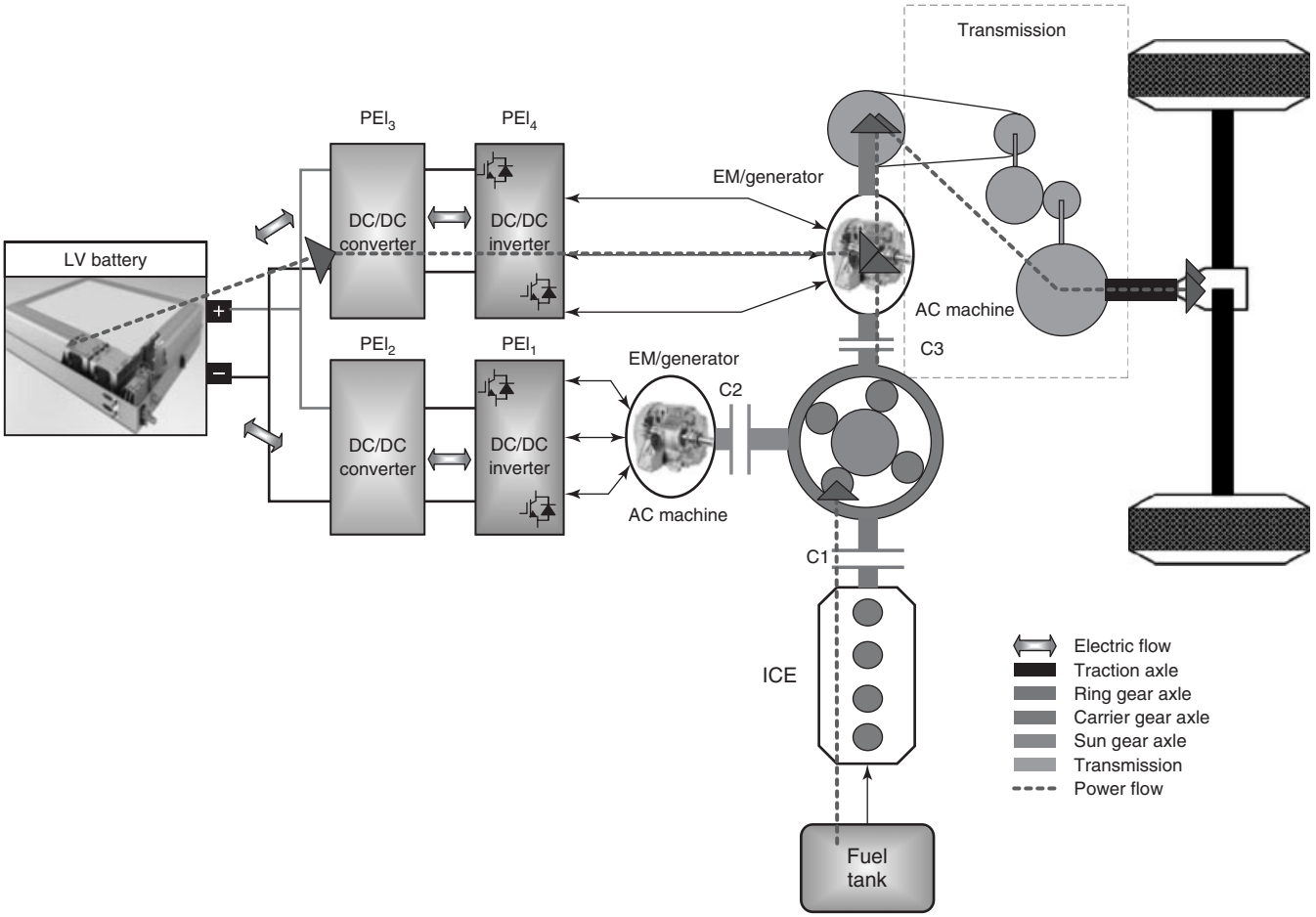


Figure 15. Operational mode 4: hybrid drive.

3. *Mode 3* (*electric mode*): in this mode, the ICE and generator are turned off and the vehicle is only supplied by the batteries (Figure 14).
4. *Mode 4* (*hybrid drive*): the traction power is supplied by the ICE and the batteries. This mode is used when the power requirement is high (Figure 15).
5. *Mode 5* (*engine traction and battery charging*): the ICE provides the energy needed to charge the batteries and drive the vehicle (Figure 16).
6. *Mode 6* (*battery charging and no drive*): in this mode, the generator transfers mechanical power from the ICE to electrical power to charge the batteries (Figure 17).
7. *Mode 7* (*regenerative braking*): during braking or deceleration, the ICE is turned off and the traction motor is operated as generator to charge the battery. Because of the speed coupling, during regenerative braking, the speed of both generators should be such that the ICE speed is zero, or the ICE must be unclutched (Figure 18).
8. *Mode 8* (*hybrid battery charging*): both the ICE and the traction axle provide power to charge the battery (Figure 19).

5 POWER FLOW CONTROL STRATEGIES

The power flow control strategies consist essentially in using the vehicle in a combination of the operational modes to obtain or optimize the objectives for a HEV specified in the previous chapter (see Series Hybrid Electric Vehicles (SHEVs)). A series–parallel HEV can combine the features of the SHEV and the parallel HEV for its power flow control strategies. In an SHEV, the ICE operating point can be set independently of the vehicle speed, but the power flow from the ICE to the drive axle has numerous power conversion steps, accumulating losses at each conversion step. This can be disadvantageous in some driving conditions. In a parallel HEV, there exists a direct

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

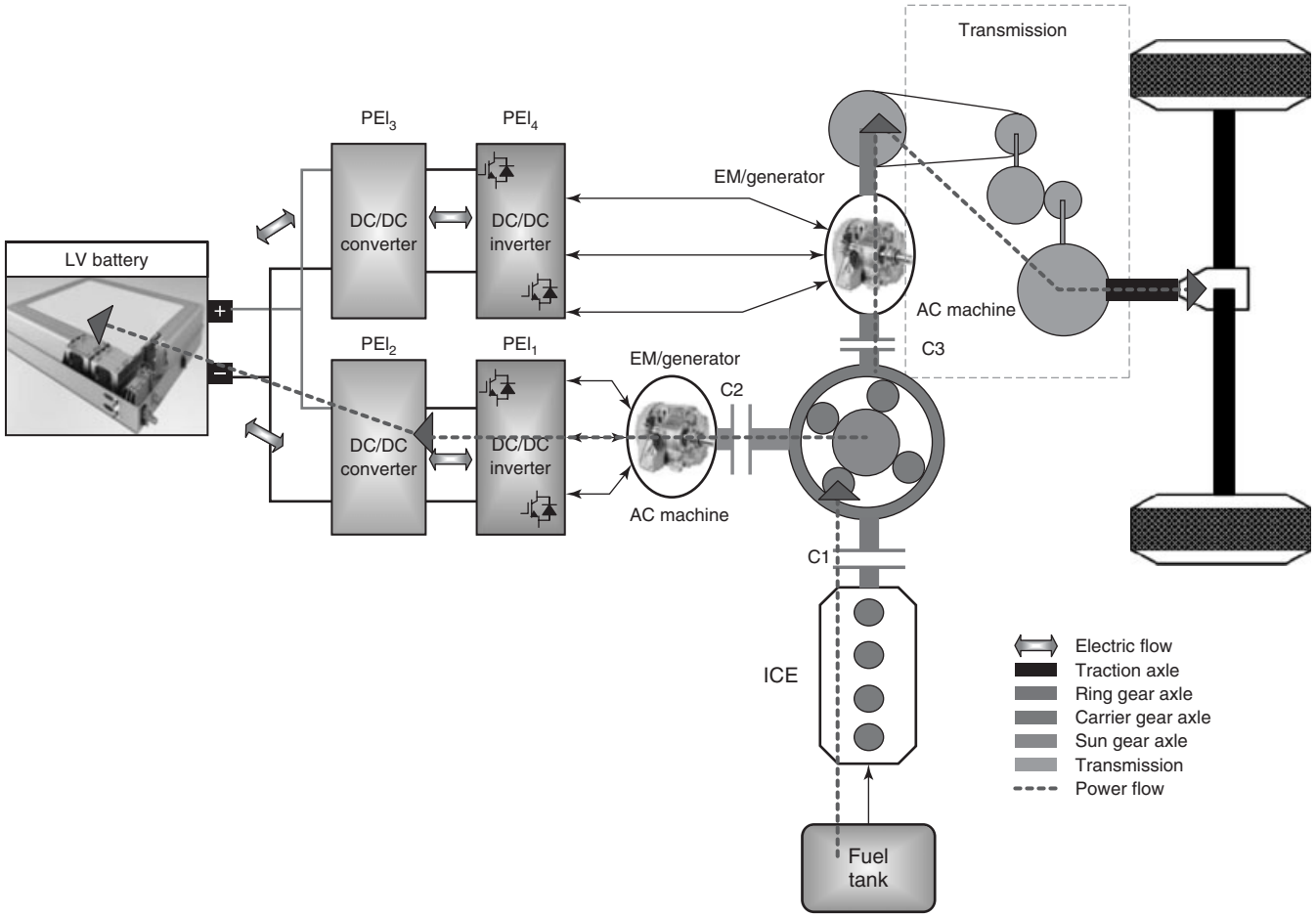


Figure 16. Operational mode 5: engine traction and battery charging.

mechanical coupling between the ICE and the drive axle. As a result, the power from the ICE can be directly transmitted to the drive axle with less power conversion steps than that in an SHEV. This is likely to result in a higher overall efficiency compared to an SHEV if the driving conditions allow the ICE to work in an operating point with a high efficiency. But the direct mechanical coupling also fixes the ICE speed according to the vehicle speed such that the overall efficiency can be lower than that of an SHEV if the driving conditions force the ICE to operate in operating points with low efficiency.

Series–parallel HEVs have the ability to combine these features of an SHEV and a parallel HEV. The series–parallel HEV's speed and torque coupling make it possible to let the ICE work in a fixed working point, regardless of the vehicle speed (Van Mierlo and Maggetto, 2000, 2001), but still has a direct mechanical coupling between the ICE and the drive axle.

One approach is to keep the ICE operating in its highest efficiency point. Keeping the ICE in a constant working point, that is, the working point corresponding to the lowest fuel consumption, does not necessarily mean the lowest fuel consumption of the complete drivetrain. While the average power is delivered by the ICE, acceleration peaks are covered by the EM (Section 4.1, mode 4) and with power demands lower than the average power, the surplus of ICE power recharges the battery (Section 4.1, mode 5). This can result in frequent charging and discharging and possibly lower overall efficiency.

Another strategy is to target maximum overall efficiency. As indicated in Section 3, the operational mode 2 has multiple paths for the power flow, and overall efficiency is determined by the power requirement (driving conditions), power distribution, and the power flow pathway. By calculating the power distribution that minimizes overall losses for each torque and speed value (driving conditions), maximum overall efficiency can be obtained (Van Mierlo and Maggetto, 2001).

In every control strategy, the control of the state of charge (SoC) of the ESS must be incorporated and limited

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

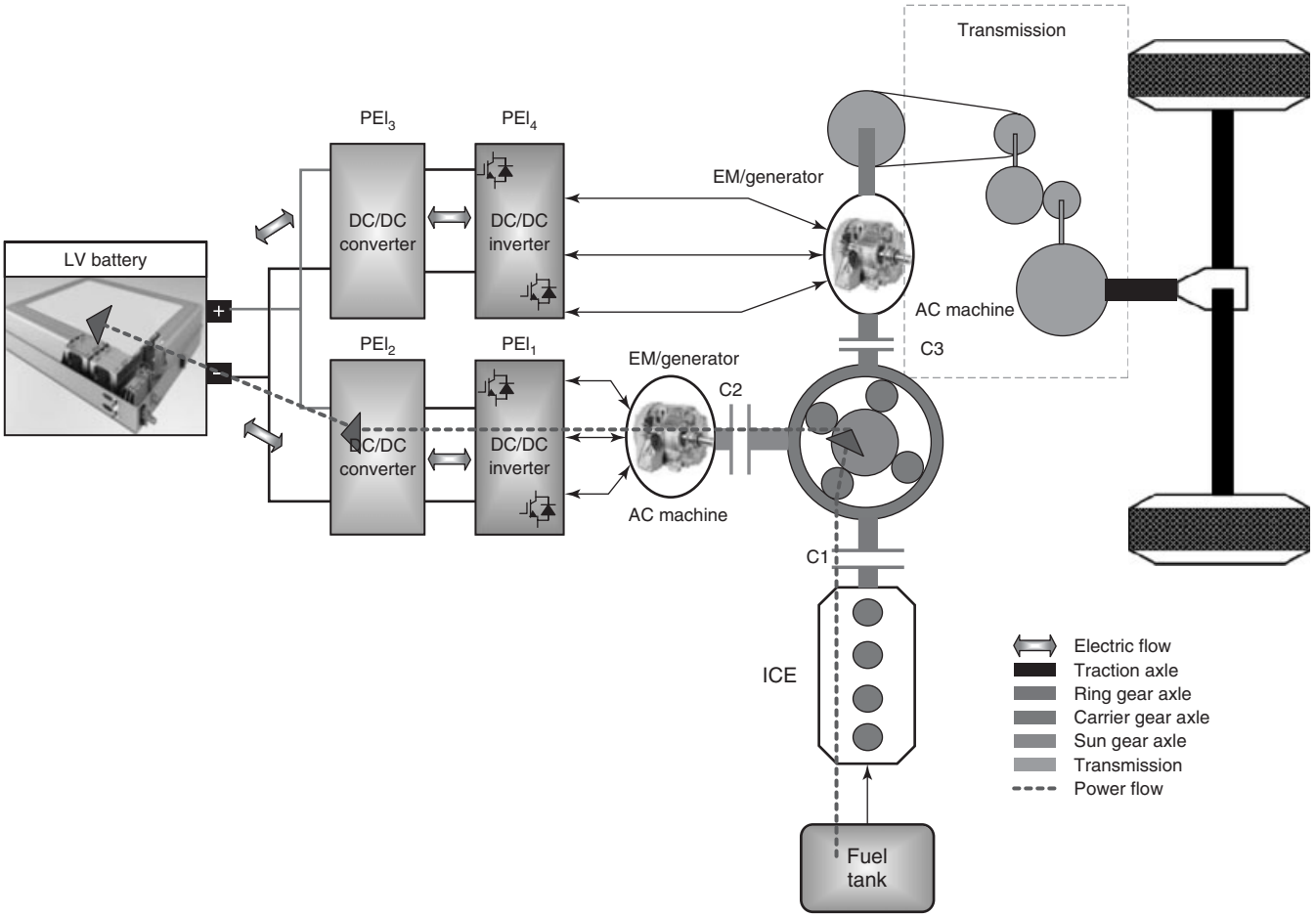


Figure 17. Operational mode 6: battery charging and no drive.

battery capacity can have an impact on the strategy (Van Mierlo and Maggetto, 2001, 2004; Van Mierlo et al. 2004; Salmasi, 2007; Barrero, Van Mierlo, and Tackoen, 2008). Battery capacity ranges from close to 1 kWh for a hybrid with almost no all electric range (AER), to a multiple or even an order of magnitude higher than that of the capacity for a plug-in hybrid vehicle, depending on the target AER (Abkemeier and Mize, 2012; Axsen, Burke, and Kurani, 2008; US Department of Energy, 2010).[1] The ESS SoC operating range in the case of a battery is generally between 80% and 40% charged, where 20% is considered as a minimum SoC to limit the battery degradation. To maintain the SoC, it can be necessary to impose a power demand for battery charging in certain drive cycles. This limits the freedom in which other strategies, such as that mentioned earlier, can be applied as the power distribution will depend on not only the driving conditions but also the power demand for battery charging. An SoC-maintaining strategy will provide battery power in the case of high or low power demand. The high power demand will be provided by the battery and the ICE together (Section 4.1, mode 4), while in the case of a low power demand, only the battery will provide the required power and the ICE is switched off (Section 4.1, mode 3). For moderate power demands, the ICE will drive the wheels and charge the battery through the generator with a charging power as a function of SoC (Section 4.1, mode 5) (Van Mierlo and Maggetto, 2001). SoC-maintaining strategies are applied in vehicles with limited battery capacity, such as the Toyota Prius, as well as in vehicles with considerable battery capacity (such as plug-in hybrid vehicles such as the Opel Ampera/Chevrolet Volt) when the battery depletion is near to its limit or at the driver's request. Figure 20 illustrates the SoC-maintaining strategy for a hybrid vehicle with limited battery capacity and a plug-in hybrid vehicle.

As was explained in the previous chapter (see Series Hybrid Electric Vehicles (SHEVs)) for SHEV, for series–parallel HEV, the power flow strategies can also generally be classified into two kinds: (i) rule-based control strategy and (ii) optimization based control strategy.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

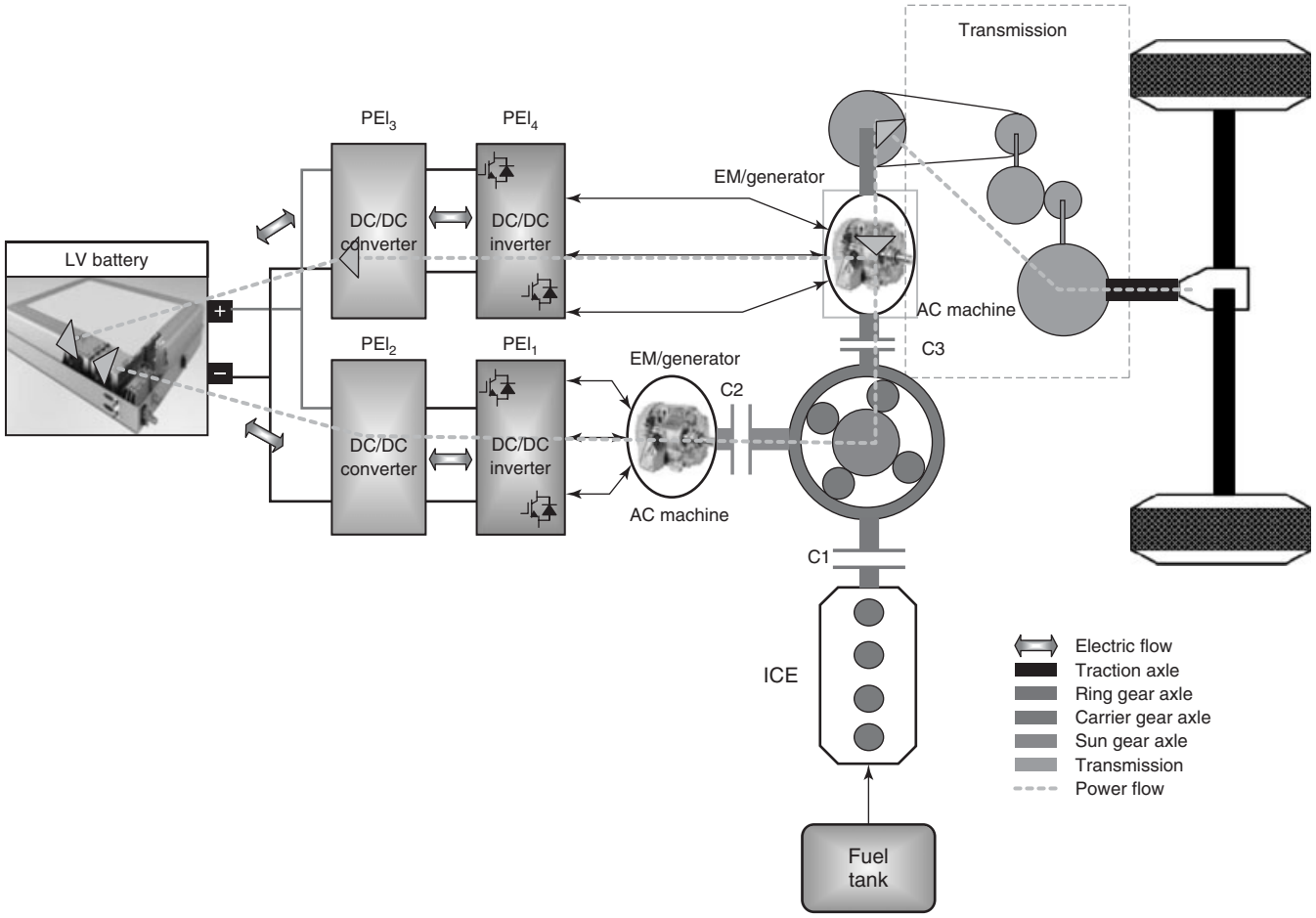


Figure 18. Operational mode 7: regenerative braking.

6 CONCLUSIONS

This chapter has presented an overview of series–parallel HEV powertrains, which are known as *CHEVs*. The main concept of the series–parallel is presented, followed by the introduction of the different operating modes of the series–parallel HEV. In addition, the transmission systems such as CVT and EVT are introduced as they are essential for state-of-the series–parallel HEV. As can be observed from the series–parallel HEV architectures, the series–parallel HEV has the features of both series and parallel topologies. Compared to series and parallel HEVs, the series–parallel HEVs can provide a good solution for better fuel economy and low emissions for both city and highway driving cycles. The Toyota Prius, Lexus, Opel Ampera/Chevrolet Volt, and Ford Escape are some examples of series–parallel HEVs, which are commercially available.

LIST OF ABBREVIATIONS

Series–parallel HEVs	series–parallel hybrid electric vehicles
Parallel HEVs	parallel hybrid electric vehicles
SHEVs	series hybrid electric vehicles
CHEVs	combined hybrid electric vehicles
ESS	energy storage system
ICE	internal combustion engine
HEVs	hybrid electric vehicles
EM	electric motor
CVT	continuous variable transmission
EVT	electric variable transmission
C1 & C2	clutches
SoC	state of charge
DoF	degree of freedom
SCIM	squirrel-cage induction machine
PM	permanent magnet
AER	all electric range

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

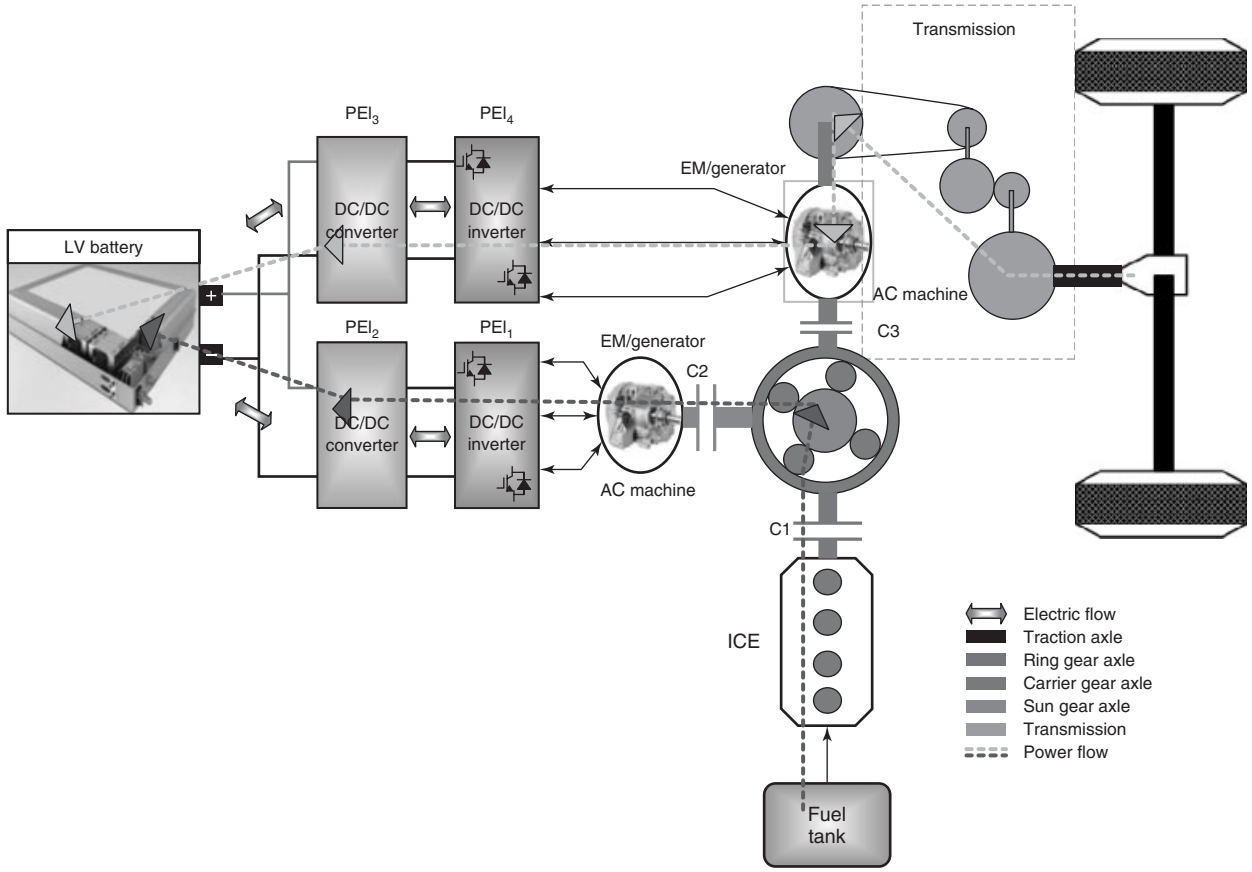


Figure 19. Operational mode 8: hybrid battery charging.

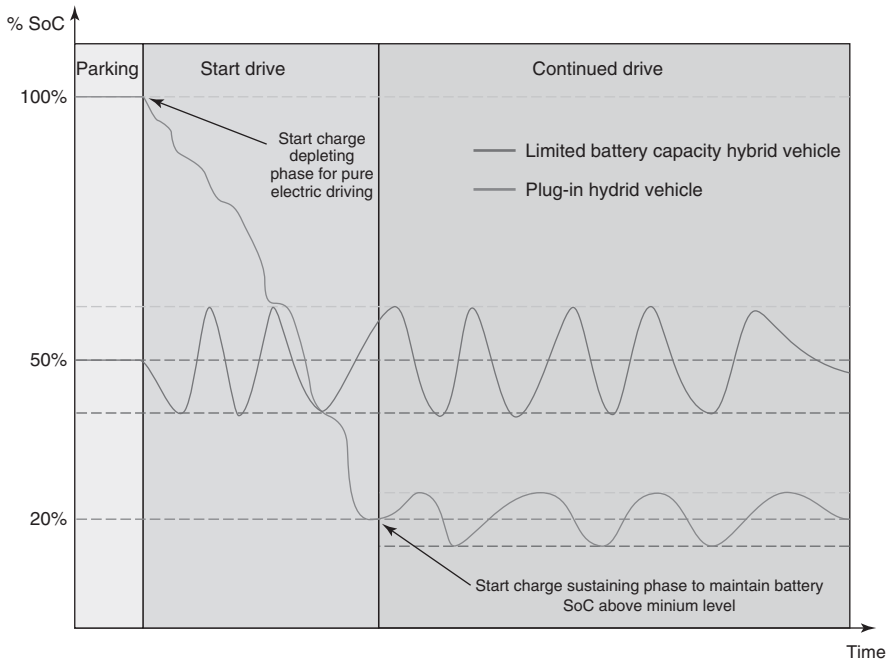


Figure 20. Evolution of SoC over time for a hybrid vehicle with limited battery capacity and a plug-in hybrid vehicle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ENDNOTES

1. http://www.opel.be/nl/Showroom/Ampera/eBrochure.aspx, November 2011.

REFERENCES

Abkemeier K. and Mize A. (2012) Hybrid and Electric Vehicles—The Electric Drive Captures the Imagination. *Implementing Agreement for Co-operation on Hybrid and Electric Vehicle Technologies and Programmes, Annual report of the Executive Committee and Task 1 over the Year 2011*.

Axsen, J., Burke, A., and Kurani, K. (2008) *Batteries for Plug-in Hybrid Electric Vehicles (PHEVs): Goals and the State of Technology Circa 2008*, Institute of Transportation Studies University of California UCD-ITS-RR-08-14, Davis, CA.

Bayindir, K.C., Gozukucuk, M.A., and Teke, A. (2011) A comprehensive overview of hybrid electric vehicle: powertrain configurations, powertrain control techniques and electronic control units. *Energy Conversion and Management*, **52**, 1305–1313.

Barrero, R., Van Mierlo, J., and Tackoen, X. (2008) Energy savings in public transport. *IEEE Vehicular Technology Magazine*, **3** (3), 26–36.

Chan, C.C. (2007) The State of the Art of Electric and Hybrid, and Fuel Cell Vehicles. *Proceedings of the IEEE, Special issue on Electric, Hybrid and Fuel Cell Vehicles*, vol. **95**, no. 4.

Chan, C.C., Bouscayrol, A., and Chen, K. (2010) Electric, hybrid, and fuel-cell vehicles: architectures and modeling. *IEEE Transactions on Vehicular Technology*, **59** (2), 589–598.

Chen, Y., Quan, L., and Zhu, X. (2011) An Overview of Double Power Flow Motor Used in HEVs. *International Conference on Electrical Machines and Systems*, Beijing, China 2011, 4 pp.

Ehsani, M., Gao, Y., and Emadi, A. (2010) *Modern Electric, Hybrid Electric and Fuel Cell Vehicles*. ISBN: 978-1-4200-5398-2

Gao, Y. and Ehsani, M. (2006) A torque and speed coupling hybrid drivetrain architecture, control, and simulation. *IEEE Transactions on Power Electronics*, **21** (3), 741–748.

Grammatico, S., Balluchi, A., and Cosoli, E. (2010) A series-parallel hybrid electric powertrain for industrial vehicles. *2010 IEEE Vehicle Power and Propulsion Conference (VPPC)*, Lille, France.

Hoeijmakers, M.J. (2003) Electromechanical Converter. WIPO 2003, WO 03/075437 A1.

Hoeijmakers, M.J. and Ferreira, J.A. (2006) The electric variable transmission. *IEEE Transactions on Industry Applications*, **42** (4), 1092–1100.

Hoeijmakers M.J. and Rondel, M. (2004) The Electric Variable Transmission in a City Bus. *35th Annual IEEE Power Electronics Specialists Conference*, pp. 2773–2778.

Kessels J.T.B.A. and van den Bosch P.P.J. (2009) Integrated Powertrain Control for HEVs with EVT. *Vehicle Power and Propulsion IEEE Conference 2009*, pp. 376–381.

Liu, J. and Peng, H. (2008) Modeling and control of a power-split hybrid vehicle. *IEEE Transactions on Control Systems Technology*, **16** (6), 1242–1251.

Miller, J.M. (2006) Hybrid electric vehicle propulsion system architectures of the e-CVT type. *IEEE Transactions on Power Electronics*, **21** (3), 756–767.

Salmasi, F.R. (2007) Control strategies for hybrid electric vehicles: evolution, classification, comparison, and future trends. *IEEE Transactions on Vehicular Technology*, **56** (5), 2393–2403.

US Department of Energy (2010) 2010 Toyota Prius-0462 Hybrid BOT Battery Test Results. Energy Efficiency and Renewable Energy.

Van Mierlo, J. and Maggetto, G. (2000) Views on hybrid drivetrain power management strategies. The 17th World Battery, Hybrid and Fuel Cell Electric Vehicle Symposium EVS-17, Montréal, Canada.

Van Mierlo, J. and Maggetto, G. (2001) Vehicle simulation programme: a tool to evaluate hybrid power management strategies based on an innovative algorithm. *Journal of Automobile Engineering*, **215**, 1043–1052.

Van Mierlo, J. and Maggetto, G. (2004) Innovative iteration algorithm for a vehicle simulation program. *IEEE Transactions on Vehicular Technology*, **53** (2), 401–412.

Van Mierlo, J., Van den Bossche, P., and Maggetto, G. (2004) Models of energy sources for EV and HEV: fuel cells, batteries, ultracapacitors, flywheels and engine-generators. *Journal of Power Sources*, **128** (1), 76–89.

Xizheng, G., Xuhui, W., and Xu, L., *et al*. (2009) Vibration Reducing in the Process of Engine Start Stop for Electric Variable Transmission. *IEEE 6th International Power Electronics and Motion Control Conference, 2009*, pp. 2001-2004

FURTHER READING

Chau, K.T. and Wong, Y.S. (2002) Overview of power management in hybrid electric vehicles. *Energy Conversion and Management*, **43**, 1953–1968.

Emadi, A., Lee, Y.J., and Rajashekara, K. (2008) Power electronics and motor drives in electric, hybrid electric, and plug-in hybrid electric vehicles. *IEEE Transactions on Industrial Electronics*, **55** (6), 2237–2245.

Shen, C., Shan, P., and Gao, T. (2011) A comprehensive overview of hybrid electric vehicles. *International Journal of Vehicular Technology*, **2011** Article ID 571683, 7 pp.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto205
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

UHC and CO Formation and Models

Miroslaw L. Wyszynski
University of Birmingham, Birmingham, UK

1 Sources of UHC and CO in an Engine 1
2 Models 2
3 Carbon Monoxide Oxidation Kinetics 6
4 Kinetics of Hydrocarbon Combustion 7
5 Typical Engine Emissions 9
6 Speciation of Hydrocarbon Emission 9
7 Summary 14
References 14

1 SOURCES OF UHC AND CO IN AN ENGINE

Unburned hydrocarbons (UHCs) and carbon monoxide (CO) are products of imperfect, that is, either partial or incomplete, combustion. As shown in Figure 1, other products of partial or incomplete combustion can include aldehydes and particulate matter. In addition, oxides of nitrogen are some of the most important by-products of high temperature combustion. Particulate matter is dealt with in Particulate Formation and Models and oxides of nitrogen in NO_x Formation and Models.

Ideally, products of stoichiometric combustion should be those of complete combustion as shown in Figure 1, while products of combustion of lean mixtures would also include unused oxygen. However, it is known that products of combustion in engines of a charge which on average is lean will include some UHCs and CO. This is partly due to the fact that for increasingly lean mixtures (even homogeneous), there will be an increasing occurrence of misfires as the mixture becomes leaner (Figure 2).

The other important reasons for UHC and CO emissions are (i) that mixtures are rarely completely homogeneous, (ii) that there are regions in the combustion chamber where temperatures will be much lower than average, and (iii) that some of the new combustion technologies, particularly the homogeneous charge compression ignition (HCCI) and similar technologies, are designed to lower the combustion temperature, largely in order to reduce nitrogen oxide emissions.

The first factor, that is, nonhomogeneity of mixtures was always the main source of emissions of particulate matter (that includes absorbed and adsorbed UHCs) from diesel engines. Recently, this problem also includes spark ignition (SI) engines with fuel injection, particularly, the gasoline direct injection engines. Problems with fuel droplet size distribution, spray penetration, evaporation, and mixing are increasingly solved by increasing the fuel injection pressures.

The second factor, that is, high production of UHCs and CO in regions of the combustion chamber where there is significant quench of flame or significant enrichment of the mixture, is best illustrated by the existence of regions such as shown in Figure 3.

A good example of evidence that UHC emissions in spark-ignited engines are caused among other factors by quenching near the walls was provided in an experiment by Tabaczynski, Heywood, and Keck (1972), with an explanation and results presented in Figure 4, quoted here after Ferguson (1986).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

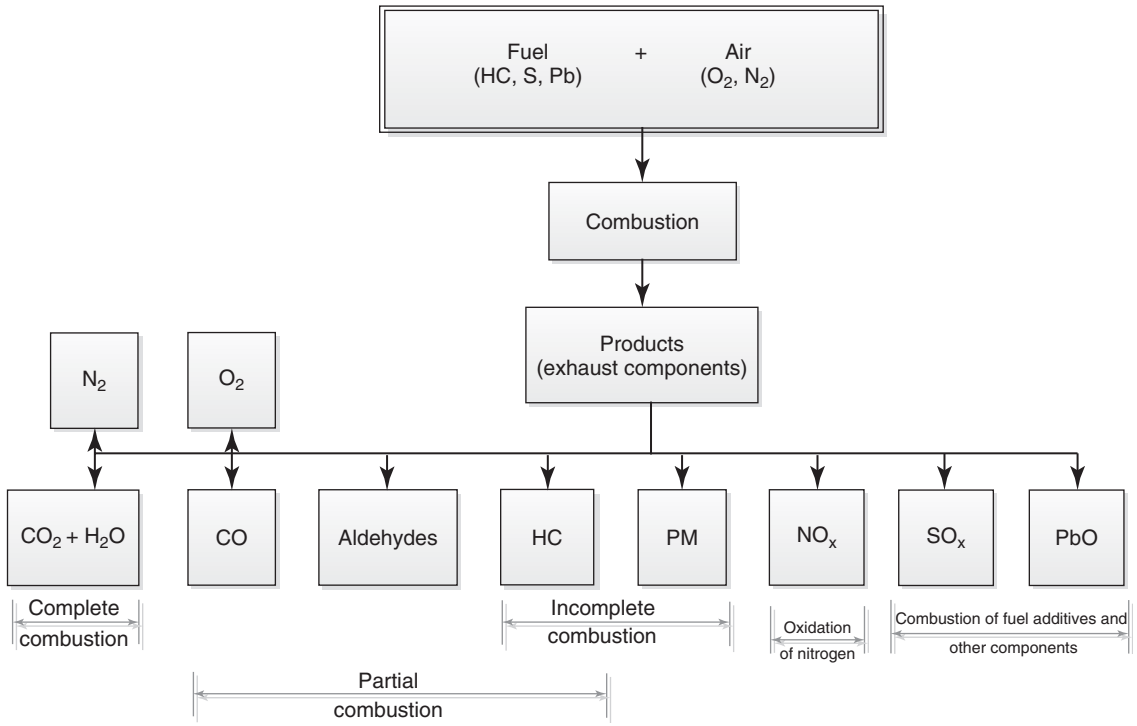


Figure 1. Typical exhaust gas components. (Adapted from Rychter and Teodorczyk, 2006. © Wydawnictwa Komunikacji i Łączności Sp.z.o.o.)

2 MODELS

There can be several levels of modeling of the quantity and formation mechanisms of UHC and CO, starting with (i) stoichiometric models that may include simple chemical equilibria for CO and NO*x*; through (ii) zero- and quasi-dimensional thermodynamic models of engine cycles with simple or detailed chemical kinetics for CO and numerous hydrocarbons (HCs); and ending with (iii) hugely computing-intensive computational fluid dynamics (CFD) models, including engine thermodynamic processes and detailed chemical kinetics in each cell of the domain; (iv) less-demanding simplified kinetics in CFD; or (v) probabilistic models.

2.1 Stoichiometric models with chemical equilibria

Low temperature (below 1000 K, as in the exhaust) equilibrium products of combustion of HC or oxygenated HC fuels with air at different mixture strengths yields mainly water and carbon dioxide as products, alongside the untouched nitrogen from air, as quoted below after Ferguson (1986), pp. 109–110.

For rich mixtures, there is some CO in the combustion products, and for high temperatures, formation of nitrogen oxides (covered in NO_x Formation and Models) has to be considered.

$$\phi\varepsilon C_\alpha H_\beta O_\gamma N_\delta + (0.21O_2 + 0.79N_2) \rightarrow \nu_1 CO_2 + \nu_2 H_2O + \nu_3 N_2 + \nu_4 O_2 + \nu_5 CO + \nu_6 H_2$$

where ε is the stoichiometric fuel/air ratio in molar terms,

$$\varepsilon = 0.21\left(\alpha + \frac{\beta}{4} - \frac{\gamma}{2}\right)$$

for example, for CH_4

$$\varepsilon = 0.21\left(1 + \frac{4}{4}\right) = 0.21 \times 2$$

ϕ is the fuel/air equivalence ratio (which is the reciprocal of air excess ratio λ, thus $\phi = 1/\lambda$), and molar coefficients ν_i give the product composition.

For lean and rich combustion, one can reasonably approximate as a first attempt:

$$\text{lean}: \quad \phi < 1, \quad \nu_5 = \nu_6 = 0$$

$$\text{rich}: \quad \phi > 1, \quad \nu_4 = 0$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

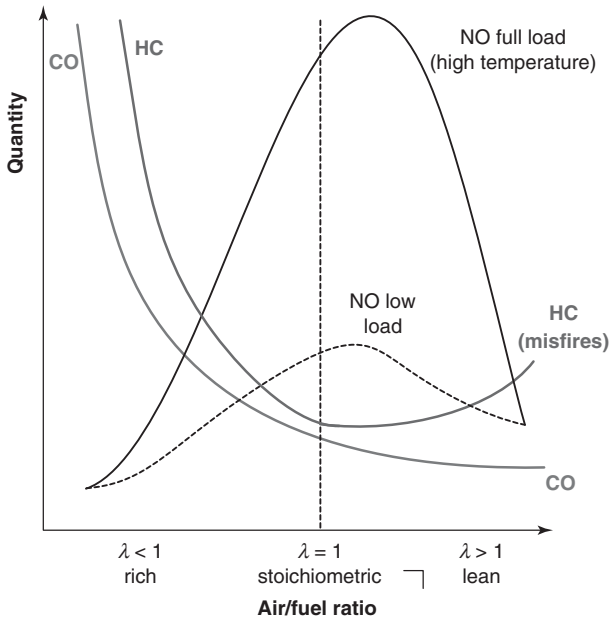


Figure 2. Relative quantities of emissions at different values of air excess coefficient λ.

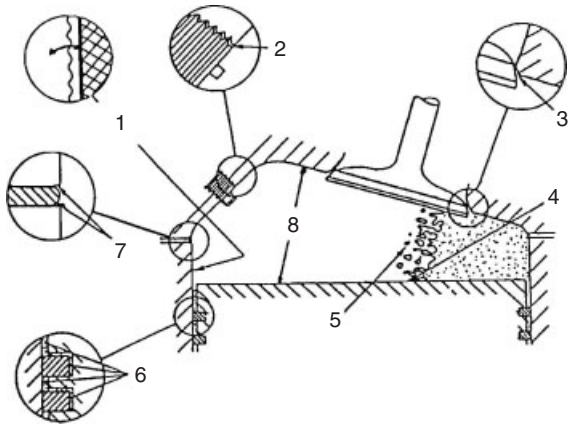


Figure 3. Sources of unburned hydrocarbons and carbon monoxide in a spark ignition engine (1) adsorption/desorption in/from lubricating oil, (2) space around the thread of spark plug/fuel injector, (3) space around the seat of inlet/exhaust valve, (4) flame quench near the walls, (5) flame quench in the area of flame front, (6) species trapped between the piston rings, (7) gaps around the head gasket, (8) adsorption/desorption from deposits on walls. (Adapted from Rychter and Teodorczyk, 2006. © Wydawnictwa Komunikacji i Łączności Sp.z.o.o.)

For the lean and stoichiometric cases, one can determine the product composition from atom balance as given in Table 1.

Rich combustion conditions are assumed as reasonably rich, that is, when there is enough oxygen so that solid carbon does not form, and all oxygen is divided in products between carbon dioxide and CO. In such cases, an important factor to consider is the equilibrium of the so-called *water gas reaction*.

For the rich case, the equilibrium constant K_{wg} of the water gas reaction:

$$CO_2 + H_2 = CO + H_2O$$

will give one more equation needed to determine the product composition:

$$K_{wg} = \frac{\nu_2 \nu_5}{\nu_1 \nu_6}$$

Molar coefficient of CO in the products ν_5, when the combustion reaction is written for 1 mol of air (as above) can then be calculated from:

$$\nu_5 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

where:

$a = 1 - K_{wg}$

$b = 0.42 - \phi\varepsilon(2\alpha - \gamma) + K_{wg}[0.42(\phi - 1) + \alpha\phi\varepsilon]$

$c = -0.42\alpha\phi\varepsilon(\phi - 1)K_{wg}$

The equilibrium constant fitted to the JANAF table data is given by:

$$\ln K_{wg} = 2.743 - \frac{1.761}{\left(\frac{T}{1000}\right)} - \frac{1.611}{\left(\frac{T}{1000}\right)^2} + \frac{0.2803}{\left(\frac{T}{1000}\right)^3}$$

where T is in degrees Kelvin (absolute temperature), with the fit applying to $400 < T < 3200$ K.

Having calculated the ν_5 coefficient for CO, the basic six components of combustion products (without taking into account the high temperature formation of nitrogen oxides) can now be explicitly determined from the available equations, and are summarized in Table 1.

Table 1. Low temperature (until 1000 K) combustion products (Hires *et al.*, 1976), as quoted below after Ferguson (1986), pp. 109–110.

i	Species	Lean and stoichiometric combustion $\phi \le 1,\ \lambda \ge 1$	Rich combustion $\phi > 1,\ \lambda < 1$
1	CO_2	$\alpha\phi\varepsilon$	$\alpha\phi\varepsilon - \nu_5$
2	H_2O	$\beta\phi\varepsilon/2$	$0.42 - \phi\varepsilon(2\alpha - \gamma) + \nu_5$
3	N_2	$0.79 + \delta\phi\frac{\varepsilon}{2}$	$0.79 + \delta\phi\frac{\varepsilon}{2}$
4	O_2	$0.21(1 - \phi)$	0
5	CO	0	ν_5
6	H_2	0	$0.42(\phi - 1) - \nu_5$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

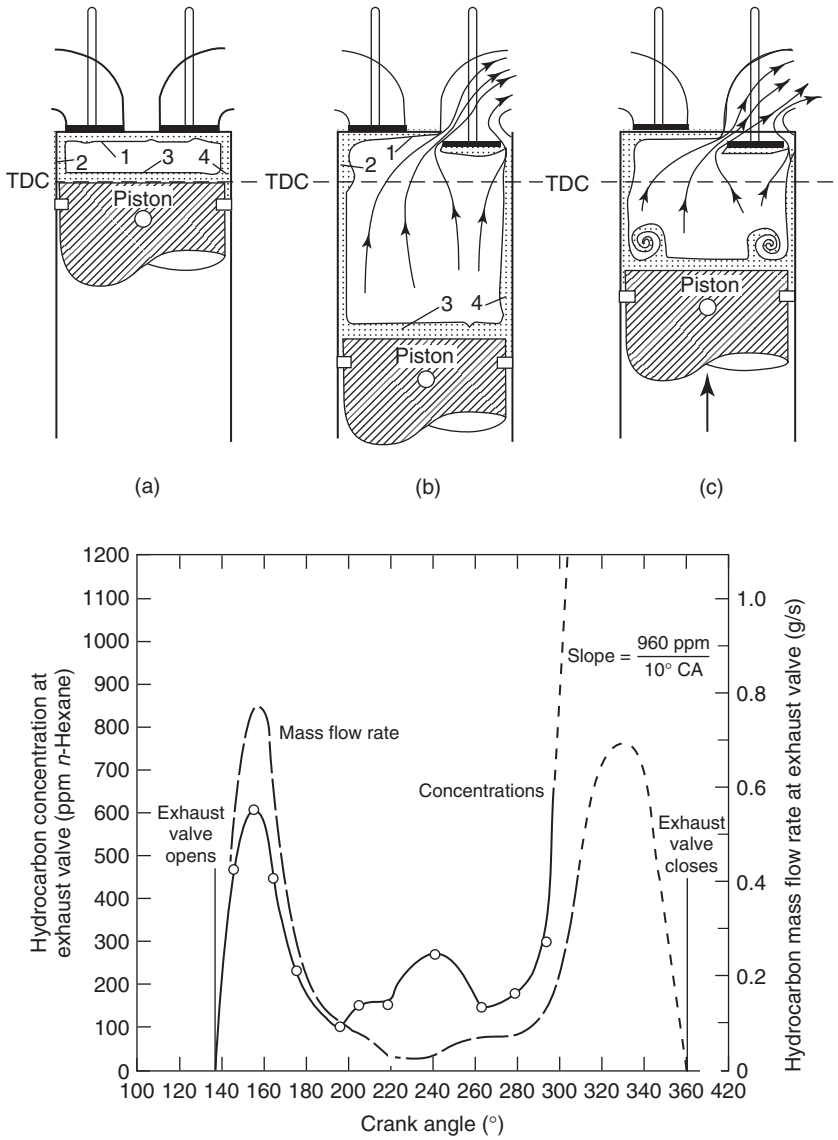


Figure 4. Important processes in hydrocarbon emissions. Unburned fuel is near the walls at (1), (2), and (3) and in the piston crevice area (4) as a result of flame quenching and absorption by oil layer (a). During the expansion in work, stroke hydrocarbons are desorbed and spread near the wall (b). During the exhaust stroke (c), hydrocarbons from regions (1) and (2) exit the cylinder with the main flow, and those from regions (3) and (4) roll up into a vortex which leaves at the end of the exhaust stroke. This is confirmed in the time-resolved measurement of hydrocarbon flow rate and concentrations during the opening of exhaust valve, as shown in the second diagram—large flow on opening the exhaust valve at 140–170° CA, just before BDC, and large flow again toward the end of the exhaust stroke near TDC at 300–340° CA.

Equilibrium combustion products involving consideration of many more chemical species and wider range of temperatures are now routinely performed using public domain software such as STANJAN (this seems to recently have gone commercial, but there are plenty of earlier free versions circulating in the engine community) or some other programs currently available online. There is an open-source code CANTERA http://code.google.com/p/cantera/, http://www2.galcit.caltech.edu/EDL/public/codes.html, and links to several equilibrium programs can be found at http://www.gaseq.co.uk/eqlinks.htm.

These programs are based on seeking a minimum of the Gibbs function, which is the basic thermodynamic condition of chemical equilibrium. Number of moles of individual atoms in the reactants (for combustion in engines, these normally include C, H, N, and O) and all the expected species in products have to be declared. The iterative solution then can proceed assuming

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

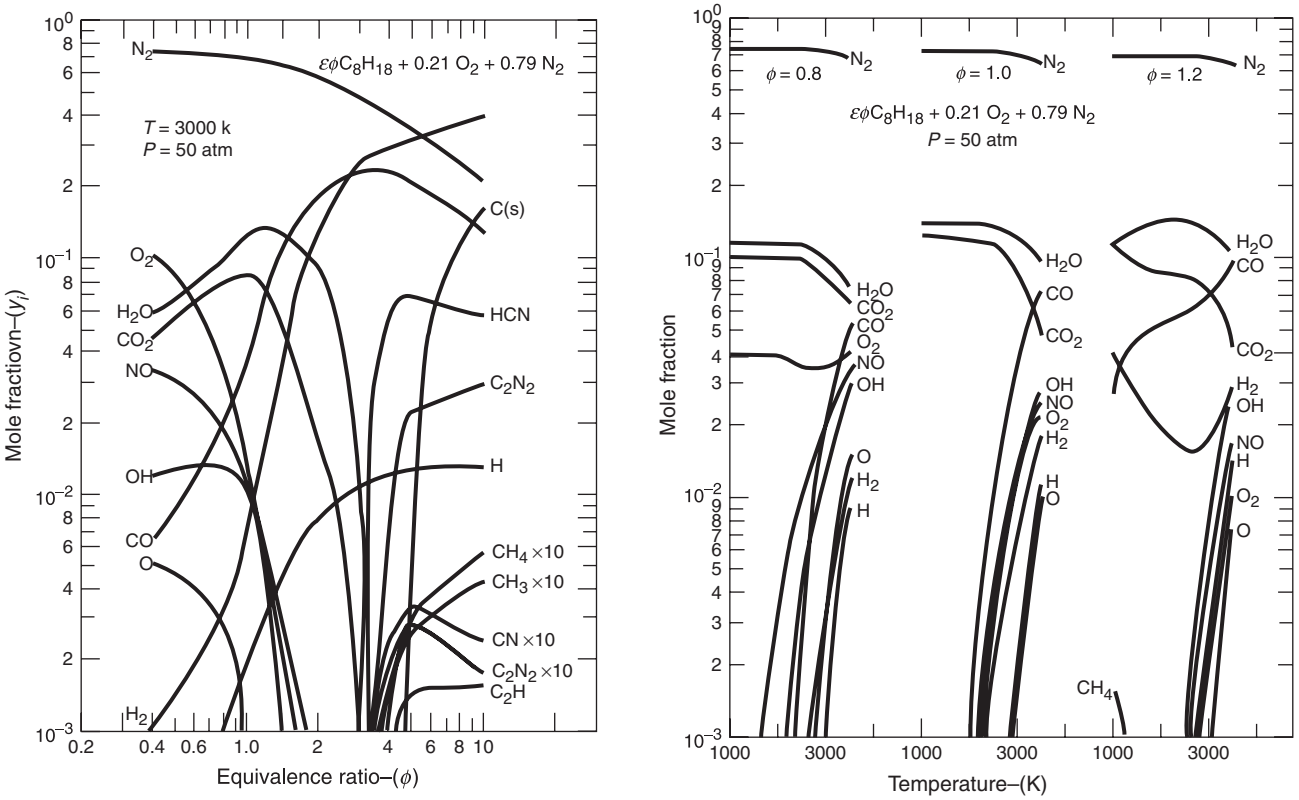


Figure 5. Equilibrium products for combustion of isooctane with air. (a) Temperature 3000 K and pressure 50 atm for a range of equivalence ratios, and (b) Pressure 50 atm, three values of equivalence ratio, and a range of temperatures in each case.

two of the defining thermodynamic properties such as temperature, pressure, enthalpy, and so on. The solutions shown in Figure 5, reproduced from Ferguson (1986) for combustion of isooctane with air are shown at (i) temperature 3000 K and pressure 50 atm for a range of equivalence ratios, and (ii) pressure 50 atm, three values of equivalence ratio, and a range of temperatures in each case.

Bearing in mind that vertical axes (mole fractions) are presented in Figure 5 on logarithmic scales, it is quite clear from Figure 5a that even at pressure and temperature typical for combustion in engines, there is very little CO predicted for lean mixtures (with equivalence ratio 0.4, it is 0.6%). For stoichiometric mixtures, CO is predicted around 3% and then it rises quickly to 30% at rich equivalence ratio of 2 and even to 30% at high equivalence ratios of 6 and beyond. UHCs appear with predicted equilibrium fractions over 0.1% for rich mixtures of equivalence ratio equal 3 and beyond. In addition, for very rich mixtures with equivalence ratio of 5 and beyond, there is predicted solid carbon C(s) forming. This partly explains the formation of soot in the rich regions of mixture, particularly around the core of evaporating and mixing liquid fuel sprays.

Figure 5b shows a very similar picture in a different configuration: for a lean mixture (equivalence ratio $\phi = 0.8$), the equilibrium-predicted CO starts to rise beyond 0.1% only at temperatures around 2200 K; for a stoichiometric mixture (equivalence ratio $\phi = 1.0$), this happens at temperatures over 2000 K and the fractions reached at higher temperatures are now larger; while for even a moderately rich mixture (equivalence ratio $\phi = 1.2$), several percent of CO are predicted at temperatures starting from 1000 K.

One HC predicted at noticeable equilibrium quantities is methane CH4 dropping to below 0.1% for temperatures rising above 1000 K (notice that most of the HCs predicted for very rich mixtures at 3000 K in Figure 5a are shown with values amplified 10 times).

2.2 Zero- and quasi-dimensional thermodynamic models of engine cycle with simple or detailed chemical kinetics for carbon monoxide and numerous hydrocarbons

Kinetics submodels take into account the progress of combustion reactions in time (as opposed to chemical

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

equilibrium models discussed above which assume that sufficient time is available for the reactions to reach their equilibria). In case of some very fast reactions, such as CO oxidation, there is little difference between the equilibria- and kinetics-controlled compositions. Quantities of some other species, particularly the oxides of nitrogen, have to be computed using the kinetics-controlled approach, as the time of residence at high temperatures is hugely important. Formation of oxides of nitrogen is covered in NO_x Formation and Models.

Reaction rates are usually described for the various reactions using the Arrhenius type relations:

$$k = C\,p^a\,T^b\,\mathrm{e}^{-E/RT}$$

where:

p is pressure, T is absolute temperature,
E is activation energy, and R is the gas constant
C, a and b are coefficients defined for the given reaction.

All these are in units defined in the individual data for a given reaction.

This approach requires the knowledge (or assumption based on an informed model) of the mechanisms of reactions for combustion of different fuels. These mechanisms can be hugely complicated and involve hundreds or even thousands of intermediate steps. Thus, the efforts of modelers for the last decades have been directed at two approaches: (i) generating (from first principles of physical chemistry) and setting up comprehensive chains of reactions that can be used in computer-intensive modeling, where computing time and effort is not of primary concern, and (ii) simplifying these models so that they contain as few reactions as possible without losing too much accuracy. The latter approach is useful in the combustion and emissions submodels of the thermodynamic models of engine cycles and in some more advanced CFD models of reactive flows within engine combustion chambers.

3 CARBON MONOXIDE OXIDATION KINETICS

CO oxidation is a very fast reaction, much faster than the time scale of gas-dynamic processes in the combustion chambers of piston engines. Because the detailed mechanisms of combustion are not well known, usually one can assume that concentrations of CO are determined by the equilibrium calculations as presented above.

3.1 Single global equation

If kinetics calculations for CO are to be conducted, it is recommended to use the following data for CO kinetics representing various burners and reactors, fuels, equivalence ratios, and pressures, which have been considered to be adequately correlated for practical purposes over the temperature range 840–2360°K by the global equation (Howard, Williams, and Fine, 1973):

$$-\frac{\mathrm{d[CO]}}{\mathrm{d}t} = k_0[\mathrm{CO}][\mathrm{O_2}]^{1/2}[\mathrm{H_2O}]^{1/2}\exp\left(-E/RT\right)$$

where [CO], [O_2], [H_2O] are molar fractions, with

$$k_0 = 1.3 \times 10^{14}\ \mathrm{cm^3/(mol \times s)}, \text{and } E = 30\ \mathrm{kcal/mol}$$

3.2 Comprehensive reduced 14-reaction model

A comprehensive 36-reaction kinetics model for CO combustion was developed and implemented into the ISIS thermodynamic engine model by Raine (2001). He has also presented a reduced 14-reaction scheme shown in Table 2.

The forward rate coefficients are presented in the form:

$$k_\mathrm{f} = A \times T^n \times \exp\left(-\frac{B}{T}\right),\ \text{where } B = \frac{E_\mathrm{a}}{R_\mathrm{u}}$$

where A is the pre-exponential factor, n is the temperature dependence exponent, E_a is the activation energy, and R_u is the universal gas constant.

Units are:

$$[A] = \left[\frac{\mathrm{cm^3}}{\mathrm{mole \times s \times K}}\right] \text{ and } [B] = [\mathrm{K}]$$

Values of the kinetic coefficients have been taken from Peters and Rogg (1993), where further details such as third body (M) efficiencies can also be found.

The backward reaction rates can be calculated through determining the equilibrium constants:

$$k_\mathrm{b} = \frac{k_\mathrm{f}}{K_\mathrm{eq}}$$

with equilibrium data taken from CHEMKIN (Kee, Rupley, and Millar, 1990) and JANAF (Chase *et al*., 1985) sources.

Initial CO concentration is calculated based on all carbon from the fuel being present as CO, even for stoichiometric and lean conditions.

The results of the CO kinetic modeling by Raine suggest that at very lean mixtures and with a homogeneous charge, the bulk gas will be only a small contributor to engine-out CO emissions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Coefficients in the Arrhenius type equations for the reduced 14-reaction model of carbon monoxide combustion.

Reaction	A	n	B
$H + O_2 \rightleftharpoons OH + O$	1.990e + 14	0.000	8460
$O + H_2 \rightleftharpoons OH + H$	5.110e + 04	2.670	3160
$H_2 + OH \rightleftharpoons H_2O + H$	1.020e + 08	1.600	1660
$4OH + OH \rightleftharpoons O + H_2O$	1.500e + 09	1.140	0
$H + O_2 + M \rightleftharpoons HO_2 + M$	2.000e + 18	−0.800	0
$H + HO_2 \rightleftharpoons OH + OH$	1.690e + 14	0.000	40
$H + HO_2 \rightleftharpoons H2 + O_2$	4.280e + 13	0.000	10
$OH + HO_2 \rightleftharpoons H_2O + O_2$	2.890e + 13	0.000	−250
$CO + OH \rightleftharpoons CO_2 + H$	4.390e + 06	1.500	−373
$OH + H + M \rightleftharpoons H_2O + M$	2.210e + 22	−2.000	0
$H + H + M \rightleftharpoons H_2 + M$	6.530e + 17	−1.000	0
$H + HO_2 \rightleftharpoons H_2O + O$	3.010e + 13	0.000	866
$CO + HO_2 \rightleftharpoons CO_2 + OH$	1.500e + 14	0.000	11910
$HO_2 + O \rightleftharpoons OH + O_2$	1.800e + 13	0.000	−204

Adapted with permission from Raine 2001. © Oxford University.

However, it is likely that with nonhomogeneous charge, the bulk gas will be a significant source of CO emissions from rich zones.

4 KINETICS OF HYDROCARBON COMBUSTION

For HCs, the situation is much more complicated for reasons of their complexity (normally typical gasoline or diesel fuel will have up to several dozen or several hundred components) and the multi-step reactions with mechanisms that can involve several hundred reactions even for simplest HC such as methane. For practical purposes of modeling combustion of HCs in engines and employing these models in a thermodynamic model of engine processes (including emissions of UHCs), there is a need to create a much simplified model, preferably involving just one global reaction. An early attempt is represented by the following rate equation, applicable to alkane HCs quoted here after (Rychter and Teodorczyk, 1990):

$$-\frac{\mathrm{d}[C_nH_m]}{\mathrm{d}t} = 6 \times 10^4 \times \gamma\, T\, p^{0.2} \exp\left[-\frac{12,200}{T}\right] [C_nH_m][O_2]^{0.5}$$

Where T [K] is temperature, p [atm] is pressure, $\gamma = 1$ for plug flow, and $\gamma = 80$ for fully stirred flow.

More complex set of mechanisms involving two-step reactions were formed based on observations that typical HC fuels burn in a somewhat sequential manner, that is, the fuel is partially oxidized to CO and H2, which persist until all of the HC species are consumed (Westbrook and Dryer, 1984). Thus, for methane, the two reactions model would be based on:

$$CH_4 + \frac{3}{2}O_2 = CO + 2H_2O$$

$$CO + \frac{1}{2}O_2 = CO_2$$

Similar considerations for other HC fuels with addition of other reactions lead to quasi-global reaction mechanisms which combine a single reaction of fuel and oxygen to form CO and H_2 together with a detailed mechanism for the $CO–H_2–O_2$ system. A table giving the coefficients for the two-step and quasi-global mechanisms is quoted here after Westbrook and Dryer (1984) (Table 3).

These values are to be used in single-rate expressions of the type:

$$k_{ov} = AT^n \exp\left(-E_a/R_uT\right) \ [\text{Fuel}]^a[\text{Oxidizer}]^b$$

with values of E_a in kcal/mol and quantities of fuel and oxidizer as molar fractions.

More detailed models and mechanisms have been presented in numerous studies that have been published over the decades on the subject of comprehensive modeling of HC combustion (and thus also formation of UHCs) (see Fundamental Chemical Kinetics). Some more recent examples of implementing the detailed chemistry into engine models include work from Cambridge (Coble *et al*., 2011), Aachen (Pitsch and Peters, 1998), and Lawrence Livermore (Curran *et al*., 1998) groups as well as numerous others published constantly. A comprehensive review of the detailed models was published by Simmie (2003). Recent approach to measurements and modeling of UHC and CO from a light-duty diesel engine was presented by Ekoto *et al*. (2009)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Coefficients for two-step and quasi-global mechanisms of hydrocarbon oxidation.

Fuel	Two-step mechanism				Quasi-global mechanism			
	A	E_a	a	b	A	E_a	a	b
CH_4	2.8×10^9	48.4	−0.3	1.3	1.1×10^9	48.4	−0.3	1.3
CH_4	1.5×10^7	30.0	−0.3	1.3	6.9×10^5	30.0	−0.3	1.3
C_2H_6	1.3×10^{12}	30.0	0.1	1.65	9.2×10^{11}	30.0	0.1	1.65
C_3H_8	1.0×10^{12}	30.0	0.1	1.65	7.2×10^{11}	30.0	0.1	1.65
C_4H_{10}	8.8×10^{11}	30.0	0.15	1.6	6.2×10^{11}	30.0	0.15	1.6
C_5H_{12}	7.8×10^{11}	30.0	0.25	1.5	5.4×10^{11}	30.0	0.25	1.5
C_6H_{14}	7.0×10^{11}	30.0	0.25	1.5	4.8×10^{11}	30.0	0.25	1.5
C_7H_{16}	6.3×10^{11}	30.0	0.25	1.5	4.3×10^{11}	30.0	0.25	1.5
C_8H_{18}	5.7×10^{11}	30.0	0.25	1.5	3.9×10^{11}	30.0	0.25	1.5
C_8H_{18}	9.6×10^{12}	40.0	0.25	1.5	6.0×10^{12}	40.0	0.25	1.5
C_9H_{20}	5.2×10^{11}	30.0	0.25	1.5	3.5×10^{11}	30.0	0.25	1.5
$C_{10}H_{22}$	4.7×10^{11}	30.0	0.25	1.5	3.2×10^{11}	30.0	0.25	1.5
CH_3OH	3.7×10^{12}	30.0	0.25	1.5	2.7×10^{12}	30.0	0.25	1.5
C_2H_5OH	1.8×10^{12}	30.0	0.15	1.6	1.3×10^{12}	30.0	0.15	1.6
C_6H_6	2.4×10^{11}	30.0	−0.1	1.85	1.7×10^{11}	30.0	−0.1	1.85
C_7H_8	1.9×10^{11}	30.0	−0.1	1.85	1.3×10^{11}	30.0	−0.1	1.85

Reproduced from Westbrook and Dryer (1984). © Elsevier.

Such detailed chemistry simulation technologies provide enhanced robustness and predictive power for modeling of combustion. Specifically, they strive to make it possible to predict the onset of ignition, low and high temperature heat release, local extinction, knocking, exhaust gas emissions including the formation CO and UHC, and so on. However, the adoption of detailed chemistry comes at a much greater computational cost. One of the methods that make it possible to retain computational robustness and ease of use while reducing computational timescales is the proposed use of the PDF (probability density function) model based on the stochastic reactor model (SRM) (Pitsch and Peters, 1998). It offers capabilities to account for in-cylinder processes such as chemical kinetics, fuel injection, turbulent mixing, heat transfer, and so on, while retaining in-cylinder stratification of mixture composition (i.e., fuel equivalence ratio) and temperature. Results equivalent to a non-PDF/SRM detailed chemistry model can be achieved with a reduction in computational time from 28 days to 10 min. In addition, modest parallelization of computing algorithms, requiring only 8–16 processor cores, can substantially speed up simulations using detailed chemistry. Reduced chemistry models dramatically cut the computational time, but only provide trends for emissions results. Parallelization is not really beneficial with reduced chemistry models, as these models can already be solved very quickly on a single-core PC.

When it is assumed that the ignition of diesel fuel can be described by using the single component model fuel *n*-heptane, a detailed chemical reaction scheme with about 1000 reactions among 168 chemical components could be replaced by a skeletal mechanism consisting of 98 reactions and 40 components, which is still capable of describing the autoignition process and concentrations of NO. Applying a flamelet model, it has been shown (Curran *et al.*, 1998) that the influence of the flow field and in particular the turbulence on the ignition process can be described by a single parameter, the scalar dissipation rate. Since the scalar dissipation rate can easily be evaluated from CFD calculations, if equations for the mean value and the variance of the mixture fraction are solved, the approximation formula can be used to estimate the departure of ignition in diesel engines from the homogeneous ignition delay times.

The key to understanding the reaction mechanisms is a careful and accurate description of the major chain-branching reaction paths and those kinetic processes that compete with the chain-branching paths. For example, it has been shown that low temperature chemistry is very sensitive to the formation of stable olefin species and to the chain-branching (Simmie, 2003).

The ultimate goal of chemical kinetic modeling is to develop an ideal set of thermodynamic data and a "perfect" reaction mechanism that will describe all the essential details of the physical reality, specifically the combustion of a HC in the gas phase. Great strides have been made of late in the computation of rate constants but it can be an acronymic nightmare for the unwary (Ekoto *et al.*, 2009). Initiatives that provide online computations of rate constants are most welcome and are a signpost for the future. Since a variety of experimental techniques is rarely, if ever, available in one laboratory, it reinforces the notion that cooperative research is essential. Modeling just your

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

own experiments with your own mechanism is scientifically worth very little (unless there is no other data of course) and one would hope that journal editors and referees would discourage researchers from such excessive introspection (Ekoto *et al*., 2009).

5 TYPICAL ENGINE EMISSIONS

Standard emissions of CO from engines are usually measured using nondispersive infrared instruments (NDIRs), whereas emissions of total UHCs are typically measured in engine research laboratories using the flame ionization detector (FID) method or the simpler NDIR geared to measure HCs based on a propane base (C_3H_8). The FID method relies in reality on measuring carbon atoms in HCs, whereas the NDIR method must be declared and calibrated to a specific base HC (usually propane). Typically, at the same time, emissions of nitrogen oxides are quoted and illustrate the trade-offs between good oxidation at high temperature (low UHC but high NO_x) and low temperature combustion at lower loads with higher HC emissions. The example below in Figure 6 (Zhong *et al*., 2010) compares three SI fuels—gasoline, ethanol, and the new experimental biofuel—2.5 dimethyl furan studied at Birmingham. One can see that emissions of CO and HC are lowest for the oxidated fuels, while emissions of NO_x are also lower for ethanol on account of its high evaporation enthalpy and lower combustion temperatures. It can also be seen from various literature sources reporting emissions from HCCI engines that the lower temperature combustion process leads to very low nitrogen oxides emissions, but higher HCs and CO emissions. These, however, are very easy to cope with in modern exhaust aftertreatment systems.

6 SPECIATION OF HYDROCARBON EMISSION

Speciation of the numerous emitted UHCs is a different and much more difficult matter. Much more sophisticated methods of separation and identification of individual HC are needed. This can be achieved by various techniques, all of those requiring extensive experience in chemical analysis. Probably, the oldest technique is the separation of HC on a gas chromatographic column and measurement on the standard flame ionization detector (called *GC-FID*), with the difference that now individual HCs can be measured as they elute from the column. With good calibration technique, a series of standard HCs in exhaust gases can be quantified. A new technique supposedly not requiring too

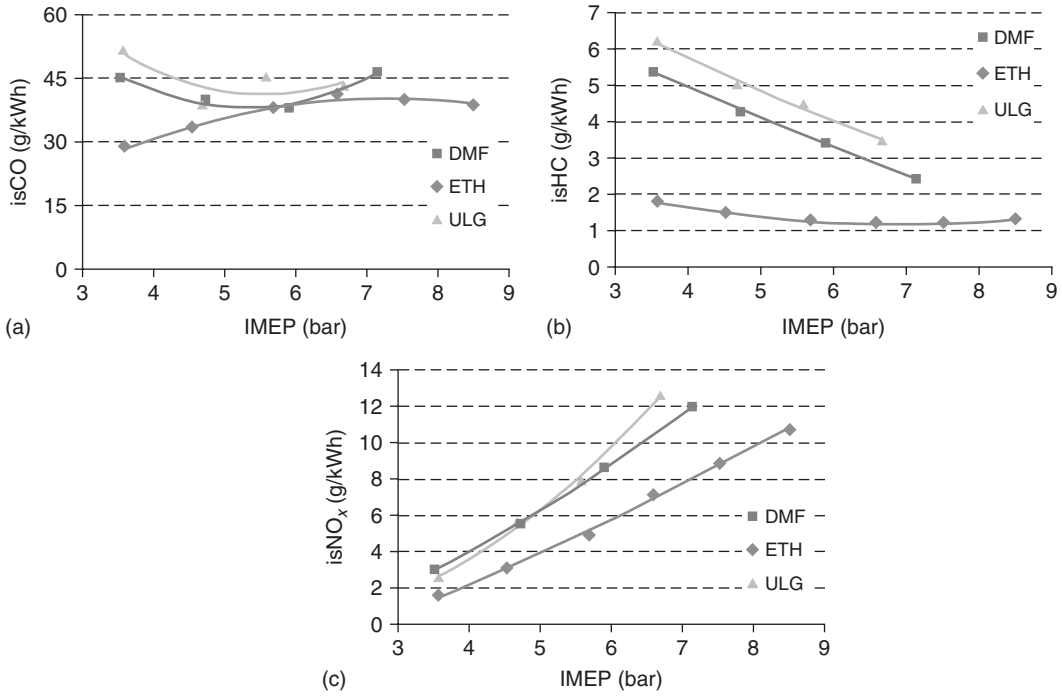


Figure 6. Indicated specific emissions: (a) CO, (b) HC, and (c) NO*x* under different IMEP values for DMF, ethanol, and gasoline at 1500 rpm and $\lambda = 1.0$. Single-cylinder pentroof head DI SI engine, at compression ratio 11.5.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

much experience is the fourier transform infrared (FTIR) detector, whereby individual spectra of HCs can be identified and with a good set of calibration gases can be quantified. These instruments have reached some maturity now and are increasingly used in different research laboratories, but still there are questions of stability, effect of water contamination, and so on. Perhaps, the most versatile method is the gas chromatograph–mass spectrometer (GC–MS) tandem, which allows identification of unknown compounds as they elute from the column on the basis of their mass spectra by comparison with vast libraries of chemical compound offered by the GC–MS vendors. After identification and selection of important species, a suitable set of calibration gases makes it possible to quantify the amounts. This method is, however, very much reliant on experience of the chemical analyst operator in selecting the GC columns and procedures and in the skill of separating and identifying the mass spectra. Heavy HCs in the exhaust, bound to particulate emissions and/or condensing from the exhaust gas, can also be separated by filtering, capturing on adsorption and absorption media, and extracting them to feed to a GC–MS procedure.

A number of the online and off-line speciation studies have been published, perhaps starting with the online GC–MS and real-time MS-only method pioneered in Birmingham in the early 1990s for measurement of individual aromatic HCs in the exhaust (Lehrle, West, and Wyszynski, 1995) and their changes under the effect of replacing some fuel with the equivalent energy fed in reformed fuel consisting largely of hydrogen and CO (Lehrle *et al*., 1993). Figure 7 presents the effects of improvement of combustion cleanliness demonstrated by the dramatic decrease in aromatic HCs selected for real-time observation using mass spectrometer fed directly with the exhaust gas (bypassing the gas chromatograph). The changes are quantified relative to the (assumed constant) argon content in the exhaust. The datum is the top left corner where no reformed gas has been introduced, and on replacement of even a small part of gasoline fuel with the equivalent calorific supply of the reformed gas fuel, one can see very large decrease in emissions of the most toxic aromatic HCs—benzene and methylbenzenes.

Recently, the particularly harmful vapor-phase and particulate-bound aromatic HCs, polycyclic aromatic hydrocarbons (PAHs), generated by a V6 gasoline engine working in SI and HCCI modes were collected and analyzed (Elghawi *et al*., 2010). All data were obtained during steady-state, fully warmed-up operation at different engine power levels (low and medium loads and mid-speed), and two different engine operation modes (SI

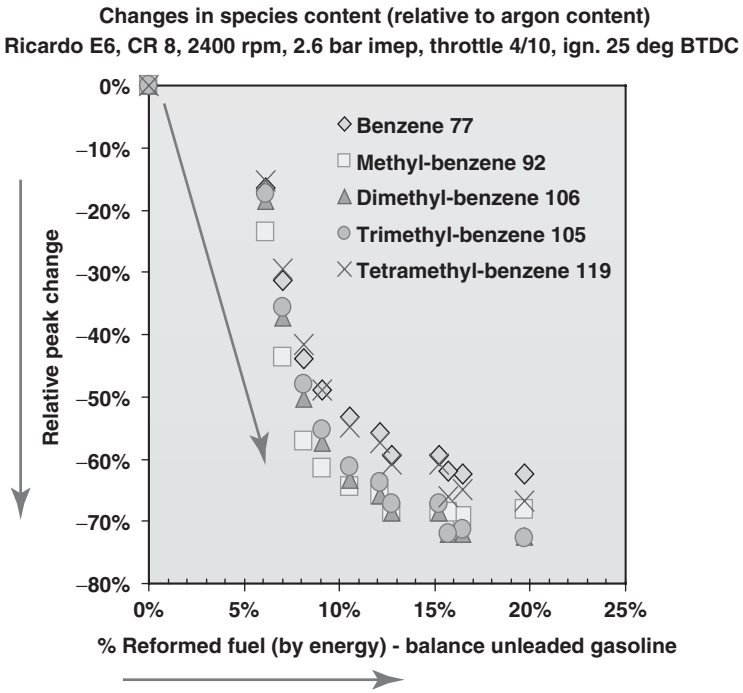


Figure 7. Effect of reformed fuel on individual HCs measured in real-time online using mass spectrometry (bypassing gas chromatograph). Gasoline operation, Ricardo E6 engine, reformed fuel: 23% H_2, 11% CO, 8.4 % CO_2, 11 % C_1–C_3 HC, and balance N_2. (Adapted from Lehrle, Wagner, and West 1993. © Lehrle, Wagner and West.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. The selected PAH compounds present in the exhaust.

Species Name	Abbreviation	Molecular Weight	Number of Rings	Molecular Formula
Naphthalene	NAP	128	2	C10H8
Acenaphthylene	ACY	152	3	C12H8
Acenaphthene	ACE	154	3	C12H10
Fluorene	FLU	166	3	C13H10
Phenanthrene	PHE	178	3	C14H10
Anthracene	ANT	178	3	C14H10
Fluoranthene	FLT	202	4	C16H10
Pyrene	PYR	202	4	C16H10
Benz[a]anthracene	BAA	228	4	C18H12
Chrysene	CRY	228	4	C18H12
Benzo[b]fluoranthene	BBF	252	5	C20H12
Benzo[k]fluoranthene	BKF	252	5	C20H12
Benzo[a]pyrene	BAP	252	5	C20H12
Dibenz[a,h]anthracene	DBA	254	5	C20H14
Benzo[ghi]perylene	BGA	276	6	C22H12
Indeno[1,2,3-cd]pyrene	IND	276	6	C22H12

Reproduced from Elghawi *et al*. (2010). © Elsevier.

Table 5. Mean PAH concentrations contained in the engine exhaust in $\mu g/m^3$ operated at different loads and 2000 rpm for the tested engine with winter grade commercial gasoline.

Compound	PAHs in HCCI Low Load	PAHs in HCCI Medium Load	PAHs in SI Low Load
NAP	740	1070	1440
ACY	70	76	78
ACE	40	45	48
FLU	BDL	26	38
ANT	BDL	33	57
PHE	BDL	17	31
FLT	BDL	9	23
PYR	BDL	BDL	11
Total	850	1276	1726

BDL—below detection limit.

and HCCI). The fuel used in this study was winter grade commercial gasoline fuel.

The vapor-phase exhaust gases were passed through stainless steel cartridges containing XAD-2 resin to capture PAHs. The PAHs were extracted from the resin with dichloromethane in an ultrasonic bath, and the obtained extracts were later analyzed qualitatively and quantitatively by GC–MS. The vapor-phase PAHs compounds observed from HCCI mode operated in low load were naphthalene, acenaphthylene, and acenaphthene only, whereas that obtained from SI mode under low load were naphthalene, acenaphthylene, acenaphthene, fluorene, anthracene, phenanthrene, fluoranthene, and pyrene.

The PAHs bound to particulates were trapped by using a complex of dilution tunnels with filter papers.

The soluble organic fractions (SOF) of the trapped particulates were separated from the insoluble fraction (ISF) with the help of ultrasonic elution, and analyzed by GC–MS method. The most abundant PAHs detected under selected operation conditions for HCCI mode was benzo[a]anthracene, followed by chrysene, then pyrene, and pursued by benzo[b]fluoranthene; in SI mode, under same operation condition, the highest PAH detected was benzo[a]anthracene followed by pyrene, benzo[b]fluoranthene, and chrysene. Probable mechanisms for the production of some of the pyrosynthetic PAH are discussed in Daniel *et al*. (2012).

Table 4 presents the list of poly-aromatic HC compounds found in the exhaust gas and selected for this study, with their molecular weight and number of aromatic rings. Table 5 presents the concentrations of the most abundant PAHs found in the exhaust gas from an engine fed with commercial grade gasoline and operated in two different modes (HCCI and SI) at two loads. It becomes clear that the HCCI mode (although it is generally believed to be prone to produce increased HC emissions) leads to greatly reduced emissions of PAH, which are lower than those from SI mode of operation even when comparing medium load HCCI with low load SI.

Another recent work (Daniel *et al*., 2012) presents the key individual HCs and carbonyls that have been identified using GC–MS and quantifies the emissions of 13 different carbonyls as specified by the California Air Resources Board (CARB) Method 1004 using high performance liquid chromatography (HPLC). The tests were conducted on a single cylinder direct-injection spark ignition (DISI) engine having compression ratio 11.5 at 1500 rpm, $\lambda = 1$, and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

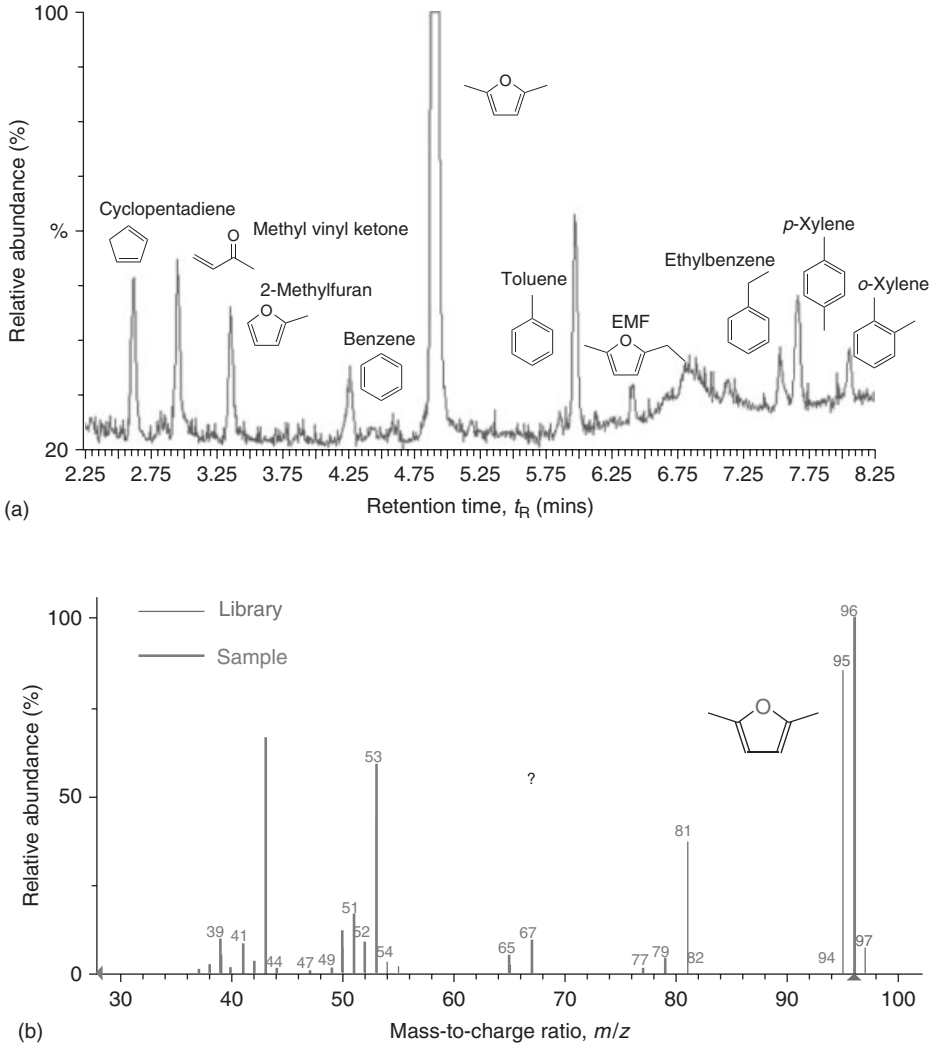


Figure 8. (a) Chromatogram of engine-out emissions using DMF (2,5-dimethyl furan as fuel. (b) Library and sample spectrum of DMF.

constant ignition timing. The fuels included 2,5-dimethyl furan (DMF), gasoline, *n*-butanol, ethanol, and methanol. For the GC analysis, the midrange HCs were identified using the mass spectra. Figure 8 shows a chromatogram of the engine-out emissions when the engine was fueled with DMF and the spectrum library used to identify the DMF. The results show that when DMF was used as fuel, unburned DMF dominates the emissions. Other main emissions include cyclopentadiene, methyl vinyl ketone, 2-methylfuran, and aromatics. There was no evidence of the emissions of linear alkanes except methane.

Figure 9 presents the individual carbonyl emissions when the engine was fueled with different fuels—2,5 DMF, gasoline, *n*-butanol, ethanol, and methanol. A carbonyl compound contains a functional group composed of a carbon atom double-bonded to an oxygen atom: C = O. A carbonyl group is present for example in the following types of compounds: aldehydes, ketones, carboxylic acids, esters, amides, and so on. A comprehensive analysis of carbonyls, their chemistry, and toxicity is beyond the scope of this work, but can readily be found in organic chemistry texts.

DMF produced the lowest overall low carbonyl (formaldehyde and acetaldehyde) emissions compared with methanol, ethanol, *n*-butanol, and gasoline and, more importantly, as seen in Figure 9, the lowest emissions of formaldehyde.

Figure 10 presents the higher carbonyl engine-out emissions for the cases of engine fueled with the three main fuels—ethanol, DMF, and unleaded gasoline. It should be noted that the levels of concentrations are an order of magnitude lower than those for low carbonyls presented in Figure 9. Here, the picture is more varied—ethanol generally produces lowest emissions and DMF is in most cases worse than gasoline. This may be caused by the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

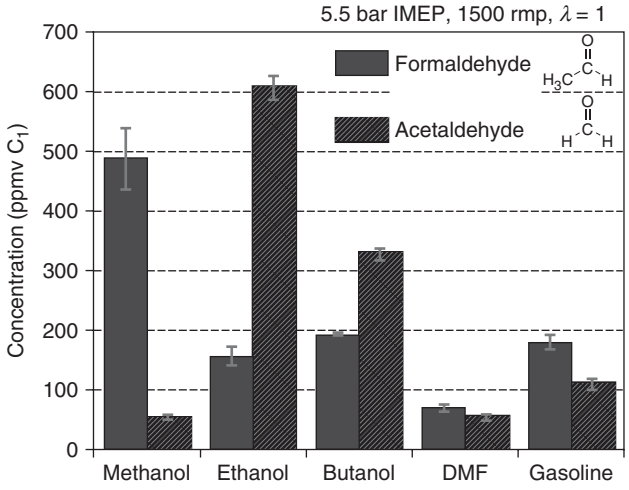


Figure 9. Formaldehyde and acetaldehyde engine-out emissions concentrations (C1-equivalent) using methanol, ethanol, *n*-butanol, DMF, and gasoline.

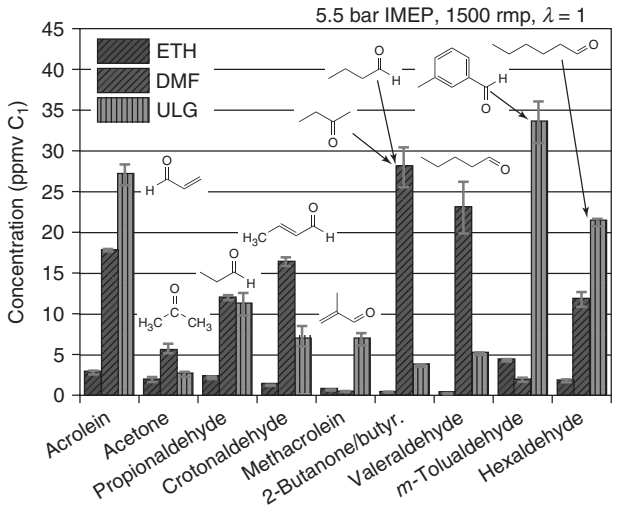


Figure 10. Higher carbonyl engine-out emissions concentrations (C1-equivalent) using ethanol, DMF, and gasoline. (Reprinted with permission from Daniel *et al*. 2012. Copyright (2012) American Chemical Society.)

composition of DMF which was not of a very high grade. An example of the variation of carbonyl proportions in engine-out emissions when fueled with the three main fuels is shown in Figure 11. It can be seen that formaldehyde and acetaldehyde are dominating the carbonyl emissions in all cases, with acetaldehyde being hugely dominant in the case of ethanol fueling.

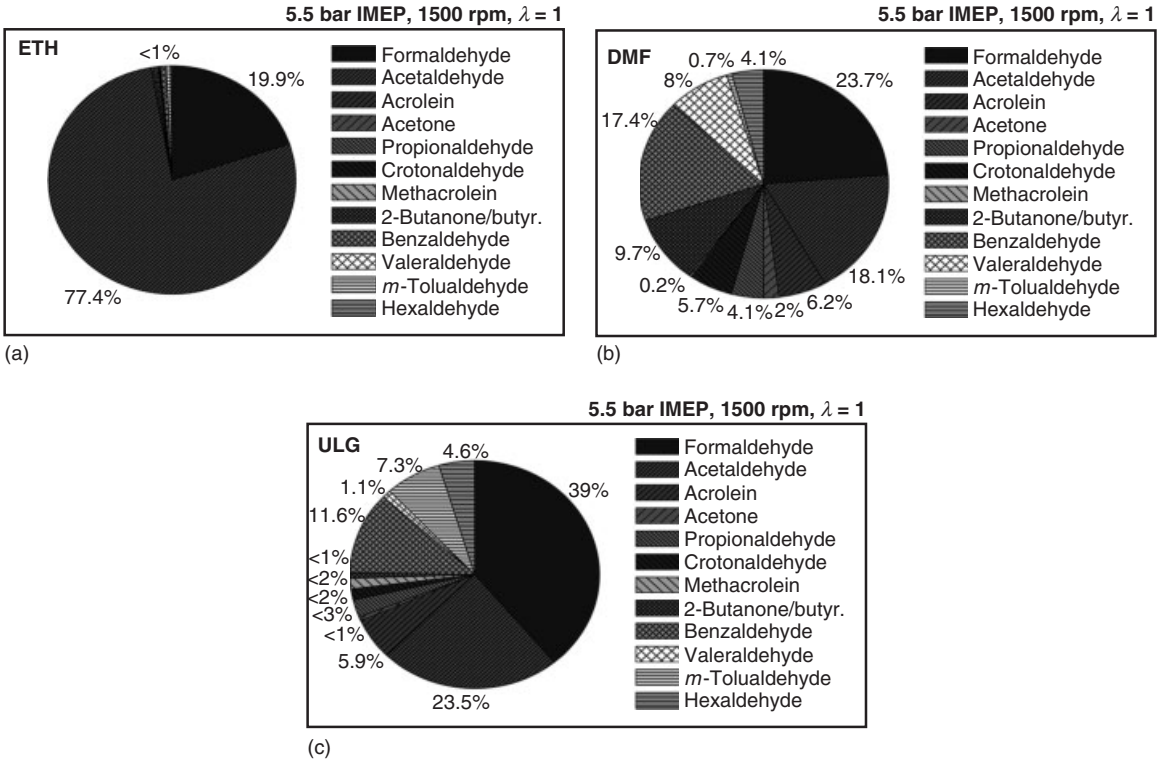


Figure 11. Variation of carbonyl engine-out emissions concentrations (C1-equivalent) using ethanol (a), DMF (b), and gasoline (c). (Reprinted with permission from Daniel *et al*. 2012. Copyright (2012) American Chemical Society.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

One more modern example of experimental study of UHC and CO emission sources is presented in Ekoto *et al*. (2009), where an optical, light-duty diesel engine operating under low load and engine speed was used, while employing a highly dilute, partially premixed low temperature combustion (LTC) strategy. The progression of in-cylinder mixing and combustion processes is studied using ultraviolet planar laser-induced fluorescence (UV PLIF) to measure the spatial distributions of liquid- and vapor-phase HC. A separate, deep-UV LIF technique is used to examine the clearance volume spatial distribution and composition of late-cycle UHC and CO. In-cylinder UHC and CO imaging highlights the differences that changes in dilution and load have on the main UHC source regions: (i) The cylinder center region contains intense near-injector fluorescence indicative of late-cycle fuel addition, whereas diffuse fluorescence is present from UHC and CO that is embedded in the surrounding fuel-lean bulk gases; and (ii) Squish volume UHC and CO principally results from the partial oxidation of lean mixture, although UHC from piston top fuel films and crevice flows is also observed (Ekoto *et al*., 2009).

7 SUMMARY

An introduction to understanding the origins of emissions of UHCs and CO from internal combustion engines, followed by a short review of the methods of modeling the formation of those pollutants has been presented. These modeling methods ranged from simple thermodynamics models involving stoichiometry and chemical equilibria through chemical kinetics of reactions to detailed chemistry of reactions, particularly for CO and some representative HCs. In the latter case, simplified reaction schemes, parallel multicore computing, and novel approaches such as probability density function analyses are particularly important to reduce the computation times while maintaining chemical accuracy and usefulness of the modeling. Some basic experimental methods of measuring engine-out emissions were introduced, while emphasis was put on determining the individual chemical species in engine-out emissions of HCs, particularly for the noxious and toxic aromatic and poly-aromatic HCs and carbonyls.

REFERENCES

Chase, M.W., Davies, C.A., Downey, J.R., *et al*. (1985) JANAF thermochemical tables. *Journal of Physical and Chemical Reference Data*, **14**(1).

Coble, A.R., A. Smallbone, A. Bhave *et al*. (2011) Implementing detailed chemistry and in-cylinder stratification into 0/1-D IC engine cycle simulation tools. SAE Technical Paper Series, **2011-01-0849**.

Curran, H.J., Gaffuri, P., Pitz, W.J., *et al*. (1998) A comprehensive modeling study of n-heptane oxidation. *Combustion and Flame*, **114**, 149–177.

Daniel, R., Wei, L., Xu, H., *et al*. (2012) Speciation of hydrocarbon and carbonyl emissions of 2,5-dimethylfuran combustion in a DISI engine. *Energy & Fuels*, **26**, 6661–6668.

Ekoto, I.W., Colban, W.F., Miles, P.C., *et al*. (2009) UHC and CO emissions sources from a light-duty diesel engine undergoing dilution-controlled low-temperature combustion, SAE Paper 2009-24-0043. *SAE International Journal of Engines*, **2**(2), 411–430.

Elghawi, U.M., Mayouf, A., Tsolakis, A., *et al*. (2010) Vapour-phase and particulate-bound PAHs profile generated by a (SI/HCCI) engine from a winter grade commercial gasoline fuel. *Fuel*, **89**, 2019–2025.

Ferguson, C.R. (1986) *Internal Combustion Engines, Applied Thermosciences*, John Wiley & Sons, New York, p. 546.

Hires, S.D., A. Ekchian, J.B. Heywood *et al*. (1976) Performance and NOx emissions modeling of a jet ignition pre-chamber Stratified charge engine. SAE Paper 760161. SAE Transactions, **85**.

Howard, J.B., G.C. Williams, and D.H. Fine (1973) Kinetics of carbon monoxide oxidation in postflame gases. Symposium (International) on Combustion. Fourteenth Symposium (International) on Combustion (1): pp. 975–986.

Kee, J.R., F.M. Rupley, and J.A. Millar (1990) The CHEMKIN Thermodynamic Data Base, *SANDIA National Laboratories Report: SAN87-8215B*.

Lehrle, R.S., Wagner, T., West, H. *et al*. (1993) Fuel reforming: potential for fuel economy and emissions, in *UnICEG Meeting, 20 December*. Coventry University, Coventry, UK.

Lehrle, R.S., West, H., and Wyszynski, M.L. (1995) On-line mass spectrometric characterisation of hydrocarbons in engine exhaust gases (paper D04593). *Proceedings of the IMechE, Part D: Journal of Automobile Engineering*, **209**, 307–324.

Peters, N. and Rogg, B. (eds) (1993) *Reduced Kinetic Mechanisms for Applications in Combustion Systems*, Springer - Verlag, Berlin.

Pitsch, H. and N. Peters (1998) Investigation of the ignition process of sprays under diesel engine conditions using reduced n-heptane chemistry. SAE Technical Paper Series, **982464**.

Raine, R.R. (2001) Thermodynamic Modelling of Combustion in a Spark Ignition Engine Report on EPSRC funded project, August 2000 to Feb 2001, Oxford University Internal Combustion Engines Group.

Rychter, T. and Teodorczyk, A. (1990) *Modelowanie matematyczne roboczego cyklu silnika tlokowego (in Polish) - Mathematical modelling of the working cycle of a piston engine*, Panstwowe Wydawnictwo Naukowe PWN, Warszawa.

Rychter, T. and Teodorczyk, A. (2006WKiL. 271.,) in *Teoria Silnikow Tłokowych (Theory of Piston Engines) - in Polish*. Series: Pojazdy Samochodowe (Automobile Vehicles) (ed. P.J. Myslowski), Wydawnictwa Komunikacji i Łączności Sp, Poland.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Simmie, J.M. (2003) Detailed chemical kinetic models for the combustion of hydrocarbon fuels. *Progress in Energy and Combustion Science*, **29**, 599–634.

Tabaczynski, R.J., J.B. Heywood, and J.C. Keck (1972) Time-resolved measurements of hydrocarbon mass flowrate in the exhaust of a spark ignition engine. SAE Transactions, **83**(Paper 72112).

Westbrook, C.K. and Dryer, F.L. (1984) Chemical kinetic model of hydrocarbon combustion. *Progress in Energy and Combustion Science*, **10**, 1–57.

Zhong, S., Daniel, R., Xu, H., *et al*. (2010) Combustion and emissions of 2,5-dimethylfuran in a direct injection spark-ignition engine. *Energy & Fuels*, **24**, 2891–2899.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto206
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Engine Performance

Anthony J. Martyr[1] and Eur Ing David R. Rogers[2]

[1]*Honorary Visiting Professor of Powertrain Engineering, University of Bradford, Bradford, UK*
[2]*AVL List GmbH, Graz, Austria*

1 Introduction 1
2 The Test Cell Structure, its Safety, and Environs 2
3 The Test Environment: Air 3
4 Combustion Air Treatment and Measurement 4
5 Fuel Treatment and Measurement 4
6 Engine Exhaust Handling 5
7 Engine Mounting 5
8 Engine Services Box 5
9 Dynamometers 6
10 Throttle Actuation 6
11 Electromagnetic Compatibility of the Test Cell 7
12 Test Cell Automation 7
13 Data Acquisition for Sensors and Transducers 10
14 Distributed I/O Systems 10
15 Integration of Intelligent Instruments 12
16 Combustion Analysis 12
17 Conditioning of Cylinder Pressure Signals 13
18 Special Purpose Test Cells 14
19 Data Management, Quality, and Security 15
References 15
Further Reading 15

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

1 INTRODUCTION

The product of engine test facilities is data; the quality and relevance of which depends on the standard of the test instrumentation, the design of the experiment (DOE), the degree to which control and measurement systems are matched and integrated, and the environment in which both instrumentation and the unit under test (UUT) operate. The test environment and the engine's loading must emulate, as closely and as repeatedly as possible, the real life or the legislatively prescribed conditions of the engine during any test sequence. The increasing industrial requirement to share and correlate experimental results across different test locations requires that all uncertainty in test variables having a material effect on engine performance and its measurement must be minimized.

The modern test facility is full of sensitive electronic devices that may have to operate in the same electromagnetic environment as powerful variable-speed electrical drive systems. This imposes an absolute requirement for the highest possible standard of electrical system layout design and detailed insulation. Problems caused by inappropriate cable routing, shielding, and the lack of equipotential (ground) bonding within the facility are difficult and time-consuming to cure postinstallation.

The international nature of engine development means that test data has to be shared and compared around the world; therefore, the subject of correlation of test results from different test facilities should be mentioned here. The empirical rules of test result correlation that the authors suggest are as follows:

- Repeatability of test results, while a good characteristic, is not a proxy for accuracy of results.

- The "golden engine" concept of checking test cell performance is a fallacy. No engine however well sealed and programmed can be used as a standard torque or gas-producing device; there are too many variables. The results of "round-robin" testing, whereby an engine is subjected to the same tests in many cells, are always disappointing. This holds true unless the engine is moved on a pallet on which it is completely prerigged with its fluid control systems and all mounted transducers, then tested with the same reference fuel and conditioned combustion air. Even then, the time wasted in chasing the causes of smaller and smaller variations in results will probably render the exercise a waste of time.
- Nothing is more important than regular and careful instrument calibration and diligence in avoiding physical or post-processing signal corruption.

2 THE TEST CELL STRUCTURE, ITS SAFETY, AND ENVIRONS

The engine test cell is subject to a wide range of safety and environmental regulations that vary only in detail worldwide. In addition to long established best operating practices, a test cell under all modern legislation is built and operated as a "hazard containment box." As an example, the European Union (EU) has enacted legislation that directly (*if sometimes unintentionally*) impacts on the design of engine test facilities, the most important of which may be the ATEX (EU 94/9/EC): (HSE) directives (Section 3). The New Machinery Directive (2006/42/EC) (European Commission) goes further with the concept of hazard containment, in that the cell structure tends to be treated as a "machine" the access to which has to be controlled, and guards within the cell (shaft guards) are treated as secondary protection.

Moreover, the walls and roof of the test cell should be of a substantial construction, providing attenuation of the engine noise at the adjoining workspaces and having at least a 1-h fire rating that is achieved by inclusion of automatic fire dampers fitted at all points of major cell penetrations (ventilation ducts).

Building codes and their local interpretation, covering the construction of engine test cells, varies around the world. In the United States, it has been a common practice to fit "blast panels" within an outside wall of the cell; whereas, in Europe, such features are only incorporated in cells using gaseous fuels; even then the panels are usually only in the exhaust ducting.

Environmental considerations on the test facility location concern both the effect of the facility on the environment and the environment on the facility. Depending on the type of testing being carried out, the location can cause deleterious effects on testing. Thus, proximity of emission laboratories to air pollution from traffic or industrial processes or noise, vibration, and harshness (NVH) facilities to major sources of vibration such as railroads is inadvisable. The impact of the facility on the local environment of a low number of engine test cells is minimal compared with the traffic densities surrounding them; however, an environmental hazard from a facility that can result in penal costs is fuel leakage into the ground water. Every precaution against such occurrences must be taken in the design and operation of bulk fuel storage and reticulation. Modern best practice dictates that buried fuel lines are of a double-walled design equipped with interstitial monitoring. Within facility buildings, fuel lines, made in welded stainless steel, are now run above head level where they can be easily monitored.

2.1 Size of test cells

Table 1 shows the typical sizes of four types of engine or powertrain test cells taken from actual European examples. Generally observed best practice is that a cell should be sized such that an operator has a clear walkway at least 1 m wide around the installed plant, although commonly this is not possible at the rear of the cell because of exhaust pipes running out of the cell. It is particularly important to leave sufficient clear space for calibration equipment, particularly for the dynamometer.

Table 1. Listing of the actual internal dimensions of different types of engine test cells.

Dimensions	Cell Category
6.5 m long × 4 m wide × 4 m high	Quality assurance (QA) test cell for small automotive diesels fitted with eddy-current dynamometer.
7.8 m long × 6 m wide × 4.5 m high	Engine control unit (ECU) development cell rated for 250 kW engines, containing work bench and some emission equipment
6.7 m long × 6.4 m wide × 4.7 m high	Gasoline engine development cell with AC dynamometer, special coolant, and intercooling conditioning
9.0 m long × 6 m wide × 4.2 high (to suspended ceiling)	Engine and transmission development bed with two dynamometers in "tee" configuration. Control room runs along 9 m wall.

(Reproduced with permission from Martyr and Plint (2012). Copyright Elsevier (2012).)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.2 Bedplates and seismic blocks

For normal development or quality assurance test cells, running engines in the light vehicle range, it is rarely thought necessary to employ vibration isolation techniques that require significant building excavation work, such as seismic blocks. The general best practice is to use cast-iron bedplates that are mounted on self-leveling "air springs" giving the combined mass system a natural frequency of around 3 Hz. Large bedplates fitted with a matrix of "tee slots" have the additional advantage over conventional bedplates of allowing the repositioning of both dynamometer(s) and the UUT to cater for different layouts, such as changing from in-line to transverse engine and exhaust layouts. Figure 3 (presented later in this chapter) will show such a facility.

2.3 Fire suppression

In spite of improved working practices over the years, fire remains a significant risk in test cells running engines, whatever their fuel. Indeed, the advent of common rail and direct injection engines has increased the occurrence of cell fires. On the detection of fire, the test cell control system should

- cut off the supply of fuel at the entry into the cell;
- shut down the ventilation system(s) and close fire dampers in the ventilation ducts;
- bring engine rotation to a stop and then shut off electrical power to the dynamometer.

Triggering of the extinguishing system often requires manual intervention in order to reduce disruptive false alarms. Systems based on CO_2 or conventional water sprinklers are not recommended anymore and several gas-based extinguishants are banned by environmental legislation. The system recommended in modern test cells is based on "water-mist," sometimes known as *micro-fog* technology. Not only do such systems have a very fast "knockdown" of fire, but importantly, they tend to wash the free-carbon particles out of the smoke and make subsequent cleaning a matter of days rather than weeks. Any chosen system will need to meet local regulation and building codes.

3 THE TEST ENVIRONMENT: AIR

The prime functions of airflow within the test cell are as follows:

- The removal of heat generated by the test equipment and the UUT. A "rule of thumb" for the airflow Q_A in cubic meters per second required to reduce the cell temperature by ΔT in degree Celsius where the estimated heat input is H_L in kilowatt is

$$Q_A = 0.84 \frac{H_L}{\Delta T} \quad (1)$$

- The prevention of the buildup of an explosive atmosphere by removal of hydrocarbon vapors. The design and operation of air-handling subsystems are prescribed under European ATEX regulations and represent worldwide best practice, as they form part of the test facility's primary safety system. This requires that the low point(s) of the test cells are purged of hydrocarbon vapor via an individual (nonsparking) purge fan for a set time before the engine is activated; this is achieved by interlocking with the cell control system. The cell must also be fitted with a hydrocarbon detection system made up of explosion-proof sniffers and connected to the primary alarm/safety system (Section 12). Note that the sniffers are the only explosion proof rated devices normally fitted, or required to be fitted in engine test cells. At a preset "alarm level" of hydrocarbons, the cell ventilation is increased, and at a higher level, the engine is shut down and a full purge sequence is activated.
- Combustion air for the engine under test is supplied by either
  - the engine ingesting the cell ventilation air or
  - a treated air supply independent of the cell ventilation system (Section 4).

To maintain consistent test results, particularly when the engine is ingesting cell air, the flow paths of air within the test cell should be consistent. The most common cause for inconsistence test results due to the variable influence of the test cell is the poorly controlled use of mobile "spot-cooling" fans. It is a best practice to control the ventilation within the test cell through balancing the inlet and outlet flows, such that the temperature within the cell is maintained below 40°C and the pressure below ambient by 50 Pa. The slightly negative pressure ensures that fumes from the cell are not forced into the control room. Using variable-speed fans on both the inlet and the outlet, the control system can control the outlet fan to control temperature and the inlet fan to control cell pressure. With this arrangement, when the air temperature rises, the speed of the outlet fans rises creating a cell pressure drop. The inlet fan then speeds up to balance the outflow. This control method ensures that during transient conditions of rising temperature, the cell pressure tends to be negative

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

rather than positive; thus, minimizing fumes being expelled into other work areas. Even with a pressure differential of only 50 Pa large, outward opening cell doors may be difficult to open and pressure surges because of badly adjusted systems constitute a hazard to staff because of doors either bursting open or effectively being locked shut.

4 COMBUSTION AIR TREATMENT AND MEASUREMENT

Variability in the content (pollutants), temperature, and pressure of the air ingested into the combustion chamber of an engine will produce commensurate variability in both power output and gaseous emissions. Therefore, high fidelity testing at most sites requires that combustion air is pretreated. *If correlation of test results is required, it is essential that all tests be carried out with common, standard conditions, at least, of air temperature and purity*.

Automotive Combustion analysis (CA) treatment units are widely commercially available that provide air at "standard conditions" of 25°C, fewer will control humidity of that air to 30% RH and fewer still will also control engine inlet air pressure. Best practice dictates that before specifying a CA treatment system, the required engine-test operating points of temperature and humidity are marked on a psychometric chart in order to describe the full operating envelope. It should be appreciated that the wider the operating envelope and the wider the range of ambient conditions are, the greater will be the capital and energy cost of the treatment module. Pressure control of CA adds a further level of complexity dependent firstly on the set point settling time following a change in engine state and secondly whether it is required to run at air pressures below local barometric.

While correction codes such as SAE J1349 are valuable for standardizing engine power outputs, to repeat or correlate engine emission measurements requires standardized air.

5 FUEL TREATMENT AND MEASUREMENT

The handling and treatment of liquid and gaseous fuels in bulk are subject to national and local safety and environmental protection regulations worldwide; facility management must take cognoscente of them. Best practice and regulation dictate that at the entry point of fuel into a test cell, fuel flow must be controlled by a "normally closed" solenoid valve that is interlocked with the emergency stop and fire alarm systems. The fuel consumption of modern engines, fitted with a variety of "spill-back" systems, cannot be measured by recording the mass or volume of liquid fuel supplied to the engine, as in the past, but rather the volume or mass of fuel leaving the metering system and before the engine's recirculation loop must be measured. A difficulty that arises is due to the heat picked up by the fuel in the pressure rail system and vapor that may have been created—both changing the volume and unit mass of the liquid. Figure 1 shows the type of circuit required to maintain temperature and degas fuel.

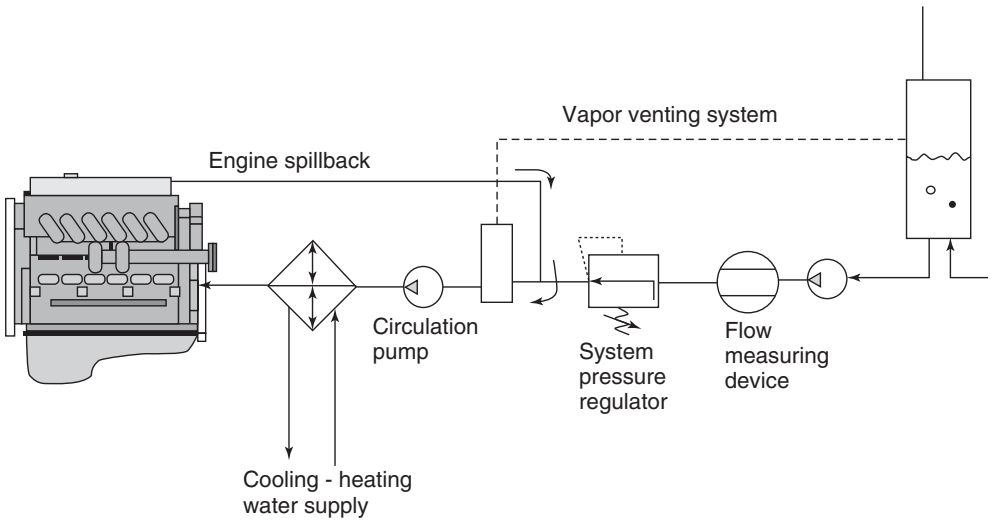


Figure 1. A circuit for measuring fuel consumption incorporating treatment for the spill back fuel from the engine. (Reproduced with permission from Martyr and Plint (2012). Copyright Elsevier (2012).)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6 ENGINE EXHAUST HANDLING

The modern automotive exhaust system with its post-combustion gas treatment modules and in one of the intended vehicle configurations has to be considered an integral part of the engine system as rigged in any test cell carrying out homologation and mapping work. The need to rig engines with a variety of exhaust configurations has led to a general increase in the size of cells. For such cases, the cell system must provide an extract duct working under negative pressure into which the last section of the exhaust pipe discharges. The fan that withdraws the exhaust, diluted with a proportion of cell air, should ideally be placed so that its discharge ducting, under positive pressure, is outside any enclosed workspace. Where a vehicle exhaust is not used, then the gasses are removed by direct coupling to pipework forming part of the cell system. This will include a short flexible section at the engine and suitable silencers (mufflers).

For both safety and operational reasons, modern best practice dictates that cells are provided with individual exhaust systems of whatever type, rather than using shared ducting. In the case of direct-connected exhaust systems, the practice is to provide oversized pipework fitted with a valve to regulate backpressure.

Finally, diluted exhaust ducts act as condensers and can produce highly corrosive condensates. The ducts must therefore be made from suitable stainless steel. Spiral wound galvanized duct must never be used in this situation. (*Note*: exhaust emission analysis equipment and operation are discussed in Exhaust Emissions).

7 ENGINE MOUNTING

7.1 Engine support mechanism

The support mechanism provided for the engine UUT depends largely on the frequency of unit change (the rig/run ratio). The requirement to obtain maximum possible use from increasingly expensive test facilities has meant that most operations prerig the UUT in support workshops on specially designed pallets (fork-lift transported) or trolleys. By providing a workshop rigging-stand that emulates the positions of critical interfaces (shaft height, exhaust flange, etc.) of test stand, unproductive installation time in the cell is minimized.

Vehicle-type engine mounts are widely used on pallet-rigged engines. However, owing to the different system dynamics, it should be noted that they do not always provide the correct degree of support or damping.

7.2 Connecting shaft

The connecting shaft is the rigging component that creates most frequent problems; shaft vibration may occur in a particular engine/dynamometer configuration when ranges of engines are tested with the same shaft. Connection shaft selection may require a detailed analysis of the system, a subject covered in detail in specialist publications such as "Engine Testing" fourth edition (see recommended Further Reading). The results of shaft failure, due to incompatibility with the torsional regime or poor quality of installation, are serious and will cause significant consequential damage unless properly controlled. Shaft guarding must be designed to retain shaft components in the case of failure and also prevent accidental human contact with rotating parts. However, it must be possible for the engine to flex on its mountings to the normal extent during running without making contact with any part of its enclosing guard; the presence of such clearance must be a standard prerun check. Modern safety practice requires that shaft guards are hinged for access but that the state of the guard system is interlocked with the cell control system.

7.3 Transducer cables

The transducer cables should be bunched and supported so far as is possible during the prerigging so that in the cell they are simply plugged into the cells transducer box. Transducer boxes (Section 8) that are close to, or above, the test engine are frequently force ventilated to keep internal temperature below 40°C.

8 ENGINE SERVICES BOX

The out of vehicle running of engines requires that the test cell is fitted with an *engine services box*, installed close to the UUT in order to provide electrical power to the engine control unit (ECU), power to the starter motor when required and other auxiliary units such as diesel heater plugs. A typical set of connections is shown in Figure 2. Note: KAP (keep alive power) terminals are for providing some ECUs with an unswitched 12 V DC supply, for preventing loss of datasets in volatile memory.

It should be noted that in addition to some ECUs requiring continuous power, it is highly advisable to fit the test cell control computer with an uninterrupted power supply (UPS) unit with a 1 h capacity.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

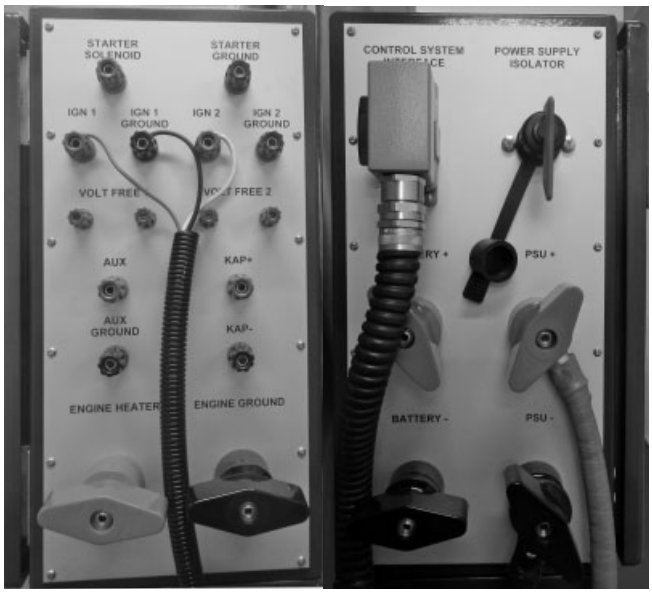


Figure 2. A composite photograph showing the two sides of an engine services box fitted within research cell testing both gasoline and diesel engines. (Photograph reproduced by permission of AVL-UK Ltd.)

9 DYNAMOMETERS

All dynamometers absorb and measure the torque produced by the connected engine while recording rotational speed of the dynamometer shaft. Thus, they can be considered as power measurement devices. Test routines require that the dynamometer both absorbs and contributes torque to the test system and is able to respond to demands for speed/torque changes quickly enough to simulate the quasi-real life drive cycles encoded in legislation and the requirements of control mapping.

AC motor-based dynamometers of low rotational inertia coupled with fast reaction pulse width modulating (PWM) drives have become the standard technology used in modern automotive test cells. It should be noted that, in addition to produce noise in the radio-frequency (RF) band (see later), AC drives contain expensive and delicate electronics; therefore, care should be taken in their position within the test facility and in their clean and reliable ventilation.

Less common, but still used dynamometers include direct current (DC), eddy-current, and water brake dynamometers. DC dynamometers are used in electromagnetic compatibility (EMC) test cells, where their lower RF signature is valued. The proven, absorb only, dry-gap, eddy-current dynamometers still retain a role in engine testing, where cost-effective running of engines under load is required such as in catalytic aging and endurance work. Water brake dynamometers are used for engines having power outputs beyond the economic range of AC machines such as use with medium speed diesels.

Torque measurement by modern dynamometers is achieved by one of the following two possible methods:

- By mounting the absorbing body of the dynamometer in trunnion bearings (cradle mounting) and measuring the torque reaction against the floor-fixed frame using a stiff solid-state strain gauge.
- By fitting a torque-measuring transducer at the dynamometer end of the shaft system connecting the engine. Such devices take the form of either a torque shaft fitted with its own bearings or a torque flange.

There are a number of subtleties in the choice of torque measurement systems but all must be capable of on-site calibration, usually using weights that apply a known force on an arm of known length.

Rotational speed is measured either by a toothed wheel and proximity pick-up, as is commonly used on eddy-current dynamometer, or, in the case of AC machines, by optical encoders. The engraved disk on the optical encoders can be produced so as to record not only instantaneous speed but also shaft position and direction of rotation.

10 THROTTLE ACTUATION

Almost all modern automotive engines are “drive-by-wire” meaning that the fueling control (for diesel engines) or “throttle” (for gasoline engines) is achieved by sending an electrical demand to the ECU. While it is possible to simulate that demand signal via an analog output from the test bed control computer in the test cell, it is rarely done this way. Owing to the complexity of this safety critical vehicle system, it is best practice and easier to use a complete vehicle accelerator pedal system directly wired into the engine loom, as in a vehicle. The pedal unit, rigidly mounted in a support frame, is operated via a Bowden cable by a fast-acting actuator under the control of the test bed controller. Setting up the zero to wide open throttle (WOT) position during engine rigging should be straightforward in a well-designed system and the shifting force should be mechanically limited. In all cases of engine rigging, particularly when intermediate linkages are used between actuator and pedal unit, backlash must be eliminated; otherwise, stable control will never be achieved.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

11 ELECTROMAGNETIC COMPATIBILITY OF THE TEST CELL

The automotive powertrain test cell contains a number of powerful emitters of electromagnetic energy encompassing a very wide frequency spectrum. Such an environment is capable of corrupting both data and control systems unless the greatest possible care is taken in the design, particularly in the installation of all cables within the test facility. Radiation in the RFs is an inherent feature of the pulse-width-modulating drives associated with AC dynamometers. This source of interference has replaced the harmonic distortion of the AC supply associated with older DC dynamometer systems and is commonly dealt with by powering the machines through a dedicated isolating transformer of the correct capacity.

The two key EMC requirements of a test facility are as follows:

- The whole building—all its services and all the installed test equipment should be connected to an equipotential grid; meaning that the earth (ground) potential is common throughout the facility and at the lowest possible impedance.
- The upmost care is paid to the layout and separation of data, control, and power cables. Signal corruption, through inductive, capacitive, or conductive coupling interference is very difficult to correct in carelessly installed sites, where underfloor trays are full of coils of over-sized or inappropriately bunched cables (Figure 3).



Figure 3. An AC dynamometer-equipped engine test cell containing a large cast-iron tee-slotted bedplate that is supported on air-springs, allowing multiple configurations of test equipment. Engine supported within a wheeled trolley allowing extensive prerigging outside the cell. Low level, rectangular "purge duct" can be seen on the right hand side of the photo next to an, circular section, exhaust dilution and extract duct. (Photograph reproduced by permission of AVL List GmbH.)

12 TEST CELL AUTOMATION

The engineer working within a test cell facility will have to understand the functions and interactions of four separate but interlinked control systems:

- the building management system (BMS);
- the test bed control and data acquisition systems;
- control systems embedded within "intelligent" devices such as AC drives and instrumentation such as combustion analysis and emission analysis equipment;
- the engine or powertrain control unit (ECU). Communication hardware with ECU can range from simply providing the "missing" digital I/O required to run the engine (wireless ignition key signal) to access required to map the control parameters.

12.1 Building management system

The BMS is a programmable logic controller (PLC)-based control system through which all of the facility services are started in a predetermined sequence, monitored, and shut down. Such systems range from the simple operation of the motor control panel, providing power to ventilation fans and fluid pumps, to highly complex building energy control systems. However, the correct interaction logic between the BMS and the test cell control system is vitally important to avoid test work being ruined by inappropriate shut downs triggered by building conditions irrelevant to the test cell in question. A key design and operational document for a test facility is the safety interaction matrix (SIM), parts shown in Table 2

The creation of a SIM is a key task in the system integration process because it requires the design team and cell operator to consider, record, and codify the system reactions in the case of every defined change of state or alarm condition of connected modules. Every SIM has to meet the operational preferences of the owner, within the restraints of safety legislation, and will be specific to the system for which it is designed.

12.2 Test bed control and data acquisition systems

The test bed control and data acquisition systems are based on powerful personal computers running specialized

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Parts of a typical safety interaction matrix that designates the interaction between control systems during normal and alarm states within the test facility.

Test cell safety matrix	Cell control system (CS)	400 V AC electrical supply	Test cell power sockets	Ventilation fans	Ventilation fire dampers	Combustion air system	AC dynamometer	Engine control	Test cell incoming fuel solenoids	Compressed air system	Other systems
Controlled by	CS	ESR	ESR	BMS	BMS	BMS	CS	ESR	ESR-DO	ESR-DO	
Event											
Main building fire alarm	No reaction	Enabled	Enabled	Manual shutdown of test cell, via "fast (regen) stop," before facility evacuation							
Test cell fire system (automatic)	Stop	Enabled	Disabled	Disabled	Closed	Disabled	Stop	Stop	Closed (off)	Closed	
Level 1 HC/CO gas alarm	Message display	Enabled	Enabled	Purge sequence	Open	Enabled	Power on	Enabled	Open (on)	Open	
Level 2 HC/CO gas alarm	Fast stop	Enabled	Enabled	Purge sequence	Open	Power off	Regen stop	Stop	Closed (off)	Open	
Manual emergency stop (remote from desk)	FAST STOP	ENABLED	Disabled	Disabled	Closed	Disabled	Regen stop	Stop	Closed (off)	Closed	
Other events											
Test cell doors opened during engine test	Fast stop or idle[a]	Enabled	Enabled	Freeze vent	Open	Enabled	Stop or idle[a]	Stop or idle[a]	Open	Open	
Test sequence stage alarm	Message display	Enabled	Enabled	Enabled	Open	Enabled	Power on	Enabled	Open	Open	
Other events											

Abbreviations: BMS, building management system; CS, (cell) control system; DO, digital out (from control system); and ESR, emergency stop relay.
[a]The action taken when the cell door is opened during engine running has to be determined by the operator via a risk analysis; the default options are either "engine stop" or "engine to idle." Ventilation freeze means that the system is held at the running status at alarm tripping.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

software. The prime functions of the test bed controller are to provide the following:

- A suitable mechanism to write and execute automated test sequences, control and record the engine's speed, and load together with all of the controlled variables provided by the cell's services control systems. Systems must be at least capable of running and recording the data required by modern legislated tests, such as the FTP75 and the NUDC drive cycles. Typically, the control loop times of modern digital systems will be >1 kHz. The data acquisition rates can be of the same order but should be adjustable to rates appropriate to the measurement being made.
- The ability to control the power produced by the engine by controlling the throttle (fueling) and load imposed on the engine by the dynamometer. This must be done so that the demanded parameters are achieved, within the capability of the engine, which may itself exhibit unstable characteristics, both during steady-state running and transient conditions between demand points. In addition, most modern controllers should be capable of direct PID or PI control of external devices such as coolant temperature control units. A suitable interface and system for connection and marshaling of sensors and transducers measuring temperatures, pressures, and other analog values around and within the UUT. Many of the raw data streams will have to be subjected to signal conditioning, such as analog to digital conversion, with the minimum possible data degradation before being transmitted via an error-tolerant high speed bus. The most common high speed bus is currently CAN (controller area network), as defined by ISO 11898.
- Recording (storage) of computer data in a suitable electronic file format for subsequent data analysis, display and retransmittal, and unique labeling of each data channel from transducer to stored data. Note that without a consistent channel identification convention being imposed by test engineers, over all cells involved in a common exercise, there is no possibility for realistic comparison or correlation of test results.
- Suitable interfaces for appropriate "intelligent" measurement devices, their remote control, monitoring, and data recording.
- A calibration system for all of the directly coupled transducers that provides a certification audit trail that meets the requirements of an ISO 9001 quality system.
- An interface that meets the conventions laid down by ASAM (Association for Standards in Automation and Measurement). These interface specifications, such as ASAP-3, ASAM MCD3, are widely adopted in automotive industry.
- A primary programmable safety system interfacing with any BMS fire and gas alarm circuits. The continuous monitoring system should be designed to react to at least two levels of alarm state: warning level and shut down level. The primary alarm levels should be set to protect the integrity of the equipment and the secondary alarms (test stage specific) to protect the UUT or the integrity of the test sequence.
- A mechanism for management and storage of test sequences and cell settings to allow for reuse and traceability of test sequences and cell settings.

At the leading edge of development work, the hardware and software forming powertrain test cell control and data acquisition systems are the signature product of a few internationally known companies. Such systems are of necessity, complex and continually updated as user requirements and experience is embodied in updates. The choice of which system to use has important implications to the efficiency of the facility. However, training in the use of such complex and expensive systems, both at basic operator and advanced engineer levels, is a fundamental requirement if full return on the investment is to be made.

12.3 Engine or powertrain ECU

The engine or powertrain ECU that controls all modern units requires calibration (mapping) in the engine's development phase such that the powertrain performance meets the requirements of the chosen user profile while at the same time meeting all the relevant legislative requirements. In addition to the basic task of controlling ignition and fueling over the full range of operational and environmental conditions within a permitted legislative envelope, the ECU also has to be loaded with the logic required by the on-board diagnostic (OBD) systems.

Once calibrated, the production version of the ECU sets a problem for all test engineers downstream of the calibration cells that have to run the engine when some of the ex-vehicular components required by the software are lacking. In these situations, various levels of communication with the ECU are required. The ECUs used in vehicles are made by a limited number of manufacturers worldwide who do not permit access to the lowest levels of the embedded software within, even to their OEM customers. Therefore, calibration engineers, who require the ability to read and write data to the control maps of the ECU, have to work with specially prepared, development versions of the ECU and use one of the software tools supported by various specialist companies.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The software tool is generally known as an *ECU application system* and allows access to measurement and calibration data from within the ECU processor and memory for the purpose of online optimization of calibration data variables. For the calibration task, the development ECU must be equipped with two important features:

- Additional memory for handling calibration variables during run time, known as *CalRAM*;
- An interface that will allow manipulation of these variables online, as well as access to measurement dataset variables from within the ECU processor. This is the interface to the application system and is defined by ASAM standards (ASAM) (MCD 1 and 2).

The development ECU also facilitates "switching off" of production-related functions—for example, operation of the ECU outside the connected vehicle network, disabling of on-board diagnostic functions, and disabling of related functions within the ECU (allowing "flat" maps to start the calibration process).

The physical interface (or connection) to the development ECU can be executed with one of the following two methods:

- *Emulator*—an additional processor board, often installed inside the ECU case, that has direct access to the ECU memory and data bus. This device has additional calibration memory on-board, as well as a high speed interface to the application system via a signal processor.
- *CCP* (*CAN Calibration Protocol*)—using an additional driver in the development ECU processor, allows access to measurement and calibration data via a CAN port on the ECU (either the existing vehicle CAN port or a dedicated calibration CAN port). The ECU must also be equipped with additional memory space, or "ring-fenced" memory, for manipulation of calibration data online, that is, CalRAM.

The emulator is more difficult and expensive to implement; however, where many levels of map and measurement values need to be manipulated, the emulator provides more memory and much faster communication to the application system. It is therefore more often applied to development ECUs that will be employed earlier in the development process of the UUT. CAN-based interfaces have memory and bandwidth limits, but are cheaper and simpler to implement, and are therefore more appropriate to the later stages of the development process.

13 DATA ACQUISITION FOR SENSORS AND TRANSDUCERS

One of the primary requirements for the test bed system is to acquire and store data from a wide range of types of sensors and transducers on and around the UUT. Table 3 lists the common measurements made in automotive test cells and the type of transducers used.

Most, but not all, of the transducers listed can be calibrated against a reference signal by the operators of the test facility as is required by any quality certification system.

Many operators who give a great deal of thought about the specification of individual transducers may tend to ignore the inherent changes to such specification wrought by the systems in which they are installed. Calibration has to include the complete measurement chain, from transducer to the input of the data store.

14 DISTRIBUTED I/O SYSTEMS

The low levels of many transducer outputs in their path to recording medium are vulnerable to corruption by external electromagnetic fields. Therefore, it is necessary that such signals are conditioned, digitized, converted, or amplified as close to the transducer as possible. The conditioned signal can then be passed via a high speed communication link, such as CAN-bus or Ethernet based interfaces to the data acquisition computer external to the cell. Signals are best marshaled and conditioned within the cell, within a "transducer box" into which run the multiple transducer cables or pressure lines. While the transducer box must be close to the engine as the source of the signals, it must also be protected from overheating by either positional shielding or/and forced ventilation.

The transducer box can be considered as one node in a distributed I/O system. Distributed I/O provides a very flexible, scalable concept for test data acquisition. It would normally consist of modules that are "daisy chained" together via a digital bus; the system can be extended by simply adding more modules to the I/O chain. Each module may have multiple input channels and its own local processing capability and intelligence, allowing it to measure the value from the connected sensor, convert it to a physical value by applying the correct calibration information, and then transmit the value back to the bus "host" with a defined response time, or data acquisition frequency. All data values are acquired with a time stamp and can hence be synchronized with the overall system data acquisition frequency.

In addition to analog signals, the test bed controller needs to be capable of reading and producing digital signals;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. A listing of the most common measurement channels used within the engine test system together with the type of signal produced.

Measurement	Principle applications	Method or transducer	Signal produced
Time interval	Rotational speed	Tachometer Single impulse trigger, starter ring gear, and pickup Optical encoder	Incremental pulses, analog sine wave or digital square edges
Force, quasi-static	Dynamometer torque	Strain gauge load cell	Analog signal, voltage, or current, to be amplified and linearized
Force, cyclic	Stress and bearing load investigations	Attached strain gauges plus wireless transmitters Strain gauge transducer Piezoelectric transducer	Analog signal, often complex, for detailed analysis May need additional signal conditioning/processing
Pressure	Fluid flow systems, gas or liquid	Liquid manometer Bourdon gauge Strain gauge pressure transducer	Analog voltage or current. Digital, pulse width modulated or variable frequency, also digital message (via CAN) for intelligent sensors
Pressure cyclic	In-cylinder, inlet, and exhaust events, fuel rail	Toughened strain gauge transducer Capacitive transducer Piezoresistive or Piezoelectric transducer.	Analog signals, used in conjunction with external signal conditioning to provide analog voltages for measurement
Position	Throttle and other actuators	Mechanical linkage Linear variable displacement transducer (LVDT) Rotary optical encoder	Analog voltage or current as a linear function of position
Displacement cyclic	Needle lift Crank top dead center (TDC)	Eddy-current or hall effect transducer, Inductive or Capacitive transducer	Requires external conditioning system to produce analog voltage for processing (absolute position not always necessary)
Acceleration	NVH investigations Shaft balancing	Strain gauge accelerometer Piezoelectric accelerometer	Analog signal, often complex, for processing and detailed analysis
Temperature	Fluids, air, and gases, mechanical components, in-cylinder	Liquid in glass Several types of Thermocouple PRT Thermistor Electrical resistance Optical pyrometer	Analog as a function of temperature, each type needs specific conditioning for amplification and linearization suitable for measurement

Table 4. Giving typical numbers of measurement channels specified for modern powertrain test cells in three generic types of industrial facilities.

Channels and I/O interfaces for measurement and data acquisition (excluding engine and dynamometer controls)	Base systems for QA and typical tier 1 and 2 testing	Engine calibration testing	Advanced powertrain development testing
Digital inputs (inc. PW and frequency)	16	16	32 or more
Digital outputs	16	16	32 or more
Analog inputs	Until 100	100–250	250 upward
Analog outputs	8–12	8–16	16–32
Interfaces to intelligent measurement devices (hybrid and digital)	4	8	8–16
Typical data acquisition speed (fastest channel sampling rate)	1–10 Hz	10–100 Hz	From 100 Hz to >1 kHz

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

typical applications include simple state bits (logic 0 or 1), binary input ports, counter–timers, pulse width, and duty cycle signals. Table 4 shows the typical I/O requirements for a range of test automation systems.

The choice of transducer, the speed at which its data is sampled and the accuracy to which it is digitized, should always be appropriate to the use. High speed acquisition and high resolution digitization of the signal from a K-type thermocouple is at best a waste of computing power and produces an allusion of precision beyond reality. The transducers serving those channels connected to primary safety systems, such as engine oil pressure, or critical control systems, such as the coolant temperature control unit, must be the most robust available (e.g., a PRT rather than K-type thermocouple).

15 INTEGRATION OF INTELLIGENT INSTRUMENTS

In addition to the analog and digital signals acquired directly from the transducers and interfaces described earlier, test cell automation has to support various special measurement devices within a test environment. These are devices with some local intelligence and are capable of data acquisition, signal conditioning, and calculation of derived data. The interface between the control computer and the device generally performs a remote control and data transfer function, communicating with the device via a digital interface. A commonly used interface is the so-called AK serial protocol that uses RS232 as the physical hardware layer. There are several other communication interfaces used and technology development in this area is in a state of evolution.

Measurements that were taken 10 years ago by long established techniques and basic instrumentation are now the subject of advanced study using increasingly complex digitally controlled devices, examples being:

- *Engine blow-by measurement* by instruments capable of detecting and integrating pulsating or inverted blow-by flows.
- *Oil consumption measurement*, which can involve instruments capable of detecting and measuring low levels of nuclear isotopes in exhaust gasses to calculate lubricating oil combustion in near real time.
- *Emission devices—gaseous and particulate*. These are sophisticated devices, containing several internal measurement modules using complex technologies covered elsewhere in Exhaust Emissions. The task of the test controller is to trigger their internally programmed processes when they are required to measure, purge, or calibrate themselves. Emission measurement units of the type used in research work have self-monitoring capabilities and are capable of flagging operation and error states. All this control and information need handling over an interface, as well as the transfer of data values, either online or as statistical values after a measurement procedure; therefore, a reliable digital interface is required. Several technologies can be encountered including general-purpose interface bus (GPIB, a parallel interface technology), RS232 (serial), and, increasingly, TCPIP (standard network technology).

16 COMBUSTION ANALYSIS

High speed measurement of combustion pressure-related parameters, in the crank-angle domain, is a common requirement employed throughout the engine development process. An overview of engine speed and cylinder pressure-related measurements is discussed in the following sections. We will consider the practical aspects of using this measurement technique later in the engine development process at the test bed. More details can be found in specialist publications such as “Engine Combustion Pressure Measurement” (see recommended Further Reading). Detailed analysis of the pressure data for information such as heat release rate is also discussed in Pressure and Heat Release Analysis. In order to implement any CA work, the engine must be prepared with additional instrumentation specific to this task, namely, an engine angle encoder and a combustion pressure sensor.

16.1 The angle encoder

This instrument facilitates recording of engine absolute position, from which the instantaneous engine volume can be derived. Its specification has to be carefully matched with the requirements of the system into which is to be integrated. The most commonly used encoder for CA applications generates square wave signal, on an incremental basis, giving the crank degree marks (a fixed number of marks per revolution) and a reference signal (one mark per revolution).

The encoder, mounted directly to the engine, operates in an environment with significant heat and vibration. In extreme situations, such as for some high performance engine test applications, the usual etched glass encoder disk has to be replaced by one made of metal. Mounting has to be carried out with care and precision so that the dynamic effects of misalignment are avoided.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

16.2 The engine combustion pressure sensor

The ECPS has to be mounted in the engine cylinder(s) to enable pressure data sampling and acquisition until 1 MHz. ECPS is manufactured with different properties to meet a wide range of test and engine monitoring applications that no sensor will fulfill. There will always be trade-off between durability, temperature range, and accuracy. An example is the knock measurement task; this will require a sensor of high durability and natural frequency rather than the highest pressure measurement accuracy.

The modification of a cylinder head required for the installation of a pressure sensor can present considerable practical challenges to a machine shop and must not be underestimated. In order to help with this task, some suppliers provide cutting tools specifically for the job. It must also be remembered that the installation should not affect the combustion chamber with respect to volume, charge motion, or thermal properties, neither must permanent installation compromise cooling fluid nor oil flow passages in the engine cylinder head.

An important part of the specification of any ECPS is the maximum temperature that the measuring element can withstand. Many water-cooled units have been produced, while the maximum temperature range of uncooled units continues to rise. A popular misconception is that water-cooled transducers must be more accurate and stable; this misconception is based on legacy issues with quartz piezoelectric transducers. Cooling of transducers requires extra equipment, and does not always mean greater accuracy or stability. The movement of water mass around the sensor element inside the transducer can create noise and crosstalk. Note also that cooled transducers are generally larger and more difficult in integrate in modern combustion chambers of production engines. Uncooled transducers are generally smaller but can have lower sensitivity.

The temporary installation of an ECPS can be achieved by using special, uncooled, or hybrid spark-plug sensors or those sized to use diesel glow-plug ports, but such installations generally trade-off convenience with ultimate accuracy (Figure 4).

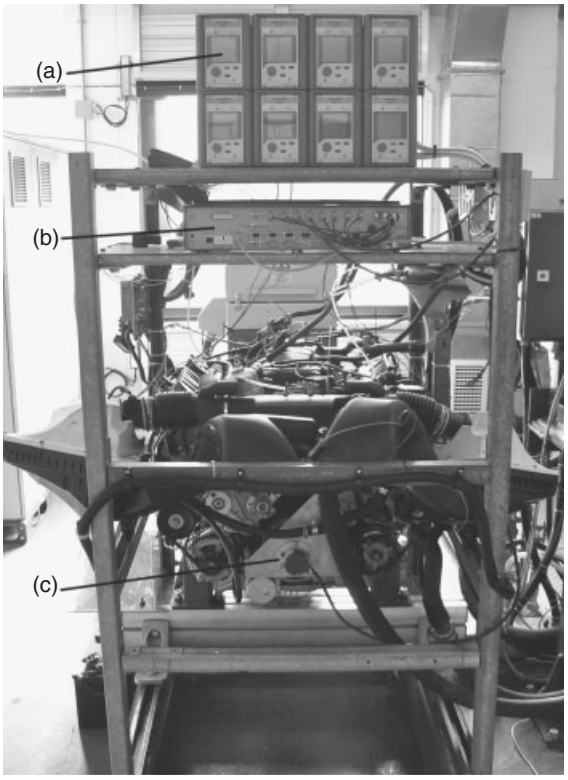


Figure 4. View of an engine fully rigged for combustion analysis work with the listed instrumentation shielded from direct engine heating by frame mounted shelving. A = signal conditioning rack and charge amplifiers; B = combustion pressure measurement and analysis system; and C = engine front pulley mounted encoder. (Photograph reproduced by permission of University of Bradford.)

17 CONDITIONING OF CYLINDER PRESSURE SIGNALS

The raw signal from engine combustion pressure transducers is not generally suitable for direct connection to the data acquisition device. The signal often needs a combination of conversion, amplification, and filtering in order to be usable. These functions are carried out by specific signal conditioning hardware, which can be physically integrated into the measurement system, but is more often positioned externally, close to the transducer.

Specific details of amplifier technology, operation, and integration are beyond the scope of this chapter and covered in detail within general texts (Randolph, 1990; Lancaster, Krieger, and Lienesch, 1976; Brown, 1967; Davis and Patterson, 2006; Kuratle and Balz, 1992) and proprietary literature from manufacturers. However, some general comments to consider in the interest of producing high quality data are listed.

- As a rule of thumb, the amplifier rack should be mounted as close as possible to the transducer, but it needs to be kept in at a stable temperature, so unnecessary switching on and off should be avoided and they must be shielded from heat sources.
- Interface panels and connectors for all cabling should be maintained clean as per laboratory instrumentation standards.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- Cable routing must be considered very carefully, avoiding all other power and signal cables in the test environment.
- The amplifier setting must be optimized with respect to sensor sensitivity and expected signal range. This will provide the optimum digital conversion quality.

18 SPECIAL PURPOSE TEST CELLS

The majority of this chapter has covered the choice, integration, and some of the uses of hardware installed within engine test cells designed for general research work on automotive engines. Many of the components and practices of these cells are also appropriate to the more common types of specialist test cell used in engine and powertrain development. The following sections cover special constructional and operational features of cells designed for climatic, NVH, or EMC research.

18.1 Climatic engine test cells

Climatic engine test cells range between those designed to run legislative "cold-start" test sequences, currently requiring the UUT to be at $-7\,^{\circ}$C, to those capable of exposing the UUT to temperature ranges of $+50\,^{\circ}$C to $-40\,^{\circ}$C. Such thermal changes demand considerable electrical energy for refrigeration plant and a great attention to the detailed construction. It is most advisable to only use refrigeration contractors with the appropriate industrial experience for the construction of such facilities, as those without will not appreciate or cater for the fluctuating energy outputs within the cell. While whole vehicle climatic chambers reduce the temperature of the complete cell, it is possible for engine testing to encase the UUT within a thermally insulated "tent" while circulating cold air around and cold fluids within the engine. Such systems greatly simplify the cell construction and reduce the energy required.

18.2 Anechoic cells

Anechoic cells of significantly different construction are required for EMC and NVH testing. In both cases, the site location has to be chosen so that there is a minimum of the type of background "noise"' in the location that would interfere with the measurements being made.

18.2.1 Anechoic cells for NVH

Anechoic cells for NVH work on engines are normally of a semianechoic (also called *hemi-anechoic*) design that can be thought of as simulating the sound regime of a vehicle sitting on a road in still air in the middle of the flat Arizonian desert. Full anechoic chambers are most commonly met in the smaller sizes required for component testing. The key specification of such cell is their "cut-off" frequency, which is the lowest frequency that sound can be emitted at the center of the cell while still obeying the inverse square law of sound energy decay. Typically, for modern engine test cells, this is around 60 Hz. These cells are fitted with sound-absorbing linings, the most common types of which take the form of cones whose apex protrude into the cell. Such cells require specialist designers experienced in dealing with the particular design and construction restraints that include

- the need to keep ventilation airflow as quiet as possible by the use of high volume, low velocity ducts whose entry termination is below the last row of cone lining and outlet is above or in the cell top corners.
- fluid control valves that create random noise, as they operate, have to be outside the chamber, which increases the thermal inertia of the system and gives control problems.
- the shaft connection to the engine needs to bridge an unusually large gap to maintain the engine position near the cell center. A pedestal bearing is usually mounted just inside the cell in order to divide the shaft to the dynamometer (outside the cell) into two sections.
- color CCTV using zoomable and steerable cameras need to be part of the cell system, as no (sound reflecting) windows can be included.

18.2.2 Anechoic cells for EMC work

Anechoic cells for EMC work are rarely constructed simply for engines but rather they are sized for whole vehicles. This is because the interference caused by the engine and powertrain and their vulnerability to electromagnetic energy sources can only be tested correctly as part of the total system of modules and their interconnecting wiring looms. The cell structures have to form a Faraday cage and they are lined with either or both:

- closed cell polyurethane foam molded into a steep pyramidal shape impregnated with carbon;
- ferrite tiles that are made from a sintered iron/nickel material usually measuring $100 \times 100 \times 6$ mm thick.

The need for certified EMC test facilities increases with the new generation of hybrid and electrical vehicles producing a whole range of new potential interference problems with their variable-speed drive systems.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

19 DATA MANAGEMENT, QUALITY, AND SECURITY

The product of all the expensive equipment within an automotive test facility is data and the empirically gained experience of its staff. In order to derive the maximum value from test data, it has to be protected from loss and corruption and be made readily available, in an appropriate form, to a range of authorized users. The greater the degree of homogeneity in the data sets, the easier will be the task of later correlation and research. This requires discipline to be imposed on the DOE on channel naming, storage of raw data, and auditing of postprocessing work.

The chosen software of the data acquisition and postprocessing suites can impose such discipline, but it must also be a part of the quality assurance procedures of the test facility. Differential access to levels of test controller software is a sensible precaution to prevent casual alteration of alarm limits; similarly, the veracity of original (raw) test data should be protected through protected archiving from post-processing.

There are a few major suites of software produced by companies, such as AVL List GmbH and MTS®, designed specifically for handling the work-flow and data-flow of automotive test facilities, but there is no substitute for diligent and supportive management of trained staff.

REFERENCES

Amann, C.A. (1985) Classical combustion diagnostics for engine research. SAE Paper No. 850395, Society of Automotive Engineers, Warrendale, PA. http://www.saedigitallibrary.org/content/technical-papers/.

ASAM Association for Standardisation of Automation and Measuring Systems. http://www.asam.net/ (accessed 3 July 2013).

Brown, W.L. (1967) Methods for evaluating requirements and errors in cylinder pressure measurements. SAE Paper No. 670008, Society of Automotive Engineers.

Davis, R.S. and Patterson, G.J. (2006) Cylinder pressure data quality checks and procedures to maximize data accuracy. SAE Paper No. 2006-01-1346, Society of Automotive Engineers, Warrendale, PA.

Den Hartog, J. (1985) *Mechanical Vibrations*, Dover Publications Inc., Mineola, NY. http://store.doverpublications.com. ISBN-13: 978-1406734812

European Commission. New Machinery Directive: Directive 2006/42/EC of European Commission.

HSE ATEX and explosive atmospheres' (Directive 99/92/EC). http://www.hse.gov.uk/fireandexplosion/atex.htm (accessed 3 July 2013).

Kuratle, R.H. and Balz, M. (1992) Influencing parameters and error sources during indication on internal combustion engines. SAE Paper No. 920233, Society of Automotive Engineers, Warrendale, PA.

Lancaster, D.R., Krieger, R.B and Lienesch, J.H. (1976) Measurement and analysis of engine pressure data. SAE Paper No. 750026, Society of Automotive Engineers, Warrendale, PA.

Randolph, A.L. (1990) Methods of processing cylinder pressure transducer signals to maximize accuracy. SAE Paper No. 900170, Society of Automotive Engineers, Warrendale, PA.

FURTHER READING

Martyr, A.J. and Plint, M.A. (2012) *Engine Testing*, 4th edn, Elsevier. ISBN-13: 978-0-08-096949-7

D.R. Rogers (2010) Engine combustion: pressure measurement and analysis. SAE International. ISBN: 978-0-7680-1963-6.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto207
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Trends—Compression Ignition

Zissis Samaras, Ilias Vouitsis, and Savas Geivanidis

Aristotle University, Thessaloniki, Greece

1 Introduction 1
2 Technology: The State of the Art and Future Trends of CI Engines 2
3 Summary and Conclusions 15
Acknowledgments 15
References 16

1 INTRODUCTION

On the evening of 29 September 1913, Rudolf Diesel (born in 1858), George Carels (an industrialist and Diesel's close friend, who would stand by him in times of need), and his chief engineer Luckmann sailed from Antwerp bound for Harwich on the Great Eastern Railway steamer *Dresden*. It would be a beautiful, calm, and clear night. The route was ideal because the English port lay only a few miles from Ipswich where a factory visit had been scheduled before the first annual meeting of Consolidated Diesel Engine Manufacturers in London on October 2. The three men had a pleasant dinner aboard and retired to their staterooms about 10 o'clock. It was agreed to gather for an early breakfast before going ashore the next morning. The ever punctual Diesel did not appear as planned, and in due course his stateroom was checked. They saw his bed turned down but not slept in, all personal baggage intact, his watch by the bedside, and keys hung from his handbag. A search of the ship before allowing passengers to disembark found no trace of him (Cummins, 1993).

One century after this fatal night, the compression ignition (CI) engine (popularly known by his name) is the standard solution for almost all heavy-duty applications, either commercial or military (buses and trucks, ships, engines for industry, power generation, construction, and agriculture). This is due to its reliability, durability, and fuel efficiency. In recent years, owing to the rapid increase in specific output with simultaneously very high torque when used with a turbocharger, it has become very popular for passenger cars (PCs) and sport vehicles as well, especially in Europe (Figure 1). Modern diesel PC offers an outstanding driving experience, while at the same time consuming very little fuel.

Future CI powertrains for both light- and heavy-duty applications have to fulfill a set of requirements determined by a combination of social, economic, environmental, technological, political, and infrastructural constraints (Figure 2). The importance of each component depends on the application but the general scheme is the same. The technological evolution of all engines in the near future will be driven mainly by environmental requests: "near zero" "toxic" emissions by the 2030 time frame and substantial reductions of greenhouse gas (GHG) emissions. Figure 3 shows the main technology trends for the light-duty powertrain over the next 20 years.

In the following, the technology status and future trends of CI engines are discussed. We begin with fuel injection equipment (FIE) and proceed with turbocharging and downsizing, flexible engine systems, friction losses and thermal management, advanced combustion modes and reduction of emissions, alternative fuels, and hybrids. We end with a brief comparison with spark-ignition (SI) engines and with the conclusions.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

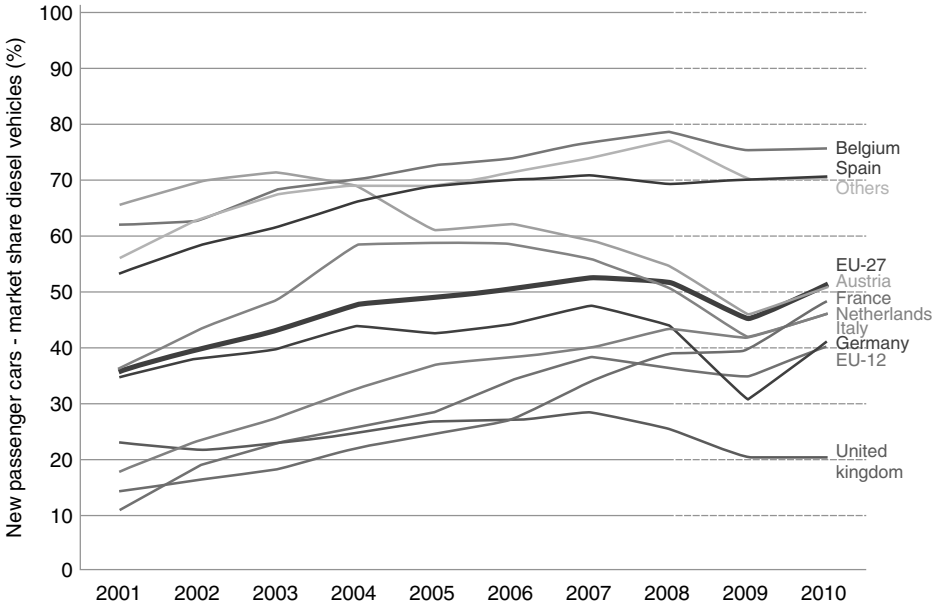


Figure 1. New passenger cars in European Union: market share diesel vehicles by Member State. (Reproduced by permission of Association des Constructeurs Européens d'Automobiles. Also from The International Council on Clean Transportation, 2011. Reproduced with permission.)

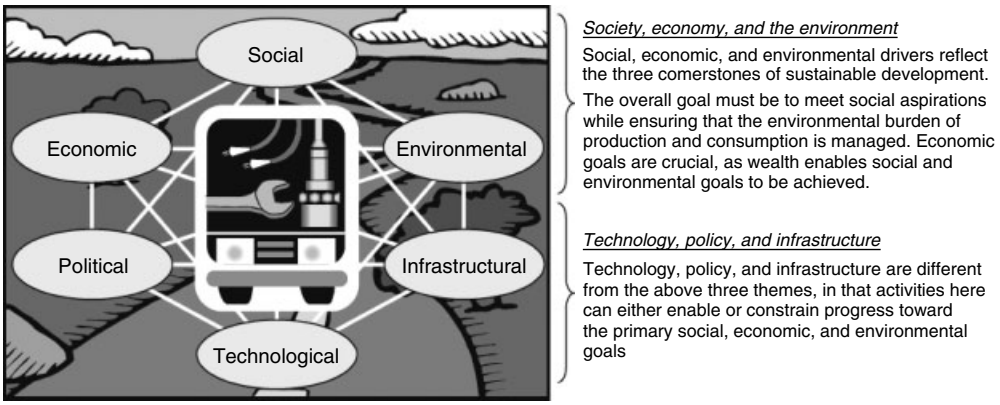


Figure 2. Trends and drivers that influence road transport system. (Reproduced with permission from Institution for Manufacturing Education and Consultancy Services (IfM ECS), 2000. © Institution for Manufacturing Education and Consultancy Services.)

2 TECHNOLOGY: THE STATE OF THE ART AND FUTURE TRENDS OF CI ENGINES

CI engine system design is a highly specialized technical field, which integrates the performance of various subsystems and properly matches them with optimization approaches to achieve precise and accurate system (Xin, 2011). It covers injection system design, thermodynamic cycle simulation and air system design, valvetrain system design, engine friction, brake performance, combustion, heat rejection and cooling, emissions, and electronic controls.

2.1 Fuel injection

Fuel injection affects heavily the process of mixture formation and combustion as well as the emission formation. Modern high speed direct injection (HSDI) CI engines are capable of reaching levels of performance and refinement that were implausible some years ago, mainly because of the availability of very advanced fuel-injection

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

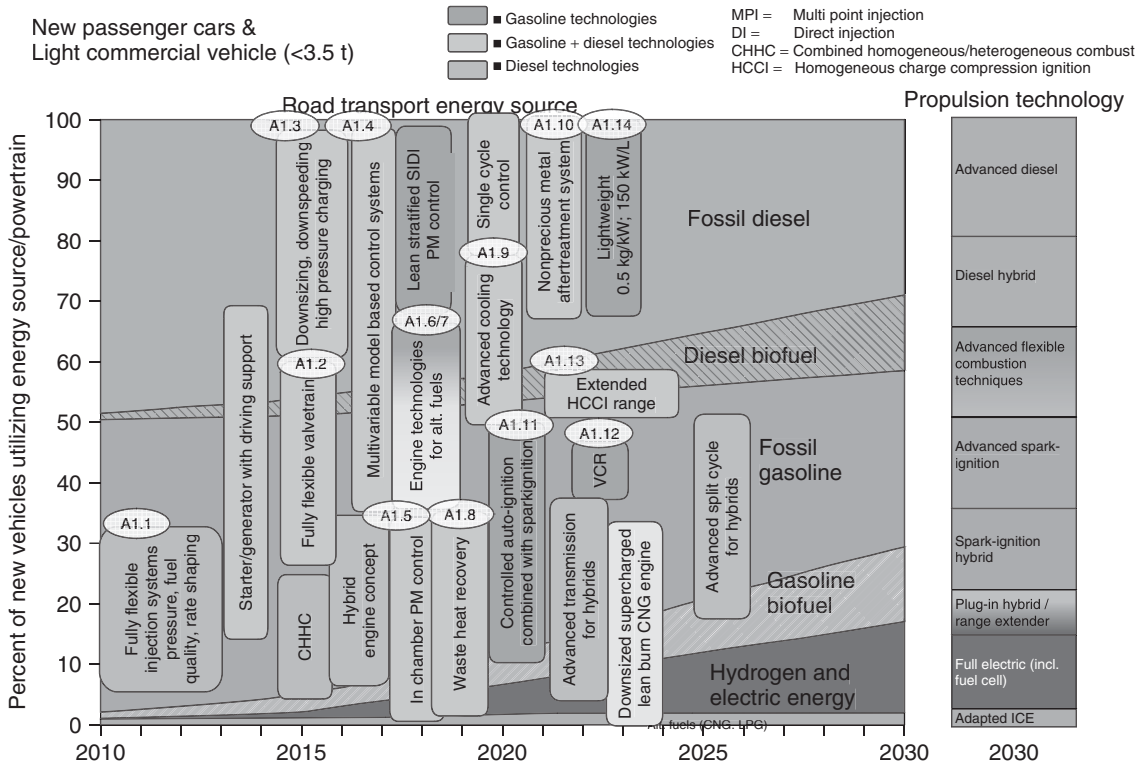


Figure 3. Technology trends and research needs for light-duty ICE powertrains to 2030. (Reproduced by permission of ERTRAC.)

technologies. In a high pressure common rail (HPCR) fuel injection system, pressure generation and fuel injection are separate events; the injection pressure is not dependent on engine speed, allowing thus highly flexible injection strategies at pressures above of 2000 bar and up to five injections per cycle per cylinder, allowing significant reductions in engine-out emissions, noise, and fuel consumption (Dober *et al*., 2008; Kimberley, 2005).

Modern FIE offers the possibility of performing multiple injections (Figure 4). In this case, the injection event consists of three parts, pre-, main, and postinjections. The multiplicity of injections gives the possibility for lower engine noise by reducing explosiveness and vibration and simultaneous reduction of particulate and nitrogen oxide (NO*x*) emissions by optimizing injection timing and quantity for variations in fuel quality, cold starting, and so on (Mittal *et al*., 2009; Baumgarten, 2006; Mahr, 2002).

The latest generation common-rail FIE incorporates piezoelectric-controlled injectors (Figure 5) that are capable of extremely rapid switching and as such the possibility of shorter injection durations, very small injection quantities, and multiple injections per cycle. The piezo element can be made to act directly on the nozzle needle without the need for an intermediate hydraulic system, unlike Figure 5,

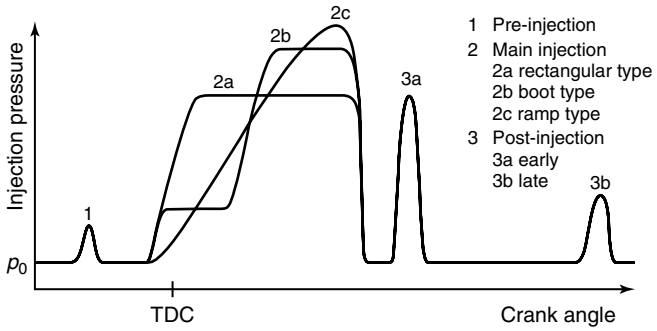


Figure 4. Schematic diagram of a multiple injection. (Reproduced from *Mixture Formation in Internal Combustion Engines*, 2006, G. Baumgarten. With kind permission of Springer Science+ Business Media.)

allowing more precise control of the degree and length of the time the injector is open. This allows the option of injection rate shaping as an alternative to the usual multiple injection method used so far, which places severe mechanical demands on the injection system. For the direct acting piezo design, the piezo actuator simultaneously acts as a sensor by reporting the precise position of the nozzle needle to the electronic control unit. The result is a self-contained

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

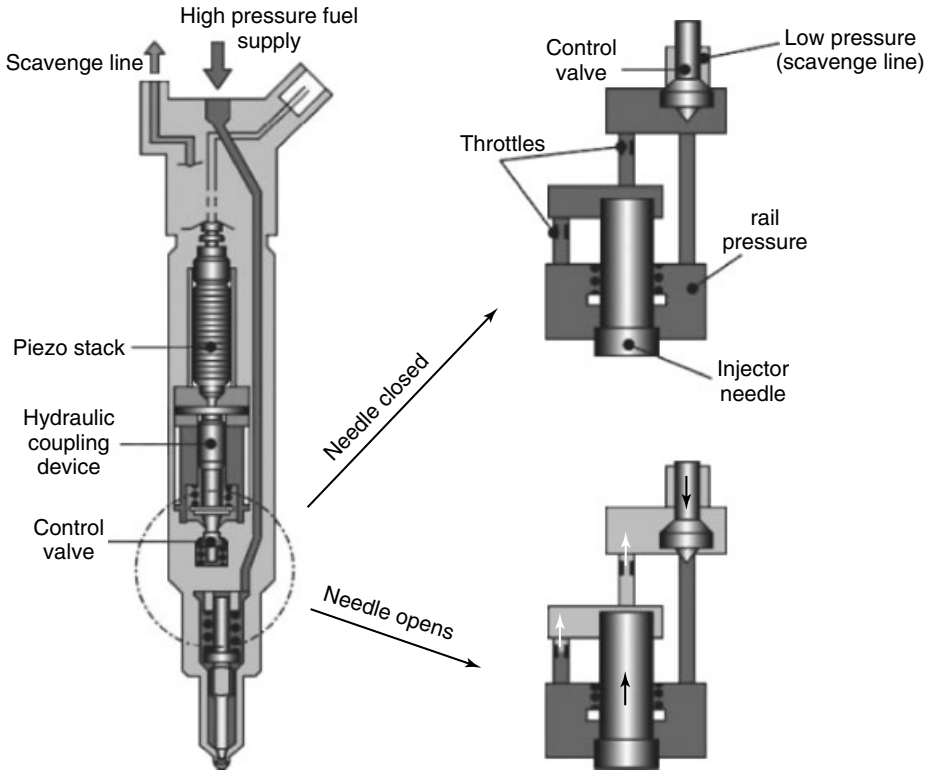


Figure 5. Piezo in-line injector from Bosch. (Reproduced from Mixture Formation in Internal Combustion Engines, 2006, G. Baumgarten. With kind permission of Springer Science+Business Media.)

flow rate control system; the control unit detects any variations in flow rate while the vehicle is being driven and individually adjusts each injector. This produces the best possible figures for NO_x emissions and fuel consumption. Improvements to the hydraulic design will optimize the performance of the already proven technology behind the millions of piezo common rail systems currently in use, enabling the fuel return flow to be reduced by 40%. This measure alone will reduce the CO_2 emissions from a typical diesel PC by 1 g/km. In addition, a fuel pump with a lower delivery rate is needed, which reduced costs still further.

In future, very flexible high pressure fuel injection systems with multiple injection and rate shaping capabilities as well as increased injection pressures, matching of the fuel spray by means of variable nozzle geometry and multihole nozzles and optimized control strategies are necessary in order to realize the optimum rate shaping and injection timing for each single point of the engine map and to get the best compromise between emission trade-off and fuel consumption. The more the injection pressure is increased, the more the efficiency of the injection system itself becomes important for achieving a low overall fuel consumption for an engine.

2.2 Turbocharging

The purpose of turbocharging has in the past been to increase the power-to-weight ratio of the engine. By increasing the amount of air available for the combustion process, more fuel can be burned effectively. Nowadays, the primary goal to use turbocharging for heavy-duty applications is still to raise the power-to-weight ratio; however, it is more and more used as a help to optimize the engine to obtain lower emissions in order to meet future emission legislation while maintaining or even improving fuel efficiency. Turbochargers and details of their performance and operation are discussed in detail in Intake Boosting. The turbocharger itself has become a mature product. Improvements on the turbomachine will, therefore, be rather incremental and the focus for development has turned more toward the turbocharger application. New complex systems to improve exhaust energy utilization over a wide engine operating range will be more frequently incorporated into the engine design such as variable geometry turbines (VGTs), multistage turbocharging, electric turbocompounding, or various additional mechanically driven compressors (Eichhorn, Boot, and Luijten,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2010; Brockbank, 2009; Plianos and Stobart, 2008; Knecht, 2008; Uchida, 2006; Hopmann and Algrain, 2003).

2.2.1 *Variable geometry turbines*

The VGT (Figure 6) is an effective approach to the turbocharger to engine matching problem. By varying the inlet turbine geometry, it is similar to having a finite range of turbine sizes in one unit. At low engine speeds, the VGT is kept closed to raise the pressure upstream the turbine, thus the isentropic energy from the exhaust increases. At high engine speeds, the inlet area is increased to avoid overboosting and high engine backpressure. The inlet area is mainly adjusted in two ways, by a sliding nozzle ring mechanism or by pivoting nozzle blades. The sliding nozzle ring mechanism keeps the nozzle blades fixed and changes the inlet area because of an axial movement of the sliding wall. By pivoting the nozzle blades, the area between the blades changes as well as the vane angle.

2.2.2 *Two-stage turbocharging*

Two-stage turbocharging (Figure 7) is an effective way of overcoming the limitations imposed on boost pressure by current compressor materials and to improve transient response. In a two-stage turbocharging system, an intercooler can be used between the compressor stages, which reduces inlet air temperature to the high pressure stage, thus it is possible to reach higher boost pressures with the standard impeller material. Increased air density also implies that a smaller compressor can be used and that less power is needed to reach the desired pressure raise. In a sequential turbocharger arrangement, two or several different sized turbochargers can be used where the flow can be switched, so that the best-suited turbocharger for that operational point is used. Sequential turbocharging can improve low end torque, response time, and/or operational range.

2.2.3 *Electric turbocompounding*

A further development is the use of turbochargers with electric motor/generators on the turbocharger shaft (Figure 8). This may offer (Knecht, 2008)

- increased boost pressure at low engine speeds where exhaust flow to drive the turbine is limited;
- maintained turbocharger speed during gearshifts;
- reduced turbocharger speed under high engine power conditions, using the electrical machine to apply a restraining torque to that exerted by the turbine;
- recovered exhaust energy using the electrical machine in generator mode, converting excess shaft power to electrical power.

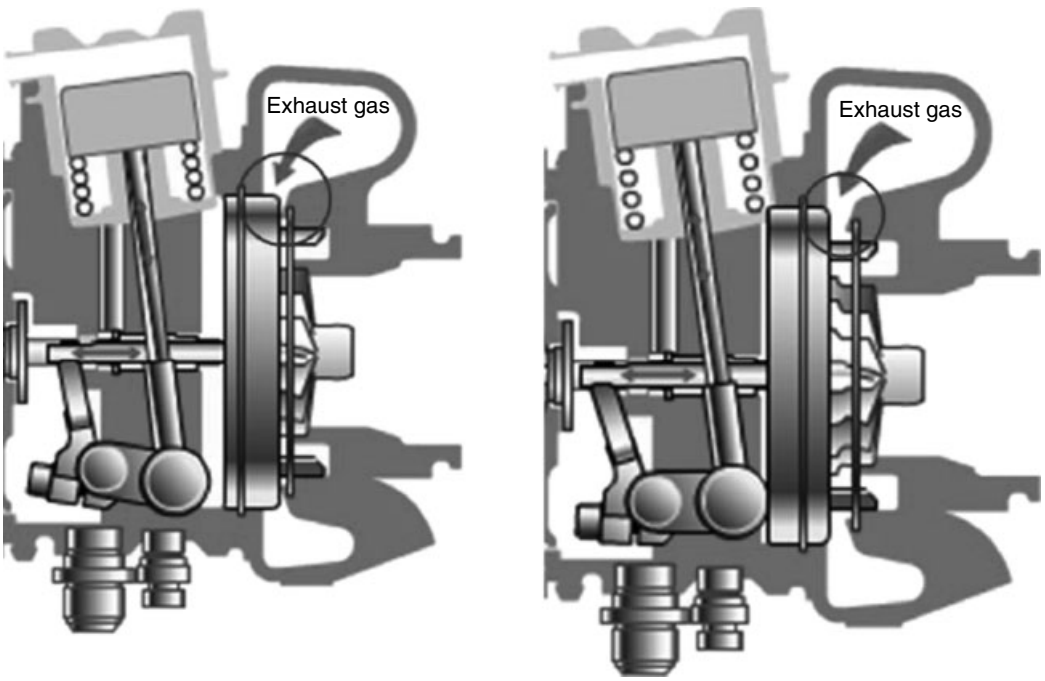


Figure 6. Variable geometry turbine, small and large areas. (Reproduced from Cummins Turbo Technologies, 2009. © Cummins Turbo Technologies.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

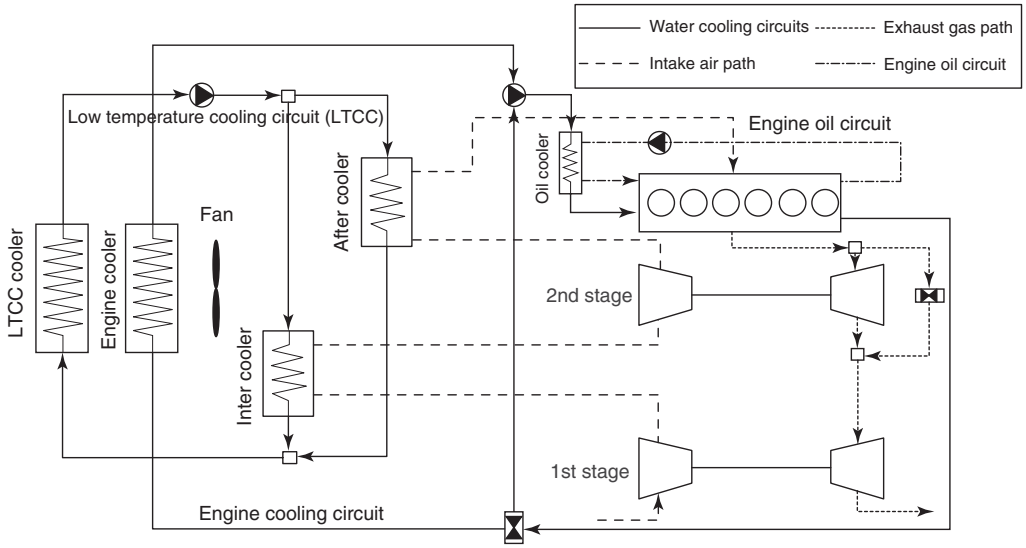


Figure 7. Two-stage turbocharging. (Reproduced from Knecht, 2008. © Elsevier.)

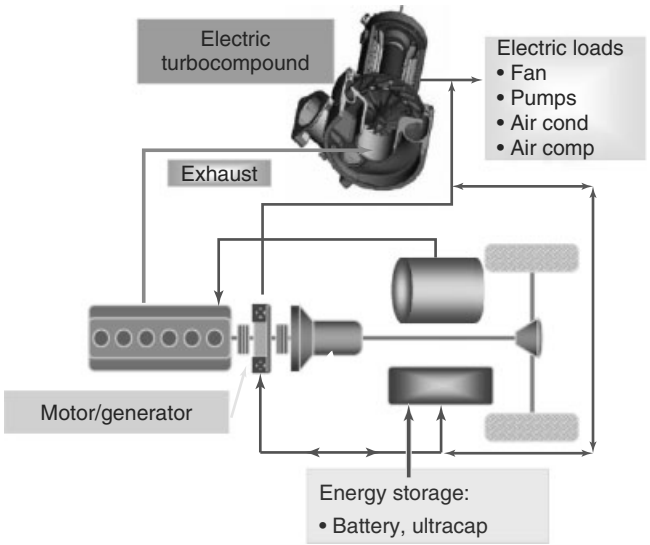


Figure 8. Electrical turbocompounding. (Reproduced by permission of Anthony Greszler, AB Volvo.)

2.3 Downsizing, downspeeding, and cylinder deactivation

Engine downsizing is the most important indirect fuel-economy measure in the powertrain. Downsizing means the reduction of displacement—either by making the specific cylinder displacement smaller or by reducing the number of cylinders. Compared with the reference engine, the downsized engine has a smaller displacement, which reduces friction, moving mass, and thermal loss. The increased specific maximum power because of engine downsizing is made possible by the appropriate redesigning of the air system (turbocharger and intercooler), combustion chamber (piston and cylinder head), and FIE (high pressure pump, injector and nozzle, and sensors and actuators).

By downspeeding, the engine is operated at low speeds but with higher torque (same power produced). The speed change can be achieved by changes to the gear ratio. Recent research (Ostrowski *et al*., 2012; Narayanan, 2011) has shown that downspeeding proves to be efficient. The fuel economy improvement is a result of three main factors: reduction of friction, relative heat transfer, and increase of fuel conversion efficiency.

Cylinder deactivation (Flierl *et al*., 2012) is a method used to create a variable displacement engine that is able to supply the full power of a large engine under high load conditions as well as the fuel economy of a small engine for cruising. By deactivating half of the cylinders, the remaining active cylinders operate at twice the load that the engine would normally operate at if all cylinders were active. This reduces the pumping losses and improves fuel consumption. Engines with cylinder deactivation can be found in several vehicles under various trade names such as multiple displacement system (MDS) and active fuel management (AFM). To date, cylinder deactivation has been applied to V6, V8, and V12 engines.

Strongly interlinked with downsizing is engine lightweighting (ERTRAC, 2010). The latter, however,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

presents substantial challenges for ultrahigh power-to-weight ratios. The higher the specific power output of an engine will be, the more the demand for high cylinder pressure exists. This in turn requires very high strength for components such as the crankcase, connecting rod, piston, and bearings, which currently does not exist in conventional turbocharged engines. Cylinder pressures of around 200 bar require special designs which in turn increase weight. The challenge is to make a step-out improvement compared to today's engine. The use of exotic materials and very smart design layouts, which are considerably different from today, is inevitable. Lightweighting will be especially important in hybrid vehicles to offset additional battery weight.

2.4 Flexible engine systems

Although a significant number of engine valve-actuation systems including cam-based and camless mechanisms have been already introduced by several researchers and companies, only a few types of these systems (mainly cam based) have been actually employed on commercial vehicles because of the liability, durability, and cost issues. Despite the fact that cam-based valve systems offer more reliable and durable functionality, the camless valvetrains can vary valve lift and timings to a greater extent compared to the cam-based types (Zheng, 2007). Several systems that allow flexible control are being assessed: variable valve actuation (VVA), variable swirl systems (VSSs), and variable compression ratio (VCR).

VVA gives the opportunity to have specific lift profiles for every operating condition and thus gives greater efficiency and power across a wider range of engine speeds. Camless-based VVA strategies offer great adjustability of valve timing and lift. The drawback is that these systems are often complex and expensive and mainly used in laboratory settings. Several different approaches have been proposed including electromagnetic, electrohydraulic, and electropneumatic systems (Figures 9 and 10) (Heinzen, Gillella, and Sun, 2011; Turner *et al*., 2004; Tai and Tsao, 2002; Schechter and Levin, 1996). VSSs fulfill the need of CI combustion systems for variable swirl levels dependent on load and speed (Kawashima, Ogawa, and Tsuru, 1998). This is true for conventional fuel systems with load- and speed-dependent injection pressure, but also for advanced common rail systems where there is freedom to choose the most appropriate injection pressure almost independently from the operating conditions.

VCR was originally developed for SI applications; however, it seems promising for CI engines as well

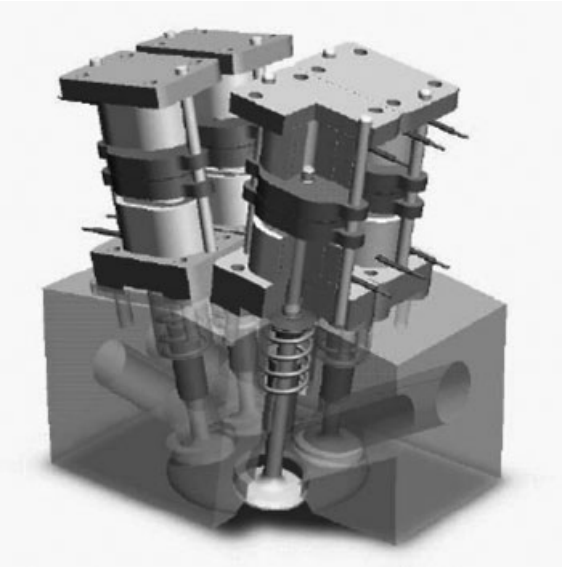


Figure 9. Electromagnetic fully flexible valve actuator. (Reproduced by permission of Engineering Matters, Inc.)

(Pesic, Milojevic, and Veinovic, 2010). A VCR engine is able to operate at different compression ratios, depending on the particular vehicle performance needs. The VCR engine is optimized for the full range of driving conditions, such as acceleration, speed, and load. At low power levels, the VCR engine operates at high compression to capture fuel efficiency benefits, whereas at high loads during turbocharging, its compression ratio is lower than that of a conventional engine, so that the temperature and pressure conditions remain at acceptable levels. One approach for realizing a VCR engine is with a variable combustion chamber volume enabled through a secondary piston in the cylinder head (e.g., Erlandsson *et al*., 1998). This design needs supercharging, because implementation becomes more difficult in engines with four valves per cylinder technology. Another possibility for the VCR engine is the repositioning of the cylinder or the cylinder head that has been put into practice by SAAB—Saab variable compression—SVC engine) (Bergsten, 2001). As an alternative to VCR, some approaches use late intake valve closing (LIVC) to further reduce the effective compression ratio (Patton, Manuel, and Gonlazez, 2010; Dec, 2009). LIVC has the advantage of allowing the effective compression ratio to be rapidly adjusted as part of the control system, while a high expansion ratio is preserved for good cycle efficiency.

VSS and VCR increase the engine complexity significantly and their cost/benefit is under investigation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

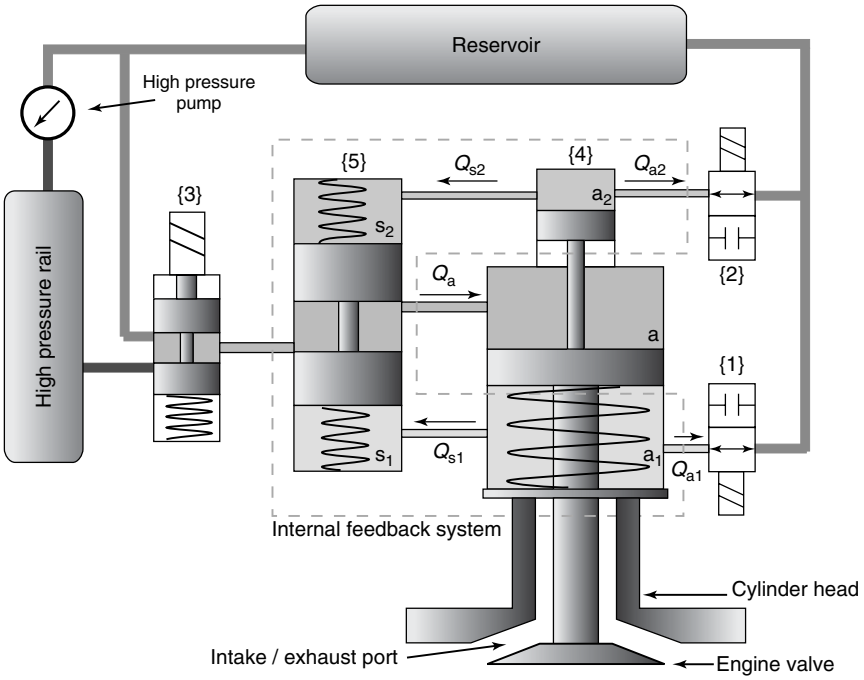


Figure 10. Electrohydraulic valve actuation system with internal feedback. (Reproduced from Heinzen, Gillella, and Sun, 2011. © Elsevier.)

2.5 Engine friction losses reduction

An area of interest to all engine developers is internal friction reduction. Insight into reduction of friction at part load (PC) or all loads (trucks) is increasing. Newer engines are showing a reduction of 40% in wall to piston ring friction by introducing new honing processes allowing the piston ring pretension to be reduced by 50%.

Losses in driving auxiliary systems are tackled by the introduction of variable cooling liquid and oil pumps that allow the precise balance between pump volume/pressure and engine condition to be controlled and thus not "waste" energy driving them. Better control of cooling liquid pumps also allows more advanced heat management. This allows the engine to warm up more quickly, which leads to a CO_2 emission reduction.

Current and future technologies needed to reduce friction and wear focus on (Automotive Technology Centre (ATC), 2011; Fenske *et al.*, 2009)

- development and application of mechanistic models of friction losses to predict parasitic losses as a function of engine and tribological conditions;
- evaluation/screening the potential of candidate surface treatments and additives to reduce boundary friction under lab conditions prototypical of engine environments;
- development of advanced surface treatments and additives that provide low friction and are robust, reliable, and durable;
- development of tools and theories to understand and model complex physical phenomena in tribological environments;
- development of advanced laser surface treatments;
- demonstrate efficiency improvements in multicylinder engines.

2.6 Thermal management

2.6.1 Advanced cooling and electric water pump

Advanced automotive thermal management systems can effectively maintain the desired temperature in internal combustion engines for enhanced performance (Figure 11). Automotive cooling systems can be upgraded to computer-controlled servo-motor actuated components rather than the conventional wax-based thermostat valve, mechanical water pump, and viscous clutch radiator fan (Chalgren and Barron, 2003). The latter action decouples the water pump and radiator fan from the engine crankshaft. Hence, the problem of having over/undercooling, due to the mechanical coupling, is solved, and parasitic losses reduced, which arose from operating mechanical components at high rotational speeds. The adjustment of thermal system operation

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

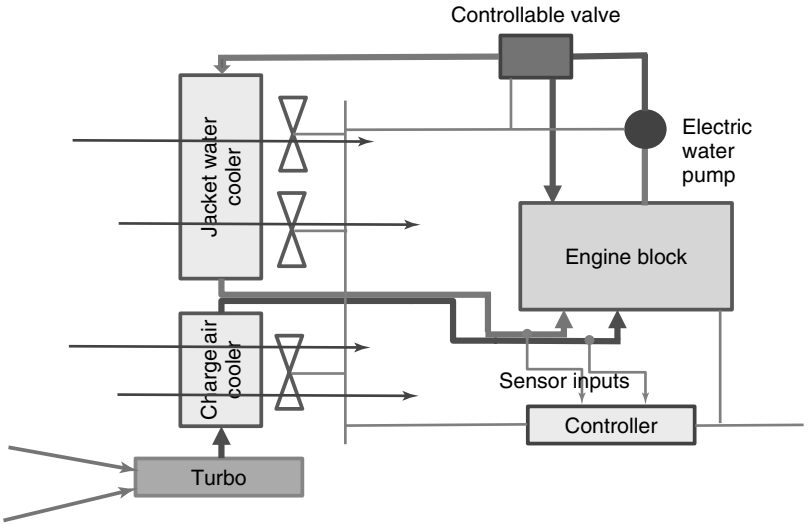


Figure 11. Advanced cooling system that features a controllable valve, electric water pump, variable-speed fan, engine block, radiator, and sensors.

per driving condition can reduce fuel consumption, parasitic losses, and tailpipe emissions during transient and steady-state operation (Page, Hnatczuk, and Kozierowski, 2005; Chalgren and Barron, 2003).

Waste heat recovery systems (WHRSs) have been already used as possible solutions for improving the efficiency by converting the high amount of heat wasted in exhaust gases into usable mechanical or electrical energy. Different methods have been recommended, such as, thermoelectric, absorption refrigeration system, and organic Rankine cycle (ORC). In particular, the latter is often proposed as a very powerful means and it is extensively investigated by the automotive manufacturers (Edwards *et al.*, 2012). Various configurations have been presented for heat recovery using ORC. Some of them are simple, preheat, and regenerated Rankine cycles. The amount of recovered heat and, consequently, cycle output power for these configurations depend on their characteristics (Figure 12).

2.7 Advanced CI combustion modes

To achieve the needs of further emissions reduction, improved efficiency, and cost, significant research is currently being focused on alternative forms of CI combustion. Advanced combustion modes attempt with in-cylinder approaches to achieve high fuel efficiency and to either meet fully the emission standards, thus avoiding the need to use aftertreatment or, at the very least, lower the performance demands required from aftertreatment systems, thus reducing their cost and complexity (Jääskeläinen, 2010). These advanced combustion systems carry numerous names such as homogeneous charge compression ignition (HCCI) and premixed charge compression ignition (PCCI) that may or may not accurately reflect the combustion process. HCCI was one of the early CI combustion concepts to attract attention. As the name implies, the goal of early HCCI work was to achieve as homogeneous an air–fuel mixture as possible before ignition—much the same as in a conventional SI engine. This can be achieved by injecting fuel either in the intake port or directly in the cylinder and allowing sufficient time between injection and ignition to achieve complete mixing of air and fuel. The charge then autoignites, as it is heated by the compressed gases—no spark or other means of forced ignition is used. In order to address many of the challenges posed by HCCI such as limited load range, controllability, and knocking, a number of other concepts have evolved from this homogeneous charge approach; in many cases, charge stratification was introduced. As the term *HCCI* may no longer accurately describes many of these systems, the term *low temperature combustion* (*LTC*) can be used as a general term to refer to these and other advanced combustion concepts, because the primary goal is to lower combustion temperatures to advantageously alter the chemistry of NO_x formation while achieving air–fuel mixture conditions that also avoid soot formation.

As stated earlier, the underlying objective of LTC is to keep in-cylinder temperatures low (less than approximately 1900 K) through volumetric energy release via autoignition of dilute air–fuel mixtures, as opposed to the high temperatures created by, for example, flame propagation through stoichiometric mixtures in conventional SI engines (Foster,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

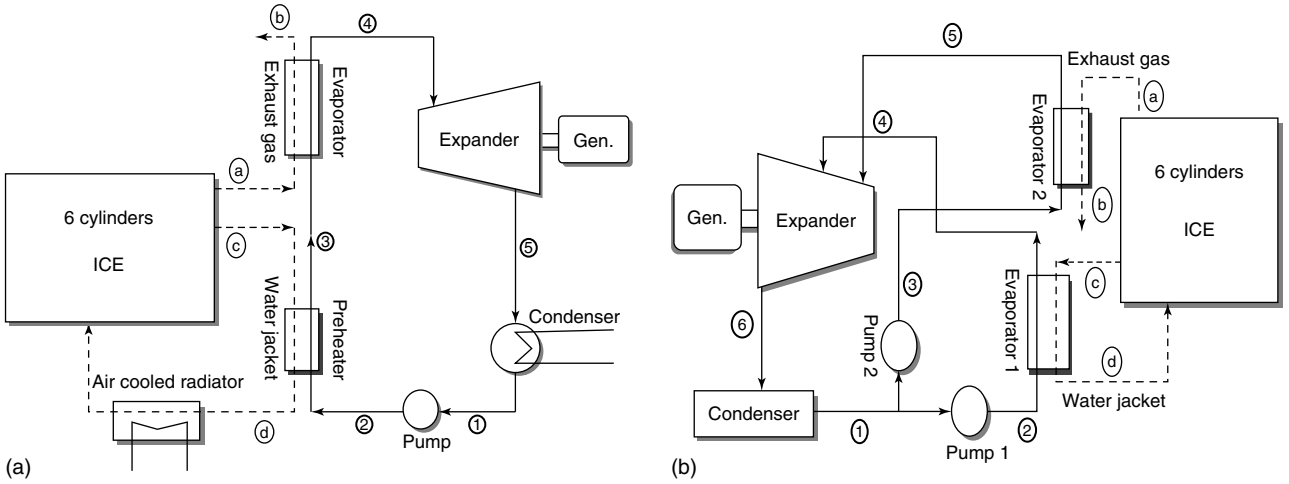


Figure 12. Configurations of Rankine cycle for waste heat recovery: preheat (a) and two stage (b). In the first one, a preheater was used for waste heat recovery from the coolant. The working fluid was preheated in this heat exchanger then entered the evaporator and was converted to saturated vapor by absorbing heat from the exhaust gas. After that, it entered the expander and generated mechanical power by expanding. Owing to working fluid mass flow rate limitation, this configuration is unable to absorb the total heat released by the coolant. Thus, an air-cooled heat exchanger should be used to reduce the temperature of the coolant before returning to the engine. In the second one, the working fluid flows in two different stages with different pressures after leaving the condenser. Low pressure stage relates to heat recovery from the coolant and the other stage relates to heat recovery from exhaust gas. Using this configuration, the total wasted heat of the coolant can be recovered and there is no need for an air-cooled heat exchanger. (Reproduced by permission of THERMAL SCIENCE, International scientific journal, ISSN 2334-7163.)

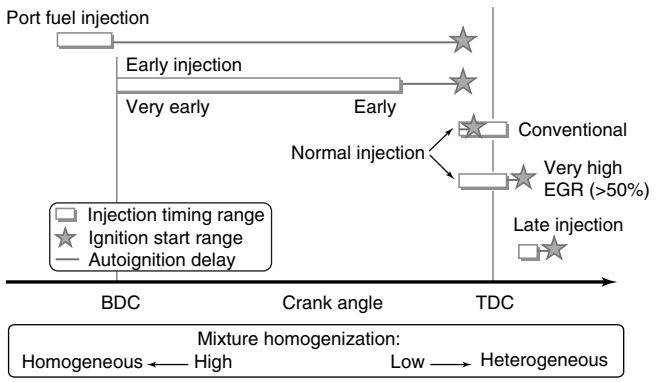


Figure 13. Classification of low temperature combustion strategies. (Reproduced with permission from Dieselnet, 2013. Source: Dieselnet.com © Ecopoint Inc., 2013.)

2012). LTC utilizes high levels of dilution (air or EGR) combined with flexible fuel injection strategies and strongly turbulent in-cylinder conditions to create conditions for reducing overall temperatures to levels below the critical temperature for NO*x* formation, and at the same time, air–fuel mixture conditions for reducing smoke formation. LTC strategies can be classified as (i) port mixing, (ii) early direct injection, and (iii) late direct injection. Some classification further breaks down the late injection systems into those that use "conventional" injection timing just before TDC and those with injection occurring after TDC. These strategies are summarized graphically in Figure 13. LTC has already demonstrated a significant potential to reduce PM and NO*x* emissions at part-load operation without major increase of fuel consumption and/or combustion noise excitation. However, an increase in hydrocarbons (HCs) and carbon monoxide (CO) emissions has been observed (Dec, 2009). Further, at higher loads, and especially at full load, it is challenging to achieve low temperatures. One approach to achieve high loads requires the development of new control strategies for transitioning between the LTC modes and a more conventional combustion mode at higher loads, such as SI. This approach will require conventional aftertreatment at high load and introduces the challenge of making the combustion mode transitions completely transparent to vehicle driver (Schulte and Wirth, 2010). Another approach to high loads that is emerging is the use of elevated levels of turbocharging with LTC. This approach does not require a combustion mode change but does drive the need for boosting/air-handling/EGR systems that effectively provide the right combination of the boost pressure and EGR rate under all load and speed conditions. Note also that because LTC depends on autoignition, the methods used to achieve it are dependent on the autoignition characteristics of the fuel being used. The various advanced CI combustion modes are discussed in more detail in Diesel and Diesel LTC

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Combustion and Advanced Compression-Ignition Combustion for Ultra-Low NO_x and Soot.

2.8 Reduction of emissions

The major challenge for the CI engine is meeting the tough near-future emission targets. For example, Euro VI standards for PCs are 0.08 g/km for NO*x*, 0.005 g/km for particle mass (PM), and 6×10^{11} particles/km for particle number (PN), and Euro VI standards for heavy-duty vehicles (HDVs) are 0.4 g/kWh for NO*x* and 0.01 g/kWh for PM (Dieselnet, 2012) at affordable cost, while further improving, or at least maintaining the diesel typical fuel economy advantages.

In contrast with SI engines, there is not just one key technology (such as the three-way catalyst for gasoline engines) that would provide sufficient potential to achieve these future standards. To overcome the inherent PM–NO_x trade-off, an integrated system approach is required. It includes two main areas:

- combustion system developments to reduce engine-out emissions;
- emission control technologies using exhaust gas aftertreatment.

2.8.1 Combustion system developments

These involve achieving highly premixed combustion and LTC in the directions already discussed in a previous section. Engines that utilize the conventional combustion mode at higher loads and LTC mode at moderate to light loads (mixed mode) are being investigated. Mixed mode operation combines the high efficiency, high load capabilities of conventional mode with the high efficiency, low emission capabilities of LTC mode to overcome deficiencies of aftertreatment system at low loads while dealing with the limited operating range of LTC.

Along with these developments, robustness control will become a critical issue (Galindo *et al*., 2011). For this reason, control system developments will be essential. This is likely to be achieved through a combination of new sensor technologies and improved processing capabilities. New sensors will provide direct feedback indicating combustion characteristics to the control system and, when coupled with model-based control of the air and EGR systems, will enable adaptive control of the engine variables. Such systems also provide improved onboard diagnostic (OBD) capability. This technology is at an early stage and further developments are required. Advances in model-based control will be an essential enabler for low NO*x* strategies, which operate much closer to engine combustion limits.

2.8.2 Emission control technologies

The exhaust emission standards for new cars have effectively required fitment of a diesel particulate filter (DPF) in the exhaust of PCs since 2009, when the Euro V standard came into force. Further, DPFs are expected to be necessary to meet the HDV Euro VI standard. DPFs usually remove 99% of the particles (Johnson, 2008). Details of PM control technologies are discussed in Solid/Condensed Phase Aftertreatment Systems.

Exhaust gas aftertreatment is one way to avoid the target conflict between NO*x* emissions and fuel consumption. Euro IV legislation launched the introduction of selective catalytic reduction (SCR) for NO*x* control by the majority of European HDV manufacturers. An alternative to SCR is an adsorber catalyst (lean NO*x* trap—LNT), which adsorbs NO*x* in "lean" exhaust gas conditions and desorbs it in short periods with "rich" exhaust gas (Johnson, 2008). Figure 14 gives an overview of emission reduction technologies for heavy-duty applications in Europe and the United States. Details of NO*x* and other gas phase emission control technologies are discussed in auto130.

While DPF (as well as diesel oxidation catalysts—DOCs) can be considered state of the art and an integral part of the CI engine, similarly to direct injection and turbocharging, this does not apply to NO*x* reduction catalysis. The development of LNT and SCR systems has not yet been successfully completed for light-duty diesel engines. SCR technology is much more mature, has a higher NO*x* reduction potential because of its wider temperature window with high conversion rates, and is less dependent on precious metal usage. The major SCR disadvantages result from the current need for onboard storage of the reductant, currently an aqueous urea solution, and the infrastructure, which is needed for the urea supply. In general, it is desirable to exploit possible synergies in the exhaust line. Typical examples of such synergies are the DOC/DPF CRT system, the "passive" LNT/SCR system, and the integration of DOC and/or deNO*x* functionality inside the DPF (Koltsakis *et al*., 2007, 2012). Figure 15 shows the structure of an advanced exhaust aftertreatment system and Figure 16 a model-based comparison of DeNO*x* performance of different exhaust system layouts. Advanced coating technologies employing multiple zones or layers may further contribute to the clever design of such systems. Designing and evaluating advanced multifunctional systems requires good understanding of the complex underlying phenomena and requires much effort, time, and experimental work, rendering the use

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

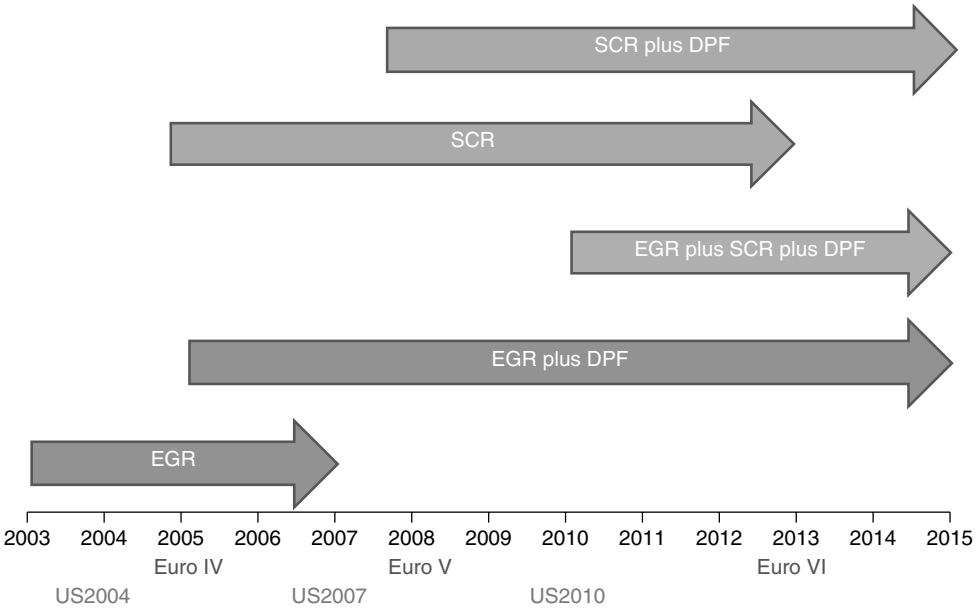


Figure 14. Overview of emission reduction technologies for heavy-duty applications in Europe and the United States.

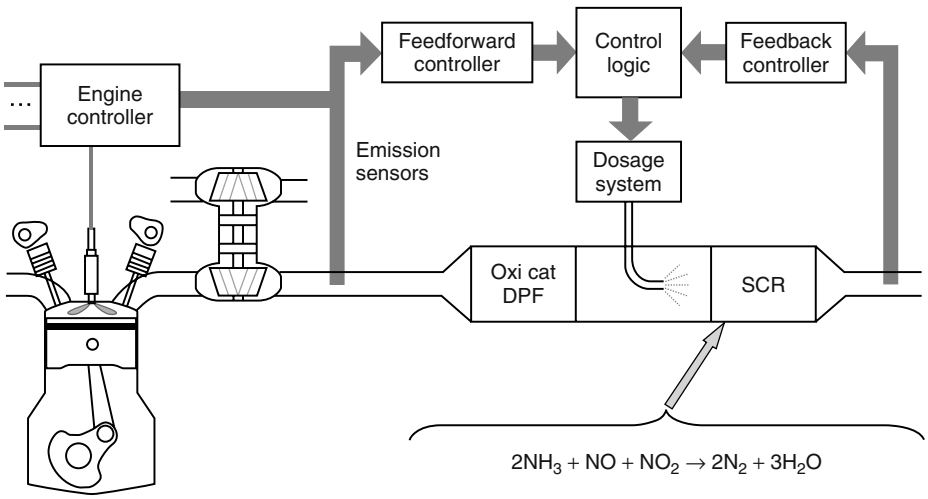


Figure 15. PM filter and SCR system. The pollutant emission sensors are used to close an emission-control feedback loop. (Reproduced from Guzzella, 2009. © Elsevier.)

of advanced simulation tools a necessity. Improved SCR technology with improved reduction agents can reduce the cost of NO*x* aftertreatment systems. Nonprecious metal aftertreatment systems are the hope for future cost and resource savings. However, the pollutant transformation efficiency still needs intensive research work.

An important issue and challenge that arises from the applications described earlier is the increased fraction of NO_2 in the total tailpipe NO*x* emissions that are forecast to be measured at roadside. Bearing in mind the upcoming tightening of the NO_2 air quality limits and the steady increase of traffic volumes, excesses of the NO_2 limits have to be expected to an increasing extent during this decade.

2.9 Alternative fuels

Fuel robustness is a key for the future development of the engine system. Because of the unique, inherent energy efficiency and operating characteristics of the CI engine, diesel is well positioned as the fuel of choice. Introduction

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

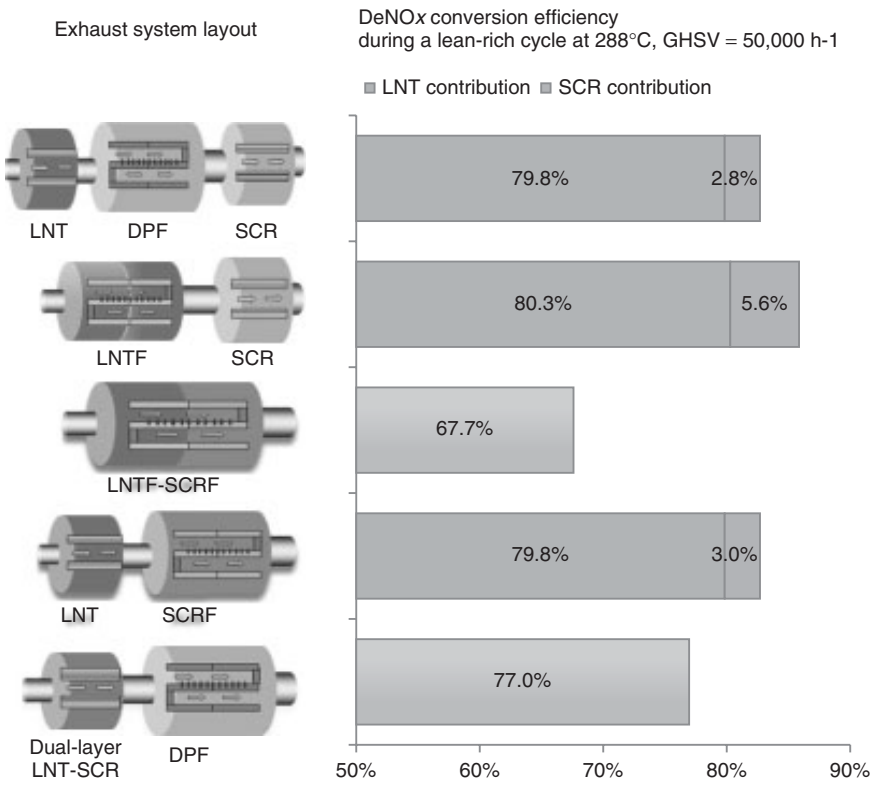


Figure 16. Model-based comparison of DeNO*x* performance of different exhaust system layouts, combining lean NO*x* trap (LNT) and selective catalytic reduction (SCR) functionalities, in wall-flow and flow-through (single-layer or dual-layer) devices. The SCR uses the NH_3 produced during the regeneration phase of the LNT. Definitions: LNTF: DPF with LNT coating, SCRF: DPF with SCR coating. (Reproduced by permission of Exothermia S.A.)

of ultralow sulfur diesel fuels has been a central part of the new clean diesel system designed to meet present and future emissions standards. In Europe, fuel sulfur levels were reduced by 97% from 2000 (300 ppm) to 2009 (8 ppm).

Renewable fuels represent an opportunity to diversify fuels available for CI engines and reduce demand for conventional petroleum-based diesel fuel. CI engines are uniquely capable of operating on a range of renewable fuel feed stocks (algae, biomass, soy-based biodiesel, palm oil, and so on). Most new and existing diesel vehicles and equipment are compatible with biodiesel or renewable diesel fuel blends at ratios ranging from 5% to 20% depending on manufacturer warranties.

Owing to the increasing pressure of rising oil price, decreasing fossil reserves, and the imperative to minimize CO_2 emissions, future fuels will be characterized by the addition of or the substitution by synthetic components (International Energy Agency (IEA), 2010; European Commission (EC), 2006). Contrary to fuels distilled from crude oil, synthetic fuel offers the possibility to be tailored. Parameters relevant for engine combustion can be swayed with purpose, aiming toward lower emissions while optimizing the combustion in terms of efficiency and noise. Synthetic fuels such as gas-to-liquid (GTL), coal-to-liquid (CTL), and biomass-to-liquid (BTL) can be used directly in current engines, as their characteristics are even more favorable than today's diesel fuel. The development of process technologies for synthetic diesel, such as Fischer–Tropsch (FT) and hydrogenation of vegetable oils (HVOs) and of animal fats, enables a wider panel of combustion approaches; thanks to the specific fuel properties of these blend components, especially higher cetane number and low aromatics content. It should be noted, however, that with the increasing share of alternative fuels and vehicles, fuel supply chains diversify and emissions tend to occur further upstream; hence, the responsibility of car manufacturers should shift to maximizing tank-to-wheel efficiency. In addition, as well-to-tank emissions become increasingly relevant with alternative fuels, the capability to reduce emissions and hence responsibility shifts from car manufacturers to fuel and energy suppliers—including refineries and utilities—and these actors are accountable for the carbon content of their respective fuels. All the above need to be done in an efficient, cost-effective, and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

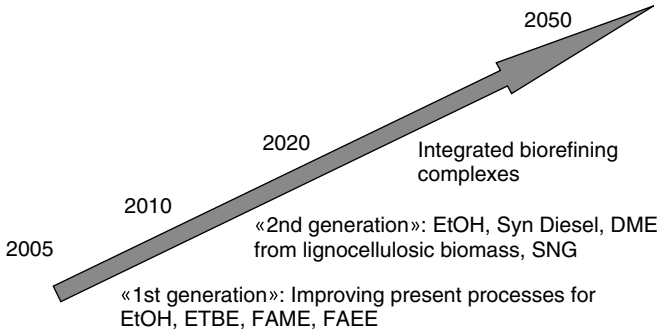


Figure 17. European Union's anticipated future roadmap. (Reproduced from EC, 2006 © European Communities.)

appropriate manner, such that instruments correspond to proper actors and set suitable incentives. Flexible and comprehensive solutions to regulate all GHG emissions and provide a level playing field across all fuels and technologies are currently being discussed.

Figure 17 shows the European Union's anticipated future roadmap for fuels. Also see Fuels for Engines and the Impact of Fuel Composition on Engine Performance for more extensive discussion on the characteristics of conventional and alternative fuels.

2.10 Hybrids

It is well understood that major reductions in GHG emissions in light-duty vehicle transportation sector are required to achieve significant reduction of the total GHG emissions in the mid-term future (Sandy Thomas, 2012; McKinsey & Company, 2007). The alternative vehicle options considered in this framework include hybrid electric vehicles (HEVs) and plug-in hybrid electric vehicles (PHEVs). By combining an internal combustion engine and an electric motor, HEVs have attracted increasing interest by the automotive industry. They are considered in the short term as the most viable alternative propulsion system for significantly improving fuel efficiency and emissions without sacrificing traditional vehicle performance criteria.

CI HEVs and PHEVs are not as common as SI HEVs, which are the mainstream choice. This is due to pollutant legislation issues (in the United States and Japan) and to cost advantages of advanced CI vehicles in Europe (Berggren *et al*., 2009), as well as due to construction nonlinearities. Some manufacturers in the European Union, however, do offer CI HEVS and PHEVs options (Volkswagen Golf TDI, Mercedes-Benz E-Class E300 BlueTEC, Peugeot 3008 Crossover HYbrid4, and Volvo V60 plug) and seem to be more interested in the long term, considering the concept as the very best of what an electric and diesel car can offer: very low fuel consumption and CO_2 levels, combined with long range and high performance.

2.11 Comparison to future SI engines

The development of both SI and CI engines is driven by meeting emission legislation while at the same time maintaining or even improving the efficiency of the CI engine. In spite of specific technical bottlenecks, advanced CI technologies offer better fuel economy over conventional SI technology under all driving conditions, and with no detriment to performance. Besides fuel economy and lower CO_2 emissions, advantages of CI engine include improved performance and towing, and high torque at low engine speed giving "fun-to-drive" characteristics. However, competing developments of advanced SI technologies are likely to erode diesel advantages.

SI engines despite their present drawbacks with respect to their diesel counterpart still offer two fundamental advantages:

- a robust and low cost emission control technology;
- an intrinsic cost advantage.

During the coming decade, the SI engine technology will be the major contributor to the CO_2 emissions reduction of the European vehicle fleet through

- downsizing by means of turbocharging and efficiency improvement by means of electronic valve control and DI of fuel (the feasible reduction in cylinder displacement is up to 40%, with a corresponding benefit in fuel consumption and CO_2 emissions of up to 20%);
- the progressive exploitation of their intrinsic capability to burn low carbon content fuels ranging from natural gas to hydrogen (if energy policy promotes these fuels into the marketplace, SI systems tailored to these fuels will emerge).

Regarding customer expectations, trends are very similar for CI and SI engines, especially in the volume segments and markets. Customers are focused on total cost of ownership, which is determined by such factors as price, resale value, fuel consumption (and fuel price), and maintenance cost, as well as safety, reliability, and durability. In current and future CI engines, there is almost a chemical factory on board and that comes at a price. At the same time, customer expectations regarding "fun to drive" are still increasing. This translates into a continuation of the "power and torque race" to further improve vehicle performance and drivability.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

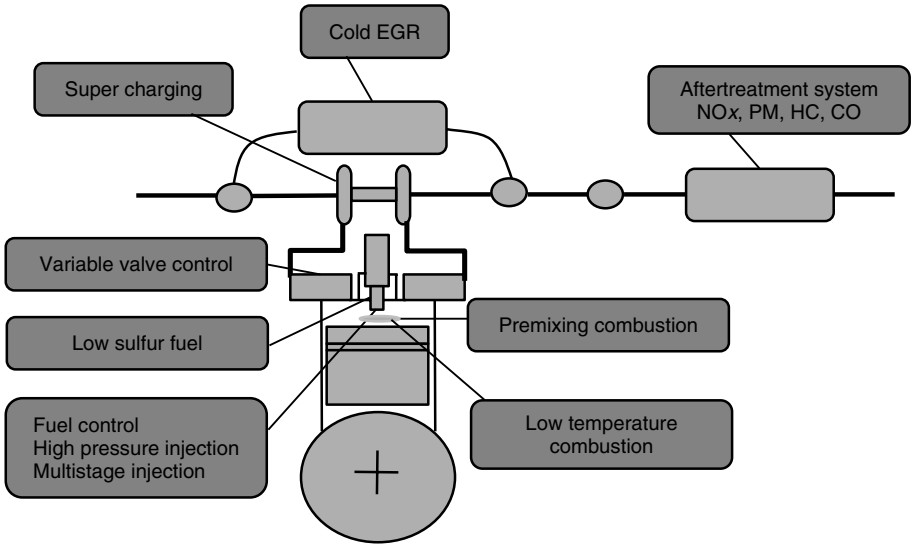


Figure 18. Advanced technologies in CI engine.

Last but not least, possible constraints in regional diesel fuel supply may also brake further expansion of the diesel market. There is indeed a limit as to how far the refinery fraction for diesel can be pushed, and Europe is already a net importer of diesel (and a gasoline exporter to the United States). If the United States begins to drive more CI vehicles, then the shortage of diesel in Europe will become even greater. As a large proportion of the diesel fuel is used by trucking to transport goods that are essential for society, it may well happen that the use of diesel for private motor cars will decline in the future (Lindstedt, 2012).

This growing imbalance in the diesel/gasoline demand ratio, particularly in Europe, initiates the need of additional out-of-the-box options. Research results generally indicate that a special fuel is not needed for advanced combustion CI engines and that future, but still practical hardware, will enable engines to be quite tolerant of a range of fuel properties (e.g., Rose *et al*., 2010). This indicates that multifuel concepts are increasingly necessary, including gasoline-in-CI concepts (Weall and Collings, 2009; Manente *et al*., 2009). Recent practical developments (such as the Mazda Skyactiv system) may be considered as stepping stones in this direction.

3 SUMMARY AND CONCLUSIONS

CI engine has made an outstanding development in Europe by giving the customer at the same time better fuel economy and more driving pleasure. From the society viewpoint, it has provided significant oil savings and CO_2 emission reduction and it has complied with even more stringent emission regulations. This evolution has been supported by a sustained technological improvement.

There are many advanced technologies for high performance engines (most of these are summarized in Figure 18). There are two ways to proceed: optimization and robustness. Flexible FIE, optimized turbocharging, variable geometry engine, reduction of engine friction losses, optimized thermal management, highly premixed and lower temperature combustion, and advanced and high sulfur resistance aftertreatment systems are candidates for the development of a highly efficient and robust CI engine, with near-zero emissions. In the next 20 years, technology innovations such as LTC are expected to play a role in the advancement of CI engines.

Increasing environmental awareness and oil supply tension put CI engines in a challenged position. On the one hand, fuel price increase and developments presented in this chapter can be considered as opportunities. However, on the other hand, diesel fuel shortage and future stringent emission regulations could make cost to customer benefit balance less favorable.

ACKNOWLEDGMENTS

We are indebted to Professor David Foster (University of Wisconsin—Madison) and Dr. Dennis Siebers (Sandia National Laboratories) for their careful and extensive review of our paper. Their corrections, comments, and suggestions for both the literary and scientific parts of our work (with emphasis on the part on LTC) were extremely useful for the overall improvement of our manuscript.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

REFERENCES

Association des Constructeurs Européens d'Automobiles (ACEA) (2011) The automobile industry, http://www.acea.be/images/uploads/files/20110921_Pocket_Guide_3rd_edition.pdf (accessed 20 April 2012).

Automotive Technology Centre (ATC) (2011) *Future of Automotive Powertrains Trends and Developments in advanced powertrain technology*, http://www.acemr.eu/uploads/media/Trendstudy_ACEMR_PowerTrains.pdf (accessed 18 September 2013).

Baumgarten, G. (2006) *Mixture Formation in Internal Combustion Engines*, Springer-Verlag, Berlin, Heidelberg.

Bergsten, L. (2001) SAAB variable compression SVC—variability and control (in German). *MTZ, Motortechnische Zeitschrift*, **62**(6), 424–431.

Brockbank, C. (2009) Application of a variable drive to supercharger & turbo compounder applications. SAE Technical Paper 2009-01-1465.

Chalgren, R. and Barron, L. (2003) Development and verification of a heavy duty 42/14V electric powertrain cooling system. SAE technical paper 2003–01–3416.

Cummins, C.L., Jr, (1993) *Diesel Engines: From Conception to 1918*, vol. 1, Carnot Press, Wilsonville, Oregon.

Cummins Turbo Technologies (2009), http://www.holset.co.uk/mainsite/files/2_4-Turbocharger%20Fundamentals.php (accessed 20 April 2012).

Dec, J.E. (2009) Advanced compression—ignition engines—understanding the in—cylinder process. *Proceedings of the Combustion Institute*, **32**, 2727–2742.

Dieselnet (2012) Emission standards, http://www.dieselnet.com/standards/ (accessed 22 April 2012).

Dieselnet (2013) LTC applications, http://www.dieselnet.com/tech/engine_ltc_app.php (accessed 28 August 2013).

Dober, G., Tullis, S., Greeves, G., *et al.* (2008) The impact of injection strategies on emissions reduction and power output of future diesel engines. SAE Technical Paper 2008-01-0941.

Edwards, S., Eitel, J., Pantow, E., *et al*. (2012) Waste heat recovery: the next challenge for commercial vehicle thermomanagement. SAE Technical Paper 2012-01-1205.

Eichhorn, R., Boot, M., and Luijten, C. (2010) Throttle loss recovery using a variable geometry turbine. SAE Technical Paper 2010-01-1441.

Engineering Matters, Inc (2005) Electromagnetic fully flexible valve actuator.

Erlandsson, O., Lundholm, G., Söderberg, F. *et al*. (1998) Demonstrating the performance and emission characteristics of a variable compression ratio, Alvar-cycle engine. SAE Technical Paper 982682.

ERTRAC (2010) Research and innovation roadmaps. Implementation of the ERTRAC strategic research agenda 2010, http://www.ertrac.org/pictures/downloadmanager/6/50/ertrac-researchinnovation-roadmaps_60.pdf (accessed 15 April 2012).

European Commission (EC) (2006) Biofuels in the European Union. A vision for 2030 and beyond. Final report of the Biofuels Research Advisory Council, http://www.biofuelstp.eu/downloads/biofuels_vision_2030_en.pdf (accessed 02 May 2012).

Fenske, G., Ajayi, L., Erck, R., *et al*. (2009) Overview of Friction and Wear Reduction for Heavy Vehicles. 2009 US DOE Hydrogen Program and Vehicle Technologies Program. Annual Merit Review and Peer Evaluation Meeting. May 18–22, Washington.

Flierl, R., Lauer, F., Breuer, M., and Hannibal, W. (2012) Cylinder deactivation with mechanically fully variable valve train. *SAE International Journal of Engines*, **5**(2), 207–215.

Foster, D.E. (2012) Low Temperature Combustion—A Thermodynamic Pathway to High Efficiency Engines. 122nd meeting of the National Petroleum Council, Washington DC, August.

Galindo, J., Serrano, J.R., Guardiola, C., *et al*. (2011) An on-engine method for dynamic characterisation of NO_x concentration sensors. *Experimental Thermal and Fluid Science*, **35**(3), 470–476.

Greszler, A. (2009) Heavy duty diesel engine and powertrains for 2010 and beyond. SAE 2009 Government/Industry Meeting, http://www.sae.org/events/gim/presentations/2009/anthonygreszler.pdf (accessed 20 May 2012).

Guzzella, L. (2009) Automobiles of the future and the role of automatic control in those systems. *Annual Reviews in Control*, **33**, 1–10.

Heinzen, A., Gillella, P., and Sun, Z. (2011) Iterative learning control of a fully flexible valve actuation system for non-throttled engine load control. *Control Engineering Practice*, **19**(12), 1490–1505.

Hopmann, U. and Algrain, M. (2003) Diesel engine electric turbo compound technology. SAE Technical Paper 2003-01-2294.

Institution for manufacturing Education and Consultancy Services (IfM ECS) (2000) Foresight vehicle technology roadmap. Technology and research directions for future road vehicles, http://www.ifm.eng.cam.ac.uk/ctm/trm/resources.html (accessed 15 April 2010).

International Energy Agency (IEA) (2010) Sustainable production of second-generation biofuels. Potential and perspectives in major economies and developing countries.

Jääskeläinen, H. (2010) Low temperature combustion. Dieselnet (accessed 23 December 2012).

Johnson, T.V. (2008) Diesel emission control in review. SAE Technical Paper 2008-01-0069.

Kawashima, J.-I., Ogawa, H., and Tsuru, Y. (1998) Research on a variable swirl in take Port for 4 valve high speed DI diesel engine. SAE Technical Series 982680.

Kimberley, W. (2005) Bosch pursues diesel technology. *Automotive Design & Production*, **117**(12), 12–13.

Knecht, W. (2008) Diesel engine development in view of reduced emission standards. *Energy*, **33**, 264–271.

Koltsakis, G., Bollerhoff, T., Samaras, Z., and Markomanolakis, I. (2012) Modeling the interactions of soot and SCR reactions in advanced DPF technologies with non-homogeneous wall structure. SAE Technical Paper 2012-01-1298.

Koltsakis, G., Haralampous, O., Tsinoglou, D., *et al*. (2007) First Conference: MinNOx—Minimization of NOx Emissions through. Exhaust Aftertreatment. Berlin.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Lindstedt, G. (2012) Threat of diesel rationing in Europe. Energy Bulletin, http://www.energybulletin.net/stories/2012-03-19/threat-diesel-rationing-europe (accessed 16 May 2012).

Mahr, B. (2002) Future and Potential Diesel Injection Systems. *THIESEL 2002 Conference on Thermo- and Fluid-Dynamic Processes in Diesel Engines*.

Manente, V., Johansson, B., Tunestal, P., and Cannella, W. (2009) Effects of different type of gasoline fuels on heavy duty partially premixed combustion. SAE Technical Paper 2009-01-2668.

McKinsey & Company (2007) A portfolio of power trains for Europe: a fact-based analysis: the role of battery electric vehicles, plug-in hybrids and fuel-cell electric vehicles, http://ec.europa.eu/research/fch/pdf/a_portfolio_of_power_trains_for_europe_a_fact_based__analysis.pdf (accessed 05 May 2012).

Mittal, M., Zhu, G., Stuecken, T., and Schock, J. (2009) Effects of Pre-lnjection on Combustion Characteristics of a Single-Cylinder Diesel Engine. *Proceedings of the ASME, International Mechanical Engineering Congress and Exposition*.

Narayanan, A.P. (2011) Downspeeding the diesel engine—a performance analysis. Master's Thesis in Automotive Engineering. Division of Combustion, Chalmers University of Technology, Göteborg, Sweden.

Ostrowski, G., Neely, G., Chadwell, C., *et al*. (2012) Downspeeding and supercharging a diesel passenger car for increased fuel economy. SAE Technical Paper 2012-01-0704.

Page, R.W., Hnatczuk, W., and Kozierowski, J. (2005) Thermal management for the 21st century—improved thermal control & fuel economy in an army medium tactical vehicle. SAE Technical Paper 2005-01-2068.

Patton, K.J., Manuel, A., and Gonlazez, D. (2010) Development of high-efficiency clean combustion engines designs for SI and CI engines. GM Powertrain Advanced Engineering, June, http://www1.eere.energy.gov/vehiclesandfuels/pdfs/merit_review_2010/high-eff_engine_tech/ace036_patton_2010_o.pdf (accessed 24 May 2012).

Pesic, R.B., Milojevic, S.T., and Veinovic, S.P. (2010) Benefits and challenges of variable compression ratio at diesel engines. *Thermal Science*, **14**(4), 1063–1073.

Plianos, A. and Stobart, R. (2008) Modeling and control of diesel engines equipped with a two-stage turbo-system. SAE Technical Paper 2008-01-1018.

Rose, K., Cracknell, R., Rickeard, D. *et al*. (2010) Impact of fuel properties on advanced combustion performance in a diesel bench engine and demonstrator vehicle. SAE Technical Paper 2010-01-0334, doi:10.4271/2010-01-0334.

Sandy Thomas, C.E. (2012) How green are electric vehicles?. *International Journal of Hydrogen Energy*, **37**, 6053–6062.

Schechter, M.M. and Levin, M.B. (1996) Camless engine. SAE Technical Paper 960581.

Schulte, H. and Wirth, M. (2010) Internal combustion engines for the future, http://essaysforstudent.com/print.html?essay=77790 (accessed 20 April 2012).

Tahani, M., Javan, S., and Biglari, M. (2012) A comprehensive study on waste heat recovery from internal combustion engines using organic Rankine cycle. Thermal Science OnLine-First Issue 00, http://thermalscience.vinca.rs/pdfs/papers-2012/TSCI120114052H.pdf (accessed 02 May 2012).

Tai, C. and Tsao, T. (2002) Control of an Electromechanical Camless Valve Actuatuor. *Proceedings of the American Control Conference*, Anchorage, AK May 8–10.

The International Council on Clean Transportation (ICCT) (2011) Campestrini, M. and Mock, P. European vehicle market statistics, http://www.theicct.org/sites/default/files/publications/Pocketbook_LowRes_withNotes-1.pdf (accessed 20 April 2012).

Turner, C., Babbitt, G., Balton, C. *et al*. (2004) Design and control of a two-stage electro-hydraulic valve actuation system. SAE Technical Paper 2004-01-1265.

Uchida, H. (2006) Trends of the turbocharging technology. *R & D Review of Toyota CRDL*, **41**(3), 1–8.

Weall, A. and Collings, N. (2009) Gasoline fuelled partially premixed compression ignition in a light duty multi cylinder engine: a study of low load and low speed operation. SAE Technical Paper 2009-01-1791.

Xin, Q. (2011) Overview of diesel engine applications for engine system design - part 1: systems engineering and rational considerations of product R&D organization design. SAE Technical Paper 2011-01-2181.

Zheng, L. (2007) Camless variable valve actuation designs with two-spring pendulum and electrohydraulic latching. SAE Technical Paper 2007-01-1295.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto208
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Induction Motor Drives

Ming Cheng[1], Zheng Wang[1], and Yuk Sum Wong[2]

[1] *Southeast University, Nanjing, China*
[2] *State Grid Energy Research Institute, Beijing, China*

1 Introduction 1
2 Structure and Operation Principle 2
3 Steady-State Performance 4
4 Power Inverters 7
5 Control Strategies of IM Drive for EV 15
6 Conclusions 21
Related Articles 22
References 22
Further Reading 22

1 INTRODUCTION

An induction (or asynchronous motor) is a commutatorless AC motor in which all electromagnetic energy is transferred by inductive coupling from a primary winding to a secondary winding, the two windings being separated by an air gap. The history of construction and application of induction motors (IMs) in electric drive dates back as far as over 100 years. At present, IM drives are the most mature technology among various commutatorless motor drives. The IM drives offer the advantages of robust structure, low cost, high reliability, and being free from maintenance as compared with the other types of electric motor drives. These advantages are particularly important for electric propulsion in electric vehicle (EV) applications. Hence, IM drives are popular choices for traction applications. However, as compared with permanent magnet (PM) motors, IMs have lower efficiency and less torque density.

There are two types of IMs, namely, wound-rotor and squirrel-cage motors. Owing to the high cost, need for maintenance, and lack of sturdiness, wound-rotor IMs are less attractive than their squirrel-cage counterparts, especially for electric propulsion in EVs. Hence, squirrel-cage IMs are loosely named as IMs.

An inverter is used to convert the energy from the batteries to the motor so that the desired torque can be delivered for a given driving condition at a certain speed. Advanced control methodologies, such as vector control and direct torque control (DTC), are popular in IM drives for electric propulsion applications.

IMs used for electric propulsion are principally similar to that used for other industrial applications. Nevertheless, these IMs should be specially designed with reference to the requirements of electrical drivetrain in EVs. Laminated thin silicon cores should be used for the rotor and stator to reduce the iron loss, whereas copper bars are preferred for the squirrel cage to reduce the winding loss. Reasonable high voltage low current motor design should be employed to reduce the cost and size of the power inverter, although the voltage level of the motor is limited by the number, weight, and type of EV batteries. High speed operation should be adopted to minimize the motor size and weight, although the maximum speed of the motor is limited by the bearing friction and windage losses. Low stray reactance is also necessary to favor flux-weakening operation. Concerning motor performances for EV operation, high torque at low speeds, low torque at high speeds, and instantaneous overloading capability are desired for hill climbing, highway cruising, and vehicle overtaking, respectively (Chan and Chau, 2001). The key differences between

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Table 1. Key differences between EV IMs and standard industrial IMs.

Items	EV IMs	Industrial IMs
Ambient temperature (°C)	−40–+125	−20–+40
Coolant temperature (°C)	75–150	<40
Mounting base	Mobile	Stationary
Protection	Salt, spray, explosive	No special
Acceleceration/ decceration	Frequent	Most stable
Motor size and mass	Tight limited	Relaxed
System voltage	Isolated and varied	Stable network
Acoustic noise level	Very low	Low
Speed range (rpm)	0–15,000 or higher	<3600

Data taken from "Electric Machine Systems for Powertrains of New Energy Vehicles" presented in the 2nd China International New Energy Vehicle Forum 2013 by Dr. William CAI.

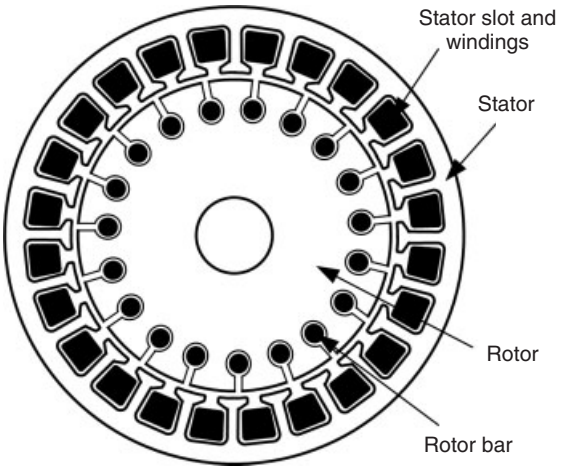


Figure 1. Cross section of an induction motor.

EV traction IMs and standard industrial IMs can be summarized as shown in Table 1.

2 STRUCTURE AND OPERATION PRINCIPLE

2.1 Structure

A cross section of an IM is shown in Figure 1. An IM is basically composed of two main parts, namely the stator and the rotor, between which there is a uniform air gap. Both the stator and the rotor are made of laminated silicon steel with thickness of 0.35, 0.5, or 0.65 mm. The laminated steel sheets are first stamped with slots and are then stacked together to form the stator and the rotor, respectively.

There are some additional components to make up the whole machine: the housing that encloses and supports the whole machine, the shaft that transfers torque, the bearing, an optional position sensor, and a cooling mechanism (such as a fan or liquid cooling tubes) (Mi, Masrur, and Gao, 2011).

It should be emphasized that the robust mechanical design and reduction of noise and vibration are indispensable for a traction motor because exposure to temperature change and external impact can result in mechanical fatigue on motor structure, and abnormal sound and movement of motor during driving can make people uncomfortable. In addition, it is required that a traction motor be of a compact size owing to the limitation in the installation space.

As shown in Figure 1, slots in the inner periphery of the stator are inserted with three-phase windings. The turns of each winding are distributed such that the current in the winding produces an approximately sinusoidally distributed flux density around the periphery of the air gap. The three-stator windings are displaced from each other by 120 electrical degrees in space around the air gap.



Figure 2. Rotor squirrel cage of an induction motor.

The most common types of IM rotors are the squirrel-cage rotors in which aluminum bars are cast into slots in the outer periphery of the rotor. The aluminum bars are short-circuited together at both ends of the rotor by cast aluminum end rings, which can also be shaped as fans (Ehsani, Gao, and Emadi, 2009). Figure 2 shows a rotor squirrel cage.

2.2 Operation principle

Figure 3a shows a simplified two-pole three-phase IM. The concentrated full-pitch coils, displaced from each other by 120° in space, are shown by coils a–x, b–y, and c–z, which may be considered to represent distributed windings producing sinusoidal magnetomotive force (MMF) waves on the magnetic axes of the respective phases.

The three-phase windings are excited by three-phase AC sinusoidal current as

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

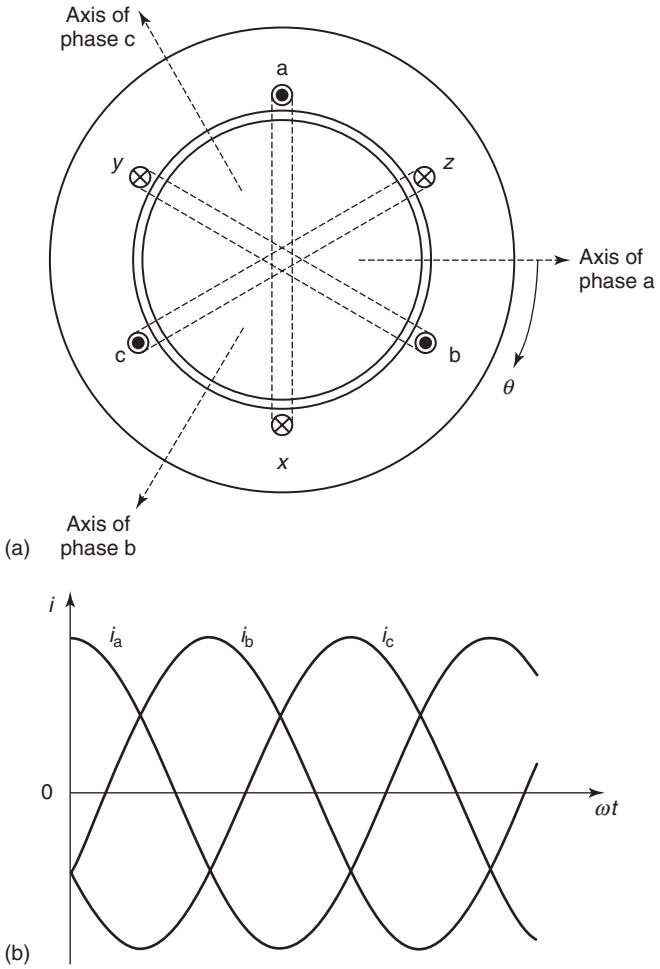


Figure 3. Simplified two-pole three-phase stator winding current: (a) spatially symmetric three-phase stator windings; (b) phase currents.

$$\left.\begin{aligned} i_a &= I_m \cos(\omega t) \\ i_b &= I_m \cos(\omega t - 120^\circ) \\ i_c &= I_m \cos(\omega t + 120^\circ) \end{aligned}\right\} \quad (1)$$

where I_m is the amplitude of the phase currents and ω the angular frequency of the current. The instantaneous currents are shown in Figure 3b. Each of the three-phase currents will generate an MMF, which is a space vector. As the three windings are located 120° from each other in space along the inside surface of the stator, the MMF generated by each phase can be written as follows:

$$\left.\begin{aligned} f_a &= F_{m1} \cos(\omega t) \cos(\theta) \\ f_b &= F_{m1} \cos(\omega t - 120^\circ) \cos(\theta - 120^\circ) \\ f_c &= F_{m1} \cos(\omega t + 120^\circ) \cos(\theta + 120^\circ) \end{aligned}\right\} \quad (2)$$

where $F_{m1} = \frac{2}{\pi}\frac{wk_{w1}}{p} I_m$ is the amplitude of the space-fundamental MMF of each phase, θ the spatial angle measured with respect to the rotor magnetic axis, w the turn number of phase winding in series, k_{w1} the fundamental winding factor, and p the number of pole pairs. Use of a common trigonometric identify permits Equation 2 to be rewritten in the form

$$\left.\begin{aligned} f_a &= \frac{1}{2}F_{m1}\cos(\omega t - \theta) - \frac{1}{2}F_{m1}\cos(\omega t + \theta) \\ f_b &= \frac{1}{2}F_{m1}\cos(\omega t - \theta) - \frac{1}{2}F_{m1}\cos(\omega t + \theta + 120^\circ) \\ f_c &= \frac{1}{2}F_{m1}\cos(\omega t - \theta) - \frac{1}{2}F_{m1}\cos(\omega t + \theta - 120^\circ) \end{aligned}\right\} \quad (3)$$

The three MMFs can be summed to form

$$f = f_a + f_b + f_c = \frac{3}{2}F_{m1}\cos(\omega t - \theta) \quad (4)$$

Equation 4 indicates that the magnetic field is rotating along the inner surface of the stator with the frequency of the angle velocity ω, and its magnitude is $3/2F_{m1}$.

Figure 4 graphically shows the stator MMF vectors at $\omega t = 0$, $\omega t = \pi/3$, and $\omega t = 2\pi/3$. At the moment $\omega t = 0$, the current i_a is at its maximum value I_m, and i_b and i_c are both $I_m/2$ in the negative direction, as shown by the dots and crosses in Figure 4, indicating the actual instantaneous directions. The corresponding MMFs are shown by the vectors F_a, F_b, and F_c, respectively, where $F_a = F_{m1}$ drawn along the magnetic axis of phase a, whereas F_b and F_c are both of magnitude $F_{m1}/2$ drawn in the negative direction along the magnetic axes of phase b and phase c, respectively. The resultant, obtained by adding the individual contributions of the three phases, is a vector of magnitude $F = \frac{3}{2}F_{m1}$ centered on the axis of phase a. It represents a sinusoidal space wave with its positive peak centered on the axis of phase a and having an amplitude 1.5 times that of phase a contribution alone. At a later moment $\omega t = \pi/3$, the currents i_a and i_b are a positive half maximum, and the current i_c is a negative maximum. The resultant MMF vector has the same amplitude as at $\omega t = 0$, but it has rotated clockwise by 60 electrical degrees in space. Similarly, at $\omega t = 2\pi/3$, the current i_b is a positive maximum, and i_a and i_c are a negative half maximum; the same resultant MMF vector is again obtained, but it has rotated clockwise by 60 electrical degrees still further and is now aligned with the magnetic axis of phase b. Obviously, as time passes, the resultant MMF wave retains its sinusoidal form and amplitude, but rotates progressively along the air gap. Thus, the net result is an MMF wave of constant amplitude rotating at a uniform angular velocity (Fitzgerald, Kingsley, and Umans, 2003).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

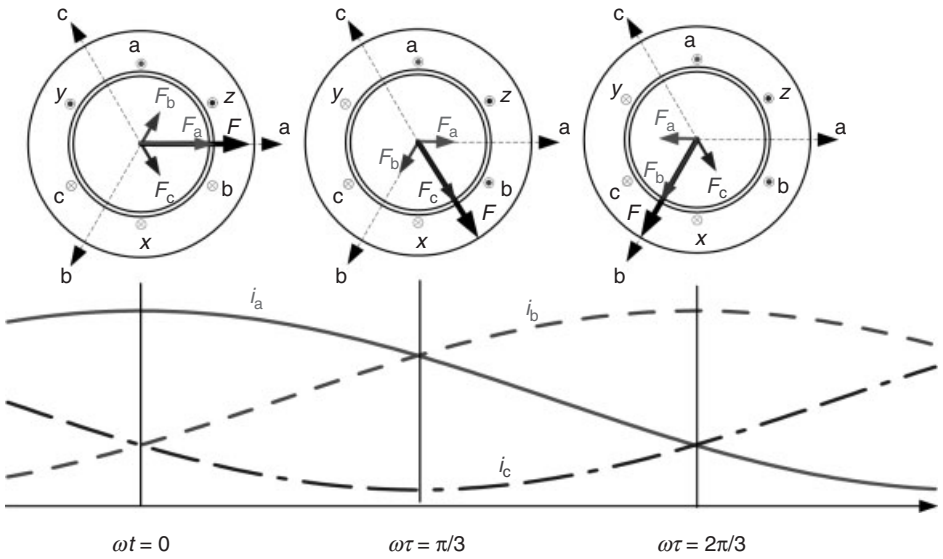


Figure 4. Rotating magnetic field produced by three-phase currents.

As $\omega = 2\pi f$, the rotating speed of the field will be the same as the supply frequency: f revolutions per second or $n_s = 60f$ revolutions per minute (rpm). Noting that the above derivation is based on one pair of poles, a more generic equation for the field speed (or synchronous speed) of an induction machine can be given as $n_s = \frac{60f}{p}$ and

$$\omega_s = \frac{2\pi n_s}{60} = \frac{\omega}{p} = \frac{2\pi f}{p} \tag{5}$$

If the rotor is turning at the steady-state speed of n rpm in the same direction as the rotating stator field, then the relative speed between the stator rotating field and the rotor is $n_s - n$. This difference between synchronous speed and the rotor speed is commonly referred to as *slip* of the rotor, which is more usually expressed as a fraction of synchronous speed

$$s = \frac{n_s - n}{n_s} \tag{6}$$

where s is slip. The rotor speed can be expressed in terms of the slip and the synchronous speed as

$$n = (1 - s)n_s \tag{7}$$

Assuming initially that the rotor is stationary, the rotating stator field will induce an electromotive force (EMF) inside the rotor bars of the squirrel cage. A current is therefore formed inside the rotor bars through the end rings. The reaction between the rotating stator MMF and the rotor currents will generate a torque on the rotor. If the torque is large enough, the rotor will start to rotate. The rotor accelerates until the magnitude of induced rotor current and torque balances the applied load. As rotation at synchronous speed does not result in any induced rotor current, an IM always operates slower than the synchronous speed. Typical slips of IMs are within 1–3% (Mi, Masrur, and Gao, 2011). When the load increases, the speed drops and the slip increases enough to create sufficient torque to turn the load, and vice versa. For this reason, IMs are sometimes referred to as *asynchronous motors*.

The essential character of an induction machine is that it is created solely by induction rather than being separately excited as in synchronous or DC machines or being self-magnetized as in PM motors.

An induction machine can be operated as an induction generator for regenerative braking in EV applications.

3 STEADY-STATE PERFORMANCE

3.1 Equivalent circuit

As the three phases are symmetrical, the equivalent circuit can be derived for one phase, with the understanding that the voltages and currents in the remaining phases can be found simply by phase shift of 120°. An IM is simply an electrical transformer in the magnetic circuit, in which the stator winding and the moving rotor winding are separated by an air gap and the transformer convention is used, as shown in Figure 5a. It is worth noting that the rotor and the stator quantities will have different frequencies except when the rotor is stationary.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

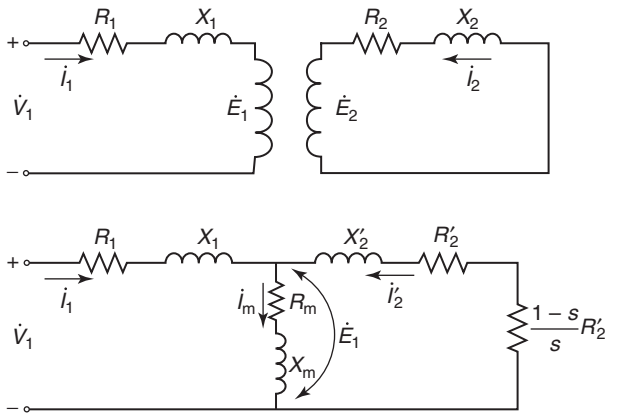


Figure 5. Equivalent circuit of an induction motor. (a) Stator and rotor circuits; (b) T-equivalent circuit.

The voltage equation of the stator and rotor circuit can be written as

$$\begin{aligned}\dot{V}_1 &= \dot{I}_1 R_1 + jX_1\dot{I}_1 + \dot{E}_1 \\ 0 &= \dot{I}_2 R_2 + jX_2\dot{I}_2 + \dot{E}_2\end{aligned} \tag{8}$$

where $\dot{V}_1$ is the stator phase voltage, $\dot{I}_1$ and $\dot{I}_2$ the stator and rotor currents, respectively, R_1 and R_2 the phase resistance of stator and rotor respectively, $X_1 = \omega_1 L_1$ and $X_2 = \omega_2 L_2$ the leakage reactance of stator and rotor respectively, and $\dot{E}_1$ and $\dot{E}_2$ phase EMF of stator and rotor respectively.

The circuit shown in Figure 5a is inconvenient for general use because of the presence of two frequencies, namely ω_1 in stator and $\omega_2 = s\omega_1$ in rotor. If both sides of the second equation in Equation 8 is multiplied by k and divided by s, it yields

$$0 = \left(k^2\frac{R_2}{s}\right)\frac{\dot{I}_2}{k} + j\left(k^2\frac{X_2}{s}\right)\frac{\dot{I}_2}{k} + \frac{k\dot{E}_2}{s} \tag{9}$$

By using the following, $R'_2 = k^2 R_2$, $X'_2 = k^2 X_2/s = k^2 s\omega_1 L_2/s = k^2\omega_1 L_2$, $I'_2 = X_2/k$, and $E'_2 = kE_2/s$, Equation 9 becomes

$$\begin{aligned}0 = \frac{\dot{I}'_2 R'_2}{s} + j\dot{I}'_2 X'_2 + \dot{E}'_2 &= \dot{I}'_2\left(R'_2 + \frac{1-s}{s}R'_2\right) \\ &+ j\dot{I}'_2 X' + \dot{E}'_2\end{aligned} \tag{10}$$

which suggests a single frequency equivalent circuit operating at stator frequency ω_1. By choosing k such that $\dot{E}_1 = \dot{E}'_2$, the equivalent circuit can be redrawn as shown in Figure 5b, where $\frac{1-s}{s}R'_2$ is the equivalent electromechanical power resistance representing the mechanical load, X_m the magnetizing reactance, and R_m the equivalent magnetizing loss resistance. The rotor primed variables are said to be referred to the stator side.

3.2 Torque characteristics

In the circuit shown in Figure 5b, for a given voltage supply, the current of the circuit can be derived as

$$\dot{I}_1 = \frac{\dot{V}_1}{R_1 + jX_1 + (R_m + jX_m)//(R'_2/s + jX'_2)} \tag{11}$$

To simplify the analysis, we can neglect $R_m + jX_m$. Under this assumption, $I_1 = I'_2$. Thus, the electromagnetic power transferred from the stator to the rotor is

$$\begin{aligned}P_{em} &= mI_1^2\frac{R'_2}{s} = \frac{mV_1}{(R_1 + R'_2/s)^2 + (X_1 + X'_2)^2}\frac{R'_2}{2} \\ &= \frac{mV_1^2}{(R_1 + R'_2/s)^2 + (X_1 + X'_2)^2}\left[R'_2 + \frac{1-s}{s}R'_2\right] \\ &= p_{Cu2} + P_{mech}\end{aligned} \tag{12}$$

where m is the number of phases. Note that electromagnetic power or rotor power can be divided into two parts: the first term p_{Cu2} is the rotor copper loss and the second term P_{mech} is the total mechanical power on the shaft. Thus, the electromagnetic torque of the motor can be written as

$$T_{em} = \frac{P_{mech}}{\omega_m} = \frac{m}{\omega_m}\frac{V_1^2}{(R_1 + R'_2/s)^2 + (X_1 + X'_2)^2}\frac{1-s}{s}R'_2 \tag{13}$$

where ω_m is the angular velocity of the rotor. Figure 6 shows the torque–slip characteristics of an IM with fixed voltage and frequency, in which the normal motoring region, the generating region, and the braking region are illustrated. In normal motor operation, the rotor revolves in the direction of rotation of the magnetic field produced by the stator currents, the speed is between zero and synchronous speed, and the corresponding slip is between 1.0 and 0. In the region of $0 < s < s_m$, where s_m is the slip corresponding to the maximum torque, the torque increases approximately linearly with the increase in slip until reaching its maximum at $s = s_m$, then it decreases with the further increase in the slip. At $s = 1$, the rotor speed is zero and the corresponding torque is the starting torque, which is less than its torque at $s = s_m$. In the region of $s > 1$, the rotor torque is positive and decreases further with the increase in slip, and the rotor speed is negative, according to Equation 7. Thus, in this region, the operation of the motor is reverse braking. The induction machine operates as a generator if its rotor is driven above synchronous speed, corresponding to the slip region of $s < 0$ or forward braking region. One such operation mode is that a vehicle goes

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

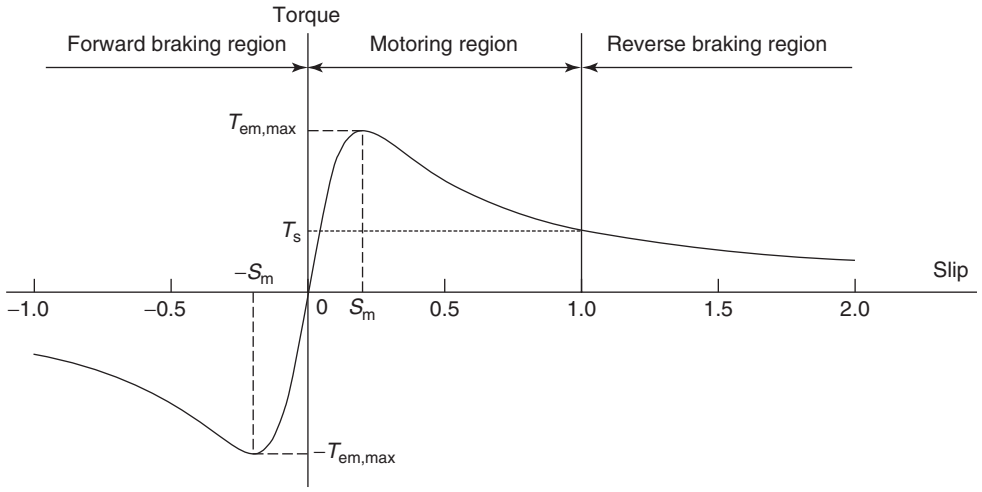


Figure 6. Torque–slip characteristic of induction motor with fixed stator voltage and frequency.

down a long slope and drives the induction machine speed higher than the synchronous speed.

Obviously, the torque–speed characteristic of an IM at fixed voltage and fixed frequency is not appropriate to vehicle traction applications because of the low starting torque, limited speed range, and unstable operation in the range of $s > s_m$, where any additional disturbing torque in the load will lead the machine to stop as the torque decreases with the speed decreasing characteristically. The high slip also results in high current, which may cause damage in the stator windings. Actually, the fixed voltage and frequency IM are usually operated in the narrow slip range of $0 < s < s_m$. Thus, for traction application, an IM must be controlled by an electronic inverter to provide proper torque–speed characteristics (Zhu and Howe, 2007).

3.3 Efficiency and losses

In the motor operation mode, the active power drawn from the supply can be expressed as

$$P_1 = mV_1I_1\cos\varphi_1 \tag{14}$$

where m is the phase number and φ_1 the phase angle between the stator voltage and the current. This power cannot be fully transmitted to the mechanical power on the shaft because of various losses in the IM. There are five types of losses in induction machine, namely,

(i) copper loss in the stator winding $p_{Cu1} = mI_1^2R_1$, which is the power dissipated as heat in the stator windings;
(ii) magnetizing loss in the stator iron (or core loss or iron loss) $p_C = mI_m^2R_m$, which is the power dissipated as heat in the stator core;
(iii) copper loss in the rotor winding $p_{Cu2} = mI'^2_2R'_2$, which is the power dissipated as heat in the rotor cage;
(iv) windage loss p_{fw} due to the rotation of the rotor and frictional loss in the bearing; and
(v) additional losses p_{ad}, which cannot be accounted for by the above components, also called *stray load loss*.

The power balance equations are

$$P_{em} = P_1 - p_{Cu1} - p_C \tag{15}$$

$$P_{mech} = P_{em} - p_{Cu2} \tag{16}$$

$$P_2 = P_{mech} - p_{fw} - p_{ad} \tag{17}$$

where P_2 is the output power to the load connected to the shaft. The power flow in an IM is shown in Figure 7. The

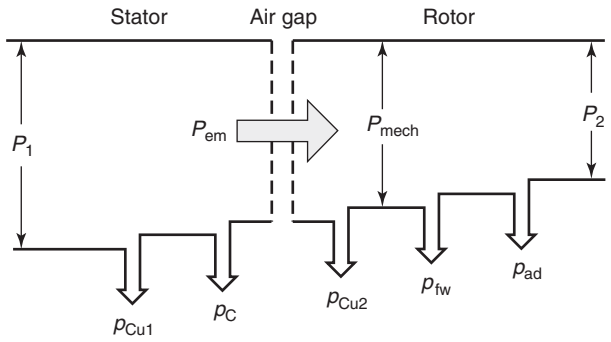


Figure 7. Power flow in an induction motor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

efficiency of an IM can be expressed as

$$\eta = \frac{P_2}{P_1} = \frac{P_2}{P_2 + p_{Cu1} + p_C + p_{Cu2} + p_{fw} + p_{ad}} \quad (18)$$

It should be noted that laminated silicon steel sheets were traditionally designed for use at low frequencies (50 or 60 Hz), and today's traction drives in modern HEV typically operate at about 6000–15,000 rpm. With four-pole machines, the operating frequency is 500 Hz. Some traction motors operate at frequencies as high as 800–1200 Hz. As eddy current loss and hysteresis loss are proportional to frequency or the square of frequency, the core loss will be significant at high frequencies. In order to keep the core loss within a reasonable range, the magnetic flux in the iron core has to be relatively lower than those used in low speed motors, and the thickness of the silicon steel sheets may have to be reduced as well.

Moreover, the IM used in vehicle traction applications is operated by inverter, which can result in harmonics in its voltage and current. These harmonics will introduce additional losses in the winding and iron core of stator and rotor. As is well known, the eddy current loss can be doubled in many IMs because of the pulse-width–modulation (PWM) supply. These additional losses may cause excessive temperature rise, which must be considered during the design and analysis of IMs (Mi, Masrur, and Gao, 2011).

In a word, more losses are generated in traction IMs than industrial IMs because of their high current density, increased harmonic components of current, and so on. This may cause insulation to become brittle and apply high mechanical stresses on the bars and end rings because of differential thermal expansion, if a motor is not properly cooled and operates above the threshold temperature of insulation with long time. This phenomenon can lead to motor failure. Particularly, it is known that insulation life decreases by 50% for every 10°C increase above the threshold temperature. Therefore, a totally enclosed liquid cooled system is usually adopted in traction IMs, and liquid cooling channels are located in the housing. Heat generated in the stator core and stator coil is cooled directly through cooling liquid in housing, and heat in the rotor part is also cooled through heat exchange of internal air (Kim *et al*., 2012).

4 POWER INVERTERS

The evolution of power inverter topologies follows that of power devices, aiming to achieve high power density, high efficiency, high controllability, and high reliability. The IM drives are necessary to offer the torque–speed requirements of the EV driving profile without involving variable gearing or gearbox. Besides, the EV motor drive should provide the capability of bidirectional power flow to recover the regenerative braking energy. The wish list of the power inverters driving EV motor includes efficiency over 95%, power density over 3.5 W/cm^3, switching frequency over 10 kHz, dv/dt below 1000 V/μs, zero electromagnetic interference (EMI), zero failure before the end of the vehicle life, and redundance with limp-home mode (Chau and Wang, 2005).

4.1 Voltage source inverters

The three-phase two-level voltage source inverter has been widely used in industry for many different applications. As shown in Figure 8, this power inverter is composed of six switches, with an antiparallel free-wheeling diode for each switch. The switches can be metal–oxide–semiconductor field-effect transistor (MOSFET) or insulated-gate bipolar transistor (IGBT) devices, depending on the power and voltage ratings of the converter. The two-level power inverters transform the fixed DC link voltage to a three-phase AC voltage with variable magnitude and frequency for satisfying the operating requirements of IM for EV.

The modulation strategies are used for the two-level voltage source inverters (VSIs) to transform the fixed DC voltage into the desired AC voltages. The sinusoidal pulse-width modulation (SPWM) and the space vector modulation (SVM) are the two widely used modulation strategies. Figure 9 shows the principle of the SPWM for the two-level VSIs. The switching signals of the three legs are generated by comparing the three-phase sinusoidal modulating signals v_{mA}, v_{mB}, and v_{mC} with a common triangle carrier v_{cr}. The waveforms v_{AN} and v_{BN} are the switching signals for upper switches in legs A and B. The conduction of the lower switches is complimentary to that of the upper switches in the same legs. v_{AB} is the line waveform between phases

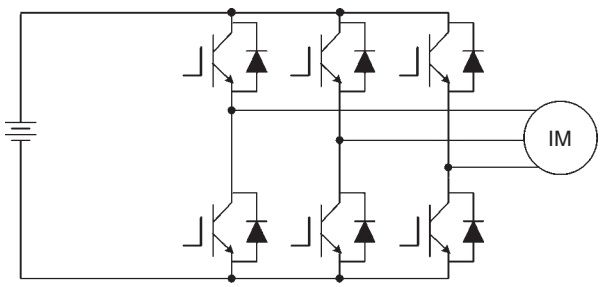


Figure 8. Configuration of three-phase two-level voltage source inverter.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

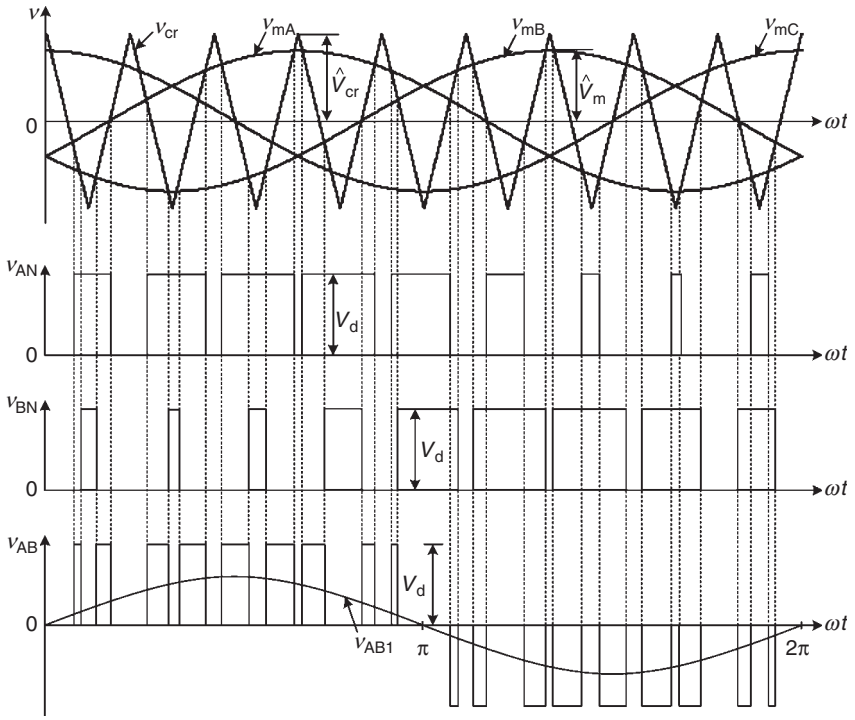


Figure 9. Principle of SPWM.

A and B. v_{AN} and v_{BN} in Figure 9 also show the two-level switched waveforms in each phase. The fundamental-frequency component in the inverter output voltage can be controlled by the amplitude-modulation index m_a, which is defined as the peak value of the modulating wave $\widehat{V}_m$ to the peak value of carrier wave $\widehat{V}_{cr}$. The frequency-modulation index m_f is defined as the frequency of the modulating wave to the frequency of the carrier wave.

Another popular modulating strategy for the three-phase two-level VSI is SVM. Figure 10 shows the vector diagram

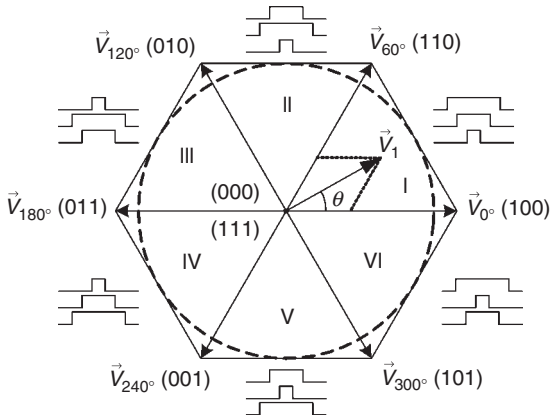


Figure 10. Vector diagram of SVM for two-level voltage source inverter.

of the SVM. The SVM is one of the real-time modulation techniques and is widely used for digital control of VSIs. There are eight switching states of the three upper-leg switches: $\vec{V}_0 = (0,0,0)$, $\vec{V}_1 = (1,0,0)$, $\vec{V}_2 = (1,1,0)$, $\vec{V}_3 = (0,1,0)$, $\vec{V}_4 = (0,1,1)$, $\vec{V}_5 = (0,0,1)$, $\vec{V}_6 = (1,0,1)$, and $\vec{V}_7 = (1,1,1)$. The three lower-leg switches have the complementary switching states. In each switching period, the sequence of the switching states is $\vec{V}_0(T_0/4) \rightarrow \vec{V}_A(T_1/2) \rightarrow \vec{V}_B(T_2/2) \rightarrow \vec{V}_7(T_0/2) \rightarrow \vec{V}_B(T_2/2) \rightarrow \vec{V}_A(T_1/2) \rightarrow \vec{V}_0(T_0/4)$, where $\vec{V}_A$ and $\vec{V}_B$ are the active switching states. The active switching states correspond to the two adjacent basic voltage vectors between which $\vec{V}_1$ locates. The dwell time for the switching vectors essentially represents the duty-cycle time (on-state or off-state time) of the chosen switches during a sampling period T_s of the modulation scheme. The dwell time calculation is based on "voltage-second balancing" principle. That is the product of the reference voltage $\vec{V}_1$ and the sampling period T_s equals the sum of the voltage multiplied by the time interval of chosen space vectors. The dwell times T_1, T_2, and T_0 are computed by:

$$T_1 = \frac{2|\vec{V}_1|T_s}{\sqrt{3}|\vec{V}_A|}\sin(60^\circ - \theta) \tag{19}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

$$T_2 = \frac{2|\vec{V}_1|T_s}{\sqrt{3}|\vec{V}_B|}\sin\theta \quad (20)$$

$$T_0 = T_s - T_1 - T_2 \quad (21)$$

where θ is the phase angle of $\vec{V}_1$ and T_s the switching period.

4.2 Multilevel power inverters

Compared to the two-level VSI, the multilevel power inverters can offer some unique advantages that are particularly beneficial to EV. Namely, they can generate near-sinusoidal voltages with only fundamental-frequency switching, produce less EMI, are suitable for high power motor drives, and fit well with battery-powered EV where floating DC sources are naturally available. There are three main types of multilevel inverters, namely the neutral point diode-clamping (NPC) type, the flying-capacitor (flying-source) type, and the cascaded H-bridge (CHB) type.

Figure 11 shows the configuration of the three-level NPC inverter. Each inverter leg is composed of four active switches with four antiparallel diodes. On the DC side of the inverter, the DC bus capacitor is split into two, providing a neutral point Z. The diodes connected to the neutral point, namely D_{z1} and D_{z2} are the clamping diodes. When the middle two switches in each leg conduct, the phase is connected to the neutral point through one of the clamping diodes. The terminal voltage of the corresponding phase is zero. Such switching state is represented by O. When the upper two switches conduct, the inverter terminal voltage is $+E$, and the switching state is represented by P. When the lower two switches conduct, the inverter terminal voltage is $-E$, and the switching state is represented by N. The switches S_{1A} and S_{3A} operate in a complementary manner, whereas the switches S_{2A} and S_{4A} operate in a complementary manner. The same operating principle works for phase B and C.

The SVM for the three-level NPC inverter is more complex. There are a total of 27 possible combinations of switching states (Feng *et al*., 2004). Based on their magnitudes (length), the voltage vectors can be divided into four groups: zero vector with magnitude of zero, small vectors with magnitude of $V_d/3$, medium vectors with magnitude of $\sqrt{3}V_d/3$, and large vectors with magnitude of $2V_d/3$. Figure 12 shows the vector diagram of the

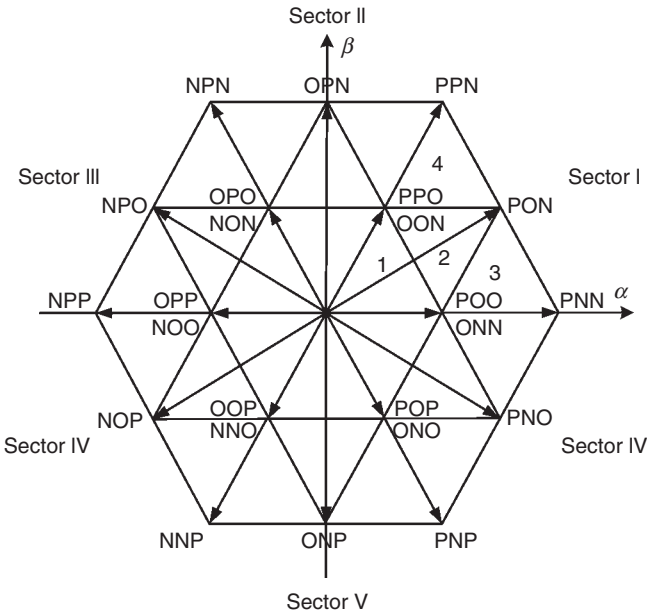


Figure 12. Vector diagram of three-level NPC inverter.

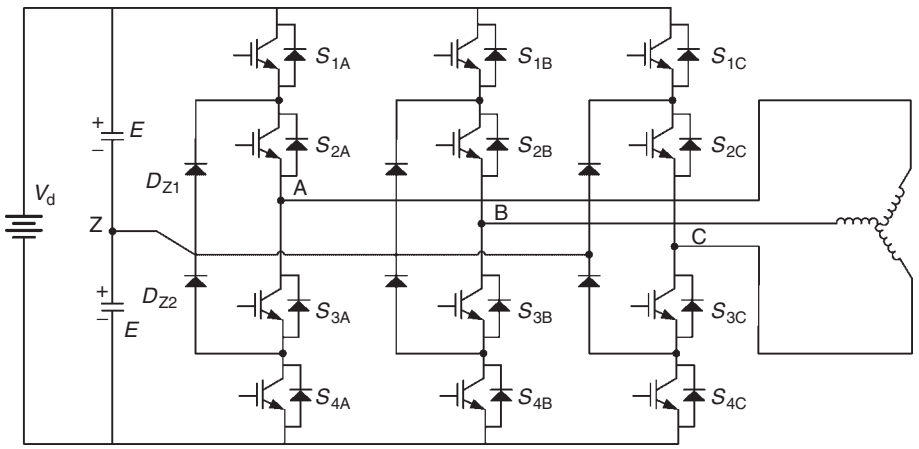


Figure 11. Configuration of three-level NPC inverter.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

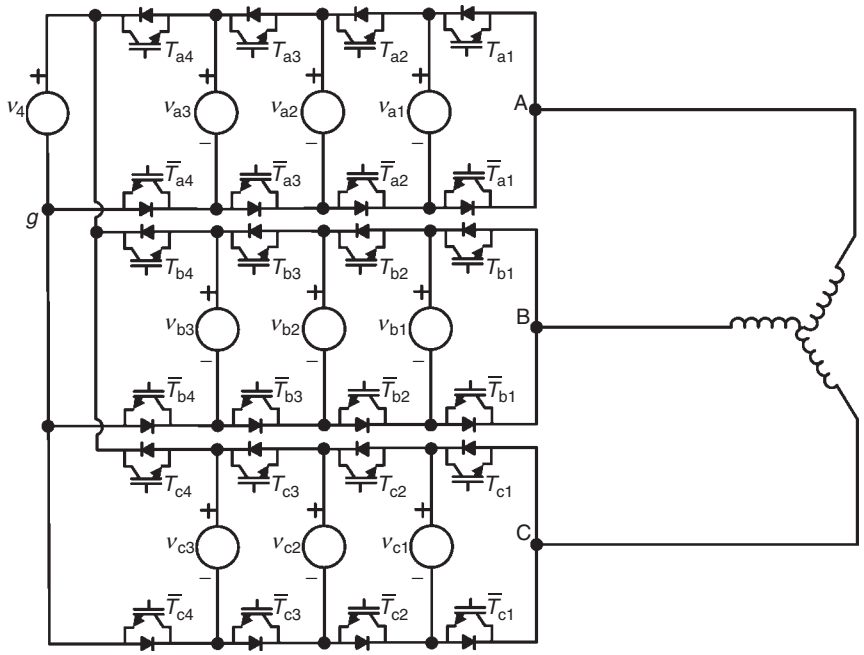


Figure 13. Configuration of a four-cell flying-capacitor multilevel inverter.

three-level NPC inverter, where the space is also divided into six triangular sectors (I–VI), each of which can be further divided into four triangular regions (1–4). For NPC inverter, the reference vector can be synthesized by three nearest stationary vectors. The dwell time calculation is also based on "voltage-second balancing" principle. The unbalance of the DC capacitor voltages can be mitigated by tuning the dwell time of small vectors.

Considering a large amount of battery sets existing in the EV, the multilevel inverter topologies with separate DC links are attractive. Figure 13 shows the configuration of a kind of four-cell flying-capacitor multilevel inverter. By replacing the floating capacitors with batteries and employing the full binary combination schema (Kou, Corzine, and Familiant, 2002), the number of voltage levels and the power quality of this inverter can be further increased. For conventional flying-capacitor multilevel inverter, the voltages of floating sources are $1:2:3:4$, and there are only five voltage levels, namely 0, $E/4$, $E/2$, $3E/4$, and E. The value of E refers to the DC voltage of v_4 in Figure 13. As the voltage step for the flying-capacitor voltages is uniform, the capacitors connected in series could provide the required flying sources. Conversely, the distribution between voltages of floating sources v_{a1}, v_{a2}, v_{a3}, and v_4 in Figure 13 is designed unevenly, in such a way that 16 voltage levels are available in the phase or line-to-ground voltages. The 16 voltage levels are 0, $E/15$, $2E/15$, $3E/15$, $4E/15$, $5E/15$, $6E/15$, $7E/15$, $8E/15$, $9E/15$, $10E/15$, $11E/15$, $12E/15$, $13E/15$, $14E/15$, and E. There are other options of voltage ratios for the binary-combination-schema-based flying source multilevel inverter. For example, $1:3:7:15$, $8:12:14:15$, and $1:5:13:15$. The batteries with different voltages in EV can provide the floating sources in Figure 13.

Figure 14 shows another multilevel inverter for IM drive of EV. It is a cascaded H-bridge-based inverter that consists of five floating DC sources per phase (Tolbert, Peng, and Habetler, 1999). This inverter takes not only the unique advantages of multilevel inverters but also the use of identical H-bridge inverter units, thus enhancing modularity and manufacturability. The CHB-based inverter requires separate DC links, and the battery sets are very suitable to provide such separate DC links. The multilevel carrier-based modulation technique is used widely for the CHB-based multilevel inverters. The multilevel carrier-based modulation includes the phase-shifted and the level-shifted PWM schemes. The multilevel inverter with m voltage levels requires $(m-1)$ triangular carriers. In the phase-shifted multi-carrier modulation, all the triangular carriers have the same frequency and the same peak-to-peak amplitude, but there is a phase shift between any two adjacent carrier waves. The shifted phase is $2\pi/(m-1)$. The modulating signal is usually a three-phase sinusoidal wave with adjustable amplitude and frequency. The switching signals are generated by comparing the modulating waves with the carrier waves.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

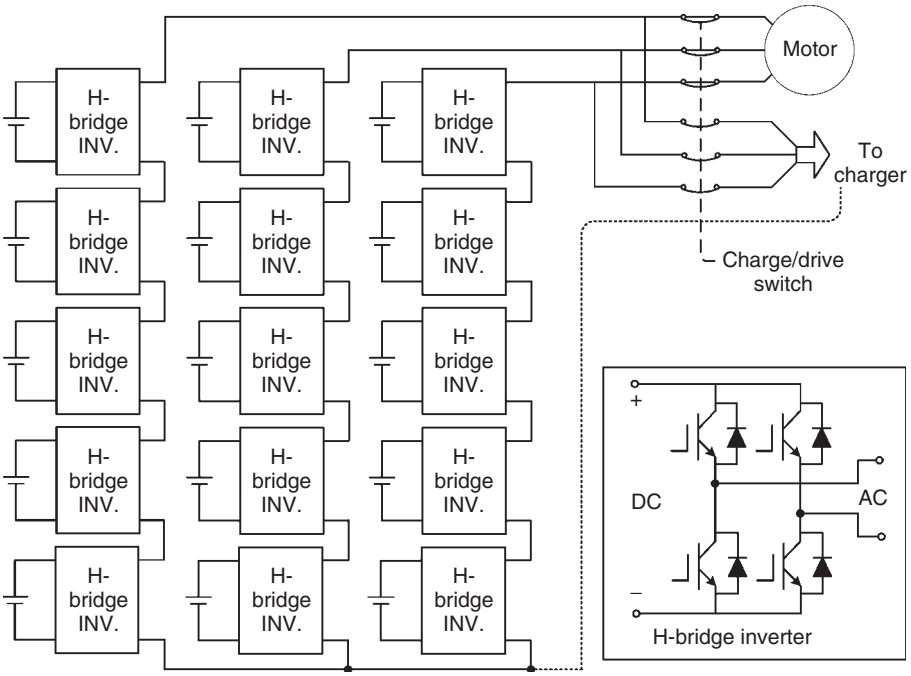


Figure 14. Configuration of CHB-based inverter.

The level-shifted modulation for CHB inverter uses $(m-1)$ level-shifted triangular carriers to construct an m-level CHB inverter (Malinowski *et al*., 2010). The $(m-1)$ level-shifted carriers all have the same frequency and amplitude. They are vertically disposed and the bands they occupy are contiguous. There are three level-shifted multi-carrier modulation schemes: in-phase disposition (IPD), where all carriers are in phase; alternative phase opposite disposition (APOD), where all carriers are alternatively in opposite disposition; and phase opposite disposition (POD), where all carriers above the zero reference are in phase but in opposition with those below the zero reference.

4.3 Soft-switching power inverters

In order to reduce the issues resulting from the high value of dv/dt such as the EMI and common-mode voltages in EV motor drives, soft-switching power inverters are invented to implement the zero-voltage transition (ZVT) or the zero current transition (ZCT) for the power switches. Figure 15 shows the configuration of a three-phase star-configured resonant snubber-based inverter, which is a kind of ZVT inverter for IM drives of EV (Lai, 1997). The resonant snubber-based inverter proposes to produce the zero voltage across the switching device using the resonant branch,

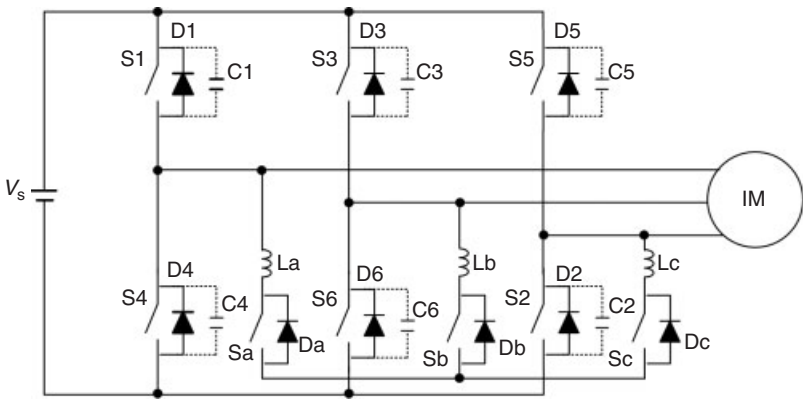


Figure 15. Configuration of three-phase star-configured resonant snubber-based inverter.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

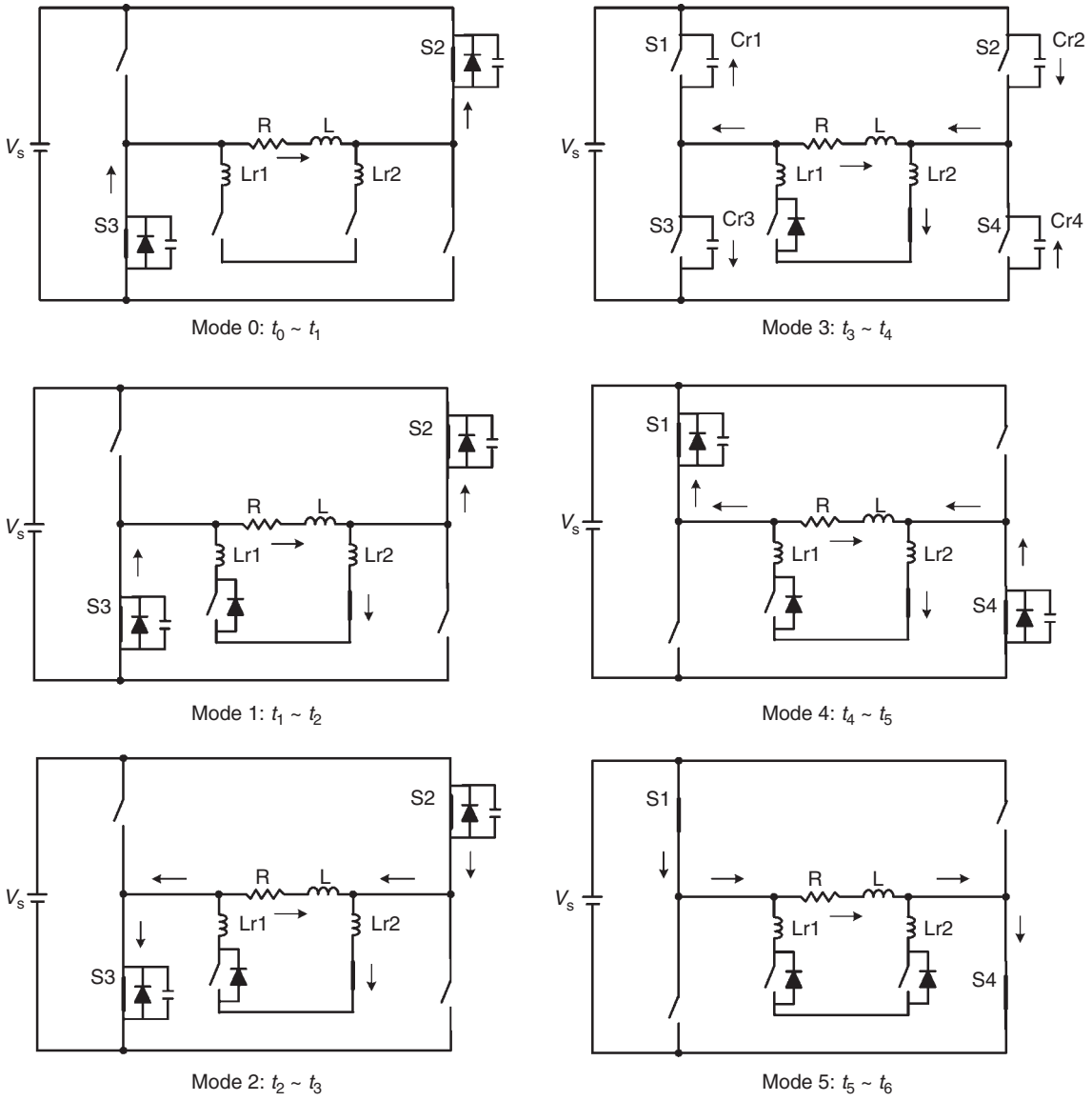


Figure 16. Working modes of phase-to-phase configuration of star-configured resonant snubber-based inverter.

which includes an auxiliary switch, a resonant inductor, and the stray capacitance of the switching device.

Figure 16 plots the working modes of the phase-to-phase configuration within the half cycle of positive load current. The initial condition is that a positive load current is freewheeling diode through D2 and D3, and the switched S2 and S3 remain on. At the time of t_1, the resonant switch S_{r2} is turned on. The resonant inductor current I_{Lr2} increases linearly. Thus, the current in switches S2 and S3 declines gradually and becomes zero at t_2. The resonant inductor current I_{Lr2} keeps increasing and exceeds the load current after t_2. Then, the devices S2 and S3 are turned off. Owing to the existence of the capacitors C_{r2} and C_{r3}, the zero-voltage turn-off operation is available for the main switches S2 and S3. After the turn-off of S2 and S3, the capacitors C_{r2} and C_{r3} are charged. Meanwhile, C_{r1} and C_{r4} are discharged to zero. The resonant current keeps decreasing, and the load current flows through D1 and D4. Thus, the switches S1 and S4 can be turned on at the zero-voltage condition. When the resonant inductor current is equal to the load current, the load current flows through the switches S1 and S4. The resonant current deceases until it becomes zero, and it is blocked by the diode across S1.

Figure 17 shows the configuration of a three-auxiliary-switches-based ZCT inverter (Li, Lee, and Boroyevich, 2003). The auxiliary switches functions to initiate resonance

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

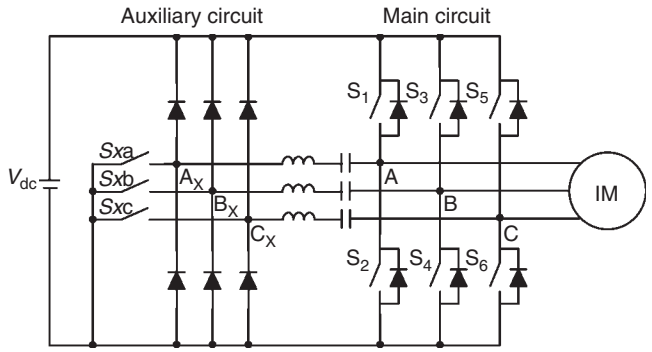


Figure 17. Configuration of three-auxiliary-switches-based zero current transition inverter.

and provide the flowing paths for the resonant current. Figure 18 shows the working modes of one-phase leg of the three auxiliary switch ZCT inverter. The operation of the circuit is not symmetrical for different load directions. The negative load current condition is used for exemplification. Therefore, the load is connected through S_2 and D_1. The initial status is that the load current flows through D_1. The turn-on of the auxiliary switch S_{xa} makes the L_x and C_x resonant. When the resonant current arrives at the peak value, close to the load current, the current flowing through D_1 approaches to zero at t_1. The turn-on of S_2 at this moment can avoid the diode reverse current and reduce the diode switching loss. The gating signal for S_1 should be removed before t_1. Otherwise, the shoot-through fault will occur. The resonant current i_x keeps decreasing to zero at t_2. After that moment, i_x is driven to the negative direction and D_{xa} conducts. Thus, the switch S_{xa} achieves the zero-current turn-off condition. When the resonant current i_x goes back to zero at t_3, the diode D_{xa} turns off and is blocked. The auxiliary circuit continues to resonate. The resonant current is then diverted to the diode D_{ca}. When the positive resonant current returns to zero, the diode D_{ca} turns off naturally and the load current flows through the switch S_2.

When S_2 is ready to turn off, the auxiliary switch S_{xa} is turned on at t_5. When the resonant current increases to the load current, the current through S_2 decreases to zero, and S_2 can be turned off at zero current condition. The resonant current continues to increase, and the surplus current flows through the diode D_2. After arriving at the peak value, the resonant current decreases back below the load current. D_2 thus turns off naturally. The load current flows only through the resonant tank and charges the resonant capacitor linearly. When the resonant capacitor voltage is charged to the DC link voltage, the diode D_1 begins to conduct. Then, the main switch S_1 can be turned on. At t_9, the resonant current i_x decreases to zero, and the resonant current will be carried by the diode D_{xa}. The resonant switch S_{xa} can be turned off with zero current. The resonant current will go back to zero and is blocked by the diodes D_{ca} and D_{xa}. Then, the load current flows through D_1.

4.4 Current source inverters

Unlike VSI, the current source inverter (CSI) adopts the inductor as the energy storage element in the DC link. Figure 19 shows the configuration of the CSI for IM drive of EV (Su and Tang, 2011). The CSI provides the following merits. First, it allows the shoot-through operation, where the upper and the lower switch in the same leg conduct meantime. Second, it can boost the DC link voltage and provide higher AC output voltage, which is suitable for the extended speed operation of EV motor. Third, the filter capacitors for current commutation of CSI can provide better output voltage waveforms.

For CSI, SVM is a very suitable modulation strategy as it offers high flexibility and controllability. Two switches are allowed to conduct at the same time in the CSI. One is in the top half of CSI and the other is in the bottom half of CSI. Therefore, they are totally nine switching states. Among them, six are active vectors and three are zero vectors. Figure 20 shows the vector diagram of the CSI (Wang *et al.*, 2012a, 2012b). Similar to the vector diagram of SVM for VSI, $\vec{I}_1$ to $\vec{I}_6$ are the active vectors and $\vec{I}_7$ to $\vec{I}_9$ are the zero vectors. The difference is that there is 30° shift between the vector diagram of CSI and that of VSI. The calculation of the dwell time for different vectors is also based on the "current-second balancing" principle to synthesize the reference current vector $\vec{I}_{ref}$. There are different switching sequences available for the SVM of CSI. Figure 21 shows a kind of switching sequence that has better THD performance under the same switching frequency (Wang *et al.*, 2012a, 2012b).

Figure 22 plots four operation modes of the CSI-based EV motor drives (Su and Tang, 2011). The CSI and the motor are represented by a voltage source V_{in}. In mode I, the switches S_a and S_b are both turned on. The battery charges the DC inductor. In mode II, the switch S_a is turned on while the switch S_b is turned off. On the contrary, the switch S_b is turned on while the switch S_a is turned off in mode III. In mode IV, both the switches S_a and S_b are turned off. In the low speed motoring operation region, the *V–I* converter operates between mode I and mode II or mode I and mode III. The DC link current is regulated by the *V–I* converter. In the low speed regenerative operation region, the converter operates between mode II and mode IV or mode III and mode IV. For the high speed motoring operation of EV, the switches S_a and S_b are kept on, and the DC link current is regulated by the CSI side. On the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

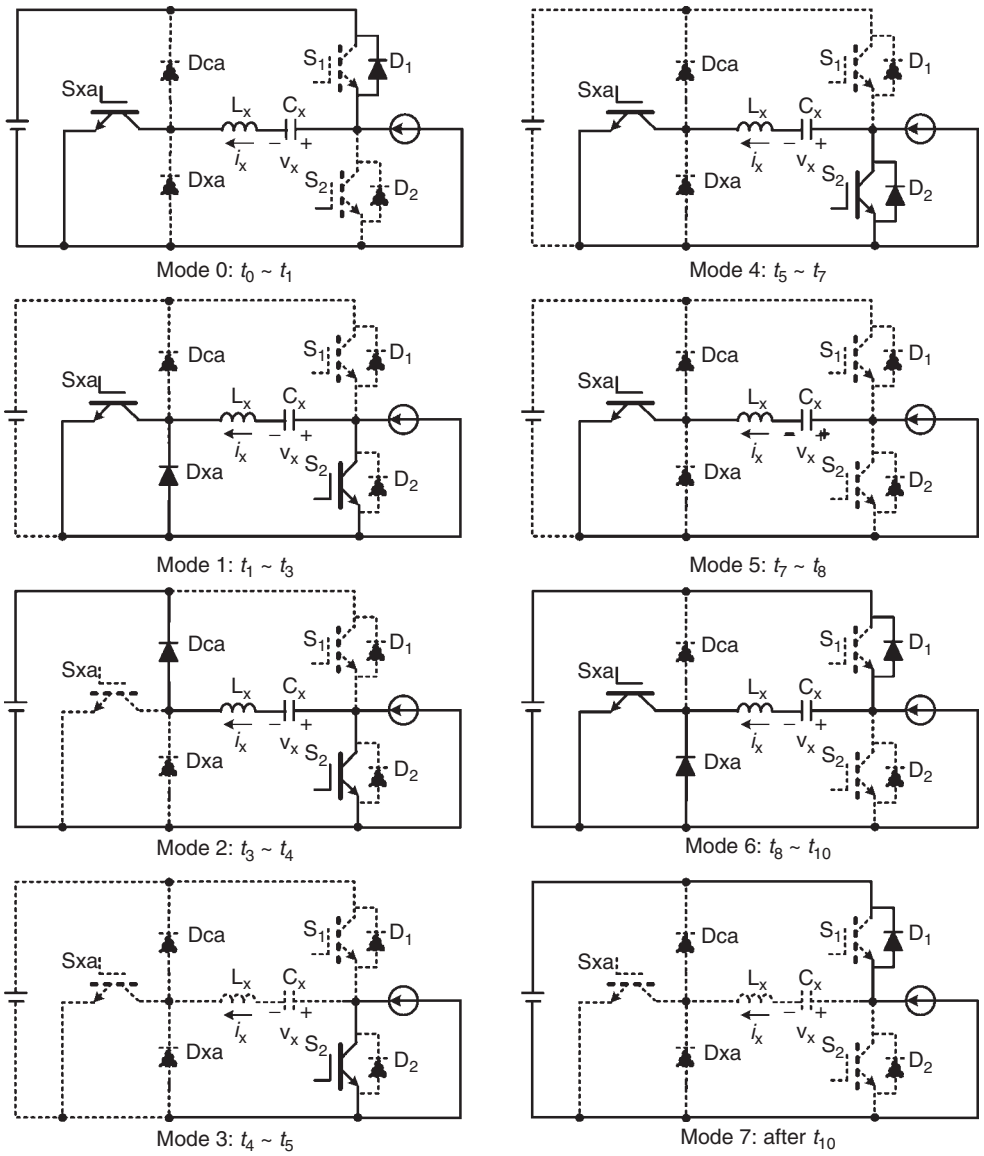


Figure 18. Working modes of one phase of three-auxiliary-switches-based zero current transition inverter under negative load current condition.

contrary, the switches S_a and S_b are kept off for the high speed regenerative operation of EV. Under this condition, the DC link current is also regulated by the CSI side.

4.5 Z-source inverters

The Z-source inverter provides a unique impedance network as shown in Figure 23. In the DC link, a two-port network consisting of a split-inductor L_1 and L_2 and capacitors C_1 and C_2 connected in X shape is employed. The unique feature of the Z-source inverter is that the output AC voltage can be of any value between zero and infinity regardless of the DC link voltage. Thus, the DC/DC converter used to boost the DC link voltage for extending operation of EV can be removed. Besides, the shoot-through operation induced by the EMI noise is allowed for the Z-source inverter. Different from the traditional PWM for VSI, a shoot-through zero states are inserted into the switching states. Hence, the equivalent DC link voltage to the inverter can be boosted effectively (Peng, 2003).

The relationship between the equivalent DC link voltage V_i and the input DC voltage V_o is (Peng, 2003)

$$\frac{V_i}{V_o} = B \tag{22}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

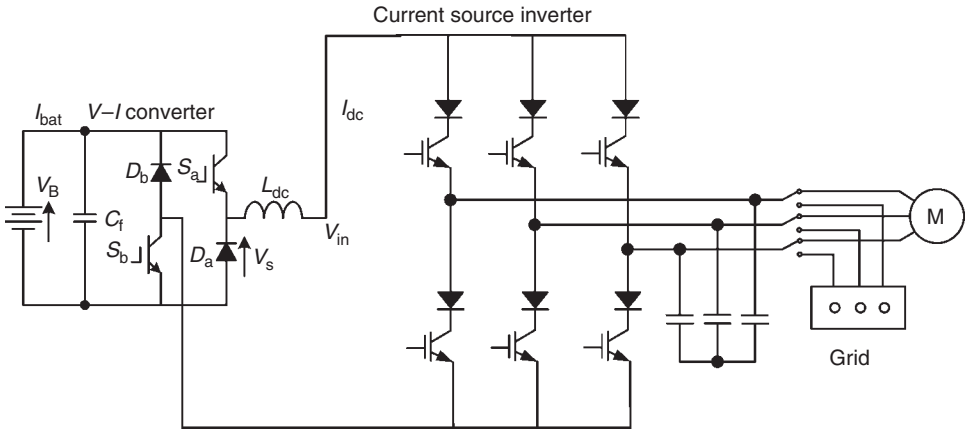


Figure 19. Configuration of current-source-inverter-based motor drive.

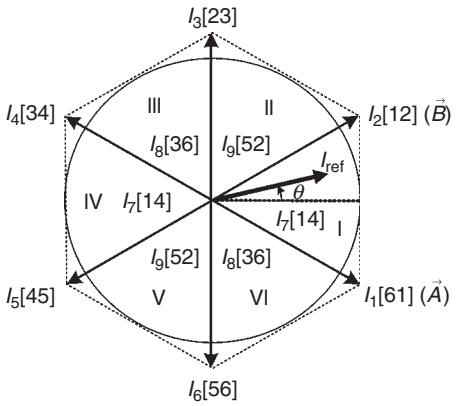


Figure 20. Vector diagram of current source inverter.

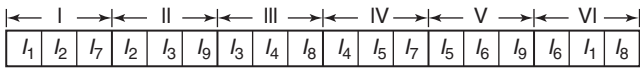


Figure 21. Switching sequence of current source inverter.

where B is the boost factor and is determined by:

$$B = \frac{1}{[1 - 2(T_{SC}/T)]} \tag{23}$$

where T_{SC} is the shoot-through time interval within a switching interval T. As $B \geq 1$, the boost operation is implemented.

Figure 24 gives 12 operation modes of the bidirectional Z-source inverter (Dehghan, Mohamadian, and Yazdian, 2010). Modes 1 and 2 correspond to the shoot-through operation of IM drive, where the switch S_7 is turned off, and its antiparallel diode is reverse biased. The DC link on AC side is short-circuited in modes 1 and 2. Modes 3 and 4 correspond to the zero states of IM drive, where the zero switching states are given at the inverter side. The DC link on AC side is thus open circuited. For modes 5, 6, 7, and 8, the inverter has an active vector for VSI, and its output current is positive. That means the active power is delivered from the battery at DC side to the IM at AC side. For modes 9, 10, 11, and 12, the inverter also has an active vector, but its output current is negative, which indicates the active power is delivered from the AC side to the DC side. This condition corresponds to the regenerative operation of IM drive in EV.

5 CONTROL STRATEGIES OF IM DRIVE FOR EV

Speed control in induction drive systems is more complex than that in DC drive systems because the IMs suffer from nonlinearity of the dynamic model with coupling between direct and quadrature axes. The basic equation of speed control of IM is governed by:

$$n = n_s(1 - s) = \frac{f}{p}(1 - s) \tag{24}$$

Thus, the motor speed can be controlled by f, p, and/or s.

As shown in Figure 25, the characteristics of the induction drive system can be divided into three operating regions (Chau and Wang, 2011). The first region is called the *constant-torque region* in which the motor can deliver its rated torque for frequencies below the rated frequency. In the second region, called the *constant-power region*, the slip is increased to the maximum value so that the stator current remains constant and the motor can maintain its

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

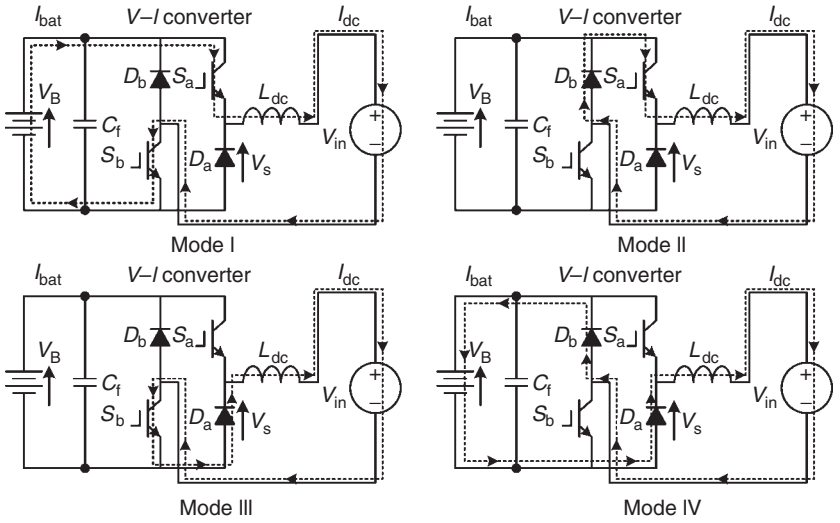


Figure 22. Operation modes of current-source-inverter-based induction motor drive for EV.

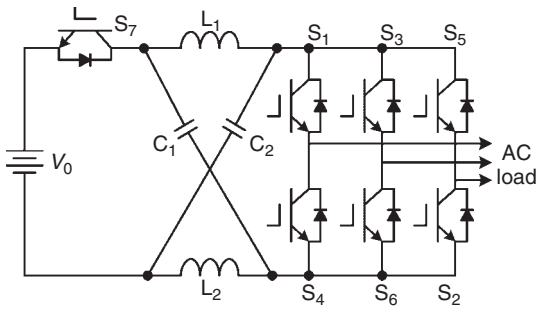


Figure 23. Configuration of bidirectional Z-source inverter.

rated power. In the high speed region, the third region, the slip remains constant while the stator current decreases. Thus, the torque capability declines with the square of speed.

5.1 Variable-voltage variable-frequency control

The variable-voltage variable-frequency (VVVF) control is a kind of open-loop control. Figure 26 shows its block diagram (Chau and Wang, 2011). This strategy is based on the constant volts/hertz control for frequencies below the motor-rated frequency and variable-frequency control with constant-rated voltage for frequencies beyond the rated frequency. For very low frequencies, voltage boosting is applied to compensate the difference between the applied voltage and the induced EMF because of the stator resistance drop.

In the VVVF control, the motor speed reference is constructed as a ramp shape to prevent the electrical and mechanical shock to the speed control system. There is no speed feedback for the VVVF control, and the motor speed is assumed to follow the output frequency. With the given voltage reference V and the frequency reference f, the PWM signals are provided for the power inverters to feed the IM. The current is usually fed back for the VVVF control mainly for protection.

As the VVVF control employs the open-loop control, it has the following drawbacks. First, the torque performance at low speed is poor because of the significant drop of stator voltage. Second, this control cannot provide high dynamic performance and torque response. Although the slip compensation strategy has been employed to regulate the frequency reference f to adapt to the change of load, the dynamic performance of the VVVF control is still worse than that of the closed-loop control strategies. Third, the IM drive with VVVF control cannot provide good operating performance under the shock loads, which are usually required by EVs. Therefore, the VVVF control strategy is less attractive for high performance IM drive systems.

5.2 Field oriented control

The field oriented control (FOC), also known as *vector control*, is proposed for IM drives to emulate the DC motor control. By using the proper field orientation, the stator current could be decomposed into a flux-producing component and a torque-producing component. Then, the two components can be controlled separately, which is very similar to that of DC motor drives. The FOC can be classified into three categories based on the synchronous

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

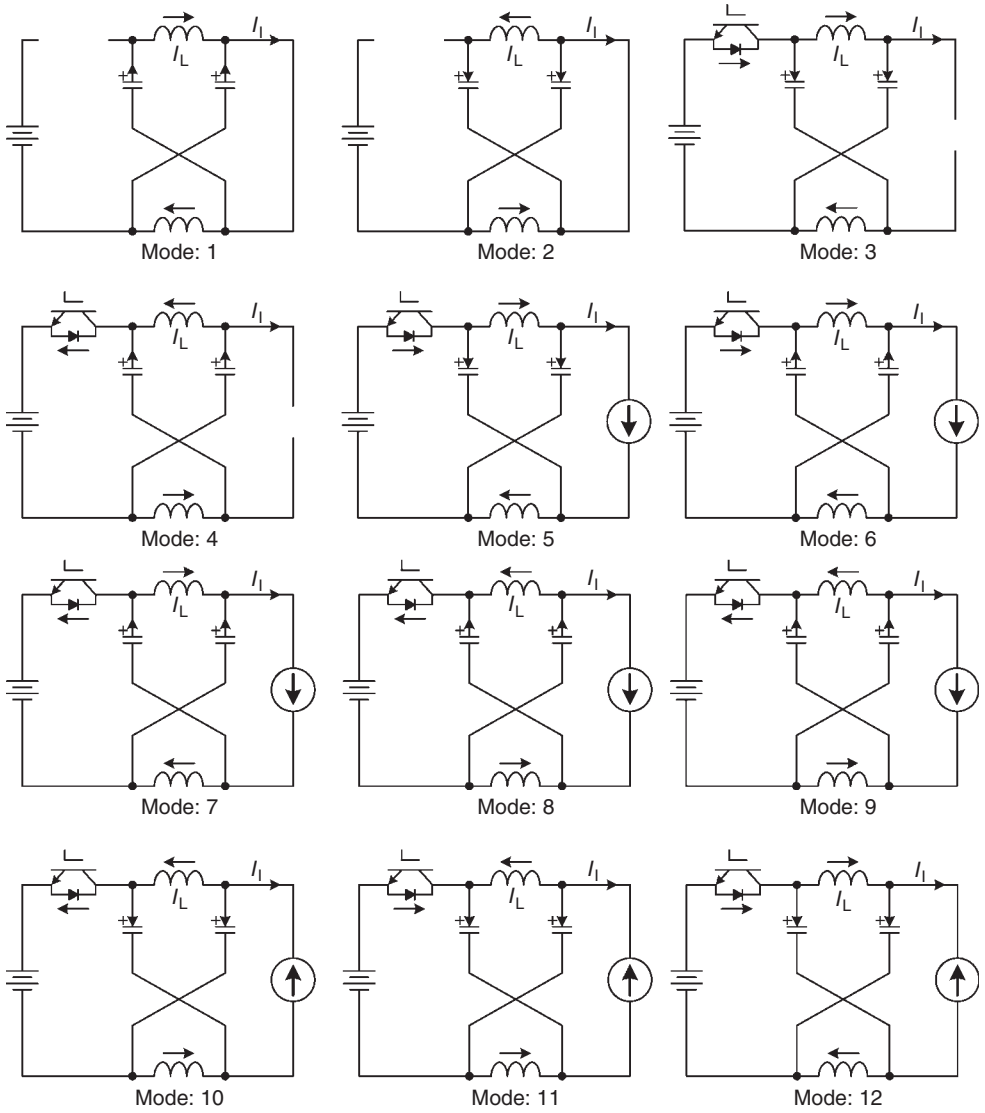


Figure 24. Operation modes of bidirectional Z-source-inverter-based induction motor drive for EV.

frame choice: the stator flux orientation, the air-gap flux orientation, and the rotor flux orientation. Among the different FOC strategies, the rotor flux orientation is most widely used. The principle of the rotor flux orientation is to align the *d*-axis of the synchronous reference frame with the rotor flux vector of IM.

Figure 27 shows the block diagram of FOC for IM drive (Chau and Wang, 2011). By using FOC, the mathematical model of IMs is transformed from the stationary reference frame ($\alpha-\beta$ frame) to the general synchronously rotating frame ($x-y$ frame) as shown in Figure 28. Thus, at steady state, all the motor variables such as supply voltage v_s, stator current i_s, rotor current i_r, and rotor flux linkage λ_r can be represented by DC quantities. When the x-axis is purposely selected to be coincident with the rotor flux linkage vector, the reference frame (d-q frame) becomes rotating synchronously with the rotor flux as shown in Figure 29, where i_{sd} and i_{sq} are the d-axis component and q-axis component of stator current, respectively. Hence, the motor torque T can be obtained as:

$$T = \frac{3}{2} p \frac{M}{L_r} \lambda_r i_{sq} \tag{25}$$

where M is the mutual inductance per phase and L_r the rotor inductance per phase. As λ_r can be written as Mi_{sd},

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

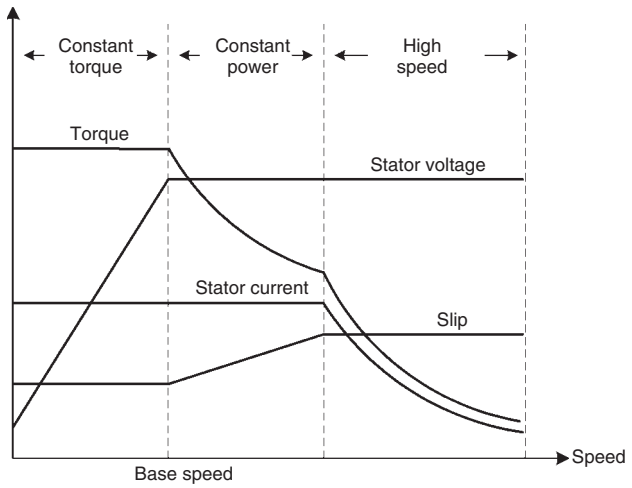


Figure 25. Characteristics of induction motor.

the torque equation can be rewritten as:

$$T = \frac{3}{2} p \frac{M^2}{L_r} i_{sd} i_{sq} \tag{26}$$

This torque equation is very similar to that of separately excited DC motors. Namely, i_{sd} resembles the field current I_f, whereas i_{sq} resembles the armature current I_a. Thus, i_{sd} can be considered the field component of i_s, which is responsible for establishing the air-gap flux. On the other hand, i_{sq} can be considered the torque component of i_s, which produces the desired motor torque. Therefore, by using this FOC, the motor torque can be effectively controlled by adjusting the torque component, and the field component remains constant. Hence, the IM drive can offer the desired fast transient response similar to that of separately excited DC drive systems. In order to attain the above FOC, the rotor flux linkage vector is always aligned with the *d*-axis. This criterion of the decoupling condition can be attained through slip frequency ω_{slip} control as given by:

$$\omega_{slip} = \frac{R_r i_{sq}}{L_r i_{sd}} \tag{27}$$

where R_r is the rotor resistance for each phase.

Since the advent of FOC, a number of methods have been proposed for implementation. Basically, these methods can be classified into two groups, namely direct FOC and indirect FOC. The direct FOC requires calculating the rotor flux and the motor speed with the measured motor voltages, currents, and the IM model. In contrast, the indirect FOC determines the rotor flux angle from the measured rotor speed and calculated slip angle based on motor parameters. Figure 30 shows the block diagram of the indirect FOC for IM drive (Chau and Wang, 2011). Although the FOC has been widely used for high performance induction drive systems, it still suffers from some drawbacks. Particularly, the rotor time constant L_r/R_r (which has a dominant effect on the decoupling condition) changes severely with the operating temperature and magnetic saturation, leading to deteriorate the desired FOC. In general, there are two ways to solve this problem. One way is to perform online identification of the rotor time constant and accordingly update the parameters used in the FOC controller. The other way is to adopt a sophisticated control algorithm to enable the FOC controller insensitive to motor parameter variations.

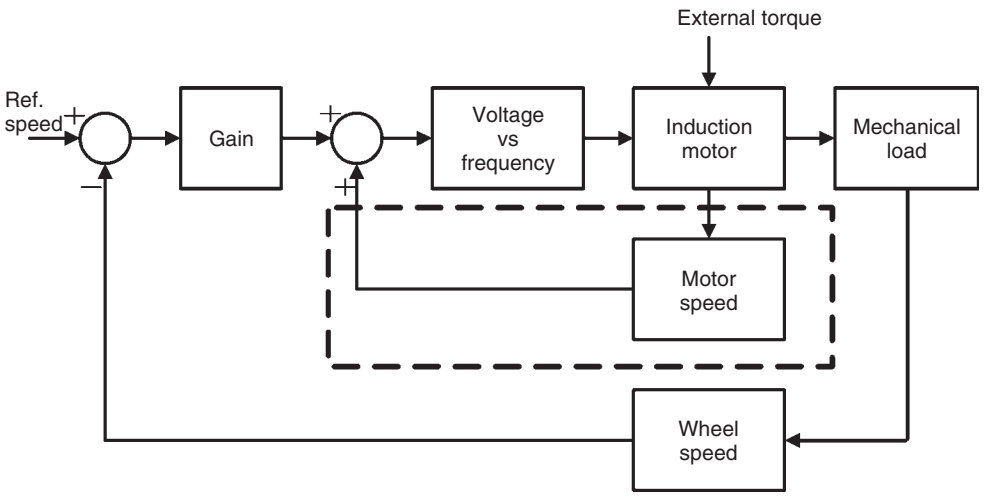


Figure 26. Block diagram of VVVF control.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

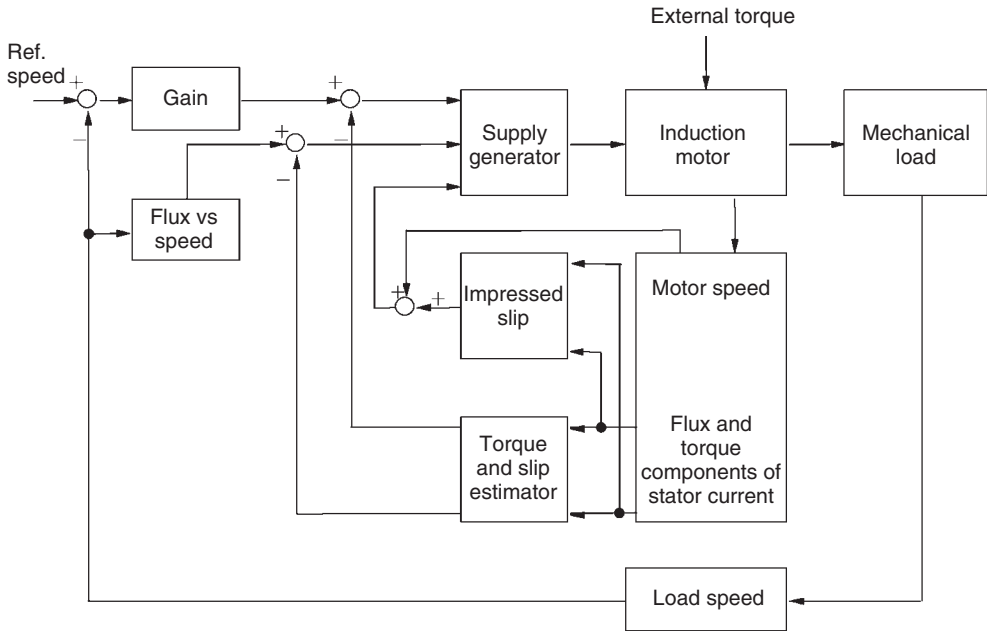


Figure 27. Block diagram of FOC for induction motor drive.

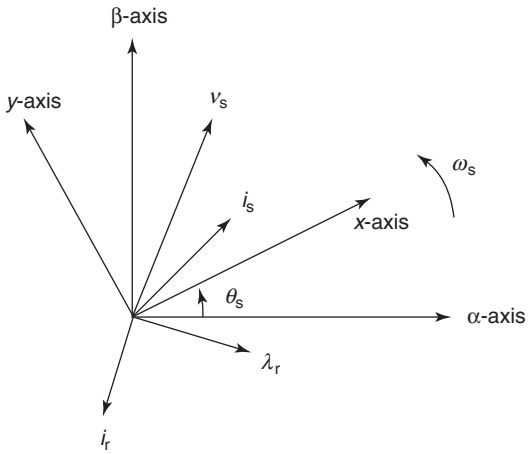


Figure 28. x–y frame rotating synchronously in general.

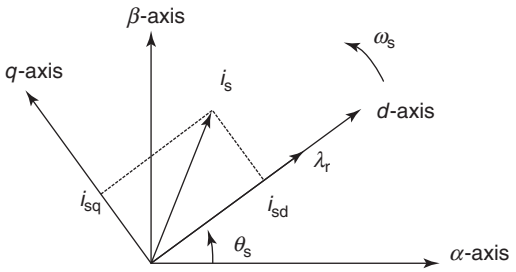


Figure 29. Principle of FOC with rotor flux orientation.

5.3 Direct torque control

The DTC is another advanced control strategy, which is widely used for IM drives (Buja and Kazmierkowski, 2004). It has the merits of simple control algorithm, fast torque response, easy digital implementation, and robust operation. The electromagnetic torque for an IM can be expressed as:

$$T = \frac{3P}{2}\frac{L_m}{\sigma L_s L_r}\lambda_s \lambda_r \sin\theta_T \tag{28}$$

where λ_s and λ_r are the magnitude (peak value) of the stator and rotor flux linkage vectors $\vec{\lambda}_s$ and $\vec{\lambda}_r$ and θ_T the angle between the stator flux and the rotor flux. Figure 31 shows the different reference frames in IM where d, q axes align with rotor flux $\vec{\lambda}_r$, x, y axes align with stator flux $\vec{\lambda}_s$, and α, β axes are stationary with α axis aligning with armature winding a. θ_T thus becomes the angle between $\vec{\lambda}_s$ and $\vec{\lambda}_r$.

In the stator flux control, the relationship between the stator flux $\vec{\lambda}_s$ and the stator voltage vector $\vec{v}_s$ is given as follows:

$$\frac{d\vec{\lambda}_s}{dt} = \vec{v}_s - R_s\vec{i}_s \tag{29}$$

The stator voltage $\vec{v}_s$ is the output voltage of the inverter, which can be controlled by the reference vector $\vec{v}_{ref}$ in the SVM. As $\vec{v}_{ref}$ can be controlled by the power inverter, the proper combination of the switching vectors could adjust the magnitude and angle of $\vec{\lambda}_s$ dynamically. The DTC is to control the electromagnetic torque by

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

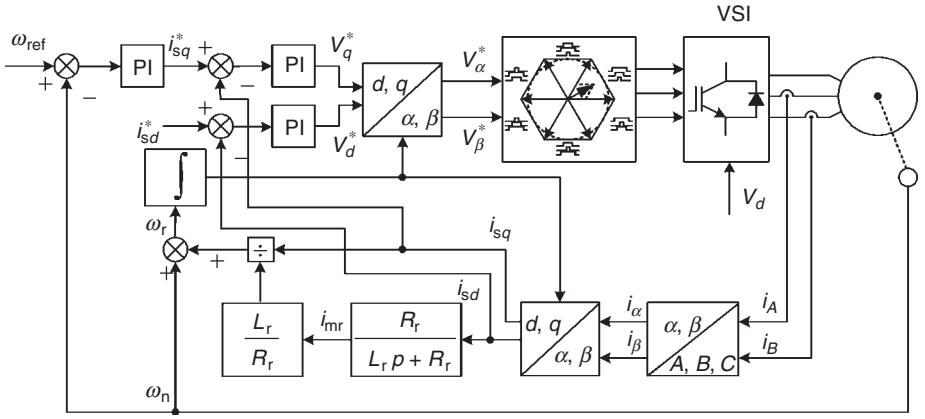


Figure 30. Block diagram of indirect FOC for induction motor drive.

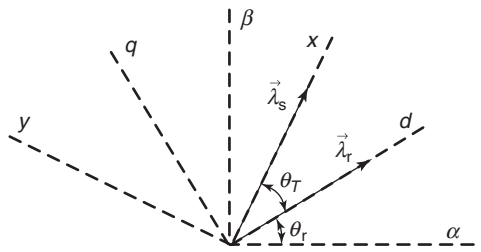


Figure 31. Different reference frames of induction motor.

adjusting the torque angle θ_T while keeping the magnitude of the stator flux at a constant value. Figure 32 shows the block diagram of DTC. The measured voltages and currents are transformed into α, β axes. Then, λ_s can be estimated with u_α, u_β, i_α, and i_β. Thus, the electromagnetic torque T is calculated with the estimated $\vec{\lambda}_s$ and the measured stator currents i_α and i_β. The reference torque T^* is generated by the speed controller. The magnitude and angle of the stator flux $\vec{\lambda}_s$ is calculated as:

$$\lambda_s = \sqrt{(\lambda_{s\alpha})^2 + (\lambda_{s\beta})^2} \tag{30}$$

$$\theta_s = \text{arctg}\left(\frac{\lambda_{s\beta}}{\lambda_{s\alpha}}\right) \tag{31}$$

As shown in Figure 32, the switching status in DTC is determined by the sector where λ_s locates and by the outputs of two hysteresis loops for torque and flux. Figure 33 shows the principle of choosing switching status for DTC. Figure 33a corresponds to the acceleration condition by defining the anticlockwise direction as

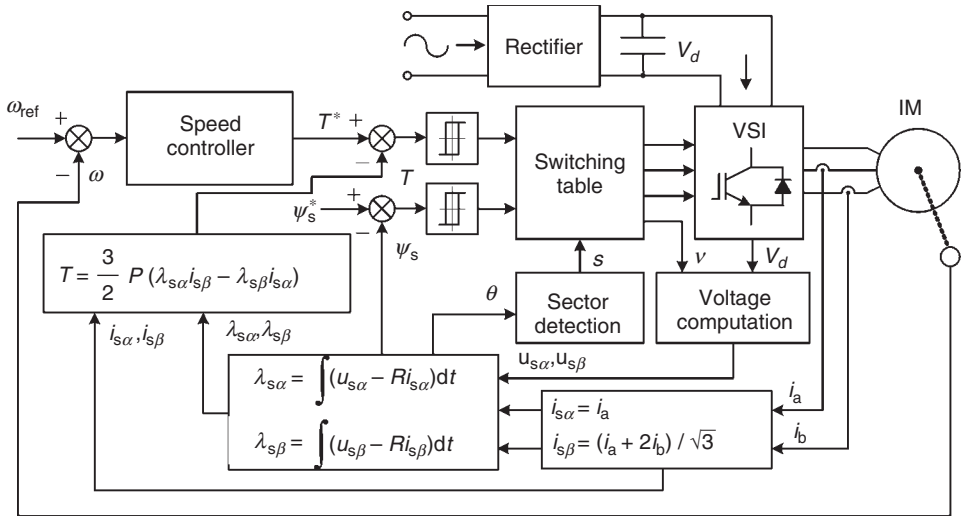


Figure 32. Block diagram of DTC for induction motor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

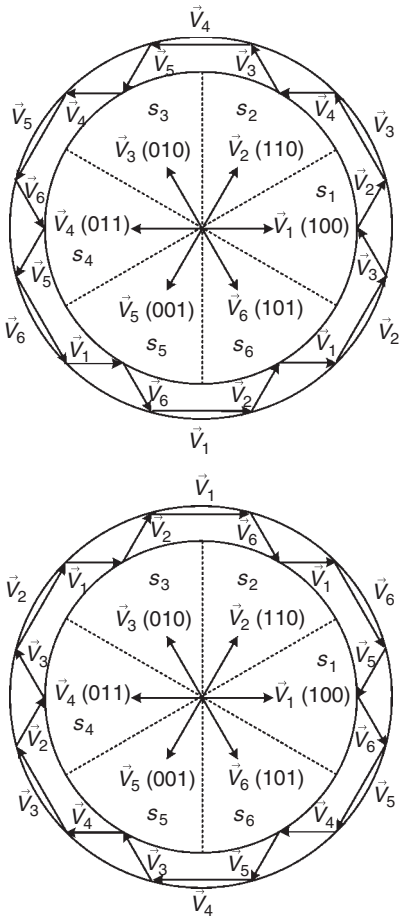


Figure 33. Choice of switching status for DTC: (a) acceleration condition; (b) deceleration condition.

the rotating direction of IM. Accordingly, Figure 33b plots the deceleration condition. The impact of the stationary voltage vectors $\vec{V}_1$ to $\vec{V}_6$ on $\vec{\lambda}_s$ and θ_T can be observed in Figure 33. The stationary plane for the stator flux $\vec{\lambda}_s$ is divided into six sectors from I to VI. For example, four switching vectors, namely $\vec{V}_2$, $\vec{V}_3$, $\vec{V}_5$, and $\vec{V}_6$ have the impact in sector I. The electromagnetic torque increases by selecting the vectors $\vec{V}_2$ and $\vec{V}_3$. On the contrary, the vectors $\vec{V}_5$ and $\vec{V}_6$ will reduce the torque and decelerate the motor. For the impact on the flux, the vectors $\vec{V}_2$ and $\vec{V}_6$ can enlarge the magnitude of stator flux $\vec{\lambda}_s$, whereas the vectors $\vec{V}_3$ and $\vec{V}_5$ can reduce the stator flux in sector I. The choice of the vectors for change of the torque and flux in other sectors are similar as in Figure 33.

It should be mentioned that the zero vectors of voltage source converter have no impact on the torque and flux. The two zero vectors, namely 000 and 111, could be used under the condition where the torque does not need changing. Thus, the smaller torque ripple can be available.

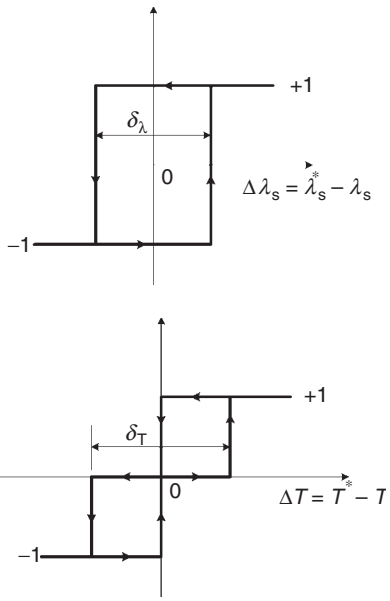


Figure 34. Characteristics of hysteresis loops: (a) flux loop; (b) torque loop.

The corresponding flux and torque hysteresis loops are shown in Figure 34, where δ_λ and δ_T are the boundaries for the flux and torque hysteresis loop, respectively (Wu, 2006). The key for choosing zero vectors is to minimize the switching times when the active vectors are switched. For example, the zero vector 111 is used when the torque does not change but the flux is increasing in sector I. On the contrary, the zero vector 000 is used when the torque does not change and the flux is decreasing in sector I. The choice of zero vectors in other vectors is similar.

6 CONCLUSIONS

The characteristics of IM drives and advanced control strategies are extensively discussed. The IM drive is the most mature technology among various commutatorless motor drives. The IM drives offer the advantages of robust structure, low cost, high reliability, and being free from maintenance, which are particularly important for electric propulsion in EV applications.

The power inverter for IMs is designed to achieve the torque–speed requirements of the EV driving profile without involving variable gearing or gearbox. Moreover, it should provide the capability of bidirectional power flow to recover the regenerative braking energy. Five mature power inverter topologies, VSIs, multilevel power inverters, soft-switching power inverters, CSIs, Z-source inverters, are elaborated and evaluated for EV applications.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Speed control of IM drive systems is more complex than that of DC drive systems because IMs suffer from nonlinearity of the dynamic model with coupling between direct and quadrature axes. Three speed control strategies, VVVF control, FOC, DTC, are also discussed and evaluated.

It should be noted that system efficiency optimization of IM drive is extremely desirable for EV because of limited energy. Hence, efficiency optimization control of IM drive, incorporated with various speed control strategies, have attracted more attention (Li, Zhang, and Cui, 2010; Sergaki and Moustaizis, 2011).

RELATED ARTICLES

General Requirement of Traction Motor Drives

REFERENCES

Buja, G.S. and Kazmierkowski, M.P. (2004) Direct torque control of PWM inverter-fed AC motors – a survey. *IEEE Transactions on Industrial Electronics*, **51** (4), 744–757.

Cai, W. (2013) Electric machine systems for powertrains of new energy vehicles. Presented at the 2nd China International New Energy Vehicle Forum 2013. May 27th-29th, Shanghai, China.

Chan, C.C. and Chau, K.T. (2001) *Modern Electric Vehicle Technology*, Oxford University Press, Oxford.

Chau, K.T. and Wang, Z. (2005) Overview of power electronic drives for electric vehicles. *HAIT Journal of Science and Engineering B*, **2**, 737–761.

Chau, K.T. and Wang, Z. (2011) *Chaos in Electric Drive Systems – Analysis, Control and Application*, Wiley-IEEE Press, Singapore.

Dehghan, S.M., Mohamadian, M., and Yazdian, A. (2010) Hybrid electric vehicle based on bidirectional Z-source nine-switch inverter. *IEEE Transactions on Vehicular Technology*, **59** (6), 2641–2653.

Ehsani, M., Gao, Y., and Emadi, A. (2009) *Modern Electric, Hybrid Electric, and Fuel Cell Vehicles: Fundamentals, Theory and Design*, 2nd edn, CRC Press, Boca Raton.

Feng F.D., Wu B., Wei S., and Xu D (2004) Space Vector Modulation for Neutral Point Clamped Multilevel Inverter with Even Order Harmonic Elimination. *In Canadian Conference on Electrical and Computer Engineering*, Niagara Fall, Canada, 1471-1475.

Fitzgerald, A.E., Jr Charles, K., and Umans Stephen, D. (2003) *Electric Machinery*, 6th edn, The Mcgraw-Hill Companies, Inc, New York.

Kim Byunghwan, Lee Jeongho, Jeong Youngho, and *et al*. (2012) Development of 50 kW Traction Induction Motor for Electric Vehicle (EV). *In IEEE Vehicle Power and Propulsion Conference*, Seoul, Korea, 142-147.

Kou, X., Corzine, K.A., and Familiant, Y.L. (2002) Full binary combination schema for floating voltage source multilevel inverters. *IEEE Transactions on Power Electronics*, **17** (6), 891–897.

Lai, J.S. (1997) Resonant snubber-based soft-switching inverters for electric propulsion drives. *IEEE Transactions on Industrial Electronics*, **44** (1), 71–80.

Li Ke, Chenghui Zhang, and Naxin Cui (2010) Comparative Study of Induction Motor Efficiency Optimization Control Strategy for Electric Vehicle. *In Proceedings of the 8th World Congress on Intelligent Control and Automation*, Jinan, China, 1882-1887.

Li, Y.P., Lee, F.C., and Boroyevich, D. (2003) A simplified three-phase zero-current-transition inverter with three auxiliary switches. *IEEE Transactions on Power Electronics*, **18** (3), 802–813.

Malinowski, M.M., Gopakumar, K., Rodriguez, J., and Pérez, M.A. (2010) A survey on cascaded multilevel inverters. *IEEE Transactions on Industrial Electronics*, **57** (7), 2197–2206.

Mi, C., Masrur, M.A., and Gao, D. (2011) *Hybrid Electric Vehicles: Principles and Applications with Practical Perspectives*, John Wiley & Sons Ltd, Chichester.

Peng, F.Z. (2003) Z-source inverter. *IEEE Transactions on Industry Applications*, **39** (2), 504–510.

Sergaki E.S. and Moustaizis S.D. (2011) Efficiency Optimization of a Direct Torque Controlled Induction Motor Used in Hybrid Electric Vehicles. *In International Conference on Electrical Machines and Power Electronics*, İstanbul, Turkey, 398-403.

Su G.J., and Tang L (2011) Current Source Inverter Based Traction Drive for EV Battery Charging Applications. *In IEEE Vehicle Power and Propulsion Conference*, Chicago, USA, pp. 1–6.

Tolbert, L.M., Peng, F.Z., and Habetler, T.G. (1999) Multilevel converters for large electric drives. *IEEE Transactions on Industry Applications*, **35** (1), 36–44.

Wang, Z., Wu, B., Xu, D., and Zargari, N.R. (2012a) A current source converter based high-power high-speed PMSM drive with 420-Hz switching frequency. *IEEE Transactions on Industrial Electronics*, **59** (7), 2970–2981.

Wu, B. (2006) *High-Power Converters and AC Drives*, Wiley-IEEE Press, Piscataway.

Zhu, Z.Q. and Howe, D. (2007) Electrical machines and drives for electric, hybrid, and fuel cell vehicles. *Proceedings of the IEEE*, **95** (4), 746–765.

FURTHER READING

Novotny D.W., Lipo T.A., and Jahns T.M. (2009) Introduction to electric machines and drives. Wisconsin Power Electronics Research Center, University of Wisconsin-Madison.

Piotr, W. (2011) *Dynamics and Control of Electric Drives*, Springer, Berlin.

Wang, W., Cheng, M., Wang, Z., and Zhang, B. (2012b) Fast switching direct torque control using a single DC-link current sensor. *Journal of Power Electronics*, **12** (6), 895–903. http://en.wikipedia.org/wiki/Induction_motor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto210
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

CVT Control—System Integration, Ratio Choice, Shift Strategy and Dynamics, Adaptive Features, Engine Calibration, Electric Motor Assist

Takashi Shibayama
JATCO Ltd., Fuji City, Japan

1 Introduction 1
2 CVT Control System 1
3 Ratio Changing Control of CVT 4
4 Shift Schedule of CVT 7
5 CVT Ratio Changing Control Strategy 8
6 Other Ratio Changing Control of CVT 10
7 Engine and CVT Integrated Control 11
8 CVT Clamping Force Control 11
9 CVT Control in HEV System 12
10 Control for CVT with Auxiliary Shifting System 13
11 Chain CVT Control 13
12 Control of Toroidal CVT 16
Acknowlwdgments 19
References 19
Further Reading 19

1 INTRODUCTION

In order to understand the basic concept and outline of continuously variable transmission (CVT) control, following topics are described. (i) CVT control system, (ii) ratio changing control of a CVT, (iii) shift schedule of CVT, (iv) CVT ratio changing control strategy, (v) other ratio changing control of CVT, (vi) engine and CVT integrated control, (vii) CVT clamping force control, (viii) CVT control in hybrid electric vehicle systems, (ix) control for CVT with auxiliary shifting system, (x) chain CVT control, and (xi) control of toroidal CVT. CVT ratio control with a stepper motor, CVT shift feeling improvement, and automatic engine braking by CVT ratio change control are also discussed.

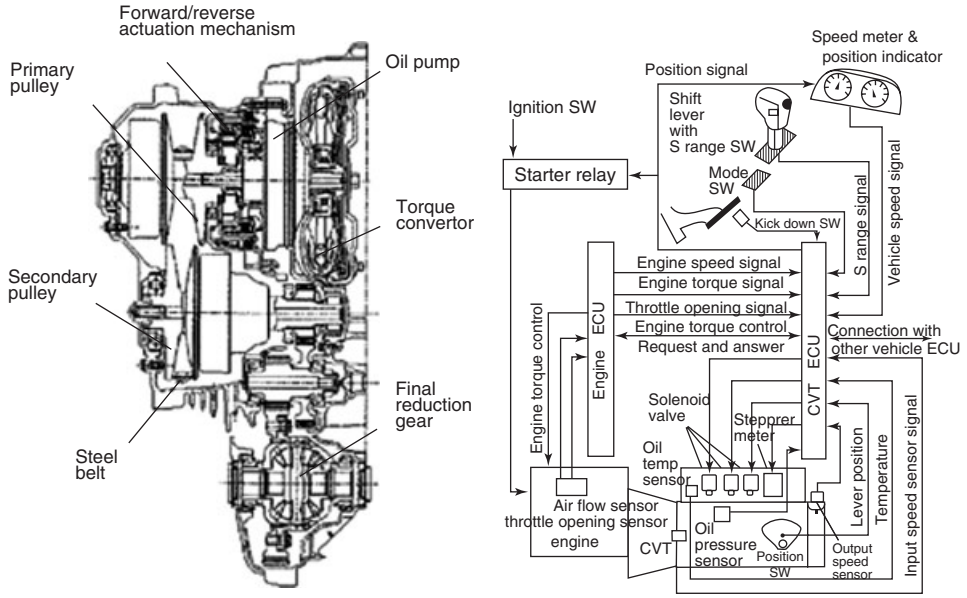
Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 CVT CONTROL SYSTEM

Figure 1a shows a cross section of a typical belt CVT transaxle (Abo *et al*., 2003). Engine output is connected to the torque converter, which acts as a launching device. The forward/reverse actuation mechanism changes torque direction, and then by the adjustment of the primary and secondary pulleys, and steel belt system, a continuously variable ratio change is realized. The wheel driving torque is then derived through the final reduction gear.

Figure 1b shows a typical CVT control system, which has several sensors, actuators, and communicators. The most important sensor is the pressure sensor by which CVT ECU (electric control unit) can detect belt clamping force, which should be enough to prevent belt slipping.

Figure 1c shows a typical device combination for CVT control. The control system is similar to that of an Automatic Transmission (AT), which is described in (see AT Control—Actuation Methods & System Integration, Gear Choice, Gear Shift Strategy & Process, Adaptive Features) The main differences are the pressure sensor and stepper motor for controlling CVT ratio changing.



Category	Device	Major act of the device
Sensor	Pressure sensor	To detect damping force of pulley
	Engine speed sensor	To detect torque converter slipping ratio
	CVT input speed sensor	To detect speed ratio of CVT
	CVT output speed sensor	To detect gear ratio of CVT
	Vehicle speed sensor	To detect wheel slip
	Position sensor	To detect P.R.N.D.L opening range
	S range SW	To make special shift control, ex. sporty or normal
	Mode SW	To switch mode, ex. economy or normal
	Kick down SW	To detect drive's high intension for acceleration
	Engine throttle opening	To detect drive's intension for acceleration
	Engine airflow meter	To detect engine output torque
	Temperature sensor	To detect CVT Oil temperature to know viscosity of oil
	Other sensors	It depends on the purpose of some additional special control Ex. Hydraulic pressure sensor, SW, g sensor, ect.
Actuator	Stepper motor	To control CVT speed ratio
	Solenoid value	To control CVT hydraulic pressure
	Engine torque adjusting device	To adjust engine torque
	Indicator lamp	To indicate P.R.N.D.L range status to driver
	Starter relay	To allow engine starting during P.n range only
Communicater	Engine—CVT communicator	To agree engine torque adjusting
	Other vehicle ECUs—CVT communicator	To make communication with ABS, ASCD, NAVT, ect. for some additional special control
	*Recently both above communications are usually done by CAN network	

(c)

Figure 1. A typical CVT control system. (a) Cross section of typical belt CVT. (b) Typical CVT control system diagram. (c) Devices for CVT control. (Reproduced by permission of Jatco, Ltd.)

There are many kinds of CVT ratio changing controls. Some CVTs have stepper motors to control speed ratio with ratio control valves mechanically connected to the stepper motor. Some use pressure control devices for the ratio control of primary and secondary pulleys. Stepper motor control systems are popular traditional systems for CVT ratio changing control, so the example figure shows a stepper motor.

Communication between the engine and the CVT is very important for keeping enough torque capacity on the belt and the pulley systems preventing belt slipping. In addition, communication between engine and CVT, by engine torque control, helps engine speed changing during CVT ratio changing.

Vehicle CAN (controller area network) brings good communication with other vehicle ECUs such as ABS

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

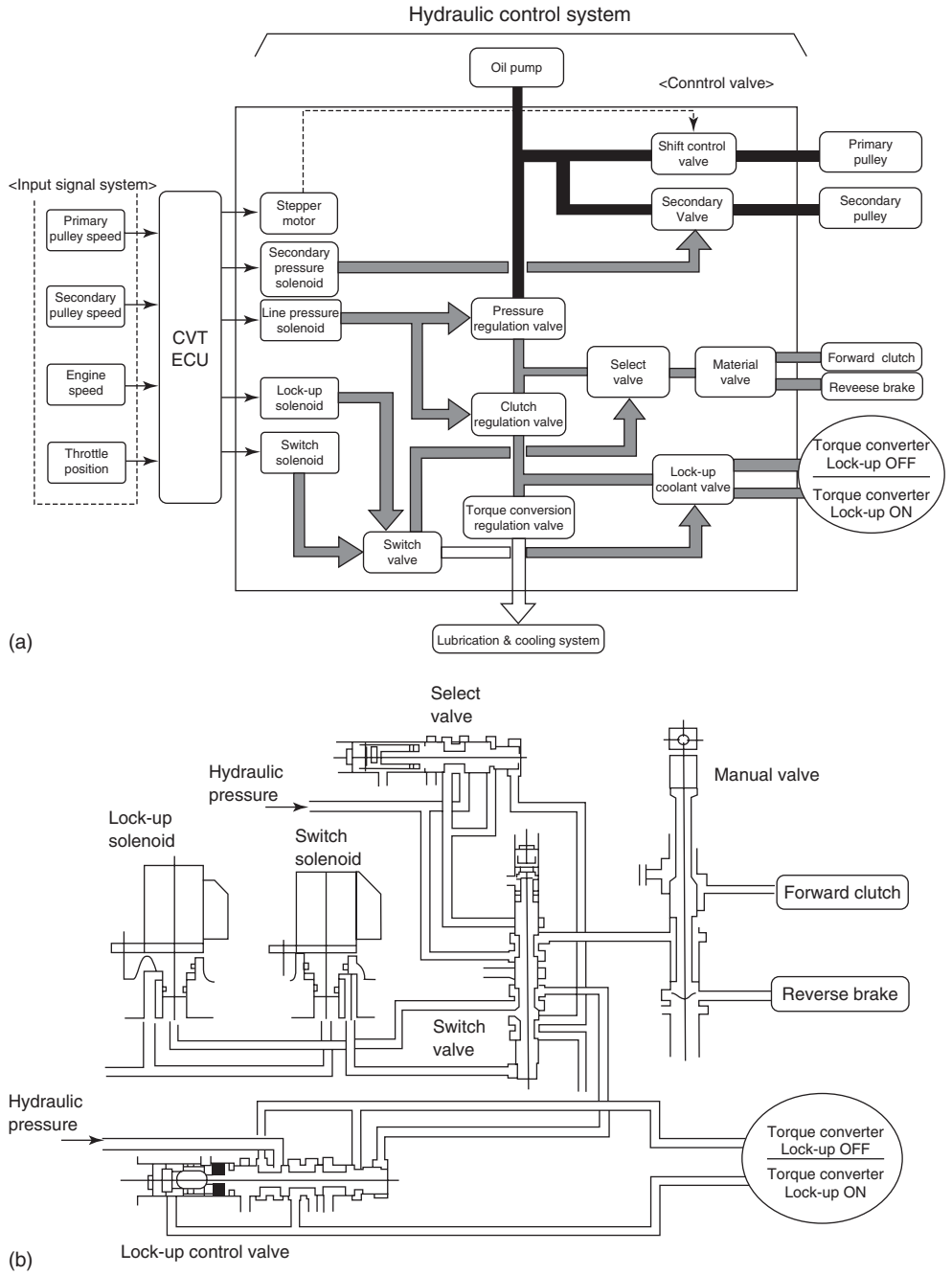


Figure 2. A typical CVT hydraulic control system. (a) Typical CVT hydraulic control diagram. (b) Control circuit for clutches. (Reproduced by permission of Jatco, Ltd.)

(antilock brake system) and NAVI (automotive navigation system). The CVT ECU can modify the ratio changing schedule based on this information.

Figure 2a shows CVT hydraulic control diagram (Sugano *et al*., 2003, p. 22). Hydraulic power is supplied by an oil pump and its flow is delivered to the belt and pulley system, the forward and reverse clutches, and the torque converter with lockup clutches. Clutches are controlled by hydraulic and solenoid valves, as shown in Figure 2b.

In a CVT, the clutch function is very simple: that is, forward and reverse only make engagements at very low or vehicle stop conditions. On the other hand, the torque converter lockup clutch makes engagements at a variety of vehicle speeds. Therefore, usually, pressure

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

control linear solenoid valves are shared and "multiplexed" between garage shifting (shifting from P,N to D,R at vehicle stop condition) control and torque converter lockup clutch control. A selected s/w valve will switch the circuit from garage shifting clutch to torque converter clutch after launching.

3 RATIO CHANGING CONTROL OF CVT

As is described in detail in Article (see The variable pulley CVT), to prevent belt slipping and to keep some speed ratio, clamping force control is important. Ratio change is made by clamping force control. Figure 3a shows primary and secondary clamping forces.

Figure 3b shows a typical chart for indicating clamping force ratio to keep a required speed ratio. As shown, this changes with transmitting torque.

On the line shown in Figure 3b, the CVT does not change ratio, but maintains a constant speed ratio. However, if clamping force is changed on the primary or secondary pulley, the clamping force balance is broken and speed ratio change begins.

For example, to make shifting from A to B, the shift controller will increase primary pulley clamping force to high, then the speed ratio begins to change from A to B [i.e., toward over drive (OD)]. The shifting speed depends on how large the primary clamping force is and the hydraulic flow determines how fast the pulleys can move.

This is the basic concept of CVT shifting control.

In order to provide such clamping force control, several kinds of clamping force generation systems and control systems have been invented.

1. clamping by hydraulic power;
2. clamping by electric motor power;
3. clamping by loading cam.

Sometimes, these methods are combined together.

Figure 3c shows the principle of a loading cam. By this cam force, the transmitting torque can be obtained without any hydraulic or electric power. Some CVTs use this mechanism.

Hydraulic control is the most popular way. However, there are two approaches

1. Direct control:

Primary and secondary hydraulic pressures or flows are controlled independently by solenoid valves.

2. Stepper motor and ratio control valve system:

In this system, stepper motor provides speed ratio changing control, and mechanical linkage provides feedback to the ratio control valve.

There are two approaches to the direct control system: primary control system or dual control system. Figure 4a shows a typical example of primary pressure direct control system. Primary pulley actuating pressure is controlled for ratio changing. The line pressure control solenoid valve provides line pressure, which is applied to the secondary pulley, whereas the primary pulley pressure is controlled by the primary control solenoid valve. The difference between primary and secondary pressures can be made by the line pressure control solenoid valve and the primary control solenoid valve. The controller calculates the speed ratio detecting pulley speeds from speed sensors. Feedback control to primary pulley pressure is adapted to control the speed ratio.

Figure 4b shows another typical direct control system of pulley actuating pressure for ratio changing. It is almost the same as the system described earlier, the difference being the secondary pressure control method. A secondary control solenoid valve is added. By this additional valve, the controller can make secondary pulley pressure lower than its feed pressure (line pressure). As a result, operation during low engine torque conditions such as cruising can be accomplished with low clamping force and low friction loss in the belt and pulley systems for good fuel economy.

However, in this system, parts count and system cost increase. Speed ratio is again calculated by the controller by detecting primary and secondary speeds. Using the calculated speed ratio, the controller provides feedback control of both pulley pressures.

Figure 4c shows the stepper motor and ratio control valve system with a mechanical linkage (Abo *et al*., 2003; Singh *et al*., 2003). One end of the linkage is connected to stepper motor, which makes linear movement step by step. Another end of the linkage is connected to mechanical sensor, which is always slipping on the primary pulley back surface. If the primary pulley strokes to the right or left as described in the figure, the linkage end will move right or left along with the primary pulley movement.

In the control from UD (under drive) to OD, the stepper motor moves the linkage to the right, and the ratio control valve provides oil flow charge to primary pulley chamber, which initiates the ratio changing toward OD. When the OD shift is completed, the primary pulley back surface will move to the left and make the ratio control valve to stop charging the primary pulley chamber.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

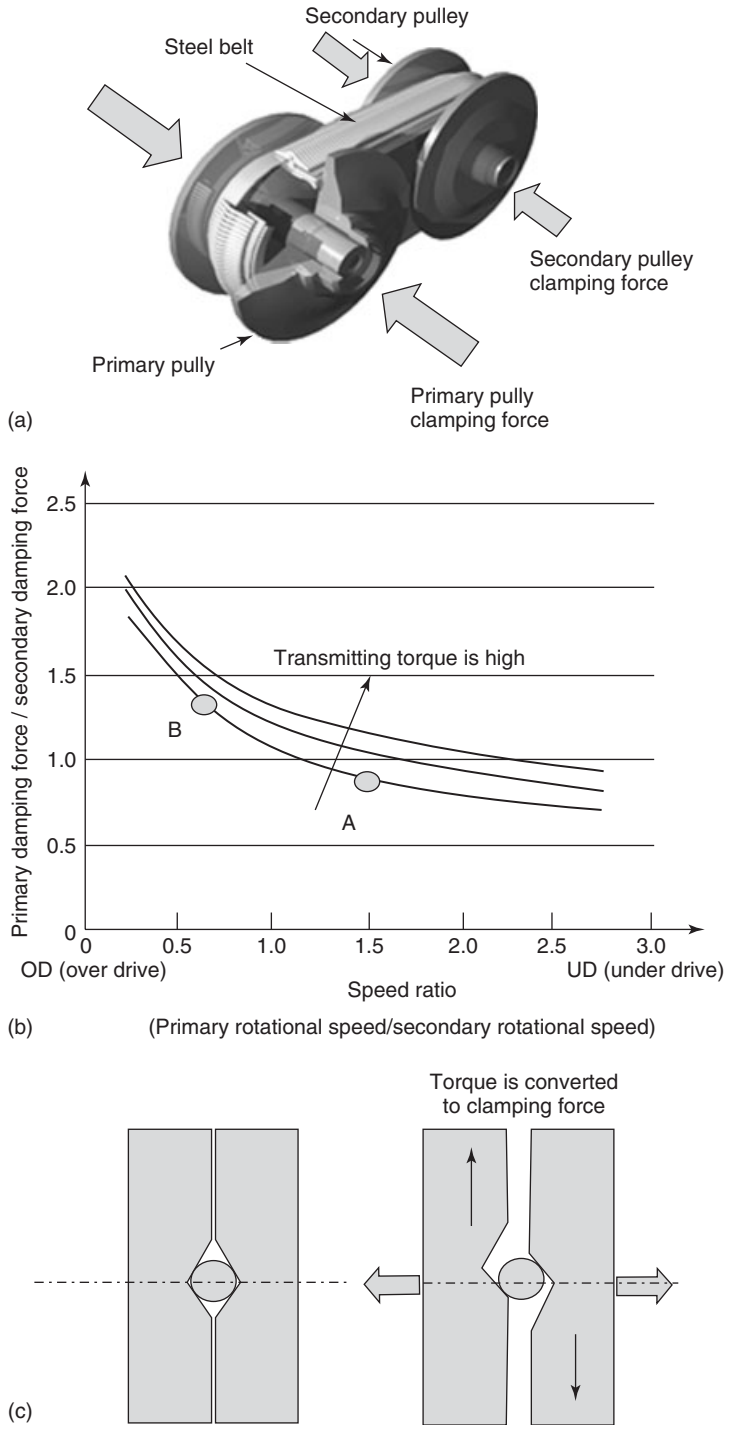


Figure 3. CVT ratio changing control by the clamping force control. (a) Clamping force. (b) Primary/secondary clamping force ratio characteristics. (c) Principle of loading cam. (Reproduced by permission of Jatco, Ltd.)

In this way, the linkage works as a mechanical sensor for hydraulic feedback control. The controller detects the speed ratio by speed sensors and makes adjustments of the stepper motor movement. In such a way, this system makes both mechanical and electrical feedback adjustments, thereby providing a more stable ratio changing system. Calibration is easier with this system than a direct control system, so many CVT designs adopt this control approach.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

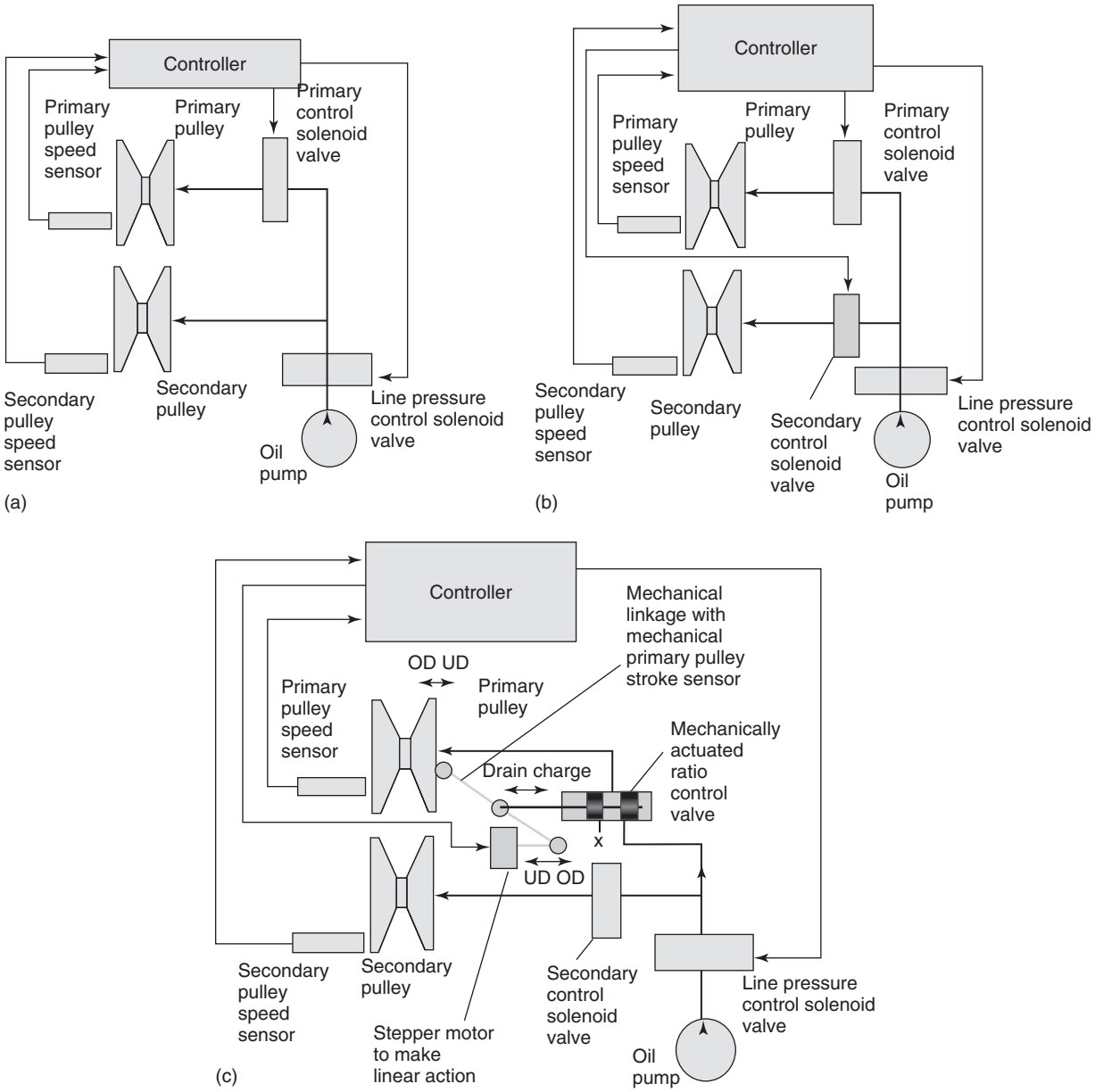


Figure 4. Various hydraulic systems for CVT ratio changing control. (a) Direct control system (primary control system). (b) Direct control system (dual control system). (c) Stepper motor and ratio control valve system. (Reproduced by permission of Jatco, Ltd.)

In CVT development, ratio speed change and accuracy are evaluated, as they have a large influence on vehicle response and drivability. Figure 5a shows various evaluation points for shift control performance. Various values are measured to evaluate the difference between target and actual ratios.

Rapidness of ratio changing is a critical evaluation point because it sets the requirement for oil pump capacity, and size of hydraulic channels and valves. These factors can influence space and cost issues as well as oil pump loss issues, which is a very important factor for fuel economy.

Figure 5b shows the explanation. Because pulleys are actuated by hydraulic flow and pressure, high capacity pumps, large valves, and flow channels need to be provided to ensure rapid response of the pulley system. The most severe operating condition need to be evaluated to determine those specifications. Kick down acceleration or rapid deceleration conditions often becomes the most severe condition.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Figure 5c shows such an evaluation test example. At some defined vehicle speed, the driver presses the brake pedal strongly to give a rapid slowdown condition. With poor system performance, the speed ratio change will lag to the vehicle deceleration rate, and the subsequent relaunch of the vehicle will have slower acceleration because of not having completed the shift and achieved the required UD ratio. Usually, the result of such an evaluation requires a larger pump, and the designer often realizes a trade-off conflict between fuel economy and pick-up performance immediately following a slowdown event.

To solve such a trade-off situation, many ideas have been invented and are being studied. For using an electric motor--driven pump or loading cam, there are two approaches. Figure 5d shows another idea of reusing hydraulic power and flow. In shifting, oil flows from one pulley to another and can be controlled via a switching valve.

4 SHIFT SCHEDULE OF CVT

Figure 6a shows a typical shift schedule of a CVT. It is different from that of automatic transmission. The vertical axis is primary pulley speed instead of throttle opening. Throttle opening is a parameter; if throttle opening is held constant, primary pulley speed will move along the throttle opening line during acceleration. If the throttle pedal is kicked down, the primary speed will increase according to the established shift schedule. Multiple throttle opening shift lines can exist between OD and UD lines. Within this area, designers can make any line freely because a CVT can make any speed ratio between OD and UD. In a torque converter lockup condition, the primary pulley is equal to the engine speed; therefore, shift schedule decides engine speed changing behavior directly.

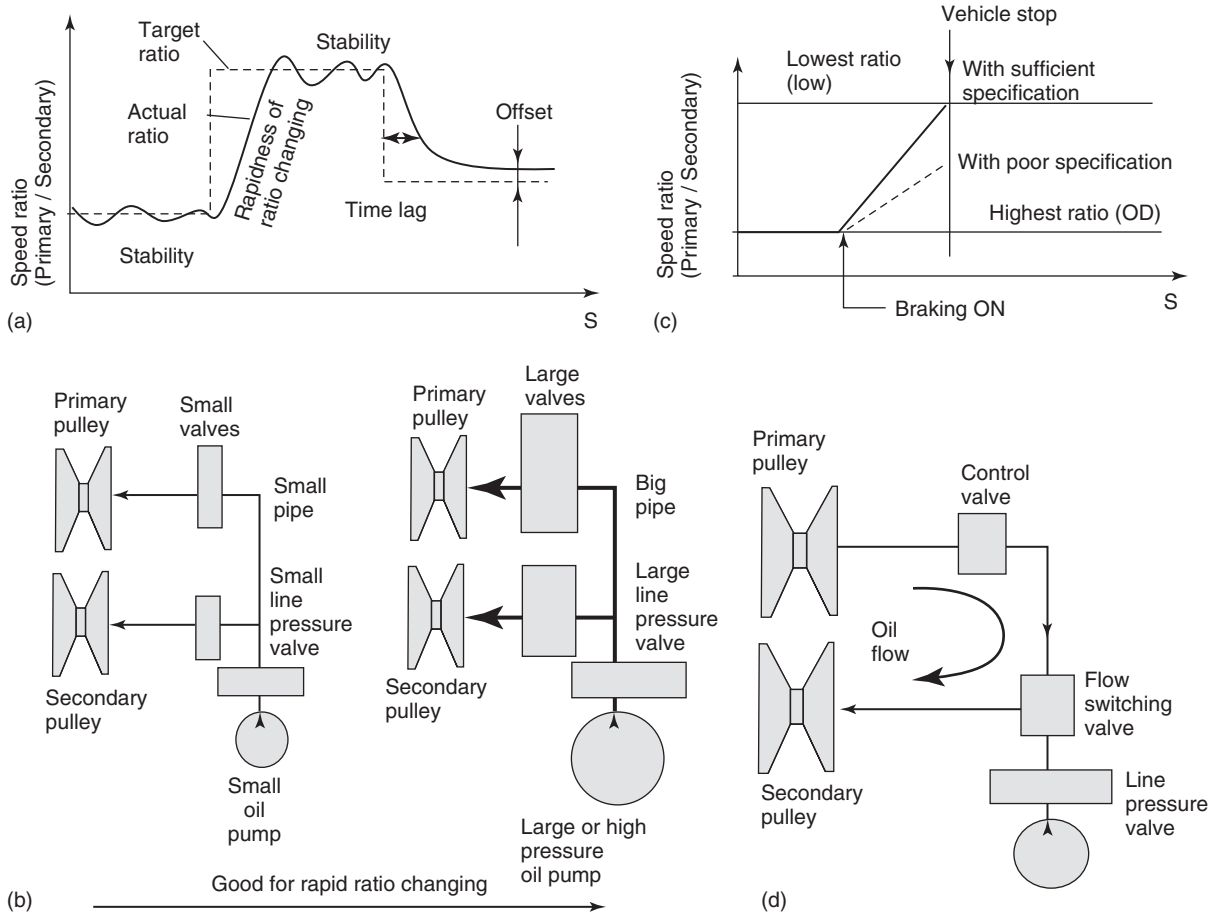


Figure 5. Typical evaluation points for CVT ratio changing control performance. (a) Evaluation points of shift control performance. (b) Improving point for rapidness of ratio change. (c) Immediate vehicle slowdown evaluation. (d) Improvement concept of reusing hydraulic power and flow. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

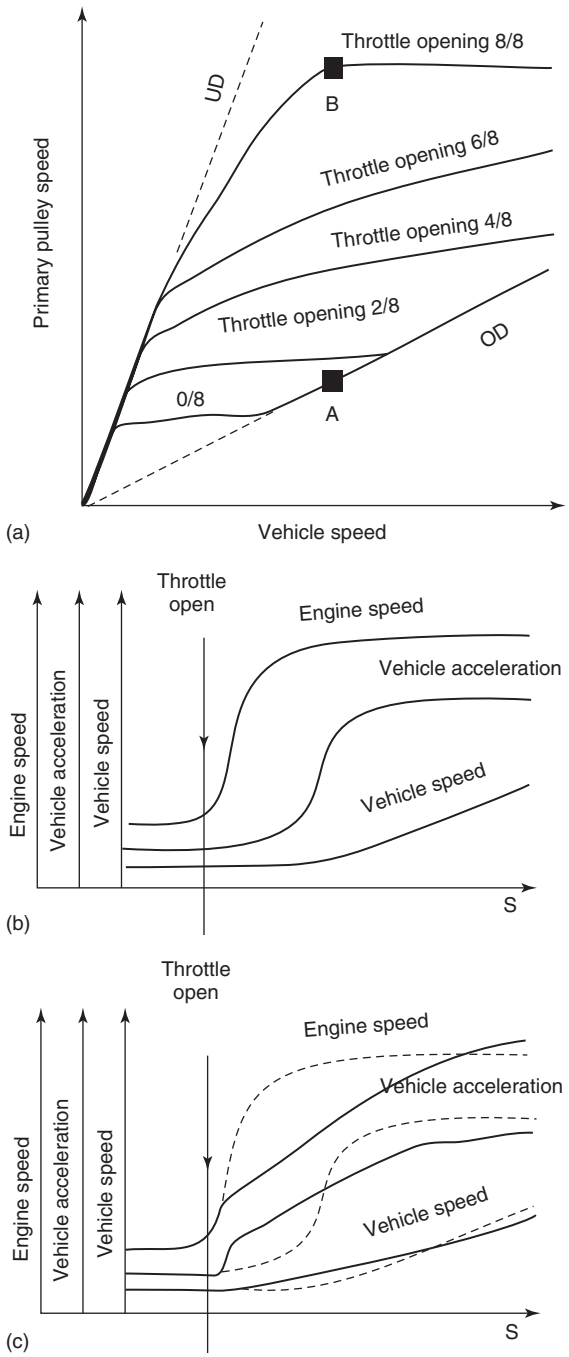


Figure 6. A typical CVT shift schedule and shift feeling improvement. (a) Typical shift schedule of CVT. (b) Rubber band feeling. (c) Improvement strategy of CVT down shift feeling. (Reproduced by permission of Jatco, Ltd.)

5 CVT RATIO CHANGING CONTROL STRATEGY

In CVT shifting control, prevention of the rubber band feeling is important. Figure 6b explains the meaning of "rubber band feeling." When a driver steps on the throttle, the CVT makes a rapid ratio change and the engine speed will increase rapidly. However, a too rapid engine speed increase provides a bad feeling to drivers, like both ends of a rubber band. Only engine speed increases, but vehicle speed does not increase concurrently and vehicle acceleration does not happen during this engine speed increase because of engine inertia consuming available output torque. After engine speed goes high, acceleration will come down. That feeling is really like stretching a rubber band. In making shift schedules and shift controls, software is often improved to avoid such rubber band feeling.

Figure 6c shows the improved control. Engine speed rising up is restrained by the control to enable earlier vehicle acceleration. The vehicle response feeling becomes much better. Furthermore, vehicle speed increasing with the increasing engine speed brings a good feeling to the driver as if the vehicle is really accelerating well. By such a control approach, linear acceleration feeling can be enjoyed by the driver.

Theoretically, such acceleration is not the fastest acceleration, but this control approach provides a better feeling to the driver. Therefore, such an approach has become very popular and is adopted by many CVTs.

Some CVTs adopt another shift schedule approach as shown in Figure 7a to obtain good acceleration feeling similar to an AT. In this control, the CVT makes ratio changes like a step AT. Rhythmical engine speed rising and falling repeatedly brings a sporty and active feeling to the driver. Such control is effective especially at wide-open throttle conditions. Therefore, usually, such an approach is adopted for wide-open throttle acceleration only. For light acceleration conditions, first priority is put on fuel economy in the shifting control strategy.

For best fuel economy, the first priority is the good operation point of the engine. As is shown in Figure 7b, the CVT control makes speed ratio adjustments to achieve the optimal operation point of the engine as much as possible. Because the CVT can make continuous speed ratio changes, it can adjust better than any stepped transmission. That is one reason why the CVT has fuel economy benefits compared with the step AT.

Another reason why the CVT has a fuel economy benefit is that an earlier lockup of the torque convertor can be employed. Torque converter lockup control can be made at lower vehicle speeds, just after launch. Therefore, the energy loss at the torque converter is less in a CVT. In step ATs, such an early lockup leads to hard NVH (noise, vibration, and harshness) issues when the driver steps into the throttle at low speed. However, in a CVT, it can make rapid ratio changes to avoid NVH issues. Such early

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

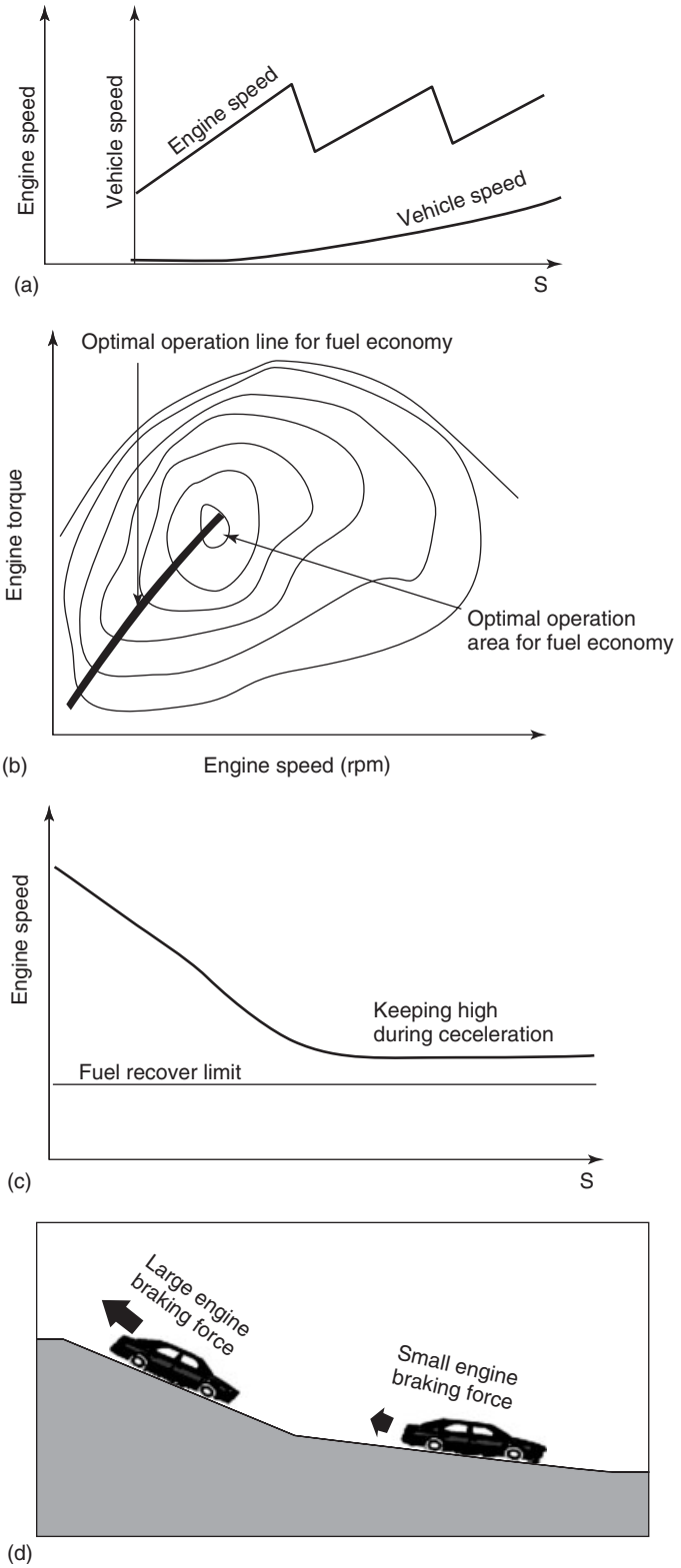


Figure 7. Various CVT ratio changing control to achieve good performance. (a) Discrete acceleration control. (b) Strategy to obtain good fuel economy. (c) Strategy to keep fuel cut control. (d) Automatic engine braking. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

lockup control is one of the important strategies in CVT control.

Much longer utilization of fuel cut control is one more benefit of CVT. Figure 7c shows the explanation of the strategy. During deceleration, fuel can be saved by adopting fuel cut control by the engine controller. By continuous precise speed ratio adjustment ability of CVT, engine speed can be kept to a high enough level beyond the fuel recovery limit where engine stall would occur. By controlling its speed ratio continuously, a CVT can make a much longer fuel cut duration than a step AT.

In such a deceleration condition or on a downhill road, the CVT can automatically adjust engine braking. Figure 7d shows the strategy of automatic engine braking adjusting control. By detecting unusually large accelerations compared with a normal flat road, the controller can detect a downhill condition. Steepness can be calculated by watching vehicle acceleration, and adequate speed ratio can be determined by the CVT controller. On a steep downhill road, large engine braking force is obtained by shifting to more low, and on a slight slope downhill road, slight engine braking force is obtained by shifting less low. Here again, the benefit of CVT is smooth ratio changing. The CVT controller can provide a very slow ratio change at the beginning of automatic engine braking adjustment. If a step AT controller makes a similar automatic engine braking control, a sudden shift down will be occur and downshift shock will happen.

6 OTHER RATIO CHANGING CONTROL OF CVT

The automatic engine braking control described earlier is one of the adaptive control features of a CVT. There are many other adaptive control features, similar to those of a step AT control, which were described in Article 5.7. For various road conditions and characteristics of driver

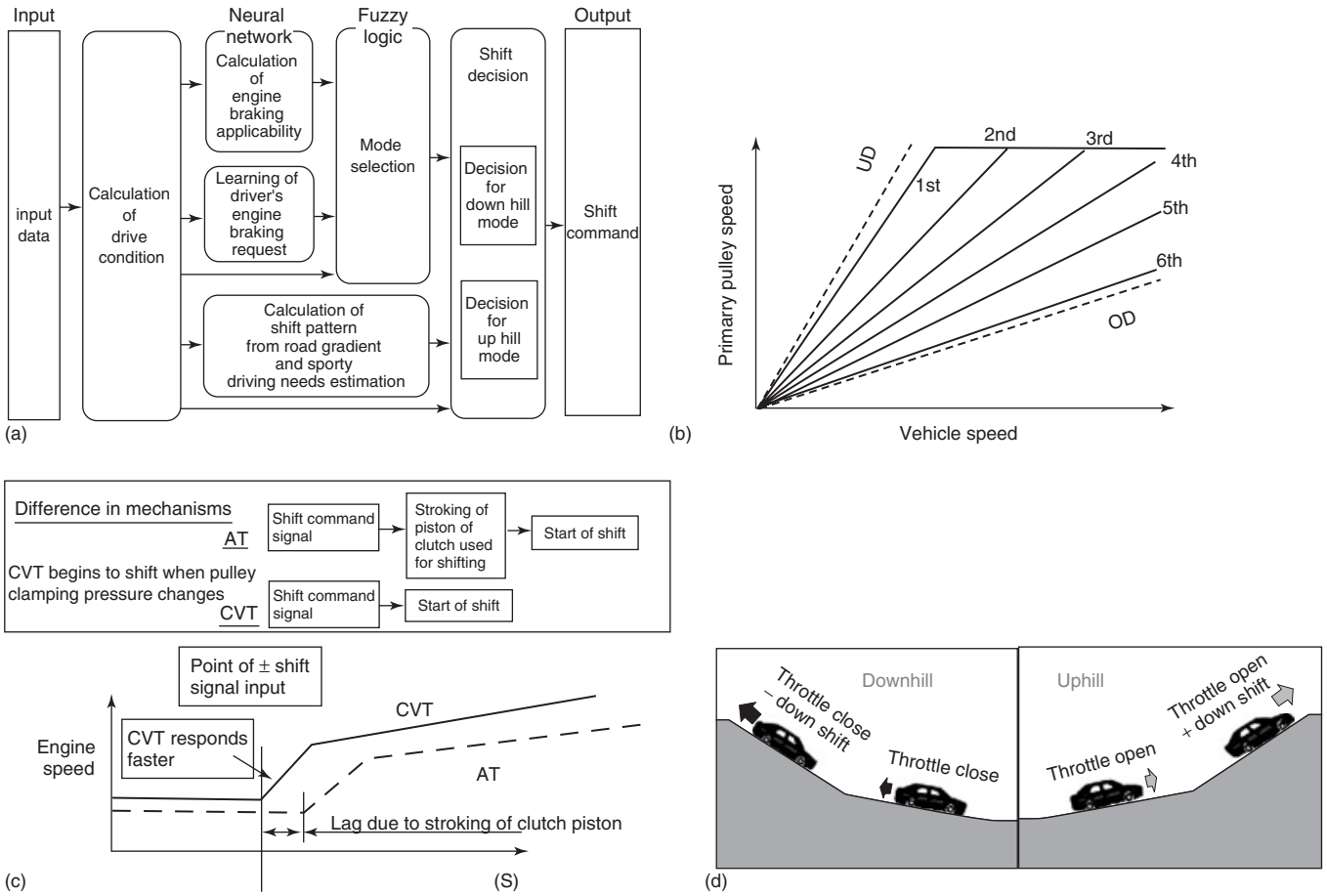


Figure 8. Other ratio changing control of CVT. (a) An example of adaptive shift schedule control diagram. (b) Typical manual shift schedule for CVT. (c) Comparison of initial shift response. (d) Downhill- and uphill-integrated controls. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

operation, various shift schedules can be determined by the computer. Figure 8a shows an example of a block diagram of an adaptive shift schedule control.

Manual shift control can also be adopted, similar to a step AT. Figure 8b shows a typical manual shift schedule of a CVT. In such a manual mode, speed ratio is kept at the same value until another shifting change signal comes. Like a step AT, the CVT controller changes the speed ratio step by step.

For such a manual shift mode, quick response to shifting switch operation is very important. From this standpoint, the CVT system has theoretically a good feature. Figure 8c explains that in detail. When shifting is executed, the engaging clutch of a step AT has to make clutch stroke at first. On the other hand, the CVT can begin shifting immediately after deciding ratio change is desired. Therefore, theoretically initial response is different. CVT does not have an initial lag by clutch stroking like a step AT, so the CVT provides a quick manual shifting system. Gear ratio selection is free; therefore, in a CVT, making 8 speeds, 9 speeds, or manual shift control can be easily realized.

In manual mode, some protection control must be adopted. For example, engine speed will not go into overspeed zone, as automatic upshifting will occur. When vehicle speed decreases to a point at which the engine may stall, automatic downshifting will occur. This is also effective for the next launch to obtain enough driving force.

As is described in Article 5.7, the switch for manual mode has a lot of possibilities. However, the basic concept is exactly the same as an AT providing a +SW (steering wheel) and −SW is popular. Sometimes, a SW or paddle shift levers are provided.

The CVT has another benefit when the driver makes shift commands very frequently; that is, the CVT can make repeated shifts as frequently as the driver's commands without lag time, because there are no friction clutches making frequent disengagements and engagements, causing friction material heating issues.

7 ENGINE AND CVT INTEGRATED CONTROL

As is similar to the description of AT-integrated control in Article 5.7, engine and CVT-integrated control is often adopted. Engine control participates in adaptive controlling. Figure 8d shows an example of such an integrated control. Here, the engine throttle opening is controlled to provide optimal driving or deceleration forces for downhill or uphill driving. Another integrated control is engine torque control during up shifting or down shifting. As for step AT control, engine torque control helps engine speed changing during shifting.

8 CVT CLAMPING FORCE CONTROL

Clamping force control is important for both fuel economy and reliability of belt system. As shown in Figure 9, there

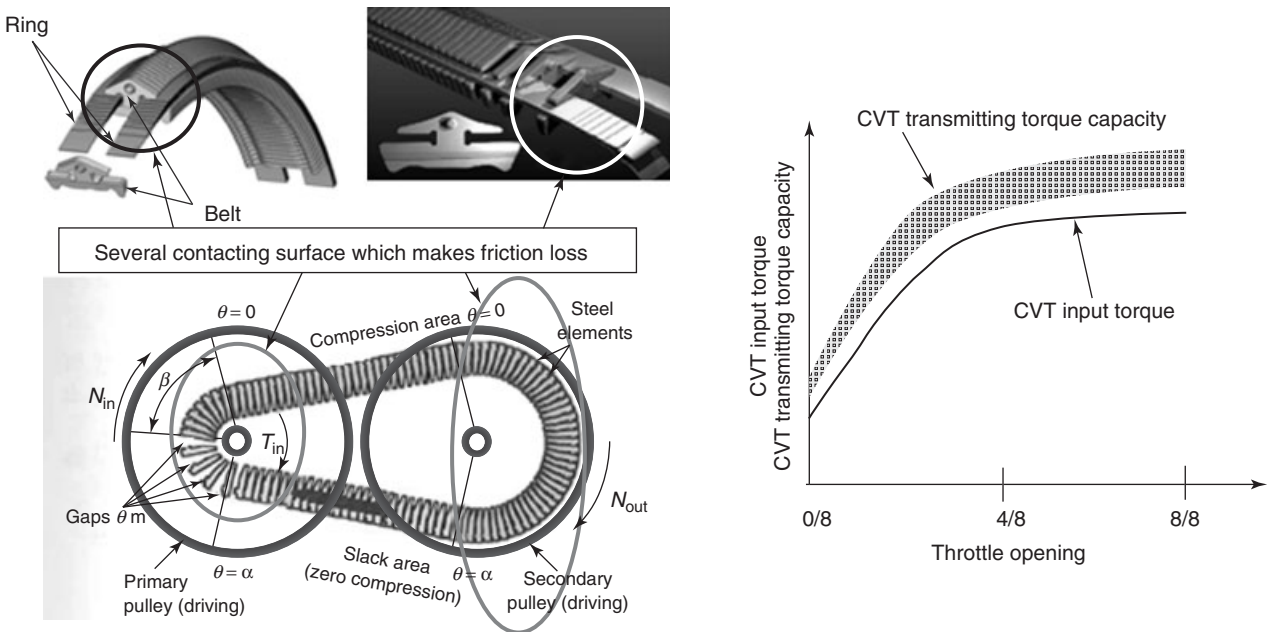


Figure 9. CVT clamping force control. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

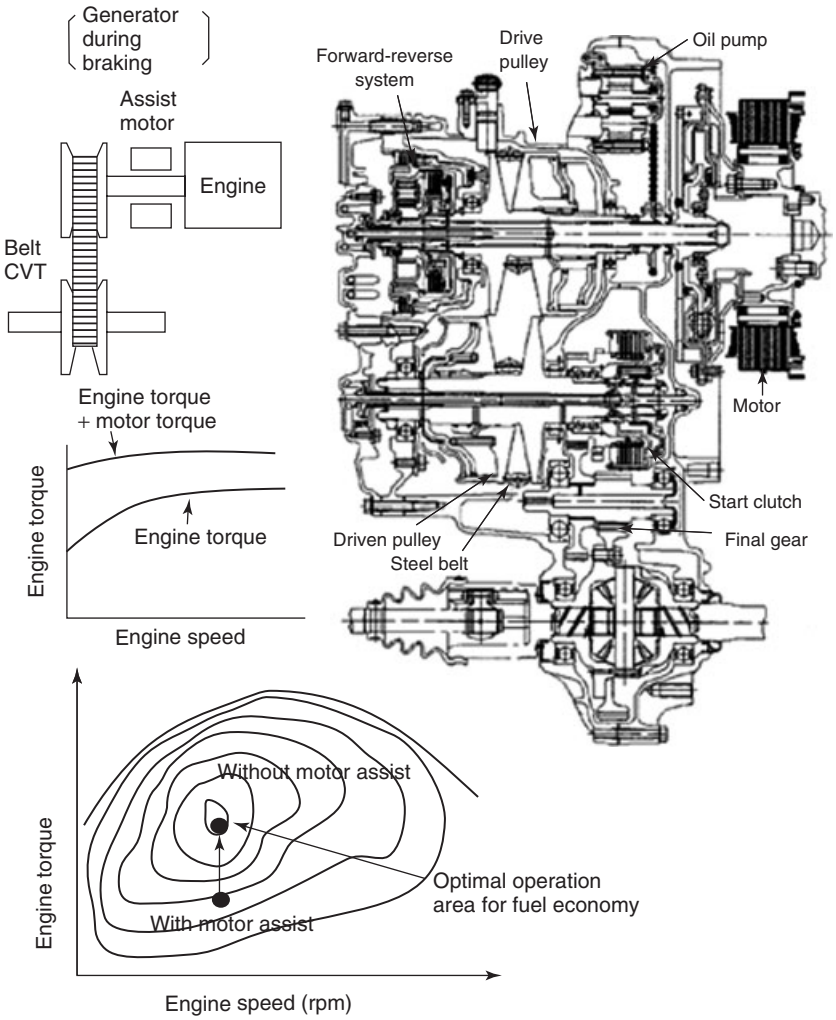


Figure 10. Honda CVT ratio changing control for HEV. (Reproduced by permission of Honda, Japan © Honda Motor Co., Ltd.)

are several contacting surfaces in a belt and pulley system at which friction loss is generated. The loss will increase as clamping force is increased. Therefore, clamping force should be kept as low as possible. On the other hand, torque transmitting capacity is related to clamping force, so clamping force needs to kept high enough to transmit torque without the risk of belt slip. That is trade off of clamping force control, if clamping force is not high enough, belt slipping will be caused. In actual CVT control, some margin is adopted on hydraulic pressure to provide enough clamping force to prevent belt slip. Engine torque signal is used for the clamping force control, along with the hydraulic pressure sensor. If hydraulic pressure cannot be achieved for some reason, the CVT controller will request the engine controller to reduce torque output. The clamping force should also be increased during engine braking because braking torque is added to the belt and pulley system. The CVT controller can detect various situations and make various control adjustments to keep optimal clamping force.

9 CVT CONTROL IN HEV SYSTEM

To obtain excellent fuel economy in a hybrid electric vehicle (HEV) system, a CVT is very effective in obtaining optimal operating condition. Figure 10 shows the Honda HEV system that has an electric motor and a belt CVT system (Ota and Eguchi, 2007, p. 183). The electric motor will act as power assistance during acceleration or as a generator during deceleration. In this system, motor torque and the CVT speed ratio are controlled to obtain optimal engine operating condition for achieving good fuel economy. For example, during slight acceleration, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

motor assist torque is not generated to make the engine torque higher as is shown in the figure. Thereby, the engine operation condition is moved to go into optimal operation area. CVT speed ratio is adjusted to obtain the driving force that the driver demands, and keeps these engine and motor states.

Honda is also adopting motor and engine integrated control. Motor assist torque is generated to make engine torque higher during high acceleration. During low speed cruising, the engine becomes idling and the valves of the engine stops their movement to reduce pumping losses, and only the motor torque is used for the vehicle driving.

10 CONTROL FOR CVT WITH AUXILIARY SHIFTING SYSTEM

Recently, a CVT with an auxiliary shifting system has been released to the market. Figure 11 shows the structure of this CVT (Nakagawa, 2012, pp. 23–30). It has a 2-speed shifting mechanism in D range. By adopting that system, ratio coverage has become much wider than a conventional CVT and pulley size has become smaller, thereby enabling weight reduction and stirring loss reduction because of fewer rotating parts in the oil sump. In this CVT, up shifting of the belt pulley system and down shifting of the auxiliary shifting system are simultaneously controlled.

Figure 11 shows such a simultaneously shifting control. At the same time the auxiliary shifting system makes an up shift, the CVT variator system makes a downshift. Therefore, the driver does not notice any ratio disturbance at this shift point. By this control, a continuous shifting feeling can be maintained.

11 CHAIN CVT CONTROL

Up to this section, the belt CVTs have been used for the explanation for CVT control because the belt CVT is the most popular system. However, there are other

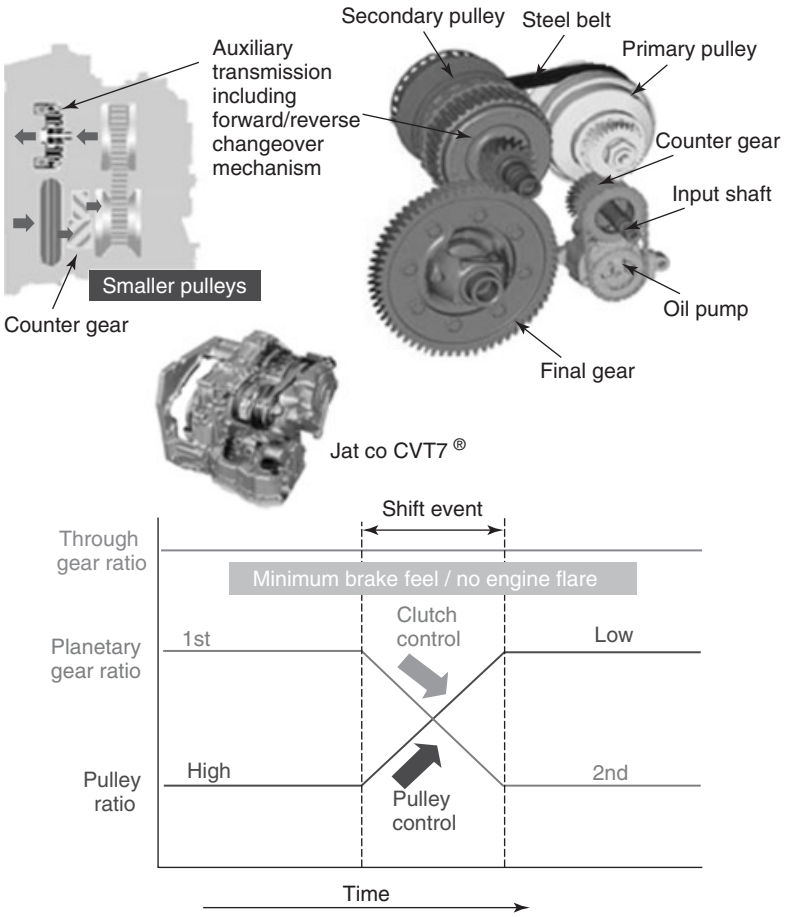


Figure 11. Control for CVT with auxiliary shifting system. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

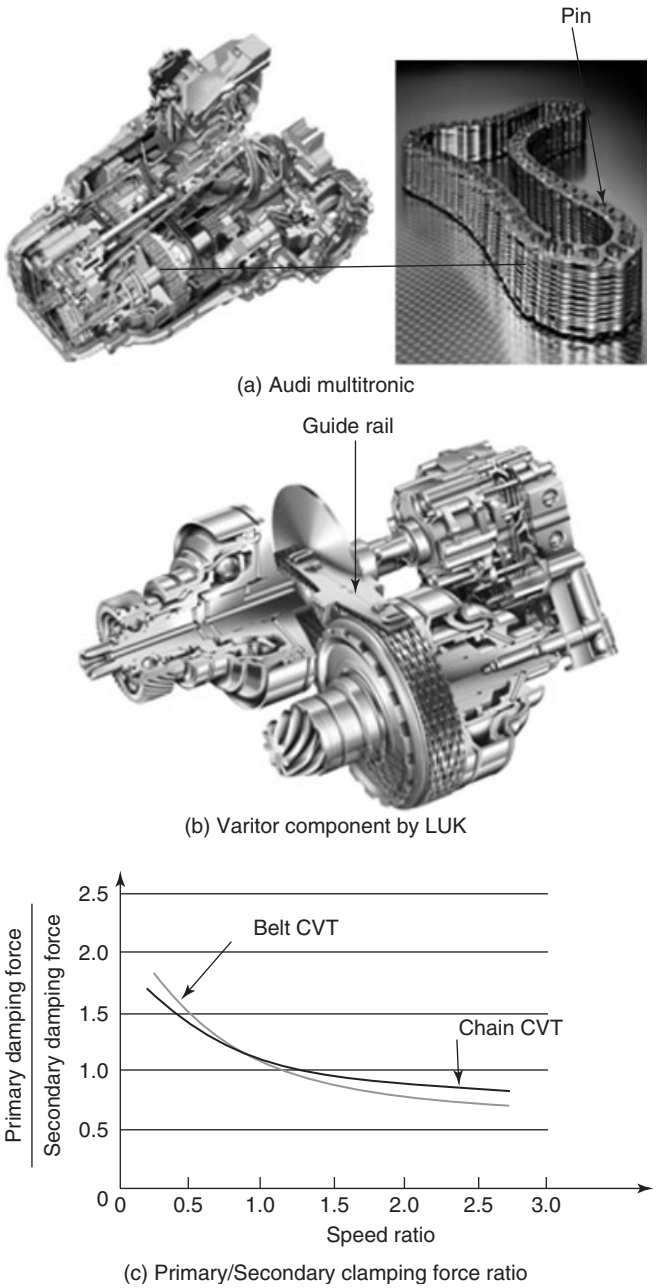


Figure 12. Typical chain CVT and required clamping force characteristics. (a) Audi Multitronic®. (Reproduced by permission of LuK (UK) Ltd, Schaeffler Group.) (b) Variator component by LUK. (Reproduced by permission of LuK (UK) Ltd, Schaeffler Group.) (c) Primary/secondary clamping force ratio. (Reproduced by permission of Jatco, Ltd.)

popular CVTs that use a chain system. Figure 12a shows a typical chain CVT, which is also described in Article 3.5 (Englisch, 2010, pp. 139–151). This is Audi Multitronic®.

Figure 12b shows a variator component supplied by LUK. A chain is used instead of a steel belt. The chain has pins and the end of pins is in contact with the surface of the pulleys. Basically, the control approach is similar to the belt CVT. Clamping forces of pulleys are controlled to affect speed ratio changes. Because the contact between pin and pulley is completely different from the contact between belt element and pulley of steel belt, clamping force characteristic is not same as Figure 6.

Figure 12c shows the difference between the characteristics of a belt CVT and that of a chain CVT. Of course, the calibrator of the chain CVT has to be aware of this difference. As is shown in Figure 12b, the guide rail is used to reduce the string vibrations of the chain.

The LUK variator system adopts several unique hydraulic controls. Figure 13a shows the hydraulic system. The cylinders to obtain basic clamping force and to adjust the ratio changes are separated from each other. The ratio change control is achieved by the adjusting pressures 1 and 2. As the figure shows, the adjusting pressures are applied to the small cylinder of the pulley.

To achieve a rapid ratio change, it is best to have such a small chamber that responds rapidly because of reduced oil flow demand.

On the other hand, basic clamping force is obtained by the clamping cylinder as is shown in the figure. The applied clamping pressure is very important, for the fuel economy, because the oil pump loss depends on the clamping pressure magnitude.

To adjust and obtain optimal clamping pressure, the torque sensor system is used. Figure 13b and c shows the system. The torque sensor system consist ball, ramp, and a clamping pressure control orifice. The clamping pressure is adjusted by the orifice in order to apply the adequate pressure values according to the input torque.

The dual stage torque sensor control systems reduce the clamping pressure in OD as is shown in the Figure 13b.The oil passage for OD is open to exhaust when the pulley moves to the OD position, and the clamping pressure is then provided by another pressure chamber, thereby reducing the pressure value. In the OD pulley position, the clamping pressure will be reduced as is shown in the figure.

Figure 13c shows a further advanced system. The variable torque sensor control consists of balls and variable ramps as is shown in the figure. Clamping pressure is adjusted and reduced continuously as is shown in the figure. The clamping pressure is lower than that of dual stage torque sensor systems, thereby better fuel economy can be achieved by this system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

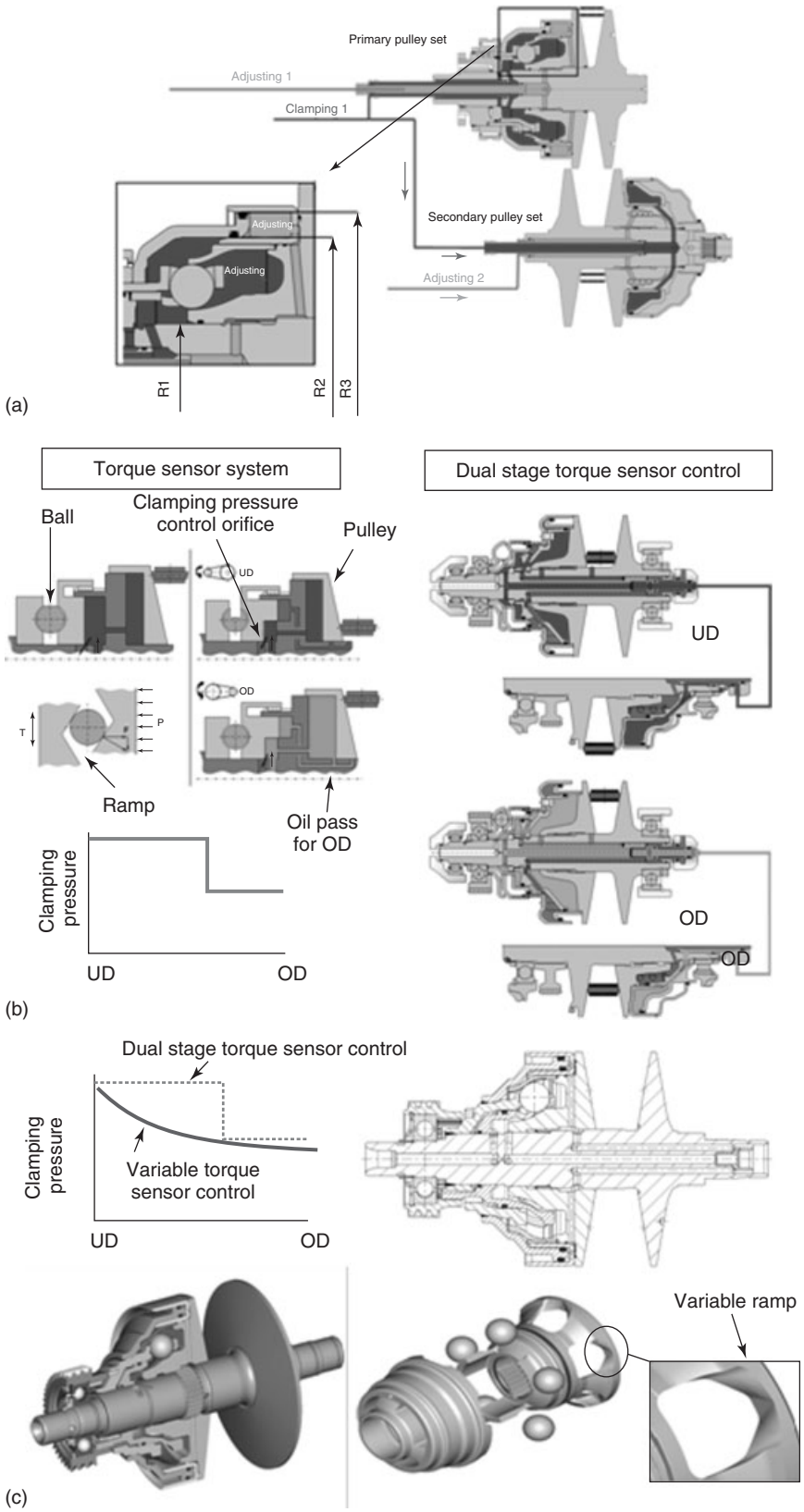


Figure 13. Various control systems for chain CVT. (a) Control systems for LUK variator. (b) Dual stage torque sensor control system. (c) Variable torque sensor control system. (Reproduced by permission of LuK (UK) Ltd, Schaeffler Group.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

12 CONTROL OF TOROIDAL CVT

In Article (see Traction drive CVT), the toroidal principle and related techniques have been described. Here, control of the toroidal will be explained.

Figure 14 shows a cross section of a toroidal CVT and its schematic model. It is the first toroidal CVT model that was sold in the market. In this CVT, clamping force for the power roller was provided by a loading cam (Shimanaka *et al*., 2001; Hibi, Sumi, and Takeuchi, 2002; Kawaguchi

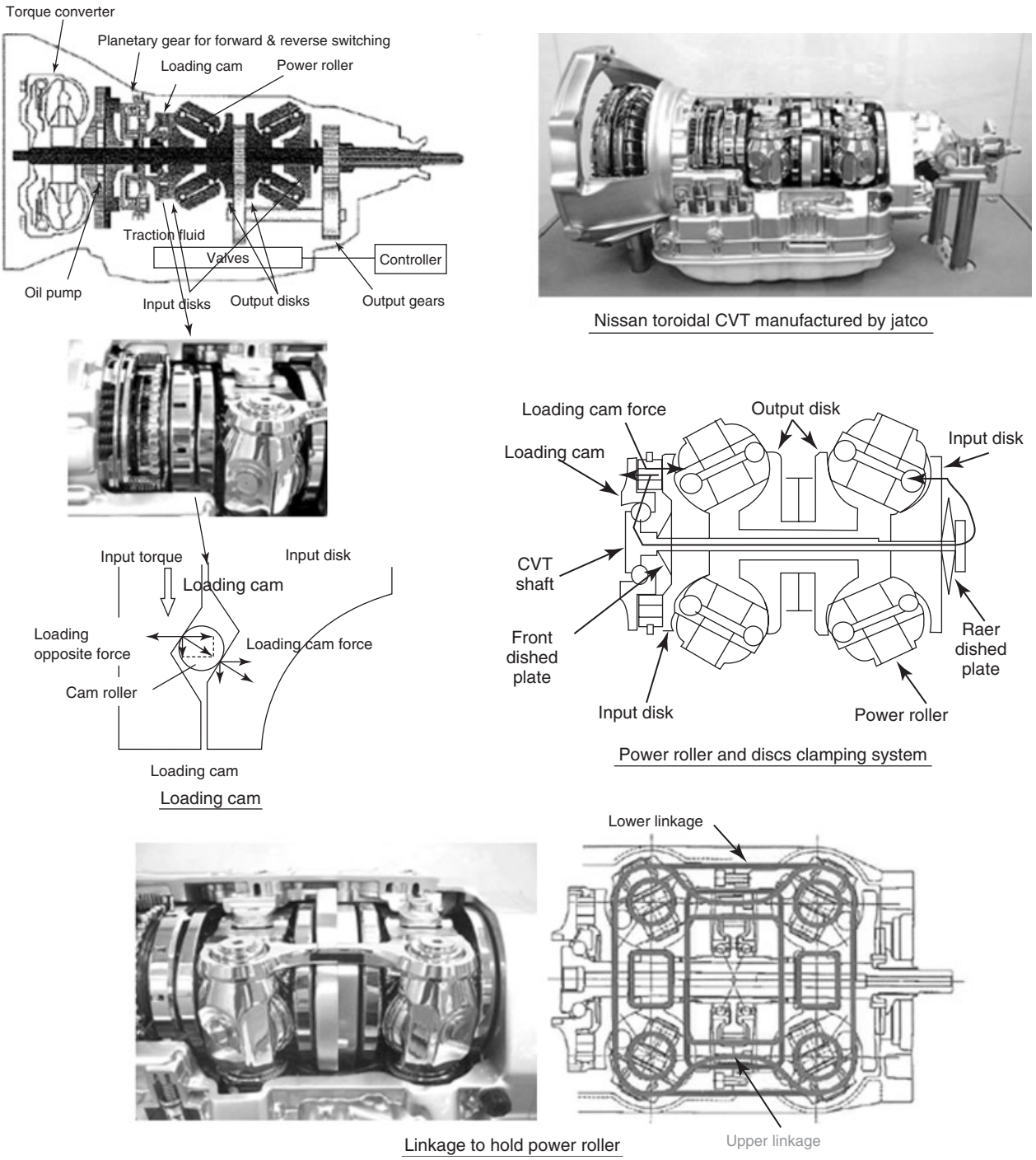


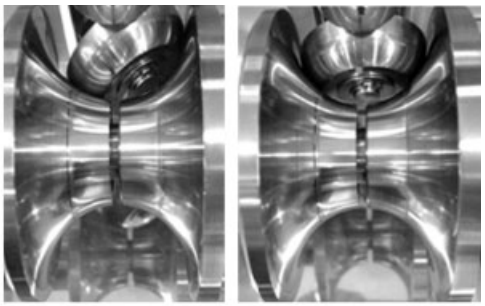
Figure 14. A cross section of a toroidal CVT and its schematic model. (Reproduced by permission of Jatco, Ltd.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

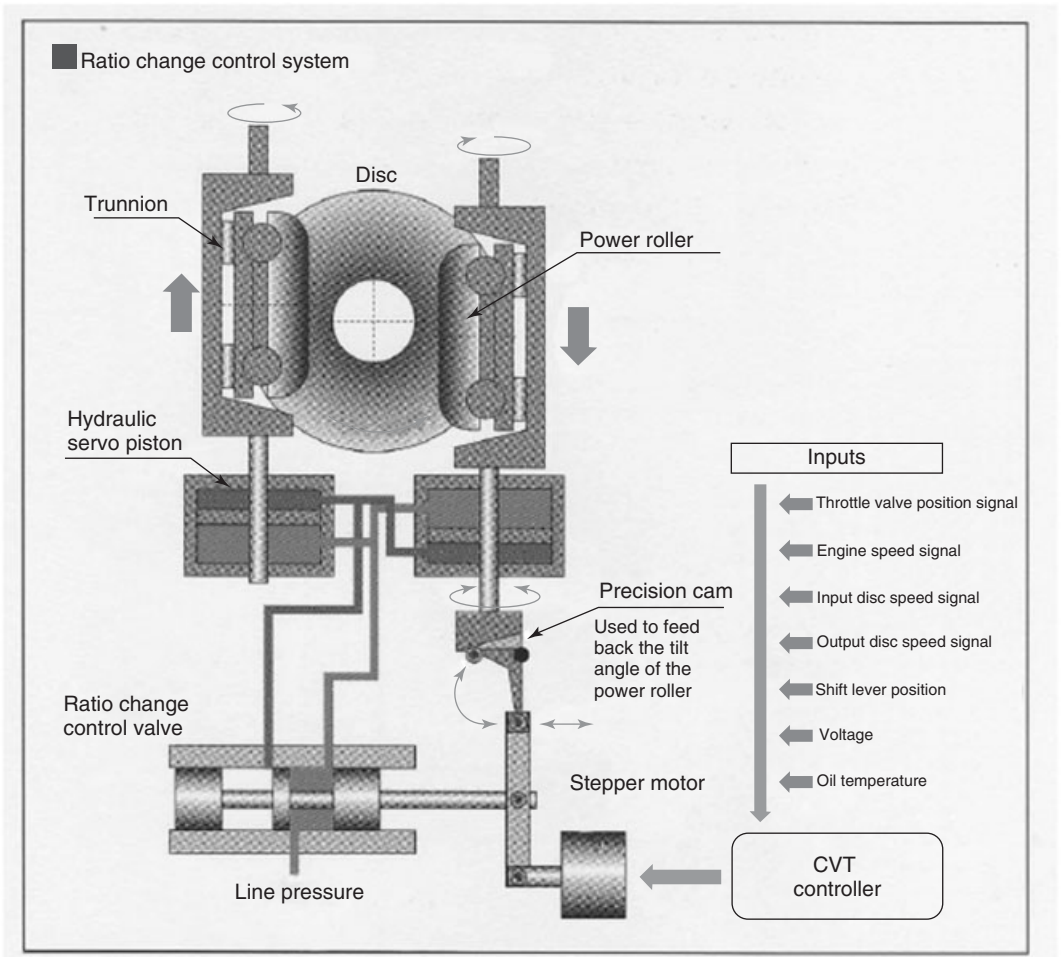
et al., 2000; Toyoda, Imanishi, and Inoue, 2010, pp. 67–72). Figure 14 shows the loading cam profile and its picture. When engine torque comes from the engine, the loading cam converts that torque and makes an axial force to obtain provide force for the discs and power rollers. As is shown in figure, this axial force is transmitted through the CVT shaft, and all power rollers and discs in the two cavities are clamped by the same force. Therefore, in this system, high hydraulic pressure as with the belt CVT is not necessary. Clamping force can be generated



(a)



(e)



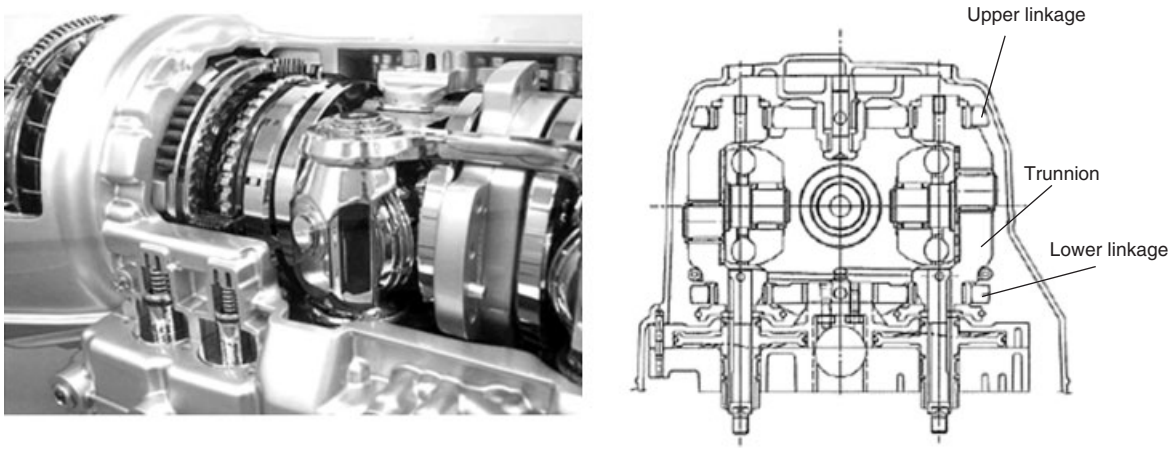
(b)

Figure 15. Ratio changing control principle of toroidal CVT. (a) Power rollers and disc. (b) Speed ratio control system. (c) Trunnion actuating system. (d) Speed ratio control by tilting control. (e) Speed ratio changing of troidal CVT. (Reproduced by permission of Jatco, Ltd.)

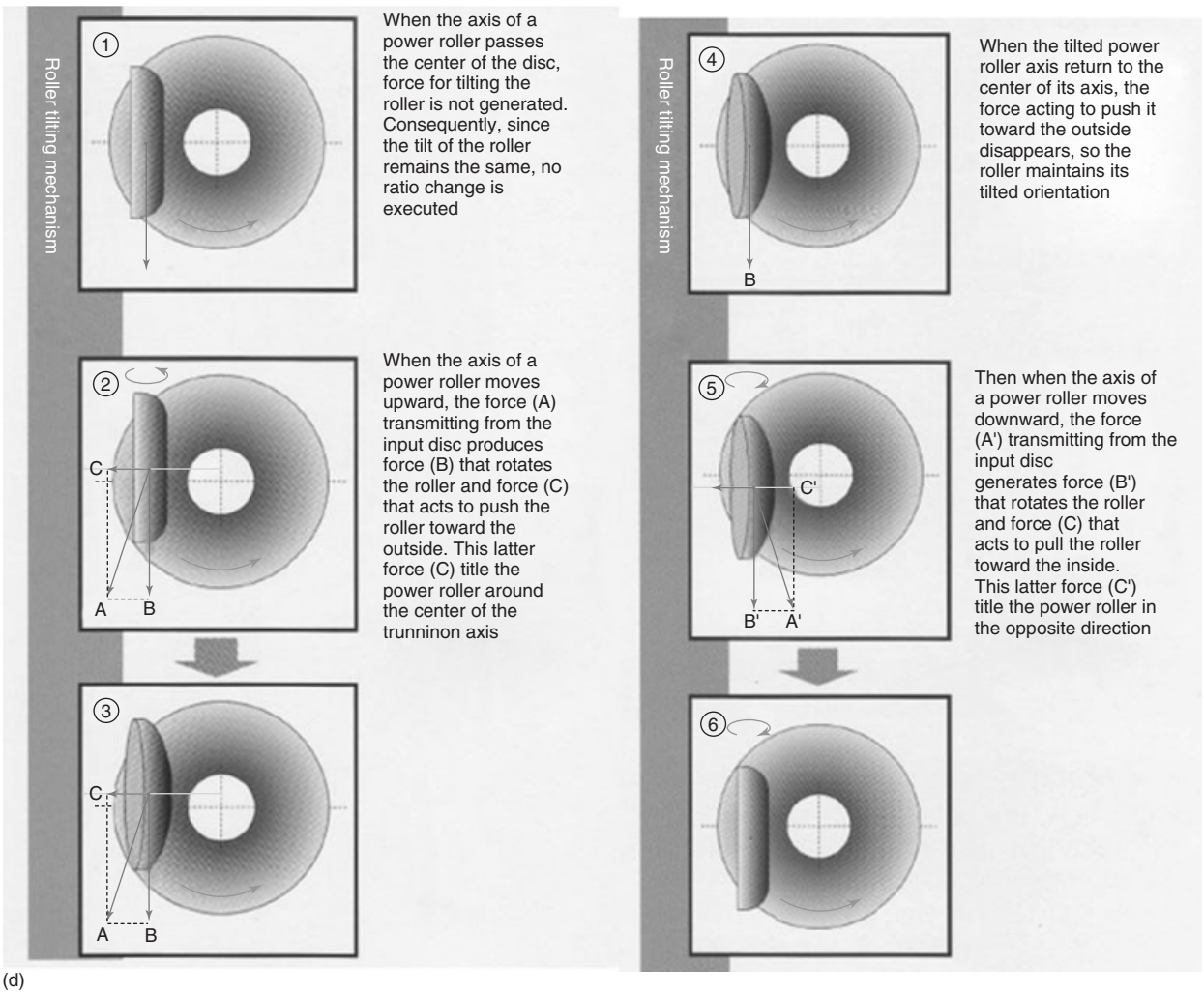
Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

mechanically. Here, as is shown in Figure 14, linkage provides assistance for keeping the four power rollers in position.

Ratio changing control of toroidal CVT is slightly different from belt CVT control. Tilting trunnion control is adopted. Figure 15a shows roller and disc contact and



(c)



(d)

Figure 15. (*Continued*)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Figure 15b shows the control system. The trunnion is moved up and down by hydraulic pressure. A hydraulic actuated piston is installed as is shown in Figure 15c and these pistons move trunnions up and down during shifting as is shown in Figure 15d. The detail control is indicated in the figure. When light movement up or down is made to the trunnion, self-rotating force in a caster action to rotate the axis of the power roller is caused at the contacting point between the power roller and the disc, because the contact point is moved slightly from the rotating axis. As the power roller makes a self-rotating behavior, ratio changing rapidness is sufficiently fast Figure 15e.

ACKNOWLWDGMENTS

The author wishes to thank Mr. Andreas Englisch and Dr. Markus Baumann from Schaeffler Group LUK Gmbh & Co.oHG, Mr. Akihito Oohashi, Mr. Yorinori Kumagai from Honda R&D Co., Ltd., and Prof. Nicolas Vaughan for their very helpful advice to improve the contents of this subscription, and also would like to thank all colleagues in JATCO and JATCO US.

REFERENCES

Abo, K., Sugano, K., Shibayama, T., and Hayasaki, K. (2003) Development of new-generation belt CVTs with high torque capacity for front-drive cars. SAE Technical Paper 2003-01-593.

Englisch, A. (2010) CVT high value and high performance. 9th Schaeffler SYMPOSIUM, pp. 139–151.

Hibi, T., Sumi, Y., and Takeuchi, T. (2002) Improvement of toroidal CVT fuel economy. Jatco Technical Review No. 3, pp. 66–76.

Kawaguchi, A., Maruyama, N., Takeuchi, T. *et al*. (2000) Introduction of JR06E toroidal CVT. Jatco Technical Review No. 1, pp. 87–92.

Nakagawa, Y. (2012) Introducing the technologies of JATCO CVT7 featuring an auxiliary transmission. Jatco Technical Review No. 11, pp. 23–30

Ota, S. and Eguchi, T. (2007) Development of CIVIC hybrid CVT system. Society of Automotive Engineers of Japan, Inc., International Congress on Continuously Variable and Hybrid Transmissions, CVT•HYBRID 2007 Yokohama, ISBN 978-904056-01-1, p. 183, JSAE Technical Paper 2007 4573.

Shimanaka, S., Okada, K., Hibi, T. *et al*. (2001) The new technology used in toroidal CVT. Jatco Technical Review No. 2, pp. 56–61.

Singh, J., Berger, K., Mack, P., *et al*. (2003) General motors 'VTi' electronic continuously variable transaxle. SAE Technical Paper 2003-01-0594.

Sugano, K., Abo, K., Hirabayashi, Y. *et al*. (2003) Development of new-generation belt CVTs with high torque capacity for front-drive cars. Jatco Technical Review No. 4, pp. 17–30.

Toyoda, T., Imanishi, T., and Inoue, E. (2010) *Fuel economy improvement items with toroidal continuously variable transmission*. CVT2010 CVT•HEV International Conference, pp. 67–72.

FURTHER READING

Jatco (2000–2012) Jatco Technical Review No. 1, 2000, No. 11, 2012.

Society of Automotive Engineers (1962a) Design Practices, Passenger Car Automatic Transmissions, AE-1 (Advances in engineering vol. 1).

Society of Automotive Engineers (1962b) Design Practices, Passenger Car Automatic Transmissions, AE-2 (Advances in engineering vol. 2).

Society of Automotive Engineers (1973) Design Practices, Passenger Car Automatic Transmissions, AE-5 (Advances in engineering vol. 5).

Society of Automotive Engineers (1994) *Design Practices—Design Practices, Passenger Car Automatic Transmissions*, 3rd, AE-18 (Advances in engineering vol. 18).

Society of Automotive Engineers (2012) *Design Practices, Passenger Car Automatic Transmissions*, 4th, AE-29 (Advances in engineering vol. 29).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto211
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Historical Overview of Electronics and Automobiles: Breakthroughs and Innovation by Electronics and Electrical Technology

Mitsuharu Kato
DENSO Corporation, Kariya, Japan

1 Introduction 1
2 Adoption: Initial Period 3
3 Microcomputers 5
4 ECU Configuration and Hardware 6
5 Semiconductor Sensors (INA, 1988) 7
6 Software Storage Program Memory 8
7 ECU–ECU Networks and Fusion of ECUS (KATO, 2010) 9
8 The Information Age (KATO, 2010) 10
9 The Future of Collaboration 11
Related Articles 12
References 12
Further Reading 13

1 INTRODUCTION

This chapter describes the relationship between automobiles and electronics. Figure 1 shows an outline of this history.

1.1 Vehicles (Casey, 2008)

The fledgling automotive technology that sprung up in Germany in the 1880s took shape as an industrial product in the form of the Model T Ford released in 1908, which went on to sell a total of 15 million units. According to Ford's internal records, the Model T was the first vehicle to be mass produced on a conveyor belt. This was state-of-the-art production line technology for the time and its basic premise is still in use today. In contrast, however, cars have completely changed. In 1908, cars were made up entirely of mechanical parts. Soon after, light bulbs began to be used in headlamps, and starter motors were introduced. Subsequently, automotive technology remained based on mechanical systems, supplemented by a small range of traditional electrical technologies, such as light bulbs, the starter, battery charger, battery, and spark plugs. At the time, the term *electronics* referred to vacuum tubes. By today's standards, the production of 15 million vehicles is an astounding feat and demonstrates the impressive evolution of mechanical engineering since the industrial revolution.

1.2 Semiconductors (Shockley, 1950)

The first practical semiconductor devices were invented in 1947, 40 years after the Model T. William Shockley was responsible for confirming the characteristics of diodes and amplification. Although already known based on solid-state physics and quantum mechanics theory, the practical verification of these characteristics was a major turning point. Soon after, p-type and n-type semiconductors were combined to form a rectifying diode, and the first transistor with an amplification function was developed. These new devices were achieved as a result of the advances in solid-state physics and quantum mechanics research that occurred in the twentieth century. In 1958, the integrated circuit (IC)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

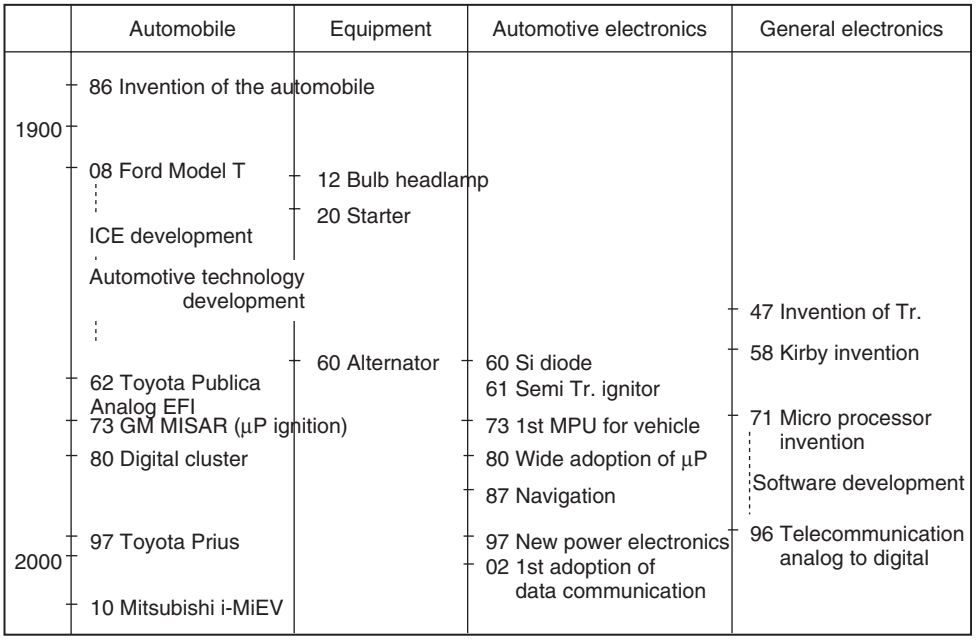


Figure 1. Automobiles and electronics (history). (Reproduced with permission from Kato, 2010 © Denso Corporation.)

that used multiple transistors was invented and patented by Jack Kilby. This invention heralded the launch of analog and digital ICs. The semiconductor device invented by Shockley in 1947 was the bipolar transistor, which used the amplification effect of a p-n junction. This was followed by the development of analog and digital ICs that used bipolar devices, which combined pnp and npn transistors. Operational amplifiers consisting of these devices in combination represented a major breakthrough as the basic components of analog ICs. In comparison to bipolar transistors, the practical application of metal-oxide-semiconductor (MOS) transistors took time due to the instability of the MOS interface. However, once these characteristics were stabilized, the simplicity of the MOS structure enabled large-scale integration (LSI) devices to enter mainstream use. Since then, the number of transistors mounted on a single IC has increased exponentially. Figure 2 shows the upward trend of device integration. As described by Moore's law, integration has roughly quadrupled every three years. This has been accomplished by the development of increasingly fine semiconductor manufacturing technologies and innovations in device structures.

1.3 Fusion of megatechnologies in the twentieth century

The vehicle is an example of a megatechnology, that is, a product in which many technologies intertwine to create a new easy-to-use technology that has the potential of helping realize the dreams of its users (Kato, 2010). As vehicles became more and more widespread, exhaust emissions created issues such as pollution, typified by the infamous smog of Los Angeles. The emissions regulations that were introduced in 1980 could not be met by conventional technology, and depended on the development of electronics, another megatechnology of the twentieth century. Figure 3 shows the principle configuration of the system developed at the time to meet these regulations.

The configuration shown in the figure is required to ensure complete combustion by optimizing the air/fuel (A/F) ratio and ignition timing, providing feedback control through O_2 sensors, and the like. This system is not entirely electronics-based. It also utilized the evolution of materials technologies such as three-way catalysts and the feedback system shown in the figure to greatly reduce emissions. It is no coincidence that the accelerated global spread of the automobile was sparked by the practical application of semiconductor technologies and the collaboration between automotive mechanical and electronics engineering to meet such regulations. This system may only represent a tiny step forward by today's standards, but it would have been unfeasible without the contribution of electronics. In this way, the negative effects on the global environment caused by the mass production and popularization of cars from 1910 were halted by the advent of electronic controls. These controls that overcame the limits of mechanical technology became feasible through the 1970s up until

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

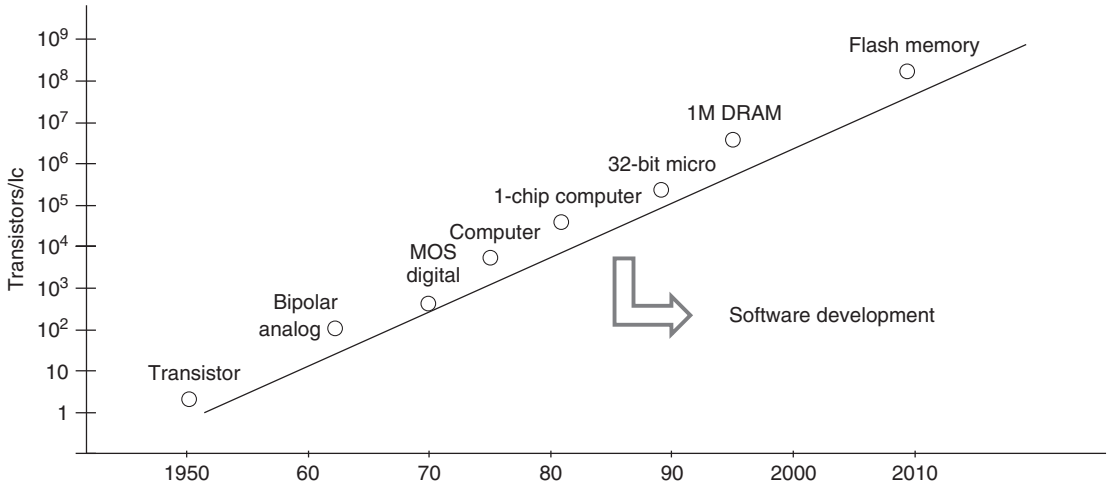


Figure 2. Integration trends.

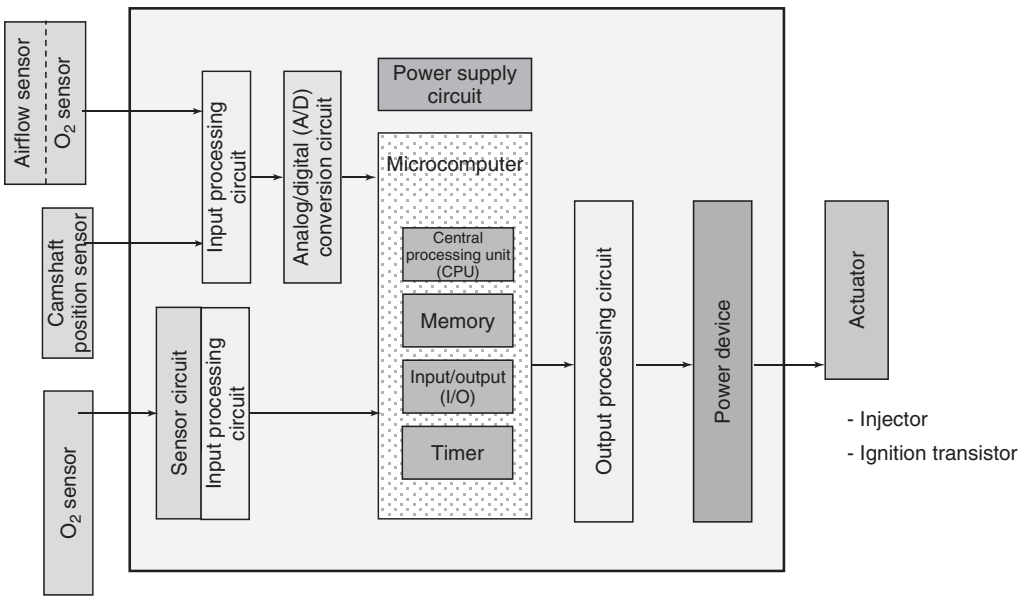


Figure 3. Basic ECU configuration.

around 1980. Modern electronics can be seen as a product of this era, in which solutions were presented to meet such regulatory requirements. This was the trigger for full-scale application of electronics to automobiles, and it inspired the development of various vehicle controls and products. In this era, the megatechnologies of automobiles and electronics finally came together.

2 ADOPTION: INITIAL PERIOD

The invention of semiconductor diodes enabled the development of solid-state rectifiers, which was a significant breakthrough because it allowed the adoption of more efficient alternating current (AC) generators in place of direct current (DC) generators. Figure 4a shows a rectifying diode. The principle of the rectifying function has not changed in the last 50 years. As shown in Figure 4b, the first transistors were used for ignition. Conventionally, a mechanical contact was used to shut off the current of the ignition coil. However, the high voltage generated by the coil would cause an arc, which limited the size of the generated voltage. Transistor ignition was capable of shutting off the current without using a contact. It enabled the generation of higher voltages and resulted in the development

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

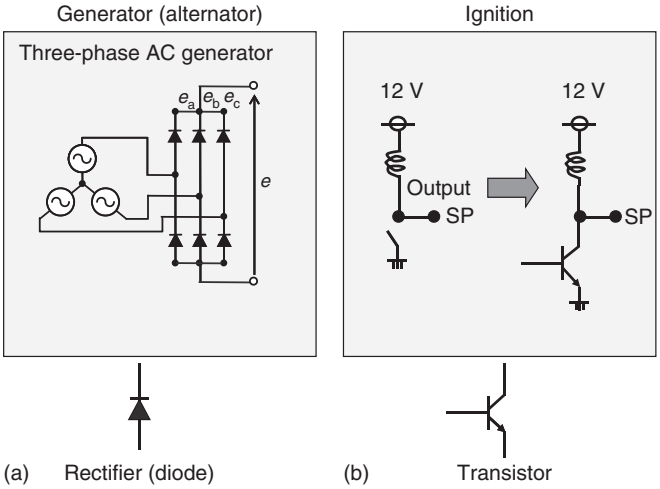


Figure 4. (a, b) Initial period of diode and transistor development (beginning of 1960s).

of gasoline engine ignition systems with high combustion efficiency.

2.1 Adoption to electronic fuel injection (EFI) (Manger, 1998)

The popularization of electronic circuits through analog circuits and analog computers, which combine operational amplifiers and make use of their functions, started in household appliances. However, when this trend began in the 1970s, such circuits were quickly adopted in vehicles as well. The first full-scale adoption of electronic circuits was in A/F ratio control systems that measured the amount of intake air in a gasoline engine with an airflow sensor and determined the optimum amount of fuel to be injected. This had been conventionally performed by mechanical systems that used the principle of atomization by a carburetor. In these systems, an amount of fuel commensurate to the airflow was mixed in the manifold and dispatched to the cylinders. In contrast, electronic control enabled the cylinder to combust the correct A/F ratio for the intake air volume. A/F ratio control was a major breakthrough because it allowed the calculation and control of engine conditions beneficial to ignition and torque output. The configuration of this system is shown above in Figure 3. Even today, the basic engine control unit (ECU) configuration of sensors, control circuits, actuator drivers, and a power source has not changed. Sensor signals include inputs such as the crankshaft, airflow, air temperature signals, and the like. This system controls parameters, such as the time that power is applied to the injectors, to determine the amount of fuel injection. Initially, this control used analog control ICs. Digital control was introduced after the invention of the microcomputer.

2.2 ECU development (Numazawa, 1998)

The development of a practical microcomputer opened the floodgates for the application of electronics to cars. Figure 5 shows the history of this application. Meter and body system microcomputers started to be adopted from around 1980. The first utilization in meter systems was in the form of digital meters that used fluorescent display tubes. Subsequently, higher performance systems started to be used as liquid crystal meters became more available. One particular use of high performance microcomputers was for

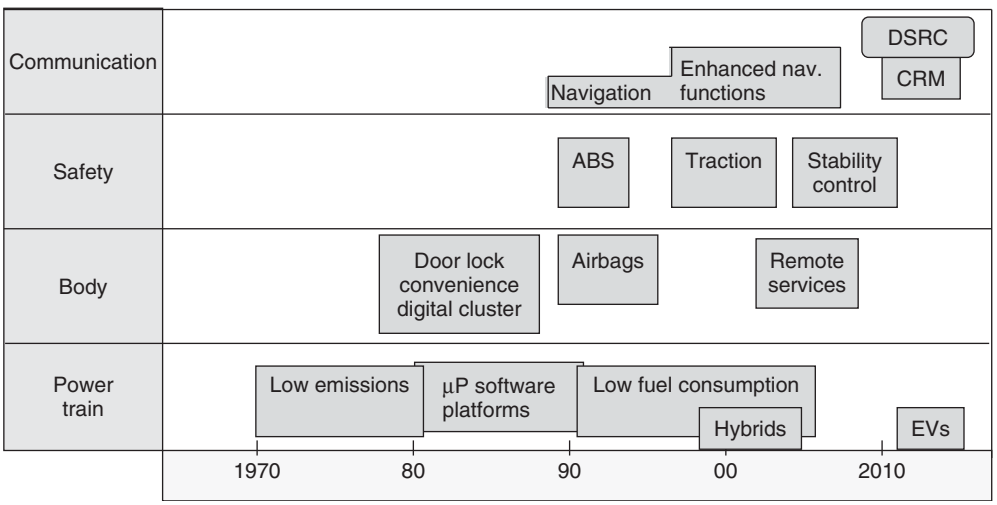


Figure 5. History of automotive electronics.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

information display. Later in the 1980s, microcomputers were adopted in safety controls. This started with seat belt interlock functions, and then spread to airbags and antilock brake systems (ABSs). The development of these systems at this time all depended on microcomputer technology.

3 MICROCOMPUTERS

Intel developed the first 4-bit microprocessor in 1971. The first automotive application was in GM's MISAR ignition control system in 1973, which used a 10-bit microprocessor. In Japan, 12-bit microcomputers were used in fuel injection systems. Although the microcomputers invented by Intel are now available in 4-, 8-, 16-, and 32-bit configurations, the first microprocessor was only available in anomalous 10- and 12-bit configurations. This was because the performance of a general purpose microcomputer was insufficient. These anomalous bit lengths were adopted to increase calculation speeds through the use of integrated instruction and data sets. Although use of these 10- and 12-bit microcomputers was limited, the same basic configuration is still present. In particular, the free running timer and compare register configuration has become a basic circuit block for real-time control. A significant breakthrough was the use of special timers to compensate for inconsistencies between the microcomputer calculation speed and the speed required for control. The use of such circuit blocks to reduce the load of the microcomputer remains unchanged today.

3.1 Microcomputer configuration

The first microcomputers required various externally attached parts, such as ROM, RAM, I/O, and the like, centered on a CPU. A watchdog timer was also required to prevent microcomputer overrun. Modern microcomputers are now realized on a single chip with in-built ROM, RAM, I/O, and program monitoring functions.

3.2 Various types of microcomputer

In addition to fields that depend on control speed, microcomputers have also been developed to meet needs for programmability and diversity. One dominant example is the 1-bit microcontroller (see the image and configuration in Figure 6). The 1-bit microcontroller was cutting-edge technology when the cost of 4- and 8-bit microcomputers was too high for use with single functions and developers were consolidating peripheral parts onto a single chip. Although the 1-bit configuration limited their coefficient of performance, these microcontrollers could be used to control event chains in single bit increments. As a result, these were ideal for monitoring and controlling door

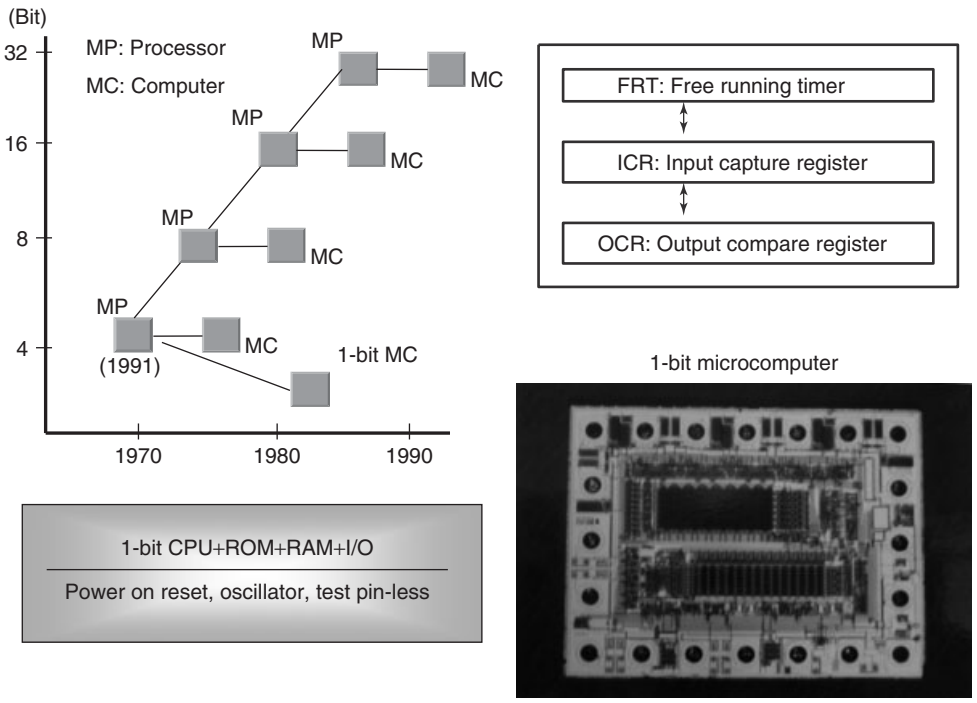


Figure 6. Microprocessor history.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5
www.it-ebooks.info

opening/closing and switch on/off functions. 1-bit microcontrollers consolidated all peripheral parts onto a single chip and have evolved into today's embedded microcomputers. Mainstream microcomputers have adopted 16- and 32-bit configurations. Innovations unique to the automotive industry such as easy-to-program architectures and free running timers mean that the automobile has created a separate field of microcomputers.

4 ECU CONFIGURATION AND HARDWARE

Although the configuration of ECU hardware has not changed from that shown in Figure 3, the shape of functional parts has changed greatly in accordance with the cyclical changes of semiconductors. Figure 7a shows that early ECUs used printed circuit boards with through holes, which were used to attach parts such as capacitors, resistors, and ICs packaged with leads. The ECU in Figure 7a uses three analog ICs. Subsequently, parts become more compact and devices were mounted directly onto the surface of the circuit board. Excluding connectors, these boards had no through holes (Figure 7b). This figure shows an ECU with a 32-bit microcomputer. The vast increase in parts mounted on the circuit board is easily noticeable. As the performance of microcomputers increased, vehicles began to use the latest microcomputer technology.

4.1 Features of on-board ECUs (Kato, 2010)

The defining feature of automotive ECUs is their harsh operating temperature range, which may reach as high as 125°C in the engine compartment and fall as low as −40°C in winter. This is vastly different from the operating temperature range of household appliances. Figure 8 lists the features of automotive ECUs. To operate in these conditions, although ECUs in the occupant compartment use conventional printed circuit boards, ECUs in the engine compartment tend to use ceramics. Actuators have been integrated onto electronic circuits from the start, and devices used with ceramic circuit boards have adopted a bare chip configuration. In particular, systems such as generators and ignition devices that have always been installed in the engine compartment adopt special packaging. One of the characteristics of automotive ECUs is the high recovery rate of defective parts. As shown in Figure 8, if a defect occurs in the market, a system has been developed that returns the defective ECUs to the manufacturer through the dealer. This enables the location of the defect to be identified and helps understand issues, particularly

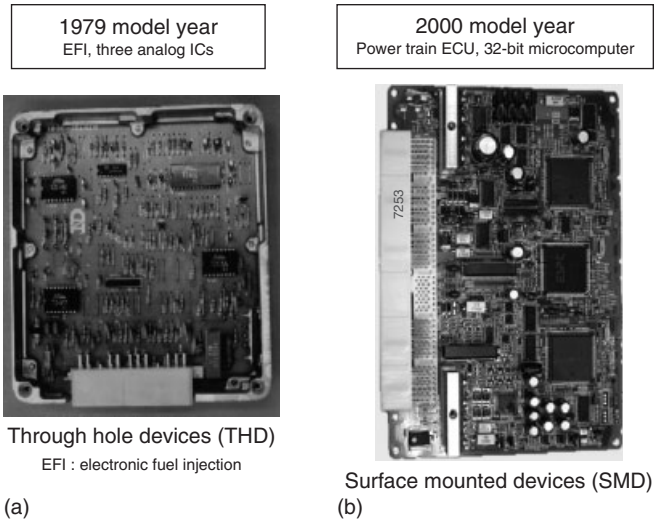


Figure 7. (a, b) Photographs of ECUs.

with semiconductors. Lessons from these steps have been reflected in improved semiconductor structures and manufacturing methods. As a result, this system has also been applied to household appliances as well as automotive semiconductors, thereby greatly helping improve the quality of the semiconductor industry as a whole.

4.2 Electromagnetic compatibility (EMC) and noise of on-board ECUs

One of the consequences of the proliferation of automotive electronics is the increase in the number of noise sources (Kato, 2010; CISPR 12; CISPR 25; ISO7637-1; ISO7637-2; ISO11451; ISO11452). As indicated in Figure 9, typical noise sources include radio frequency noise from the ignition system, surges, and electrostatic discharges. Power wiring voltages vary from 4.5 V used by 12-V systems when the starter is in operation to 17 V. Signal wiring may generate overvoltages of 200 V or more when switching parts with embedded inductance. Therefore, during ECU design, diodes and the like must be provided to protect electronic components from malfunction or breakage due to voltages transmitted through power or signal wiring. In recent years, the amount of high frequency noise emitted from ECUs has increased as microcomputer operation has speeded up. This is a potential cause of audiovisual (AV) system malfunction. Another major issue is ECU malfunction due to a reduction in signal levels caused by greater semiconductor integration. These issues are referred to as *electromagnetic interference (EMI)* and *electromagnetic compatibility (EMC)*, respectively. Measures for EMI include reducing the strength of electromagnetic

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

(1) Operating environment conditions
- Operating temperature range:
  −40°C to +85°C or +125°C
- Humidity: 60% to 90% RH
- Vibration resistance

(2) Traceability
- First-in first-out system on ECU production line
- Lot control of ECU production line
- Lot control of parts
- Supply chain control of parts
- Recovery system for defective ECUs
- Analysis system for defective ECUs
- Feedback to improve part quality

(3) 100% nondefective parts and the building of quality with responsibility
- Guarantee of 100% defective-free parts
- Adoption of systems to build quality with responsibility in ECU production (i.e., that prevent defective parts from moving to the next production process)

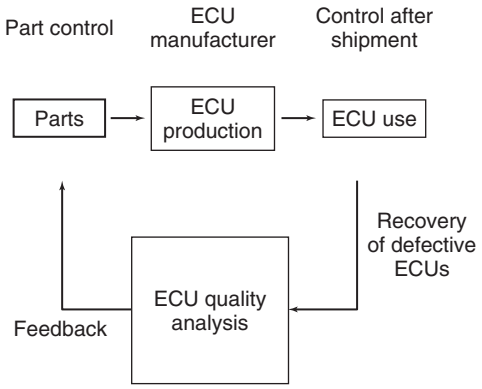


Figure 8. Features of automotive electronics.

Type		Source	Effects
Ignition system radio noise		• Spark plugs • Distributor	• Obstruction of communication with general devices (regulatory compliance) • Electronic devices: Malfunction
Surge	Low frequency	• Alternator • Ignition coil • Relays • Solenoids, motors	• Electronic devices: Malfunction • Electronic devices: Breakage
	High frequency	• Relay contacts • Motors	
Electrostatic discharge		• Between occupants and vehicle • Static electricity from workers on assembly line	• Electronic devices: Malfunction • Electronic devices: Breakage

Figure 9. Sources and effects of ECU electrical and radio noise.

wave emissions as a part of electronic circuit design. EMC is a potential case of electronic circuit malfunction, and robust design measures are required to prevent electromagnetic interference. Various international test standards have been established with respect to these issues. Figure 10 shows some relevant examples. EMI and EMC will play an increasingly significant role when improving the performance of electronic products in the future.

5 SEMICONDUCTOR SENSORS (INA, 1988)

Semiconductors have another critical role in electronic systems as sensors. A glance at early electronic fuel injection (EFI) control shows that the airflow sensor played a key part. Rudimentary airflow sensors measure the angle of a damper to gauge the airflow. These evolved into sensors that measured the Karman vortex and then into the indirect measurement of intake air by monitoring the manifold pressure. The use of pressure sensors has greatly reduced the size of these systems. A dedicated semiconductor pressure sensor was developed to measure the manifold pressure, which signaled the start of the groundbreaking technique of semiconductor micromachining. The semiconductor sensor shown in Figure 11 uses the Piezo resistance effect. This provided the stimulus for the development of various sensors using different semiconductor properties (Figure 12). Semiconductors are now used in pressure, acceleration (*G*), photo, and magnetic sensors. Silicon plays a key role as the base material for micromachining. Angular acceleration sensors have

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Evaluation items			International standards	Use scenarios
Immunity	Static electricity		ISO 10605	On contact
	RF immunity		ISO 11452-2, -3, -4	Operation in strong electrical fields
Emissions	Transient characteristics	Load dump	ISO 7637-2	Positive voltage surge when battery terminals are disconnected
		Field decay	ISO 7637-2	Negative voltage from inductive load of alternator
	Radiant emissions		ICISPR 25	Other devices in the occupant compartment (radio waves to AM/FM radio, TV)
	Conductive emissions		ICISPR 25	Other devices in the occupant compartment (radio waves to AM/FM radio, TV)

Figure 10. EMI and EMC evaluations for ECUs.

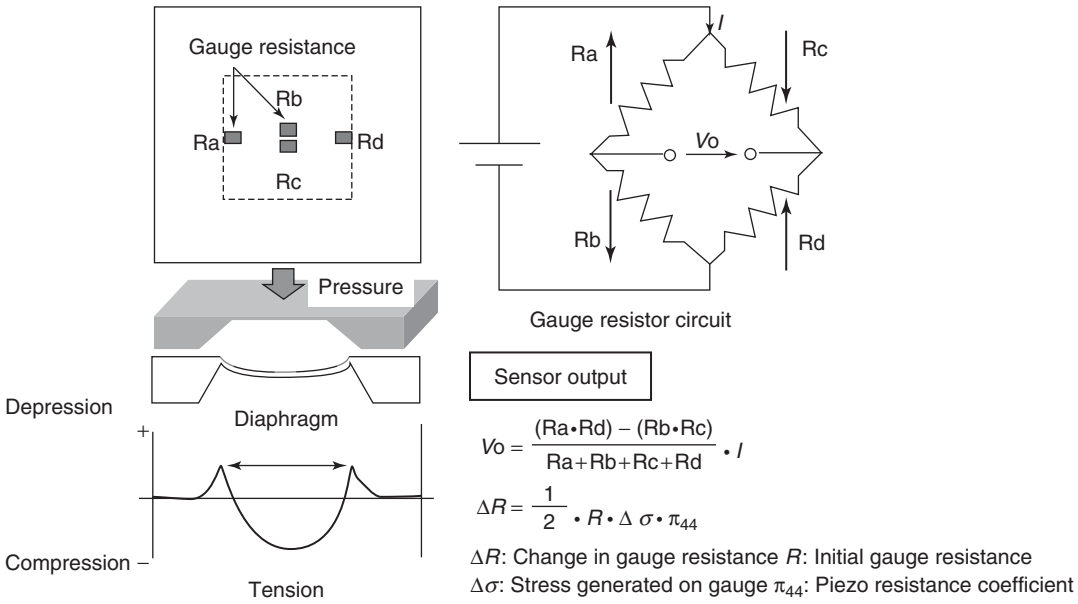


Figure 11. Pressure sensor. (Reproduced with permission from Kato, 2010 © Denso Corporation.)

also been practically adopted. Magnetic sensor devices that mount thin film magnetic resonance (MR) devices on an IC are a crucial part of proximity sensors that identify minute vector changes in magnetic fields. The feasibility of control systems depends on the enhancement of sensor performance as well as the microcomputer processing capability. For example, engine control systems are now capable of fine controls using linear sensors that monitor the oxygen concentration in emissions. Vehicle attitude control uses *G* and yaw rate sensors. The detection and control of states using sensors involves calculation by a microcomputer, and feedback of the results to actuator outputs.

6 SOFTWARE STORAGE PROGRAM MEMORY

The configuration of program memory has switched from mask ROM to EPROM, EEPROM, and electronically rewriteable flash memory. The usability of program memory has changed substantially. Writeable technology for program memory using masks required the details of the program to be determined a few weeks before the IC was packaged. Therefore, technology was developed to push the program writing step further back in the semiconductor manufacturing process to minimize the program creation lead time. The program still required four to six weeks,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

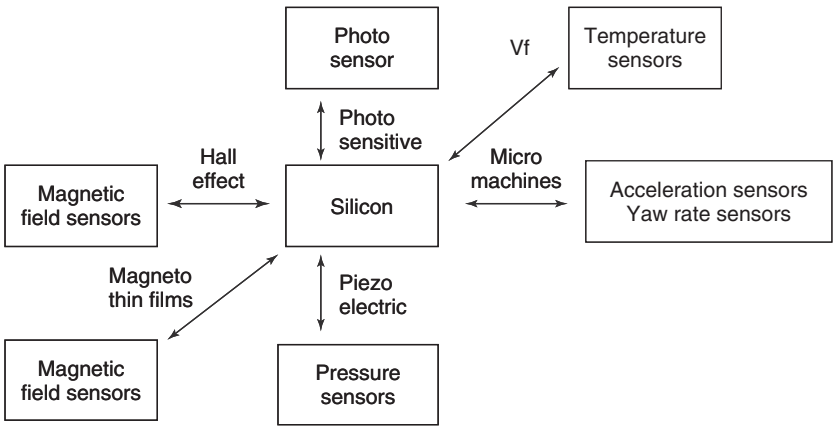


Figure 12. Semiconductor sensors.

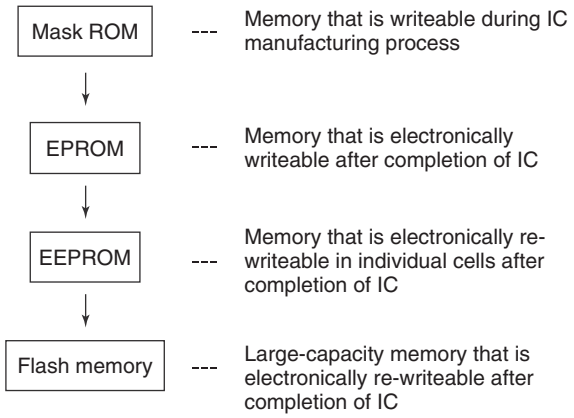


Figure 13. Program memory.

which became a major issue once software control became the mainstream control method. For this reason, the appearance of EPROM technology was a major breakthrough as it allowed the contents of the ROM to be electronically written. Electronic programming could then be set as the final step of ECU manufacture. Semiconductor technology continued to evolve and perfect the basic technologies for writing programs into memory electronically. Flash memory is a good example of this. It can be used to update ECU programs through a service connector while the ECU is mounted in the vehicle. This has helped further accelerate the pace of vehicle control development (Figure 13).

6.1 Importance of software technology

High performance ECUs that use microcomputers have given rise to new system controls, which have in turn spurred the development and introduction of even higher performance microcomputers. The scale of software has increased as a result of more sophisticated microcomputers and cheaper memory, making system controls more complicated. Consequently, the quality of software has become more important. Software program languages have evolved from assembly language to advanced languages such as C and C++. Programs now consist of between several tens of thousand to several hundreds of thousand lines of code. Advanced languages are a prerequisite for software creation productivity. Software developers are facing a growing burden, which means that software architecture is playing a more central role and a critical issue has become how to improve reusability rates. A capability to build architectures with individual software modules that have independent characteristics is the most important current issue for improving both productivity and quality. From a systems standpoint, detailed studies of control specifications are more important than ever, and software developers are driven by the need to build fault-tolerant systems (Figure 14).

7 ECU–ECU NETWORKS AND FUSION OF ECUS (KATO, 2010)

The number of wire harnesses (W/H) has increased in accordance with the spread of electronic systems. As a result, W/H weight and complexity is on the increase. This increase in W/H is the direct consequence of the importance of linking ECUs. Multiplex communication has also become more important as the number of W/H has exceeded 1500. Historically, ECU communication began with the controller area network (CAN) used by engine systems and spread with the adoption of J-1850 body control systems. More recently, low, medium, and high speed CAN systems have been adopted, and FlexRay has

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

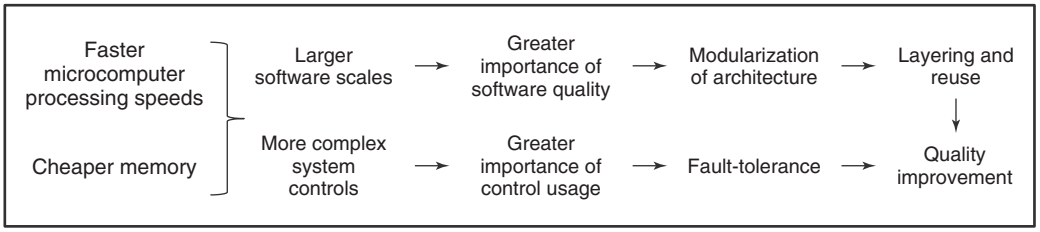


Figure 14. Software trends.

emerged as a feasible direction for sensor systems. Standard multimedia systems include Media Oriented Systems Transport (MOST). Although such systems are capable of eliminating several hundred W/H, increasing system complexity is causing the number of W/H to increase. ECU fusion is one basic method of suppressing system complexity and W/H increase. This aims to consolidate similar ECUs, thereby reducing their number as well as the connecting W/H. ECUs can be categorized into five main types: powertrain systems, chassis control systems, body systems, entertainment systems, and systems related to external information. However, the best way to consolidate each ECU type remains the key issue. Possible bases for combining ECUs include standard functions, differences in optional functions, and differences in development cycles. Furthermore, the growing popularity of hybrid vehicles is driving even greater ECU complexity by making control systems more complicated, further diversifying the information that must be communicated between ECUs, increasing the driving assistance functions that can be performed by connecting with external networks, and the like. The development of fault-tolerant systems is becoming even more crucial.

8 THE INFORMATION AGE (KATO, 2010)

In the 1990s, the application of electronic systems became further oriented toward information. Although the first navigation system was actually developed in 1987, the full-scale adoption of navigation systems began at the end of the 1990s. This was due to the spread of accurate position detection using GPS technology, the development of digital maps on compact discs (CDs), and the improvement in microcomputer performance. Figure 15 shows these trends. In addition, display technology was enhanced by the development of the liquid crystal display (LCD). As a result, it became possible to show other information rather than just maps on a multipurpose screen, such as information captured by rearview cameras.

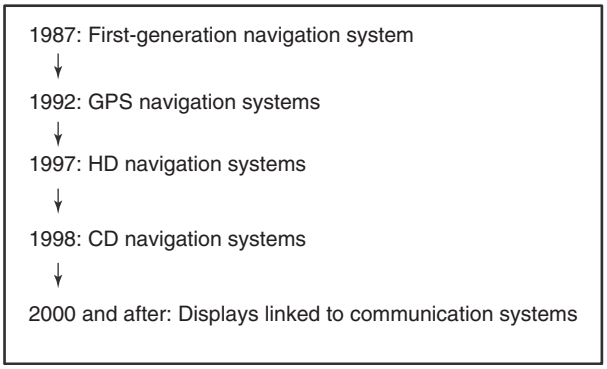


Figure 15. Information device trends.

The importance of multipurpose displays has increased further. The digitalization of communication has significantly speeded up rates of information transmission. As a result, communication technologies have become an indispensable aspect of vehicles in the 2000s. In addition to destination information as typified by navigation systems, emergency notification systems such as OnStar from GM and eCall from Mercedes have also been developed. Toyota's G-Book, Nissan's CARWINGS, and Honda's Internavi are other specific examples of enhanced communication systems. These developments highlight the growing significance of information communication in twenty-first century. Further progress will see the permanent connection of vehicles to communication systems and the adoption of more innovative communication applications.

8.1 The age of ITS

Improving the efficiency of communication, information, and traffic flow is a key part of modern vehicle development. Helping alleviate the congestion and negative environmental impacts caused by the popularization of the automobile and reduce the number of traffic accidents require comprehensive solutions. In the future, it is possible that automated vehicles will work as an ideal solution. Major short-term issues include reducing accidents by

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

alleviating driver burden and congestion avoidance. The navigation system plays an indispensable role in addressing these points. The identification of the vehicle's current location by GPS and vehicle speed sensors, and detailed maps that include traffic information can help guide the driver to target destinations, minimize congestion, and reduce energy consumption. Navigation software exceeds half-a-million lines of code, the most of all automotive systems. This field is likely to become even more important in the future to help improve usability and save energy, as well as to provide detailed road information and safety assistance.

8.2 Expansion of automotive communication technologies

Communication technologies will play an even more important role in the future. Radio waves, which were originally used for AM and FM radio systems, have also been adopted for remote entry systems and tire pressure monitoring systems (TPMSs), as well as for external communication functions such as the Vehicle Information and Communication System (VICS), electronic toll collection (ETC), and mobile telephones. Figure 16 shows the history and status of radio wave usage in Japan. Effective radio wave technologies have already evolved from AM modulation, FM modulation, time division multiple access (TDMA), and code division multiple access (CDMA) protocols. In the future, the long-term evolution (LTE) protocol is expected to be adopted. These advances have been made feasible by the greater integration and processing speed of semiconductor devices. In addition to helping boost the performance of microcomputers, the evolution of semiconductor technology has also contributed greatly to the advancement of communication technology, that is, the effective use of finite frequency resources. In the future, it is expected that 1-GHz broadband technology will become even more sophisticated and spread to vehicles by improving communication quality.

9 THE FUTURE OF COLLABORATION

Until now, control system ECUs and electronics have been able to overcome the limitations of mechanical technology. In turn, this has encouraged the evolution of mechanical systems, a synergistic effect that has accelerated the evolution of automotive controls. Figure 17 illustrates these trends. Taking ABS as an example, microcomputers spurred the evolution of hydraulic systems. Compact and high performance oil pumps were then integrated with electronic circuits, which have been further integrated with actuators. These systems have diverged from household applications where devices are mounted on printed circuit boards to form new derived technology. Packaging technology has advanced to the stage where devices can be mounted on ceramic boards with excellent heat resistance. Although this is only a single example, the growing complexity of controls is likely to require the development of even more

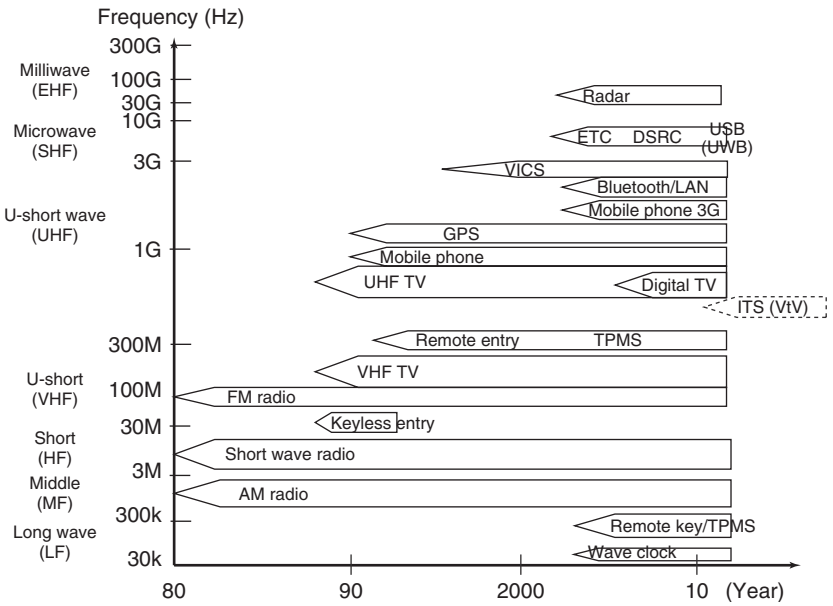


Figure 16. Wave frequencies and automotive applications. (Reproduced with permission from Kato, 2010 © Denso Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

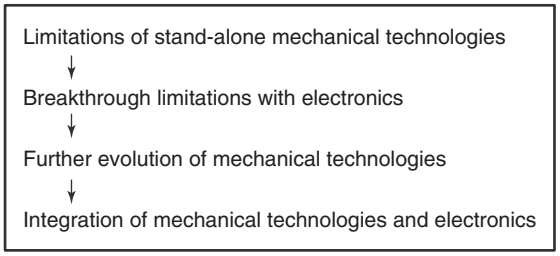


Figure 17. Collaboration trends (1).

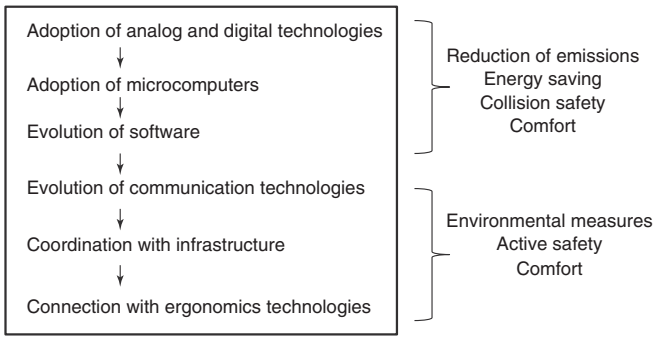


Figure 18. Collaboration trends (2).

unique technology. One possibility is the electronic correction and reduction of mechanical tolerances that exist in individual torque converters to improve the efficiency of automatic transmissions. It should be possible in the future to electronically adjust individual mechanical tolerances. This will require the integration of mechanical products and electronic circuits, which is dependent on the development of new packaging methods. Furthermore, although technology applications started with analog and digital technology, microcomputers have taken center stage and larger scale software has evolved. It is likely that wireless communication technology will become more advanced and find greater applications in the fields of environmentally friendly technology, safety, comfort, and the like. At the same time, this is likely to trigger further discussion about communication reliability. The introduction of CO_2 regulations as vehicles become even more widespread, safety concerns, and growing demands for active safety systems will usher in an age where image processing and recognition technologies, and technologies that incorporate aspects of ergonomics are as important as simple control systems. Figure 18 illustrates these likely trends. Thus, the role of electronics, software, and algorithms will continue to increase, placing even greater emphasis on collaboration between automotive and electronics technologies.

RELATED ARTICLES

Micro, Mild and Full Hybrid
Mechanics of Contacting Surfaces
ECU Technologies from Components to ECU Configuration
Diversification of electronics and electrical systems and the technologies for the integrated systems
ECU Design and Reliability
Microcomputers and Related Technologies: Enlargement of Software Size, Algorithms, Architectures, Hierarchy Design, Functional Decomposition, and Standardization
In-Vehicle Network
Semiconductor Sensors (3): Optical Sensors
Various Types of Sensors
Semiconductor Sensor (1): Sensors for Power Trains
Ceramics sensors for power train ECU systems
Semiconductor Sensors (2): Sensors for Safety Systems
Engine ECU Systems
Chassis ECU (Vehicle dynamics, ABS)
Generators and Charging Control
Body ECU (airbag)
Body ECU Cluster
Car Navigation
Telecommunications

REFERENCES

Casey, R. (2008) *The Model T, A Centennial History*, The Johns Hopkins University Press, Baltimore.

CISPR 12: *Vehicles, boats, and internal combustion engine driven devices – radio disturbance characteristics – Limits and methods of measurement*.

CISPR 25: *Radio disturbance characteristics for the protection of receivers used on board vehicles, boats, and on devices – Limits and methods of measurement*.

Ina, O. (1998) Recent Intelligent Sensor Technology in Japan. In Automotive Electronics: The 1980s - A Collection of Landmark Technical Papers. SAE International, 831–839.

ISO7637-1: *Road vehicles – Electrical disturbances from conduction and coupling – Part 1: Definitions and general considerations*.

ISO7637-2: *Road vehicles – Electrical disturbances from conduction and coupling – Part 2: Electrical transient conduction along supply lines only*.

ISO11451: *Road vehicles – Vehicle test methods for electrical disturbances from narrowband radiated electromagnetic energy*.

ISO11452: *Road vehicles – Component test methods for electrical disturbances from narrowband radiated electromagnetic energy*.

Kato, M. (2010) *Automotive Electronics Illustrated*, Nikkei Business Publications, Inc, Tokyo.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Manger, H. (1998) Electronics Engine Controls in Europe. In Automotive Electronics: The 1980s - A Collection of Landmark Technical Papers. SAE International, 465–469.

Numazawa, A. (1998) Automotive Electronics in Passenger Cars. In Automotive Electronics: The 1980s - A Collection of Landmark Technical Papers. SAE International, 789–802.

Shockley, W. (1950) *Electrons and Holes in Semiconductors, With Applications to Transistor Electronics*, D. Van Nostrand, New York.

Scholl, H. (1998) Electronics Applications to the Automobile by Robert Bosch. In Automotive Electronics: The Early Years - A Collection of Landmark Technical Papers: The Early Years. SAE International, 521–537.

Shah, P. (1998) Programmable Memory Trends in the Automotive Industry. In Automotive Electronics: The 1990s and Beyond - A Collection of Landmark Technical Papers. SAE International, 93–104.

FURTHER READING

Kitano, T. (1998) The Status of Automotive Electronics in Japan. In Automotive Electronics: The Early Years - A Collection of Landmark Technical Papers. SAE International, 377–395.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto212
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ECU Technologies from Components to ECU Configuration

Yukihide Niimi
DENSO Corporation, Kariya, Japan

1 Introduction 1
2 ECU Outline 1
3 ECU Installation Locations and Operation Conditions 2
4 ECU Examples 5
5 ECU Configuration and Component Parts 6
6 ECU and Microcomputer Trends 13
7 The Future of ECU Technologies 15
Related Articles 15
References 15

1 INTRODUCTION

Since the 1960s, when the first radios and stereos equipped with transistors began to be installed in cars, electronics were seen very much as optional equipment. The full-scale introduction of electronics for vehicle control began in 1970 with the proposal of the Clean Air Extension Act (popularly referred to as the *Muskie Act* after its main sponsor, the United States Senator Edmund Muskie) to regulate harmful emissions (NO*x*, HC, and CO) from the engine. To comply with this Act, there was a trend for conventional carburetor fuel supply systems to be replaced with electronic fuel injection systems, which used injectors to control the amount of fuel injected into the engine intake system. This fuel system was controlled by an electronic control unit (ECU).

As electronic fuel injection systems became the mainstream means of supplying fuel, engine control systems began to evolve steadily. Initially, control simply judged the cleanliness of emissions to determine the optimum fuel injection amount. Once ECUs were introduced, automakers realized that various controls could be performed efficiently. Electronic circuits evolved from analog to digital (microcomputer control), and various controls were added to the engine ECU. These included ignition timing control, knock control, idling speed control, transmission control, and so on. At the same time, electronic control also spread quickly to functions such as the air conditioner, ABS (anti-lock braking system), and airbags.

Currently, controls are categorized in accordance with the field, such as powertrain, air conditioning, body, and dynamic and safety controls. There are many different control system variations within these categories. This chapter describes the internal structure and parts used in ECUs, using the engine ECU as a typical example.

2 ECU OUTLINE

Most car electronics contain an electronic control computer called an *ECU* (Kato, 2010a). This ECU functions to process inputs from sensors and the like to drive motors or other kinds of actuators. ECUs communicate with each other using the in-vehicle local area network (LAN). ECUs contain sensor and switch input processing circuits, analog–digital (AD) conversion circuits, microcomputers, a power supply, output processing circuits, and a communication circuit for interacting with other ECUs. The main components are semiconductors. Figure 1 shows a typical

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

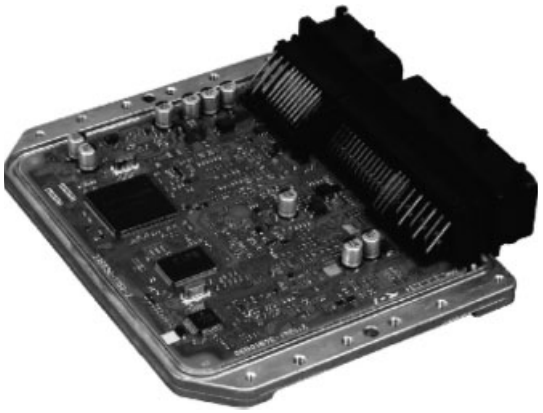


Figure 1. Typical ECU.

printed circuit board type ECU and Figure 2 shows the general functional blocks.

Each block in the ECU has the following roles. The power supply provides a stable voltage to each block (5, 3 V, and the like). It is connected to the battery (12 V) in the engine compartment. The power supply block must be highly accurate as it also uses the reference voltage of the AD conversion circuit (see below).

The input processing circuit converts digital input signals into signal levels that can be inputted into the microcomputers.

The AD conversion circuit (also known as the *AD converter*) converts analog input signals into digital values that can be inputted into the microcomputers. The microcomputers calculate the control amounts from each input signal and control the output signals in accordance with the results of these calculations. The output processing circuit converts the output signals of the microcomputers into signal formats capable of driving the actuators, amplifies voltages, and so on. The communication circuit contains a communication driver that converts data output from the microcomputers into communication signals that comply with communication standards. It also includes a communication receiver that converts signals transmitted from other ECUs into signal levels that can be input to the microcomputers.

3 ECU INSTALLATION LOCATIONS AND OPERATION CONDITIONS

3.1 ECU installation locations

The simple term *ECU* covers a wide range of ECU types that differ depending on the installation location. The structure of each ECU is designed in accordance with its installation environment (Kato, 2010b).

Electronic control systems comprise three types of devices: (i) an ECU that governs the system control, (ii)

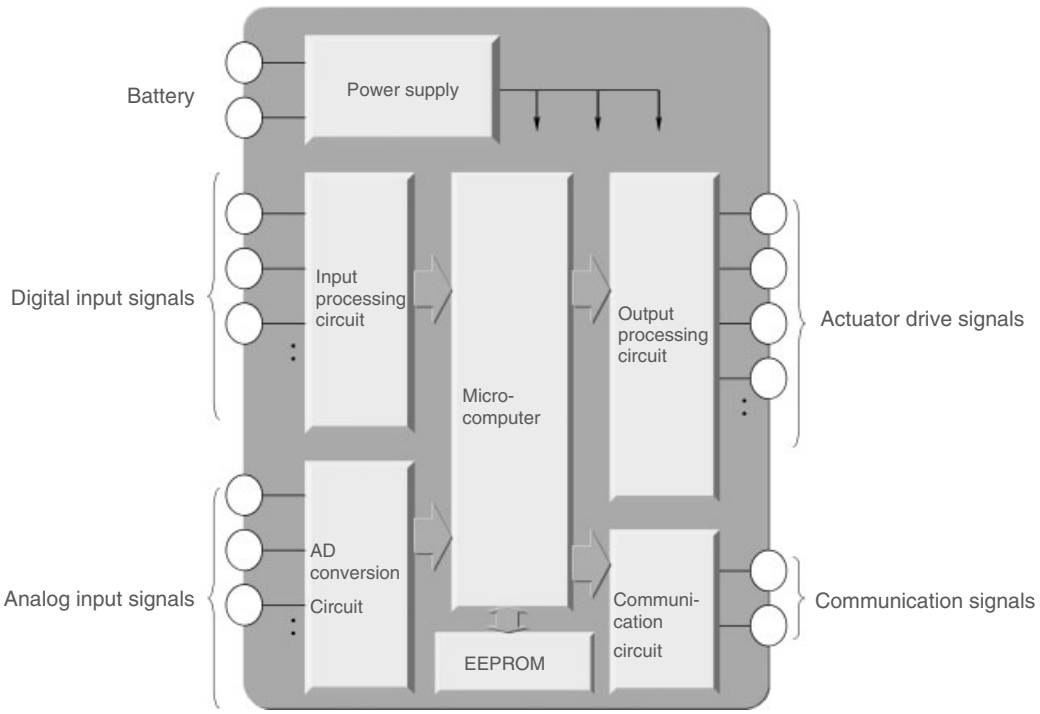


Figure 2. ECU functional block diagram.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

sensor groups that input the states and information relevant to the control into the ECU, and (iii) actuator groups that ultimately achieve the targeted control. Of these device types, the sensors and actuators have to be installed in certain locations to achieve the function. In contrast, the installation locations of ECUs have changed over the course of history. When the electrification of vehicles started, electronic parts were particularly vulnerable to heat, humidity, and vibration. For this reason, ECUs were installed inside the occupant compartment. Subsequently, as the reliability of circuit board-based electronic parts and packaging technology for electronic parts improved, it became possible to install ECUs inside the engine compartment, despite its severe heat, humidity, and vibration, and even directly attach ECUs to the engine itself. By making the appropriate considerations at the design stage, it is now possible to install ECU hardware anywhere in the vehicle. This is known as the *free-installation concept*.

However, ECUs, sensors, and actuators have to be connected together electrically using wire harnesses. It is clearly preferable to ensure that these wire harness connections are as short as possible. Short wiring lengths have the following merits.

- lower material costs, enabling cheaper wire harnesses;
- less copper used, enabling vehicle mass reduction;
- lower electrical resistance, enabling thinner wires to be used;
- less susceptibility to the effects of induced noise caused by wiring impedance, which is generated depending on the frequency of the signal. Shorter wires also emit less noise.

For these reasons, the transmission ECU is installed as close to the transmission as possible. In some vehicles, it may actually be installed inside the transmission itself. Furthermore, a growing number of electronic parts related to engine control are installed directly on the engine. Ignition induction coils are installed on the plug hole to minimize ignition noise by making the high voltage wiring as short as possible. Impedance is a restriction for the electric driver unit (EDU) that is used to drive the injectors in a direct injection gasoline engine, because of the large current flow and considering operation speed. Accordingly, these units are installed close to the engine in the engine compartment with as short a wiring length as possible.

The installation location is often determined from a functional standpoint. For example, acceleration (G) sensors are located in multiple locations throughout the vehicle in the airbag system. As the airbag inflator has to deploy multiple airbags, it is located close to the center of the vehicle. In addition, the navigation system display has to be in a position that is easily visible, which restricts its installation location. Recently, as electronic circuits become more compact, motor control circuits are becoming extremely small. These compact controllers are now being installed and integrated into the actual motors, further driving reductions in size. Car electronics are installed in locations that facilitate the target function. Installation methods are categorized in accordance with the installation location (Figure 3). Each installation

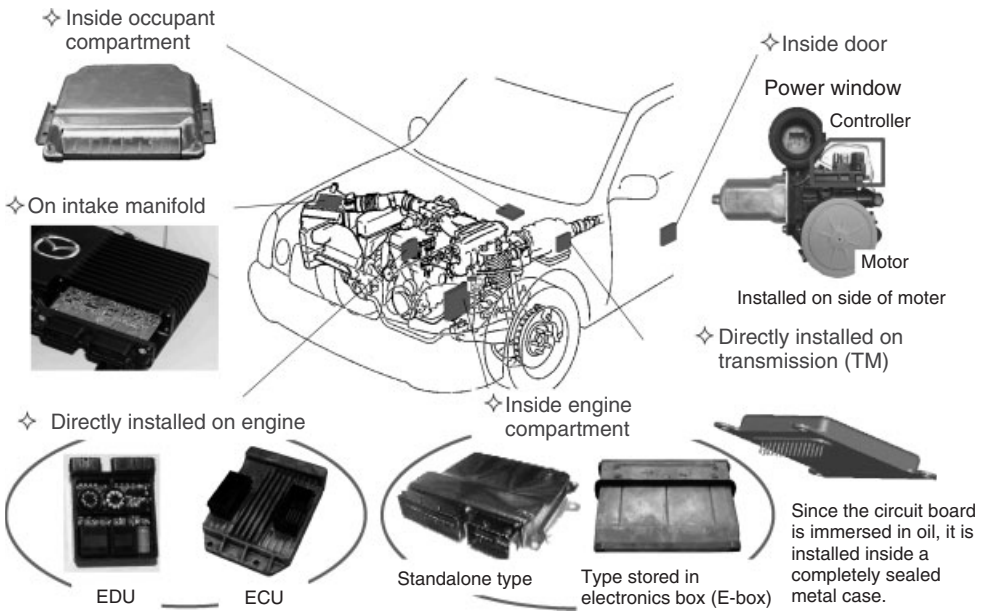


Figure 3. ECU installation location and method.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Typical maximum humidity exposure of automotive parts

Location	Maximum humidity
Engine compartment (close to engine)	38°C 95% RH
Seats	66°C 80% RH
Close to side doors	38°C 95% RH
Close to dash board	38°C 95% RH
Floor sheet	66°C 80% RH
Rear deck	38°C 95% RH
Luggage compartment	38°C 95% RH

Typical maximum temperature exposure of parts in engine compartment

Location	Maximum temperature (°C)
Engine cooler	120
Engine oil	120
Transmission oil	150
Intake manifold	120
Exhaust manifold	650
Alternator intake air	130

Typical maximum temperature exposure of parts in occupant compartment

Location	Maximum temperature (°C)
Top of dash board	120
Bottom of dash board	71
Floor	105
Rear deck	117
Headlining	83

Figure 4. Installation environments inside vehicle.

location has a different environment and ECUs are designed considering that environment. Figure 4 shows the typical environmental conditions of each installation location in a vehicle.

3.2 ECU operation conditions

Table 1 shows the different categories of operation conditions.

ECUs related to basic vehicle performance (the engine and dynamic ECUs) and ECUs related to occupant safety (the airbag ECU) operate under harsh conditions. In contrast, the operation conditions for the navigation system ECU are not set at the same level. This is because enabling an ECU to operate under harsher conditions than necessary generates excess cost and because the electronic circuits used in complex controls such as the navigation system are not yet able to operate in the harshest environments.

3.2.1 Operation conditions of parts used in ECU

Each part has different operating conditions (Kato, 2010b). Table 2 shows the temperature conditions for an electrolytic capacitor.

As the operation environment of automotive ECUs is harsher than those for home electronics, electrolytic capacitors are required to operate under more severe conditions. That difference is reflected in the higher purchase price.

Table 1. ECU operation conditions.

Application	Installation Location	ECU Type	Voltage (V)	Temperature (°C)	Vibration (G)
Engine	Engine compartment	Engine ECU	6–16	−30 to +95	4.4
Dynamics	Occupant compartment	ABS ECU	6–16	−30 to +80	4.4
Safety	Occupant compartment	Airbag ECU	6–16	−30 to +80	4.4
Body	Occupant compartment	Navigation system ECU	8–16	−30 to +80	4.4

Table 2. Operation conditions of electrolytic capacitor.

For Automotive Use		For Home Electronics
Engine ECU	Navigation System ECU	
−40 to 125°C	−40 to 105°C	−40 to 85°C

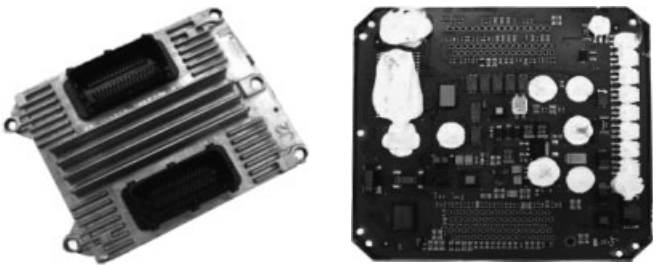
Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4 ECU EXAMPLES

4.1 Examples of printed circuit board-based ECUs

4.1.1 Engine ECU

The engine ECU is a standalone package and has only one internal control circuit board. As the number of control functions increased, some engine ECUs were developed with multiple boards. However, these functions have come to be collected together as integrated circuit (IC) devices, and modern engine ECUs have returned to a one-board configuration. Figure 5 shows an example of this type of ECU. It has a waterproof structure and its outside shape is designed to dissipate heat generated from the inside.



(a) (b)

The ECU case is provided with fins to improve heat dissipation. The white plastic portions in Figure (b) are made from a material called a thermally conductive gel, which is designed to improve heat transmission. These design points improve heat dissipation from the ECU and allow installation on the engine

Figure 5. Example of engine ECU. (a) Outside appearance of ECU. (b) Circuit board inside ECU.

4.1.2 Navigation system ECU

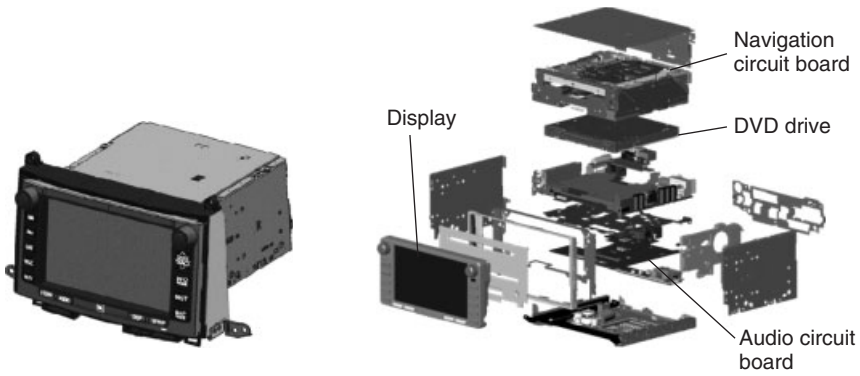
The navigation system ECU arranges components such as a display, DVD drive, and hard disc drive (HDD) in a compact package on the circuit board. Figure 6 shows a basic 2DIN type navigation system. Although it contains a built-in audio system and multiple circuit boards, the navigation system functions are collected on the same board. The navigation system uses a microprocessor with high functionality. To ensure efficient heat dissipation, a metal case is used that is punched with holes for ventilation.

4.2 Hybrid ECU

Hybrid ECUs are compatible with high functionality and large currents. In this type of ECU, bare chips are packaged on multilayered laminated ceramic circuit boards with high thermal conductivity. Figure 7 shows the example of a diesel engine hybrid ECU (Kato, 2010b). It contains a microcomputer, hybrid ICs, and power devices. Hybrid ECUs are usually adopted to meet requirements for size and weight reduction, and environmental resistance.

As shown in Figure 8, a hybrid ECU is structured with a control portion, power portion, large parts, and a connector. It uses multilayered laminated ceramic circuit control boards with high thermal conductivity, a bare chip type of chip scale package (CSP), conductive adhesive connections for the power ICs and chip parts, gold (Au) wire bonding, and silicon gel sealing.

Figure 9 shows an electric drive unit (EDU). A typical example of EDU application is as an injector drive controller. EDUs use a thick multilayered laminated ceramic circuit board. The ICs are bare packaged to help reduce the size of the circuit board. The largest parts on the



(a) (b)

Component parts such as the DVD drive and display, and control boards such as the navigation audio system boards are collected into a 2DIN space

Figure 6. Example of navigation system. (a) Outside appearance of ECU. (b) Internal configuration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

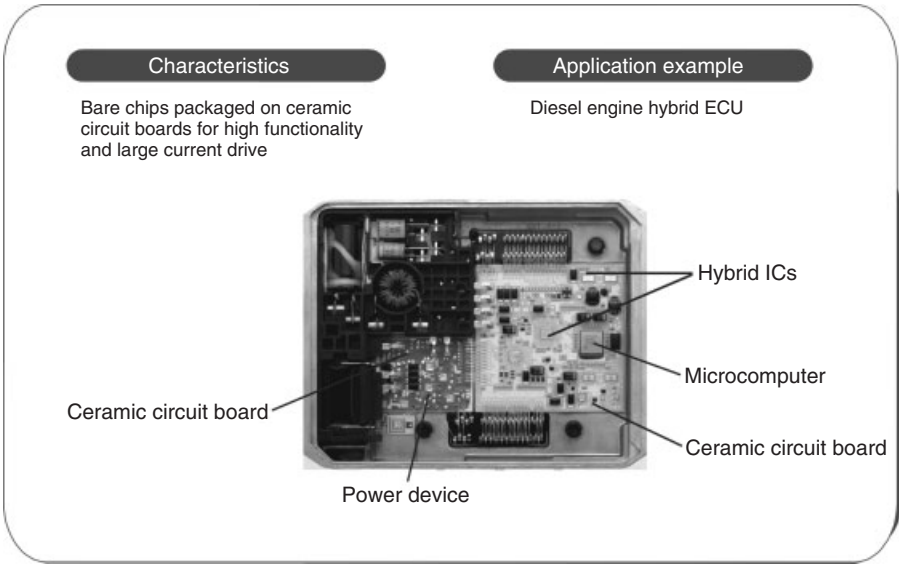


Figure 7. Characteristics and application example of hybrid ECU.

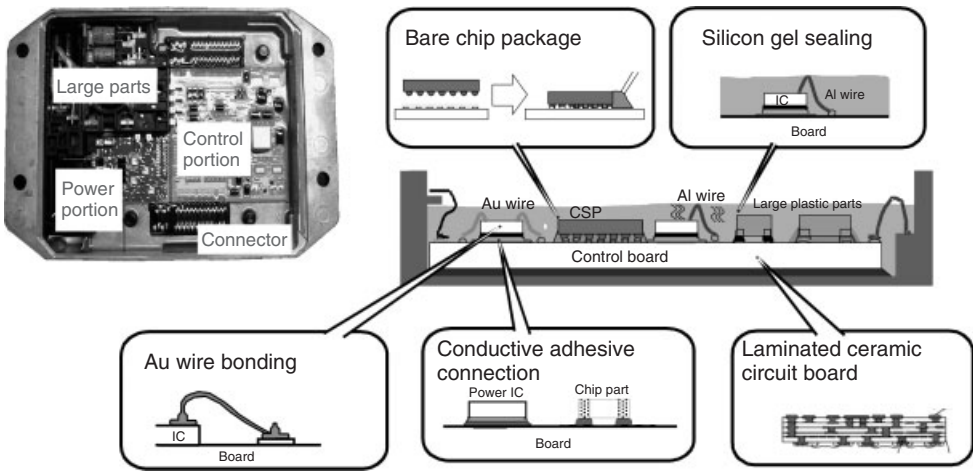


Figure 8. Hybrid ECU structure.

circuit board are the capacitors, which store accumulated energy for driving devices, and coils. EDUs are designed to be resistant against heat, vibration, and water. The bare chips are sealed in silicon gel to protect the bare packaged chips against the external environment.

5 ECU CONFIGURATION AND COMPONENT PARTS

5.1 ECU configuration

In broad terms, an ECU consists of a circuit board, a connector, and a case that protects these two portions from the external environment (Figure 10). Active, passive, and functional parts are arranged on the circuit board. Figure 11 shows the configuration of these parts. ECUs are increasingly being installed inside the engine compartment and on the engine itself, as well as in the occupant compartment. Figure 12 shows some typical examples of each.

5.2 Case

ECUs located inside the occupant compartment are open because water exposure is not considered likely. In contrast, ECUs installed inside the engine compartment or on the engine are designed with a waterproof case. The cases of ECUs attached to the engine are also designed to dissipate heat considering the harsh temperature environment. Most

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

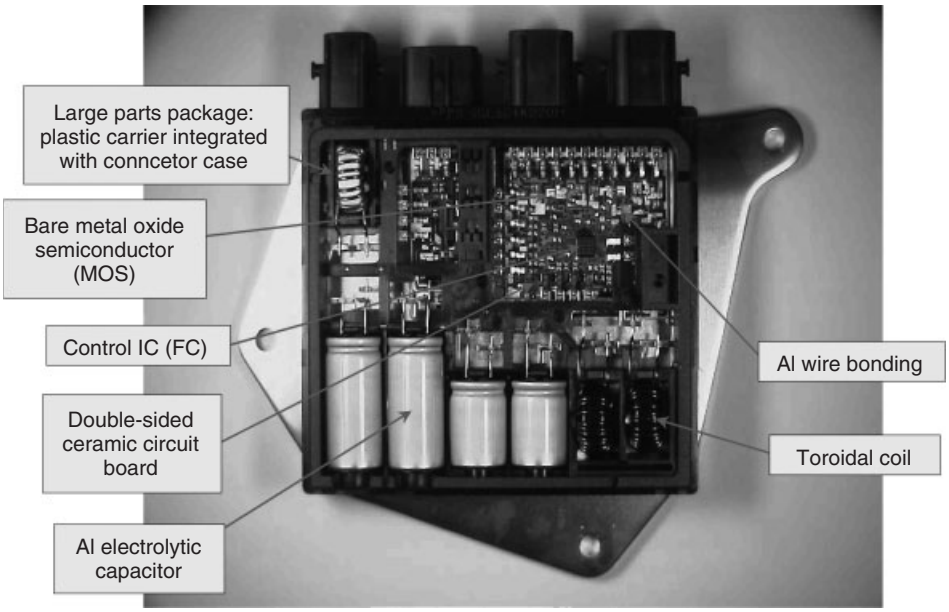


Figure 9. Electric driver unit.

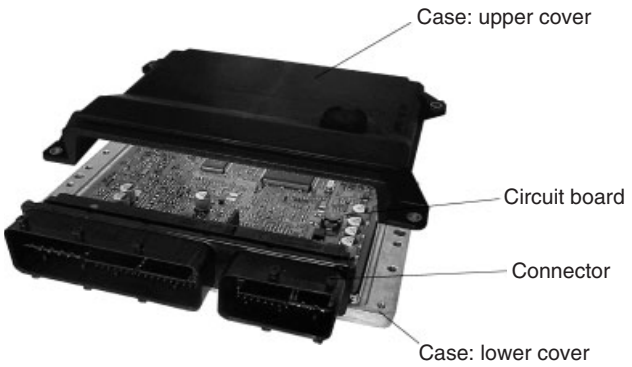


Figure 10. Configuration of typical ECU (waterproof type).

cases are configured in two sections with an upper and a lower part.

5.2.1 Non-waterproof case

The primary role of a non-waterproof case is to protect the circuits from the impact of external shocks. Non-waterproof cases that are not required to dissipate heat are often made of plastic. When heat dissipation (natural cooling) is required, metal cases are used, including Al die-cast, Al-stamped, and Fe-stamped cases.

5.2.2 Waterproof cases

Circuit assemblies in which the circuit board and connector are soldered together must be protected from water using upper and lower cases that have been designed accurately in accordance with the dimensional requirements. Accuracy is particularly required for the joining case surfaces, and shape around the connector, as these areas are closely related to waterproof performance. A waterproof sealing material is used to fill the gap between the cases. This material is generally a silicon gel with properties that do not change under temperatures as high as 150°C.

5.2.3 Waterproof type high heat dissipating cases

Cases installed in the engine compartment must be designed with resistance against heat, vibration, and water exposure. In addition to the waterproof design described earlier, the mechanical properties of the case material must be considered to withstand the vibration applied to the ECU from the engine. Most cases in engine compartment are also shaped to dissipate heat. This shape must be determined considering the surrounding airflow and recent developments use simulations to improve the efficiency of the shape design process.

5.3 Connector

Unlike conventional electrical products, which only have to perform a single function, ECUs are used for control based on signals measured by sensors with respect to a control target. ECUs identify the state of the control target, calculate the final control value with this built-in microcomputer, and transmit signals to drive actuators. Consequently, ECUs require a large number of signal

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

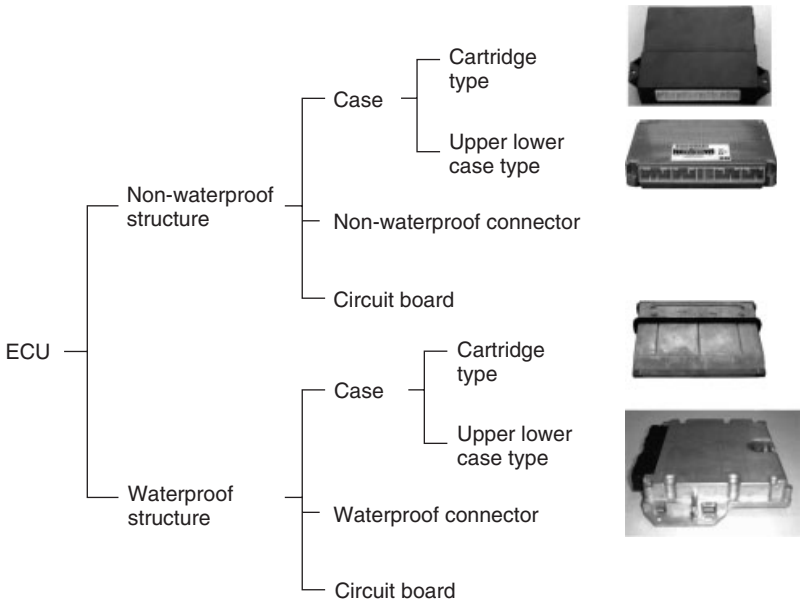


Figure 11. ECU component parts.

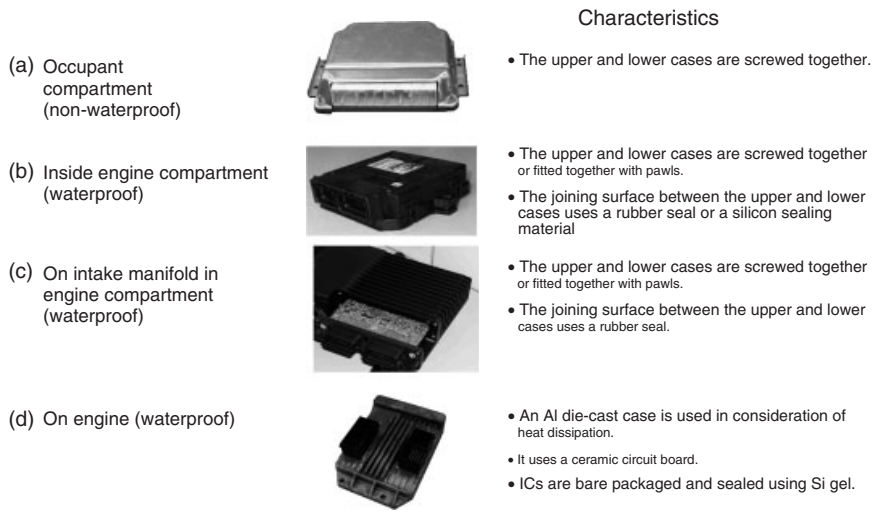


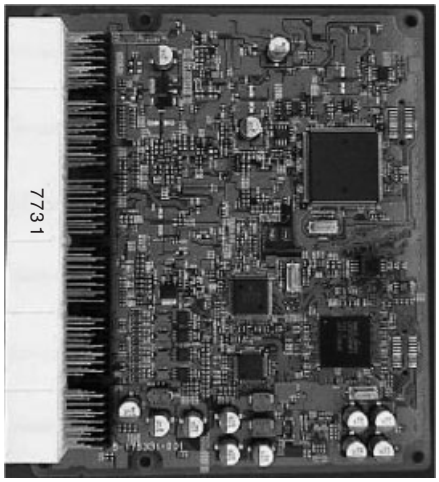
Figure 12. Appearance of ECU according to installation location.

terminals, such as sensor input terminals, actuator drive output terminals, power supply terminals, and terminals for communicating with other ECUs.

The part that collects these terminals into a single group is called the *connector* (Figure 13). The number of terminals in the connector is increasing as ECU functions become more sophisticated. Although advances in compact packaging technology have reduced the size of the circuit board, there has been no progress in making the connector smaller because of restrictions related to the diameter of the externally connected wires (i.e., the wire harnesses).

Waterproof and non-waterproof connectors are available. Waterproof connectors tend to be larger than non-waterproof connectors. Furthermore, non-waterproof connectors are divided into blocks of around 20 terminals. This is related to the maximum number of wires that can pass through a hole in the wall separating the engine and the occupant compartments (about 200). In contrast, waterproof connectors are designed with a split terminal layout, which is divided between wire harness groups used in the engine compartment and wire harnesses that transmit signals within the occupant compartment. In this

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Non-waterproof type connector

Figure 13. Connector circuit board inside ECU.

case, a 200-terminal connector often has two blocks of about 140 terminals for transmitting signals to the engine compartment and about 60 terminals for the occupant compartment. Figure 14 shows the housing (circuit board side) and plug (vehicle wire harness side) of a waterproof and a non-waterproof connector.

5.4 Circuit board

Figure 15 shows the two types of circuit board used in an ECU, which are fabricated from epoxy resin and ceramic, respectively. Circuit parts are installed on the circuit board and the packaging format of these parts differs greatly between the circuit board types. More specifically, IC parts are bare packaged on ceramic circuit boards because this type of board is more expensive.

The type of the circuit board also affects the forms of the packaged parts. The use of through-holes to package parts has given way to the use of surface mounted devices (SMDs). This is because electronic parts are becoming smaller because of constant demands to install more functions within a smaller case. The merits of SMD can be outlined as follows.

- Compared to insertion, eliminating the through-hole enables more efficient use of the circuit board surface, increasing the available area for packaging. Therefore, a circuit board with the same number of parts can be made smaller.
- The soldered portion of the through-hole can be eliminated, reducing the size of the soldering land pattern and increasing the density of the wiring pattern.
- Eliminating the through-hole simplifies the circuit board fabrication process, which helps to reduce costs.

5.4.1 Epoxy resin printed circuit boards

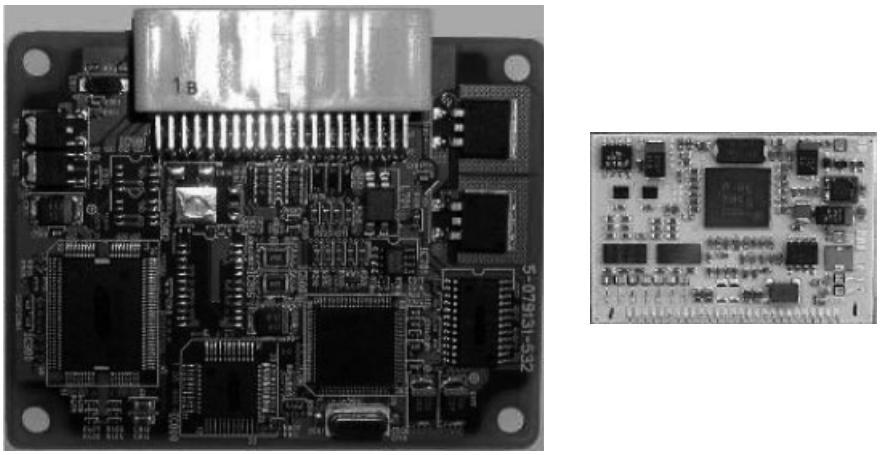
Automotive applications widely use laminated circuit boards (Japan Institute of Electronics Packaging, 2006) based on glass epoxy resin. This type of circuit board is widely categorized into through-hole circuit boards (double-sided, 4-layer, and 6-layer), blind via hole (BVH) circuit boards (a type of through-hole circuit board), and build-up circuit boards that have a build layer on both sides and a conventional layered structure in the center (Figure 16). Build-up circuit boards that have a single build layer on the sides and a core 4-layer center structure are called *1-4-1 build-up circuit boards*. Build-up circuit boards are used in the navigation system and other systems that require high functionality with dimensional restrictions, as well as parts where small size is important. 1-4-1 build-up circuit boards have also begun to be adopted in engine



(a) (b)

Figure 14. (a) Non-waterproof and (b) waterproof connectors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) (b)

The photo shows the same circuit packaged onto an epoxy resin and a ceramic circuit board. Ceramic circuit boards are considerably smaller because the ICs are bare packaged

Figure 15. (a) Epoxy resin and (b) ceramic circuit boards.

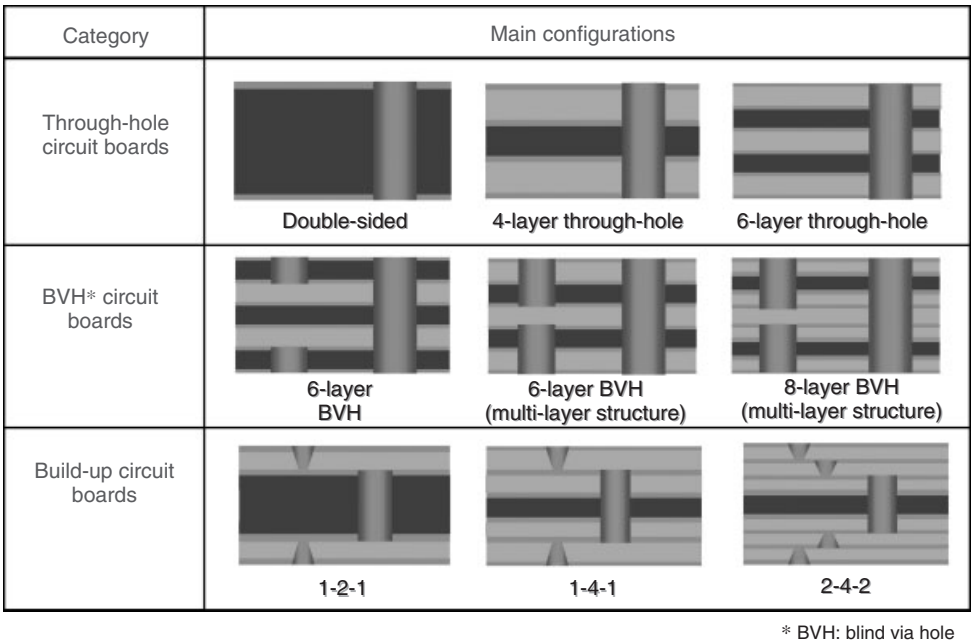


Figure 16. Categories and layer configurations of epoxy resin circuit boards.

ECUs installed in the engine compartment. Through-hole circuit boards are particularly widely used in ECUs that have severe restrictions on size.

5.4.2 Ceramic circuit boards

Types of ceramic circuit boards include single-sided, double-sided, thick multilayered laminated, and sheet-laminated. The type used depends on the application. Single-sided to thick multilayered laminated ceramic circuit boards are widely used in control circuit boards for driving actuators. These control circuit boards also include built-in power devices for driving the actuators, with an emphasis on heat dissipation performance and size reduction. Sheet-laminated circuit boards are used in ECUs that operate under harsh temperatures and vibrations, such as those directly installed in the engine. These circuit boards include bare packaged ICs such as microcomputers and power

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

devices. The configuration of sheet-laminated circuit boards is useful in achieving high wiring densities and size and weight reduction.

5.5 Categories of circuit parts

The parts (Japan Institute of Electronics Packaging, 2006) installed on a circuit board can be categorized in terms of function and shape. From the standpoint of function, these parts can be divided into general electronic and functional parts (Table 3). General electronic parts can be further categorized into passive parts, functional parts, connection parts, and conversion parts. Functional parts include individual semiconductors, ICs, and hybrid ICs. From the standpoint of shape, categories include inserted, surface mounted, and bare chip parts (Figure 17).

5.5.1 Inserted parts

These parts include axial lead parts, radial lead parts, irregular lead parts, and various types of packages, such as single inline packages (SIPs), dual inline packages (DIPs), and pin grid arrays (PGAs).

5.5.1.1 Axial lead parts. These parts have a transverse shape with two lead wires in the direction of the device. Lead wires are generally made of copper with a round cross section and use solder plating. These parts are inserted onto the circuit board and the lead legs are bent into an "L" shape. Typical axial lead parts include carbon film resistors, cylindrical ceramic capacitors, diodes, and so on.

5.5.1.2 Radial lead parts. These parts have a vertical shape with two or three lead wires protruding in parallel. Radial lead parts may be formed by processing the leads of axial lead part. As radial lead parts are created to match the pitch of the insertion holes in the printed circuit board, there is no need to bend the lead legs after insertion. In addition, the vertical shape of these parts means that a narrow interval can be set between leads, allowing denser packaging than with axial lead parts. More radial lead part types are available than axial lead parts, and these include Al electrolytic capacitors, transistors, ceramic filters, crystal oscillators, and the like.

5.5.1.3 Irregular lead parts. These are parts that cannot be categorized as either axial or radial lead parts. The shape, layout, and number of wires can be freely set, which means that there are various types of irregular lead parts. Typical examples include connectors, switches, variable resistors, and the like. Transistors and other devices that have an amplification action include ICs, power devices, and so on.

5.5.2 Surface mounted parts

These can be categorized into general electronic parts and active parts. General electronic parts can be further categorized by shape into flat, cylindrical, and irregular types. Flat parts are currently the mainstream type of surface mounted parts. Active parts can be categorized according to package type. These include mini-mold packages, small outline packages (SOPs), quad flat packages (QFPs), ball grid arrays (BGAs), and CSPs.

Table 3. Circuit part function categories.

Part Categories		Functions	Applications
General electronic parts	Passive parts	Controls voltage and current without changing the input signal characteristics	Resistors, capacitors, etc.
	Functional parts	Changes the input signal characteristics such as the frequency and time axis	Crystal oscillators, LC filters, etc.
	Connection parts	Mutually connects and switches parts and circuit devices	Switches, connectors, etc.
	Conversion parts	Converts input signals into different forms of energy	Speakers, sensors, etc.
Active parts	Individual semiconductors	Has active functions such as input signal amplification control, and memory. Standalone devices	Transistors, diodes, LEDs, etc.
	ICs	Collects and integrates multiple semiconductors.	Analog ICs, microcomputers, etc.
	Hybrid ICs	Collects active, passive, and film devices on a printed circuit board, and has active functions	Thick film hybrid ICs, thin film hybrid ICs, etc.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

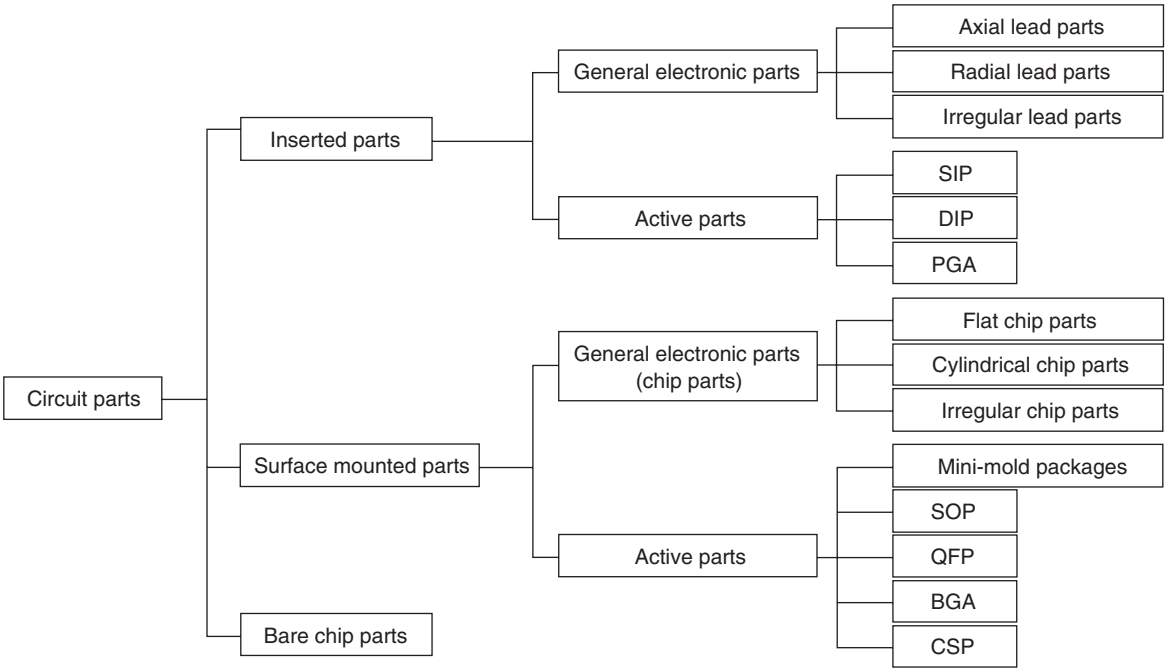


Figure 17. Circuit part shape categories.

5.5.2.1 Chip parts. Surface mounted general electronic parts are called *chip parts*. Flat chip parts are rectangular. These parts generally have no leads and have portions referred to as *terminals* that are located on the bottom or side surface of the solid shape. Fixed resistors, ceramic capacitors, laminated inductors, and the like use printed or laminated ceramics materials, which are formed to match the shape of the chip. These are formed into the flat shape by molding or enclosing the rectangular chip in a case. Typical flat chip parts include tantalum electrolytic capacitors, molded wound inductors, and the like.

Unlike flat chip parts, cylindrical chip parts have no directionality and can be used effectively in a package. However, the cylindrical shape is a disadvantage when packaging on a surface with various other parts. For this reason, there are not many parts with this shape. Irregular chip parts are those parts that are neither flat nor cylindrical. Typical examples include Al electrolytic capacitors, inductors, and the like.

5.5.3 Surface mounting semiconductor packages

5.5.3.1 Mini-molded packages. These packages enclose individual semiconductors such as transistors and diodes. The lead wires protrude to the outside of the package.

5.5.3.2 Small outline packages. SOPs are compact DIPs, with the leads processed for surface mounting. The lead pitch for automotive applications is between 1.27 and 0.5 mm, but consumer applications may have a lead pitch of 0.4 or 0.3 mm.

5.5.3.3 Quad flat packages. QFPs have lead wires that protrude in four directions from the package. QFPs are designed for high density packages with more leads than SOPs. The lead pitch for automotive applications is between 0.8 and 0.5 mm, but consumer applications may have a lead pitch of 0.4 or 0.3 mm.

5.5.3.4 Ball grid arrays and chip scale packages. Providing connection terminals for more than 500 pins on a QFP results in a large package. A countermeasure for this is to provide terminals on the back surface, opposite to the package front surface where the semiconductor devices are installed. The name BGA is derived from the round solder balls of these terminals. CSPs have the same structure as BGAs, but have a smaller package of virtually the same size as the semiconductor devices. CSPs attached with solder balls in the rewiring semiconductor fabrication process performed are sold under the names wafer process packages (WPPs) or wafer level packages (WLPs).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6 ECU AND MICROCOMPUTER TRENDS

6.1 ECU evolution

Figure 18 illustrates how ECUs have evolved. Figure 18a (Kato, 2010a) shows an electronic fuel injection device ECU with an analog circuit, which was used before the practical application of microcomputers. This is an analog ECU formed by discrete devices and used in 1973. This evolved into ECUs comprised of analog ICs, such as that shown in Figure 18b. This ECU is from the DIP era of ICs. Figure 18c is a digital engine ECU configured around a microcomputer. The custom microcomputers used in 1978 adopted software that was programmed using a 12-bit assembly language. The program used read-only memory (ROM), which was built in as a wafer during the microcomputer fabrication process. The photo shows the most complex engine ECU installed in 1983 models. Figure 18d is a modern powertrain ECU with two 32-bit microcomputers. Excluding the connectors, all the ECU parts used in 2006 were mounted on the surface. The software is written in C language. It uses flash memory and can read information from external terminals. As this shows, in addition to the scale of ECUs, the component parts, package shapes, and software languages have all evolved with the times.

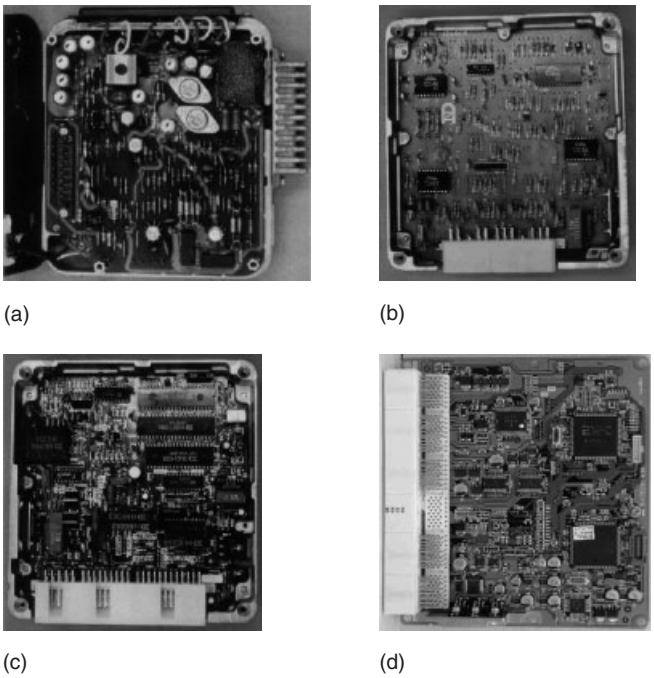


Figure 18. ECU evolution. (a) Analog ECU using discrete devices (1973 model). (b) Analog ECU using custom ICs (1979 model). (c) Digital 12-bit microcomputer (1983 model). (d) Digital 32-bit microcomputer (2006 model).

6.2 Microcomputer evolution

6.2.1 Function

This section describes the basic configuration of a microcomputer used in an ECU (Figure 19).

As the configuration of automotive microcomputers is no different from general-purpose types, a detailed description of general microcomputer characteristics is omitted in this section.

ROM has been adopted in recent years for regulatory compliance, and memory in modern microcomputers is configured to allow changes (i.e., flash memory) in block units within the ROM.

The timer controller performs time and time interval control. These devices have a compare function that changes an existing output when the set interval matches the internal timer value, and a capture function that memorizes the input time interval of external signal edges. The compare function is used to determine injection signal output timings and the like. The capture function is used to measure the engine speed signal input timing in combination with the interruption controller.

The communication interface is used to interact with input/output expansion ICs inside the ECU and with other ECUs. A growing number of microcomputers have an in-built CAN (controller area network) function for compatibility with the in-vehicle LAN.

6.2.2 Microcomputer evolution

The ROM capacity of microcomputers has rapidly increased to comply with new and more stringent regulations. This trend is likely to continue in the future, necessitating further miniaturization and increases in speed. Microcomputer types have also changed from 8- to 16-, to 32-bit configurations to meet requirements for greater processing speeds. As the pace of development has speeded up, the time to market introduction has grown shorter. In accordance with this trend, the type of ROM has switched from mask ROM to one-time programming (OTP) to flash ROM. Figure 20 shows an example of an engine control microcomputer. The details are described elsewhere.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

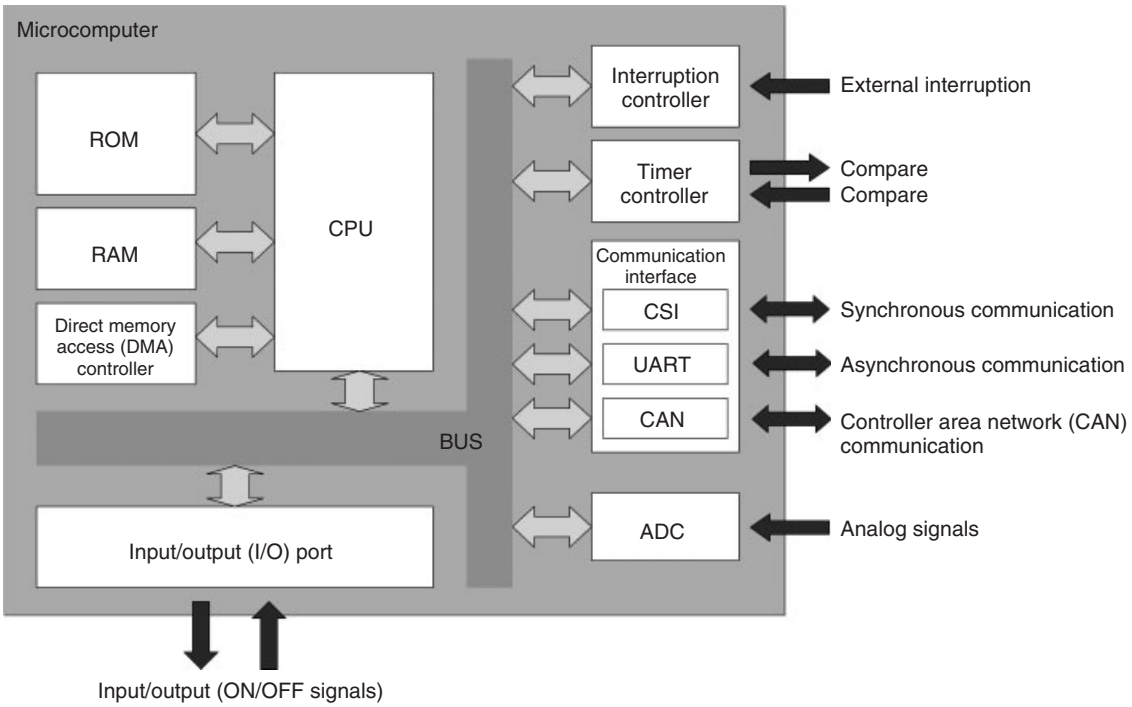


Figure 19. Internal microcomputer configuration.

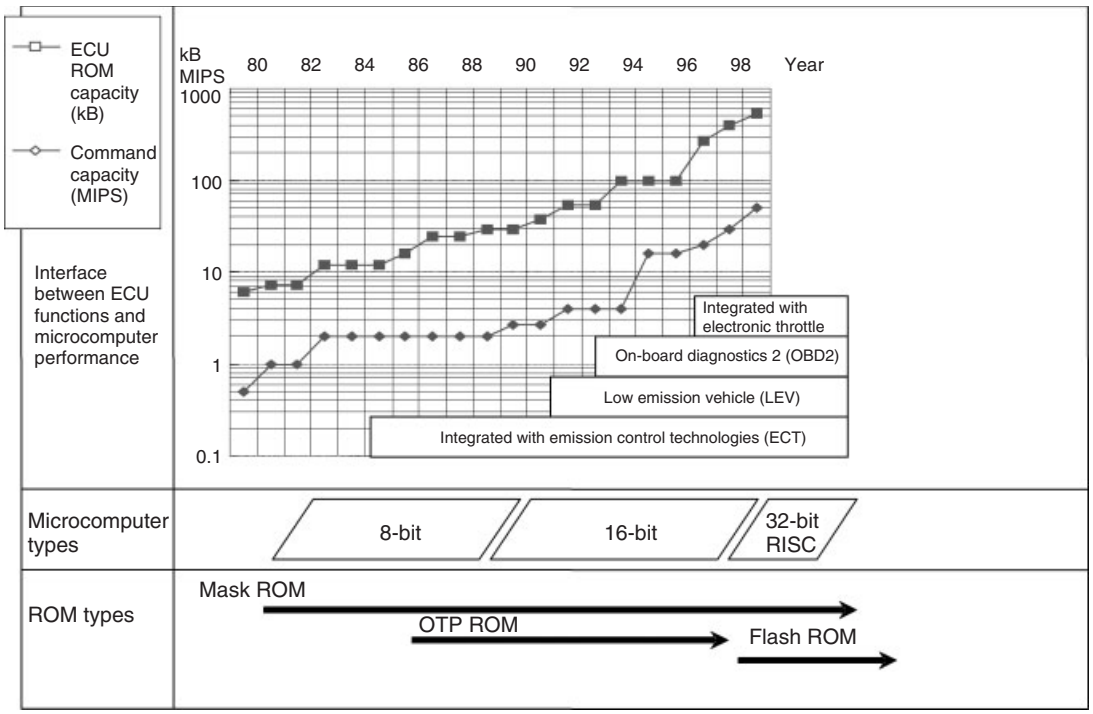


Figure 20. Engine control microcomputer evolution.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6.2.3 Software trends

Engine control by microcomputer began in the period from 1970 to 1980 in response to emissions regulations. Since then, microcomputers have spread to the transmission, airbags, navigation system, automatic air conditioner, and so on. As car electronics particularly the software is responsible for adding value to modern vehicles, this trend is likely to continue in the future. In the 1970s, the length of an engine control ECU program was about 4,000 lines of code. As a result of the recent explosive growth of electronic controls, which is rooted in the growing demands to improve the environment by reducing emissions and fuel consumption and to enhance vehicle safety through systems such as the airbags and seatbelts, a luxury vehicle may contain more than 80 ECUs that are larger in scale and far more sophisticated than ever before. Excluding the navigation system ECU, the total amount of program (software) code now exceeds 7 million lines, leading to a huge jump in developmental work hours.

6.3 Other ICs

Custom ICs designed to combine required functional blocks, such as operational amplifiers, comparators, and power supplies, and semi-custom ICs that make changes to existing blocks have also been developed. In the same volume, custom ICs are most expensive, followed by semi-custom ICs, and the lowest cost general-purpose ICs. The specifications of custom and semi-custom ICs are generally not disclosed to parties other than the customer. Another type of IC mixes digital and analog circuits on a single semiconductor chip.

In accordance with the configuration of the ECU, various single-chip custom ICs have been adopted that combine power devices with function blocks such as the power supply and in-vehicle LAN. As power devices are a cause of heat generation, these custom ICs use power packaging with a high heat dissipation performance instead of standard packaging to reduce the size of the ECU.

7 THE FUTURE OF ECU TECHNOLOGIES

This chapter described ECUs, focusing on their installation environment, operation conditions, common structures, component parts, and circuit parts, using the engine ECU as a typical example. ECUs contain sensor and switch input processing circuits, AD conversion circuits, microcomputers, a power supply, output processing circuits, and a communication circuit for interacting with other ECUs. This chapter detailed the parts provided on the circuit boards of ECUs, categorized in accordance with function and shape. It also included a brief outline of ECU trends.

In the past, ECUs operated independently, but ECUs today communicate with each other, and this has evolved into a networked system of ECUs. Recently, cooperation and integration with network applications have been seen. In the future, these systems will develop into cooperative autonomous systems that autonomously respond to the user's needs. So, as you can see, ECUs will continue to become more complex and larger in scale.

In addition, the ability to develop ECUs for which software and hardware in a short time frame is becoming crucial. It also goes without saying that functional safety must be considered at all times. We are in an age in which it is necessary to create frameworks for developing ECUs that can respond to a diverse range of needs.

RELATED ARTICLES

Historical Overview of Electronics and Automobiles: Breakthroughs and Innovation by Electronics and Electrical Technology
Diversification of electronics and electrical systems and the technologies for the integrated systems
ECU Design and Reliability
Manufacturing: An Introduction to Production Technology, SCM, and Quality Assurance
Microcomputers and Related Technologies: Enlargement of Software Size, Algorithms, Architectures, Hierarchy Design, Functional Decomposition, and Standardization
In-Vehicle Network
Engine ECU Systems

REFERENCES

Japan Institute of Electronics Packaging (2006) *Handbook printed circuit technology*, 3rd edn, Nikkan kogyo Shinbun, Tokyo.

Kato, M. (2010a) *Automotive Electronics: Systems*, Nikkei Business Publications, Inc., Tokyo.

Kato, M. (2010b) *Automotive Electronics: Basic Technologies*, Nikkei Business Publications, Inc., Tokyo.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto213
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Diversification of Electronics and Electrical Systems and the Technologies for Integrated Systems

Yukihide Niimi
DENSO Corporation, Kariya, Japan

1 Introduction 1
2 Layering, Structuring, and Functional Allocation 3
3 Introduction of Common Standardized Parts and Reuse 5
4 International Standardization 7
5 Conclusions 8
Related Articles 9
References 9
Further Reading 9

1 INTRODUCTION

Until recently, cars were a collection of mechanical parts. Electronic control began in the 1970s with the engine, and now extends to the brakes, steering, suspension, instrument cluster, climate control, airbags, door locks, power windows, and the navigation system. In fact, virtually every device used in a modern vehicle is subject to electronic control. The human body works as a metaphor for an electronic control system. Sensors play the roles of the eyes and ears, actuators replace the hands and feet, and the brain becomes the electronic control unit (ECU). Signals measured by the sensors are transmitted to ECUs, where they are used to calculate control signals. The signals to and from the ECU are transmitted via wire harnesses and the like. An ECU consists of hardware such as a microcomputer, peripheral ICs, a power supply, printed circuit board, and connectors, as well as software that controls the calculation procedure.

A typical example of electronic control is the electronic control system in a gasoline engine. The engine ECU functions as the brain to calculate the fuel injection amount and ignition timing. As shown in Figure 3, the engine control ECU is connected with the in-vehicle network to exchange information with other ECUs. For example, the engine coolant temperature is transmitted to the instrument cluster for display on the temperature gage. Another example is in vehicles with adaptive cruise control (ACC), which functions to maintain a set distance with the vehicle ahead. If the vehicle-to-vehicle distance increases, the in-vehicle network is used to request the engine ECU to accelerate. In addition, when the cooling capacity of the climate control is insufficient in hot weather while the vehicle stops, the climate control ECU transmits a signal to the engine ECU to increase the engine speed. For details of automotive control systems, see Robert Bosch GmbH (2008, 2011), Denton (2004), Jurgen (1999), Mizutani (1992).

Figure 1 illustrates the change from independent standalone vehicle systems for engine control, brake control, climate control, and the like to integrated vehicle control systems that perform complex functions in cooperation with multiple control systems. There is also a growing trend for coordinated control with infrastructure outside the vehicle. Conventional vehicle electronics systems consisted of collections of individual ECUs installed separately to perform independent control. However, to achieve the type of complex and versatile control required by modern vehicles, these ECUs are being connected to the in-vehicle network (Figure 3). As the number of connected ECUs

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto214
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

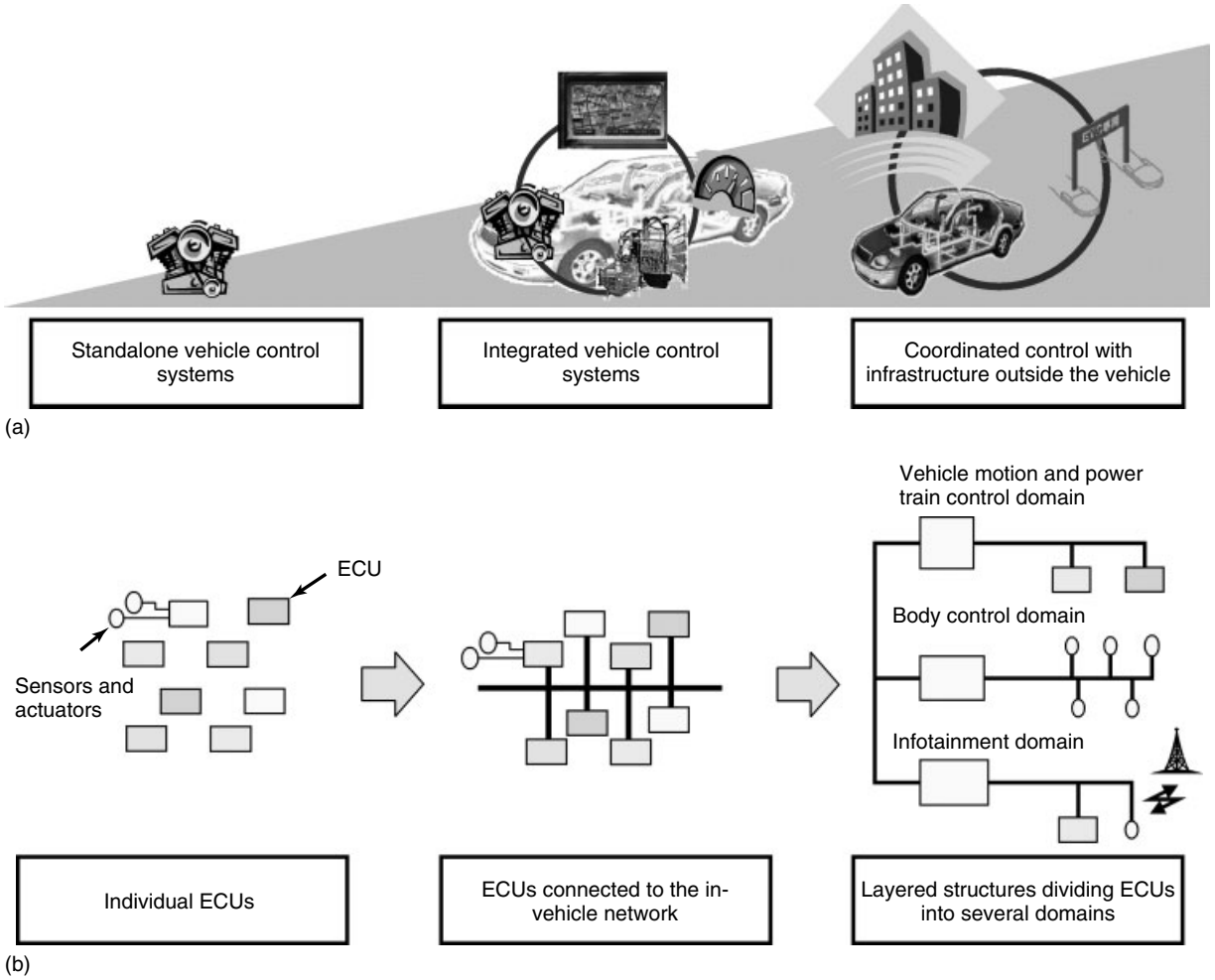


Figure 1. Evolution of automotive electronics systems. (a) Vehicle control systems. (b) Vehicle electronics systems.

increases, it is likely that layered structures will be created that divide these ECUs into several domains (Schäuffele, 2005a).

In the 1970s, the length of an engine control ECU program was about 4000 lines of code. As a result of the recent explosive growth of electronic controls, which is rooted in the growing demands to improve the environment by reducing emissions and fuel consumption and to enhance vehicle safety through systems such as the airbags and seat belts, a luxury vehicle may contain more than 80 ECUs that are larger in scale and far more sophisticated than ever before. Excluding the navigation system ECU, the total amount of program (software) code now exceeds 7 million lines. Figure 2 shows the trends for ECUs installed in luxury vehicles since 1995 (Schäuffele, 2005b).

Consequently, it is becoming more difficult to secure enough space to install electronic system components, particularly ECUs. As a result, ECUs are being made smaller and integrated with sensors and actuators. The commonization and standardization of parts, including software for configuring electronic control systems is also becoming more prevalent to help achieve major improvements in development efficiency.

The control systems of modern vehicles mutually connect the control functions of the power train, chassis, and body (i.e., the fundamental driving, turning, and stopping functions of a vehicle) with service functions that support the driver through information and communication. By doing so, modern vehicles are safer, more comfortable, easier to use, and more environmentally friendly than ever before. The scale of such systems is likely to increase in the future in response to the needs of society, which means that their functions will become even more sophisticated and versatile. New control, information, and communication technologies will be required to achieve such large-scale systems, and it will be necessary to design the optimum systems while ensuring high quality and reliability for both hardware and software.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto214
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

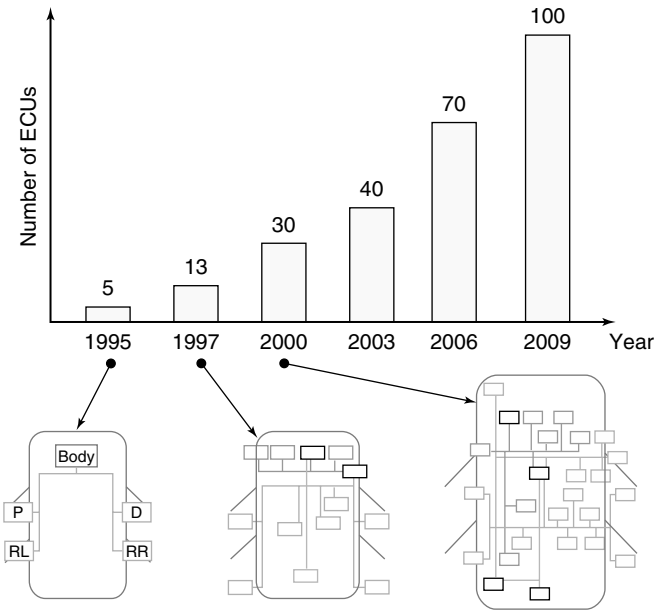


Figure 2. Number of ECUs installed in luxury vehicle.

2 LAYERING, STRUCTURING, AND FUNCTIONAL ALLOCATION

As shown in the example in Figure 3, the ECUs in a modern vehicle are connected to an in-vehicle network. The free exchange of data between the ECUs as well as between the sensors and actuators enables the vehicle system to determine which ECUs should be allocated with which functions from the standpoint of overall optimization. This stands in contrast to the conventional method in which the processes performed by the ECUs were automatically determined for each function.

However, as the number of ECUs in a vehicle increases, it is becoming more difficult to secure enough installation locations. Individual functional ECUs of standard functions that have no variations depending on the region of sale or vehicle grade are gradually giving way to integrated control. The integration of similar or mutually closely related functions is one way of reducing the number of ECU cases, ECU power supplies, component devices, connectors, and wire harnesses required. In other words, reducing these parts is a way of reducing costs. Under this strategy, nonstandard functions, new functions that have not yet found universal application, and functions for certain regions or vehicle grades are not subject to integration. These remain carried out by individual functional ECUs.

As the number of functions in each vehicle increases, another issue that has emerged is how to organize functional relationships and reconfigure the functional structure in the vehicle as a whole. In other words, this means determining how to most efficiently distribute functions throughout a vehicle.

Figure 4 analyzes each type of support function installed in a vehicle. This example shows three layers (the objective control layer, state control layer, and equipment control layer). The relationship between the three layers is as

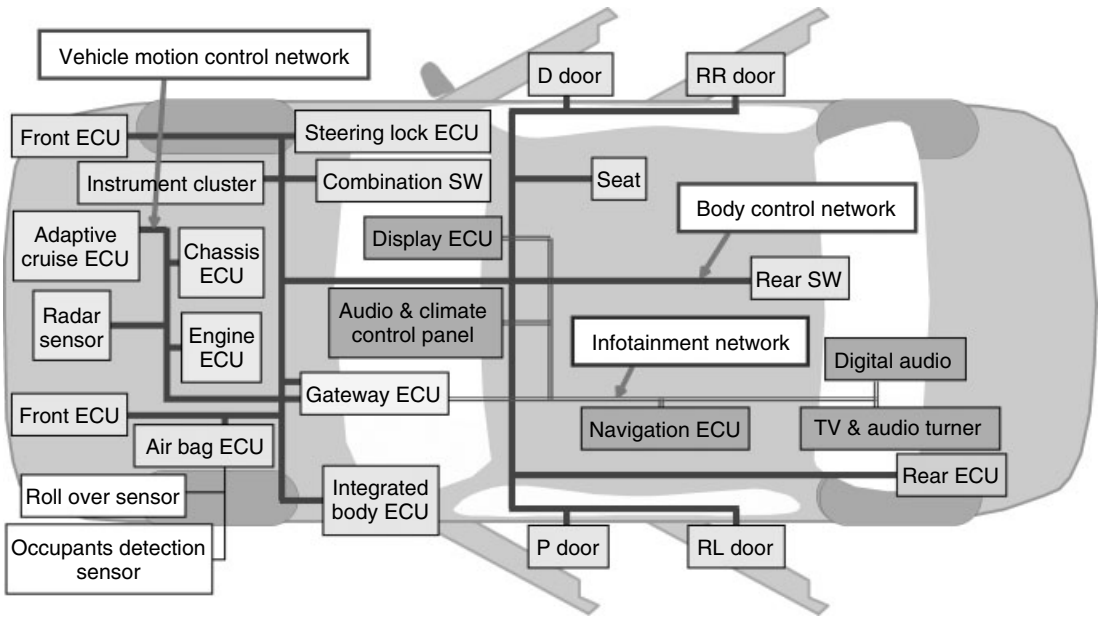


Figure 3. Example of in-vehicle network.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto214
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

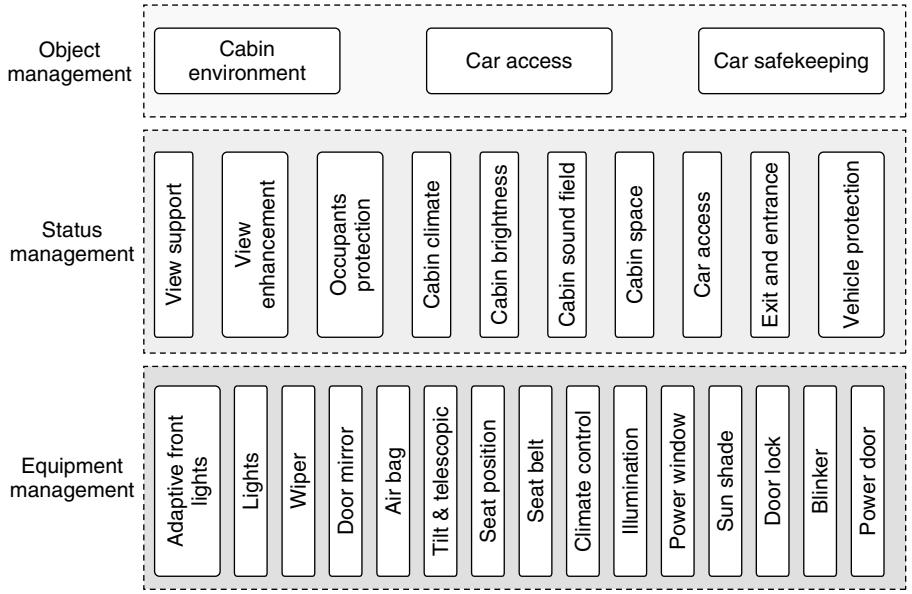


Figure 4. Example of layering of support functions.

	Role of layers	Interfaces of layers
Objective management layer	To obtain target value of required state to achive objective	Interface of objective management and state management: target value of required state
State management layer	To select the optimum equipment for achieving the state requested by the objective management, and to request drive to that equipment	Interface of state management and equipment management: target value of equipment control
Equipment management layer	To drive the actuators based on the required equipment control	

Figure 5. Role and interfaces of layers.

follows. The upper layer acts as the objective for the lower layer, and the lower layer acts as the means for the upper layer. As Figure 5 shows, the role of the objective control layer is to judge the required state and to set a target value that achieves the objective. The role of the state control layer is to select the optimum device for achieving the state requested by the objective control, and to request drive to that device. The role of the device control layer is to drive the actuators based on the required device operation.

For example, electronic keys are one type of support function that helps the driver enter the vehicle. If the driver approaches the vehicle carrying something bulky, the security control detects the key, temporarily deactivates the warning mode system, unlocks the door, and illuminates the interior with foot lamps. The driver can then open the door simply by gripping the door handle. Then, when the driver acts to close the door, the system drives the door closer and sets the driver's seat to the optimal position. Other examples are the climate control that controls the temperature and humidity of the vehicle interior to ensure the comfort of the occupants, and the view support functions such as the windshield wipers, mirrors, and glass defoggers.

Figure 1 illustrated that the scale of electronics systems will continue growing, particularly in luxury vehicles.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto214
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Layering and structuring is a key theme in these vehicles. In the future, a domain controller is likely to be provided for each system to enable sophisticated process function integration of these systems. Another likely trend is the integration of input/output processes with sensors and actuators, regardless of vehicle type. As these trends progress, it is also likely that the in-vehicle network will become layered with small-scale local networks between ECUs, sensors, and actuators, networks within domains for each system or domain, and a backbone network between domain controllers (Bertram *et al*., 1998; Dieterich and Schröder, 1998; Lang *et al*., 2008; Stolz, Kornhaas, and Krause, 2010).

3 INTRODUCTION OF COMMON STANDARDIZED PARTS AND REUSE

At present, one of the largest issues in vehicle development is how to establish an efficient process for designing the myriad of electronics systems concurrently and in a short space of time. This process has to cover a range of vehicle types from mass-market vehicles, which contain only an electronic instrument cluster and engine control ECU, to luxury vehicles that have large-scale electronics systems. This is achieved by maximizing the reuse of parts that have already been designed and developed through commonization and standardization, and newly designing only those parts that are required to achieve new functions. These two types of parts can then be combined to create versatile automotive electronics systems.

A structure of commonized and standardized parts and technologies may be called an *electronics platform*. Vehicle platforms generally refer to shared body and chassis designs. For computers, although the term generally relates to hardware or operating systems (OSs), the meaning and objective are the same from the standpoints of organizing systems, structures, commonized and standardized parts, and technologies to enable the efficient combination of a wide range of variations.

3.1 ECU software structure

This section describes the structure of the software in the ECU as a practical example. Figure 6 shows the relationship between the hardware and software in an ECU, and the software structure.

The electronic control system in an ordinary vehicle consists of sensors, actuators, and ECUs. Each ECU contains hardware such as a microcomputer, peripheral ICs, a power supply, and connectors, as well as software that performs the control procedure.

Conventionally, software was created in a structure that fitted with the characteristics of each individual ECU. More recently, software has been divided into two major categories: the application portion and the common portion, which does not rely on the vehicle type or function. Both are increasingly adopting structures that group together reusable parts. In principle, this structure can be used to form real-time OS, local area network (LAN) communication, device driver, and diagnostic driver software that allows common use across functions (such as engine control, brake control, steering control, and so on), vehicle models, and microcomputer types. It also facilitates the reuse of application component units.

In addition, creating the software for each ECU using the same structure may also enable the reuse of particular application components across ECUs. For example, application components installed in the engine ECU may be used in the climate control ECU. When adding a new function, only the application component that achieves that function has to be created. As a result, the function can be placed in the software platform with minimal changes to the platform, enabling fast and efficient development.

3.2 Other electronic platforms (commonized and standardized parts and technologies)

Obviously, the hardware parts of an ECU also require some form of platform. In the initial phase of electronic system adoption, it was acceptable to make optimal designs for each specific application. However, today, it is becoming increasingly important to achieve the standardization and common use of virtually all ECU hardware parts, such as microcomputers, power supply ICs, communication circuits, connectors, passive components, and so on.

Figure 3 illustrated that modern vehicles have an in-vehicle network system that connects ECUs by communication and exchanges data over the communication paths. This system enables cooperative control of vehicle behavior to assist the driver. For example, by sharing sensor data and using the navigation ECU to predict corners in advance, the engine ECU can command the transmission to generate engine braking by automatically changing gear. This in-vehicle network is utilized by something called a *communication platform*. This refers to a standard configuration of ECU groups connected to the network, a standard communication procedure, standard drivers, and standard interface signals for communicating between the ECUs.

This type of coordinated navigation system control means that the functional structure of the vehicle electronics system is becoming more complex. For example, in the case of longitudinal vehicle motion control, requests for engine

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto214
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

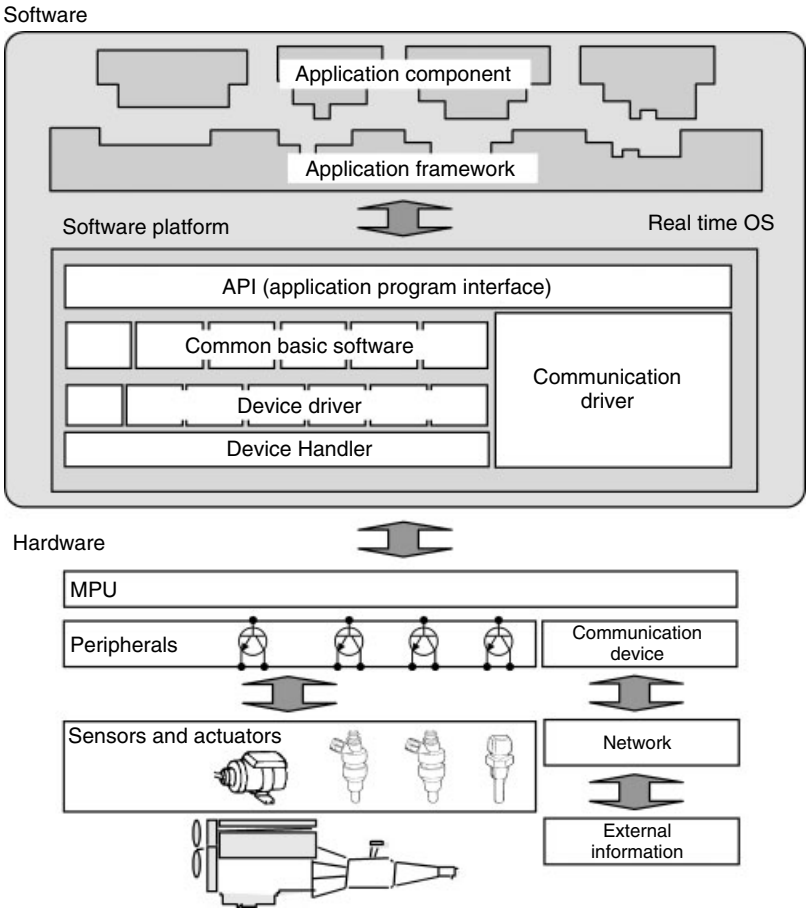


Figure 6. Software structure.

braking do not involve simply the conventional aspects of acceleration and braking. Other ECUs that are relevant to this control include the ACC ECU, which uses millimeter wave sensors and cameras to maintain a set distance to the vehicle ahead, the vehicle stability control (VSC) ECU that ensures stable vehicle handling on rough or dangerous road surfaces, the parking assist ECU that helps the driver parallel or reverse park, the precrash ECU that detects the potential of collision with objects, and the like. Torque commands from all these ECUs are transmitted through the in-vehicle network. Arbitration control prioritizes these signals is an indispensable part of the system and is an important element in the control platform.

Figure 7 shows a vehicle motion control platform. If the vehicle ahead is moving away, the ACC ECU will issue an acceleration request. Similarly, if the vehicle-to-vehicle distance is narrowing, the ECU will issue a deceleration request. The vehicle motions control platform factors in requests from other ECUs (e.g., a deceleration request from the precrash ECU will be prioritized) and calculates a target torque by prioritizing and judging information about the vehicle state to determine how the vehicle should react. This torque is then distributed to the engine and brakes. If gradual deceleration is required, it will reduce the engine request. If sudden deceleration is required, it will issue a braking torque request to the brakes.

The vehicle motion control platform also facilitates the reuse of functions such as these with other functions. By adopting the control structure shown in Figure 7, the high-speed ACC function (i.e., its control logic) developed for use on highways can also be used in principle without modification in vehicles with six-cylinder gasoline or diesel engines, or in hybrid or electric vehicles, as well as in the original four-cylinder gasoline engine vehicle. The development of new functions, such as full-speed ACC that includes start and stop control, can be separated from the development of other engines and carried out independently at the same time. Control platforms are also required for prioritizing vehicle behavior in the lateral and vertical directions as well as longitudinally (AUTOSAR, 2009).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto214
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

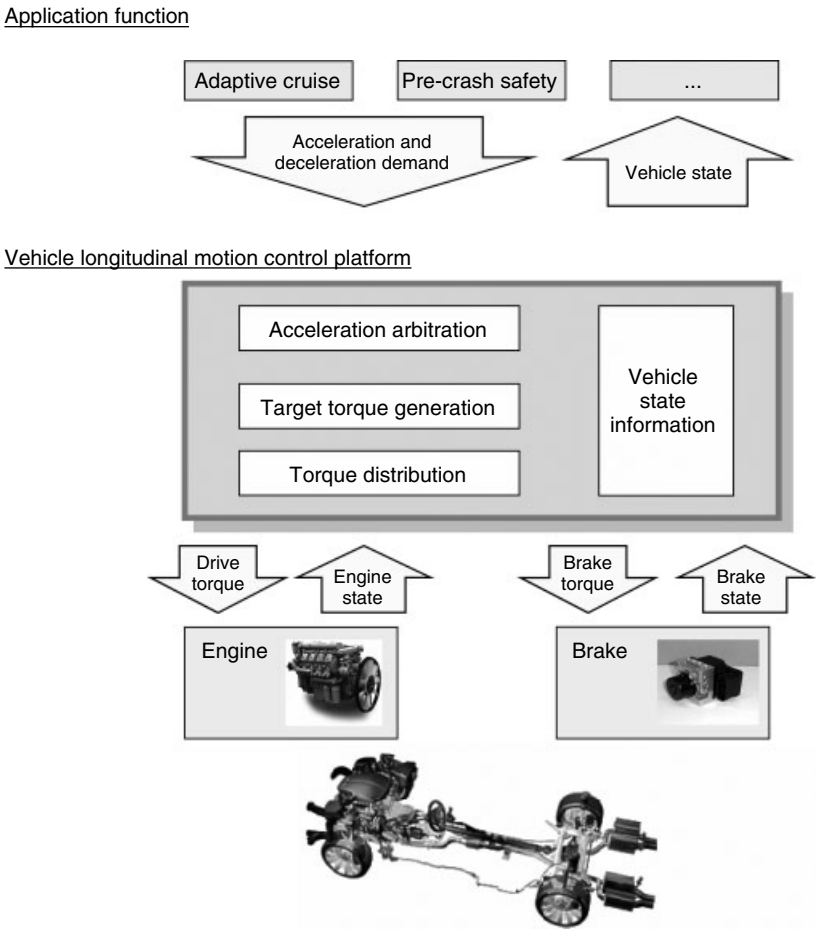


Figure 7. Structuring of vehicle motion control.

Another example of a control platform is the system for prioritizing displays of information to the driver. Since a vehicle only has one driver, it is not possible for that person to simultaneously process large amounts of information. A luxury vehicle has functions to display more than 10 different types of warnings as well as the operation states of various devices. This control platform prioritizes the displays in accordance with the constantly changing driving conditions and shows them in the proper sequence.

This section has described the details of electronic platforms (i.e., commonized and standardized parts and technologies) through various examples. In addition, electronic platforms are a wide-ranging concept in development that also encompasses vehicle power supply systems, types of vehicle control structures, and tools for designing ECUs. These platforms are also seen as the key concept for efficiently developing high quality automotive electronics systems that will continue evolving into the future. All electronic platforms have an interdependent relationship, and it will be necessary to find optimal solutions for the best way to combine such platforms into automotive electronics systems.

4 INTERNATIONAL STANDARDIZATION

The parts and technologies that are being popularized and standardized as described above contain many examples that require standardization on a global basis, irrespective of the industry standards of each country. A typical example is communication procedures, which are a basic technology for constructing the in-vehicle network. This has been the subject of international standardization efforts for some time. These efforts have recently spread to in-vehicle OS, ECU software structures, common basic software, the languages used for specifications, functional safety design methods, design and development tools, evaluation methods, and so on.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto214
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Conventionally, standardization in this field has been carried out internationally through the International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). Although these organizations dominated standardization efforts in the past, the speed of technological and market changes means that recent efforts have mainly been carried out through consortiums that make contracts with the main stakeholders to eliminate time-consuming deliberations on a country level. Although consortiums are in charge of establishing new standards that require speedy response, items that have legal and certification ramifications are still carried out by the ISO. These include the operation and maintenance of established standards, functional safety design methods, interfaces between ECUs and failure diagnostics tools, and procedures to dispose of airbags in scrapped vehicles.

As a typical example, the following sections describe the approach of functional safety design methods, which are formulated by AUTOSAR (automotive open system architecture) and established as ISO standards. Other main consortiums are LIN (local interconnect network) (http://www.lin-subbus.org/) and MOST (http://www.mostcooperation.com/home/index.html), which are related to communication procedures. OSEK (http://portal.osek-vdx.org/) deals with in-vehicle OS related to software, GENIVI (http://www.genivi.org/) works to reduce costs and to shorten development lead times in the fields of multimedia, telematics, and human machine interfaces, and the Association of Standardization of Automation and Measuring Systems (ASAM, http://www.asam.net/) is related to calibration and measurement.

4.1 AUTOSAR (automotive open system architecture)

In 2003, this consortium was founded, centering on German automakers and several major parts suppliers. AUTOSAR is working to standardize software structures, common basic software, software design tools, interfaces between application components, and certification test methods. As of June 2009, there were a total of 158 participating members: nine companies with voting rights (BMW, Daimler, VW, Toyota, PSA, GM, Ford, BOSCH, and Continental), 85 premium associate members with proposal rights (including Honda, Nissan, Mazda, DENSO, Hitachi, NEC, Fujitsu, and Renesas), another 85 associate members, and 8 development members. The work of AUTOSAR is being carried out through 26 working packages.

AUTOSAR has specified the common basic software and proposals for globally standardized specifications that transcend individual functional ECUs, vehicle types, manufacturers, and countries have been made with regard to in-vehicle network drivers such as CAN (controller area network) drivers, LIN drivers, and FlexRay drivers, systems services such as the ECU State Manager and Watchdog Manager, and software models including those related to input/output. For details of AUTOSAR activities, see AUTOSAR (2008), Ataide (2007), and Heinecke *et al.* (2004).

4.2 Functional safety standards

ISO 26262, an international standard for road vehicle functional safety was published in 2011. The standard covers automotive electronic systems relating to safety, and includes subjects such as definitions of concepts for determining safety levels such as the automotive safety integrity level (ASIL), risk assessment and other safety evaluation methods, and approaches to safe lifecycles.

ISO 26262 consists of 10 parts (Gräter, 2011). Part 1 begins with a definition of terminology, part 2 covers the requirements of organizations and people related to safety functions, and part 3 describes the procedure for developing functional safety concepts after setting safety targets. Parts 4–6 describe development and design procedures for systems, hardware, and software (in that sequence). Part 7 covers procedures related to production and market servicing. Part 8 describes guidelines for configuring and managing products developed to support these procedures, guidelines for managing modifications, guidelines for verification, guidelines for managing documentation, guidelines for certifying tools, guidelines for certifying components, and the like. Part 9 deals with guidelines related to ASIL. Finally, part 10 describes reference information to be used as guidelines for parts 1–9.

5 CONCLUSIONS

This chapter described the layering and structuring methods used in basic development of complex and large-scale vehicle electronics systems of today, the standardization and reuse of common parts in these systems, and activities to develop related international standards. Layering and structuring were outlined through a description of software structures and a motion control platform for longitudinal vehicle behavior. International standardization was covered in sections about AUTOSAR and functional safety standards. We expect these technologies will contribute to the higher level of improvement of the function and the performance in the field of fuel economy and vehicle safety in the near future.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto214
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

RELATED ARTICLES

In-Vehicle Network
Vehicle Safety, Functional Safety, OBD Diagnosis
Engine ECU Systems
Hybrid Systems and High Voltage Components
ECU Chassis (Steering)
Chassis ECU (Vehicle dynamics, ABS)
Body ECU (airbag)
Chassis ECU (ACC and sensor)
Body ECU Cluster
Body and Lighting ECU (Key-less Entry, Sonar, HID, LED Usage for Lamps)
Car air conditioning and electronics: analog, digital control and zone management
Car Navigation
Telecommunications
Active Safety, Pre-collision Safety and Other Safety Products (millimeter wave, image recognition, laser)

REFERENCES

AUTOSAR (2009) *Explanation of Application Inter-faces of the Chassis Domain*, http://www.autosar.org/download/R4.0/AUTOSAR_EXP_AIChassis.pdf (accessed 10 October 2013).

AUTOSAR (2008) *Technical Overview*, http://autosar.org/download/AUTOSAR_TechnicalOverview.pdf (accessed 10 October 2013).

Ataide, F.H. (2007) Automotive open system architecture—concept, benefits and challenges. SAE Paper 2007-01-2928.

Bertram, T, Bitzer, R., Mayer, R. and Volkart, A. (1998) CARTRONIC—an open architecture for networking the control systems of an automobile. SAE Paper 980200.

Denton, T. (2004) *Automobile Electrical and Electronic Systems*, 3rd edn, SAE International, USA.

Dieterich, K. and Schröder W. (1998) CARTRONIC—an ordering concept for future vehicle control systems. SAE Paper 98C011.

Gräter, A. (2011) Safety of electric vehicles during their life cycle.. *ATZ Autotechnology*, **11**, 12–17.

Heinecke, H., Schnelle, K., Fennel, H., Bortolazzi, J., Lundh, L., Leflour, J., Maté, J., Nishikawa, K. and Scharnhorst, T. (2004) Automotive open system architecture—an industry-wide initiative to manage the complexity of emerging automotive E/E architectures. SAE Paper 2004-21-0042.

Jurgen, R.K. (1999) *Automotive Electronics Handbook*, 2nd edn, McGraw-Hill, USA.

Lang, H., Döricht, M., Preis, H., Spiegelberg, G. and Gombert, B. (2008) Vehicle architecture integration as an answer to the automotive challenges. SAE Paper 2008-01-0572.

Mizutani, S. (1992) *Car Electronics*, Sankaido Co., Ltd., Japan.

Robert Bosch GmbH (2008) *Automotive electrics and electronics*, 5th edn, Wiley.

Robert Bosch GmbH (2011) *Automotive handbook*, 8th edn, Wiley.

Schäuffele, J. (2005a) *Automotive Software Engineering Principles, Processes, Methods and Tools*, SAE International, USA, pp. 123–125.

Schäuffele, J. (2005b) *Automotive Software Engineering Principles, Processes, Methods and Tools*, SAE International, USA, pp. 16–17.

Stolz W., Kornhaas, R., and Krause, R. (2010) Domain control units—the solution for future E/E architecture? SAE Paper 2010-01-0686.

FURTHER READING

Kato, M. (2010a) *Automotive Electronics: Systems*, Nikkei Business Publications, Inc., Japan.

Kato, M. (2010b) *Automotive Electronics: Basic Technologies*, Nikkei Business Publications, Inc., Japan.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto214
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ECU Design and Reliability

Joji Yoshimi
DENSO Corporation, Kariya, Japan

1 Introduction 1
2 ECU Design 1
3 ECU Reliability 6
Related Articles 13
References 13

1 INTRODUCTION

There are increasing calls to make vehicles more environmentally friendly, safe, convenient, and comfortable. To meet these demands, onboard engine control units (ECUs) have grown more sophisticated in terms of function and performance. As a consequence, the scale of hardware and software has dramatically increased compared to the 1970s when the widespread use of automotive products with electronic applications first began. Drivers are also more sensitive to quality and now demand ECUs with higher reliability. This chapter describes the use of simulations for efficiently designing a large-scale ECU and methods for evaluating whether the ECU achieves a required level of quality.

2 ECU DESIGN

With the advancement of onboard electronics technology in recent years, luxury vehicles are now installed with around 100 ECUs, which cooperate with each other through an onboard LAN connection (Ookura, 2005). In the future, ECUs will have more functions that will also become increasingly sophisticated, and the interaction between the different functions is likely to become more complex.

2.1 ECU design process

Figure 1 shows a general design flow for an ECU. The requirement definition step consists of defining all requirements throughout the ECU life cycle relating to reliability, regulations, and the ECU manufacturing process, in addition to defining requirements for the functions and performance demanded. This can reduce reworking at subsequent steps.

At the step for allocating hardware and software functions, a combination of hardware and software is selected that meets the requirements while minimizing the ECU price and development costs, and various specifications are then created. However, the specifications are often modified many times before mass production of the ECU starts. Any need for hardware changes or software structure changes lengthens the development period. Therefore, creating specifications capable of adapting to any likely requirement changes that can be anticipated at such time helps suppress the length of the development period.

At the hardware design and software design steps, the defined requirements are incorporated into design activities based on the technologies held by each ECU manufacturer.

At the hardware/software consolidated testing step, it is confirmed that the ECU satisfies all the defined requirements. Finally, the ECU is installed in a vehicle and system verification is performed, thus completing the ECU design process.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

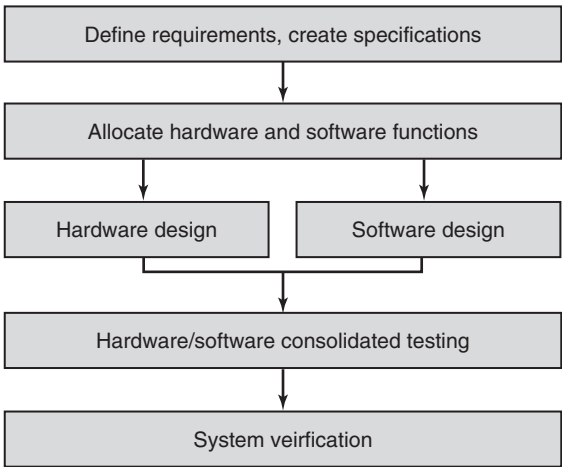


Figure 1. General design flow for ECU.

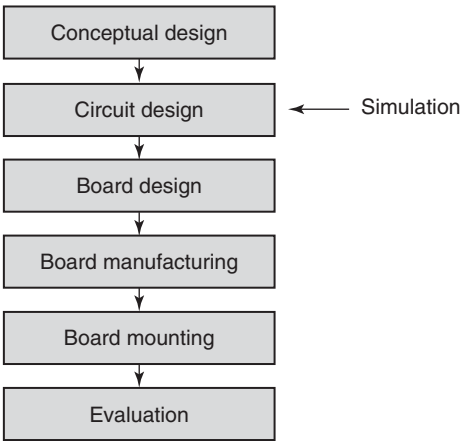


Figure 2. Design flow for printed circuit board.

2.2 ECU design support tools

As mentioned earlier, the ECU has grown in scale and become more complex. Designing and evaluating such an ECU by hand as in the past would require an enormous amount of man-hours and is not realistic. Therefore, the design support tools described below are gaining more widespread use.

2.2.1 ECU hardware development

Figure 2 shows the design flow for a printed circuit board. The process begins with conceptual design, and once the conceptual design is fixed, the process moves on to circuit design using circuit design tools. Next, operation verification is performed using an analysis tool (simulator). After circuit design, board design tools are used to set the layout and wiring for electronic components (board design), thus completing the printed circuit board design process (Kato, 2010).

2.2.2 Software development using virtual environments

Software used to be developed using actual chips. However, recent advances in simulation technology have enabled the co-validation of hardware (mainly microprocessors) and software by simulation (Keating and Bricaud, 2000). This virtual environment has the advantages of allowing software development to start before the trial production of actual chips with excellent observability. However, an amount of development time corresponding to roughly several hundred to several thousand times more than that needed when using actual chips is required (Keating and Bricaud, 2000). Accordingly, there is a trade-off relationship between the speed and precision of the simulation, which means that model development well suited to the simulation purpose and verification items is key.

2.2.3 Software verification using HILS

Instead of using actual automotive components, hardware-in-the-loop simulation (HILS) mimics (simulates) electric signals from such components (Figure 3). Software verification on the ECU is performed using signals originating from the HILS, without requiring the elaborate environment of the actual vehicle. Software-in-the-loop simulation (SILS) methods that build a simulation environment with only software on a computer and use no hardware such as an ECU or actuator have also been proposed (Kato, 2010).

2.3 Future of design support tools (CAE)

This section describes the prospects of high level design and model-based design as applications of design support tools for onboard electronics.

2.3.1 Onboard electronic system development using high level design technology

Owing to recent advances in onboard electronics, ECUs have increasingly sophisticated functions and the interaction between functions has grown more complex. In the future, this trend is expected to accelerate. As an example, the design and verification for an integrated control that requires cooperation among ECUs will become considerably more complicated. Therefore, a new ECU design environment that addresses broader concepts than individual optimization on a component level as in the past

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

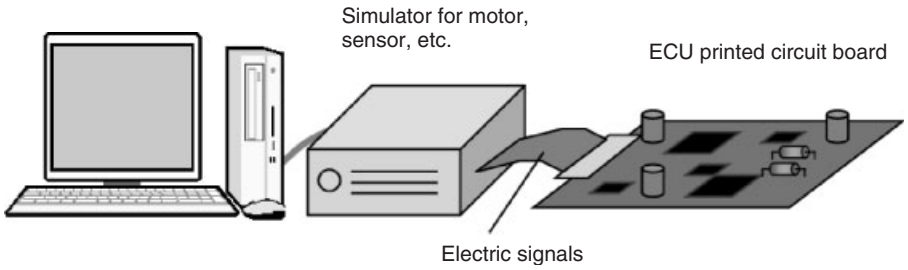


Figure 3. Example of HILS environment.

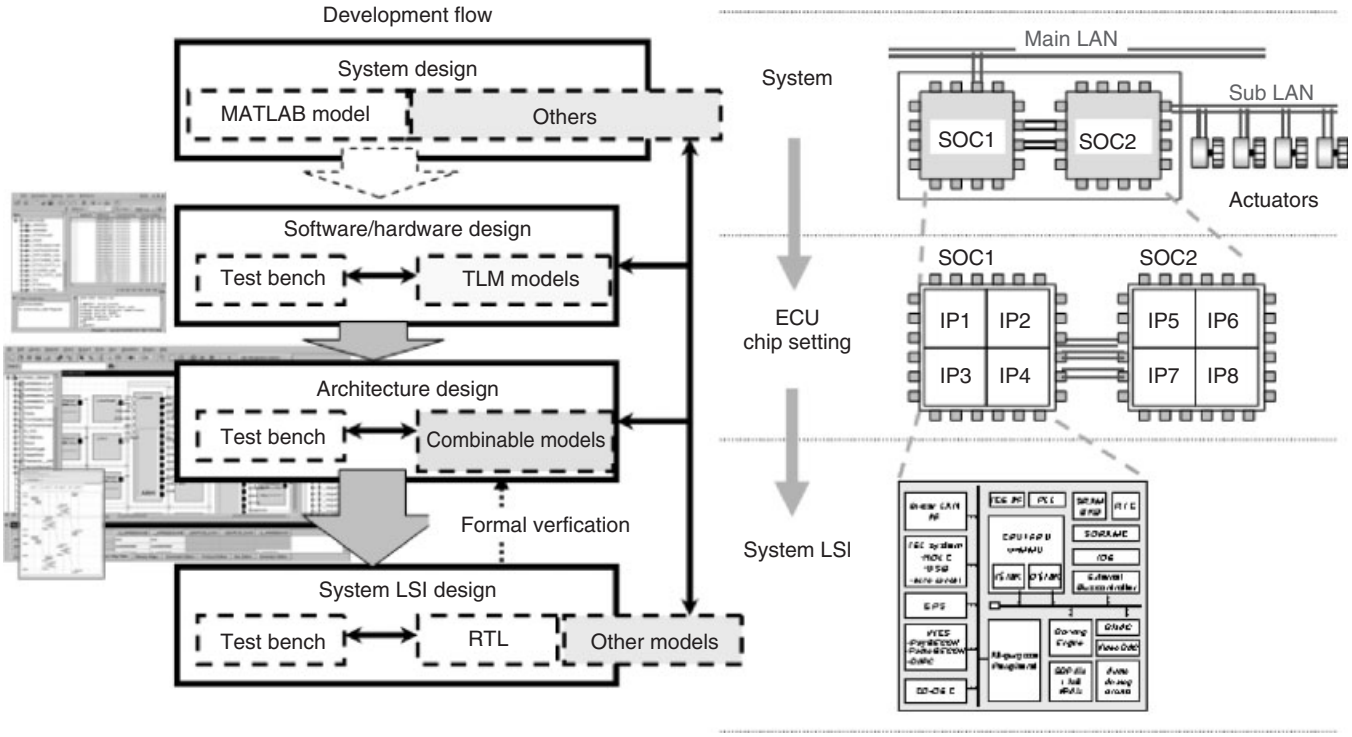


Figure 4. High level design and applicable layers.

is anticipated. One such environment gaining attention is electronic system level (ESL) design (high level design) for ECUs. In this field, the Open SystemC Initiative (OSCI) has standardized the design language SystemC and module interface TLM 2.0.

Figure 4 shows a design flow and applicable layers that apply high level design to an onboard electronic system. High level design has broad applicability ranging from onboard system operations to ICs. However, the pursuit of more precise verification lowers the abstraction level and reduces the simulation speed. Therefore, the verification items must be clarified and the abstraction level set for each applicable layer (Kato, 2010). Creating models of functional components and accelerating the speed of simulation including microprocessors are important elements in ESL design for ECUs.

The advantages of high level design in ECU design are listed below (Kato, 2010).

High level design can be considered as one type of SILS that does not require the actual device. However, high level design includes the concepts of hardware such as microprocessors and enables a system design that takes hardware into account.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

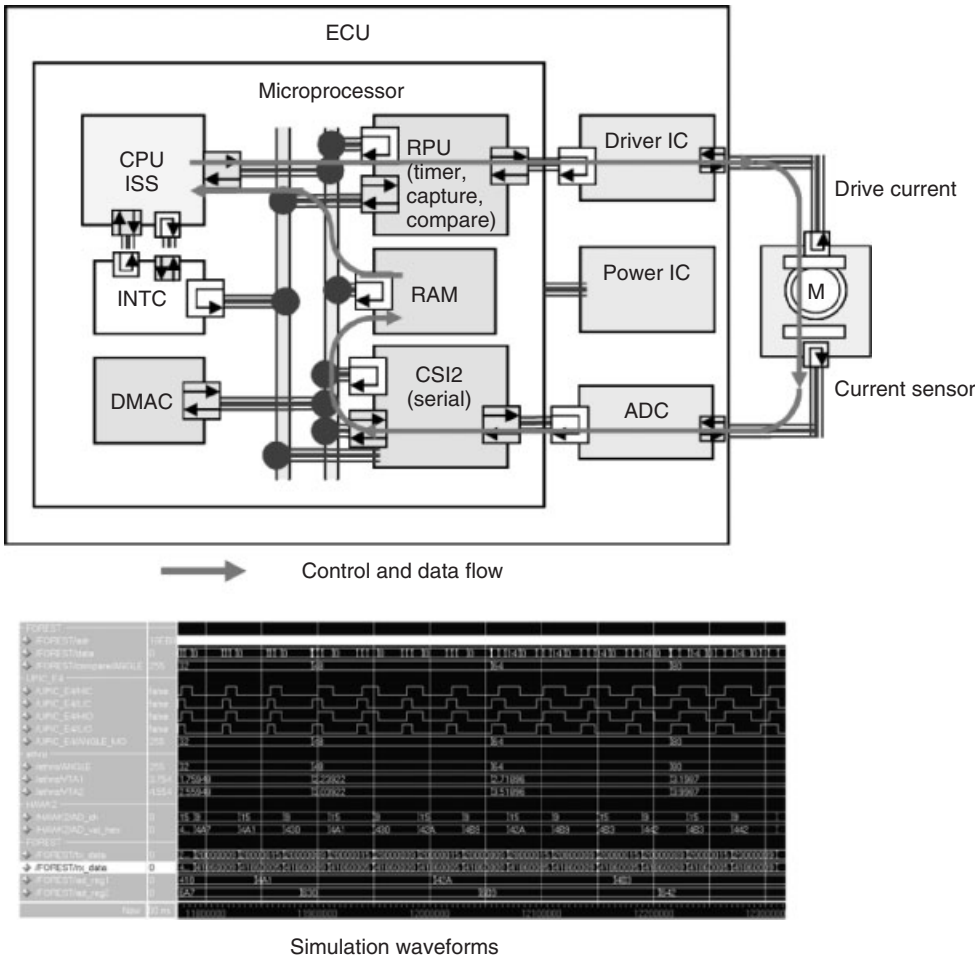


Figure 5. Example of high level design applied to ECU.

Instead of design that simply improves on the conventional ECU, various case studies of ECU architecture are possible in the virtual environment of high level design.

ECU design in the past involved mainly desk calculations, and it was necessary to incorporate a margin of error to achieve the required performance. Performance can be evaluated in a quantitative manner with high level design, which enables suitable design targets to be set and leads to cost reductions.

Problems can be discovered earlier by proceeding with verification in the virtual environment before making the actual device.

Advance software development is possible with the use of the microprocessor virtual environment.

Figure 5 shows an example of an ECU model. As the ECU, in addition to the microprocessor (CPU + peripheral circuits), models of the power IC and driver IC, as well as models of sensors and actuators connected to the ECU are required.

2.3.2 *Model-based design (linked to algorithm development)*

Physical states such as temperature and rotation angle are converted into sensor electric signals, and a feedback control is executed for actuators including the engine and motors based on these electric signals. This type of control development exists for many onboard systems.

The development process consists of developing algorithms first, and then designing the control hardware and building the system. MATLAB or the like is often used for algorithm development. An example of the flow of a model conversion that incorporates algorithms developed using MATLAB into a hardware model is described below.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

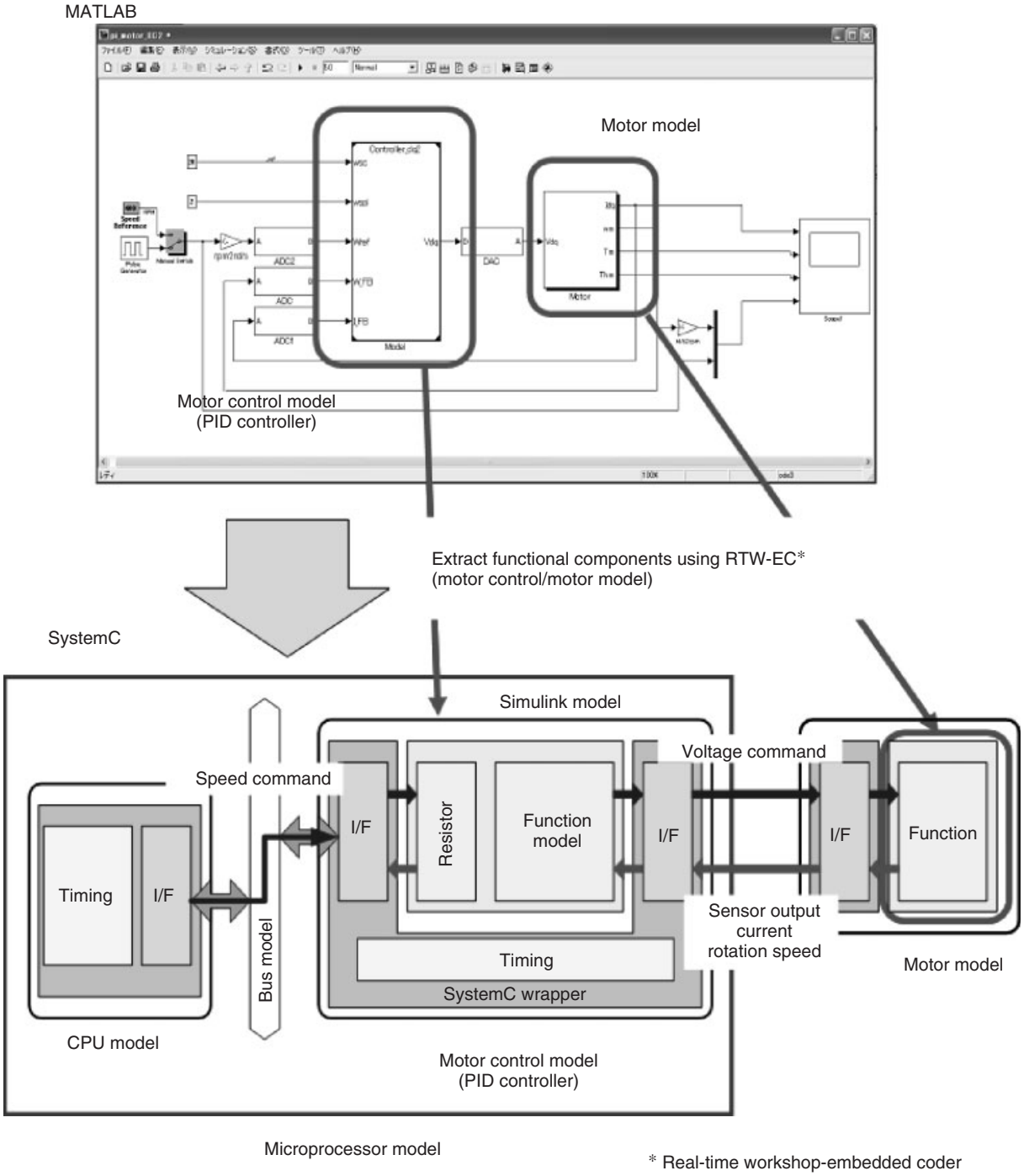


Figure 6. Example of conversion from MATLAB to SystemC model.

1. Algorithm development using MATLAB
2. Analog signal discretization
3. Function output in C code
4. Add interface and time control to C code and use as SystemC code
5. Use SystemC code circuit as high level design model

Although problems with simulation speed remain, tool vendors have recently developed simulation technology that combines MATLAB and high level design tools. Further advances in technology and application to onboard electronic systems are anticipated (Figure 6) (Kato, 2010).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3 ECU RELIABILITY

To ensure ECU reliability, first, the targeted quality and life are specified. Next, the environment in which the ECU will be used is studied. Finally, reliability testing that applies stresses corresponding to the use environment must be performed to confirm that the ECU satisfies the quality and life targets (Reliability Engineering Association of Japan, 1997).

A number of ECU failure patterns are represented by a bathtub curve in Figure 7. The failure rate tolerance and product service life shown in Figure 7 must satisfy the targeted quality and life.

The bathtub curve has an early failure period, a chance failure period, and a wear failure period based on shifts in the failure rate. The early failure period contains the highest probability of manufacturing and design defects, which means that there are many failures at the start of the period but the number gradually decreases. Screening by burn-in is effective for ensuring that defective products from this period do not enter the market (Renesas Electronics Corp, 2010). Burn-in is a method of applying stress to the ECU and ECU components over a set period to sort good and defective devices. This is explained in Figure 8 using an S-N curve.

Defective devices are normally distributed to a location away from good devices. Only defective devices fail after a fixed time (tB) following application of a stress (SB), so defective devices along with good devices can be stratified. Screening consists of focusing on the failure mode of the applicable device, and finding and eliminating design weaknesses and potential faults ahead of time through the

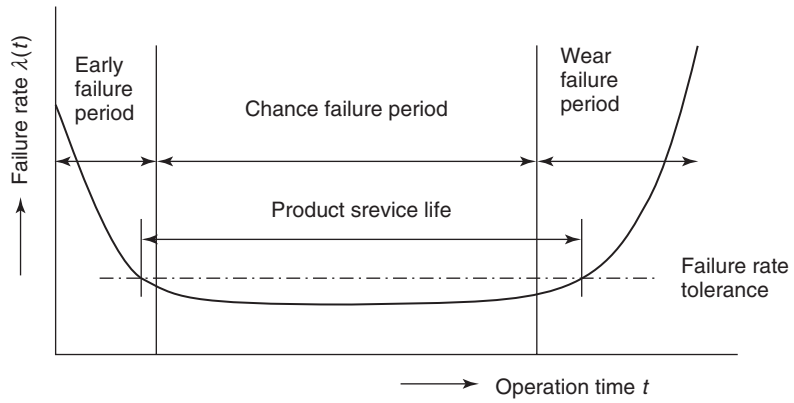


Figure 7. Bathtub curve.

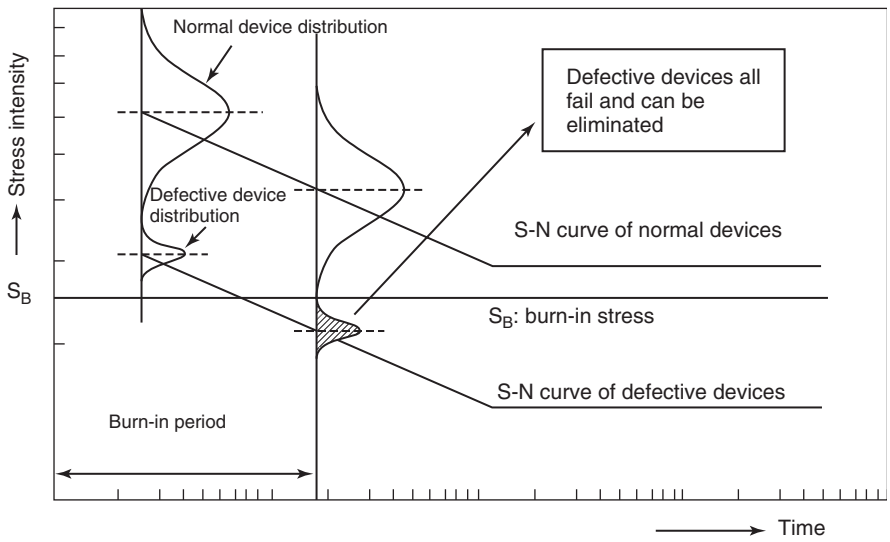


Figure 8. Burn-in.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

appropriate application of stress. For effective screening, a stress type and application time capable of exposing a failure in the defective device must be selected. However, applying excessive stress to shorten the screening time or applying stress over a long period to reliably eliminate defective devices may adversely affect the characteristics of good devices or cause failures. Therefore, it is important to select the appropriate conditions.

In the chance failure period, failures occur for any random reason and there is a fairly steady failure rate. The wear failure period is a period in which failures increase due to deterioration over time from product use. The only ways to reduce the failure rate in the chance failure period and the wear failure period are by understanding various types of failure mechanisms and continuing steady efforts to improve tolerance against failure. In particular, countermeasures for wear-related failures should ensure that a large number of defects do not occur within the warranty period.

3.1 ECU environment

Vehicles must be safe and comfortable under various environments and continue to operate normally until they are scrapped. The quality of onboard ECUs must also remain high over a long period under harsh conditions such as heat, humidity, vibration, noise, and static electricity. Table 1 shows examples of environment conditions for an onboard ECU (Kato, 2010). The values listed in the table are only examples, and the ECU mounting environment in the vehicle must be accurately understood in the course of ECU development. These values ensure the validity of the reliability assessment described later.

3.1.1 Temperature

While the engine is operating, the temperature in the engine compartment can increase until 100°C or more. The temperature is also influenced by weather conditions, and may increase until 90°C on the top of the dashboard and until 65°C in the trunk when parked under the hot sun. In addition, the temperature may alternate between high (during the day) and low (at night), and sharp temperature increases may occur due to self-heating during load driving.

Table 1. Examples of environment conditions for onboard ECU.

Temperature	−30 to 110°C: engine compartment
	−30 to 80°C: vehicle cabin
Humidity, water	95% RH or more, water immersion
Vibration	20G, 20–200 Hz: engine compartment
	4.4G, 20–200 Hz: vehicle cabin
Electricity	Power voltage fluctuation: 6–24 V
Vibration	Surge voltage from induction load
	Electromagnetic damage, static electricity
Electricity	Dust, salt, grease

Reproduced with permission from Kato, 2010 © Denso Corporation.

3.1.2 Humidity and water

This environment refers not only to outdoor humidity but also to the condensation that occurs when devices cooled by the air conditioner are subjected to outside air with high humidity, which may leak inside the circuit board and cause electrical or other corrosion. Water immersion even in the vehicle cabin must also be considered due to the possibility of spilled drinks and water entering the cabin during car washing.

3.1.3 Vibration

Driving is always accompanied by vibration, and some regions may reach until 20 G depending on the road surface and running conditions. Therefore, trouble caused by vibration must be taken into account.

3.1.4 Electricity

Various loads are connected to the vehicle battery, and the power voltage varies depending on their on/off state. At cold starts in particular, there are considerable changes in power voltage, which may range from a nominal 12 V to approximately 6 V. Operation of induction loads such as motors, solenoids, and relays, as well as current interruptions, may be accompanied by induction noise in other devices or parallel wire harnesses due to electromagnetic or electrostatic coupling.

3.2 ECU failure examples

Understanding the failure modes that occur in the ECU environment is essential to securing ECU reliability. This is because the accelerated testing described later focuses on failure modes to calculate an acceleration coefficient, and proper testing cannot be performed if the failure modes cannot be anticipated. Figure 9 shows an example of failures that occur when humidity, heat, hot/cold stresses, and vibration are applied as typical stresses. Most of these failures are exposed after a number of years in a normal use environment, so countermeasures at the design stage and sufficient evaluations are critical.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

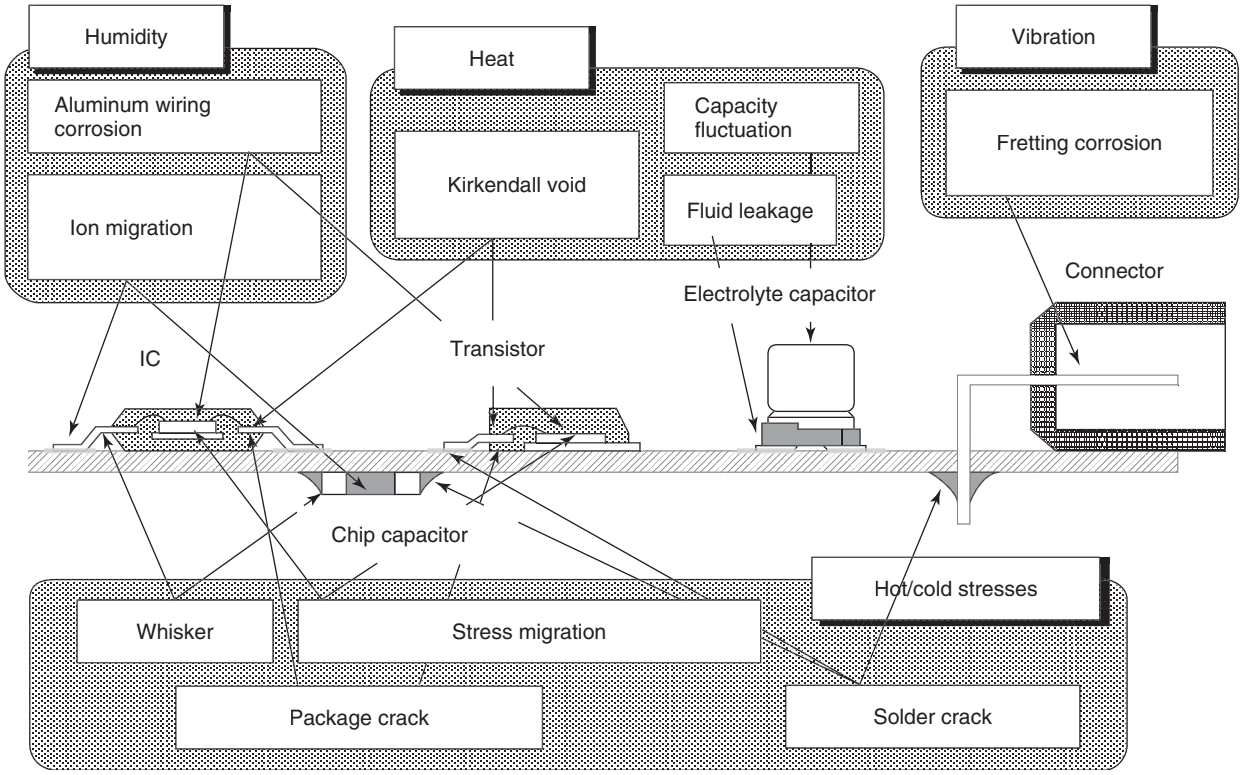


Figure 9. Examples of ECU failures caused by various stresses.

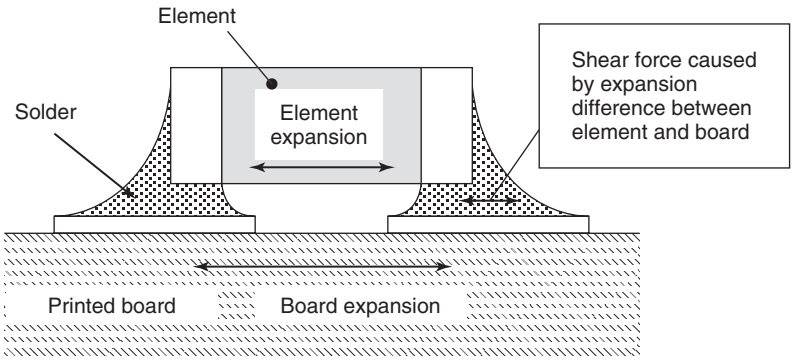


Figure 10. Solder crack mechanism.

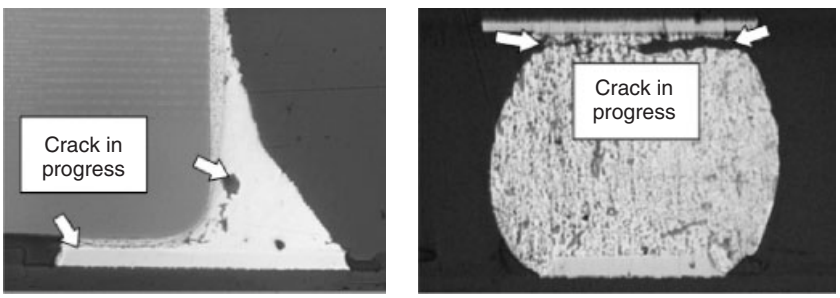


Figure 11. Solder crack examples.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

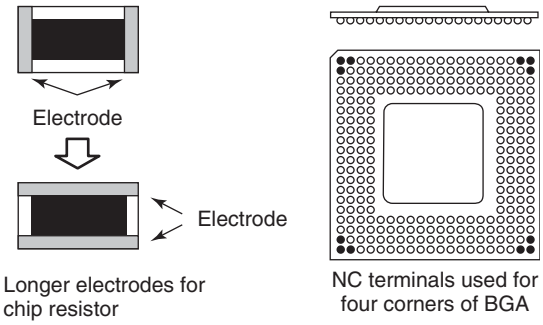


Figure 12. Solder crack countermeasures.

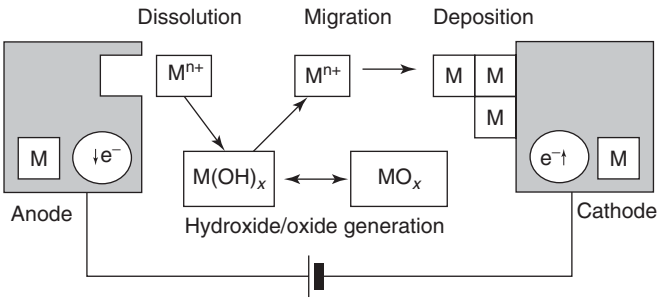


Figure 13. Ion migration mechanism.

3.2.1 Solder crack

The printed circuit board and its mounted electronic components have different coefficients of thermal expansion. Therefore, temperature fluctuations may produce stress on bonded sections (Figure 10). Such stress is absorbed by the solders, but with repeated heating and cooling, the solders may experience plastic deformation and exhibit minute cracks. The cracks may gradually grow as shown in Figure 11, and ultimately lead to an open defect.

For an ECU used in an environment with large temperature differences, to reduce hot/cold stresses, components with longer electrodes to shorten the interval between electrodes are adopted. For a BGA package IC, countermeasures such as using NC terminals without internal connections for the corner terminals subject to the maximum stress are adopted to ensure that functionality is not affected even if cracks occur (Figure 12) (Kato, 2010).

3.2.2 Ion migration

Ion migration is a phenomenon in which metal that includes electronic material ionized by the action of the electric field migrates between electrodes and is then reduced to metal again and deposited, as shown in Figure 13. The ion migration phenomenon is accompanied by visible tree-like branching called dendrites. An example is shown in Figure 14. The dendrites contain metal and are generated by the normal environment of the onboard ECU, such as temperature, humidity, electric field (voltage), and impurities (e.g., ions, contamination, and dust). These environment elements act in complex ways to form dendrites that cause insulation degradation in circuits and components.

For cases in which the ECU is used in an environment with high humidity and there are only narrow intervals of several millimeters between electrodes, countermeasures include avoiding the use of Ag, Pb, Sn, and Cu materials, blocking the ion migration path by covering the space between electrodes with a moisture-proof material or gel resistant to moisture absorption, and reducing the halogen concentration (Kato, 2010).

3.2.3 Whiskers

Whiskers are a phenomenon in which needle-like or nodule-like metal single crystals grow on a metal surface. Whiskers are known to occur from Sn plating and Zn plating. Whisker formation may lead to defects caused by short-circuiting between terminals of the circuit components. Figure 15 shows examples of whiskers found on an electrode of a chip capacitor and an IC lead in a temperature cycling test. Known causes of whiskers include formation by stress acting on the plating film, which occurs due to different coefficients of thermal expansion, intermetallics and their

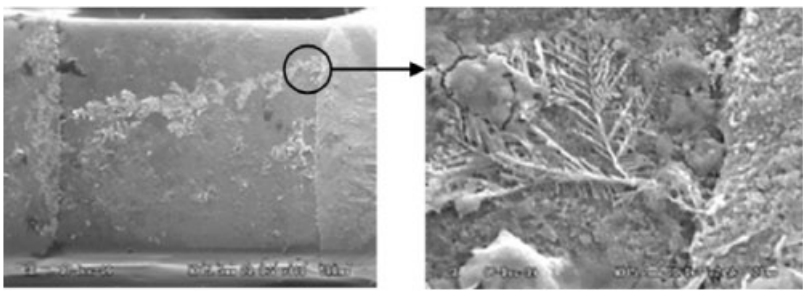


Figure 14. Dendrite formed between electrodes of ceramic capacitor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

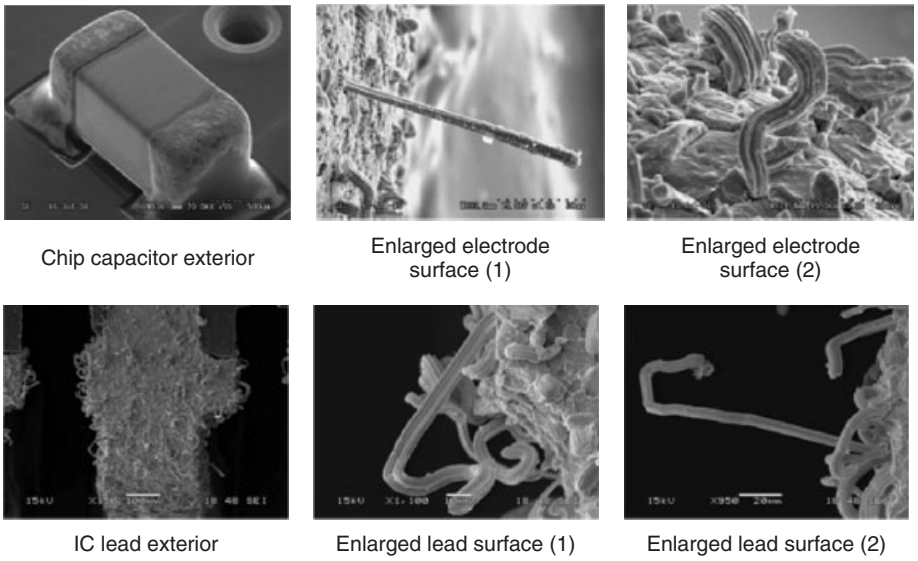


Figure 15. Whisker examples.

spread, galvanic corrosion, external stress, and the like. However, the mechanism is still not clear (Kato, 2010).

Countermeasures that have proved effective include applying a heat treatment (annealing treatment) after plating to reduce residual stress and adding a second metal to the Sn plating. Sn plating containing Pb was once often used in the industry. However, plating containing Pb can no longer be used because of the RoHS Directive, and reliability issues involving tin whiskers have arisen from the use of lead-free compliant components (Yoshino, Sanji, and Iguro, 2007). Therefore, whisker countermeasures for Sn plating are gaining attention again.

3.2.4 Kirkendall voids

Leaving the IC in a high temperature state or operating the IC in a continuous manner produces a solid phase reaction at the bonding portions of gold bonding wires and aluminum wiring. This may form a compound of gold and aluminum as a consequence and generate a disconnection defect as shown in Figure 16.

Interdiffusion at the Au–Al bond is accompanied by the formation of various types of metal compounds as shown in Figure 17. Voids known as Kirkendall voids may occur in intermetallics that are weaker than aluminum and gold, which could produce cracks after repeated cycles of heating and cooling and lead to fractures. Countermeasures include not using substances that accelerate the Kirkendall phenomenon such as bromine in the IC package resin and delaying the progress of the Kirkendall phenomenon by including palladium in the gold wires (Kato, 2010).

3.2.5 Aluminum wiring corrosion

Resin-sealed packages are widely used for semiconductor devices. However, external moisture may penetrate the package and corrode the aluminum wiring of the semiconductor chip, resulting in a failure (Figure 18).

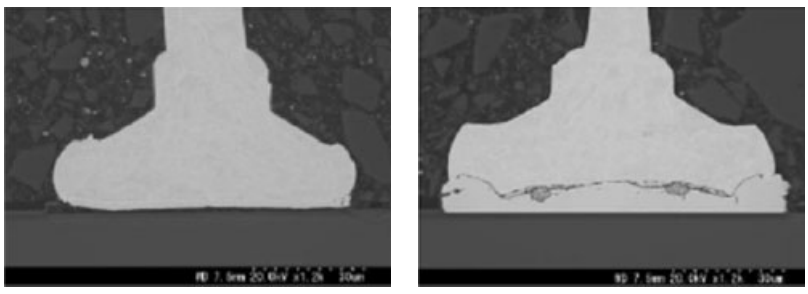


Figure 16. Kirkendall void.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

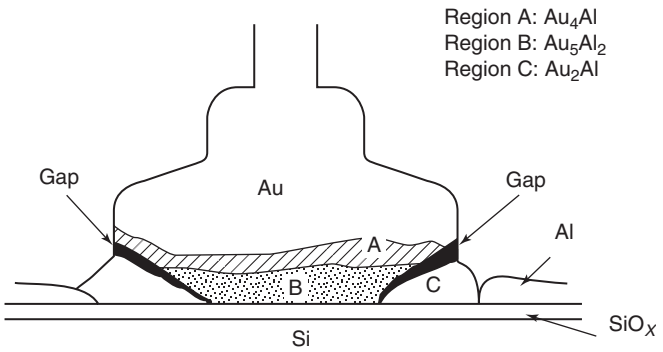


Figure 17. Alloy layers of bonding portion.

When left in dry air, aluminum develops a stable aluminum oxide (Al_2O_3) that acts as a protective film. However, an aluminum hydroxide ($Al(OH)_3$) soluble in acid and alkali is formed in the presence of moisture. If ions of impurities such as P^{3-}, Cl^-, F^-, and Na^+ are also present, corrosion occurs at exposed sections of aluminum wiring, bonding pads, pin holes in the chip protective film, cracks, and other locations. Countermeasures include adopting a highly moisture-resistant passivation film (chip protective film), lowering the moisture absorbency of the resin, and improving resin adhesion (Kato, 2010).

3.3 Reliability assessment

Before the ECU is mounted in a vehicle and shipped to market, it is evaluated by various tests and any design or manufacturing defects are corrected. To ensure effective reliability testing, it is important to accurately understand the vehicle environment and, using the accelerated testing approach described later, to select test conditions that correlate to stresses the ECU will likely receive once on the market, so that the extent of the margin with respect to the targeted life can be correctly grasped.

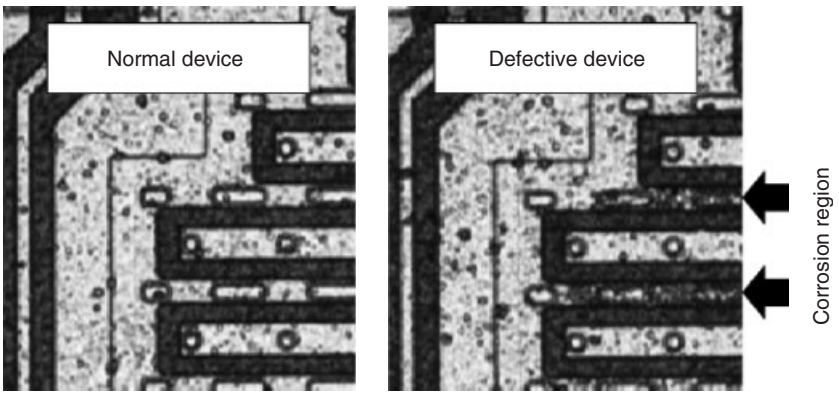


Figure 18. Aluminum wiring corrosion.

Table 2. Main reliability test items.

Test Item		Purpose
Life test	High temperature operation test	Evaluate ECU durability when subjected to electrical and thermal stress over a long period
	High temperature/high humidity operation test	Evaluate ECU durability when operating under a high temperature, high humidity atmosphere
	Humidity cycle test	Evaluate ECU durability in a temperature-varied environment that repeatedly cycles between high and low temperatures
	High humidity shelf test	Evaluate ECU durability when left in a high humidity environment
	High temperature/high humidity shelf test	Evaluate ECU durability when left in a high temperature, high humidity environment
Strength test	Load dump test	Evaluate ECU resistance to positive surge voltage applied with battery terminals removed
	Field decay test	Evaluate resistance to negative surge voltage from an alternator induction load
	RF immunity test	Evaluate ECU resistance when left under an intense electric field atmosphere
	Electrostatic test	Evaluate ECU resistance to static electricity during handling
	Vibration test	Evaluate ECU mechanical strength in a vibration environment
	Salt spray test	Evaluate ECU corrosion resistance when left in a saline environment

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.3.1 Reliability test items

Examples of general tests include ISO, IEC, SAE, and other standards. If the ECU will be used within the ranges assumed by these standards, the reliability of the ECU can be confirmed according to the test conditions set in the standards. However, car and parts manufacturers often set their own individual test items and evaluation conditions as internal standards to quickly respond to changes in onboard environments and achieve individual quality targets. Table 2 shows examples of test items (intended as only a partial list and not comprehensive). Many life tests overlap with test items performed for individual electronic components, and strength tests have many test items that concern the ECU used in an onboard environment such as the load dump test and the field decay test.

3.3.2 Accelerated testing

Utilizing the fact that applying a large amount of stress to a product shortens the life of the product, accelerated testing is used to perform a reliability assessment in a limited period of time by testing under stricter conditions than use conditions and environment conditions in the market. The acceleration coefficient A is expressed as a ratio of the test times at which a stress S_1 used as a reference and a stress S_2 used as an acceleration condition correspond to the same failure rate as shown in Figure 19.

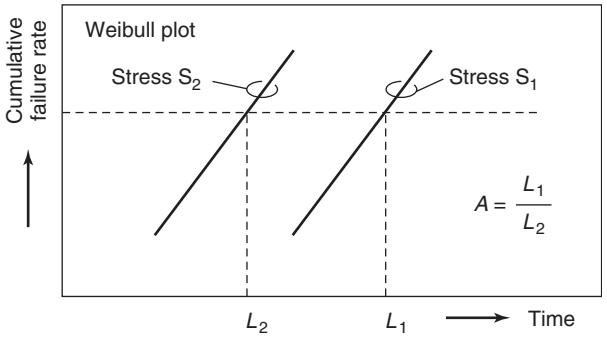


Figure 19. Calculation of acceleration coefficient.

Table 3. Representative acceleration models.

Acceleration Factor	Acceleration Model/Life Equation	Notes
Temperature	Arrhenius model $L = A \bullet \exp\left(\frac{E_a}{k \bullet T}\right)$	- Electromigration - Change in oxide film over time (TDDB) - Al electric field capacitor life L: life A: constant E_a: activation energy k: Boltzmann constant T: absolute temperature
Electric field	E model $L = A \bullet \exp(-\beta \bullet E)$ $1/E$ model $L = A \bullet \exp\left(\frac{\gamma}{E}\right)$	- Gate oxide film TDDB β, γ: acceleration factor E: electric field
Temperature difference	Modified Coffin–Manson model $L = A \bullet f^m \bullet \Delta T^{-n} \bullet \exp\left(\frac{E_a}{k \bullet T_{max}}\right)$ Temperature difference accelerated model $L = A \bullet \Delta T^{-n}$	- Solder thermal fatigue m, n: constant ΔT: temperature difference f: frequency T_{max}: maximum temperature - Al slide - Interlayer cracks
Humidity	Absolute water vapor pressure model $L = A \bullet V_P{}^{-n}$ Relative humidity model $L = A \bullet (\mathrm{RH})^{-n} \bullet \exp\left(\frac{E_a}{k \bullet T}\right)$	- Al corrosion Vp: absolute water vapor pressure RH: relative humidity
Current	Black model $L = A \bullet J^{-n} \bullet \exp\left(\frac{E_a}{k \bullet T}\right)$	- Electromigration J: current density

Reproduced by permission of JEITA: Japan Electronics and Information Technology Industries Association (2011).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Activation energy examples.

Failure Mode	Failure Mechanism	Activation Energy (eV)
Threshold voltage shift at MOSFET gate	Ionic contamination	1.0–1.4
	Slow trapping	1.0–1.5
Current leakage	Generation of inversion layer	0.5–1.0
	Channel effect	0.5
Lowered hfe	Accelerated ion migration caused by moisture	0.8
Al wiring disconnection	Al wiring corrosion	0.5–1.0
	Al electromigration	0.4–0.7
Short-circuit	Breakdown of oxide film	0.3–1.1

Reproduced with permission from Kato, 2010 © Denso Corporation.

3.3.3 Acceleration model

Table 3 shows representative acceleration model examples (Kato, 2010), and Table 4 shows examples of activation energy used in acceleration equations (Japan Electronics and Information Technology Industries Association, 2011). Using an acceleration equation enables infinite acceleration in desk calculations, but the failure mode may change depending on the stress conditions. Therefore, consideration must be given to setting the acceleration conditions within a range where the failure mode does not change.

RELATED ARTICLES

ECU Technologies from Components to ECU Configuration
Microcomputers and Related Technologies: Enlargement of Software Size, Algorithms, Architectures, Hierarchy Design, Functional Decomposition, and Standardization
Engine ECU Systems

REFERENCES

Japan Electronics and Information Technology Industries Association. (2011) *EDR-4708: Guideline for LSI Reliability Qualification Plan*.

Kato, M. (2010) *Automotive Electronics Illustrated*, Nikkei Business Publications, Inc., Tokyo.

Keating, M. and Bricaud, P.Nihon Synopsys G.K., Mentor Graphics Japan Co., Ltd. (2000) *Reuse Methodology Manual For System-On-A-Chip Designs*, Maruzen Publishing Co. Ltd, Tokyo.

Ookura, K. (2005) Semiconductor Technology for Automotive Electronics, *Denso Technical Review*, **10**(2).

Reliability Engineering Association of Japan (1997) *Reliability Handbook*, Union of Japanese Scientists and Engineers.

Renesas Electronics Corp (2010) *Semiconductor Reliability Handbook Rev.0.50*.

Yoshino, M., Sanji, M., and Iguro, S. (2007) Whisker Generation Mechanism of Lead-free Soldered Joints, *Denso Technical Review*, **12**(2).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto215
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Manufacturing: An Introduction to Production Technology, Quality Assurance, and SCM

Naoki Ueda
DENSO Corporation, Kariya, Japan

1 Introduction 1
2 ECU Production Technology and Quality Assurance 2
3 Supply Chain Management 9
4 Conclusion 10
Related Articles 10
References 10

1 INTRODUCTION

The production of automotive electronic control units (ECUs) has many requirements. First, ECUs with various forms and structures must be produced for a wide range of purposes in addition to the engine ECU. Electronics have been applied to many automotive controls, which means that vehicles are now equipped with a number of ECUs (Kato, 2010a). As a result, it is necessary to create structures that can be mounted near desired locations based on ECU function, as well as to secure installation space and reduce vehicle weight. In addition to the need to reduce size and weight, there are demands for mountability at locations with harsh environments in terms of temperature, humidity, moisture, and the like. The ECU mounting position varies depending on function and purpose. The engine ECU is often mounted inside the engine compartment because many signals originate from the engine compartment and also because the role of this ECU is to control the engine. Body system ECUs often receive signals from inside and around the vehicle cabin and are usually mounted inside the occupant compartment as a consequence. As it is critical that the combination graphic display and car navigation system can be easily recognized and operated by the driver, the ECUs for these are placed near the instrument panel. A location near the transmission is desirable for the transmission ECU. The motor ECU is preferably mounted near the motor, and integration of the ECU and its control object is also being studied.

Thus, the form and structure of the ECU varies depending on differences in usage and the installation environment. An ECU in the occupant compartment often uses a low-cost, resin-printed board with high flexibility. An ECU with a resin-printed board is often used in the engine compartment as well, because the reliability of parts and the mounting technology of parts have been improved and structure of heat dissipation also have been improved. However, for severe environments such as direct mounting on the engine, a hybrid ECU that uses a ceramic substrate capable of withstanding this type of environment is adopted (Kato, 2010b).

As explained earlier, there are ECUs with various functions, specifications, forms, and structures, and the production technology for each ECU is appropriately selected in consideration of quality and cost.

Second, a stable and high-quality ECU must be provided as necessary to various users while adapting to the increased sophistication and diversification of specifications. Therefore, it is important to ensure stable supply performance and quality over the entire supply chain, starting with material and part suppliers and ending with the manufacturer and the user.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto216
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Third, environmentally friendly products and production methods are required. There have been demands for soldering using a solder that does not contain lead (lead-free soldering) as part of measures to reduce environmentally hazardous substances in recent years. This is a huge switch for electronic products that have been manufactured with lead soldering for many years (Japan Institute of Electronics Packaging, 2006). Going lead-free has also changed the solder melting point and wettability. Various countermeasures were required in addition to material improvement, such as improving component heat resistance, component plating, circuit board artwork (AW) design, production equipment, and process controls. In the home electronics industry, the switch to lead-free has rapidly progressed because of the Restriction of Hazardous Substances (RoHS) Directive effective from July 2006 (Official Journal of the European Union L 234/44, 2011). The automotive industry is similarly subject to the End-of-Life Vehicles (ELVs) Directive, in which the European Commission has set lead-free requirements for printed circuit boards in new vehicles to be sold from 2016 (Official Journal of the European Union L 85/3, 2011) and the automotive industry is starting to take action.

2 ECU PRODUCTION TECHNOLOGY AND QUALITY ASSURANCE

Using the example of an engine ECU among the various types of ECUs, this section summarizes an example of the manufacturing process for an ECU with a resin-printed board, and an example of the manufacturing process for a hybrid ECU with a ceramic substrate to be utilized in harsh environments such as direct mounting on the engine. Approaches to and methods for production technology and quality assurance from a production standpoint and examples of countermeasures against part variations are illustrated (Kato, 2010b; Japan Institute of Electronics Packaging, 2006).

2.1 Manufacturing process for printed board ECU

The manufacturing process for a printed board ECU is explained in the following text. The engine ECU shown as an example in Figure 1 is configured from approximately 500 parts, including a microprocessor, transistor, resistor, capacitor, connector, resin-printed board, case, and cover. The manufacturing process can be roughly divided into a mounting process and a subsequent assembling process. Figure 2 shows an outline of the processes.

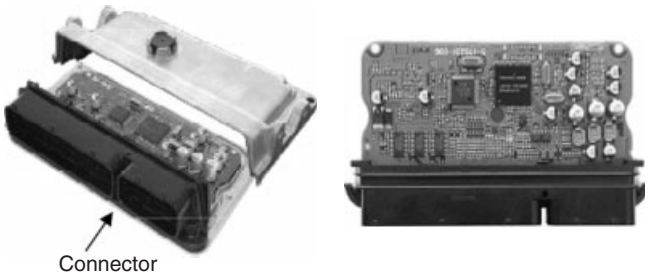


Figure 1. Example of a printed board engine ECU.

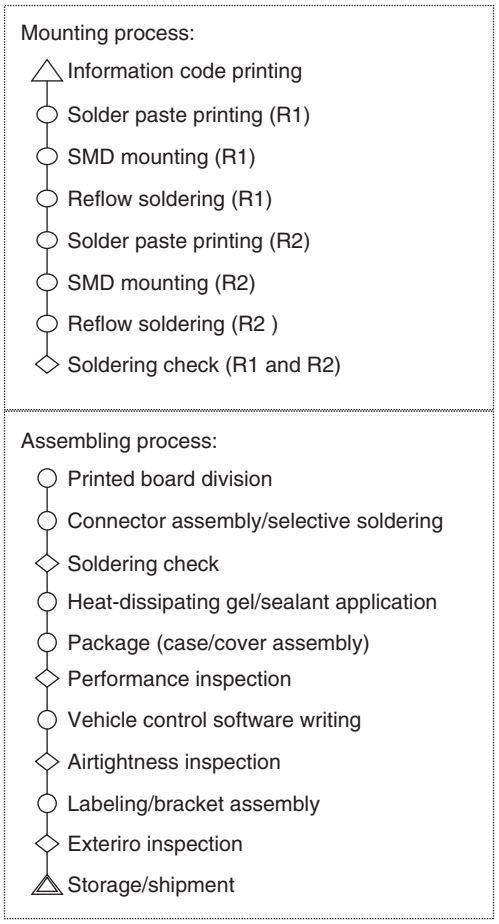


Figure 2. Example of the manufacturing process for a waterproof engine ECU.

Features of the automotive engine ECU manufacturing process include a selective soldering process because the engine ECU has a large connector that is a through hole device (THD), and include a high reliability inspection process. An ECU without a waterproof structure or with certain mounting specifications also requires an additional process for coating the entire printed board or required

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto216
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

sections with a drip-proof material to resist condensation and other humid environments. A detailed description of this process is omitted here as the main focus is on the manufacturing process for an ECU with a waterproof case structure.

2.1.1 Mounting process

Soldering is performed to mount electronic components on a printed board. To meet demands for reduced size and weight, surface mounting technology (SMT) is mainly used to mount the electronic components on both surfaces of the printed board (Japan Institute of Electronics Packaging, 2006).

2.1.1.1 Information code printing. An information code such as QR (quick response) code (ISO/IEC18004, 2006), which is a two-dimensional barcode, is used to print information such as the serial number and type number on the printed board with a laser. The information is utilized as automatic setup information for equipment and traceability, and within the assembly process for each product.

2.1.1.2 Solder paste printing. A pastelike solder material for bonding electronic components is printed on the printed board using a screen printer (Figure 3).

2.1.1.3 SMD mounting. Various types of surface mount devices (SMDs) are mounted on the printed solder paste material by a mounting machine. The code information is read to set the mounting program that corresponds to the product type and automatically perform mounting. Multiple suction nozzles are moved in succession to mount several tens of electronic components per second at high speed (Figure 3). Equipment that requires replenishing of electronic components is automatically displayed on a board, and a worker supplies the components. This operation takes advantage of the barcode information and prevents mistakes in the supply of electronic components.

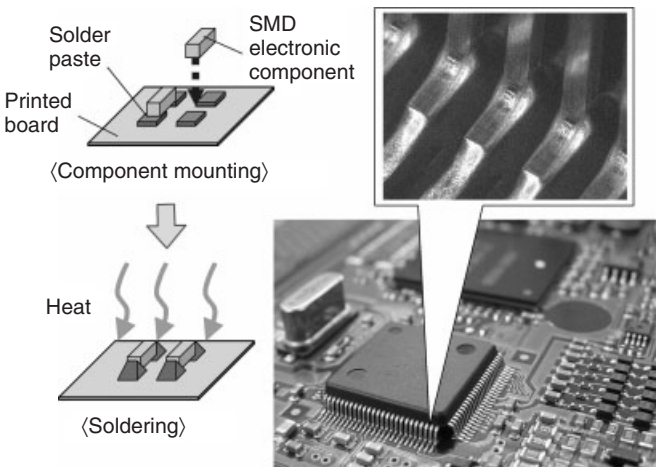


Figure 3. Solder printing, SMD mounting, reflow soldering.

2.1.1.4 Reflow soldering. Soldering is performed by melting the solder paste at approximately 230°C in a reflow (heating) oven (Figure 3). The temperature of the entire circuit board is controlled to maintain a uniform temperature and ensure a temperature that is below the upper temperature limit of the electronic components and that also melts the solder on all the components with different heat capacities. This temperature range has narrowed because of recent lead-free countermeasures, which has led to stricter temperature controls. This process completes the mounting of electronic components on one surface of the printed board. Once mounting on one surface is complete, electronic components are mounted on the other surface using the same process described earlier.

2.1.1.5 Soldering check. Next, the appearance of all soldered locations (roughly 1500 points) and the mounting condition of the components are automatically checked by an automatic viewing (optical inspection) device (Figure 4).

The principle behind the method of confirming the solder shape in a color image is explained here. The component and solder condition are judged based on the condition of this image. The component mounting process is subject to a meticulous process control to ensure that defects do not occur. However, as several hundred tiny electronic components are soldered all at once, there are rare cases of abnormal bonding as shown in Figure 4 (IPC-A-610E-2010, 2010). Therefore, all electronic components are checked so that defects are not passed on to downstream processes. The results, after the check, are fed back and used in the process control for the mounting process and to make further improvements. The checked printed board is then transferred to the subsequent assembling process.

There have been strong demands in recent years for higher functioning and smaller automotive ECUs, as well as smaller semiconductor packages with more pins. The commonly used quad flat package (QFP) form cannot satisfactorily meet these demands, and the ball grid alley (BGA) has been increasingly adopted. With this package, the solder-bonded sections are not visible after mounting on the circuit board, which means that inspection with the optical automatic viewing device used earlier is not possible. As a consequence, even tighter controls and even

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto216
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

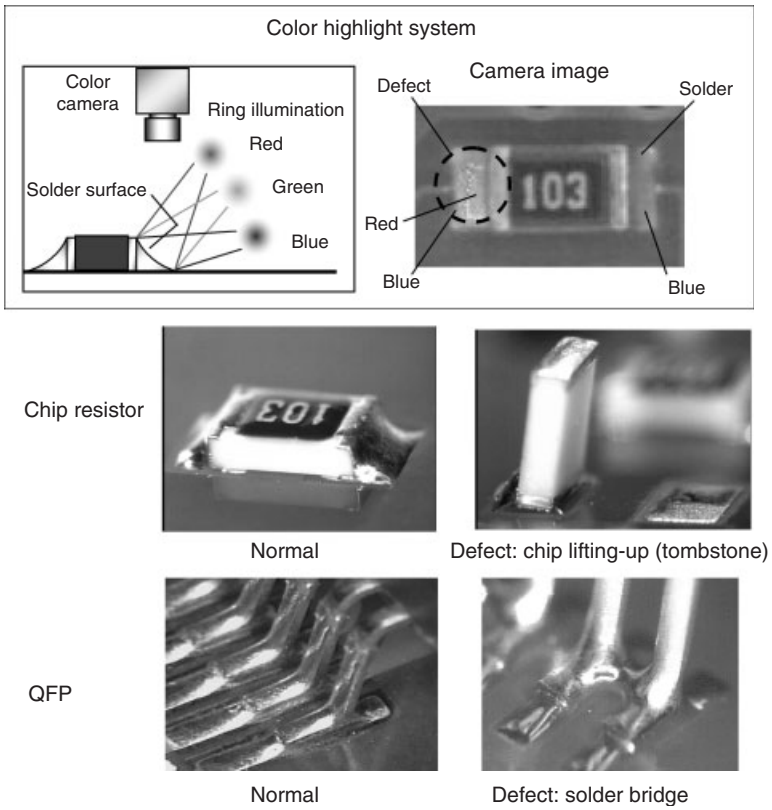


Figure 4. Soldering check by automatic viewing device.

better built-in quality from the component manufacturer up to the mounting process are required. After solder paste printing in the mounting process, an automatic viewing device is used to inspect that the solder paste has been normally printed. In addition, the solder-bonded sections may be checked by X-ray in the initial production as well as in the prototype stage (Figure 5).

2.1.2 Assembling process

2.1.2.1 Printed board division. In the mounting process, one sheet has multiple boards and the boards are cut and divided by a router or the like.

2.1.2.2 Connector assembly/selective soldering. The connector is inserted and assembled to the through hole of the printed board. For automotive ECUs such as the engine ECU, this connector alone is often a THD that is joined using a through hole. Also, as there are many pins, selective soldering is often used to enable soldering of all pins at once (Japan Institute of Electronics Packaging, 2006). First, the printed board is coated with flux to facilitate soldering. Next, it is preheated and then is on a soldering nozzle, after which melted solder is injected from below to perform selective soldering. The amount and range of the flux coating, the solder immersed state, and so forth greatly influence bonding, and process conditions are precisely set and tightly controlled for each type (Figure 6).

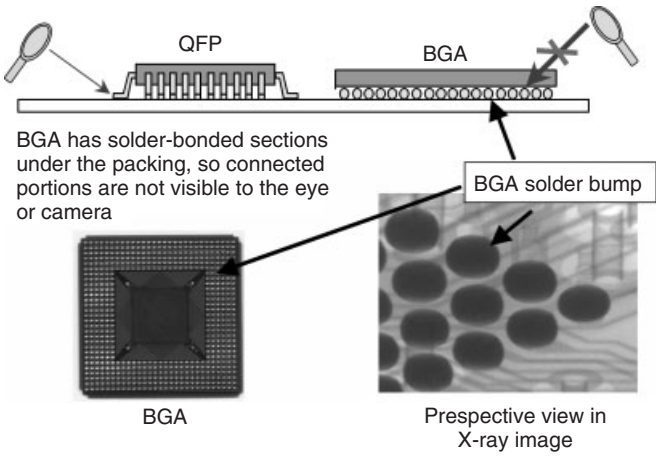


Figure 5. BGA and X-ray check.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto216
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

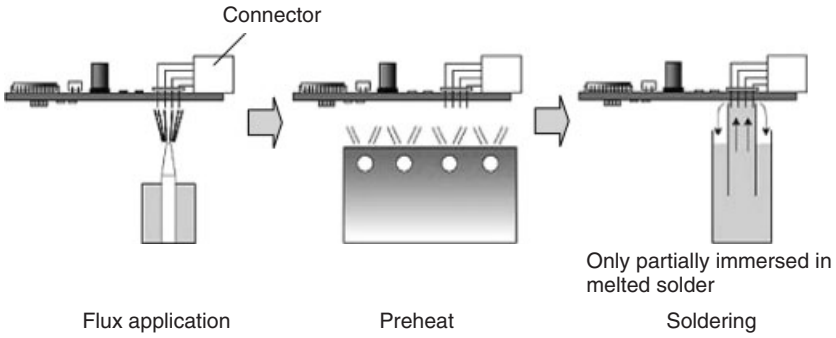


Figure 6. Selective soldering.

2.1.2.3 Soldering check. Similarly to the mounting process, the solder condition is checked by an automatic viewing device.

2.1.2.4 Heat-dissipating gel/sealant application. Heat from the electronic components must be dissipated to prevent excessive heating of the electronic components. Therefore, a heat-dissipating gel (silicone) is quantitatively applied to correspond with the layout of the electronic components on the case. To prevent water from entering the ECU interior, a sealant (silicone) is applied for hermetically sealing the outer periphery of the connector and case (Figure 7).

2.1.2.5 Package (case/cover assembly). A case and cover are attached to the printed board to package the printed board (Figure 8).

2.1.2.6 Performance inspection. To check the several hundred electronic components, as well as their assembling quality and operation as an ECU, a performance inspection is performed in which quasi-signals such as vehicle speed signals and engine speed signals are input to determine whether a set output signal can be acquired (Figure 9).

Inspections at high and low temperatures that estimate the temperature environment when mounted in a vehicle may be performed to detect any defects in the temperature characteristics of the components.

2.1.2.7 Vehicle control software writing. Programs for controlling the vehicle are automatically written based on ECU type. The programs are written in time to create an ECU capable of engine control.

2.1.2.8 Airtightness inspection. Hermetic performance is inspected to check whether the sealant applied for a

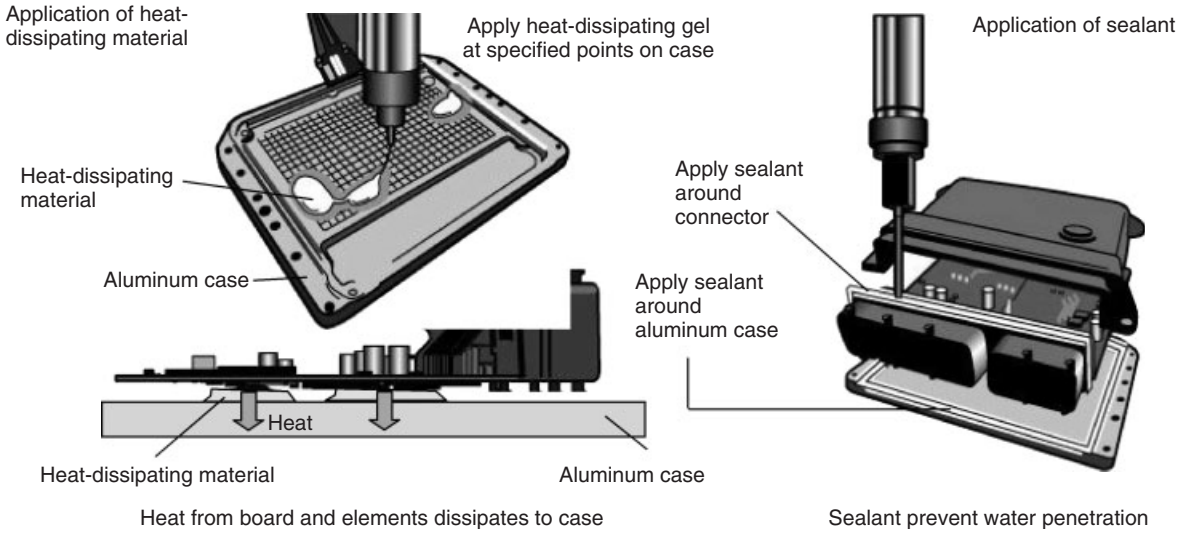


Figure 7. Heat-dissipating material, sealant application.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto216
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

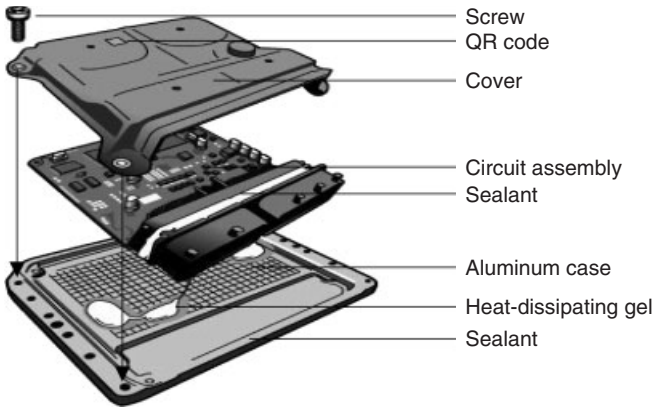


Figure 8. Package (case/cover assembly).

hermetic seal can withstand the installation environment in the vehicle.

2.1.2.9 Labeling/bracket assembly. Labels for each type are printed and attached. Brackets for attachment to the vehicle are assembled.

2.1.2.10 Exterior inspection. The exterior of the fully assembled product is inspected, including the name, labels, and the like. This completes production of the ECU.

2.1.2.11 Storage/shipment. Following completion of the inspection, the product is matched to an order and shipped in accordance with the order.

The preceding process and production technology examples are only brief summaries, and the processes may be differently ordered, omitted, or added to, depending on the product.

As mentioned earlier in one of the processes, each process uses the information code for the production progress management of each product, process controls, automation of in-process switching of product types, and the like. A traceability system is also employed so that history information can be traced to ensure thorough quality control. Although various inspections as mentioned earlier are implemented to prevent the outflow of any products with defects, quality is built into each process as a general rule.

For example, process capability is normally evaluated using the process capability index Cp (Cp = standard width/6σ, where σ is the standard deviation). However, the product data average may deviate from the standard median, and Cpk (Cpk = (data average − limit)/3σ) must be considered in this case. Cp and Cpk are used in combination to evaluate processes and implement necessary improvements. To ensure that excessive stress is not applied to electronic components or bonded portions during processing and assembling, the strain applied to the product in each process is evaluated and the positional relationships between components are reflected in the board design. Necessary improvements are also made to the processes themselves. Specific examples include the screw position and the locations of board divisions.

This section introduced some activities to ensure quality. As described earlier, stable manufacturing and supply is only possible after first building high-quality processes.

2.2 Manufacturing process for hybrid ECUs

Figure 10 shows an example of a hybrid ECU that uses a ceramic substrate. The ECU is directly mounted on the engine and used as an engine ECU. The ECU is configured from approximately 150 parts, including a microprocessor, transistor, resistor, capacitor, connector, ceramic substrate, case, and cover (Kato, 2010b).

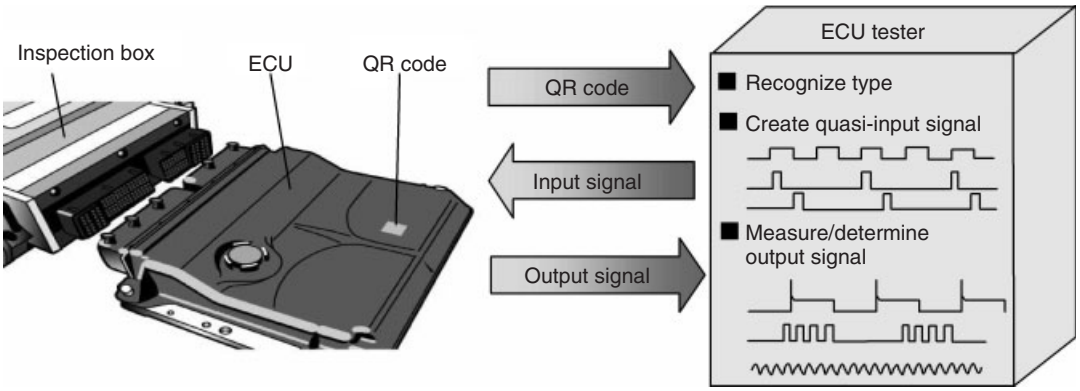
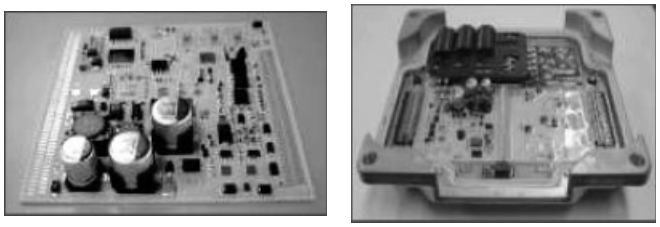


Figure 9. Performance inspection.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto216
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) (b)

Figure 10. Example of hybrid ECU directly mounted on engine. (a) Photo of exterior of components mounted on ceramic substrate. (b) Exterior of ECU product.

Directly mounting the ECU on the engine requires high heat resistance and a small size. To increase heat resistance, a ceramic substrate is used for the circuit board, and semiconductors are mounted on the circuit board by bare chip mounting.

The production technology for a hybrid ECU is outlined in the following text. Figures 11 and 12 illustrate examples of the process flow.

First, a solder material is applied to required locations on the ceramic substrate using a screen printing method. Parts such as semiconductor bare chips and chip capacitors are mounted on the solder material, and reflow soldering is performed at approximately 230°C. Flux contained in the solder material is subsequently cleaned. Fluorocarbon cleaning solvents have been used in the past as the cleaning solution, although water-based cleaning solutions are now employed in consideration of the global environment. Semiconductor bare chip components are mounted with an electrically conductive adhesive, after which gold (Au) wire bonding and aluminum (Al) wire bonding are performed to create electrical connections.

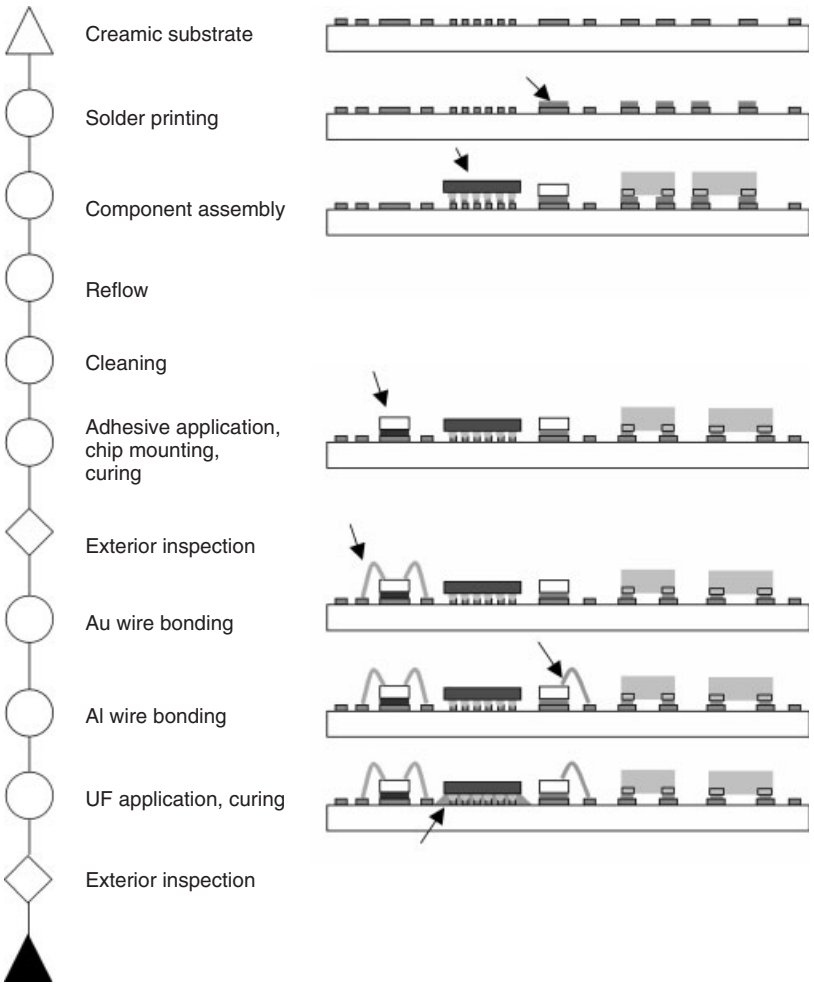


Figure 11. Example of manufacturing process flow for hybrid ECU (circuit board process).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto216
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

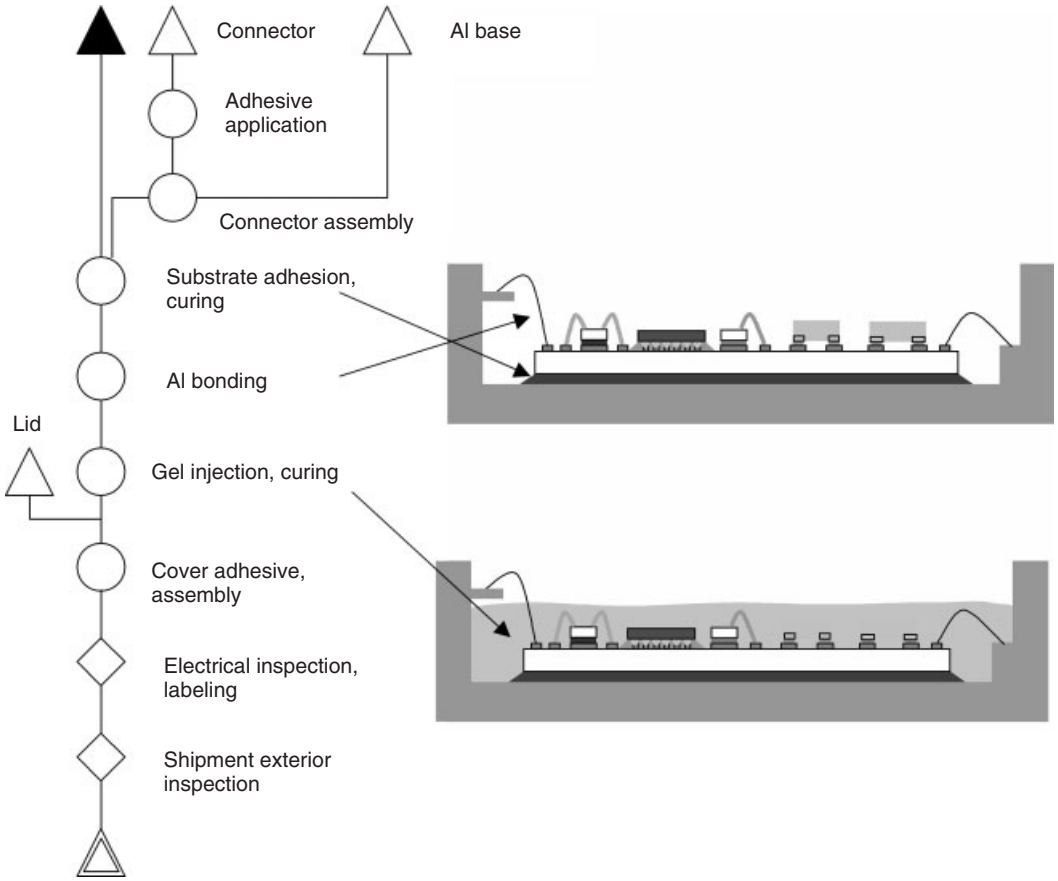


Figure 12. Example of manufacturing process flow for hybrid ECU (assembling process).

To improve the reliability of the solder-bonded portions of some components, an epoxy-based resin material [underfill (UF) material] is applied to reinforce the soldering. This completes the manufacture of one circuit board. An exterior inspection is performed at this stage, and the circuit board is then transported to the downstream process. Note that, although not described in detail here, the ECU shown in Figure 10 is configured from three boards.

The circuit board with assembled components is closely attached using adhesive to components to which a base made of metal (generally Al) and connector components are already adhered. Next, these are covered with a potting material (silicone-based gel) to protect the semiconductors and bonding wires, and a lid is then adhered. Following an electrical inspection, the product is shipped.

2.3 Comparison of hybrid ECU and printed board ECU

This section briefly compares the hybrid ECU that uses a ceramic substrate and the commonly used printed board ECU, while focusing on the merits of the hybrid ECU in particular. The hybrid ECU is advantageous in terms of having a small size, as well as good heat resistance and environment resistance (Table 1).

From the standpoint of downsizing, the hybrid ECU has an edge because of easy bare chip mounting. The printed board ECU uses molded ICs including QFPs, and the practical area required for mounting these components is, therefore, approximately five times larger than that needed for bare chip mounting. In addition, the hybrid ECU can adopt structures well suited for downsizing, for example, direct printing of resistors on the substrate and full-thickness via holes.

With regard to heat resistance, the hybrid ECU uses a ceramic substrate, that is, a sintered compact, and is extremely stable with respect to heat. Ceramic components mounted on the substrate have a matching coefficient of thermal expansion for reduced stress on connected regions and high reliability. The environment resistance of the hybrid ECU enables installation in high-temperature regions subject to large vibrations.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto216
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Comparison of hybrid ECU and printed board ECU

			Hybrid ECU	Printed board ECU
Downsizing	IC		Bare chip	Molded IC
	Resistance		Direct printing resistor + chip resistor component	Chip resistor component
	Board design		Via hole	Through hole + via hole
Heat resistance	Board glass-transition point		Approximately 1600°C (firing temperature)	Approximately 140°C (FR4: glass epoxy board)
	Coefficient of thermal expansion	Board	7 ppm/°C	16 ppm/°C
		Components	6–9 ppm/°C	6–9 ppm/°C
Environment resistance	Use environment		Direct mounting on engine	Mounting in occupant compartment or engine compartment
	Vibration resistance		30G	5G

3 SUPPLY CHAIN MANAGEMENT

Automotive ECUs are configured from a number of materials and parts, including electronic components such as microprocessors, transistors, resistors, and capacitors, as well as printed boards, connectors, cases, covers, solders, sealants, screws, and brackets. In addition, those individual materials and parts are composed of raw materials and other parts and materials. Their quality and supply support ECU production, and the management, that is, supply chain management (SCM), of the material and part supply network is extremely critical to achieve high-quality and stable production.

In the automotive industry, consolidation of the supply of raw materials and parts has advanced thus far based on the pursuit of lower costs and higher quality. Partial modularization has similarly progressed. Furthermore, there are parts that demand lower costs and higher quality of both the ECU and components, which can only be achieved by surveying the supplier processes and the entire ECU production process, and considering all processes and inspections as a whole. Instead of a pyramid supply

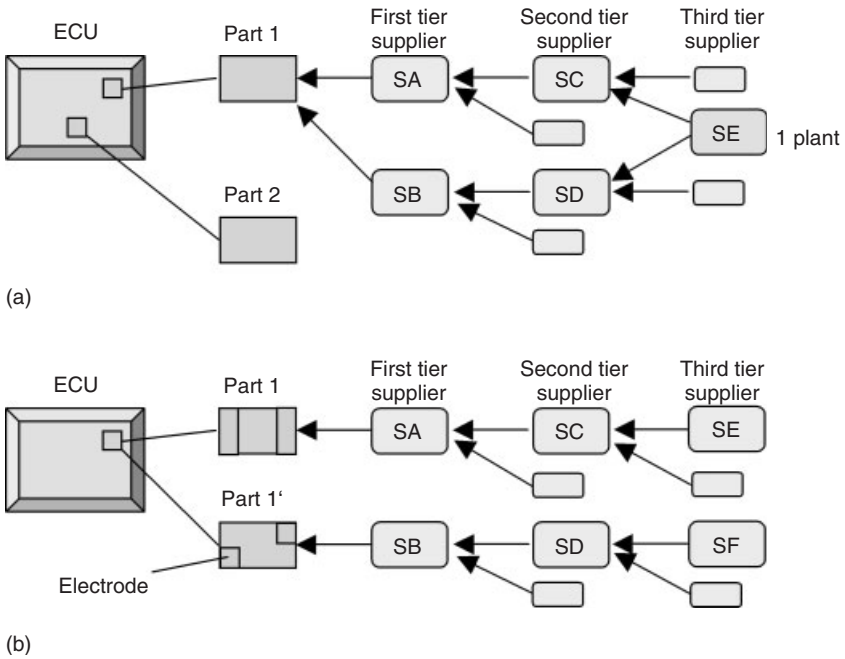


Figure 13. Supply chain. (a) Example of common raw material supplier. (b) Example of different suppliers and electrode geometry.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto216
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

configuration, it may be discovered that some materials or parts are restricted to a limited number of suppliers.

Supply risk management has become extremely important because of large-scale disasters, company risks, and other factors and is gaining more attention. As the ECU is configured from a large number of materials and parts as described earlier, it cannot be produced if the supply of even one part is interrupted. Although there are parts that can be easily substituted or are not directly linked to product competitiveness, materials and parts that sustain the competitiveness of the product in terms of quality, performance, and cost cannot be easily replaced. For this reason, restructuring and visualization of the supply chain, particularly in consideration of risk management, is necessary.

An in-depth and comprehensive examination of the supply chain must be performed. For example, as shown in Figure 13a, in the case of two first-tier suppliers, there should be no problem with the two first-tier suppliers purchasing materials and parts to manufacture their own parts from different second-tier suppliers. However, the fact that an underlying third-tier supplier supplies raw material to both second-tier suppliers and also has only one plant may be overlooked. If the third-tier supplier is affected by a disaster, the supply of raw material could be stopped. Moreover, this could happen again at any time in the future.

Parts with the same exterior shape and function cannot necessarily be easily interchanged. For example, as shown in Figure 13b, even parts with the same exterior shape and functions may have different electrodes. In this case, the geometry of the electrodes (lands) on a printed board may correspond to one part but not the other. If the other part is mounted to the noncorresponding printed board and soldered, a mounting defect such as part rotational displacement may occur. If possible, the geometry of the electrodes of printed board (lands) should be designed in advance so that either part can be mounted to printed board without any problems. This involves AW design technology, and in a wider sense, production technology.

Simply prioritizing risk avoidance by increasing inventory results in unnecessary costs and reduces product competitiveness. Instead of thinking strictly in terms of the quality, cost, and supply of materials and parts, the supply chain should be considered and strategically studied at the stages of design, part development/selection, and material development/selection.

4 CONCLUSION

There are ECUs with various functions, specifications, forms, and structures and appropriate production technology for each ECU is selected considering required quality and cost. This chapter takes the engine ECU as a representative example and gives an overview of production technology for printed board ECUs and hybrid ECUs made of ceramic substrate. It also summarizes approaches and methods for quality assurance taken during production of these.

It is important to ensure stable supply performance and quality over the entire supply chain, starting with material and part suppliers to the manufacturer itself and the users. Supplier chain management is outlined also in view of significant supply issues that have surfaced in recent years.

In the near future, integration of the ECU and its control object, that is, electro-mechanical integration, will increasingly progress. Further development of connection technology and size reduction technology will be required, together with improved production technology to achieve these goals.

RELATED ARTICLES

ECU Technologies from Components to ECU Configuration
ECU Design and Reliability
Engine ECU Systems

REFERENCES

IPC-A-610E-2010 (2010) Acceptability of electronic assemblies, association connecting electronics industries.

ISO/IEC18004 (2006) Information technology—automatic identification and data capture techniques—QR code 2005 bar code symbology specification..

Japan Institute of Electronics Packaging (2006) *Handbook printed circuit technology*, 3rd edn, Nikkan kogyo Shinbun, Tokyo.

Kato, M. (2010a) *Automotive Electronics: Systems*, Nikkei Business Publications, Inc., Tokyo.

Kato, M. (2010b) *Automotive Electronics: Basic Technologies*, Nikkei Business Publications, Inc., Tokyo.

Official Journal of the European Union L 234/44 (2011) Commission Decision of 8 September 2011, amending, for the purposes of adapting to technical progress, the Annex to Directive 2002/95/EC of the European Parliament and of the Council as regards exemptions for applications containing lead or cadmium, September 10, 2011.

Official Journal of the European Union L 85/3 (2011) Directives, Commission Directive 2011/37/EU of 30 March 2011, amending Annex II to Directive 2000/53/EC of the European Parliament and of the Council on end-of-life vehicles, March 31, 2011.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto216
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Microcomputers and Related Technologies: Enlargement of Software Size, Algorithms, Architectures, Hierarchy Design, Functional Decomposition, and Standardization

Hideaki Ishihara
DENSO Corporation, Kariya, Japan

1 Introduction 1
2 Microcomputers 1
3 Software 6
4 Adoption of Electronic System Platforms 10
5 International Standardization (OSEK and AUTOSAR) 11
6 Summary and Conclusion 12
References 13

1 INTRODUCTION

Automotive electronics have been advancing rapidly in the successful pursuit of environmental friendliness, safety, comfort, and convenience. These advancements have been achieved through the adoption of microcomputers, which now total to approximately 100 in each luxury car. Since the late 1970s, microcomputer technologies have achieved enhanced performance, affinity with real-time processing, high fault-tolerance, and the like. However, the consequent enlargement of software size has resulted in the need not only for new software architectures such as hierarchical design by functional decomposition but also for international standardization [such as OSEK (open systems and the corresponding interfaces for automotive electronics) and AUTOSAR (automotive open system architecture)]. In the twenty-first century, there is public awareness of the need for the creation of more green and dependable automotive systems, driving the need for the reduction of CO_2 emissions. There is also a general interest in achieving high road safety levels. From these points of view, this chapter describes the histories and future directions in regard to both microcomputers and software in automotive electronics, while considering semiconductor technologies and their applications.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 MICROCOMPUTERS

2.1 History of automotive microcomputer applications

The automobile and the microcomputer are two of the greatest developments of the twentieth century. The automobile enabled universal freedom of movement and the breakthrough of the microcomputer-enabled complex functions conceived by people and plays a main role in modern systems. Microcomputers have come to be used in almost every industry. This was triggered by the development of the integrated circuit (IC) after the invention of the transistor by John Bardeen, Walter H. Brattain, and William B. Shockley at Bell Laboratories, which had the effect of turning large computers into compact and portable

devices. Present microcomputers generally contain not only microprocessors but also input/output circuits and memory devices all within a single semiconductor chip. A microprocessor consists of an arithmetic unit and a control unit, which functions as a central part of microcomputer circuits. The terms *central processing unit* (*CPU*), *processor*, and *micro processing unit* (*MPU*) are also used as general synonyms for a microprocessor. In 1971, Intel Corporation created the 4004, which was the world's first commercially available microprocessor. It had a 4-bit processing capability and a clock speed between 500 and 741 kHz. A total of 2300 transistors were integrated on a chip with an area of 3 mm × 4 mm and it was manufactured using 10-μm semiconductor processes. The 4004 was developed as a joint venture between a Japanese electronics engineer called *Masatoshi Shima* and Intel Corporation based on a request from the Japanese company Busicom for use in an electronic calculator. Subsequently, microprocessor bit lengths evolved from 8 to 16 bits. From the beginning of the 1990s, 32-bit microprocessors also started to become available in certain embedded systems (Kato *et al*., 2010).

When implementing an electrical function, a key difference between the microcomputer and the hard logic is the length of the development period between the software and the hardware. Software functions can be changed in a period from several hours to several days. However, hard logic takes at least several months (i.e., at least 100 times as long) to change as different semiconductors have to be prototyped and evaluated. Furthermore, the number of software designers can be increased substantially at short notice as software education and training is not subjected to serious physical limitations. These are the main reasons why microcomputers have spread so rapidly in vehicles (Kato *et al*., 2010).

The operating voltage of microcomputers was generally 5 V, whereas the semiconductor manufacturing process was 0.5 μm or larger. Subsequently, since the beginning of the twenty-first century, the operating voltage has been decreasing, reaching 3.3 V at 0.35 μm, 2.5 V at 0.25 μm, and approximately 1 V at 90 nm. However, the operation voltage of most peripheral functions embedded in microcomputers remains 5 or 3.3 V for reasons related to sensor and actuator interfaces.

As automotive electronics have evolved, the number of microcomputers has increased to dozens in an ordinary passenger vehicle and more than a hundred in a luxury vehicle. Unlike personal computers, automotive microcomputers generally use a single chip in which the input and output circuits are integrated onto a semiconductor chip. This is to satisfy the various performance requirements of a vehicle, such as the built-in memory capacity, high real-time capabilities, strong electromagnetic compatibility (EMC), and wide operating temperature and voltage ranges. This has been enabled by the explosive evolution of semiconductor IC technology in the second half of the twentieth century. The aim of automotive microcomputers is not simply to achieve the basic moving, turning, and stopping performance aspect of the vehicle. Microcomputers are also installed in core portions of safety devices such as airbags, millimeter wave radar, and the like, antitheft, power window, and other body control devices, and intelligent transport systems (ITS) devices such as electronic toll collection (ETC), and navigation systems. These functions can now be cooperatively controlled through the local area network (LAN) (Ishihara, 2009; Tsuda, 2007).

The first use of a microcomputer in a vehicle was for engine control, prompted by the introduction of stricter emissions regulations. Microcomputers were adopted by General Motors (GM) in 1976 and by Ford in 1977 for engine controls, such as ignition timings. The automotive industry in Japan was also researching microcomputers in engine controls at virtually the same time. In 1980, Nippondenso Co., Ltd. (the current Denso Corporation) adopted a single-chip microcomputer-based engine control with a 12-bit microprocessor and an 8-input interrupt function. These microcomputers had superior processing capabilities, real-time performance, and operating temperature range than the 8-bit microcomputers provided by Intel at that time. For a long time after then, the bit length of automotive microprocessors became a logical target for developers. However, 8-bit microprocessors became the predominant microcomputer since the second half of the 1980s because of their sufficient performance for all aspects of engine control by establishing a 16-bit address range capable of pointing to locations in memory. Although mainstream modern microprocessors are 32-bit, some 8-bit devices are still used in low-cost systems (Kato *et al*., 2010).

Microcomputers entered widespread use over the second half of the 1980s. Figures 1 and 2 show two examples: the 1-bit microcomputer made by Denso for low-end electronic systems and the 8-bit microcomputer (ND8) made by Denso with a multithread structure to avoid runaway.

2.2 Requirements of automotive microprocessors

Non-multimedia automotive control systems require *hard real-time performance*. This refers to an absolute permissible maximum run time for individual software within the application. For microprocessors such as personal computers, the cache and other means are effective ways of

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

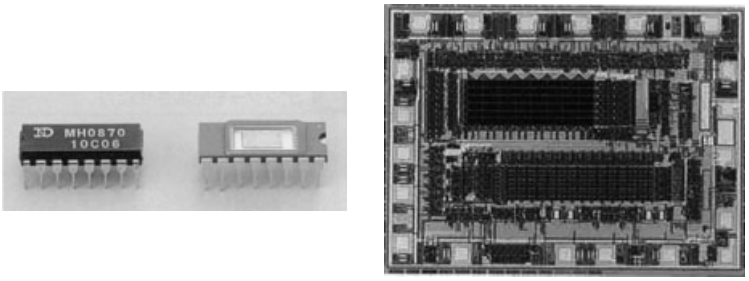


Figure 1. 1-Bit microcomputer made by Denso. (Reproduced by permission of H. Ishihara/Denso. © Denso Corporation.)

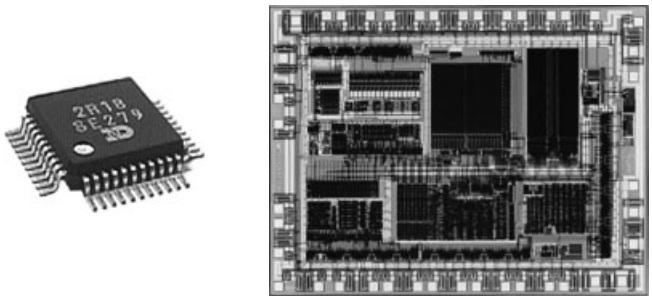
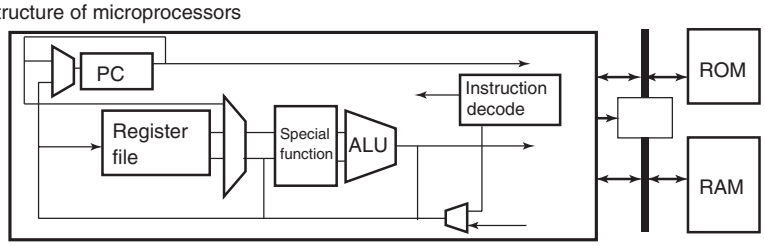


Figure 2. 8-Bit microcomputer (ND8) made by Denso. (Reproduced by permission of H. Ishihara/Denso. © Denso Corporation.)

improving performance. However, these do not satisfy the hard real-time requirements and therefore cannot be used in automotive microcomputers. Instead, memory devices are embedded in the chip to improve performance. As there is a limit to the capacity of memory that can be embedded in a chip, one key item for selecting a microcomputer is the efficiency of translating the software to machine code. This is also known as an *evaluation item* called *code size efficiency* (Figure 3) (Ishihara, 2007, 2009).

In 1995, Denso developed the Nippondenso reduced instruction set computer (NDR) machine, a 32-bit reduced instruction set computer (RISC) optimized for automotive controls. Using the RISC architecture of a general-purpose register machine (Figure 4), the development closely emphasized the relationship between the instruction set architecture and C compiler to enhance affinity with automotive control software. As a result, it was possible to greatly improve the code size efficiency in comparison with other microprocessors (Figure 5) (Ishihara, 2007; Kawamoto *et al*., 2001).

Structure of microprocessors



(a)

Evaluation items

Evaluation items		Cost	Real time	Reliability	Power consumption
Technology	Small gate count (compact design)	✓			✓
	Fast response to interruption		✓	✓	
	Code size efficiency Special instruction and compiler optimization	✓ Memory size	✓	✓ Memory devices should be embedded in the chip	✓

(b)

Figure 3. (a) Structure and (b) evaluation items of microprocessors.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

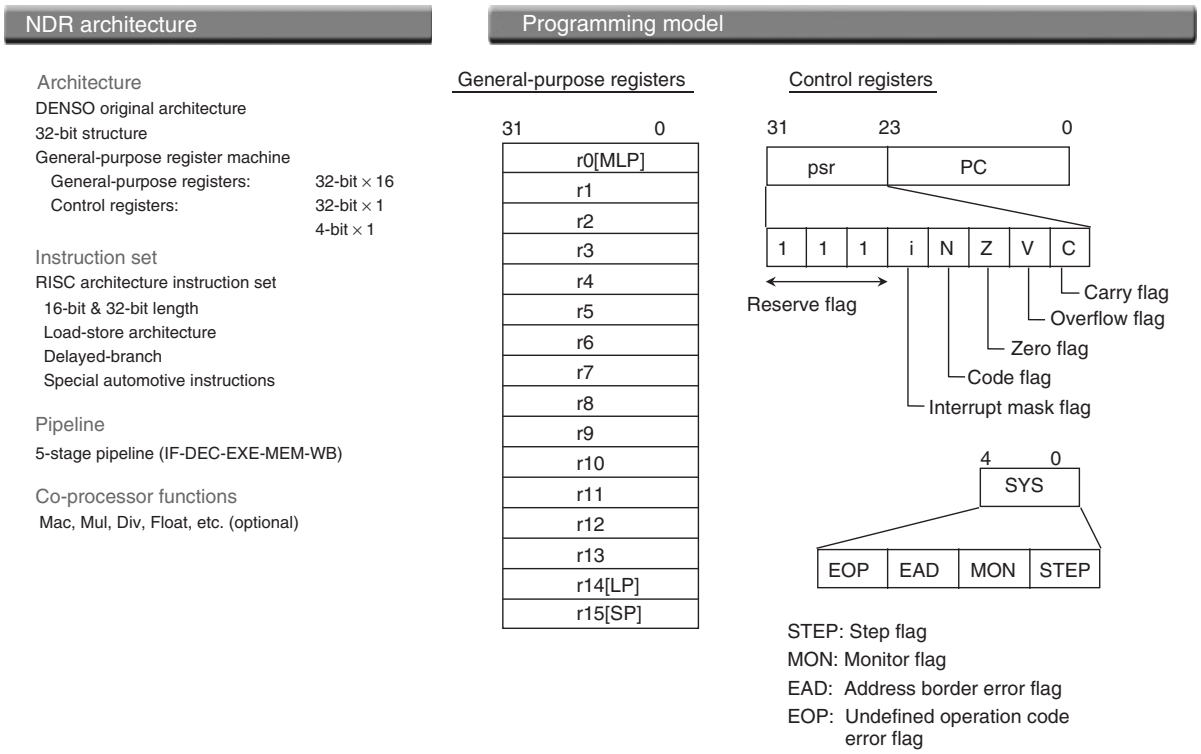


Figure 4. RISC architecture of general-purpose register machine.

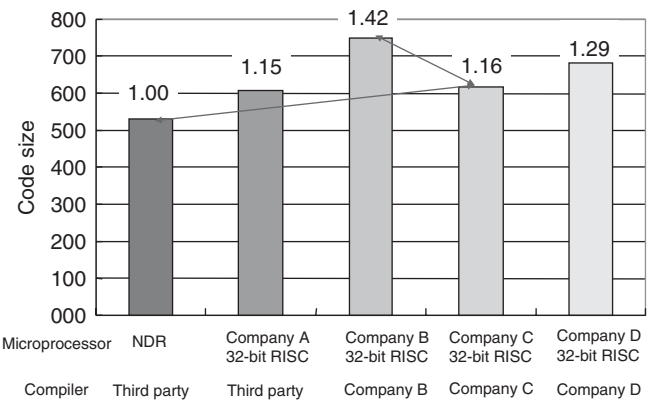


Figure 5. Affinity with automotive control software.

2.3 Internal structure of single-chip microcomputer

Figure 6 shows an example microcomputer structure.

In addition to the microprocessor, the integration of functions for special applications on single chips include embedded memory, interrupts (INT), timers, inputs [analog-to-digital converters (ADCs) and input capture], outputs [pulse width modulation (PWM)], communication functions [universal asynchronous receiver-transmitters (UARTs), serial peripheral interfaces (SPIs), and controller area networks (CANs)], and the like. Timers are used for the time management of tasks that are indispensable for real-time performance. Cooperative operations with INT are used to start periodic tasks and to measure the interval between events. ADCs are used to convert analog values from sensors to digital values that can be used by the microcomputer. PWM is capable of generating output waveforms with a modulation protocol that changes the duty ratio of the pulse wave and is used in motor control and for simple data transmission. SPI is characterized by clock-synchronized communication and is used for communication between the multiple peripheral ICs on the ECU (engine control unit) circuit board. CAN is used for communicating between ECUs. Input capture is used to measure the pulse width and interval. Other embedded functions may also include the voltage regulator required to operate the microcomputer, low voltage reset, and a watchdog timer. Recent years have also seen greater demands for extremely compact implementation as microcomputers are now being embedded within sensors and actuators. As an example of this trend, Figure 7 shows the microcomputer embedded in a power window motor (Ishihara, 2009; Ookura, 2005).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

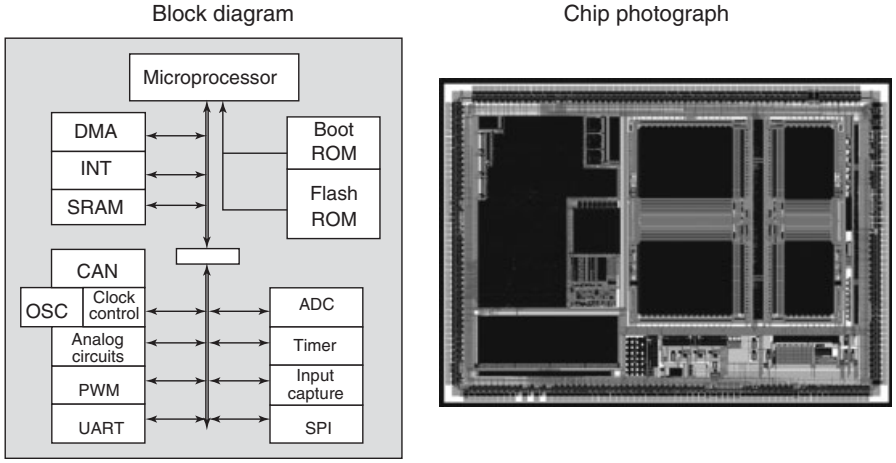


Figure 6. Example 32-bit microcomputer (body electronics).

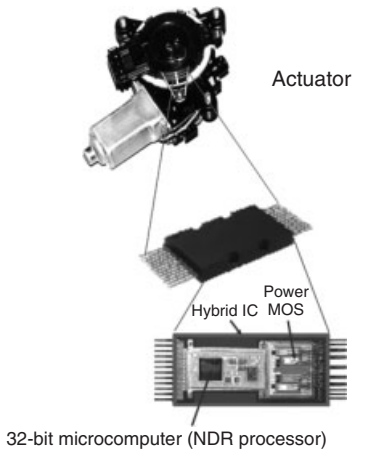


Figure 7. Example microcomputer (power window).

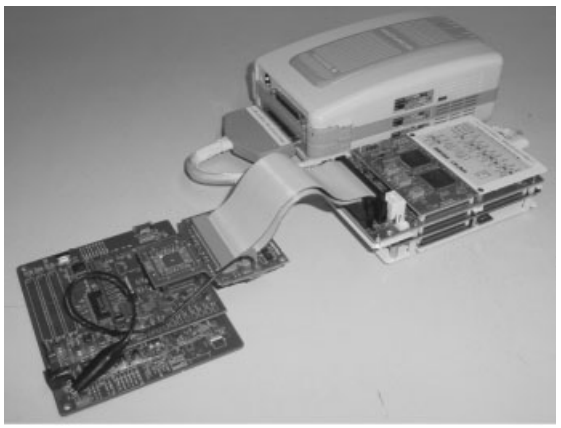


Figure 8. Example ICE.

2.4 Role of the software development environments

Use of the software development environment is an indispensable part of developing software to be embedded in microcomputers. When this software is developed in C language, the C programs are translated into assembly language by the C compiler and then further translated into machine language by an assembler and linker. The software is then written into the embedded memory of the microcomputer by a ROM (read-only memory) writer. The software is debugged by an in-circuit emulator (ICE) and/or RAM (random access memory) monitor, which confirms that the software operates according to expectations while monitoring the internal information of the microcomputer. Figure 8 shows an example ICE. As the integration density of semiconductors increases, the functions of the software development environments may be embedded into the microcomputer to confirm the software operation more efficiently.

2.5 Future prospects of automotive microcomputers

The amount of software is likely to increase dramatically in the future in response to requirements for greater environmental performance, safety, comfort, and convenience. Consequently, microprocessor performance will have to be enhanced to allow these processing requirements to be performed effectively. In the 30 years since 1980, microprocessor performance has improved by more than 1000 times. Moore's law is used by the world's semiconductor industries, computer industries, and research institutions

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

as a guideline for making predictions regarding semiconductor and computer technologies. Gordon E. Moore, the cofounder and Chairman Emeritus of Intel Corporation, first proposed the law in 1965. Since then, Moore's law has been widely discussed among researchers, resulting in its use as a guideline for the prediction of future trends in both semiconductor and computer industries. This rule of thumb indicates that the numbers of transistors on an IC will double approximately every 18 or 24 months. This has been expected to be a driving force behind the progress of microprocessor performance. However, future trends remain unclear. As current developments are approaching their limits with respect to Moore's law in semiconductor technologies, the field of computer engineering is trending toward multi-core technologies that increase the number of microprocessors on silicon. However, multi-core technologies compatible with the hard real-time processing requirements of vehicles is still at a nascent stage around the world (Ishihara, 2007, 2009).

Hard real-time refers to systems in which the processing value switches to zero immediately after a preset time restriction (i.e., a deadline) is not satisfied. These are different from soft real-time systems in which processing values gradually fall over the course of time (such as personal computers and other information devices). In other words, automotive electronics resembles those used in the aviation field and are complex temporal protection systems. However, as the current focus of multi-core technology is on increasing average and optimum performance, these systems may even crash unless the minimum performance is guaranteed. In addition, as the number of microprocessors on a chip has increased, communication between microprocessors has also increased, which makes satisfying the fundamental temporal protection requirements unpredictable. Consequently, further technological innovation is strongly required. Figure 9 shows an example of a multi-core microcomputer with enhanced temporal protection performance (Ishihara, 2007, 2009).

Furthermore, many multi-core technologies are being researched that increase the number of processors on silicon from several tens to several hundreds. This is reputed to be generally suitable for applications that require high parallel performance, such as image processing. Multi-core technologies featuring more than 100 processors have already been applied for image recognition in automotive electronics.

In contrast, automotive microcomputers are trending strongly toward single-chip solutions. However, as technologies in ICs become increasingly advanced, it will be more difficult to integrate heterogeneous devices. Additionally, system in package (SiP) technologies that use through-hole vias are becoming more widely adopted. In the future, when the adoption of single-chip technologies is difficult or small amounts of a wide range of products are required, more fields are likely to request SiP solutions that combine standardized semiconductor chips.

3 SOFTWARE

Software is a key component of automotive systems. This section describes the fundamental details and characteristics of software for automotive electronics.

3.1 Role and positioning of software

The shape and form of software is invisible even under close examination of the ECU, which is the main component of automotive electronics. The software or programs are located within the internal ROM and RAM of the microcomputer. These are interpreted by the microprocessor and comprise rows of instructions to be operated and data

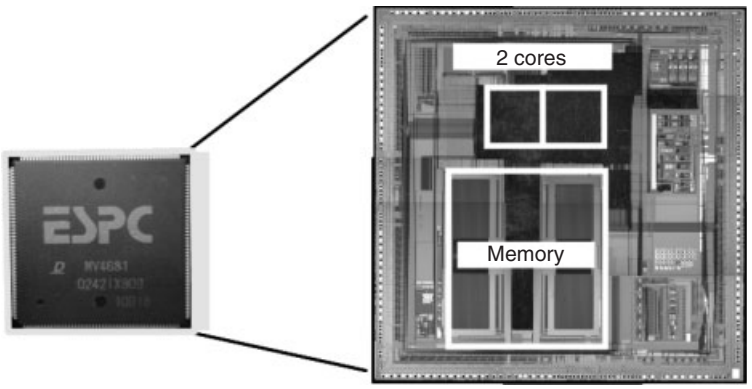


Figure 9. Example multi-core.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

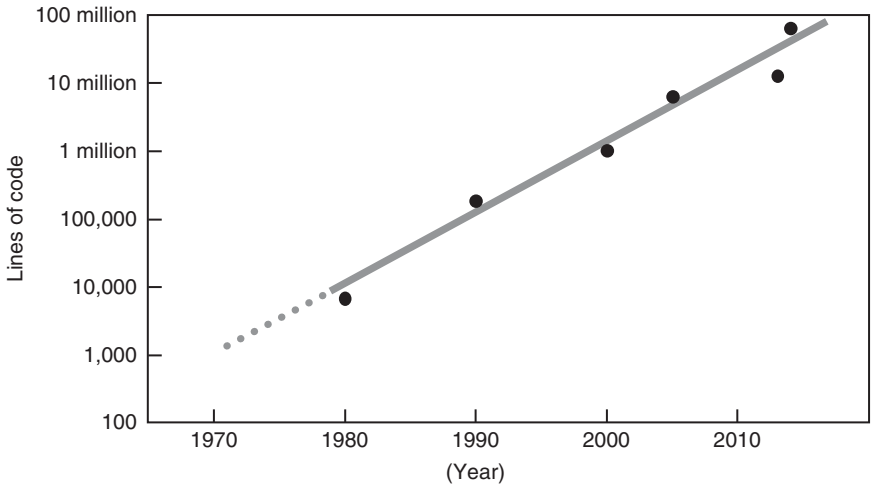


Figure 10. Trends for software installed per vehicle .(estimated from trends of Denso and other companies)

groups that represent the operation target. Software is the language used for comparison with the hardware and it defines how the hardware functions.

3.2 Increasing amount of software in automotive electronics

Initially, the adoption of microcomputers began in the 1970s with the development of engine controls. Subsequently, this trend began to spread to products used in body electronics, safety systems, and information systems.

In accordance with this growth, the number of microcomputers per vehicle increased to dozens in an ordinary passenger vehicle and more than a hundred in a luxury vehicle. From the standpoint of software, the lines of code increased from <10,000 in the 1980s to more than 10 million in 2010. This number is likely to continue growing (Figure 10) (Kato *et al*., 2010).

3.3 Algorithms

The functions of the microprocessor embedded in a microcomputer include inputting and processing external data, and outputting the results. The procedure and timing for performing this data input/process/output sequence is determined by the algorithm. As a computer operates in accordance with the instructed program, the algorithm must be written correctly to ensure that the computer operates as intended. Historically, although Euclid's algorithm is regarded as the first mathematical algorithm, other types of algorithms based on state transition rules, physical models, and hardware restrictions are also in use. In automotive electronics, the means and procedures for key

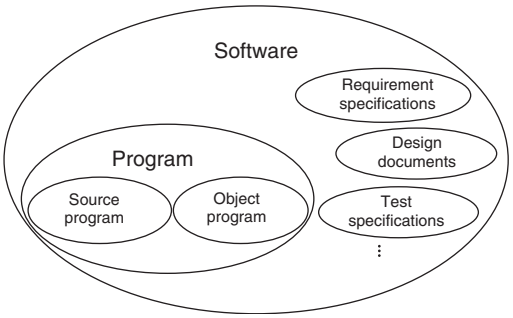


Figure 11. Software and program.

data processing are often regarded as the specifications. The software is implemented to realize the algorithm as concrete processing based on the specifications. Even if the same result is obtained, the algorithm used can result in substantial differences in the execution time required for microprocessor and the amount of data required for performing the work. The terms *software* and *program* are often used synonymously but actually represent slightly different nuances (Figure 11). A program is a set of instructions for computer processing, whereas *software* refers to all the documentation required to develop the programs in addition to the actual programs themselves (Kato *et al*., 2010).

3.4 Software architecture, functional decomposition, and hierarchical design

Software architecture refers to the basic component, framework elements of the software, and the relationship between these elements (i.e., the input/output characteristics and the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

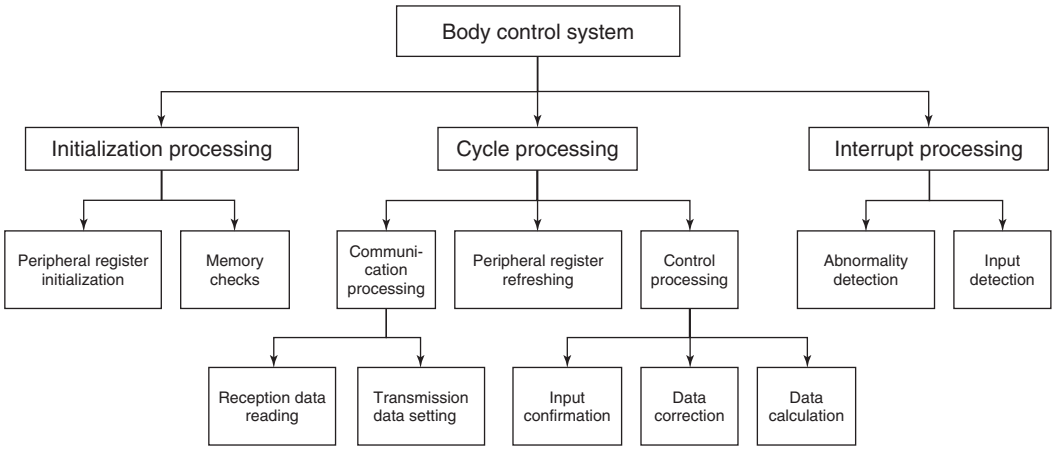


Figure 12. Example of hierarchical design in body electronics.

way one element uses another or is used). The software architecture can be used for mutual understanding and communication between software designers. In addition, the design policy is determined in the initial period of the development to ensure that the subsequent development steps proceed smoothly.

Hierarchical design by functional decomposition is used to clarify the component elements of software. In this design method, the software is divided into functional units that are assembled in a hierarchical manner (Figure 12). Hierarchical design enhances serviceability and quality by clarifying the design process.

3.5 Program categories

Programs can be categorized into system and application programs (applications). System programs can be further categorized into operating systems (OSs), device drivers, and middleware (Figure 13).

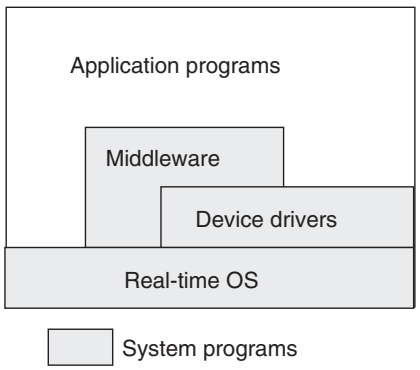


Figure 13. Categories of program.

Small-scale automotive systems are configured with interrupt processing programs, simple schedulers, and applications to reduce the processing time and memory overhead required by an OS. The OS is the fundamental program for ensuring that the computer resources are used efficiently. Automotive applications generally use a real-time operating system (RTOS). The basic functions of an RTOS include scheduling (task management), synchronization, and exclusive control. An RTOS with restricted functions may be referred to as a *kernel*. Under RTOS control, program units instructed by the microprocessor are called *tasks*. Task states are defined as run, ready, and wait. In addition, each task has a priority attribute. The RTOS transits to the run state of the task with the highest priority within the tasks that are in the ready state. This control is called the *scheduling function*. The wait state occurs when waiting for some kind of phenomenon to occur or when a resource cannot be used because it is being used by another task. The wait state continues until that phenomenon occurs or the object resource is released (Figure 14) (Kato *et al*., 2010).

The program portion that quickly carries out the minimum functions required of the interrupt processing is called the *interrupt handler*. This is regarded as a separate unit from the tasks. In the event that multiple tasks are operating the same resources, the exclusive control acquires those resources to secure the operation and excludes the operation of other tasks (Figure 15) (Kato *et al*., 2010).

The program that directly controls the hardware to allow the microcomputer to operate external data input and output is called the *device driver*. The device driver function is often achieved by a combination of the interrupt handlers and tasks. The aim of the device driver is to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

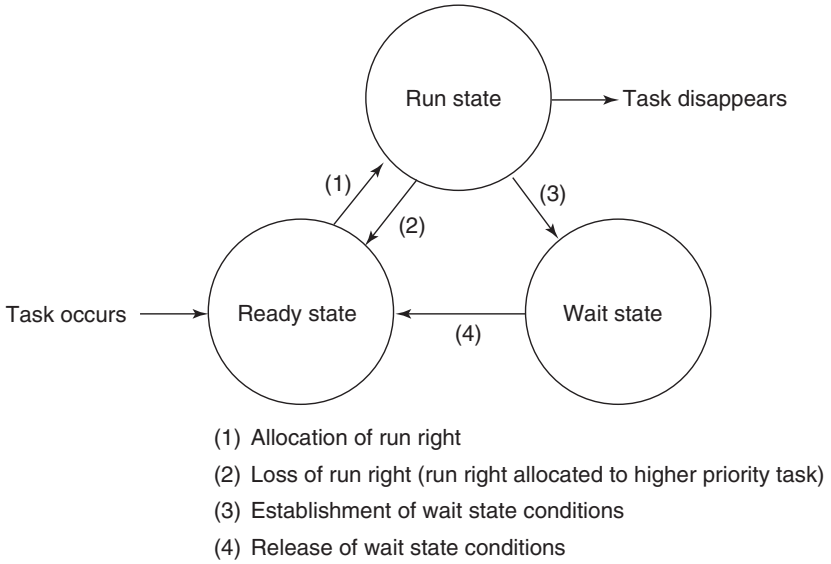


Figure 14. Task state transition.

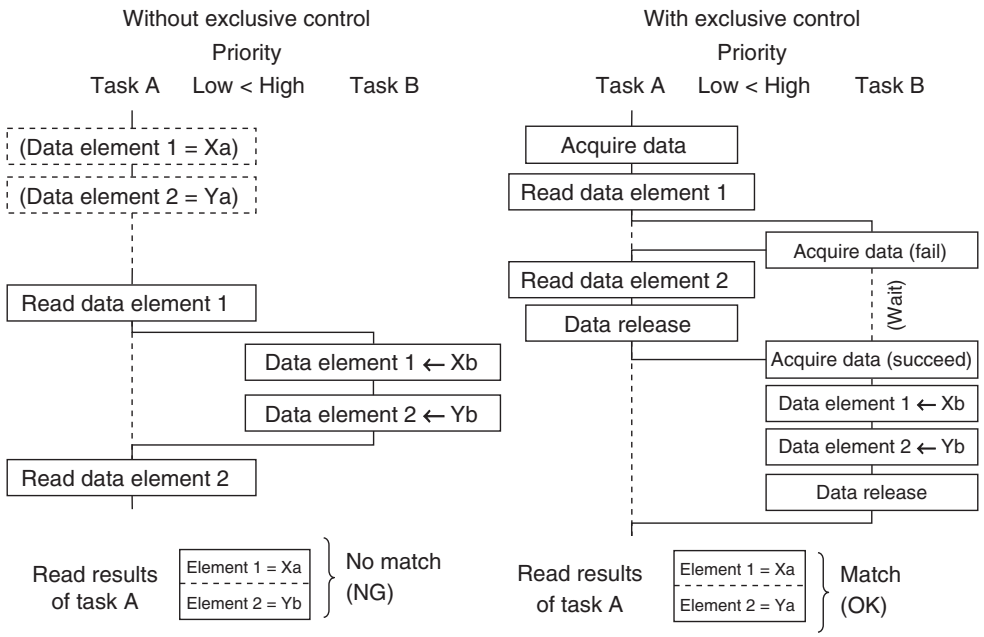


Figure 15. Securing of data operation by exclusive control.

isolate and conceal hardware differences and hardware control details from the applications. Isolating the hardware controls from the applications enables those applications to be reused with different hardware. Device drivers with the same hardware can also be used in common by different applications. In addition, the term *middleware* refers to *software* that is positioned between the RTOS and the applications.

3.6 Software development process

The series of activities involved in the development of software is referred to as the *software development process* or the *software process*. Each unit of these activities that occurs during the software process is called a *phase*. A *process model* is the general and abstract term for a software process. A

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

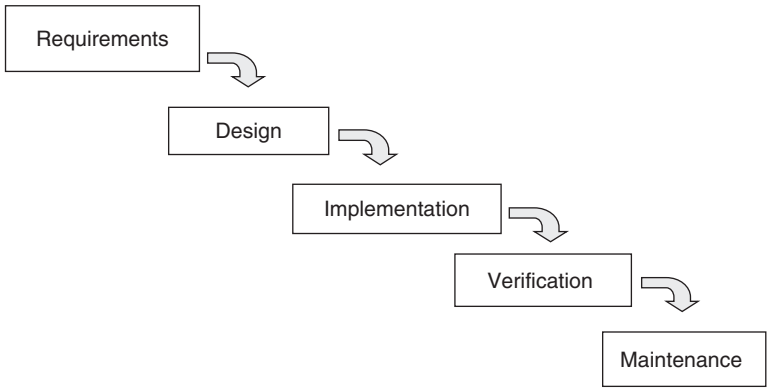


Figure 16. Waterfall model.

classic and well-known process is the waterfall model (Figure 16).

In the waterfall model, development is viewed as flowing downward through several phases throughout the software development process. In this model, the development moves on to the next phase only when the preceding phase is completed and perfected. The benefit of this process is the ease of quality control. If a problem is identified in the previous phase, it is fixable by reverting to the former phase. Therefore, it might be assumed that perfection within the given phase should ensure the perfection of the end product. However, this is often not the case. If the model itself has a problem that can only be revealed during development, perfecting each phase fails to produce the desired end result. Nevertheless, this development process is still widely used because of its ease of implementation.

The most popular process model employed in the automotive software development is the V-model, which is a modified version of the waterfall model. The V-model consists of the coding phase and the verification phase, in which the validation process is subdivided. As seen in Figure 17, the verification phase is divided into unit testing, integration testing, and system testing. Each of these serves the function of testing mounting, designing, and requirement analysis. The V-model shares weaknesses similar to the waterfall model but is superior because it emphasizes the verification phase. This ensures superior quality control of the software and is the most widely used model in the automotive industry.

3.7 Characteristics of automotive control software

The first characteristic of automotive control software is its hard real-time properties. Failure of an automotive control system to process a program within a certain deadline can have severe consequences for the vehicle.

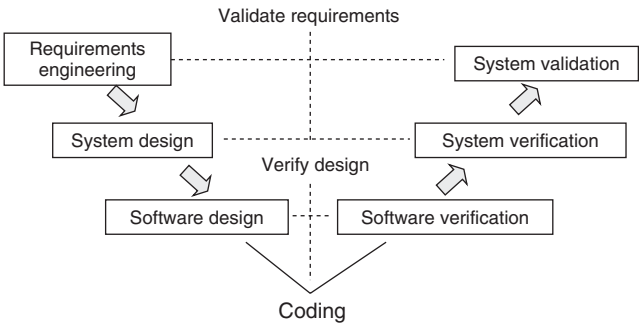


Figure 17. V-model.

The second characteristic is the importance of saving resources. As many automotive systems use single-chip microcomputers, it is crucial to implement the required functions as programs within the capacity of the chip. For this reason, there are strict requirements to save resources.

The third characteristic is the requirement for high quality. As software increases in scale and complexity, activities to ensure quality are becoming more difficult and are growing in importance.

4 ADOPTION OF ELECTRONIC SYSTEM PLATFORMS

4.1 Hierarchy of electronic systems

As automotive electronic systems have become more widespread, large numbers of ECUs, sensors, and actuators have been connected using automotive LAN. Consequently, compared with the early 1990s when the processes to be performed by the ECUs were determined for each function,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

there is now a greater need to determine the allocations of the role of each ECU from the standpoint of overall system optimization.

As the number of ECUs installed in a single vehicle increases, it is becoming clearly more difficult to secure adequate installation locations. Standard functions that do not rely on model variations (i.e., the intended region of sale, such as Asia, North America, Europe, and the like, and the vehicle grade) are being incorporated into integrated ECUs. In contrast, new nonstandard functions that are still in development as well as functions for certain regions and vehicle grades only are regarded as stand-alone functions.

As a result, function hierarchies have been created. This has generated an issue of function allocation, that is, which functions should be allotted to which hardware (i.e., ECUs, sensors, and actuators). A case of function allocation, in which multiple software and hardware items are defined in a hierarchy of related functions, is described later (Kato *et al*., 2010). For example, the electronic key is a function that facilitates the driver entering the vehicle. Even if the driver is carrying a heavy package, the electronic key detects weak radio waves emitted by the key when the driver approaches the vehicle, temporarily disengages the alarm mode, unlocks the doors, and illuminates the courtesy lights at the driver's feet. This system then drives the door opener when the driver touches the door handle. After the driver enters the vehicle and starts to shut the door, the system drives the door closer and adjusts the seat to the optimal preset position (Kato *et al*., 2010).

The air conditioning may also include functions that maintain the most comfortable environment in the vehicle for the driver by controlling the temperature and humidity in the occupant compartment, in addition to supporting the driver's field of vision through the windshield wipers, mirrors, and rear window defogger (Kato *et al*., 2010).

Future automotive electronic systems, especially in luxury vehicles, will probably become more hierarchical and structured. It is also likely that ECUs with sophisticated processing functions will become combined, while creating sensors and actuators integrated with standard input and output processing, regardless of the vehicle model. As a result, automotive network hierarchies have been created with backbone and subnetwork systems.

4.2 Standardization and reuse of electronic parts

There is a need for approaches that efficiently develop and design various automotive electronic systems in parallel and in a short period of time. This can be accomplished by standardizing and reusing as many parts as possible, and newly developing only the parts required for new systems. Such approaches must be capable of constructing various systems quickly and easily.

This approach of adopting standard and reused parts is also called an *electronic platform*. Although vehicle platforms tend to refer to the body and chassis and *computer platforms* generally refer to the hardware and OSs, the purpose of enabling the simple construction of multiple variations is the same (Kato *et al*., 2010).

Design is no longer focused on optimizing individual functions. Current trends are clearly highlighting the need to standardize and reuse many hardware parts, such as microcomputers, power supply ICs, communication circuits, connectors, passive devices, and the like.

4.3 Example of platform adoption

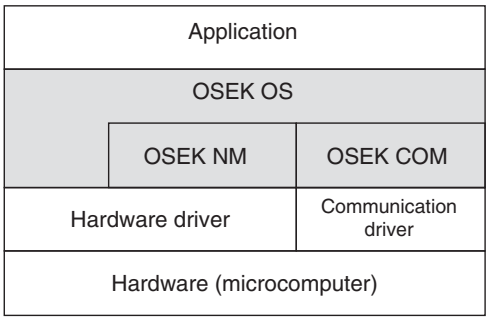
Modern vehicles use network systems to connect multiple ECUs to allow dynamic cooperative control to support the driver. For example, sensor data can be shared with corner prediction by the navigation system ECU to request the engine ECU to generate engine braking by automatically changing gear.

The longitudinal motion of the vehicle can be controlled by commands sent over the network. In addition to conventional accelerator and brake controls, engine and brake requests can be transmitted by adaptive cruise control (ACC) systems, which maintain a set vehicle-to-vehicle distance using millimeter wave sensors and/or cameras, vehicle stability control (VSC) that ensures stable vehicle behavior and maneuverability on poor road surfaces, parallel and reverse parking assist systems, precrash systems that detect the risk of a collision in advance, and so on. Such commands require arbitration systems that apply priorities to each signal (Kato *et al*., 2010). These systems are not limited to vehicles with four-cylinder gasoline engines and should be applicable in theory with six-cylinder, diesel, hybrid, and electric vehicles.

5 INTERNATIONAL STANDARDIZATION (OSEK AND AUTOSAR)

International standardization activities include those led by the International Organization for Standardization (ISO) and the International Electro Technical Commission (IEC). However, a growing number of consortiums have been set up under contract with the main organizations to respond to the speed of changes in technology and the market. However, although the consortiums are responsible for establishing new standards that require a speedy approach,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



OSEK: open systems and the corresponding interfaces for automotive electronics

VDX: vehicle distributed executive

Figure 18. Software structure of OSEK/VDX.

items related to the maintenance, management, regulatory approval, and certification of these standards are still performed by the ISO, IEC, and the like.

The following sections describe some examples of international standardization for software. As development costs increase because of the multiplying types and scale of software, activities to standardize software and their use have been prompted by the European automotive industry, leading to the determination of an RTOS for automotive control called the *OSEK/VDX* (*vehicle distributed executive*) specifications. OSEK/VDX comprises of RTOS specifications (OSEK OS), communication specifications within and between ECUs (OSEK COM), and network management specifications (OSEK NM) (Figure 18). OSEK/VDX has been certified as ISO 17356 as an international standard for automotive OS.

The AUTOSAR consortium was established in 2003 for the purpose of international standardization. It is composed of more than 100 organizations including automotive manufacturers, suppliers, and tool vendors, divided into four types of membership consisting of core members, premium members, associate members, and development members. AUTOSAR specifications have been formulated based on the OSEK/VDX specifications and defined with extremely extended functions for realizing standardization, scalability, transferability, integration, maintainability, and software upgrades in the fields of automotive electrics and electronics (E/E) architectures. The AUTOSAR run time environment (RTE) provides a software application interface consisting of system services, memory services, communication services, I/O hardware abstraction, and complex drivers (Figures 19 and 20). These enable application programs to be used independently from the different hardware configurations (Kato *et al.*, 2010).

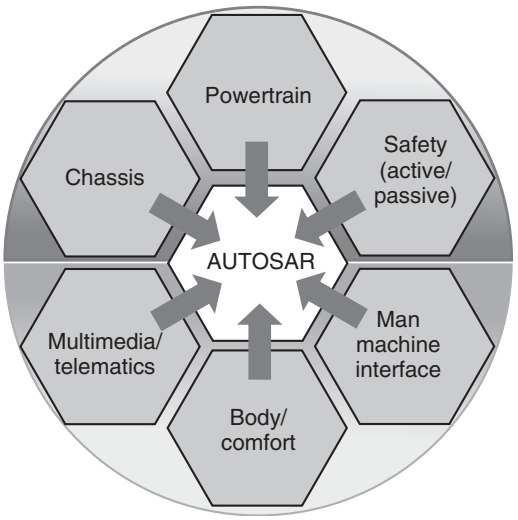


Figure 19. Field of applications in AUTOSAR.

This type of platform-based development may even form a key technological element with the potential to change the vertical integration business model into an open architecture business model.

6 SUMMARY AND CONCLUSION

This chapter has described the histories and future directions of automotive microcomputers, software, algorithms, architectures, hierarchy design, functional decomposition, and international standardization (OSEK and AUTOSAR), while considering semiconductor technologies and their applications in automotive electronics. These issues will continue to become increasingly important in the realization of further green and dependable automotive systems.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

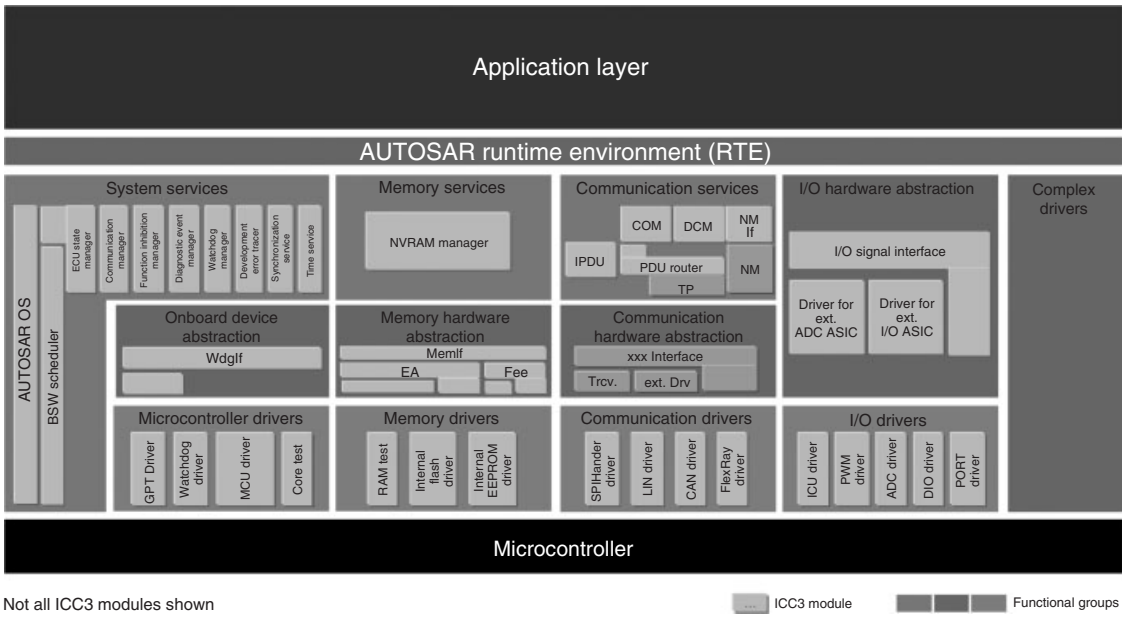


Figure 20. Software structure of AUTOSAR.

In the future, the progress of microcomputer performance and the evolution of software technologies are likely to result in advanced sensing technologies, super high speed data processing, higher integration, more efficient power electronics, and greater progress in high frequency applications such as millimeter wave and wireless functions. In addition to the need to advance these technologies, standardization across the industry will also be an important issue for pursuing development efficiency.

REFERENCES

Ishihara, H. (2007) An Overview of Automotive Electronics and Future Requirements for Microprocessors. *Microprocessor Forum 2007*, Keynote Speech.

Ishihara, H. (2009) Current issues and future prospects of automotive semiconductors *Journal of Society of Automotive Engineers of Japan*, **63** (1), 65–70.

Kato, M., *et al*. (2010) *Illustrated Car Electronics in Two Volumes*, Nikkei Business Publications, Inc., Tokyo.

Kawamoto, K., *et al*. (2001) A single chip automotive control LSI using SOI bipolar complimentary MOS double-diffused MOS *Japanese Journal of Applied Physics*, **40**, 2891–2896.

Ookura, K. (2005) Semiconductor Technology for Automotive Electronics *Denso Technical Review*, **10** (2)

Tsuda, K. (2007) Denso's Automotive MCUs. *Microprocessor Report*, MPR Newsletter.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto217
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In-Vehicle Network

Susumu Akiyama
DENSO Corporation, Kariya, Japan

1 Introduction 1
2 In-Vehicle Network 2
3 Conclusions and Future of In-vehicle Network 9
Related Articles 9
Further Reading 9

1 INTRODUCTION

Despite the origins of motor vehicles as a collection of mechanical parts, the introduction of electronic engine controls in the 1970s heralded the start of a trend that continues today. Now, electronic controls are utilized in a majority of automotive devices, including the brakes, steering, suspension, instrument cluster, climate control, airbags, door locks, power windows, navigation system, and so on.

Electronic control systems consist of sensors, actuators, and electronic control units (ECUs). To make a human analogy, sensors roughly correlate to the eyes and ears of a system, the actuators act as the hands and feet, and the ECU plays the role of the brain. In these systems, measurement signals are transmitted from the sensors to the ECUs. The measurement signals are used by each ECU to calculate control signals that are transmitted to the actuators. These signals are carried by wire harnesses and other related parts. ECUs consist of hardware such as microcomputers, peripheral integrated circuits (ICs), a power supply, and printed circuit boards, as well as software that define the algorithms used by the ECU.

A typical example is the electronic engine control system. As mentioned earlier, the role of the engine ECU is analogous to the brain as it is responsible for calculating fuel injection quantities and ignition timings. The engine ECU is connected to an in-vehicle network to enable information exchange with other ECUs. For example, engine coolant temperature data is sent from the engine ECU to the instrument cluster and displayed on the engine temperature gage. The in-vehicle network also carries signals from other ECUs to the engine ECU. For example, if the vehicle is equipped with adaptive cruise control (ACC), which functions to maintain a set distance with the vehicle ahead, the ACC ECU will send a signal to the engine ECU requesting acceleration if the distance to the preceding vehicle exceeds the set amount. Alternatively, if the vehicle is stopped in high temperatures during the middle of summer, the climate control ECU will send a signal to the engine ECU requesting an increase in the engine idling speed to supplement the cooling capability of the climate control.

Figure 1 shows the evolution of the vehicle system from an era of standalone controls (such as for the engine, brakes, and climate control) to a period of integrated control in which multiple controls cooperate to achieve a set of complex functions. It also shows current developments, in which electronic controls are starting to expand outside the vehicle and cooperate with external infrastructure. Conventionally, these types of complex and versatile functions were achieved by an automotive electronic system that grouped together individual ECUs performing individual controls. These ECUs are now connected to each other using the in-vehicle network. The number of ECUs connected to the in-vehicle network is increasing and hierarchical control structures, divided into a number of domains, are being developed.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto218
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

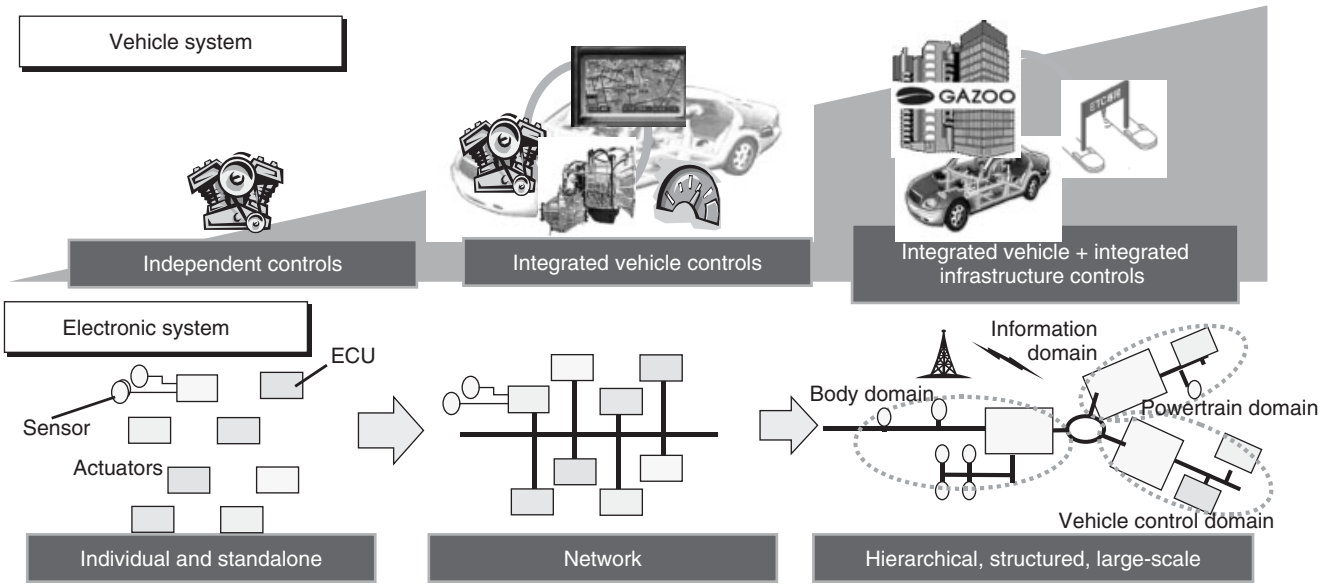


Figure 1. Evolution of automotive electronic system.

In a modern automotive electronic system, control functions for the powertrain, chassis, and body (i.e., the basic performance aspects of a vehicle responsible for driving, turning, and stopping) mutually interface with service functions that support the driver (i.e., information and communication) to achieve a high level of safety, comfort, usability, and environmental performance. The scale of systems responding to these social needs is likely to expand in the future, leading to more sophisticated and versatile functions. To achieve such large-scale systems, new control technology will be incorporated into vehicles alongside new information and communication technology. As a result, the optimum systems will have to be designed with hardware and software of the highest quality and reliability. Accordingly, the in-vehicle network will play an increasingly important role as a core component technology of the automotive electronic system.

2 IN-VEHICLE NETWORK

Typical familiar networks include data communication nets accessible through mobile devices, and the internet that can be used through computers at home or in the workplace. Vast quantities of information can be accessed by connecting to such networks at any time and in any location, enabling people to talk across huge distances. New network technology has been adopted to enable this type of communication environment. The main basic technologies behind these networks are communication devices, communication-related standards for procedures and the like, and software designed based on these standards.

As a result of the evolution of the automotive electronic system as described earlier, sensors and actuators are required to supply reliable information (data) to enable highly functional and accurate controls. Therefore, the construction of an in-vehicle network is necessary to enable information (data) exchange among sensors, actuators, and ECUs. From the standpoints of data communication efficiency and reliability, as well as cost, the in-vehicle network mainly uses multiplex communication technology with digital signals.

2.1 System structure

Figure 2 shows the typical structure of an in-vehicle network. The figure shows a total of 26 ECUs that exchange data in three mutually connected networks (the vehicle control system, body system, and information system). Exchanges among these systems are conducted through the gateway ECU. The gateway ECU switches the communication procedure, transmission speed, and frame format of the data as required by each network.

The major differences between the in-vehicle network and conventional computer-based networks used in university laboratories or design departments in corporations are as follows: the data flowing within the in-vehicle network is fixed in advance, and the connected devices are dedicated ECUs with set functions (such as engine control, door control, and current position estimation) instead of a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto218
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

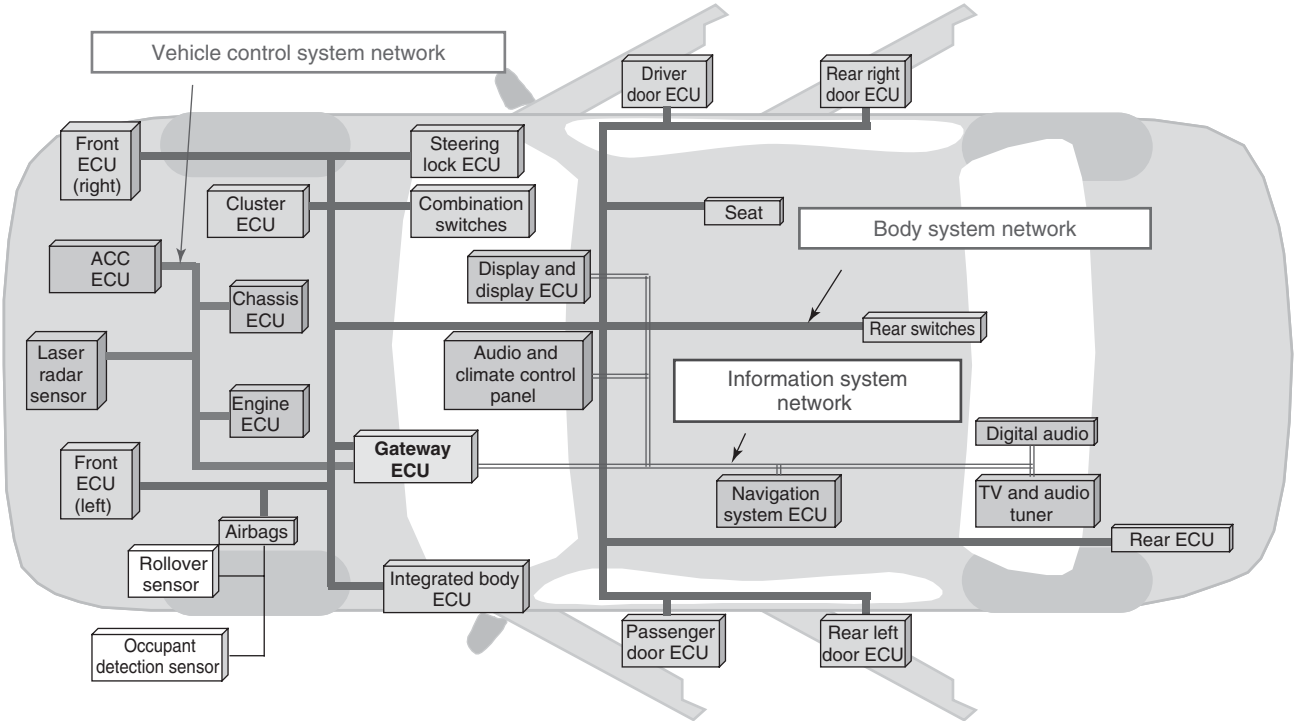


Figure 2. Example of in-vehicle network.

general-purpose computer or server with similar functions. Although the in-vehicle network is already prevalent even in mass-market vehicles, it has developed historically as a multiplex communication system to reduce wire harness complexity.

2.2 Multiplex communication systems

An automotive multiplex communication system is capable of carrying multiple types of signals through a single signal wire. This reduces the number of wire harnesses that have to be installed in a limited space and enables a larger number of functions to be performed with a smaller number of wires. Figure 3 shows the wire harnesses used throughout a vehicle and illustrates the effect of adopting multiplex communication systems for some of these harnesses. In this example, the multiplex communication system reduced the weight of the wire harnesses by 39.9% and the number of wires by 42.5%.

As an example of a body multiplex communication system introduced at the beginning of the 1980s, the control switches for the body-related control systems on the driver's side were collected and installed together. This system connected each ECU installed in the doors and body into a network to perform the following functions:

1. Door lock and unlock control
2. Vent window (i.e., triangular quarter glass) control
3. Power window control
4. Power seat control
5. Seat heater control
6. Ashtray and switch illumination control
7. Illuminated entry (i.e., the illuminated ignition lock and under-foot illumination lamps) control

2.3 Communication protocol

Early multiplex communication systems adopted optimum communication procedures that were newly developed for each model by the manufacturer. However, as the number of ECUs increased and the use of the in-vehicle network became established, several standard procedures began to be adopted around the world.

Figure 4 lists some typical network access methods. As a basic principle, an in-vehicle network transmits and receives data using one signal line. Therefore, if each ECU connected to the network starts to transmit information without coordination among the ECUs, transmitted data will collide and be lost. Consequently, rules are required to prevent data collision. These rules are called the *access method*. Access methods are broadly divided into master–slave types in which all communication is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto218
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

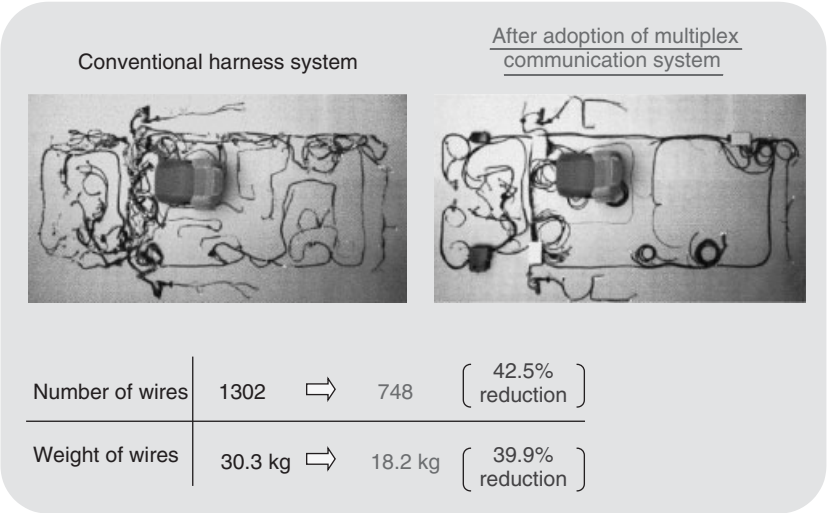


Figure 3. Effect of adopting multiplex communication system.

Access method		Characteristics	Example applications
Multi-master	CSMA/CA	Immediate transmission is possible when the bus is empty	CAN
	Token passing	Sequential transmission is possible	MOST
	Time triggered	Communication performed at set intervals	FlexRay
Master-slave		Communication performed following instructions from the master	LIN

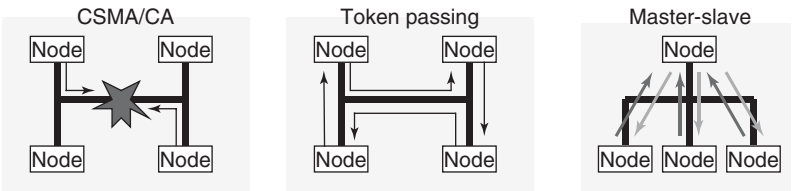


Figure 4. Access methods.

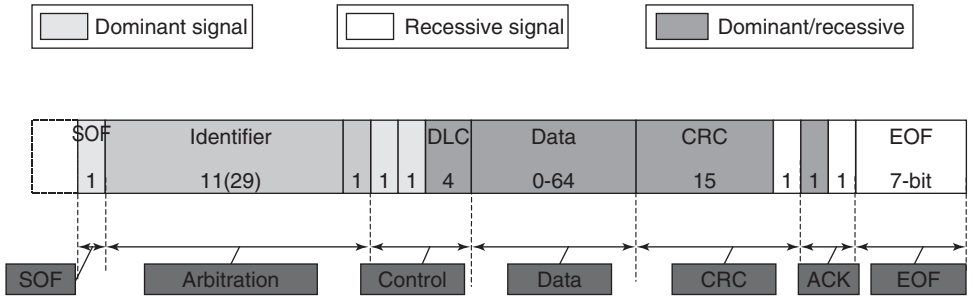
controlled by the master, and multi-master types (also referred to as *multiple access types*) in which each ECU has equal transmission rights as a master. Typical access methods used by the in-vehicle network include master–slave methods, and multi-master methods such as carrier sense multiple access with collision avoidance (CSMA/CA), time-triggered protocols, and token passing.

2.3.1 Controller area network (CAN)

Controller area network (CAN) is the most widely adopted type of in-vehicle network communication procedure. The CAN communication procedure uses CSMA/CA as the access method. Before transmission, the ECUs connected to the network check whether other signals are being sent through the communication line. Transmission starts once the communication line is empty. As a result, the ECU that accesses the communication line most quickly obtains the transmission right. As shown in Figure 5, ECUs attach an identifier (ID) to each message. If two ECUs start to transmit a message at the same time, the ECU with the smaller ID value obtains the transmission right. The ECU with the higher value ID immediately stops transmission and transits to reception mode.

This mechanism is a feature of CSMA/CA and is referred to as *arbitration*. Unlike the general carrier sense multiple access with collision detection (CSMA/CD) method, CSMA/CA has high data transmission efficiency as data is not lost in collisions, which means that lost data does

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto218
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Field	Content
Start of frame (SOF)	Expressed by one dominant bit.
Identifier (Arbitration)	Code field that expresses the details of the subsequent data and the priority of the message frame. The term arbitration also means prioritization and this field determines which message frame has priority if signal collision occurs.
Control	The DLC portion expresses the length of the data field (0 to 8 bytes).
Data	The data field contains the data (0 to 8 bytes).
CRC	Check code that detects if data in the message frame is garbled (transmission error).
ACK (Acknowledge)	A positive response. A node that has received a message normally uses the ACK field to transmit a dominant signal.
End of frame (EOF)	Comprises a 7-bit recessive signal.

Figure 5. CAN frame format.

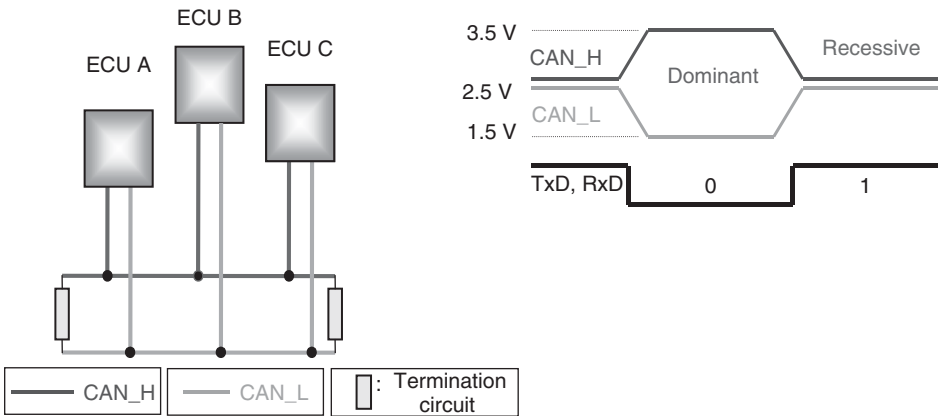


Figure 6. Communication waveform.

not have to be retransmitted. Consequently, this method is used in fields with comparatively short communication distances where the maximum delay time can be specified. In addition to the in-vehicle network, examples include networks among medical electronic devices, networks among control devices in factories, and so on.

Another feature of the CAN communication procedure is that it carefully defines the error detection, notification, and recovery process. In this way, the detection method when an error occurs and the operations after detection are clearly defined.

Figure 6 shows a waveform of a twisted pair communication cable for a high speed CAN with termination circuits at both ends. In a vehicle, communication among electronic devices is accomplished by this type of signal.

2.3.2 Local interconnect network (LIN)

Compared to CAN, which is mainly used in networks among ECUs, local interconnect network (LIN) is primarily used for low cost network systems from ECUs to sensors and actuators.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto218
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

LIN uses the master–slave access method in which the slave (usually a sensor or actuator) transmits information based on instructions from the master (usually an ECU).

LIN has a single communication cable and can use general-purpose serial communication ports normally built into microcomputers. The maximum transmission speed of LIN is 20 kbps and communication is possible without an expensive crystal oscillator using the synch field shown in Figure 8.

Figure 7 shows an example of a typical LIN system configuration. In many cases, LIN is used for subnetworks of network systems among ECUs, rather than as a standalone network. Many slave nodes are devices with single functions, such as motors and sensors.

The tasks written in each node can be described as follows. The master task is the function that manages the communication of all nodes in accordance with the LIN communication procedure. Here, it manages the

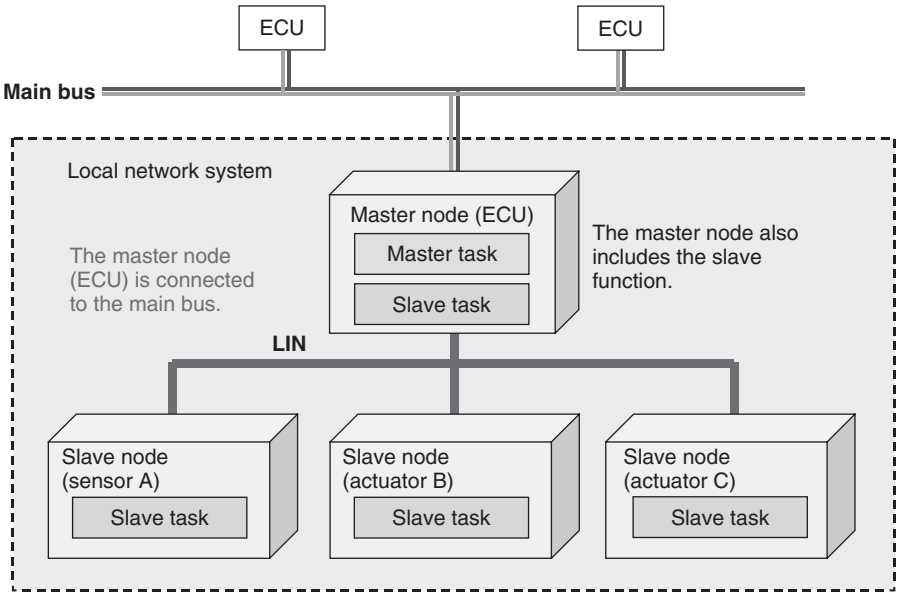
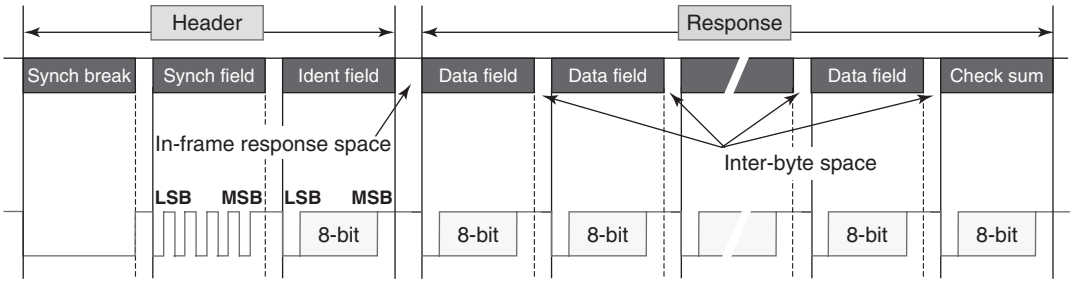


Figure 7. Example of LIN system configuration.



Field	Content
Header	Transmitted by the master task.
Response	Transmitted by the slave task.
Synch break	Field transmitted by the master task to notify the start of the LIN message frame.
Synch field	Field that enables the slave node to perform clock (transmission speed) correction.
Ident field	Field in which the master task specifies the slave task that should transmit the response.
Data fields	Fields that contain the data to be notified by the slave task in response to the master task.
Check sum	Field that enables the response of the slave task to be judged as normal or abnormal.

Figure 8. LIN frame format.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto218
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

transmission timing of all nodes. The slave task is the function that transmits data in accordance with the instructions from the master.

Every master task has a table that determines the transmission sequence of each node. This is called a *schedule table*. The master task transmits the header in sequence in accordance with this table. The data to be transmitted by the slave task in response to each ID is determined in advance (Figure 8).

As shown in Figure 8, LIN consists of one frame that comprises two blocks called the *header* and the *response*. The header is always transmitted by the master task and contains a code to express the start of the frame (the synch break), a code that teaches the slave the transmission speed (the synch field), and an ID that defines the node to transmit the data (the ident field).

The response is always transmitted by the slave task. The slave that holds the data specified by the ID (the ident field) included in the header transmits the data using the start–stop synchronization method.

2.3.3 *FlexRay*

FlexRay is a communication procedure developed for real-time controls such as X-by-wire that require high speed and reliability. It was used first in the suspension systems of luxury vehicles and is expected to be broadly adopted in the future.

The access method of FlexRay is the time-triggered protocol. All ECUs connected to the network share a clock and data is transmitted and received among the ECUs at set intervals. Therefore, this communication procedure can guarantee the arrival time of data in advance. This contrasts with CSMA, in which data arrival times depend on the busy status of the communication paths. FlexRay is also compatible with a duplex communication system, which is a far superior approach to other protocols from the standpoint of error detection, notification, and recovery.

Figure 9 shows a portion of the FlexRay communication procedure. One communication cycle consists of a static segment that always transmits data at the set time, and a dynamic segment that attempts to transmit data only when such data is present. The figure shows a duplex communication path comprising channels A and B. Frame A1 transmitted by ECU A and frame C1 transmitted by ECU C are used as the duplex communication path. In contrast, frames A2, D1, and E1 use channel A as the communication path and frames B1 and B2 use channel B. As each transmission node in the static segment is allocated, periodic transmission is secured and reliability is improved through a duplex communication system such as that illustrated by frames A1 and C1. In the dynamic segment, each frame is only transmitted when necessary and frames are not transmitted once the transmission start/stop line is exceeded. Therefore, in the example, frame C3 is not transmitted. These frames are transmitted from the next cycle.

Figure 10 shows the FlexRay frame format. Each frame consists of a header segment that includes synchronization and start frames, a payload segment with a maximum length of 254 bytes, and a trailer segment that functions as a 24-bit cyclic redundancy check (CRC) for detecting errors.

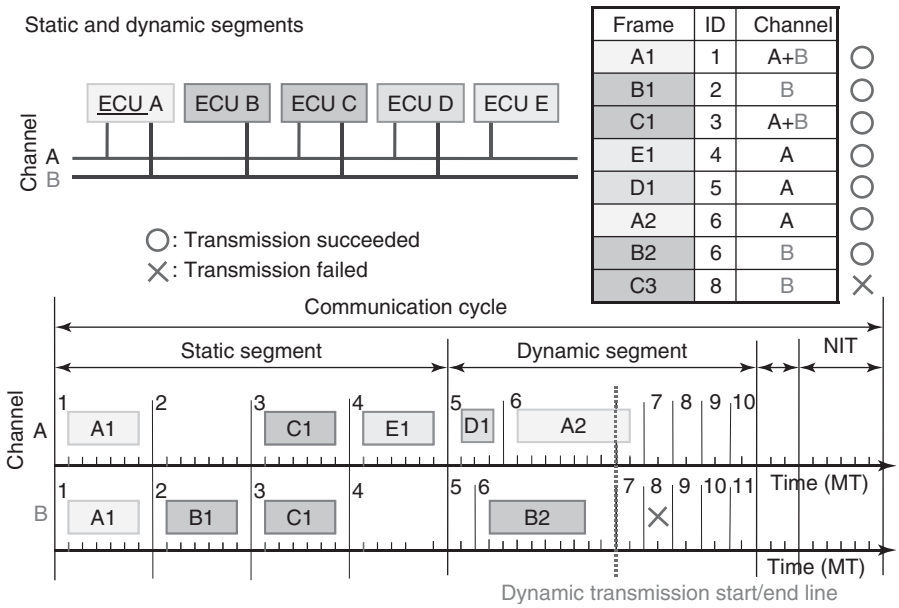


Figure 9. FlexRay communication procedure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto218
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

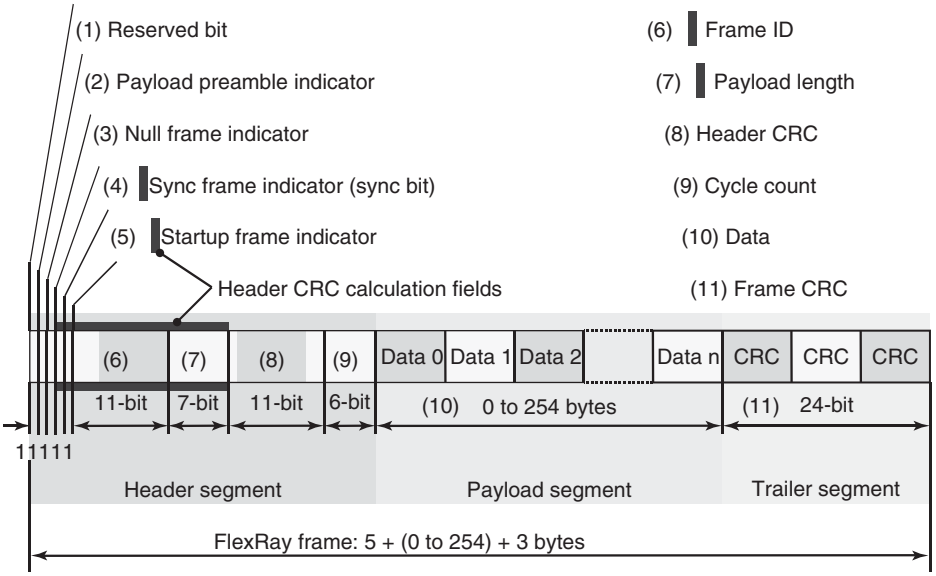


Figure 10. FlexRay frame format.

2.3.4 Ethernet

Unlike CAN, LIN, and FlexRay, which were primarily developed for automotive applications, Ethernet was originally a protocol for consumer applications developed for general-purpose LANs connecting computers, printers, servers, and the like. Owing to the demand for an in-vehicle network capable of transmitting greater volumes of data at higher speeds, the Ethernet protocol has recently begun to be applied in vehicles. Ethernet automotive applications include backbone networks that connect various domains, diagnostics, and reprogramming, as well as networks, control systems, body systems, entertainment systems, and so on for displaying moving images (see Figure 11 for some typical examples).

Figure 12 shows the Ethernet frame format for automotive applications. This protocol is capable of transmitting large volumes of data up to 1500 bytes in a single frame, which is not possible using other automotive communication protocols. Although standard communication lines for commercial applications contain two twisted pairs of cables, the development of single twisted pair cables for vehicles is also making progress.

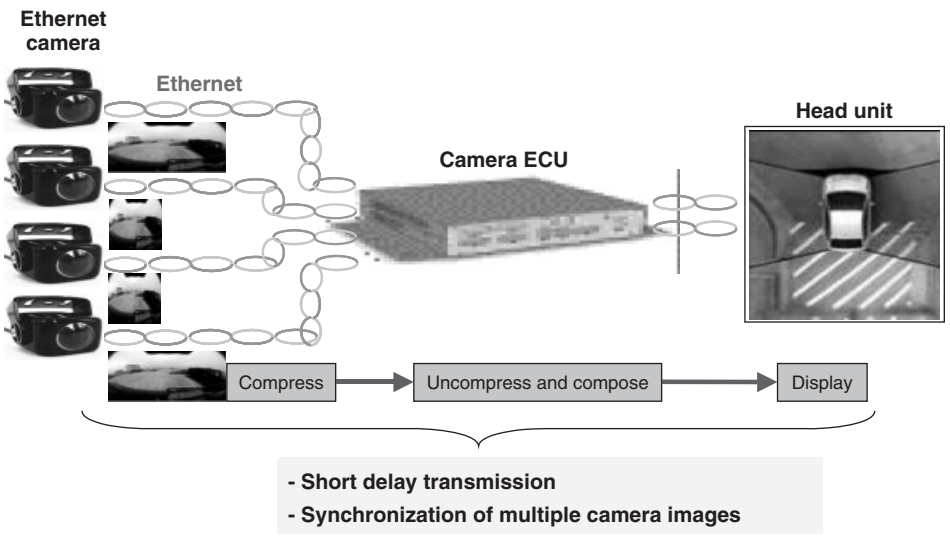
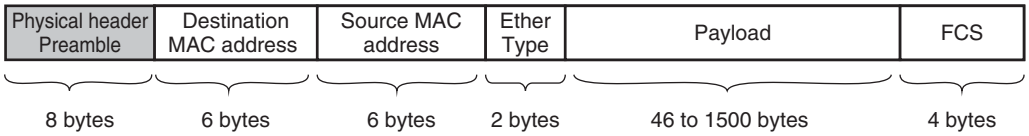


Figure 11. Peripheral monitoring system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto218
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Field	Content
Preamble	This field indicates the start of the frame and also functions for synchronization.
Destination MAC address	The address that indicates the transmission destination.
Source MAC address	The address that indicates the transmission source.
EtherType	The identifier that indicates the upper layer protocol in the data field.
Payload	The field in which the upper layer protocol packet is generally stored. Always consists of bytes in multiples of 4.
FCS	The CRC for checking errors is stored in this field.

Figure 12. Ethernet frame format.

3 CONCLUSIONS AND FUTURE OF IN-VEHICLE NETWORK

The number of onboard electronic devices has increased in accordance with the growth of electronic functions in vehicles. This trend is driving the development of the in-vehicle network. The introduction of new functions in the future is likely to increase the hierarchical levels, transmission speed, and transmission capacity of the in-vehicle network.

The current in-vehicle network is primarily used for communicating among ECUs. In the future, a similar network will also be established among ECUs and intelligent sensors, actuators, switches, relays, and motors, helping to reduce the number of wire harnesses, vehicle weight, and fuel consumption.

Furthermore, vehicles are also likely to evolve from independent units to more comfortable and safer means of transportation through two-way connections outside the vehicle. This will be accomplished through infrastructure-to-vehicle communication, vehicle-to-vehicle communication, battery charging communication, as well as the introduction of long-term evolution (LTE) wireless communication services, and so on. In these cases, the in-vehicle network will play a major role in distributing data received from external sources and collecting data from inside the vehicle.

From the standpoint of development efficiency, standardization of a larger scale automotive electronic system including the in-vehicle network is a key issue. Obviously, the in-vehicle network protocols described earlier are important key international standardization items. However, the necessary scope of standardization will expand in the future to include the types of data exchanged over a network, interfaces among controls, security guidelines for external data, and so on.

RELATED ARTICLES

Historical Overview of Electronics and Automobiles: Breakthroughs and Innovation by Electronics and Electrical Technology
Microcomputers and Related Technologies: Enlargement of Software Size, Algorithms, Architectures, Hierarchy Design, Functional Decomposition, and Standardization
Increase of ECU and Wire Harness, JB Simplify and Decrease Networks)

FURTHER READING

ISO 11898, Road vehicles–Controller Area Network (CAN).

ISO 17987, Road vehicles–Local Interconnect Network (LIN).

ISO 17458, Road vehicles–FlexRay Communications System.

Kato, M. (2010) *Automotive Electronics: Systems*, Nikkei Business Publications, Inc., Tokyo.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto218
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Application of Image Recognition Technology to Vehicles

Yukimasa Tamatsu and Naoki Nitanda
DENSO Corporation, Kariya, Japan

1 Introduction 1
2 Camera Categories 1
3 Image Recognition Technology by Onboard Cameras 2
4 Object Recognition 2
5 Recognition of Visual Environment (Visibility Recognition) 5
6 Conclusion 7
References 7
Further Reading 8

1 INTRODUCTION

Onboard cameras used in driver assistance systems can be broadly categorized into viewing cameras for displaying images from around the vehicle (e.g., rear view monitors) and sensing cameras that detect the driving environment using image recognition technology. Recent technological progress in image recognition, and the growing sophistication and lower cost of processors, has resulted in an influx of image recognition applications onto the market. This chapter outlines sensing camera systems that use image recognition technology.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto224
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 CAMERA CATEGORIES

Sensing cameras can be categorized in accordance with the object recognition method into monocular systems that use one camera and stereo systems that use two cameras (Figure 1).

A monocular camera is generally used for recognition processing based on the density pattern of the images (the pattern or the texture). Example applications are the recognition of objects such as vehicles, pedestrians, lane markings, traffic signals, and the like. Monocular cameras may also be used for dynamic stereo processes using the disparity generated by camera movement (Yamaguchi, Kato, and Ninomiya, 2006). This is a method that identifies the distance to a stationary object from two images obtained at a time delay. Although issues remain to be resolved from the standpoints of distance estimation accuracy and calculation speed, single-camera systems have the merits of lower cost and installability.

In contrast, stereo cameras are capable of calculating the distance to an object by verifying the left and right images and applying the principles of triangulation (Kanade *et al*., 1996). Recently, the semi-global matching (SGM) method (Hirschmuller, 2005) has been used to enhance the accuracy of this verification process to improve the precision of distance estimation. However, systems that require two cameras are more expensive and require larger cases than monocular cameras.

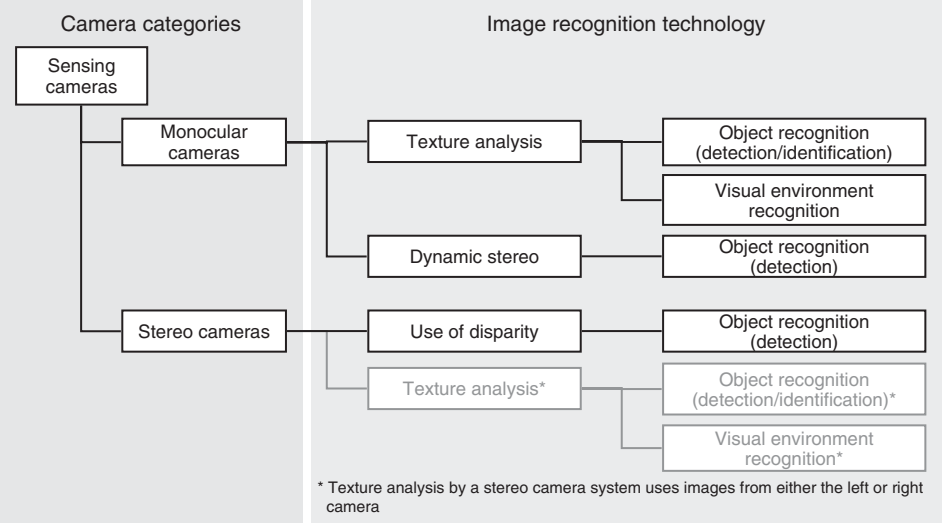


Figure 1. Sensing cameras and image recognition technology.

3 IMAGE RECOGNITION TECHNOLOGY BY ONBOARD CAMERAS

Image recognition technology that uses cameras installed in the vehicle generally employs a technology called *specific object recognition*. Specific object recognition consists of the following processes: the extraction of objects from an image that correspond to a certain target, such as vehicles, pedestrians, and the like; detection processing that determines whether the target object is present; and identification processing that judges what the detected object is. Approaches to detection processing include the detection of solid objects based on disparity using a stereo camera system, or object detection based on a certain shape or texture. Identification processing also generally uses the shape or texture of an object. In the following description, specific object recognition is simply referred to as *object recognition*.

In recent years, research into object recognition has also focused on visual environment recognition that identifies the appearance of objects. Visual environment recognition identifies objects and driving scenarios and may also be referred to as *visibility estimation*. The most common approach uses textures to estimate the appearance of an object to the driver. Although there are few cases of applications with onboard cameras, this is a key technology for realizing driving assistance systems adapted to the visual environment.

The following sections will describe the recognition of traffic signs, vehicles, and pedestrians as examples of detection objects, and visibility estimation of traffic signs and the driving environment as examples of visual environment recognition.

4 OBJECT RECOGNITION

Figure 2 shows the general process flow of object recognition. It consists of preprocessing, feature extraction, and identification and categorization. Preprocessing includes the normalization of pixel values and the like. Feature extraction refers to the effective recognition of aspects such as the brightness and color information of images, or geometrical characteristics such as corners and edges. Identification

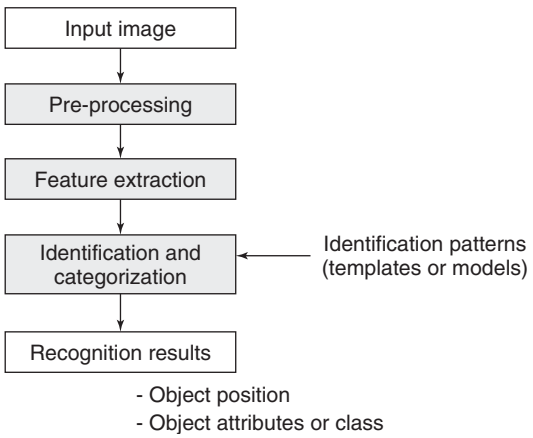


Figure 2. Process flow of object recognition.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto224
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

involves the direct comparison of the extracted features with predetermined patterns (such as templates and models), or categorization using statistical pattern recognition methods or the like, and the output of recognition results such as the position or attributes of objects.

From the standpoint of onboard applications, the identification of traffic signs and other objects that completely match a pattern or shape is generally performed by a direct comparison using templates or models. In contrast, statistical pattern recognition is generally adopted for the recognition of objects that have different patterns or shapes due to physical differences or posture but are in the same category, such as vehicles or people.

This section describes the recognition of lane markings using classical features, the recognition of traffic signs as a typical onboard application that adopts direct comparison and identification, and the recognition of vehicles and pedestrians as a typical onboard application that adopts statistical pattern recognition.



Figure 3. TSR. (Reproduced with permission from Takaki et al., 2009. © Institute of Electrical Engineers of Japan.)

4.1 Recognition of lane markings

The recognition of painted lane markings on the road surface is used in systems that warn the driver when the vehicle is about to depart its lane (such as in lane departure warning systems) and systems that prevent lane departure by controlling the steering. The contrast edge distribution between the road and painted marking from images obtained by a forward monitoring camera are applied to a road model. The positional relationship between the paint on the road and the driver's vehicle, the curvature of the road, and the like are calculated, and the results are applied in the warning or control system.

4.2 Traffic sign recognition (TSR)

The recognition of traffic signs may help prevent the driver from driving too quickly or may help reduce instances in which the driver overlooks a stop sign (Figure 3). These systems judge the type of sign in an image from its texture.

Conventional detection and identification of traffic signs uses template matching. Template matching is a method that detects objects corresponding to images in a template prepared in advanced (called *template images*) and compares this template with the input images. However, the size and color of traffic sign images captured by an onboard camera change depending on the positional relationship between the camera and the traffic sign as well as due to fluctuations in light from the sun. Consequently, the detection and identification of traffic signs by template matching requires the preparation of template images considering the possibility of enlargement, reduction, and rotation of signs, as well as fluctuation in brightness levels. Accordingly, a huge amount of template images would be required to match every condition.

Takaki *et al*. proposed a TSR method to resolve these issues (Takaki *et al*., 2009). This method is called *scale invariant feature transform* (*SIFT*) (Lowe, 1999) and it does not use template images for matching. Instead, it calculates features that do not vary in accordance with changes in rotation and scale. It then detects and identifies traffic signs by matching between these features. SIFT features can be used to recognize signs at an angle or signs that are partially blocked. This is a robust TSR method that is also compatible with changes in brightness levels as the SIFT features do not require color information.

4.3 Recognition of vehicles and pedestrians

Applications that recognize vehicles in front of the driver's vehicle (referred to below as the preceding vehicle) and estimate the distance and relative speed between it and the driver's vehicle include forward collision warning (FCW), adaptive cruise control (ACC), and automatic emergency brake systems (AEBSs) (Figure 4). Systems have also been developed that recognize pedestrians and help avoid collision accidents by indicating their position to the driver (Figure 5).

As mentioned above, the pattern and shape of vehicles and pedestrians change in accordance with physical differences and posture. Therefore, as is the case with TSR, it

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto224
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 4. Vehicle recognition.

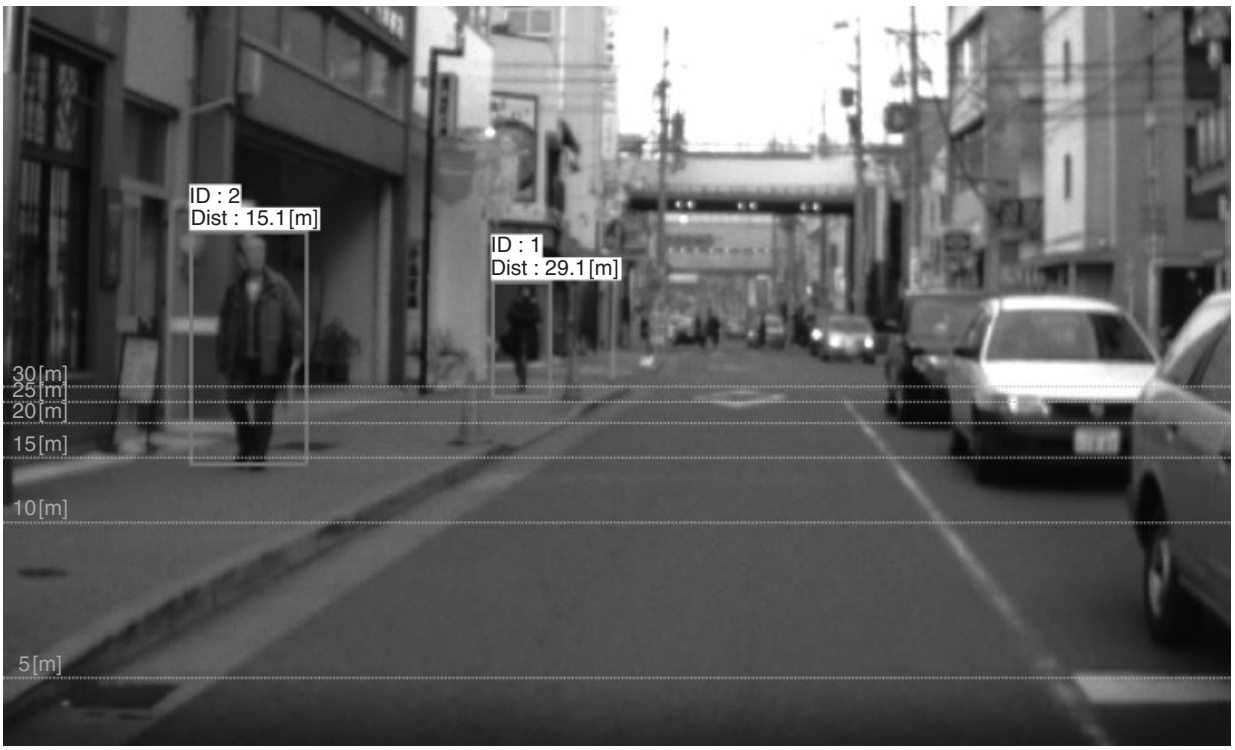


Figure 5. Pedestrian recognition.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto224
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is difficult to apply matching methods that use specific template images. For this reason, vehicle and pedestrian detection tends to use statistical pattern recognition.

In statistical pattern recognition, large volumes of sampled data are analyzed, and the features and judgment criteria useful for recognition are extracted and learned from that data. The theory of this method has progressed from multiple standpoints as new learning algorithms have been proposed from the machine learning field since the 1990s. An algorithm that has gained attention recently uses a process called *boosting* and has been adopted in various fields inside and outside onboard applications, such as household electrical appliances.

Boosting uses weak classifiers that have a low recognition rate but are simple and have a low calculation load. Weak classifier points are used to compensate each other to construct single strong classifiers. The Viola-Jones method is a typical example that uses boosting (Viola and Jones, 2001). This method constructs a strong classifier by forming a cascade of weak classifiers arranged in sequence. It stops the recognition process as soon as an object is judged not to be the target object, thereby enabling a rapid process with a high recognition rate.

In addition, high recognition performance requires optimization of the image features used by boosting, in accordance with the recognition objects. Although the Viola-Jones method uses Haar-like features, recent proposals include the use of histograms of oriented gradients (HOGs) (Dalal and Triggs, 2005), an improved version of HOG called *co-occurrence histograms of oriented gradient*s (*CoHOG*s) (Watanabe, Ito, and Yokoi, 2009), and shapelet features (Sabzmeydani and Mori, 2007). In actual fact, the features used in boosting, the method of constructing strong classifiers, and the like are optimized in accordance with the application. Boosting is used in several examples of vehicle and pedestrian recognition, as well as in facial recognition using digital cameras.

5 RECOGNITION OF VISUAL ENVIRONMENT (VISIBILITY RECOGNITION)

Recently, sensing camera technology has begun to be developed that is capable of recognizing various objects in the driving environment, such as vehicles, pedestrians, traffic signs, traffic signals, and the like. However, if the driver is informed of all detected objects, the processing capability of the driver will be overwhelmed and the recognition technology itself will become a cause of driver distraction (Pettitt, Burnett, and Stevens, 2005). For this reason, the system must judge whether to inform the driver in accordance with the state or appearance (visibility) of the object. Research is currently in progress to evaluate the visibility of objects and the environment based on this approach. This section describes the application of visibility to traffic signs and the driving environment as typical examples of visual environment recognition technology.

5.1 Visibility of traffic signs

Figure 6 shows an example of traffic sign visibility. Some signs are located in complex backgrounds and are difficult to identify at a single glance. The type of sign may also be difficult to visualize due to the light levels or fading. Types and positions of signs that can be identified instantly are referred to as having high visibility. In these cases, the driver is notified only of traffic signs with low visibility to prevent driver distraction and provide the minimum necessary information.

More specifically, the visibility of an object is an integrated property that describes how easy it is to see and how easy it is to understand. Object visibility estimation technology based on this property is being researched on following two approaches: focusing on background and contrast (Kimura, 2010) and focusing on the appearance of the object itself (Simon, Tarel, and Bremond, 2009)–(Maerz and Niu, 2003). The former uses the idea that the complexity of the background has a large impact on

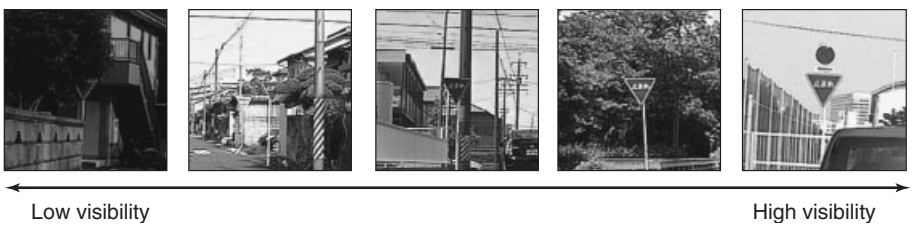


Figure 6. Visibility of traffic signs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto224
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

object visibility. This method evaluates visibility based on contrasting complexity. The latter uses the idea that the visibility of an object is affected by the brightness and its color distribution, and evaluates visibility based on the brightness of the image, the red–green–blue (RGB) components, color histograms, and the like. An onboard camera must factor in both of these two approaches when estimating the visibility of objects. This is because the positional relationship of the driver, object, and background changes as the driver's vehicle moves, because the complexity of the background is constantly changing, and because the brightness of the environment also changes as the vehicle turns and so on. In addition, as the apparent size of the object (i.e., its size in the image) also changes as the vehicle moves, the effect of time variations must also be considered.

On the basis of these characteristics of onboard cameras, Doman *et al*. have proposed a method of estimating the visibility of traffic signs using multiple temporally connected image groups (Doman *et al*., 2011). This method first extracts the traffic signs from the images captured by the onboard camera and then evaluates the visibility at that time (i.e., an instantaneous visibility value). This instantaneous value is expressed by a linear combination of five types of features: the color of the sign and background, the edges, the texture contrasts (each average color, average edge intensity, and color histogram difference), the image quality (difference with template image), and the size in the image. This instantaneous value is calculated for every captured image containing a traffic sign and the final traffic sign visibility is obtained by determining an average within a predetermined time.

5.2 Visibility of driving environment

The previous section covered object visibility. Research is being carried out that expands on these ideas to evaluate the visibility of the driving environment (Nitanda *et al*., 2011). This technology has potential application in driving assistance systems and systems that adapt to timing and intensity in accordance with visibility.

This section only covers a method that estimates the visual burden of the driver using a forward monitoring monocular camera (Nitanda *et al*., 2011). This method first calculates a saliency map (proposed by Itti *et al*.) for the images from the forward monitoring camera (Itti, Koch, and Niebur, 1998). This saliency map is a calculation model that uses physiological information to extract visual features that are different from the surrounding area as salient regions to predict the target of the driver's gaze. The use of this model in onboard camera images is equivalent to predicting the direction of gaze of someone sitting in the front passenger seat, who is looking forward without any particular purpose. This gaze movement process depends on features such as color and brightness when the observer glances at the surroundings. It is referred to as a *bottom-up process*.

In contrast, in actual driving, the driver focuses on the front of the vehicle and alters the gaze direction to obtain the necessary information when objects such as other vehicles, pedestrians, traffic signals, traffic signs, and the like appear, whose position or status has to be confirmed before a driving action is carried out. In this process, gaze movement is determined based on the purpose of the observer or the existence of problems. It is referred to as a *top-down process*.

These bottom-up and top-down processes can be used to roughly predict the driver's gaze movement and to estimate the visual burden. Figure 7 shows the system model. In this model, the gaze position of the driver is predicted by the sum of the bottom-up and top-down processes. Here, the bottom-up process expresses the physiological gaze movement of the driver, whereas the top-down process expresses the intentional gaze movement. The theory behind this system states that the gaze position should exist within one of these processes. In addition, as safe driving requires the driver to pay attention to the front of the vehicle, the physiological gaze movement must be reduced. On the basis of the idea that the suppression of physiological reactions is a way of defining driver burden, the visual burden can be estimated from the difference between the bottom-up and top-down processes.

Figure 8 shows examples of the bottom-up and top-down processes, and a calculated image that shows the differential between the bottom-up and top-down processes. This differential image is shown in Figure 8d, which shows the calculated results for visual burden as the total brightness value. Figure 8d demonstrates that the system reacts strongly to objects such as electronic shop signs that naturally attract the driver's gaze, although the driver does not need to look at them for driving. A high correlation coefficient of 0.90 has also been reported between the visual burden calculated from five types of images at night and subjective evaluations by test subjects (Nitanda *et al*., 2011).

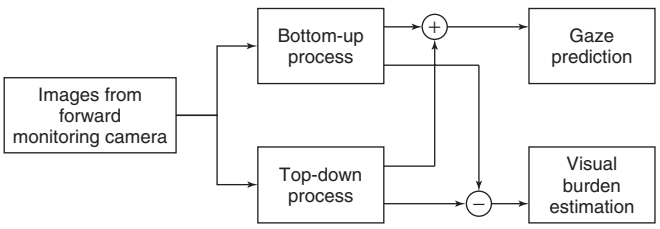


Figure 7. Gaze prediction and visual burden estimation model.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto224
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

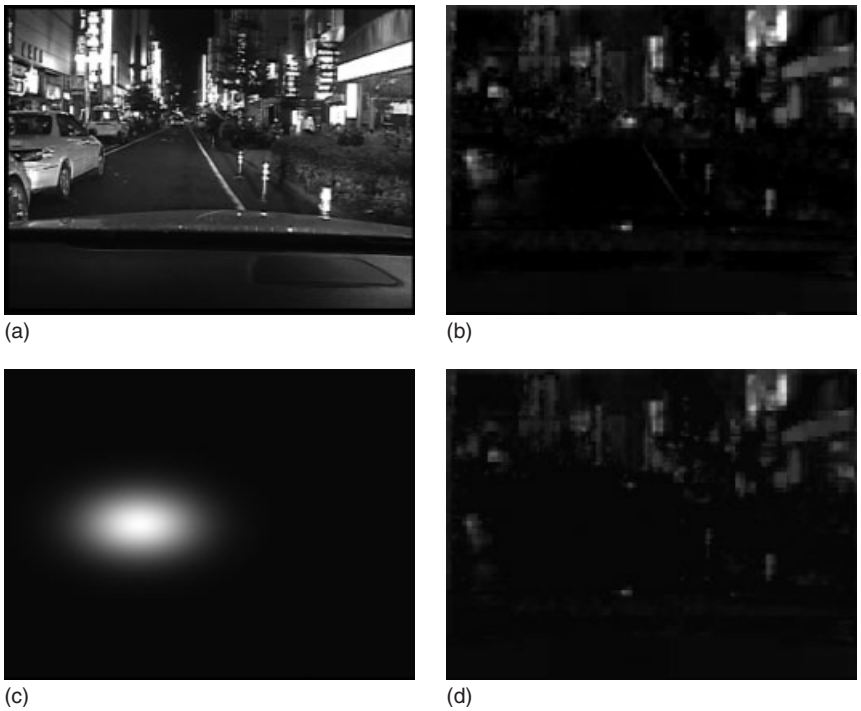


Figure 8. Visibility of driving environment. (Reproduced with permission from Nitanda *et al*., 2011. © Naoki Nitanda *et al*.)

6 CONCLUSION

This chapter has outlined typical examples of technology that senses and recognizes the circumstances around the vehicle using a sensing camera from the standpoints of object recognition and visual environment recognition. The future is likely to see the development of cooperative processes with other sensors and demands to share recognition results and create recognition databases by integrating, accumulating, and communicating sensing and recognition information. As a result, it will be necessary to develop even more sophisticated technology for sensing and recognizing the circumstances around the vehicle as vehicles grow even more dependent on information.

REFERENCES

Dalal, N. and Triggs, B. (2005) Histograms of Oriented Gradients for Human Detection. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 886–893, IEEE: USA.

Doman, K., Deguchi, D., Takahashi, T. *et al*. (2011) Estimation of Traffic Sign Visibility Considering Temporal Environmental Changes for Smart Driver Assistance. *Proceedings of IEEE Intelligent Vehicles Symposium (IV)*, 667–672, IEEE: USA.

Hirschmuller, H. (2005) Accurate and efficient stereo processing by semi-global matching and mutual information. *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 807–814, IEEE: USA.

Itti, L., Koch, C., and Niebur, E. (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **20**(11), 1254–1259. IEEE: USA.

Kanade, T., Yoshida, A., Oda, K. *et al*. (1996) A Stereo Machine for Video-rate Dense Depth Mapping and its new Applications. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 196–202, IEEE: USA.

Kimura, F. (2010) Development of a quantification method of traffic signal visibility for driver assistance. *IEEJ Transactions on Electronics, Information and Systems*, **130-C**(6), 1034–1041. Institute of Electrical Engineers of Japan: Japan.

Lowe, D. (1999) Object Recognition from Local Scale-invariant Features. *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 1150–1157, IEEE: USA.

Maerz, N. and Niu, Q. (2003) Automated mobile highway sign retroreflectivity measurement. *Final Rep. NCHRP-IDEA Project 75*.

Nitanda, N. *et al*. (2011) Method of estimating visual burden in driving environment. *Meeting on Image Recognition and Understanding (MIRU)*, 697–702, MIRU: Japan.

Pettitt, M., Burnett, G. and Stevens, A. (2005) Defining Driver Distraction. *Proceedings of 12th World Congress on Intelligent Transport Systems*, 1–12, Curran Associates, Inc: USA.

Sabzmeydani, P. and Mori, G. (2007) Detecting Pedestrians by Learning Shapelet Features. *Proceedings of IEEE Conference on*

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto224
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Computer Vision and Pattern Recognition (CVPR), 1–8, IEEE: USA.

Simon, L., Tarel, J.P., Bremond, R. (2009) Alerting the Drivers About Road Signs with Poor Visual Saliency. *Proceedings of IEEE Intelligent Vehicles Symposium (IV)*, 48–53, IEEE: USA.

Takaki, M., *et al*. (2009) Road sign recognition using SIFT feature. *The Institute of Electrical Engineers of Japan*, **129**(5), 824–831. Institute of Electrical Engineers of Japan, Japan.

Viola, P. and Jones, M. (2001) Rapid Object Detection using a Boosted Cascade of Simple Features. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 511–518, IEEE: USA.

Watanabe, T., Ito, S., Yokoi, K. (2009) Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection. *Proceedings of 3rd Pacific-Rim Symposium on Image and Video Technology*, PSIVT 2009, Tokyo, Japan, January 13–16, 2009, published in *Lecture Notes in Computer Science*, Vol. 5414, SBN: 978-3-540-92956-7, Springer: Berling, Germany.

Yamaguchi, K., Kato, T., Ninomiya, Y. (2006) Moving Obstacle Detection using Monocular Vision. *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, 288–293, IEEE: USA.

FURTHER READING

Siegmann, P., S. Lafuente-Arroyo, S. Maldonado-Basc (2005) Automatic Evaluation of Traffic Sign Visibility using SVM Recognition Methods. *Proceedings of 5th WSEAS International Conference on Signal Processing, Computational Geometry & Artificial Vision*, 170–175.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto224
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Engine ECU Systems

Yukihide Niimi
DENSO Corporation, Kariya, Japan

1 Introduction 1
2 Engine ECU 1
3 Engine Control 9
4 The Future of Engine ECU 15
Related Articles 16
References 16

1 INTRODUCTION

In modern vehicles, the engine is controlled by the engine control unit (ECU) system. This system has developed in conjunction with tightening of exhaust emissions and fuel economy regulations. In the future, these trends will likely continue. The performance requirements of the engine ECU itself will also continue to increase from the standpoints of promoting greater safety, comfort, and protecting the environment. These are the three major requirements of vehicles in recent years.

The engine control parameters include the quantity of intake air, the amount of fuel injection, the ignition timing, and the opening and closing timing of the intake and exhaust valves. All of these parameters are controlled by the engine ECU. This chapter describes the hardware and software of the engine ECU system. It also explains the fuel injection control system, ignition control system, and electronic throttle control system as representative examples of the engine controls performed by the engine ECU system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 ENGINE ECU

2.1 Overview of engine control

The engine system (Kato, 2010a) provides optimal control of combustion-related parameters in accordance with the operating conditions of the engine to improve engine output and fuel economy, and to achieve cleaner exhaust emissions. There are two types of engine control: gasoline engine control and diesel engine control. This chapter uses gasoline engine control in the explanations of the engine ECU system.

The gasoline engine control parameters include the amount of intake air, the amount of fuel injection, the ignition timing, and the opening and closing timing of the air intake and release valve. All of these parameters are controlled by the engine ECU.

Figure 1 shows a diagram of a gasoline engine control system. The system is composed of sensors and actuators, whereas the ECU is the nucleus of the engine control. The sensors detect information such as the engine speed and the depression angle of the accelerator pedal. The ECU then takes in this information, makes calculations, and controls each of the actuators, starting with the throttle valve, so that the engine is in the optimal state.

The following are the main controls of the ECU: electronic throttle control that controls the amount of intake air, idle speed control, fuel injection control that controls the amount of fuel injection, ignition timing control and knock control, variable valve timing control for the air intake and exhaust valve, and the variable cylinder management system that disables certain cylinders. Additional controls of the ECU include the evaporator control that inhibits the escape of evaporated gasoline from the fuel tank, exhaust gas recirculation (EGR) control that recirculates exhaust emissions and reduces nitrogen oxides (NOx),

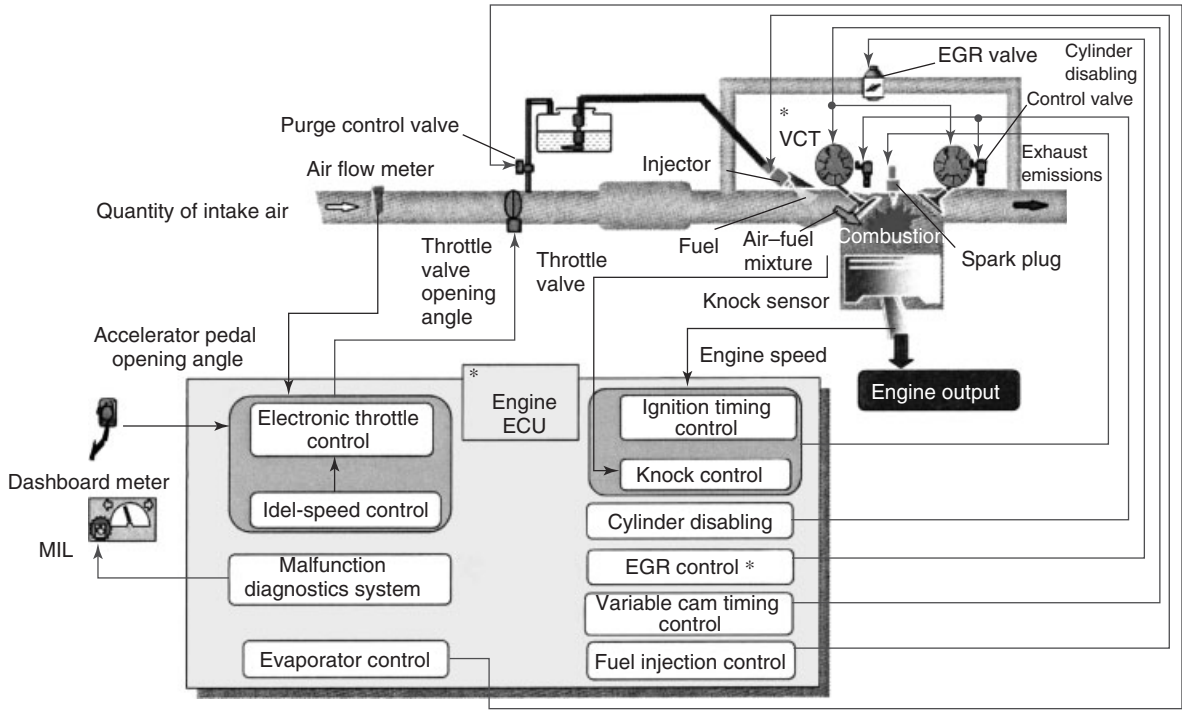


Figure 1. Electronic control system of gasoline engine.

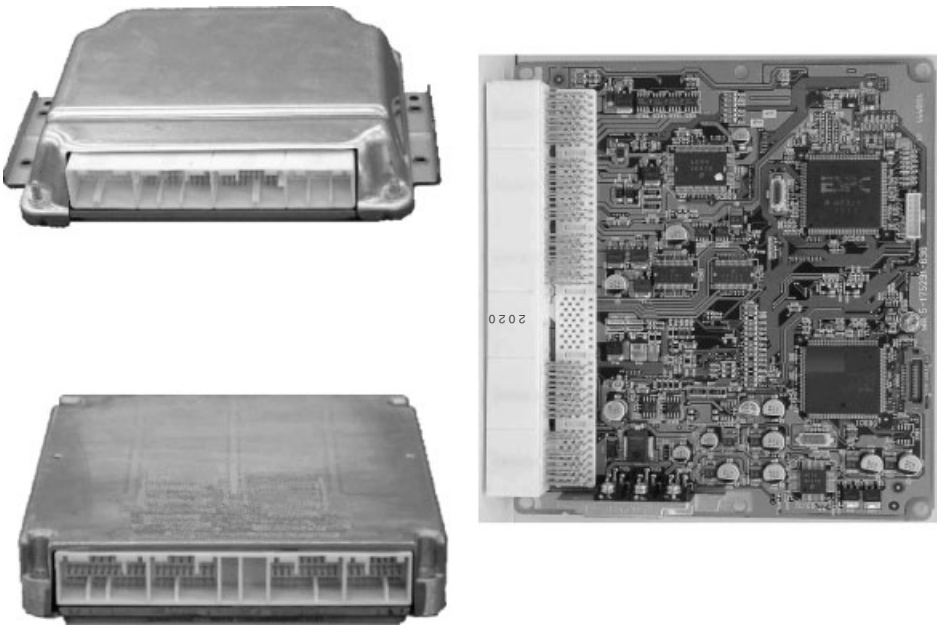


Figure 2. Interior and exterior of engine ECU.

the malfunction diagnostics system, and the like. The ECU is normally covered by a case made from aluminum alloy, with the microcomputers and various electronic parts located inside, as shown in Figure 2.

2.2 Evolution of engine ECU

This section describes the evolution of the engine ECU. Exhaust emissions regulations started to be implemented

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Engine ECU history.

	1970s	1980s	1990s, first half	1990s, second half
Regulatory trends	Start of exhaust emissions regulation	Regulations strengthened in California US fuel economy regulations	OBD regulations in California	ULEV, ZEV regulations in California, European regulations, POST 53
Engine trends	Carburetors are mainstream, engine performance sacrificed to satisfy regulations	EFI becomes widespread Engine performance improved Engine power competition	EFI is commonplace Competition to lower cost Differentiation	Development of alternative fuels, EV Lean burn engine becomes widespread Direct injection gasoline engine appears
Control method	Independent control of fuel injection	Concentrated control (Fuel injection, ignition, etc.)	Total control (Engine control + transmission)	Total control (+ throttle control)
Electronic technology	Analog IC	8- to 12-bit microcomputers	16-bit microcomputers	32-bit microcomputers
Software		Function partitioning 4 to 12 kB	Structured (C language) 32 to 96 kB	Object oriented RTOS 128 k to 1 MB

ULEV: ultra-low emission vehicle
SULEV: super ultra-low emission vehicle
ZEV: zero emission vehicle

in the 1970s. At that time, the main types of engines still used carburetors, so engine performance was sacrificed to comply with these new laws and regulations (Table 1).

Electronically controlled fuel injection became widespread in the 1980s in response to tighter emissions regulations in the United States. These regulations were satisfied using microcomputers to control fuel injection and ignition, as well as to improve engine output. By the first half of the 1990s, most vehicles were equipped with electronic fuel injection control and the ECU performed integrated control of both the engine and the transmission. In the latter half of the 1990s, the low exhaust emissions regulations concerning ultra-low emission vehicle (ULEV), super ultra-low emission vehicle (SULEV), and zero emission vehicle (ZEV) were strengthened even further. Alternative fuels and electric vehicles (EVs) were developed and direct injection gasoline engines also appeared. Integrated control by the ECU is steadily becoming more complicated. It was a common practice to install the engine ECU within the occupant compartment, such as behind the instrument panel, to protect the electronic parts from the high temperatures of the engine and the electric noise from ignition (Figure 3). However, the limitations on installation space in the occupant compartment became more severe over time, as more interior space was required to ensure passenger comfort. As a result, in recent years, the engine ECU is increasingly installed in the engine compartment (Kato, 2010b).

The conditions within the engine compartment are very severe for the electronic parts that make up the ECU. If the ECU is near the engine, it is exposed to high temperature heat sources during operation and directly receives vibrations from the engine. There is also a lot of electric noise, mainly produced by the ignition within the engine, and the presence of water spray. There are examples of the ECU being placed inside a plastic box to protect it when it is installed in the engine compartment. Some engine ECUs may be installed directly on the engine and the development of special electronic parts that can adequately withstand this severe environment is now under way.

As described in Section 2.1, the engine control ECU installation environment is becoming increasingly severe. Figure 4 compares this environment with that of other electronic parts. The vertical axis shows the required level of reliability and safety, whereas the horizontal axis shows the temperature.

Focusing on the temperature environment, it is common for household appliances, such as personal computers and televisions, to be used indoors where the temperature range is limited. However, vehicles are characterized by a much broader range of operating temperature, and it is assumed that they will be driven in outdoor environments all over the world. On the other hand, if the required level

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

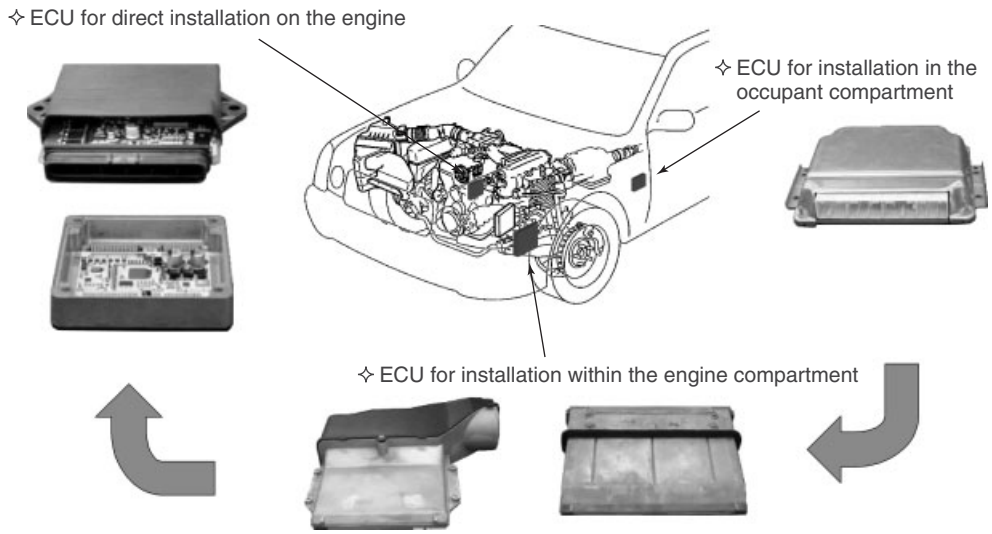


Figure 3. Changes in installation environment of engine control ECU.

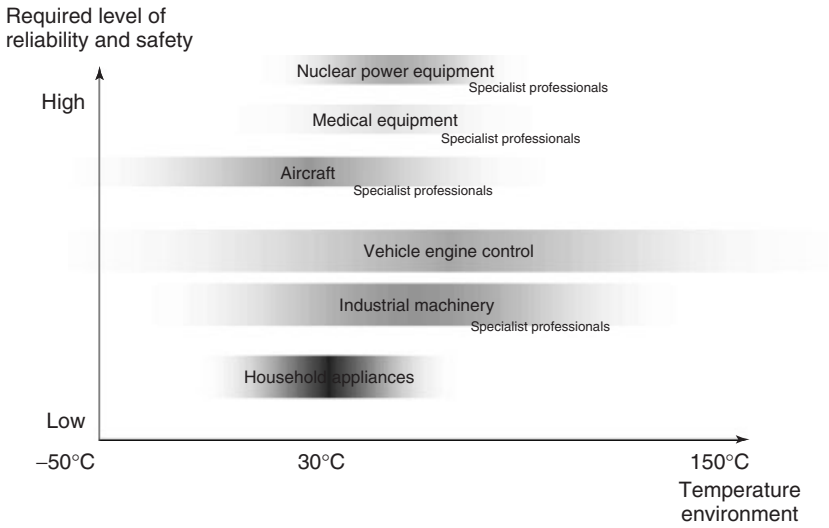


Figure 4. Required level of quality.

of reliability and safety is considered, vehicles have the potential to immediately threaten human life by not moving (not starting), going out of control, or catching fire. For these reasons, the required level of safety is quite high. Vehicles are also often used for many years, so reliability over a long operating life is another requirement. Vehicle users often do not possess any specialized knowledge or specialized skills (i.e., they are not specialist professionals). Therefore, it is crucial to consider how best to ensure the safety of users, even if they operate the vehicle in a manner that is unexpected.

2.3 Circuit configuration of engine ECU

Figure 5 shows the function blocks (Kato, 2010b) of the engine ECU. The engine ECU uses the calculations of a microcomputer to realize optimum control in response to the vehicle state. It possesses a means to receive inputs that detect the external conditions as an external interface, and a means to supply outputs that actually drive the controls. In recent years, many ECUs have also been equipped with a means of communicating with each other to promote closely coordinated control.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

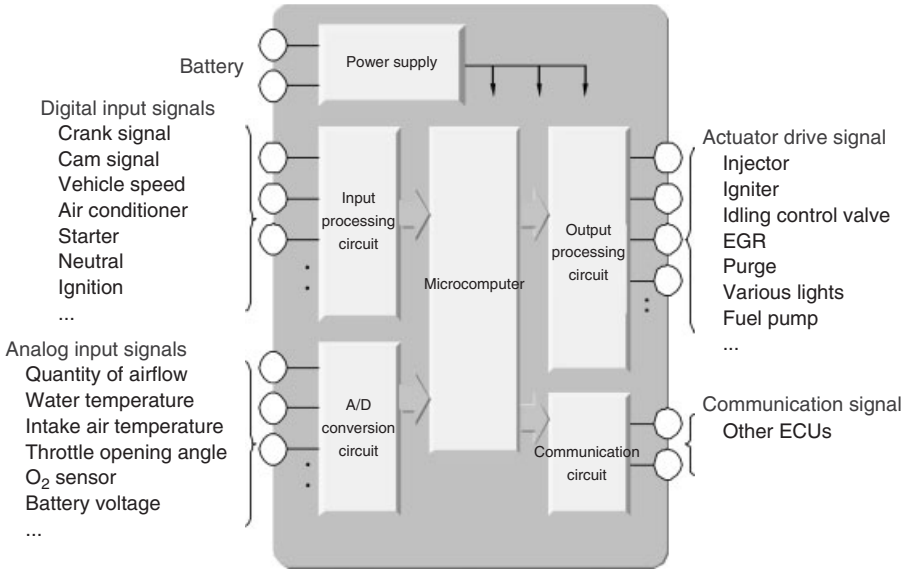


Figure 5. Function block diagram.

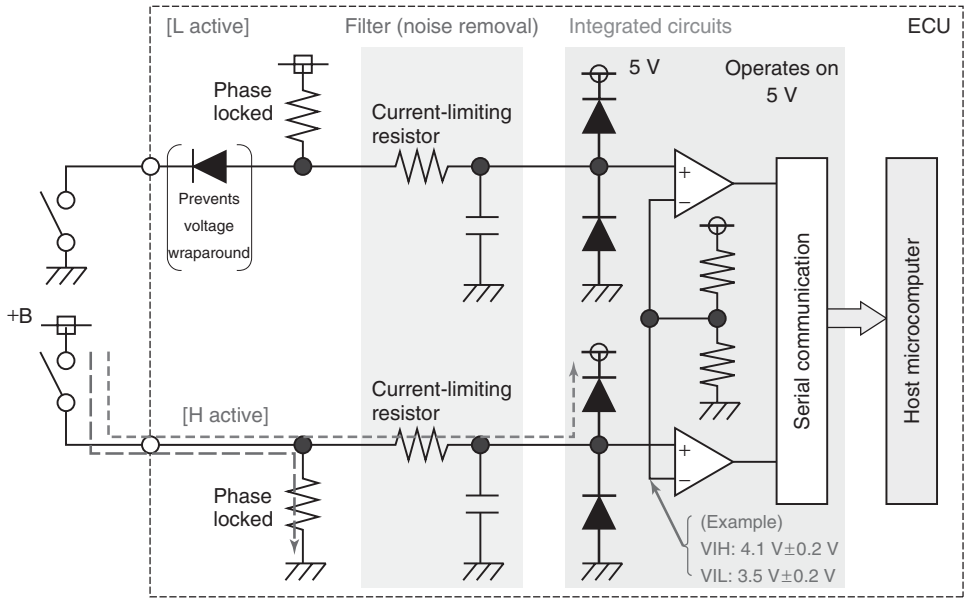


Figure 6. On and off inputs.

The following sections describe representative examples of the input processing circuit, output processing circuit, and communication circuit interface.

2.3.1 Input processing circuit

Figure 6 shows an example of a circuit used to input the "on–off" condition to the microcomputer in accordance with the switch operations of the driver.

The resistance for phase locking ensures the input voltage when the switch is open. In the case of an ECU in a poor usage environment where there is a lot of electric noise, the ECU must be equipped with a proper filter circuit and other mechanisms that prevent noise penetration. The software in the microcomputer is also protected to prevent improper operation because of noise by implementing noise removing functions and other means.

In the environment created by various devices installed on a vehicle, different input voltages exist side by side

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

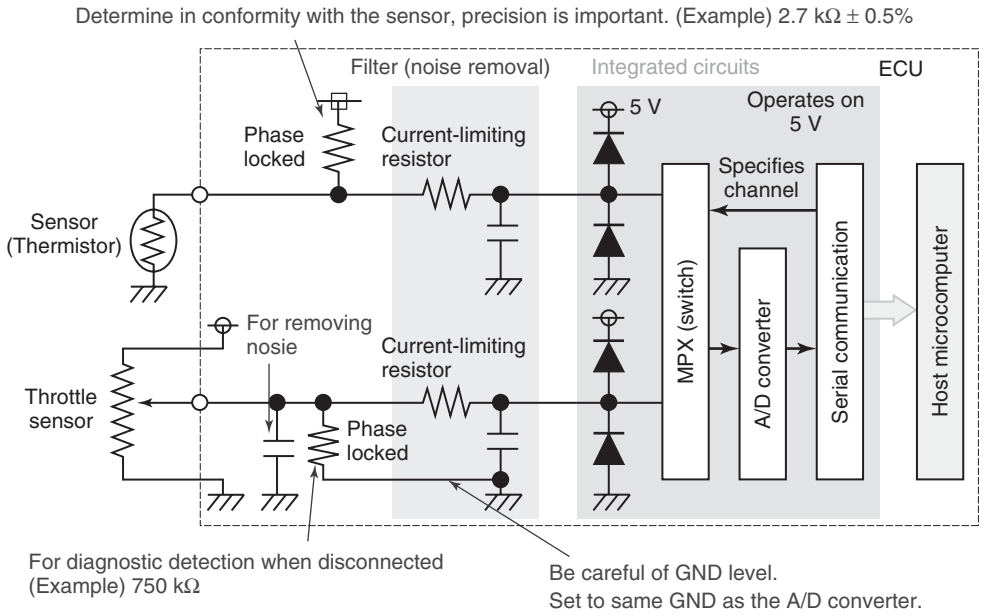


Figure 7. Analog inputs.

(such as the battery voltage and ECU internal power supply voltage). These different voltages have to be separated. Multiple inputs are collected together and then input into the microcomputer using serial communication for the purpose of handling the numerous on–off inputs. The adoption of integrated circuits (ICs) is being promoted in an effort to reduce the size of the ECU.

Figure 7 shows an example of a circuit used to input the "level" condition to the microcomputer of items such as the temperature of the engine coolant and the throttle operations of the driver.

The resistance for phase locking ensures the input voltage from the sensor. It also ensures the input voltage when the sensor is disconnected, so that abnormalities can be detected. Detecting a vehicle abnormality and then warning the driver is an extremely important function of the ECU.

Analog inputs are converted into digital signals that can be used by the microcomputer through analog/digital (A/D) conversion. In recent years, demand has grown for A/D converters with very high conversion performance (conversion speed and conversion accuracy) because of the increasing precision of control.

2.3.2 Output processing circuit

Figure 8 shows an example of the circuit that communicates with and drives the fuel injector based on the fuel injection pulse sent to the engine. The pulse is determined from the control calculations of the microcomputer.

2.3.3 Communication interface

The communication signal of the communication bus is input into the microcomputer, so that data can be exchanged with another ECU. Figure 9 shows an example of the circuit for controller area network (CAN) communication that is output from the microcomputer.

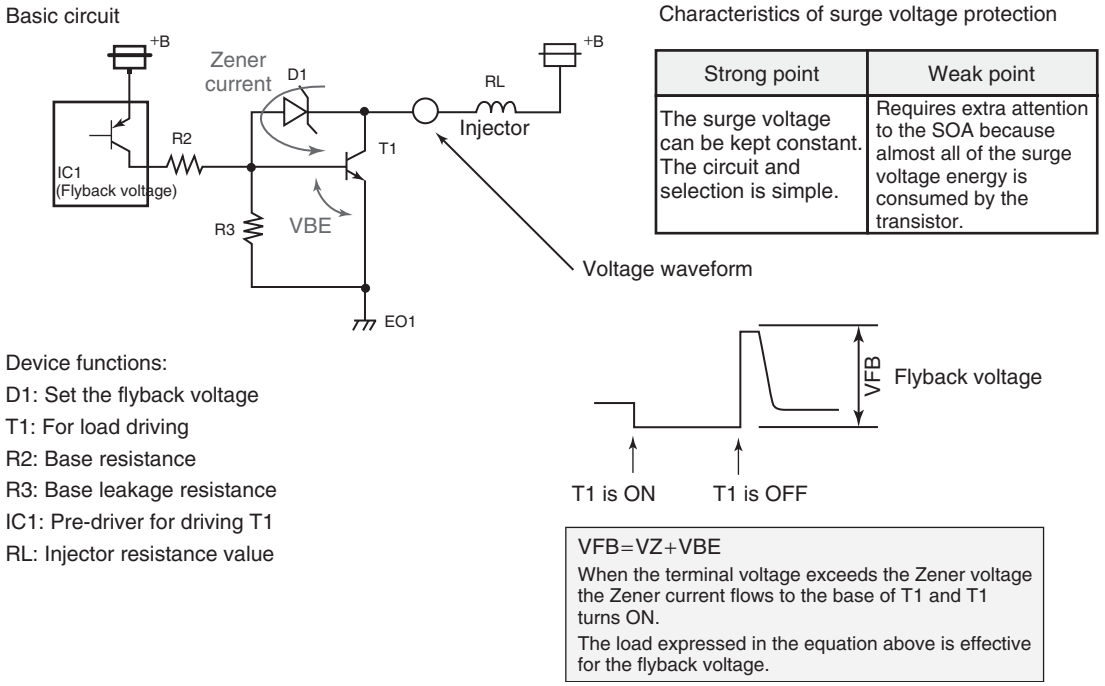
2.4 Microcomputer

The microcomputer for engine control is required to operate for a long time and must also operate under some of the harshest conditions in the world, from the high temperature regions in the Near and Middle East to low temperature regions such as Alaska. As a result, it must have an operating temperature range −40 to 125°C and a service life of 20 years or more (Figure 10).

The microcomputer is equipped with an A/D converter for sensor input processing, a highly precise timing function for injector output processing, and a serial function for use with the malfunction diagnostics tools in the marketplace. It is also equipped with a debug function to develop and check these other functions.

The large-capacity memory in the microcomputer for increased performance must be rewritable flash memory to satisfy laws and regulations. Finally, the microcomputer is required to realize this high quality and high performance at a low cost. The read-only memory (ROM) capacity of the microcomputer for engine control was

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Characteristics of surge voltage protection

Strong point	Weak point
The surge voltage can be kept constant. The circuit and selection is simple.	Requires extra attention to the SOA because almost all of the surge voltage energy is consumed by the transistor.

Figure 8. Fuel injector communication and drive circuit.

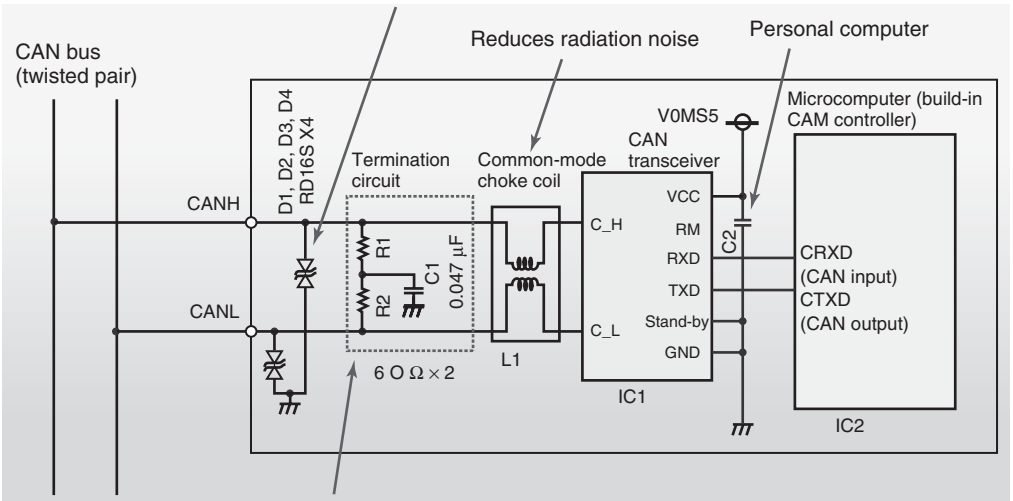


Figure 9. CAN communication circuit.

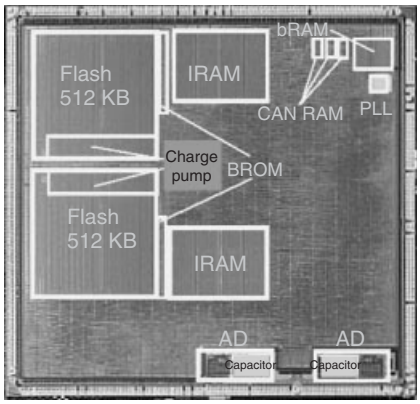
approximately 4 kB in the 1980s, but in response to the introduction and strengthening of the laws and regulations previously described in this chapter, in 2010, the capacity had increased to exceed 2 MB. This trend is fully expected to continue in the future, so microcomputers will be required to become even smaller in size and run at even faster speeds.

2.5 Software

2.5.1 Trends in engine control software

The use of microcomputers to control the engine began in the 1970s–1980s, but this trend really started to accelerate due to the introduction of regulations for exhaust emissions and fuel economy. The size of the software (Kato, 2010b)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



- High reliability
  - Mounting conditions: –40°C to 125°C
  - Service life: 20 years or more
- Highly functional
  - A/D conversion performance: Sensor input processing
  - Highly precise timing function: For injector output, etc.
  - Serial communication function: Malfunction diagnostics tool communication
  - Debug function
- High performance
  - Large-capacity memory: Flash memory
- Low cost

Figure 10. Requirements of microcomputer for engine control.

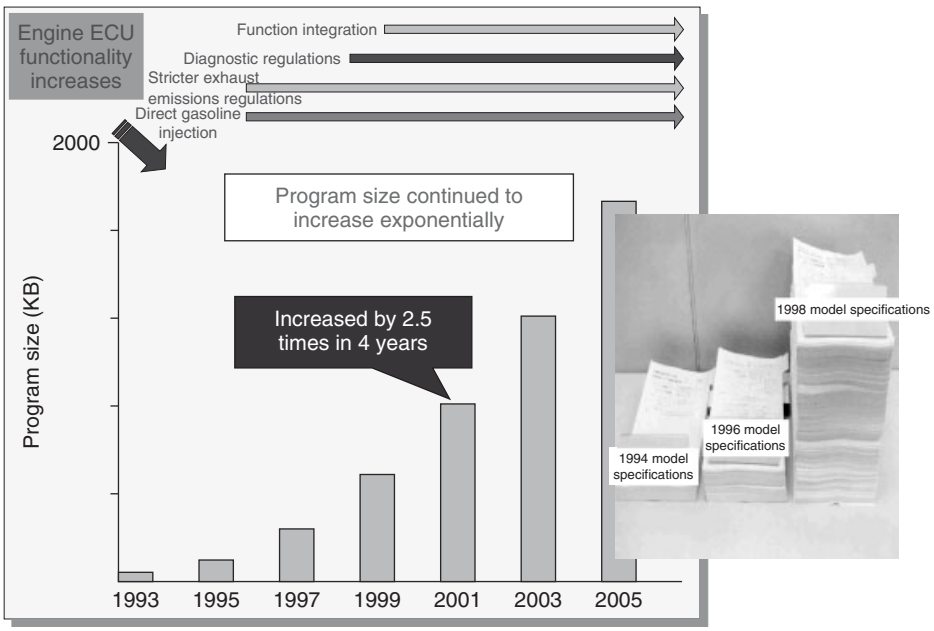


Figure 11. Increasingly sophisticated and larger scale software.

for engine controlling microcomputers has continued to grow, as described in this chapter, and the engine controls have also continued to increase steadily. Taking a representative piece of engine controlling software as an example, the program size in 2005 was 2.5 times larger than it was in 2001 and this trend is also expected to continue in the future (Figure 11).

2.5.2 *Engine control software operations*

The structure of engine control software is often divided into the application layer, which is the software that realizes the control, and the platform layer, which drives the sensors and actuators. This gives the software the flexibility to be adapted to the different regulations in each country and to optional vehicle parts. Figure 12 shows fuel injection control as an example to outline and explain the operations of the engine control ECU software.

(a) The required amount of fuel injection is calculated from the engine speed and engine load in the application layer. This amount is then transmitted to the sensor and actuator layer.
(b) The program in the platform layer that determines injector operation adjusts the required amount of fuel injection according to the injector characteristics

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

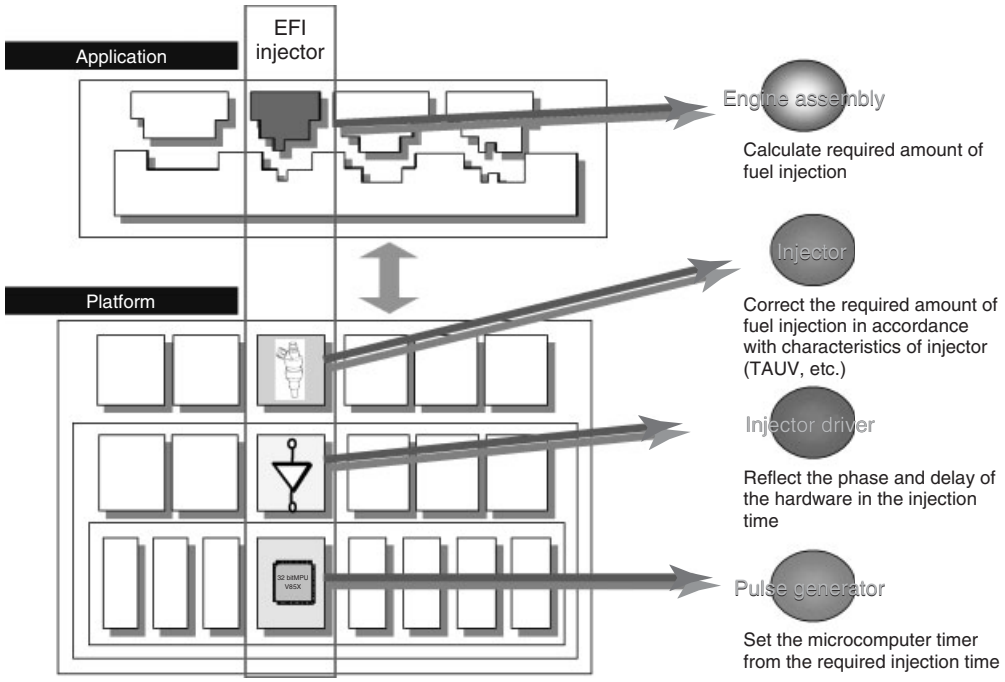


Figure 12. Outline of fuel injection software operations.

(invalid injection time and so on) and corrects the fuel injection time. It then transmits this information to the injector driver in the ECU layer.

(c) The injector driver reflects the phase and delay of the hardware in the corrected fuel injection time. This is then transmitted to the pulse generator in the CPU layer.

(d) The pulse generator sets the initial value of the microcomputer timer as the required fuel injection time. Structuring the software parts in this manner means that it is only necessary to change the pulse generator if the microcomputer is changed and the other parts can continue to be used as is. This improves the reusability of the software.

2.5.3 *Software development process*

A special characteristic of the engine control ECU software development process in comparison to that of general software development (IPA of Japan, 2008) is the use of a calibration process (Figure 13). In this calibration process, the parameters within the program are set to the optimum values and confirmation carried out to verify that the control logic was realized by the ECU. The optimum values are determined by examining a variety of different factors, such as the cleanliness of the exhaust emissions, fuel economy, engine output, and the like.

2.5.4 *Model-based development*

There are many actuators and sensors in the engine control system. If software development did not begin until all of those were completed, it would take a long time to develop the whole system. Therefore, in recent years, a method called *model-based development* (Otsuka and Oi, 2006) has been used. Stated simply, model-based development is software development based on simulation technology. Figure 14 shows the main simulation technology. The control logic can be examined using simulations before the sensors, actuators, and ECU hardware are completed. This drastically shortens the development period.

Creating a model of the problem region is a necessary prerequisite of model-based development. Block diagrams are often used in software development for controls and the MATLAB and Simulink tools of MathWorks, Inc. in the United States are the de facto standards for computer-aided software engineering (CASE) tools (Figure 15).

3 ENGINE CONTROL

This section outlines fuel injection control, ignition control, and electronic throttle control systems as three representative examples of engine control. This is preceded by an explanation of basic engine operations. In a gasoline engine, an air–fuel (A/F)mixture is created and supplied inside the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

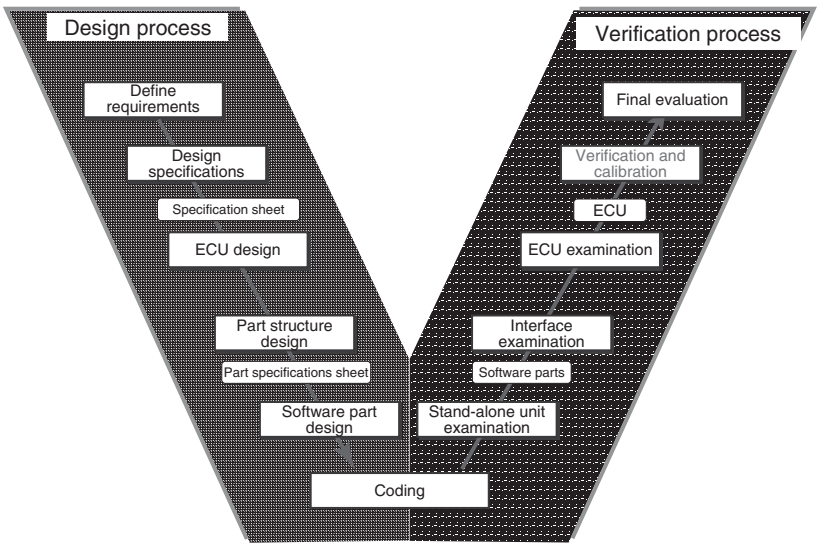


Figure 13. Engine ECU software development process.

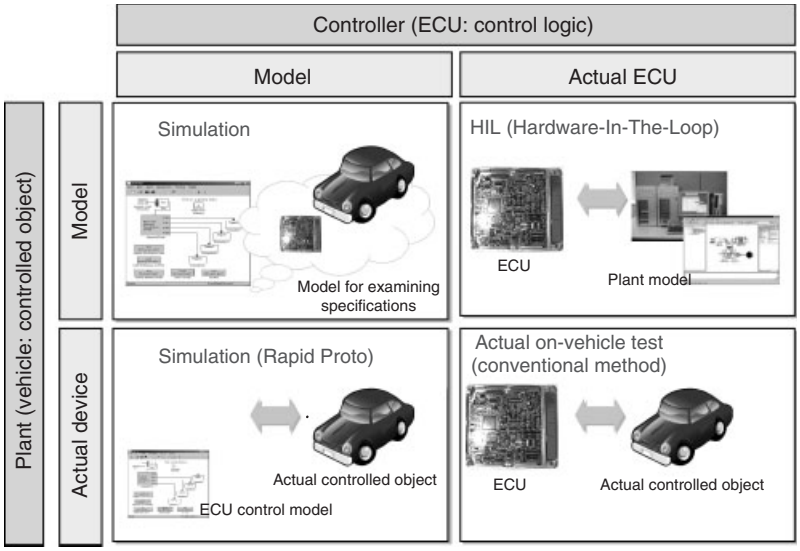


Figure 14. Main simulation technology used in model-based development.

cylinders. Heat energy is produced by causing this A/F mixture to combust. The heat energy pushes on pistons and the movement of the pistons is turned into torque, also called *rotating force*.

Figure 16 shows a diagram of a normal four-stroke engine. These four strokes are air intake, compression, expansion, and exhaust. In the air intake stroke, the piston starts at the top dead center (TDC) and then drops, causing the air intake to start. Once the piston reaches the bottom dead center (BDC) and begins to rise again, the air intake valve closes and the compression stroke begins. The A/F mixture in the cylinder is compressed until just before the piston reaches TDC again. The spark plug fires at that moment and ignites the A/F mixture. The combustion energy that is produced pushes the piston back down and this is the expansion stroke. The exhaust valve opens when the piston drops to the BDC and the exhaust stroke begins.

This means that combustion occurs once in a given cylinder for every 720° that the crank rotates in a four-stroke engine. In the case of the four-cylinder engines that are often used in compact cars, the combustion process occurs four separate times, once in each cylinder, so combustion occurs each time the crank rotates 180°.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

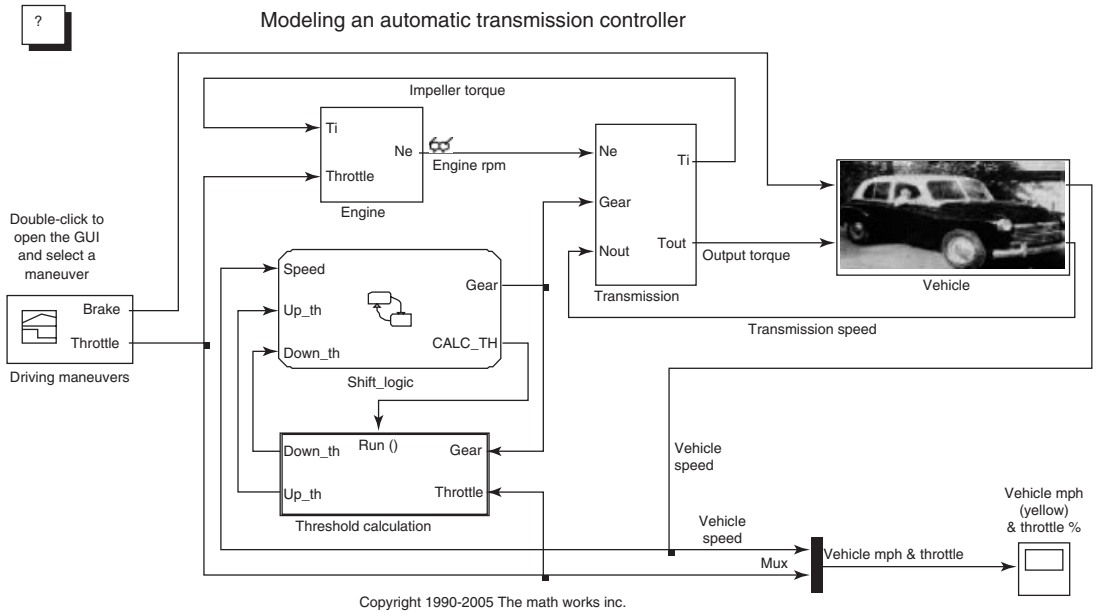


Figure 15. Example of Simulink model (automatic transmission controller).

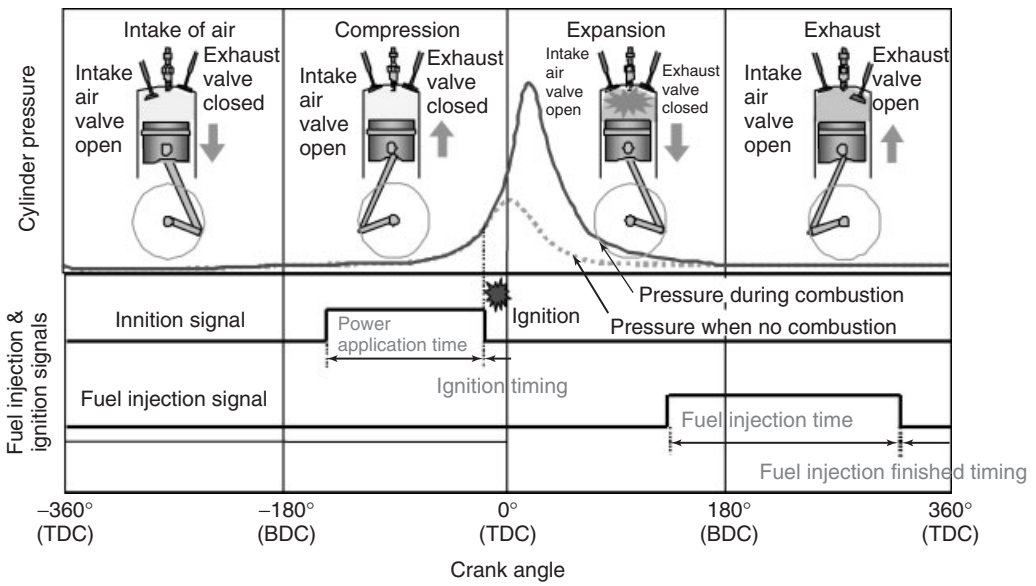


Figure 16. Combustion process of four-stroke engine.

3.1 Outline of fuel injection control

3.1.1 Controlling amount of fuel injected and injection time

This section discusses the control of the amount of fuel that is injected and the injection time in a port-injection-type gasoline engine. Figure 17 shows a general gasoline fuel injection system.

The engine ECU calculates the amount of fuel to inject as the injection time (T_i) and then sends this information to the injector. Three elements are used to determine this injection time (T_i): the basic injection time (T_p), the amount of feedback correction, and the transient correction. First, the necessary amount of fuel for a single combustion is calculated from the quantity of intake air (G_a) and the engine speed (N_e) as the basic injection time (T_p). After this,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

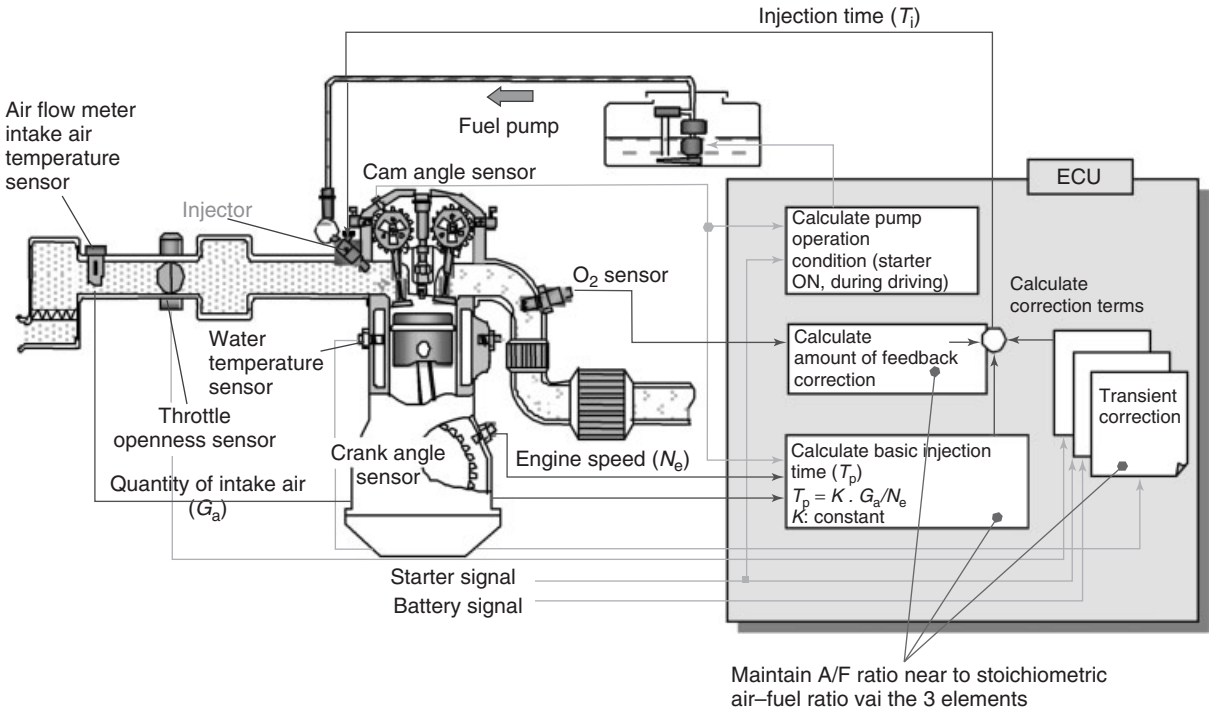


Figure 17. Fuel injection control of port-injection type gasoline engine.

the amount of feedback correction from the O_2 sensor or any difference from the target A/F ratio during the transition is added to bring the A/F ratio into line with the target stoichiometric A/F ratio.

There are also cases where the amount of fuel that is injected is increased, such as when the engine is started and when high engine output is necessary. Table 2 summarizes the concepts behind control that is performed in accordance with the operating conditions, such as before and after engine start.

3.1.2 Air–fuel ratio feedback control

The exhaust emissions from the engine contain toxic substances such as hydrocarbons (HCs), carbon monoxide (CO), and NO*x*. A catalyst is used as part of an aftertreatment process for the exhaust emissions to reduce the amounts of these toxic substances. Currently, a three-way catalyst is the most common catalyst that is used. A three-way catalyst simultaneously causes an oxidation reaction with CO and HC and a reduction reaction with NO*x*, so that they are converted into nontoxic carbon dioxide (CO_2), water (H_2O), and nitrogen (N_2), respectively. This catalyst possesses excellent conversion efficiency characteristics when the A/F ratio of the exhaust emissions (λ) is controlled and kept within a narrow range near to the stoichiometric A/F ratio (Figure 18).

Always maintaining the A/F ratio near to the stoichiometric A/F ratio ($\lambda = 1$) is necessary to keep the catalyst operating efficiently, even when the driving conditions change. The O_2 sensor installed on the exhaust manifold is used for this purpose. The O_2 sensor detects the concentration of oxygen in the exhaust emissions and the output voltage of the sensor changes greatly depending on this concentration with the stoichiometric A/F ratio as the dividing line. A judgment is made about whether the A/F mixture is rich or lean based on the output voltage that is detected and this is used as A/F ratio feedback to correct the amount of fuel injection. However, the voltage of the O_2 sensor changes abruptly with the stoichiometric A/F ratio as the dividing line. Therefore, a linear A/F ratio sensor (Figure 19) is used to realize even more highly precise feedback control. In a linear A/F sensor, the saturation current of the sensor changes linearly in response to the changes in the A/F ratio.

3.1.3 Fuel injection timing control

The fuel injection timing control switches the fuel injection start timing depending on the engine speed and the load. The fuel injection is controlled, so that it finishes before the intake air valve opens. The purpose of this timing control is to prevent the injected fuel from blowing into the exhaust manifold because of the overlap of the air

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Control in accordance with operating conditions.

Various corrections		Outlines
At start	Start injection time	Determines the basic injection time from amount of engine exhaust, water temperature, and engine speed.
After start	Increased amount for warm up correction	Increases amount when coolant temperature is low and gasoline evaporation is poor.
	Increased amount after start correction	The rate of increase is at maximum immediately after start but then gradually decreases to stabilize the engine speed.
	Transition time air–fuel ratio correction	Acceleration or deceleration is determined from the changes in the quantity of intake air and the amount is increased or decreased in accordance with the conditions to improve driving performance.
	Increased amount for high load correction	Amount is increased when the exhaust temperature exceeds the predetermined value.
	Increased amount for output correction	Corrections are made so that air–fuel ratio for maximum output is attained during high load driving.
	Air–fuel ratio feedback correction	Amount of fuel injection is increased or decreased based on the signal from the O_2 sensor. The air–fuel ratio is controlled within a narrow range near to the stoichiometric air–fuel ratio where the three-way catalyst has superior conversion performance.
	Voltage correction	Prevents fluctuations in the amount of fuel injected by the injector because of battery voltage fluctuations.
Other	Fuel cut	Fuel supply is cut off when decelerating, as torque is unnecessary, and also when throttle is fully closed.

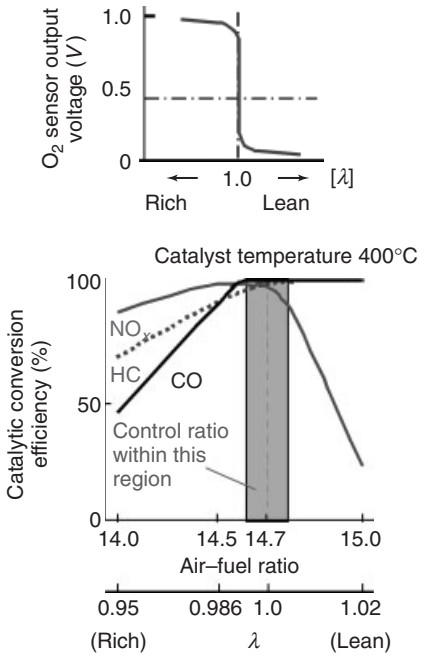


Figure 18. Output voltage of O2 sensor and catalytic conversion efficiency compared to air–fuel ratio.

intake and exhaust valves (both valves being open at the same time). This also increases the amount of time for fuel evaporation and prevents an increase in the amount of HC from unburned fuel that is produced when the fuel is ignited before it evaporates.

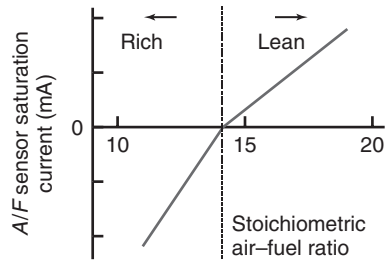


Figure 19. Linear A/F sensor output characteristics.

3.2 Outline of ignition control

3.2.1 Ignition control system

The ignition system calculates the ignition timing and power application time for the spark plugs attached to the cylinder head. It also outputs the ignition signal in accordance with the crank angle of the engine. Figure 20 shows the composition of this system. The crank angle of each cylinder is detected from the signals of the crank angle sensor and the cam angle sensor, which differentiates between the different engine cylinders. The ECU calculates the optimum ignition timing for each driving condition. It then outputs the ignition signal to the ignition device (ignition coil and igniter). The ignition device generates a high voltage based on these ignition signals and applies this voltage to the electrodes of the spark plugs. This causes

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

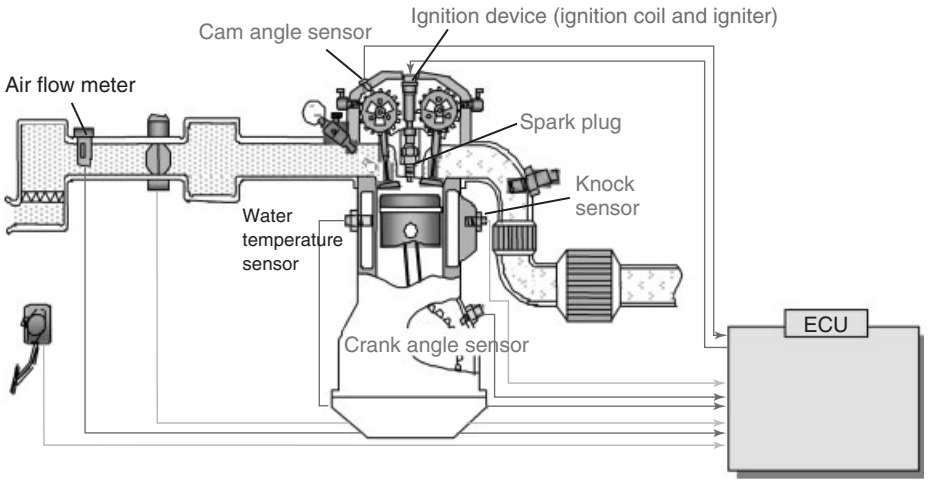


Figure 20. Ignition control system.

a spark to be formed that ignites and combusts the A/F mixture within the cylinder.

The maximum effective work due to the combustion energy is obtained when the maximum voltage is generated when the piston is slightly past-TDC. This ignition timing is called *maximum spark advance for best torque* (MBT) and the ignition timing is usually controlled, so that this condition is achieved. In the majority of engines, the piston position that attains maximum firing pressure and MBT is approximately 10°CA after top dead center (ATDC).

3.2.2 Ignition timing control

The ECU calculates the ignition signal, consisting of the ignition timing and power application time, based on the information from each sensor and then transmits this signal to the ignition device. The cylinders are differentiated by the output of the cam angle sensor, and the engine speed is calculated from the output of the crank angle sensor. Corrections are added based on the engine load and water temperature and then the ignition timing is determined. The power application time is determined by adding the correction for battery voltage to the engine speed (Figure 21).

3.2.3 Knock control

The ignition timing is advanced when the engine is in the high load region to get closer to the MBT. This increases the maximum firing pressure and causes the A/F mixture to ignite early, which produces abnormal combustion called *knocking*. The knock control system limits the degree

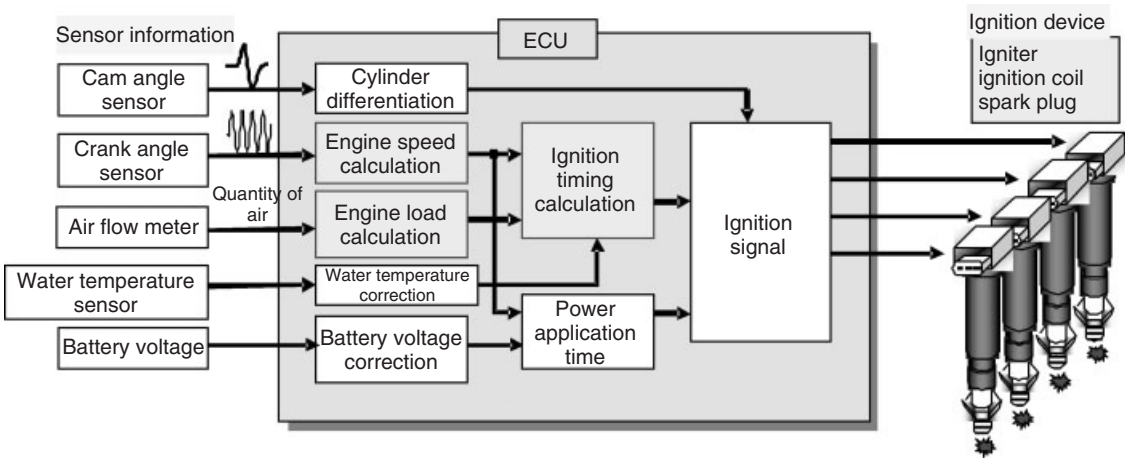


Figure 21. Content of ECU control.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of knocking to an allowable level that does not damage the engine. Allowing for some leeway and flexibility in the ignition timing is necessary in consideration of the mechanical tolerance and other factors. Ignition will occur well in advance of the MBT if knock control is not implemented and this will lead to a corresponding reduction in torque. The power and fuel economy of the engine can both be improved by incorporating knock control.

3.3 Outline of electronic throttle control

3.3.1 Electronic throttle system

Owing to ever-increasing demands for better fuel economy and lower exhaust emissions, improving the flexibility of control over the quantity of airflow is becoming more important. Each year, a growing number of vehicles are adopting an electronic throttle, which can control the throttle valve independently of the accelerator pedal operations of the driver. As this feature also is being adopted on the majority of engines currently in development, use of an electronic throttle system is becoming more and more common. Figure 22 shows a diagram of an electronic throttle system that is composed of a throttle body, accelerator pedal module, and an ECU.

The throttle body consists of a throttle valve, a motor that opens and closes that valve, and a sensor that detects the valve opening angle. A sensor that detects the depression angle of the accelerator pedal and the speed of depression is built into the accelerator pedal module.

3.3.2 Throttle opening angle control

The throttle opening angle control determines how far the throttle is opened in consideration of the acceleration, based on how far the accelerator pedal is depressed by the driver. The target opening angle is calculated from the accelerator pedal sensor. That is then compared to the actual opening angle of the throttle sensor to find the deviation in the opening angle. The necessary amount of control for the motor is also calculated (Figure 23). The motor that drives the throttle valve provides the drive by changing the duty ratio, so the amount of control is converted into the duty ratio and a current pulse is transmitted to the motor. In the case of a throttle valve that uses a DC motor, switches A and D are turned ON and the current flows in the direction of A when the valve is opened. Conversely, switches B and C are turned ON and the current flows in the direction of B when the valve is closed.

4 THE FUTURE OF ENGINE ECU

The engine control parameters include the quantity of intake air, the amount of fuel injection, the ignition timing, and the opening and closing timing of the intake and exhaust valves. All of these parameters are controlled by the engine

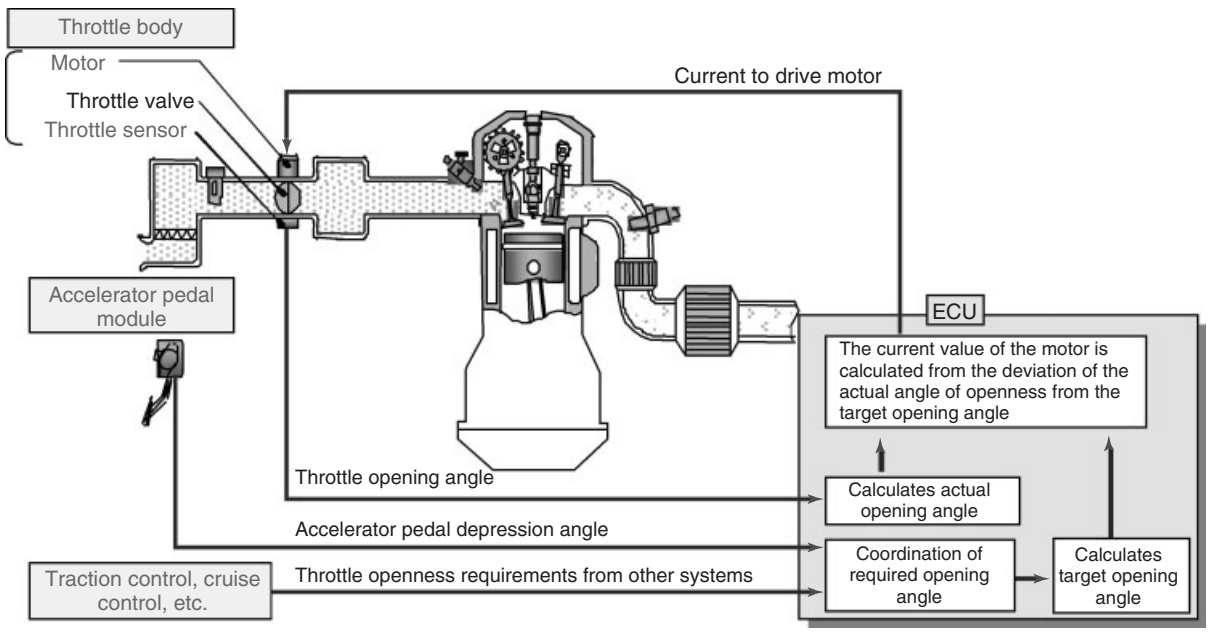


Figure 22. Configuration of electronic throttle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

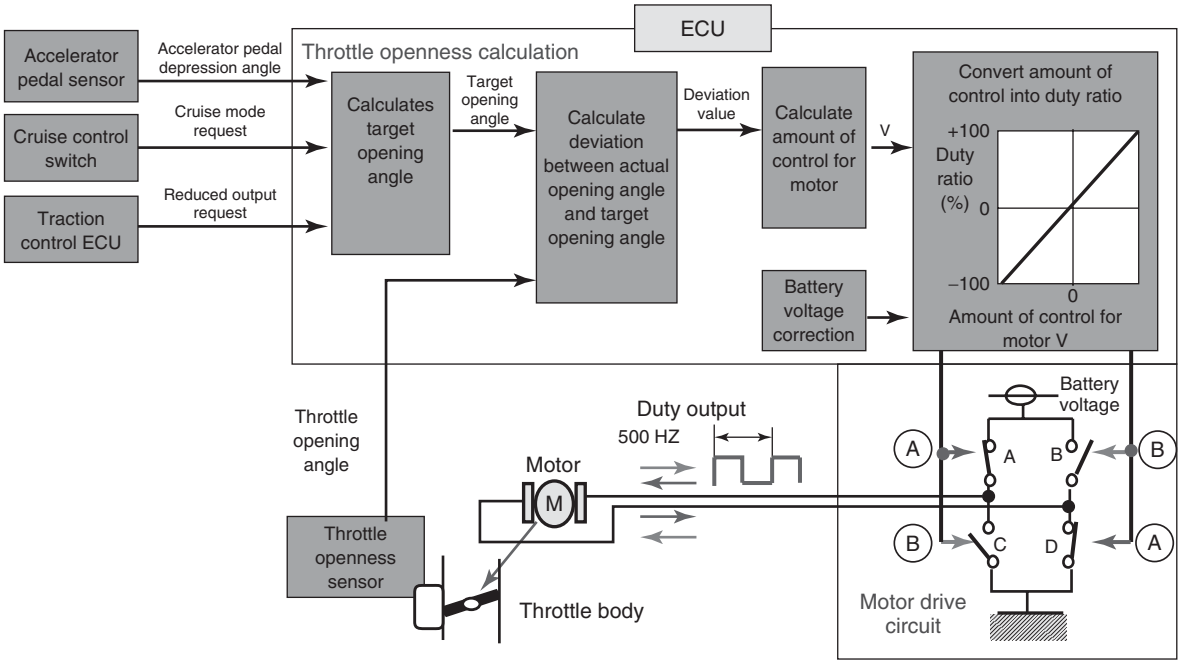


Figure 23. Throttle valve control.

ECU. This chapter described the hardware and software of the engine ECU system, as well as the fuel injection control, ignition control, and electronic throttle control as representative examples of engine controls.

In modern vehicles, the engine is controlled by the engine ECU system. This system has developed in conjunction with tightening of exhaust emissions and fuel economy regulations. In the future these trends will likely continue. The performance requirements of the engine ECU itself will also continue to increase from the standpoints of promoting greater safety, comfort, and protecting the environment. These are the three major requirements of vehicles in recent years.

The ECU system must be designed so that a fatal accident will not occur, even if the ECU system malfunctions. There has also been a great deal of attention and focus on ISO 26262, functional safety for road vehicles, in recent years.

The use of materials in the ECU system that include substances that are harmful to the environment, such as lead (Pb) and cadmium (Cd), are being abolished and demands to use materials that are renewable and recyclable are increasing to make vehicles more environmentally friendly.

There are also demands to reduce the size of the engine ECU system and install it somewhere other than in the occupant compartment in an effort to ensure more interior room and to improve passenger comfort. The recent trend is to install this ECU system closer to the engine, such as in the engine compartment or even directly on the engine. As a result, materials are needed that can withstand the more severe heat and vibration conditions of that installation environment, while also satisfying the demands for environmental friendliness.

RELATED ARTICLES

Exhaust Emissions
Engine Performance
Historical Overview of Electronics and Automobiles: Breakthroughs and Innovation by Electronics and Electrical Technology

REFERENCES

Kato, M. (2010a) *Automotive Electronics: Systems*, Nikkei Business Publications, Inc., Tokyo.

Kato, M. (2010b) *Automotive Electronics: Basic Technologies*, Nikkei Business Publications, Inc., Tokyo.

Otsuka, K. and Oi, M. (2006) Deployment of model based development for automotive software. *JSAE Journal*, **60**(6), 80–85.

Software Engineering Center, Information-Technology Promotion AgencyIPA of Japan (2008) *Embedded Software Process Reference*, Shoeisha, Tokyo.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto225
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Hybrid Systems and High Voltage Components

Masami Morikawa
DENSO Corporation, Kariya, Japan

1 History 1
2 Outline 2
3 EV System 2
4 HV Systems 2
5 Electric Motors 7
6 Inverter 9
7 Electric Power Supply System 14
8 High Voltage Auxiliary System 16
9 Future Trends 17
Further Reading 17

1 HISTORY

The history of hybrid vehicles (HVs) is part of the long history of electric vehicles (EVs). EVs first appeared in the Japanese market in extremely small numbers around 1950 as alternative energy vehicles. However, these EVs were only produced for about a year because the level of performance did not match that of internal combustion engine (ICE) vehicles. Development of EVs restarted some 20 years later around 1970, this time as environmentally friendly vehicles. The Agency of Industrial Science and Technology in the Ministry of International Trade and Industry (MITI) drew up a plan for a major EV project that was then undertaken with the cooperation of many vehicle manufacturers and components manufacturers. Unfortunately, this project also came to an end after several years due to the oil shocks and other circumstances in the 1970s. Another 20 years passed and once again EV development started up in the 1990s in response to the exhaust emissions regulations implemented in the state of California in the United States. The regulations were to take effect in 1998 and would have required every vehicle manufacturer to sell a predetermined percentage of EVs. Vehicle manufacturers worked feverishly to develop EVs to meet the new requirements. However, the regulations were not implemented as planned and once again the curtain fell on EVs after several years of limited production. In the end, their cruising range, recharging time, and cost made it difficult to seriously consider EVs as a viable substitute for conventional vehicles. They did find a niche in some limited markets (such as forklifts), however.

The vehicles that finally overcame these issues with EVs and became known for their fuel efficiency and environmental-friendliness were the HVs that went on sale in the later half of the 1990s. HVs were positioned as the vehicles of the twenty-first century and were ultimately equipped with much of the new technology developed for the EVs that were intended to comply with the Californian emissions regulations. Nickel-metal hydride (Ni-MH) batteries in particular achieved levels of performance that far surpassed that of conventional lead acid (Pb) batteries. It is not an overstatement to say that HVs would not have become a reality if not for these batteries. In addition to the batteries, the new hybrid drive systems that were proposed were also groundbreaking technology for HVs. The development and mass production of HV components made a major contribution to the improvement of EV performance and the reduction of overall costs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 OUTLINE

This chapter uses the hybrid systems adopted in the Toyota Prius and Honda Insight as typical examples. These systems utilize an electric motor to greatly reduce the energy consumed and exhaust emissions generated during start up, acceleration, and deceleration, which are the weak points of conventional ICE vehicles. These systems also resolve the cruising range and recharging time issues of EVs. Electric motor systems were originally adopted on EVs, and the HV system is a skillful integration of this EV technology with conventional gasoline engine technology. This chapter describes the types and special characteristics of HV systems based on an EV system and also explains the configuration and roles of the main HV components.

3 EV SYSTEM

3.1 EV system

Figure 1 shows the configuration of the EV system in a Toyota RAV4 EV. The EV engine control unit (ECU) controls the inverter and drives the motor in accordance with the signals from the accelerator pedal. A permanent magnet (PM) motor is used for the electric motor, whereas a Ni-MH battery is used for the main battery. The voltage is 288 V and the state of charge and discharge of the battery is monitored by the ECU, which can also calculate the remaining battery capacity as needed during driving. The main battery is recharged using a commercial power supply via the charger installed in the vehicle. The auxiliary battery is charged by converting the voltage from the main battery using a direct current (DC)–DC converter.

3.2 Driving control of EV

Figure 2 shows the driving control of an EV. The vehicle drive torque is determined by the angle of depression of the accelerator pedal, the brake signal, the shift position, and the vehicle's speed based on a drive torque map that was recorded in advance. Commands are then sent to the motor control portion to achieve this torque. For example, it is controlled as a motor during acceleration and as a generator during deceleration.

4 HV SYSTEMS

There are two ways of classifying HV systems, in accordance with their functions or their drive system. First, Figure 3 shows the different HV systems categorized according to function. A system that only has the idling stop function is referred to as an *ISS* or *micro HV*. The addition of other functions, such as acceleration assist, energy regeneration, and high efficiency engine operation, turns the system into a mild HV. The system is called a *strong HV system* if it includes an EV drive mode. CO_2 and exhaust emissions both decrease the further that the system progresses toward a strong HV. Once the vehicle becomes

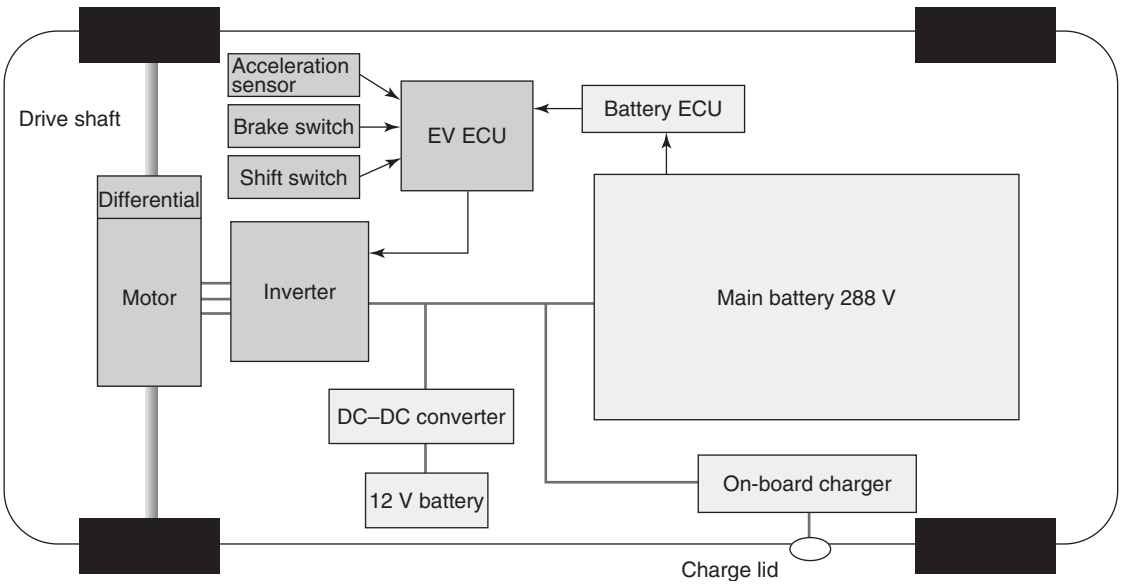


Figure 1. Configuration of RAV4 EV system.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

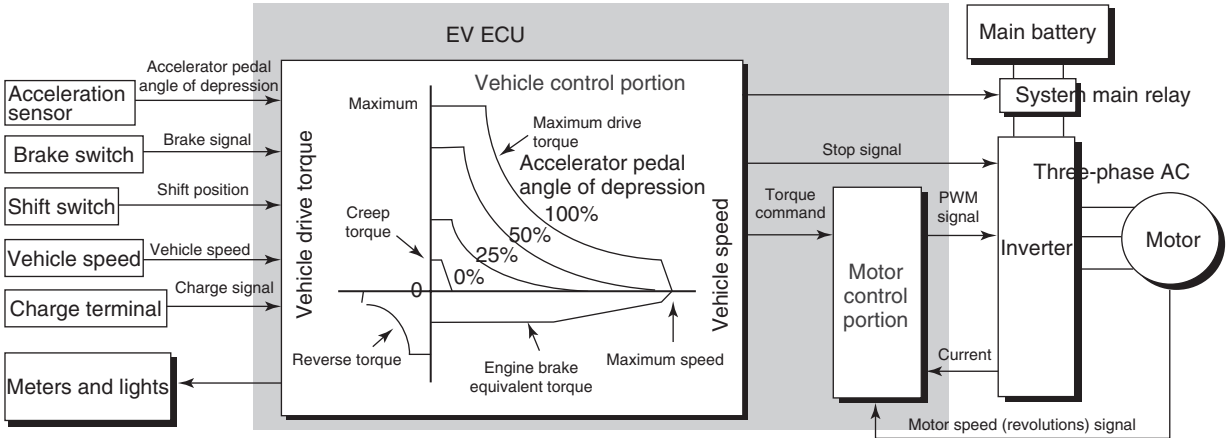


Figure 2. Driving control of EV.

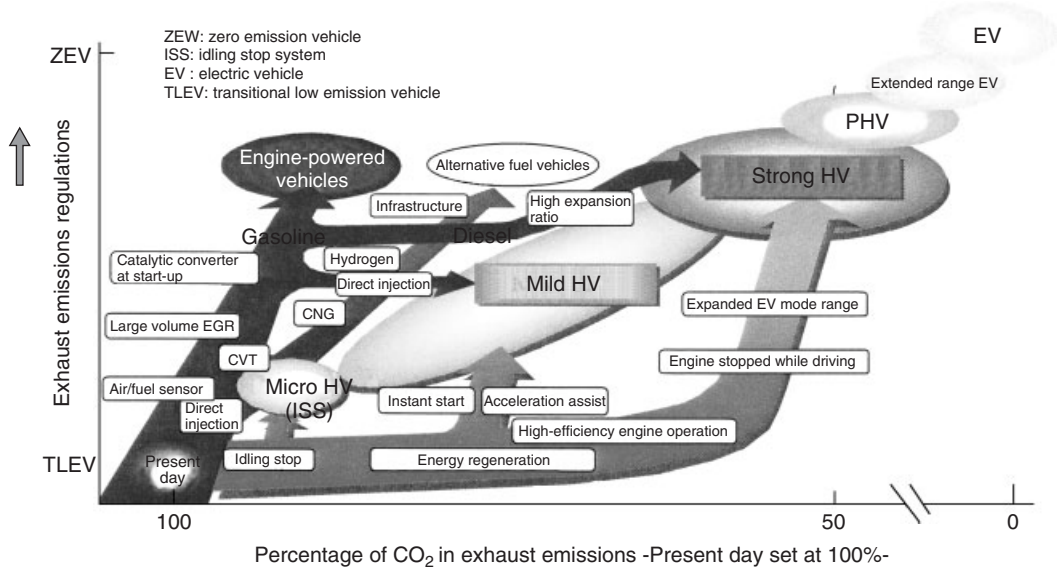


Figure 3. Reduction in CO_2 emissions by HVs and EVs.

an EV, there are zero exhaust emissions. Plug-in hybrid vehicles (PHVs) and extended range EVs are positioned somewhere in between a strong HV and an EV.

HV systems can also be classified according to their drive system. There are three main HV drive systems: series, parallel, and series parallel. The configuration and special characteristics of each of these drive system are explained in the following sections.

4.1 Series HV

Figure 4 shows the configuration of a series HV system. An engine is installed in the vehicle to drive the generator that charges the main battery. The vehicle is propelled by an electric motor and the main battery is constantly being charged.

4.1.1 Special characteristics

– The vehicle is propelled only by the electric motor so the power and size of the motor and generator are larger than in the other drive systems (based on an EV with an engine and generator added).
– All of the motive power from the engine is converted into electrical power, so the conversion efficiency (transmission efficiency) is slightly poor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

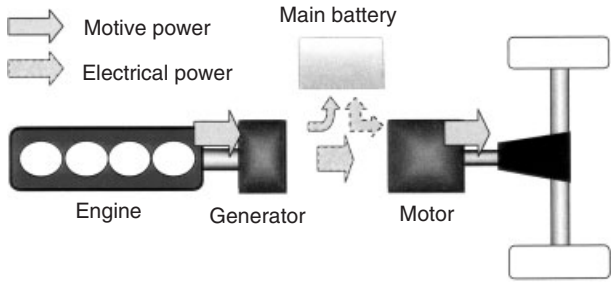


Figure 4. Configuration of series HV system.

– It is easy to control the drive power and output.
– It is comparatively easy to control the exhaust emissions because the engine speed is steady.

4.2 Parallel HV (Honda insight)

Figure 5 shows the configuration of a parallel HV system. The engine and electric motor are arranged parallel to each other and the motive power to propel the vehicle can be supplied from both of these sources. However, in the case of this example, the engine and motor form a directly connected structure, which must move simultaneously.

4.2.1 Special characteristics

– The configuration of the system is simple as an electric motor is added between the engine and the transmission of a conventional vehicle.
– The engine is used a large portion of the time, so the electric motor plays a mostly auxiliary role. The motor system itself is small as there is no EV mode and there is comparatively little change from a conventional vehicle.
– Electrical power can only be used to propel the vehicle after the battery is charged.

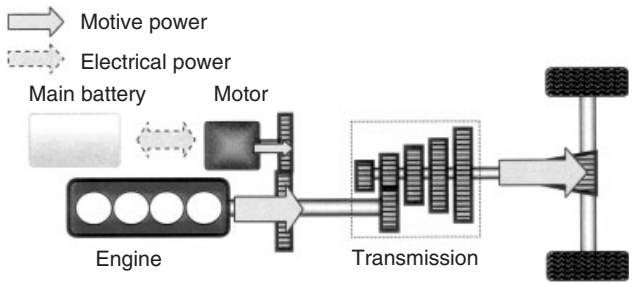


Figure 5. Configuration of parallel HV system (Honda insight).

– EV mode becomes possible if the engine and motor are separated and designed to provide drive independently. In this case, the motor system would become larger.

4.3 Series parallel HV (Toyota Prius)

Figure 6 shows the configuration of a series parallel HV system. Three power sources, the engine, MG1, and MG2 are connected by a planetary gear train. These provide drive to the vehicle in combination with each other in accordance with the driving conditions. Here, MG is an abbreviation for motor generator. Usually, these devices are referred to separately, but each is frequently used for their electrical operation function and electrical generation function, respectively, so they are referred to collectively as MG to reflect their actual usage. The role of the engine is to provide drive to the vehicle and MG1. The role of MG1 is to charge the main battery and also to provide assistance during engine start and auxiliary drive to the vehicle as an electric motor. The role of MG2 is to provide drive to the vehicle (EV mode) and any necessary assistance, as well as energy regeneration as a generator.

4.3.1 Special characteristics

– This system achieves both good fuel economy and driving performance, because it possesses the advantages of both the series HV system and parallel HV system.
– There is great potential to improve fuel economy, because the system is very efficient.
– The system and its various controls are complicated.

4.3.1.1 Configuration of HV system. Figure 7 shows how a HV is equipped with both an engine system and

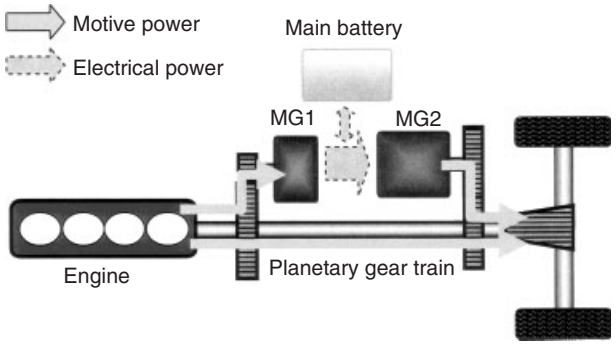


Figure 6. Configuration of series parallel HV system (Toyota Prius).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

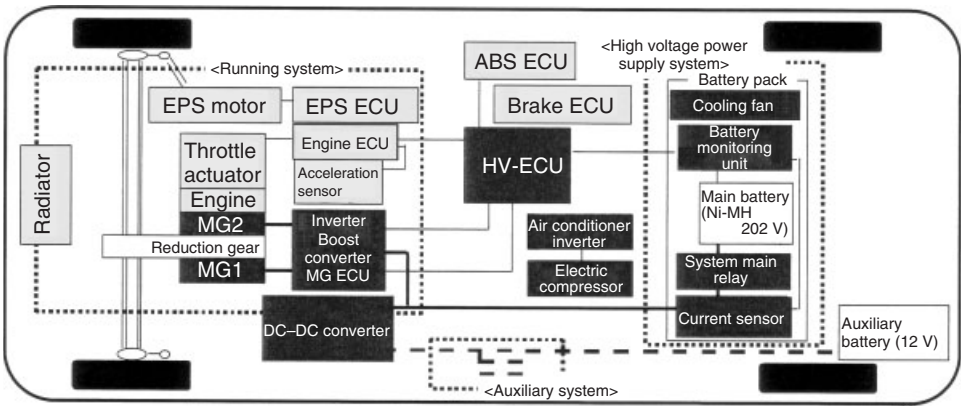


Figure 7. Configuration of HV system (Toyota Prius).

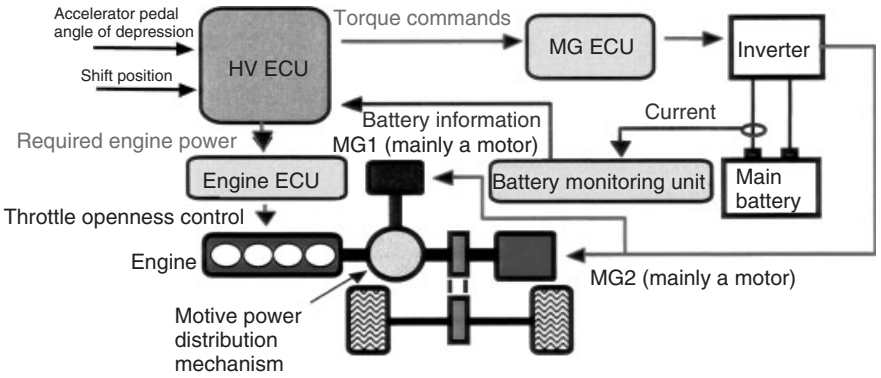


Figure 8. Configuration of HV control system.

two MG (motor) systems. The HV ECU is provided for integrated control to ensure optimum operation of the two drive systems. A Ni-MH battery is used for the main battery, and the components that compose the electric power supply system, such as the battery ECU and current sensors, are the same as those used on an EV. A DC–DC converter is used to charge the auxiliary battery and an electric compressor is used for the air conditioner, exactly the same as on an EV. One difference from an EV is that there is no battery charger.

4.3.1.2 System control. Figure 8 shows the configuration of the control system in a HV.

– The HV ECU issues commands to the engine ECU to produce the required engine power based on the requests from the driver, the state of the vehicle, and the state of charge (SOC) of the main battery. It also calculates the MG1 torque and MG2 torque and then issues the torque commands to the MG ECU. After this, the engine ECU controls the degree of openness of the electronically controlled throttle in accordance with the engine power requirements from the HV ECU.
– The MG ECU controls MG1 and MG2 via the inverter in accordance with the torque commands from the HV ECU. It also controls the drive power, so that operation is in line with these torque commands.
– The battery monitoring unit monitors the SOC of the main battery and other items. It then transmits the SOC and other battery information to the HV ECU.
– The HV ECU controls the SOC of the main battery so that the system is in the optimum state.

4.3.1.3 Regenerative braking control. Figure 9 shows the brake configuration. The motor generates power during deceleration by converting kinetic energy into electricity. This electricity is then stored in the battery. The distribution of braking force between the mechanical brake and the regenerative brake is controlled so that even the slightest increase in kinetic energy can be recovered.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

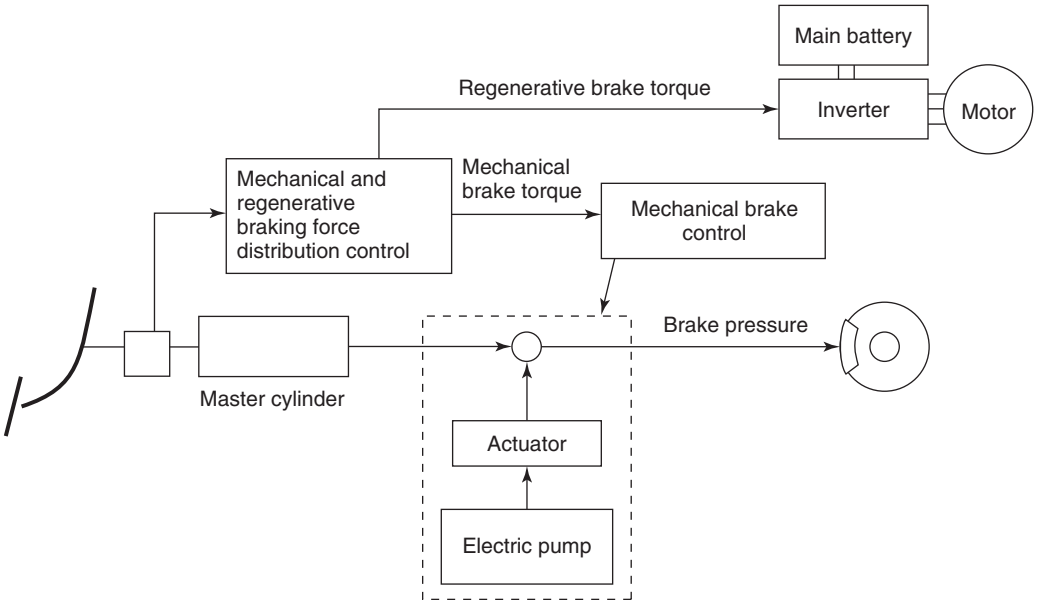


Figure 9. Configuration of HV brakes (example).

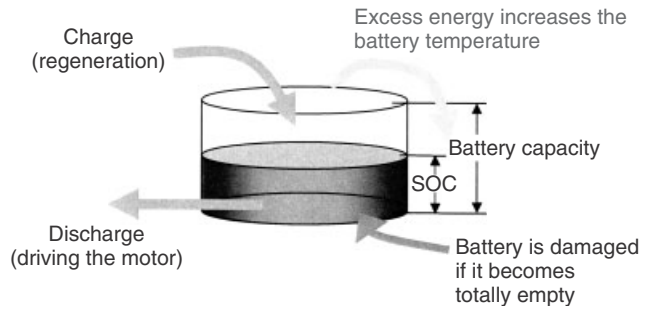


Figure 10. Concept of electrical power control.

4.3.1.4 Electrical power control: management of main battery capacity. Figure 10 shows the concept behind electrical power control. In a HV, the main battery is discharged when the MG is used as a motor, such as during EV mode. The main battery is charged when the MG is used as a generator, such as during regenerative braking. A longer EV mode is possible when the remaining main battery SOC is high, but less energy regeneration occurs. The opposite is true when the remaining SOC of the main battery is low.

Figure 11 shows how an actual battery is used. The main battery is repeatedly charged and discharged in accordance with the driving conditions of the vehicle. Properly maintaining the remaining SOC in the main battery is necessary to optimize its use. Therefore, regeneration and discharge are monitored at all times to constantly calculate the SOC.

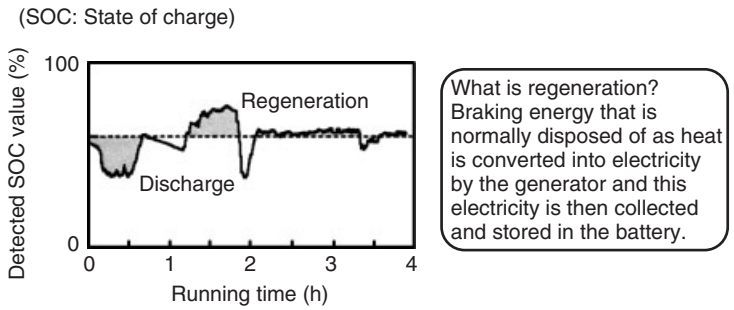


Figure 11. SOC during actual vehicle driving.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

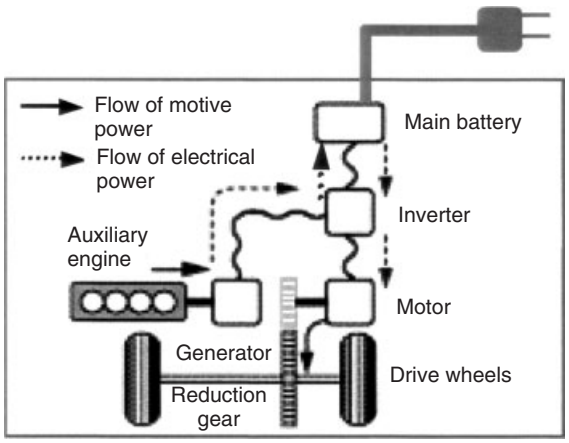


Figure 12. Configuration of extended range EV.

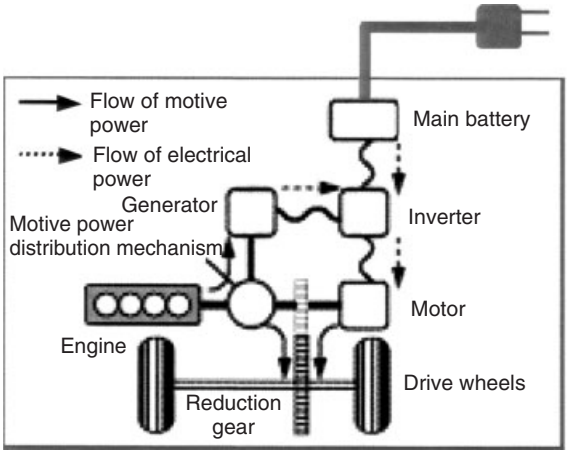


Figure 13. Configuration of PHV.

4.4 EV and HV application systems

4.4.1 Extended range EV

With this type of vehicle, an engine and generator are added to an EV to extend the cruising range. Figure 12 shows the configuration of this system and the flow of energy.

- EV mode is the basis of this system.
- The cruising range can be extended in an emergency as a series HV.
- The vehicle is equipped with the smallest engine possible.

4.4.2 Plug-in hybrid vehicle

The EV mode distance is extended by keeping the main battery of a strong HV at full charge. Figure 13 shows the configuration of this system and the flow of energy.

- A strong HV is the basis of this system and a charger is added.
- The capacity of the main battery is augmented.

5 ELECTRIC MOTORS

5.1 Types and uses of electric motors

Figure 14 shows the different types of electric motors and their applications. DC motors (wound DC motors), induction motors (squirrel-cage type), and permanent magnet motors are the types of electric motors that are most often adopted as traction motors to propel the vehicle.

5.2 Summary of HV motor requirements

The following items are the required characteristics and capabilities of a HV motor. The motor must have high power, high revolutions, and high voltage to reduce its size. The motor must be highly efficient to improve the fuel economy of the vehicle. The motor must also be reliable and largely maintenance free to be used in the system that provides drive for the vehicle. Figure 15 compares the main types of electric motors. Currently, alternating current (AC) motors have the advantage in terms of their ability to handle high voltages and to require little maintenance. In particular, the PM motor is often used due to its small size and high efficiency. Rare-earth magnets are often used in these types of motors to realize significant reductions in size and increases in power.

5.3 PM motor

PM motors can be divided into two different types depending on where the magnet is located: surface permanent magnet (SPM) and internal permanent magnet (IPM). Figure 16 shows the structure of the rotor in both these types of motors.

Currently, IPM motors are often used in HVs, because it is easy to increase their speed and to utilize reluctance torque (Figure 17). SPM motors are used in systems (such as the power steering) where vibration noise is expected to be a problem. However, there are cost and supply issues that arise when using large amounts of rare-earth magnets. Figure 18 shows a photograph of an actual motor for a HV.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Types of motors			Uses
DC motor	Wound-DC-motor		Traction motor for EVs Starter
	PM-DC-motor		Small-sized motor for vehicles (Windshield wipers, power windows, etc.)
AC motor	Induction	Squirrel-cage	General motive power motor Traction motor for EVs and HVs
		Wound-rotor	Asynchronous generator for wind power
	Synchronous	**PM**-rotor	Traction motor for EVs and HVs
		Wound-rotor	Generator for hydroelectric and thermoelectric power plants
		reluctance	

Figure 14. Types and applications of electric motors.

Requirement items	Explanation	DC motor	AC motor	
		Wound DC motor	Permanent magnet (rare-earth)	Induction
Small size, light weight	Fits in limited space, total weight is reduced	△	◎	○
High power	Acceleration from a stop and for passing - High torque at low speeds - Fixed power at a wide range of speeds	△	◎	○
High revolutions	10,000 rpm or more	△	○	○
High efficiency	Fuel economy	△ 80% to 87% efficiency	◎ 90% to 92% efficiency	△ 79% to 85% efficiency
High voltage	100 V or more	×	◎	◎
Long service life	Maintenance free - Brush never needs replacing	×	◎	◎
Low cost	Total including the control devices	×	△	

◎: excellent
○: good
△: fair
×: poor

Figure 15. Comparison of HV motor requirements.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

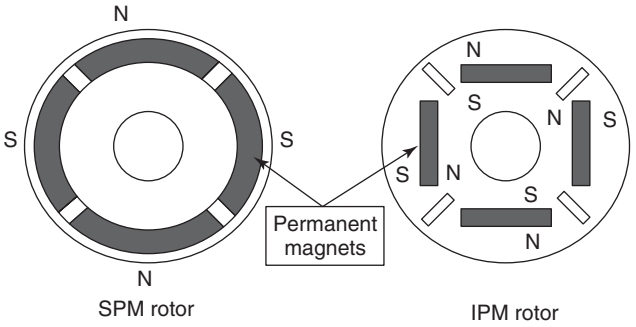


Figure 16. Cross sections of SPM and IPM rotor structures.

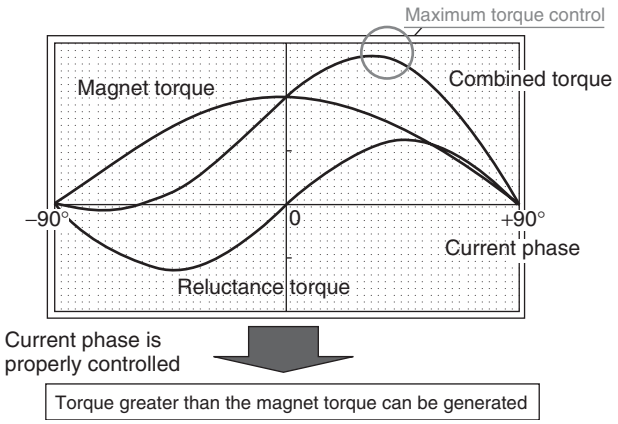


Figure 17. Utilization of reluctance torque.

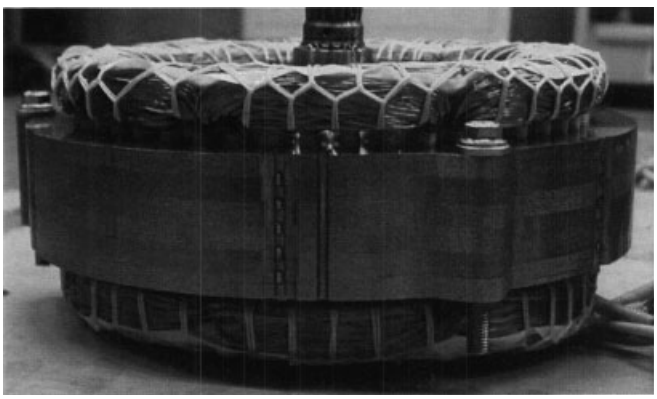


Figure 18. Electric motor for HV.

6 INVERTER

6.1 Overall configuration and functions

Figure 19 shows the configuration of an inverter. The inverter is a device that takes the DC from the main battery installed in the vehicle and converts it into three-phase AC via six bridge-linked power devices to supply it to the motor. The application of pulse width modulation (PWM) control to the power devices enables output of voltage-controlled sine wave AC and controls the torque of the motor. (PWM control is explained below.)

The necessary torque is calculated by the HV ECU from the acceleration commands indicated by the amount of accelerator pedal operation by the driver. This is then sent to the inverter as drive signals of the power devices. The torque that can be obtained changes depending on the relationship between the voltage phase and the rotor position. Therefore, the electric power is supplied at the time that will obtain the maximum torque by detecting the position of the rotor.

6.2 Principle of three-phase AC generation

This section explains the principle behind the method of creating the voltage-controlled sine wave three-phase AC (Figure 20). The sine wave voltage commands that are phase shifted by 120° are compared to triangular wave and ON and OFF signals created for the two power devices in the upper arm and lower arm in each of the three phases, U, V, and W. This allows pulse voltages to be obtained for each of the phase voltages (V_U, V_V, and V_W) in the form of sine waves that are phase shifted 120° (the average voltage of each pulse is changed into the form of a sine wave). The pulses can be changed ON and OFF by changing the amplitude of the voltage command and this in turn changes the voltage value. The current will only flow through the motor when ON in this state, but the current will also flow through the motor during the OFF period due to the flywheel diode (FWD) that is connected in parallel. This allows for uninterrupted sine wave current to be applied to the motor. The higher the frequency of the triangular wave, the more that the motor magnetic noise and current ripples can be suppressed. However, this will increase the loss of the power devices, so usually the frequency of the triangular wave is set between 5 and 10 KHz.

6.3 Power devices

The power devices are the most important components of the inverter from the standpoints of functionality and cost. One of the key points for lowering the cost is reducing the size of the elements that make up the inverter as much as possible. Reducing the loss that is generated by the elements is vital to reducing their size. Figure 21 shows that in an ideal switch, no loss is generated no matter how much current flows through it, but that the switch ON voltage in a semiconductor that is generated when current flows through it becomes ON loss.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

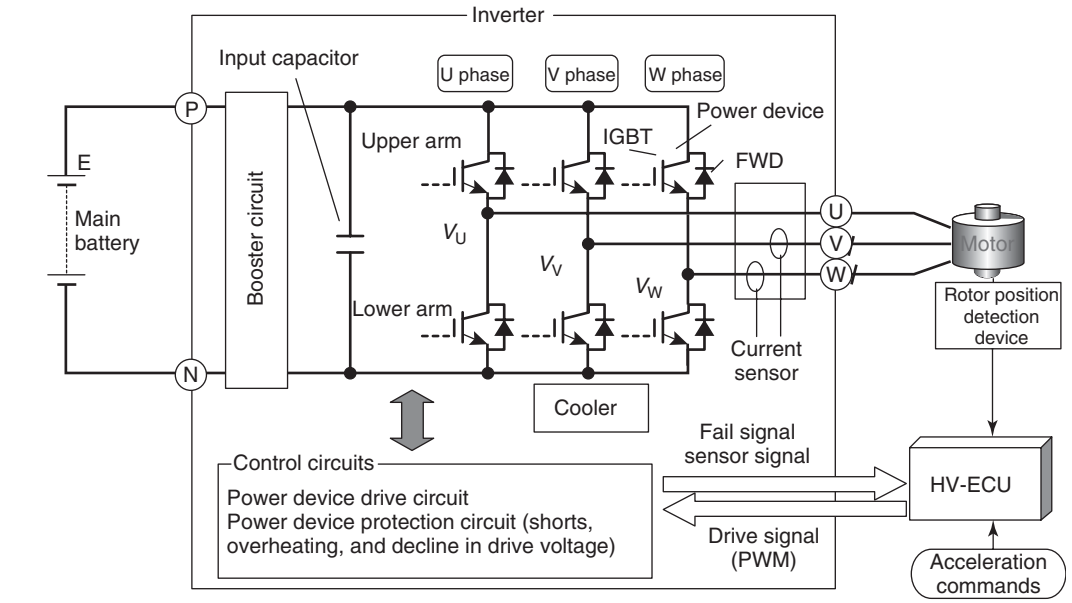


Figure 19. Configuration of inverter.

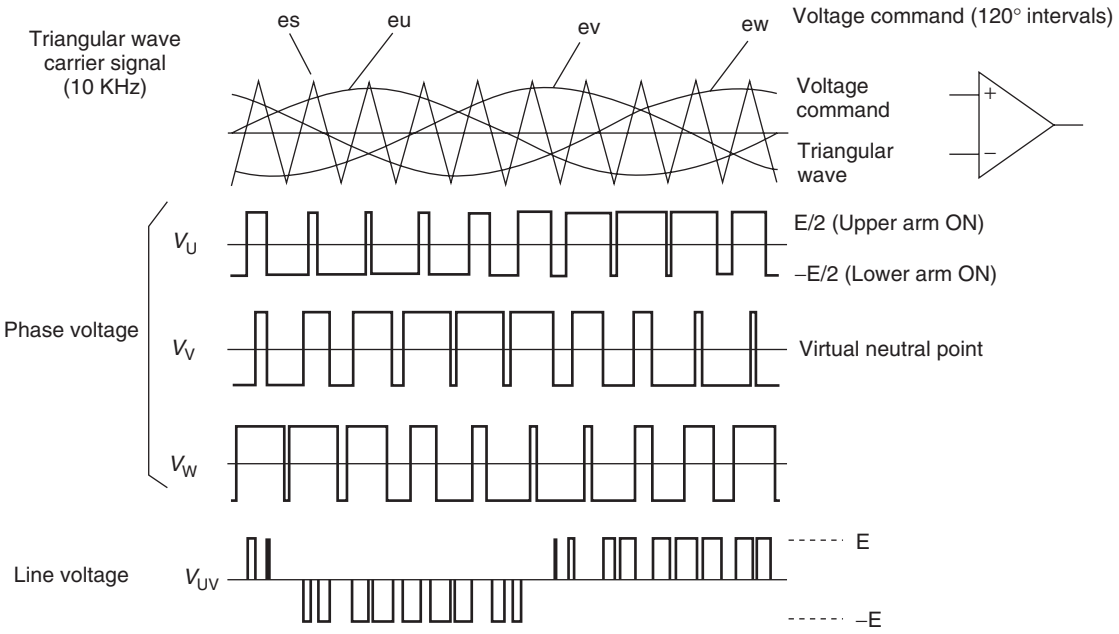


Figure 20. Principle of three-phase AC generation.

In addition, although a switch is turned ON and OFF, it does not switch to the other state instantly. This extra time (switching time) results in switching loss. There are three methods of reducing loss, as shown in Figure 21: (i) decrease the current, (ii) decrease the ON voltage, or (iii) shorten the switching time. The following sections provide explanations of these methods.

6.3.1 Decreasing current

Figure 22 shows the configuration of a booster circuit. A booster circuit is provided between the inverter and the main battery. It boosts the voltage to 650 V and supplies it to the inverter. The current flowing through the power devices can be reduced because the current in the motor

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

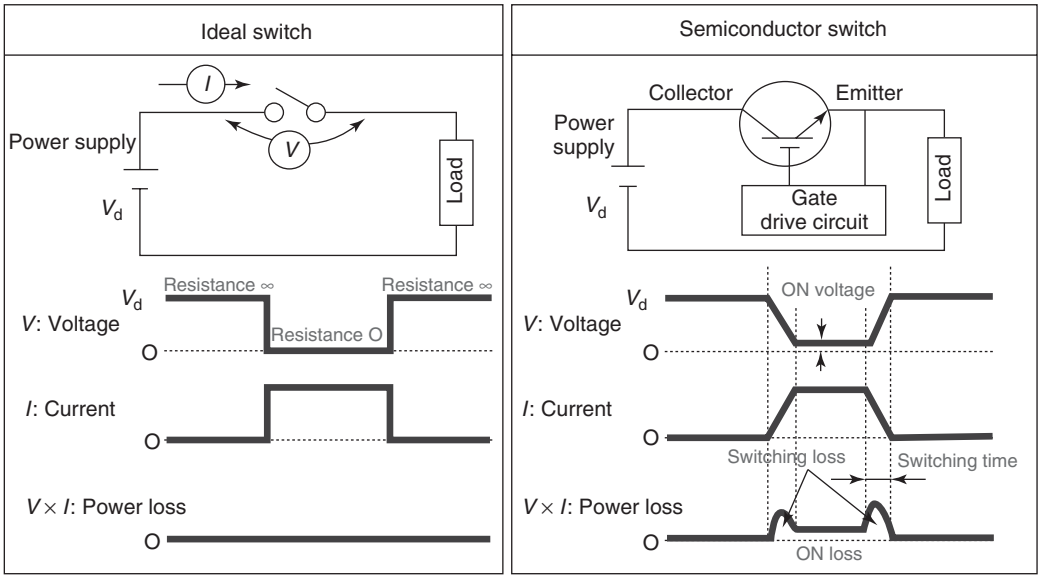


Figure 21. Ideal switch and semiconductor switch.

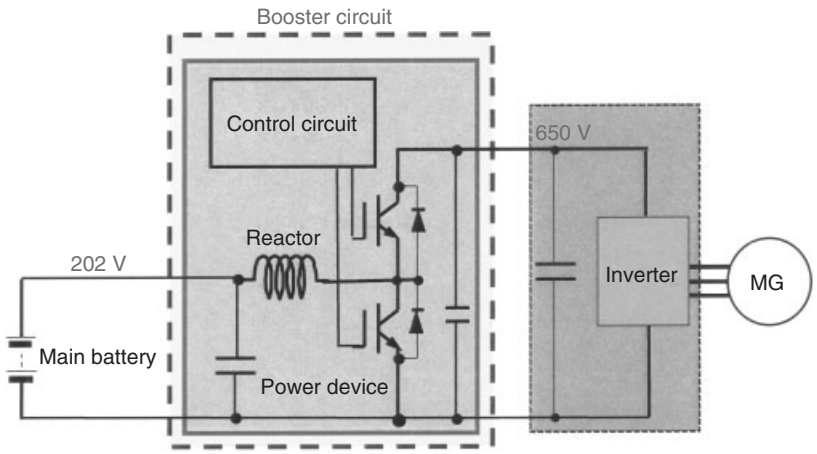


Figure 22. Configuration of booster circuit.

will decrease in proportion to the voltage. The more that the boost voltage is increased the more that the current will decrease. However, the use of very high voltage will cause devices such as the inverter and motor to increase in size due to insulation problems. Therefore, this level of voltage (650 V) is optimum for onboard devices.

The reduction in cost attained by reducing the current is sufficiently large to offset the cost of including the booster circuit, so this method of decreasing the current is often adopted. The booster circuit operates as follows. The power device on the lower side is turned ON and energy builds up in the reactor. The voltage then rises when this power device is turned OFF. This voltage is then stored in the capacitor after passing through the diode on the upper side, thereby boosting the voltage. When regeneration occurs, the power device on the upper side and the diode on the lower side are operated causing current to flow to the main battery.

6.3.2 Decreasing ON voltage

The most suitable power devices are selected because their characteristics determine the ON voltage. As a result, the power devices that can be used in an onboard inverter are insulated-gate bipolar transistors (IGBTs) or power metal oxide semiconductor field effect transistors (MOSFETs). Figure 23 shows that an IGBT has a low ON voltage

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

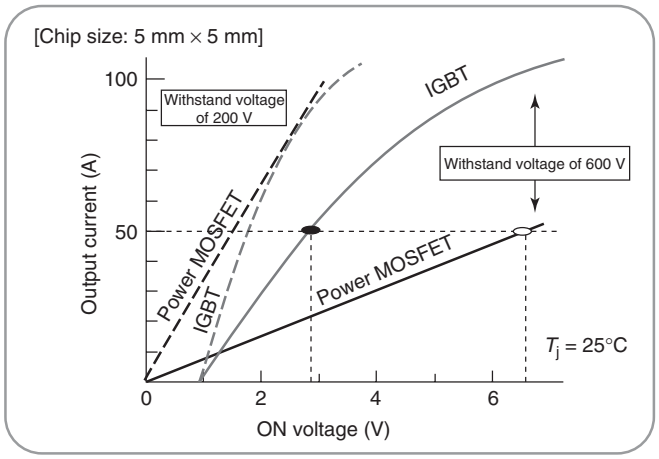


Figure 23. Comparison of IGBT and power MOSFET characteristics.

compared to other power devices with a withstand voltage of 200 V or more. Since the actual operating voltage is boosted to 650 V, the withstand voltage of the power devices must be around 1000 V, which makes IGBTs even more advantageous. IGBTs have been adopted due to the reasons explained above. However, it goes without saying that the higher the withstand voltage rises, the higher the ON voltage will also become, so the power devices and inverter must be formed from low withstand voltage components as much as possible.

6.3.3 Shortening switching time

It is easy to understand that shortening the switching time will result in a decrease in the switching loss. This can be accomplished by changing the gate resistance, as the time will decrease the smaller that the resistance becomes. However, increasing the withstand voltage of the power devices is necessary because the change in the current (d*i*/d*t*) increases and the surge voltage, ΔV, also increases. In this case, the ON loss will increase, although the switching loss has been reduced. Figure 24 shows that keeping the wiring inductance as small as possible is a key point of the design because the surge voltage is generated by wiring inductance.

In concrete terms, the capacitor should be placed as close as possible to the IGBT and the wiring should be kept short, but extensive. The + and − poles should be placed between thin insulating material and arranged as close to each other as possible as the mutual inductance effect will reduce *L*. Methods such as these can be employed to reduce the surge voltage.

There is also one other method of reducing the size of the power devices, besides reducing the loss. This is to lower the temperature of the power devices by improving the heat dissipation performance. Figure 25 shows the mounting structure and heat dissipation channels of a power device. The most common cooling method is to use water cooling and there are also cases where air cooling is used for power devices with comparatively small capacities.

The rise in the temperature of the power device, ΔT, is the product of the element loss, *P*, and the heat resistance, *R* (inverse of the heat dissipation performance). The loss can be permitted at the rate that *R* is reduced, so a small-sized element can be used.

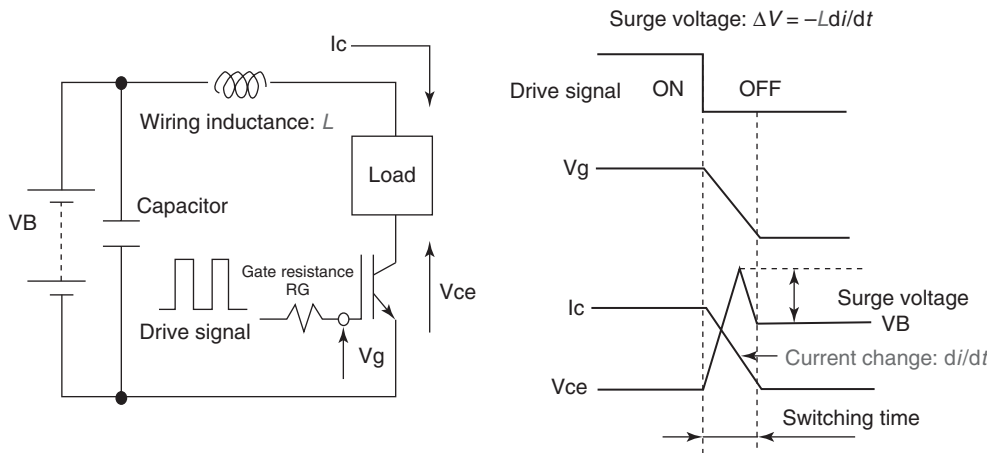


Figure 24. Principle of surge voltage generation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

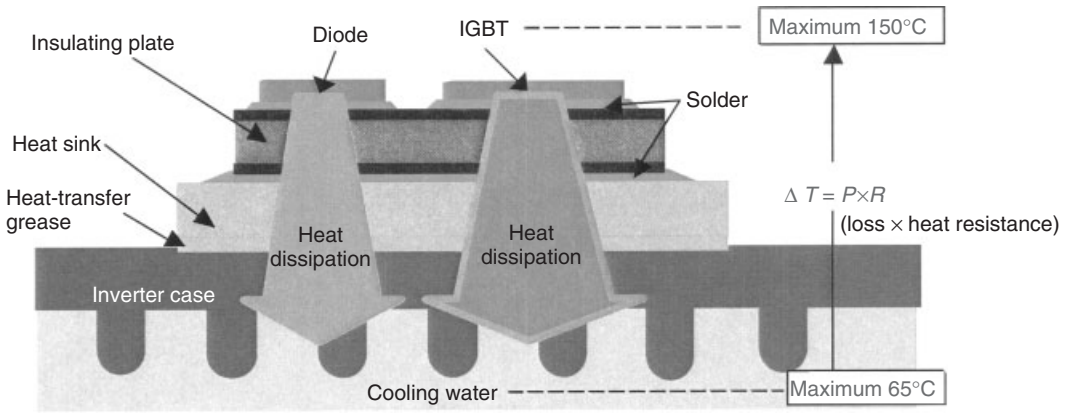


Figure 25. Power device mounting structure and heat dissipation channels.

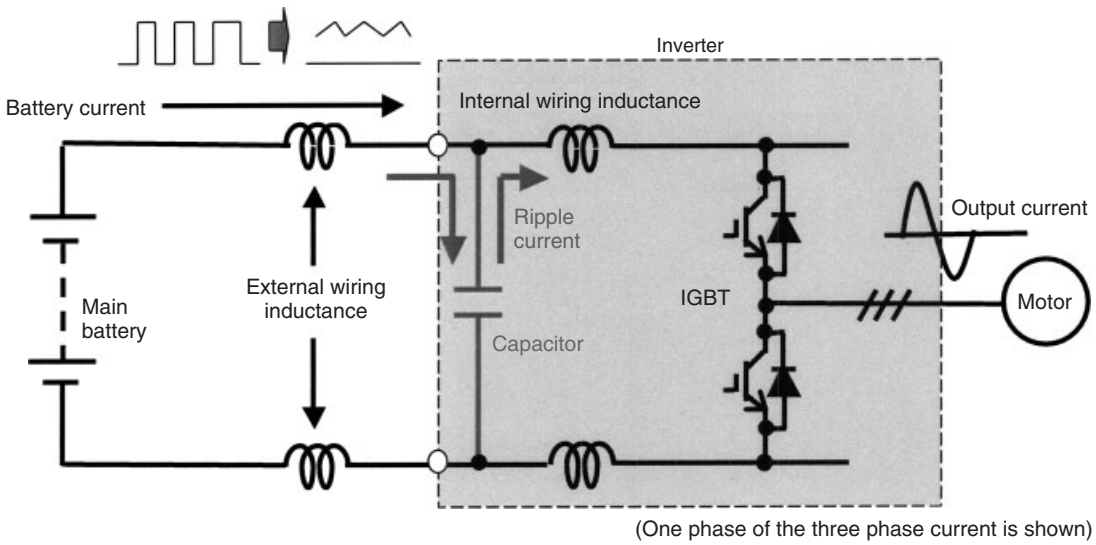


Figure 26. Capacitor role.

6.4 Capacitors

Figure 26 shows that the role of the capacitor is to smooth out the current from the main battery. As explained previously, if there is no capacitor, the current will change into the form of a pulse because of the ON and OFF operations of the IGBT. In this case, an excessively large surge voltage will be produced when the current is OFF due to the external wiring inductance and the IGBT will be destroyed instantly.

This surge voltage is suppressed through the use of a capacitor to gently decrease the current. Another issue is that when the current flows in the form of a pulse, it radiates noise with a wide frequency component from the wiring in between the main battery and the inverter. This noise can cause interference with the vehicle's radio and other devices. This noise is also suppressed by smoothing the current with a capacitor.

6.5 Control circuit

Figure 27 shows the configuration of the control circuit. The control circuit is composed of the drive circuit that drives the IGBT and the protection circuit that protects the IGBT. The low voltage system and the high voltage system are electrically insulated from each other by photo couplers and a transformer.

The drive signals are produced by vector control of the MG ECU. The protection circuit detects any overcurrents, short circuits, overheating, or drive voltage decrease in the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

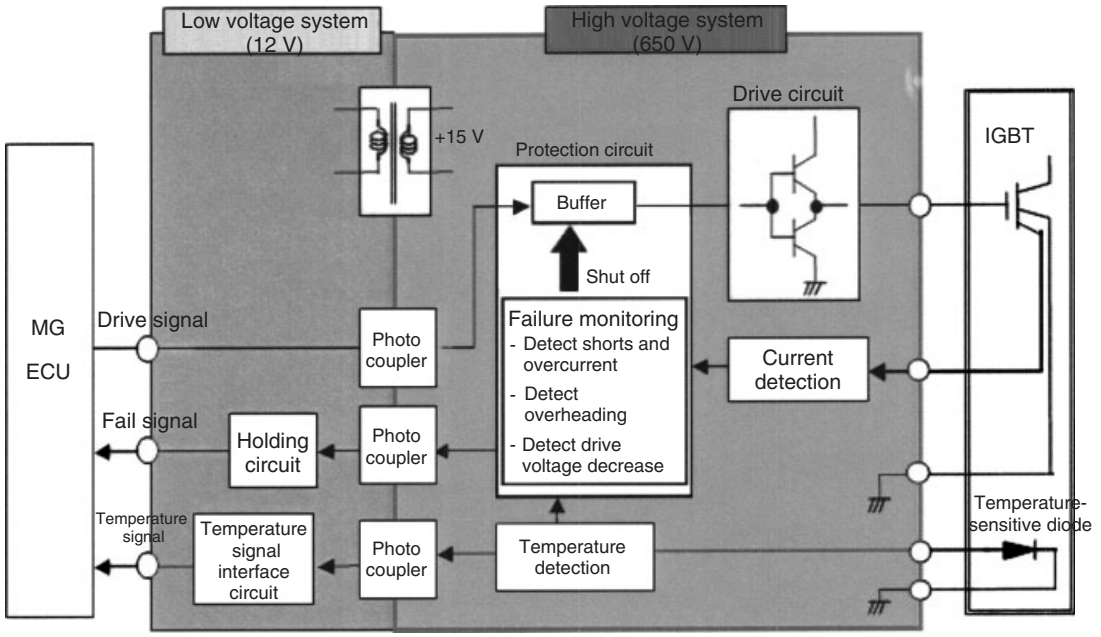


Figure 27. Configuration of control circuit.

IGBT and will shut off the IGBT when it detects any of these problems.

7 ELECTRIC POWER SUPPLY SYSTEM

7.1 Configuration

Figure 28 shows the configuration of the electric power supply in a HV. The electric power supply of a HV is composed of the following items: the main battery and the battery cooler, the battery ECU (monitoring unit) that monitors the battery, the system main relay (SMR) that mechanically supplies or shuts off the high voltage, the precharge relay and resistor for preventing inrush current, and the current sensor that senses the current being input into and output from the battery, and the like. All of these items are contained within a single case, which is also called a *battery pack*. A DC–DC converter is also used in a HV as a substitute for a conventional alternator to charge the auxiliary battery.

7.2 Main battery

There are four major points that define the main battery installed in a HV. The first point is the cycle life. The main battery in a HV is frequently and repeatedly charged and discharged. The cycle life of Pb batteries is short, so Ni-MH and lithium-ion (Li-ion) batteries that have longer cycle lives have become the mainstream batteries for use in HVs.

The second point is the energy density. A compact battery that can deliver large amounts of energy is necessary from the standpoints of fuel economy and cruising range. Li-ion batteries are superior from the standpoint of energy density (Figure 29).

The third point is power density. A battery that can produce large power is necessary to obtain faster acceleration performance. Li-ion batteries also have superior power performance, but efforts are being made to significantly improve the performance of Ni-MH batteries as well.

The fourth point is safety. In the event of an accident or unintentional overcharge or over-discharge of the battery, the battery must not catch fire. Although Li-ion batteries have superior performance, their use of combustible metallic lithium is a fire risk. Consequently, the incorporation of a protection system via the battery package and controls on the vehicle side are necessary. This makes the system more complicated, which is a weak point of Li-ion batteries.

7.3 Concept behind insulation of high voltage system

The main batteries in HVs use voltages that can exceed 200 V. At this level of voltage, a person who comes into contact with it would be electrocuted to death. Consequently, thoroughly developed insulation concepts and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

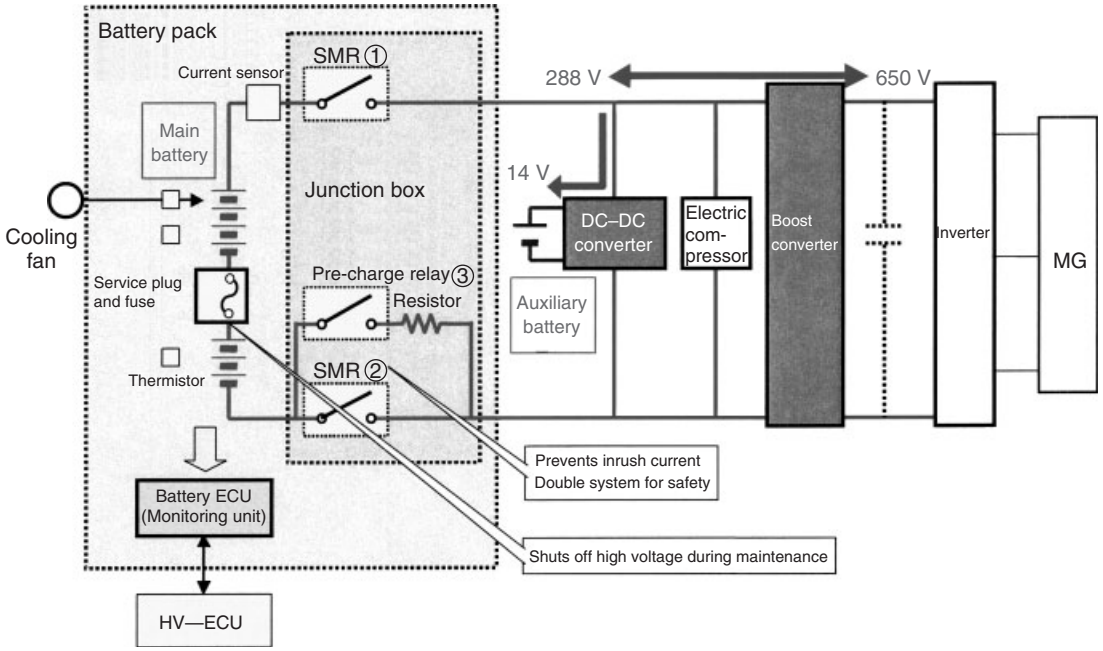


Figure 28. Configuration of HV electric power supply.

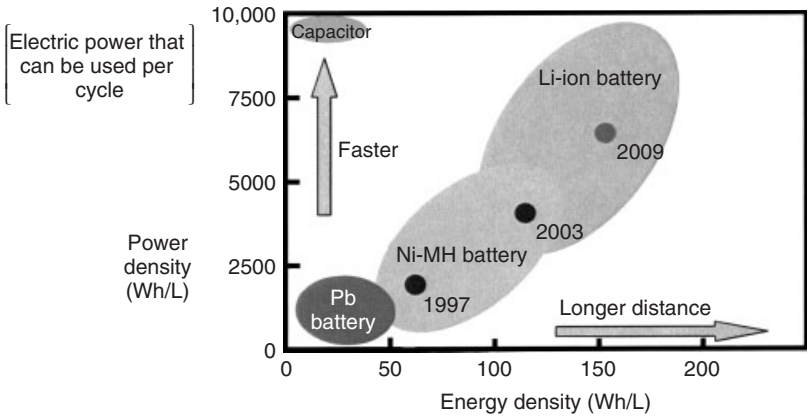


Figure 29. Energy and power densities of various batteries.

measures are a necessity. The voltages in a conventional vehicle only range from 12 to 24 V, so even if a person was shocked, it would not be a major problem. The body of the vehicle is used as the earth in this case to simplify the wiring. However, if this same way of thinking was applied to a high voltage system, then a person may be electrocuted if they accidentally touched a high voltage component. There is also the risk that an insulation failure of a high voltage device may lead to a fire. Therefore, the high voltage circuits in a HV do not use the vehicle body as the earth. Instead, the minus line is directly wired, the same as the plus line, and utilizes a floating earth. The use of a floating earth means that even if a person directly touched a high voltage component and the vehicle body, they would not receive a shock. In a further effort to improve the safety of HV, there are also some vehicles that are equipped with a warning system that will detect an insulation failure of a high voltage device or wiring and then warn the driver.

7.4 DC–DC converter

In a conventional ICE vehicle, the alternator charges the auxiliary battery. In a HV, however, the engine is stopped frequently, so using an alternator for charging would result

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

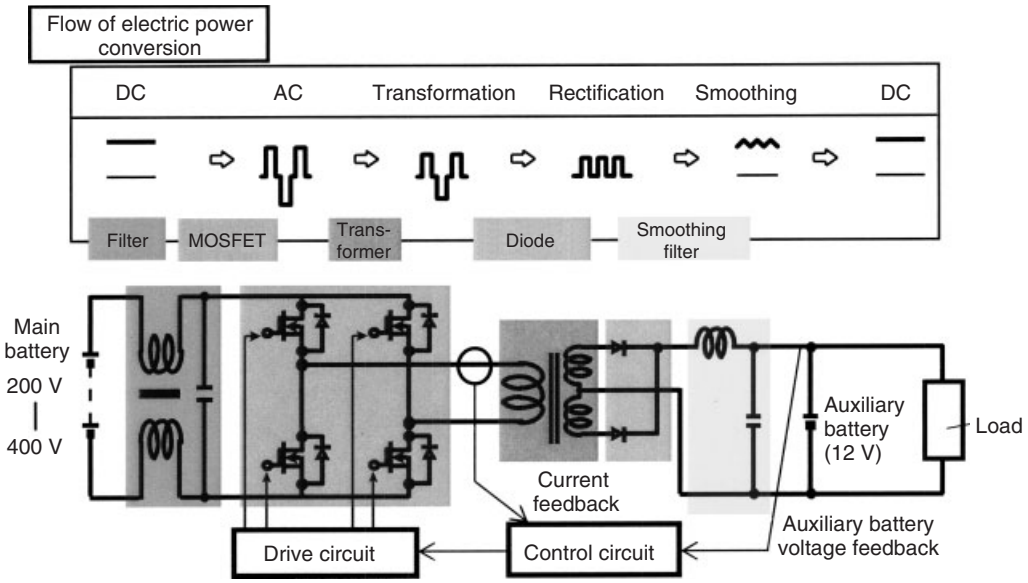


Figure 30. Principle of electric power conversion.

in voltage fluctuations. This in turn would lead to flickering lights and fluctuations in the volume of air coming from the blower. Therefore, the method used for battery charging in a HV is completely different. The voltage from the main battery is converted and this is used for charging. The device that converts the voltage from the main battery is the DC–DC converter. Figure 30 shows the principle of electric power conversion.

The high voltage DC from the main battery is subjected to high frequency switching in the power devices (MOSFETs) and then the voltage is transformed (lowered) by the transformer. This voltage is then rectified by the diode, smoothed by the smoothing filter, and converted into low voltage DC for charging the auxiliary battery. The transformer is used as a part of the safety design of the DC–DC converter to prevent the fires and shocks that would occur if high voltage was accidentally applied to the low voltage side due to some kind of malfunction. The high frequency switching allows for the size of the transformer to be reduced significantly. Figure 31 shows a photograph of an actual DC–DC converter for use in HVs.

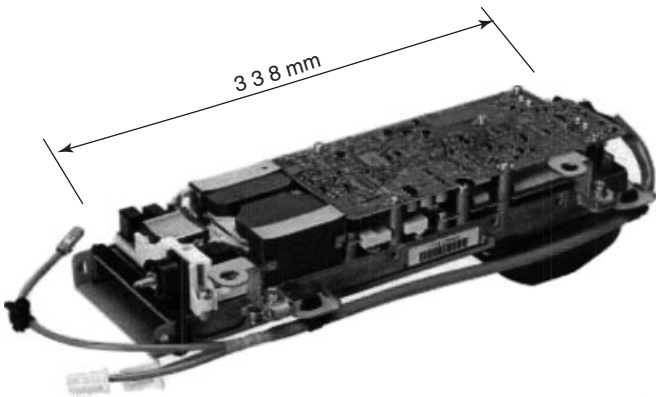


Figure 31. DC–DC converter for HV.

8 HIGH VOLTAGE AUXILIARY SYSTEM

The engine is frequently stopped in a HV, so it is necessary to examine the feasibility of converting all of the loads that are driven by the engine to electric power. A HV is equipped with a main battery so it is possible to reduce the size of systems with large load capacities by designing them to have high voltage specifications. This section describes systems in current commercial HVs that utilize these concepts.

8.1 Electric compressor system

In an ICE vehicle, the compressor is driven by the engine. In a HV, however, the electric motor drives the compressor so that the air conditioner does not turn off each time that the vehicle comes to a stop. The air conditioner places a large load on the vehicle, so this system is designed to have high voltage specifications. The system is installed in the HV in the same location as on an ICE vehicle. Therefore, efforts are being made to combine the compressor, motor,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

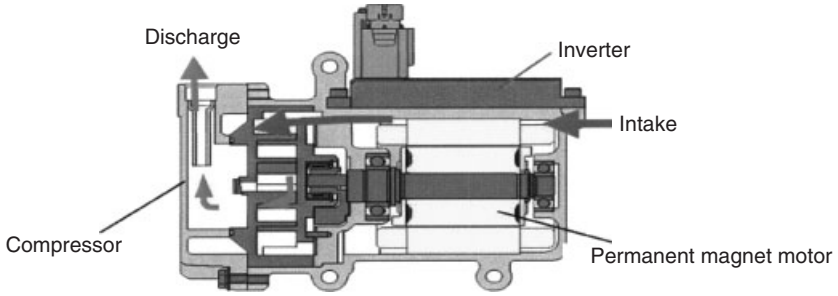


Figure 32. Structure of electric compressor.

inverter, and control ECU into a single unified structure and reduce its size. Figure 32 shows this structure.

The electric motor is a permanent magnet motor (IPM). The motor coil and inverter utilize the refrigerant of the compressor for their own cooling. Special consideration was given to how the components are mounted and radiate heat and to the water-resistant structure to satisfy all of the environmental requirements, such as vibrations, temperature, and water exposure.

9 FUTURE TRENDS

HVs have become popular and accepted by the market in the 10 years since their launch in 1997. These are groundbreaking vehicles that are environmentally friendly and have realized excellent levels of fuel economy without placing a larger burden on their users. It can be said that these vehicles changed the consciousness of their users and indicated the direction that all future vehicle design should move in. However, current HVs merely mount an electric motor system onto an existing engine system, so they suffer from fundamental problems of mounting all the necessary components and controlling the total cost. The key to further improving the popularity and widespread use of HVs is reducing the size and cost of the components that make up the electric motor system.

The major issue for HVs is the main battery. HVs were only realized once Ni-MH batteries that can endure frequent and repeated charging and discharging were developed. Currently, Li-ion batteries that possess even higher performance are being developed. It is expected that this will lead to significant reductions in size and increases in power. However, the development of batteries with even greater performance is expected and needed for use in EVs, PHVs, and extended range EVs. Electric motors and inverters are critical components of HVs. Electric motors have a long history, but there is still plenty of research that remains to be done into issues such as winding methods and magnet placement to further reduce their size. The development of new power devices for inverters is also widely expected and efforts to reduce their size through improved cooling technologies are continuing to advance.

Research into vehicle controls, such as motor and battery control, is essential to using energy as efficiently as possible. There are many other HV components. Research into the best ways of reducing their size and cost is also being pursued aggressively. The number of people now working on HVs and EVs has increased dramatically compared to 10 years ago. This of course includes those working at vehicle and component manufacturers, but it has now expanded to include active efforts by those at electronics manufacturers and various research institutes. This trend is not limited just to Japan but has become a worldwide movement. This large shift in focus and resources toward HVs and EVs will not stop and will likely only continue to grow larger in the future.

FURTHER READING

Denso Car Electronics Research Committee (2010) *An Illustrated Guide to Car Electronics* (in Japanese), Nikkei Business Publications, Inc, Tokyo.

Electric Vehicle Handbook Editing Committee (2001) *Electric Vehicle Handbook* (in Japanese), Maruzen Publishing Co., Ltd, Tokyo.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto226
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Body ECU (Airbag)

Takashi Noguchi
DENSO Corporation, Kariya, Japan

1 Introduction 1
2 Collision Mechanics 2
3 Airbags 2
4 Airbag System 6
5 Airbag Sensing System 8
6 Airbag ECU 10
7 The Future of Airbag 17
Further Reading 17

1 INTRODUCTION

Airbags are a technology that protects the occupants of a vehicle in an accident. When a vehicle collides with an object, this is referred to as the *primary collision*. The vehicle will decelerate rapidly if the object that it collides with has a heavy mass and high rigidity. However, the occupants will continue to move forward because of the existing inertia before the collision and will hit their heads and chests on items in the vehicle interior. This is referred to as the *secondary collision* and it accounts for the great majority of injuries in a vehicle accident. The same occurs in the case of a collision between a vehicle and a pedestrian or a bicyclist. The secondary collision between the person and the vehicle pillar, hood, the base of the windshield wipers, or if they hit their head on the road surface, results in much more serious injuries than the primary collision.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Collision safety devices, such as putting a cushion (i.e., an airbag) between the person and the collision object, or restraining the occupants in their seats to prevent movement due to inertia (i.e., with seat belts), reduce the injuries caused by the secondary collision. In the event of a side impact collision, the primary collision also becomes a problem depending on how far the object penetrates into the vehicle interior. However, the basic safety principles behind the devices used to protect the vehicle occupants do not change in this case. Airbags are still deployed as a cushion and seat belts help prevent the occupants from being thrown out of the vehicle. Collision safety devices have a long history. The first safety devices were the two-point seat belts installed by the Ford Motor Company on their vehicles in 1955. The airbag was invented by Yasusaburo Kobori, a Japanese, in 1963, but it was not until 1973 that it was actually installed in vehicles for the first time by General Motors (GM). In both cases the initial safety devices were mechanical in nature, but advances in electronics technologies have led to the addition of numerous new functions and improvements in performance. Airbag technology in particular has advanced and adopted electronic acceleration sensors (G sensors), pressure sensors, roll rate sensors, and the like. These sensors enable the airbags to handle a variety of different kinds of accidents, including side impact collisions and rollovers, in addition to frontal collisions.

In recent years, other safety technologies have been placed on the market, such as crash compatibility technology (which protects the occupants of the other vehicle in the event of a collision with a compact car) and airbags and active hoods (which protect pedestrians). These kinds of technologies, which protect the other party in the collision, have been advancing and evolving.

There are also other safety technologies such as fuel shut-offs, electric power shut-offs, and mayday systems in

addition to those that protect vehicle occupants directly. In the case of a collision that is strong enough to cause the airbags to deploy, the fuel shut-off closes the valve on the fuel tank after the collision to prevent any secondary accident due to a fire. The electric power shut-off performs the same role in vehicles that possess a high voltage power supply, such as hybrid vehicles and electric vehicles. The electric power is shut off in the event of an accident to prevent electric shocks when the occupants are rescued. Some vehicles are even equipped with a data communication module (DCM) that uses the band frequency of mobile phones to contact a call center with an SOS message after a collision has occurred.

2 COLLISION MECHANICS

This section provides an explanation of the mechanics of a collision using a frontal collision between two vehicles as an example. Figure 1 shows a frontal collision between vehicle A, with a mass of m_1 and a velocity of v_1, and vehicle B, with a mass of m_2 and a velocity of v_2. The hatched portions on the front of both vehicles are referred to as the *crushable zones*. These zones absorb the kinetic energy of the collision by deforming adequately and control the acceleration at the time of the collision. On the other hand, the occupant compartment is referred to as a *noncrushable zone* and the rigidity of this compartment is increased to ensure the survival space of the occupants. This *vehicle structure* is referred to by different names depending on the vehicle manufacturer, but in general it is referred to as a *collision-safe body* and it helps ensure the integrity of the vehicle.

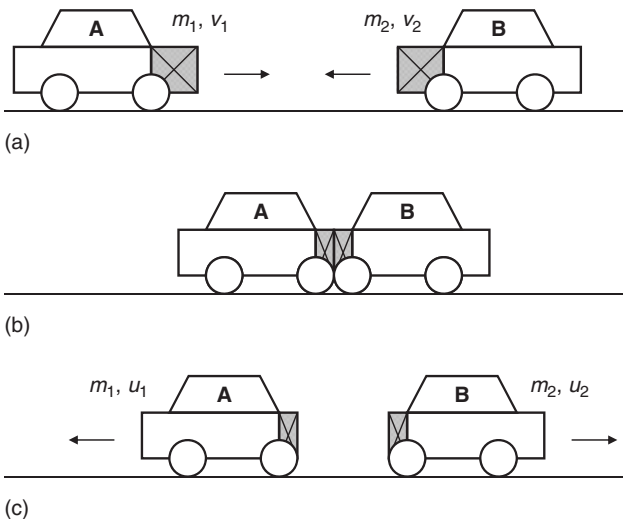


Figure 1. Frontal collision between two vehicles. (a) Before the collision. (b) Collision is complete. (c) After the collision. (Reproduced by permission of Denso Corporation.)

At time in Figure 1b, the collision is ongoing, but the velocity of the center of gravity is zero and the crash stroke (amount of crumpling) of both vehicles has reached the maximum. After this, as shown in Figure 1c, both vehicles advance in opposite directions because of rebound. Finally, the vehicles will stop moving because of the force of friction. This type of collision is referred to as an *inelastic collision* and results from the absorption of the kinetic energy by the crushable zones. The rebound coefficient (e) is approximately 0.2 in this frontal collision. The amount of change in velocity before and after the collision is divided proportionally between each of the masses involved because the momentum is conserved. Consequently, the vehicle in this collision that has a smaller mass will have a larger amount of change in velocity before and after the collision. This amount of change in velocity largely conforms to the velocity of the secondary collision and so the injuries suffered by the occupants in the vehicle with the smaller mass will be more severe.

Collision phenomena differ depending on the type of collision, but the crash stroke generally reaches its maximum from 50 to 150 ms after initial impact. If it is assumed that the rebound coefficient is approximately 0.2 and the amount of change in velocity is 60 km/h in a collision between two vehicles traveling at 50 km/h, then an average acceleration of 350 to 110 m/s^2 is applied with a peak acceleration of 700 to 220 m/s^2. Therefore, this is an extremely intense phenomenon in which a force of 10 times to 70 times a person's body weight is applied in a time of about 0.1 s.

The injuries suffered by occupants due to the secondary collision are also determined by the rigidity of the objects that the occupants collide with. The strength design of items such as the steering wheel, dashboard, and interior materials are important factors in ensuring the safety of the basic vehicle itself.

3 AIRBAGS

3.1 Introduction to airbags

Initially, airbags were installed within the steering wheel to protect the person in the driver's seat in the event of a frontal collision. This was followed by airbags in the instrument panel to protect the person sitting in the front passenger's seat. Now airbags are also installed within the seats (side airbags) and within the headlining (curtain airbags) to protect passengers in the event of a side impact

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

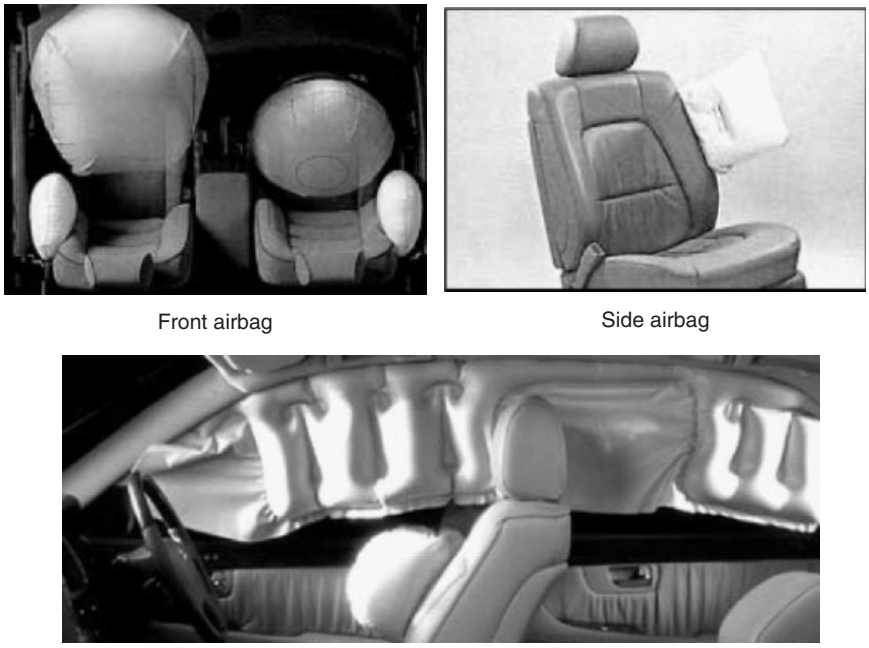


Figure 2. Examples of airbags. (Reproduced by permission of Denso Corporation.)

collision. There are also some curtain airbags that have a larger width and cover more area to prevent passengers from being thrown out of the vehicle in a rollover crash. In addition, there are also knee airbags installed in the lower portion of the instrument panel to protect the lower legs of passengers and active hoods to help protect pedestrians struck by a vehicle. Airbags to protect pedestrians are also being developed. Figure 2 shows images of typical airbags. Figure 3 shows the structure of the airbag modules for the driver's seat and the front passenger's seat airbags. The airbag for the driver's seat is installed within the steering column and the airbag for the front passenger's seat is installed within the instrument panel. The main parts of an airbag are the inflator that produces the gas and the bag itself. The airbag is folded up and the gas is injected into it causing the bag to expand and burst through the surface material when it is deployed.

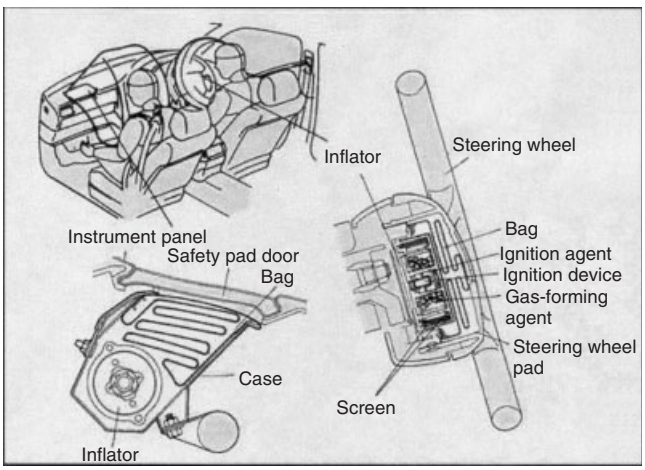


Figure 3. Structure of airbag modules. (Reproduced by permission of Denso Corporation.)

Airbags are currently installed in most vehicles. However, airbags did not become widely used after they were invented. This was because an explosive powder was used during airbag inflation and the extremely intense force that the occupants were subjected to when the airbags operated unnecessarily caused serious adverse effects. GM installed airbags as optional equipment on mass produced vehicles starting in 1973, but accidents due to malfunctioning airbags occurred and production was discontinued in 1977. The current popularity and widespread use of airbags in vehicles started in 1987 when the installation of airbags in vehicles became required by law in the US. At that time, the US government was under pressure from an active consumer movement and a law known as *FMVSS208* was implemented. Vehicles in Japan and Europe were required to be equipped with seat belts but they were not required by law to be equipped with airbags at that time. The legal requirement for vehicles

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 4. Example of 64 km/h ODB test. (From http://www.euroncap.com/results/bmw/5_series/2010/401.aspx. Reproduced by permission of Euro NCAP.)

to be equipped with airbags in the US greatly increased awareness of vehicle occupant safety issues and prompted vehicle manufacturers to compete with each other on the basis of safety features. This in turn led to a major increase in development and the advancement of widespread airbag use.

Another factor that accelerated the use of airbags in vehicles was the new car assessment program (NCAP). This program evaluates the safety performance of new vehicles, gives the tested vehicles grades, and then announces them to the public. A typical evaluation is the 64 km/h offset deformable barrier (ODB) test. In this test, the vehicle is forced to collide with a fixed barrier that has an aluminum honeycomb block attached to the front of it that simulates a collision with another vehicle. This is a very strict test that evaluates the occupant protection performance in a frontal crash between two vehicles at a relative velocity of 128 km/h. Figure 4 shows an example of this test. This test has practically become an international standard through the adoption of the JNCAP (Japan new car assessment program) in Japan, the Euro NCAP in Europe, and the ANCAP (Australia new car assessment program) in Australia. The existence of these assessment programs has spurred on the safety competition between the world's vehicle manufacturers.

Table 1 shows the evaluation results from Euro NCAP. The evaluations of the NCAP are called *star ratings* and the level of performance is expressed by the number of stars. The table clearly shows that new models of vehicles have improved safety performance in comparison to older models, even for the same vehicle by the same manufacturer, because the number of stars they received increased. This demonstrates how the benefits to vehicle users have increased thanks to this fair safety competition. In the twenty or so years since airbags have become readily available in the market on a large scale, the safety of vehicles has increased dramatically.

Table 1. Euro NCAP evaluation results.

Large family cars		
Model	Year	Rating
Renault Laguna	2001	☆☆☆☆☆
Audi A4	2001	☆☆☆☆
Honda Accord 1.8i LS	1999	☆☆☆☆
Mercedes C -class	2001	☆☆☆☆
Rover 75	2000, 2001	☆☆☆☆
Saab 9 -3	1999	☆☆☆☆
Volkswagen Passat	2001	☆☆☆☆
Volvo S40	1997	☆☆☆☆
Mitsubishi Carisma	2001	☆☆☆
Nissan Primera	1997	☆☆☆
Volkswagen Passat	1997	☆☆☆
Audi A4	1997	☆☆☆
Ford Mondeo	1997	☆☆☆
Renault Laguna	1997	☆☆☆
Vauxhall Vectra	1997	☆☆☆
Mercedes C -class	1997	☆☆
Peugeot 406	1997	☆☆
BMW 3 -series	1997	☆☆
Citroen Xantia	1997	☆☆
Rover 600	1997	☆☆
Saab 900	1997	☆☆

Reproduced by permission of Denso Corporation.

3.2 Operating principle of airbags

Figure 5 shows the operating principle of an airbag. After a collision, the front satellite sensor that is equipped on the front of the vehicle, and the G sensors that are located

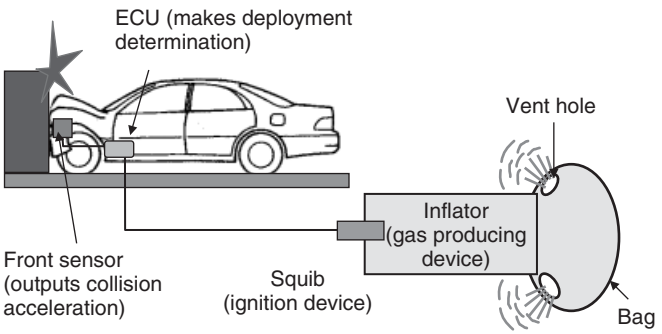


Figure 5. Airbag operating principle. (Reproduced by permission of Denso Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

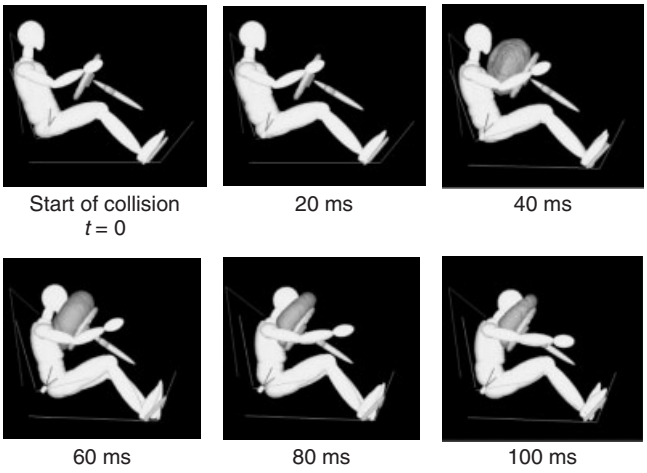


Figure 6. Deployment of airbag and occupant behavior. (Reproduced by permission of Denso Corporation.)

within the engine control unit (ECU) detect the collision. If the ECU determines that deploying the airbag is necessary, power is applied to the squib, which then ignites. The ignition agent then transfers this heat to the gas-forming agent, which produces a large amount of gas. This gas then fills the airbag, the pressure increases, and the airbag bursts out of the surface material and deploys. Holes in the airbag called *vent holes* allow the gas in the air bag to escape when the vehicle occupant collides with the bag and the pressure inside the bag rises. These vent holes prevent the pressure inside the airbag from rising too high. Figure 6 shows the behavior of an occupant at the time of a collision and the appearance of the airbag when it deploys. The scene shown in this figure is of a collision with a fixed barrier at 50 km/h. This is an extremely severe collision that is equivalent to a frontal collision between two vehicles each traveling at 50 km/h.

It takes approximately 10 ms after the start of the collision for the ECU to determine if the collision requires the airbag to deploy and ignites the squib. First the airbag vigorously bursts through the surface material and shoots out to the maximum stroke (primary expansion). Then the airbag inflates out to the sides (secondary expansion). Approximately 30 ms elapse between these two expansions. At 40 ms after the collision the secondary expansion is finished and all preparations to catch the occupant are complete. At 60 ms after the collision the occupant's head begins to drive into the airbag. The occupant is restrained as the gas is gradually expelled from the vent holes and the restraining process finishes at approximately 100 ms. The required sensing operations finish in 10 ms while the airbag's job is completed and all operations are finished in just 100 ms. Precise operation over this extremely short time to protect the life of the occupant is fully expected and this can be said to be the fundamental purpose of airbag technology.

The fact that it takes 30 ms until the secondary expansion is complete indicates how difficult it is for the airbag sensing to predict the acceleration after this 30 ms has elapsed. Of course it is best for the sensing to be completed as quickly as possible, but predicting the acceleration 100 ms later based on the acceleration after just 10 ms is necessary. The acceleration that is produced by the collision will naturally be different depending on the structure of the vehicle body and the type of collision that occurs. Therefore, the airbag sensor tuning is performed based on data obtained from crash tests of actual vehicles at a variety of different speeds and involved in various different types of collisions.

3.3 Crash tests

Figure 7 shows a crash testing facility. There is a motor behind the massive concrete wall that acts as the fixed barrier and this motor tows actual vehicles toward the barrier using a wire rope. Once the vehicle approaches near to the barrier, the wire rope releases from the latch on the vehicle and then it collides with the barrier under its own inertia. Crash test dummies that simulate occupants in the vehicle are placed on board and these dummies are fitted with various sensors so that the injuries sustained during the collision can be evaluated. Several G sensors are also attached to the vehicle and a total of over 100 channels of measurements are taken.

Figure 8 shows examples of the different types of collisions used in crash testing. There are three basic kinds of crash testing: testing at the lower limit velocity at which the airbags must operate, testing at the upper limit velocity at which the airbags must not operate, and testing at the upper limit velocity to confirm the occupant protection performance. The crash testing is carried out at a great variety of different velocities for each of the various types of collisions. There are also other non-crash tests that are performed as intentional defect tests when working on the airbag tuning, such as a rough road running test, extreme rough road test, and an abuse test. There are also non-operational mode tests. All of this testing is carried out to ensure the performance of the airbag system in the market.

There are many different kinds of crash test dummies with different body types used in the crash tests. The dummy called *AM95* is the largest and it represents American males in the 95 percentile, in other words the largest 5% of body types. The AM50 dummy is the standard

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

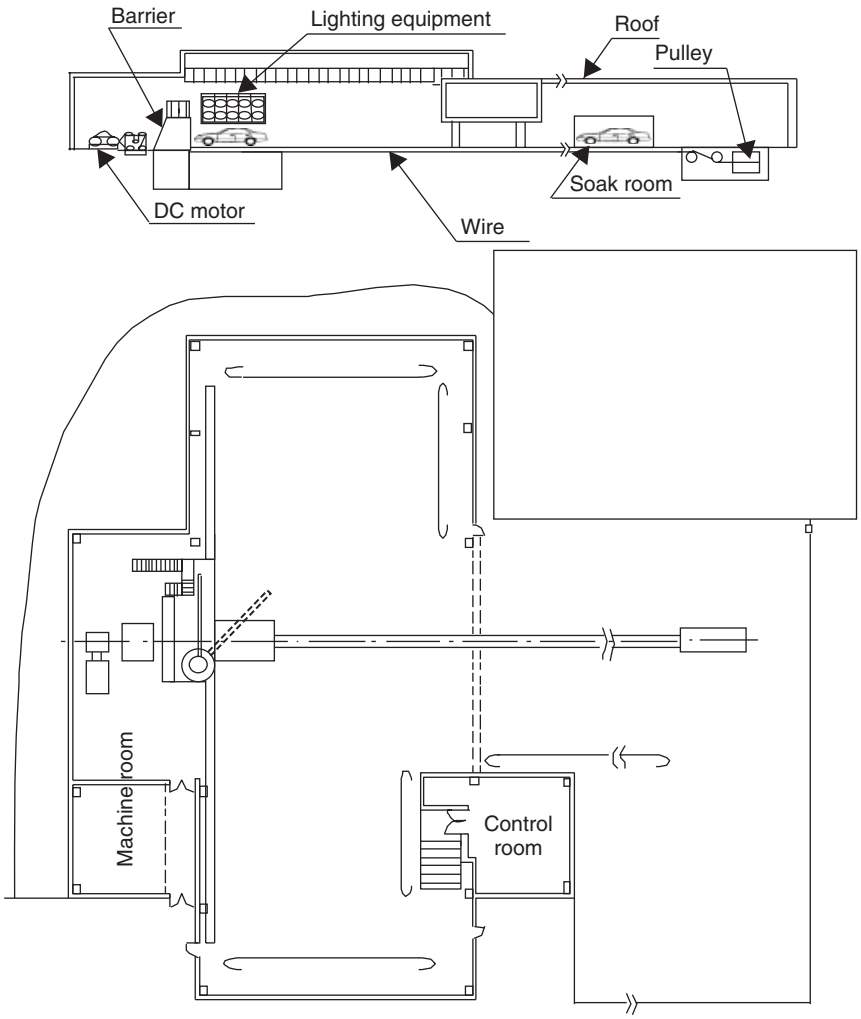


Figure 7. Crash test facility. (Reproduced by permission of Denso Corporation.)

body type for an American male and it is used in the laws and regulations, as well as for the safety ratings. The AF05 dummy simulates the body type of a smaller American woman and there are also dummies that simulate the body types of a 6-year-old child and a 3-year-old child. Figure 9 shows an image of a family of these crash test dummies.

4 AIRBAG SYSTEM

Figure 10 shows the configuration of a typical airbag system. The main parts are the satellite sensors for both frontal collisions and side impact collisions, the ECU, the airbag modules, and the seat belt pre-tensioners. The satellite sensors are located in the crushable zones to speed up the response of the system when a collision occurs. These sensors transmit the acceleration data to the ECU in a collision. The ECU is located on the floor of the center console. The ECU has two internal G sensors that have different detection axes for frontal collisions and side impact collisions. The ECU combines the data from these G sensors with the acceleration detected by the satellite sensors to make the calculations and then the determination about whether a collision is occurring. At the same time, if it determines that deployment of the airbags is necessary, the ECU will also provide current for ignition to the airbag modules and the seat belt pre-tensioners. The ECU possesses this ignition function and a malfunction diagnostic function to check the system for any malfunctions. The ECU also possesses an energy reserving function using a capacitor to ensure that the ignition current is supplied, even in the event that the battery becomes disconnected during the collision. Figure 11 shows an

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

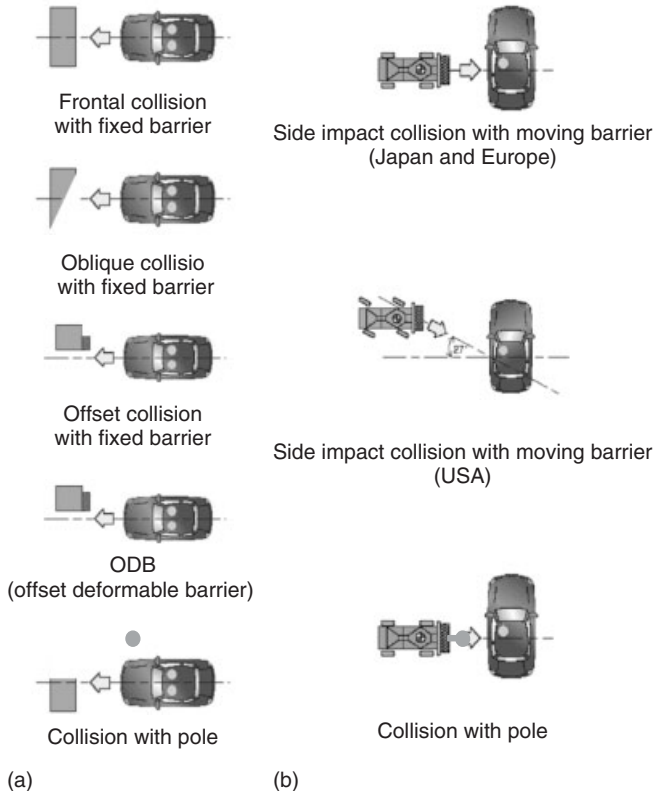


Figure 8. Types of crash tests. (a) Frontal collisions. (b) Side impact collisions. (Reproduced by permission of Denso Corporation.)

image of an ECU. Figure 12 shows a block diagram of the ignition system

The malfunction modes are broadly classified into two kinds of malfunctions: ignition and airbag deployment when it is not necessary and failure to deploy airbags when they are necessary because the ignition could not be performed. The overall design concepts of the system described below demonstrate the consideration given to preventing inadvertent airbag deployment.

- Improve the reliability of parts.
- Adopt a series redundant structure so that a single malfunction will not lead to inadvertent airbag deployment.
- The system shall be able to detect any malfunctions that may lead to inadvertent airbag deployment.
- A warning light or other means shall be used to alert the vehicle user when a malfunction is discovered and encourage them to have it repaired.
- Simultaneous processing so that the next malfunction does not lead to inadvertent airbag deployment.

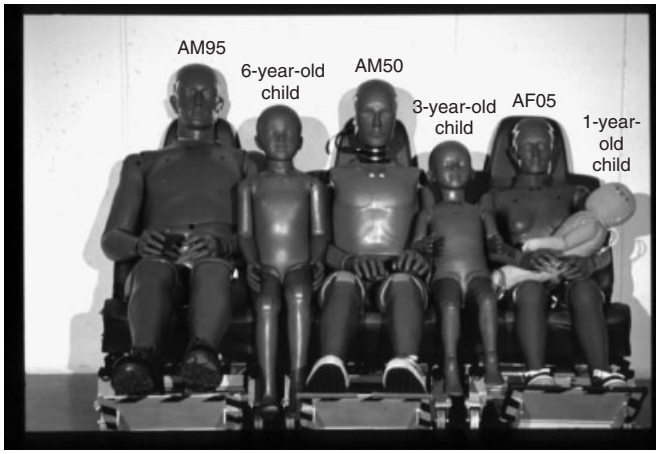


Figure 9. Family of crash test dummies. (Reproduced by permission of Denso Corporation.)

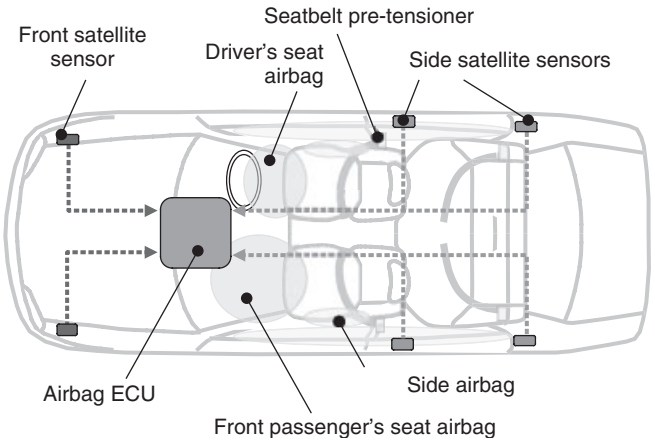


Figure 10. Configuration of airbag system. (Reproduced by permission of Denso Corporation.)

The following design considerations also aim to prevent airbag deployment failure.

- Improve the reliability of the parts.
- The system shall be able to detect any malfunctions that may lead to airbag deployment failure.
- A warning light or other means shall be used to alert the vehicle user when a malfunction is discovered and encourage them to have it repaired.
- Adopt a parallel redundant structure to handle malfunction modes that cannot be detected.

The reliability of the airbag system is ensured through the design concepts listed above. In recent years, some airbags for frontal collisions are being equipped with two squibs. These squibs may ignite simultaneously or one may ignite

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

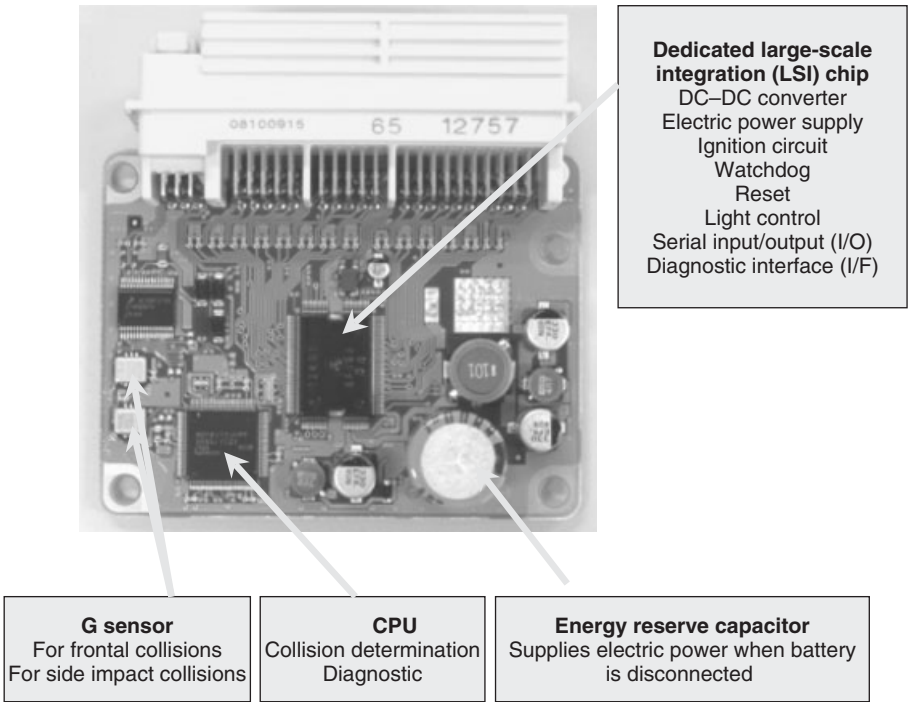


Figure 11. Airbag ECU. (Reproduced by permission of Denso Corporation.)

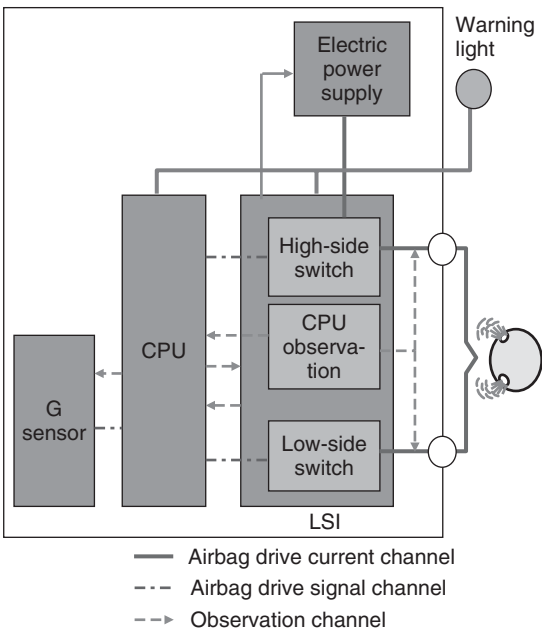


Figure 12. Configuration of ignition system. (Reproduced by permission of Denso Corporation.)

after a delay depending on the severity of the collision to control the increase of pressure within the airbag. In the case of side impact collisions, it is not uncommon for there to be 14 channels or more of squibs operating in conjunction with the curtain airbags and the seat belt pretensioners.

Consequently, all of these circuits are integrated and are mounted in the vehicle as part of a large-scale application specific integrated circuit (ASIC). This is a key part of the airbag system for realizing low cost.

The airbag system is extremely complicated and a malfunction may have a direct impact on human life. Therefore, the ECU is equipped with malfunction memory, which keeps a record of all malfunctions that occur so that they can be read out later. The NHTSA (National Highway Traffic Safety Administration) is promoting the standardization of this memory function through the event data recorder (EDR). The acceleration waveforms and the amount of change in velocity at the time of a collision are also recorded so that the kind of collision that occurred can be re-created during the investigation into a vehicle accident.

5 AIRBAG SENSING SYSTEM

The principle behind the airbag sensing system is extremely simple. If the type of collision is the same, then the higher the collision velocity, the greater the acceleration that is produced. Therefore, it should only be necessary to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

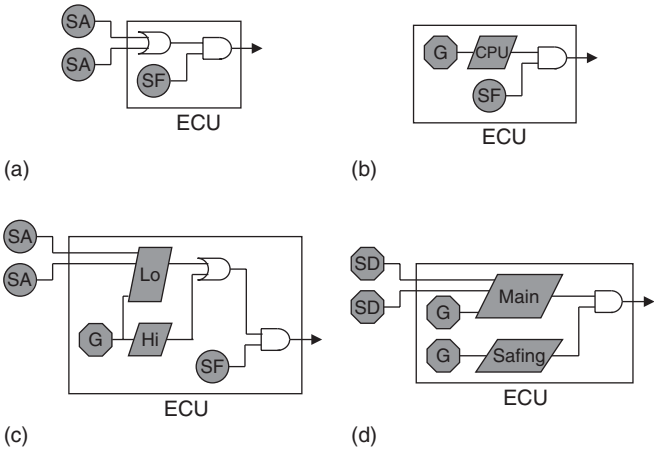


Figure 13. Sensing system evolution. (a) Electromechanical multi-point; (b) electronic single-point; (c) electronic electromechanical multi-point; (d) electronic multi-point. (Reproduced by permission of Denso Corporation.)

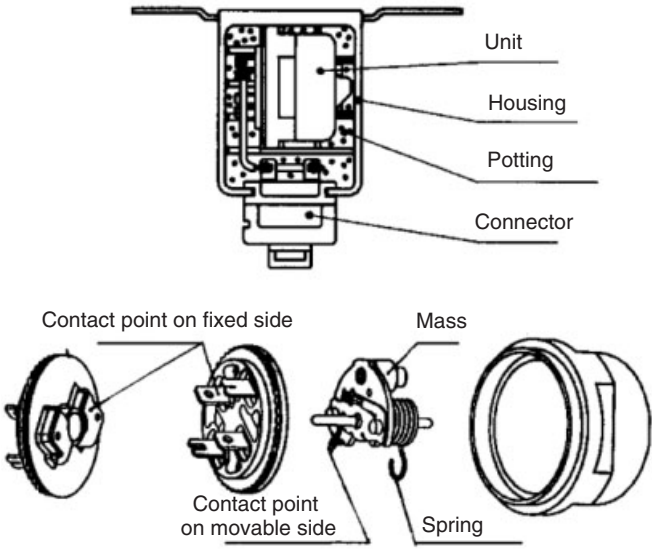


Figure 14. Electromechanical satellite sensor. (Reproduced by permission of Denso Corporation.)

compare the magnitudes of the accelerations. However, this is complicated by the fact that the vehicle deformation characteristics change when the type of collision is different. The rigidity of the vehicle crushable zones is different in a 100% full-lap frontal collision and a 40% offset frontal collision. There are cases where a higher acceleration is produced from a low velocity full-lap frontal collision than from a high velocity offset frontal collision, so sensing becomes difficult. Figure 13 shows the changes in airbag systems over the years. More details about the sensing performed by the airbag system will be presented in the following sections. This evolution in the sensing system demonstrate the history of efforts to solve this acceleration sensing issue.

5.1 Initial sensing system

Figure 13a shows the initial airbag system. The main sensors are the two electromechanical sensors at the front of the vehicle and the electromechanical safing sensor located within the ECU. An electromechanical sensor utilizes a spring-mass system mechanism. The acceleration at the time of a collision is applied to a spring-biased mass and this mass moves, closing the electrical contact. Figure 14 shows a structural diagram of this sensor. This system is designed so that ignition will not occur if the main sensors at the front of the vehicle and the safing sensor in the ECU are not all turned ON simultaneously. This forms a series redundant system. Multiple electrical contacts are provided within the sensors where the ignition current flows directly. This forms a parallel redundant system to prevent the airbag from failing to deploy.

The characteristics of an electromechanical sensor are determined by the equation of motion of the spring-mass system. The sensor has low degree of freedom for tuning, diagnostics cannot be performed on the sensor to prevent an airbag deployment failure, and it has a high cost. Therefore, this type of sensor was gradually replaced by electronic sensors and it is no longer used. Following this, sensing systems appeared in which an electronic G sensor was added within the ECU. This system was equipped with an analog arithmetic integrated circuit (IC), but the degree of freedom for tuning was low, and this system is also no longer used.

5.2 Single-point sensing

Figure 13b shows a sensing system referred to as a *single-point system*. It is equipped with one electronic G sensor and one electromechanical safing sensor. Both of these sensors are located within the ECU. The purpose of this design was to reduce the cost of the sensing system. However, the ECU was located on the floor of the occupant compartment, which is a noncrushable zone, so it transmitted the acceleration information at a slow speed. This made it difficult for this sensing system to conform to the gradually increasing performance requirements of the NCAP, and so it currently occupies only a minor position among the other types of sensors. After this, an attempt was made to use a microcomputer to calculate the output from the G sensor within the ECU and make the system capable of performing the collision determination. This was achieved through improving

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

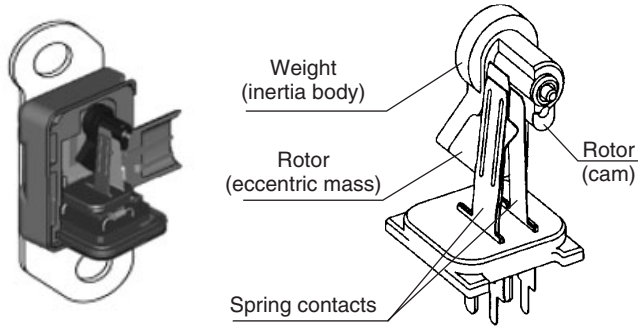


Figure 15. Electromechanical auxiliary sensor. (Reproduced by permission of Denso Corporation.)

the performance of the microcomputer and lowering the cost. It also increased the degree of freedom for tuning and made it possible to perform even more meticulous settings.

5.3 Electromechanical satellite sensing

Figure 13c shows a sensing system equipped with an electromechanical satellite sensor. The electromechanical sensor was revived, but in comparison to the one in the initial sensing system, the sensor itself is low performance and low cost. The ignition current does not flow directly through the sensor. Instead, it outputs the contact point ON/OFF signals to the CPU (central processing unit), so it is referred to as an *auxiliary sensor*. In comparison to the initial sensing system, the performance of this system was focused on the detection of large impacts coming from the front of the vehicle. This system was introduced to meet the demands for improved sensing system performance. However, it is continuing to disappear under the wave of electronic sensors that have high degrees of tuning freedom as the cost of those sensors continues to decline. Figure 15 shows a structural diagram of an electromechanical auxiliary sensor.

5.4 Electronic satellite sensing

Figure 13d shows the newest sensing system. The satellite sensors in Figure 13c have been replaced with electronic ones. The figure shows that the acceleration waveforms from the satellite sensors and the signals from the two G sensors within the ECU are sent to the CPU where calculation is made freely and the collision determination is made. The degree of freedom for tuning has also increased dramatically. Originally the electronic satellite sensors communicated with the ECU via serial communication, but currently a bus communication system called a



The G sensor and communication IC are packaged together. The detected acceleration is sent to the ECU via digital communication

Figure 16. Electronic satellite sensor. (Reproduced by permission of Denso Corporation.)

sensing bus is the mainstream method being used. This bus has special reliability requirements that are different from a general vehicle-mounted bus. For example, the communication will not be interrupted even if it contacts the earth during a collision. A bus that is exclusively for airbags is used. Figure 16 shows an image of an electronic satellite sensor.

As described in the previous sections, the airbag sensing system is now based on electronic G sensors that possess a high degree of tuning freedom to realize a higher level of sensing performance. It has evolved into an electronic multipoint system that makes a comprehensive judgment of the sensing results coming from multiple locations.

6 AIRBAG ECU

6.1 Configuration of ECU

Figure 17 shows the configuration of an airbag ECU. The ECU consists of the ignition circuit area, the collision determination area, the auxiliary power supply area, and the malfunction determination area.

6.2 Ignition circuit area

The circuits that drive the squibs that deploy the airbags receive commands from the collision determination area through the signals from each sensor. The configuration of these circuits is different depending on the number of loads (driver's seat airbag, front passenger's seat airbag, seat belt pre-tensioners, driver's seat side impact airbag, and the like), the type of ECU, and other factors. The ignition circuit is designed to have serial redundancy by installing two or more independent drive circuits to prevent the inadvertent deployment of airbags due to a single malfunction (squib earth short, drive circuit malfunction, and the like) or when a malfunction is diagnosed. The reliability of the ignition circuit is further ensured through the performance of malfunction diagnostics. Malfunction

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

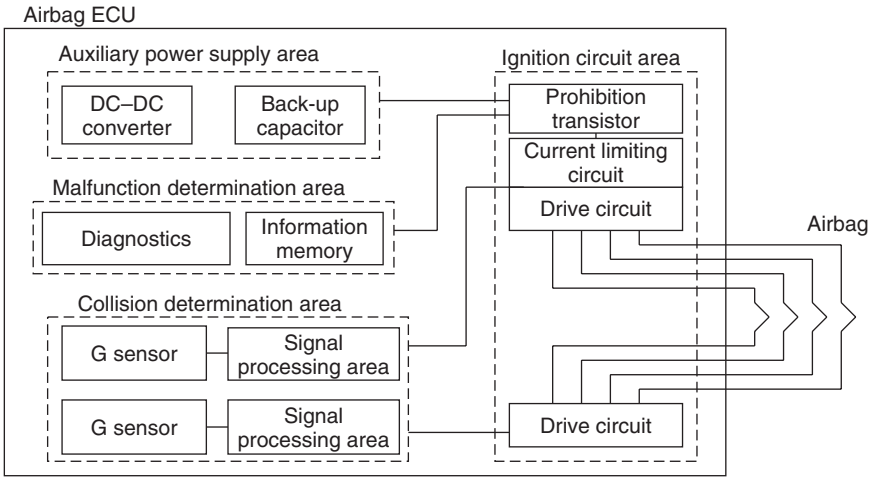


Figure 17. Schematic diagram of ECU. (Reproduced by permission of Denso Corporation.)

diagnostics help prevent airbag deployment failure and the vehicle user is encouraged to obtain a replacement part when a malfunction occurs. In the case of parts that cannot be checked through malfunction diagnostics, the reliability of the parts themselves must be ensured or they must be designed to have a parallel redundant configuration.

6.2.1 Drive circuits

These are the circuits that apply current to the squibs based on the signals from the collision determination area within the ECU. These circuits are located both upstream and downstream from the squibs and are designed to have a series redundant configuration.

6.2.2 Current limiting circuit

In the case where there are multiple squibs, as the impedance of each squib and wire is different, a squib may be short circuited by the electric power supply or earth due to a malfunction or ignition. Current will concentrate at that squib and the specified current will not be able to flow to the other squibs. The current is limited by constant current circuits and resistance shunt circuits to prevent this from happening and to ensure that the specified current flows to each squib.

6.2.3 Prohibition transistor

This is one element of a series redundant circuit. This circuit breaks the ignition circuit in the case where a malfunction occurs that may lead to inadvertent deployment of the airbags. In the past, a mechanical switch was often used for this purpose, but now this function is realized through the use of a semiconducting element. There are cases where the drive circuits also make use of this element.

6.3 Collision determination area

This determines whether it is necessary to activate the occupant protection devices (airbags and the like) by detecting the physical quantities (deceleration, amount of deformation, etc.) generated at the time of a collision. These physical quantities are detected by electromechanical and electronic sensors the collision determination is performed using those output signals in the processing circuit. The system is always composed of two or more determination pathways, which include the sensors and the signal processing circuit, to ensure the redundancy of the system.

6.3.1 Sensors

The sensors detect the physical quantities at the time of the collision and then generate the corresponding electric signals. In an electromechanical sensor, the collision determination is made in the mechanical portion and then this result (ON or OFF) is output as an electric signal. An electronic sensor also outputs an electric signal that corresponds to the physical quantities at the time of the collision. However, any components that are unnecessary for the collision determination (including noise and others) are removed and the sensor is equipped with a built-in low-pass filter for the purpose of waveform shaping. This low-pass filter also possesses an antialiasing function for digital signal processing.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6.3.2 Signal processing area

This processes the electric signals from the sensors and then determines whether the collision is strong enough to require the activation of the occupant protection devices. The collision determination has a serial configuration, comprising the main determination area and the safing area. The main determination area mostly determines the degree and severity of the collision, whereas the safing area is set so that it will not respond to normal driving or normal operations and that it will also not inhibit the main determination to ensure redundancy. In the case of a collision determination that uses electronic sensors, the determination is performed by an IC and general purpose CPU with an integrated collision determination algorithm, and by using the specific determination threshold value that is set for each vehicle.

6.4 Auxiliary power supply area

There are cases where the electric power supply to the airbag ECU declines or is cut off because of damage to the vehicle battery or because of the power line getting caught up in the vehicle body during a collision. There are also cases where it is necessary to operate the airbags even though the battery voltage has dropped due to an abnormality in the vehicle power supply system. Therefore, the ECU possesses an auxiliary power supply function to handle these kinds of situations. The auxiliary power supply area possesses enough energy to operate the airbags in approximately 100 ms in the event of a frontal collision and to operate them for 1 s in the event of a rollover crash after the ECU power supply has been cut off.

6.4.1 DC–DC converter

This circuit increases the voltage when the vehicle power supply voltage is low to ensure that there is enough electric energy to activate the squibs. The CPU and sensors of the ECU, as well as the satellite sensors possess enough voltage and capacity to be able to operate when the power supply voltage of the ECU is 6–8 V or more. The output voltage is 25–35 V to ensure that the energy reserve capacitor can efficiently store up energy.

6.4.2 Energy reserve capacitor

This is an energy storage device that enables airbag operation even if the battery or power supply cable is damaged during a collision. It is necessary to store up energy so that the collision determination area can operate, the squibs can be activated, and information can be written into the memory, or the like. This capacitor is composed of either one aluminum electrolytic capacitor or several of them. In general, the capacity is from 3000–15,000 μF. This capacitor is required to have a low equivalent series resistance because discharging large currents in a short time is necessary and it is also required to have good performance at low temperatures and after a durability test.

6.5 Malfunction determination area

This performs diagnostics to check for any abnormalities in the ECU and the system, including outside the ECU. A warning light is illuminated and the driver is notified when an abnormality is detected. Information related to the malfunction diagnostics result is recorded in the nonvolatile memory and other locations. There are cases where this information is later utilized for malfunction analysis and investigation into a vehicle accident.

6.5.1 Malfunction diagnostics

This area observes the internal workings of the ECU and the entire system as a whole. The driver is notified via a warning light when an abnormality occurs. In many cases it possesses a function that generates a code (diagnostic code) to identify the malfunctioning component via operation of the diagnostic terminal (set on the vehicle). The basic concepts behind the malfunction determination are as follows.

6.5.1.1 Inadvertent airbag deployment. In the case where the airbags deploy unnecessarily due to a malfunction (primary malfunction) and then the occurrence of another malfunction (secondary malfunction), the initial primary malfunction is detected and the driver is notified.

6.5.1.2 Airbag deployment failure. In the case where the airbags fail to deploy due to a malfunction, that malfunction is detected and the driver is notified. There are also cases where malfunction diagnostics to detect potential inadvertent deployment or failure to deploy cannot be performed. In those cases, redundant design is employed to cover those undetected items or the reliability of the parts is sufficiently ensured. The detection of a malfunction will illuminate a light for a fixed time when electric power is initially supplied to the ECU as an indicator and a means of informing the driver that the system is prepared, ready, and waiting. This also notifies the driver that the light is operating properly.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6.5.2 *Information memory*

This memory stores the results that were detected by the malfunction diagnostics function as well as other information so that the state of the system when an accident occurred can be understood to a certain extent. Usually, nonvolatile memory is used so that information can still be recorded in the memory, even if the power supply is shut off. EEPROM (electrically erasable programmable read only memory) is often used as the nonvolatile memory, but flash memory that is built into the CPU has begun to be employed recently.

6.5.3 *Procedures after malfunction detection*

Measures are often taken to prohibit the airbag ignition from operating due to a malfunction, in addition to notifying the driver by illuminating the warning light. For example, when a G sensor malfunction is detected, the output of ignition signals to the collision determination area is stopped to prevent the inadvertent deployment of the airbags.

6.6 G sensors

Figure 18 shows a conventional capacitive G sensor. The sensor head and processing circuit are contained within a package with a size of 6.2 mm by 5 mm. The sensor head measures approximately 2 mm square and is composed of a fixed electrode and a movable electrode. The position of the movable electrode changes, altering the gap between the two electrodes when the sensor head is subjected to acceleration. This acceleration is then detected as a change in the capacitance and this is converted into voltage by the sensor. The opposite is also true. The electrode can be moved by applying a voltage to the sensor, which means that the sensor can perform self-diagnostics.

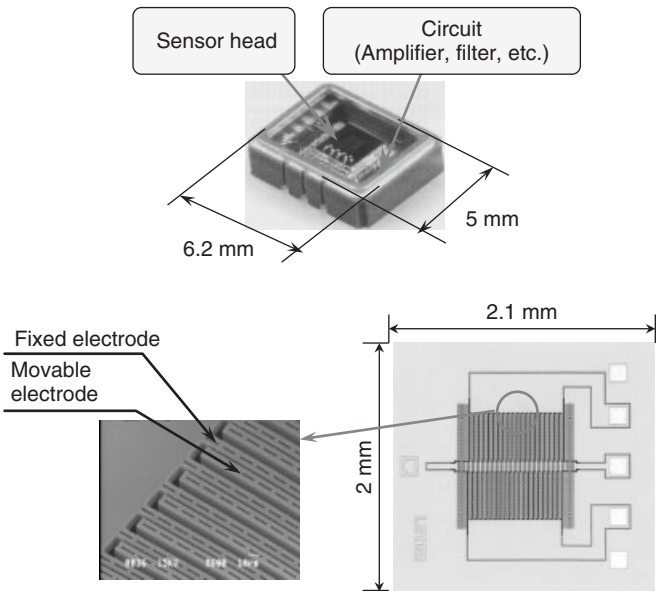


Figure 18. Capacitive G sensor. (Reproduced by permission of Denso Corporation.)

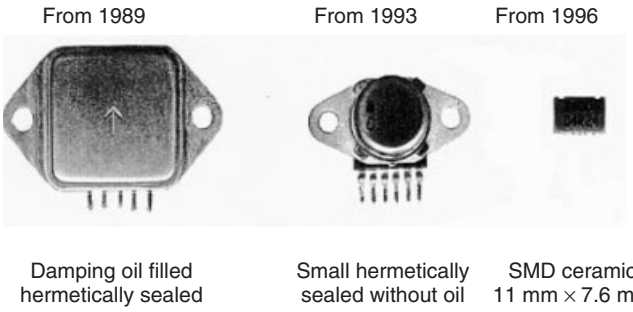


Figure 19. Evolution of piezoresistive G sensors. (Reproduced by permission of Denso Corporation.)

Figure 19 shows the evolution of G sensors. Piezoresistive sensors were used before capacitive-type sensors. The principle behind piezoresistive sensors is that a cantilever bends when it is subjected to acceleration and then this deformation is detected by a bridge circuit. The initial sensors did not have good sensitivity, so the sensor head was large and the sensors were encased in a large hermetically sealed package filled with damping oil. Following this, advances in both micro electromechanical systems (MEMSs) technology and circuit technology led to major miniaturization of the sensors down to 1/25 (projected area) of their previous size in just 7 years. The current day sensors are even smaller and are also highly precise.

6.7 Roll angle sensors

Current roll rate sensors are vibrating gyroscope-type sensors. The Coriolis force is generated in the tuning-fork-shaped oscillator when the vehicle rolls is converted into voltage and then output. AC voltage is constantly being applied to the oscillator that vibrates at a frequency of 2 kHz. When an angular velocity is applied due to the roll of the vehicle body, the apparent force that is generated is the Coriolis force and the oscillator becomes strained. The sensor then converts this strain into voltage. At the time that the actual determination of vehicle roll is made, the vehicle behavior is predicted in conjunction with the acceleration produced in the lateral direction. This means that these sensors are able to handle a variety of different roll modes.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6.8 Collision determination algorithm

6.8.1 Collision determination algorithm and tuning

The comprehensive judgment of whether to deploy the front impact airbags is based on the output from the satellite sensors in the crushable zone and the output from the G sensors in the ECU located in the noncrushable zone. The acceleration that is produced depending on the collision velocity and the type of collision is calculated and a judgment is made about whether it is necessary to deploy the airbags. The role of the collision determination algorithm is to send the ignition signals to the airbag modules if deploying the airbags is necessary. The collision determination algorithm possesses parameters that allow it to adapt to the characteristics of each vehicle. Simulations based on acceleration signals from actual crash tests are used to tune the algorithm. The calibration parameters are set after first establishing a margin in consideration of sensor error and variation in the collision. This thoroughly ensures the operation and non-operation of the airbags under real-world conditions. The necessity of air bag deployment is determined based on the specifications obtained from the vehicle manufacturers. The algorithm is tuned using the crash test data, the must fire/no fire requirements for airbag operation or non-operation, and the required response time [time to fire (TTF)] in the case of a must fire requirement. The collision determination must be completed within that amount of time.

6.8.2 Concepts behind collision determination algorithm

There are an infinite number of different types of collisions that occur in the real world, and it is impossible to test for all of these different collisions at a crash testing facility. Therefore, crash tests are conducted using limited test modes as typical examples, and the airbag sensors are brought into conformity with this data during the tuning process. Table 2 shows the typical test modes and the required TTF. These crash tests are conducted once the vehicle has reached the prototype stage. The number of crash tests that are conducted increases in the case of an important type of collision. Therefore, the airbag sensor tuning for just one model of vehicle normally requires that 50 or more crash tests be conducted using expensive prototype vehicles.

The collision determination algorithm is required to be robust enough to handle the different collision modes and velocities when it is being constructed. For example, an offset collision is normally referred to as a *40% offset* and the crash test is conducted so that the vehicle only has 40% overlap with the fixed barrier. Therefore, no crash test data exists between this 40% offset collision and the 100% full-lap frontal collision. The structural members of the vehicle have a strain rate dependency, so the acceleration that is produced depending on the velocity of the collision becomes nonlinear, even in the same type of collision. This means that an algorithm that correctly understands the collision phenomena must be constructed to ensure that the airbags will operate as intended, even in a collision mode that was not part of the crash tests. The vehicle collision waveform is composed of the waveforms of the accelerations produced by the deformation of the individual structural members and parts of the vehicle. A larger acceleration is produced at the time of a high velocity collision because of the strain rate dependency. No change in acceleration is produced in the extreme rough road mode. The algorithm must be based on the principles behind these kinds of physical collision phenomena. Stated in concrete terms, the collision waveform is based on the digital signal processing theory. The deformations of the individual structural members and parts of the vehicle comprise this collision waveform. These deformations produce pulse components. The algorithm then makes the collision determination by extracting these pulse components and combining their sizes and characteristic parameters. Consequently, the robustness of the algorithm is ensured by detecting the characteristic part deformation pulses, even if the collision lap rate is different. Figure 20 shows an actual collision waveform. The figure clearly shows how the overall waveform is composed of the combination of pulses produced by the deformations of various structures and parts.

The following section first describes the collision determination made by a floor-mounted ECU in a basic frontal collision. After this, the collision determination made in conjunction with the front satellite sensors is explained.

Table 2. Typical crash test modes.

Type of Collision	Velocity	TTF
Full-lap frontal collision	50 km/h	15 ms
	25 km/h	50 ms
	18 km/h	No fire
40% lap offset frontal collision	50 km/h	25 ms
	18 km/h	No fire
Oblique collision	50 km/h	30 ms
	35 km/h	50 ms
Pole	50 km/h	25 ms
ODB	64 km/h	25 ms
Extreme rough road	Pothole	No fire
	Driver over a bump	No fire
	Driver over a card	No fire

Reproduced by permission of Denso Corporation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

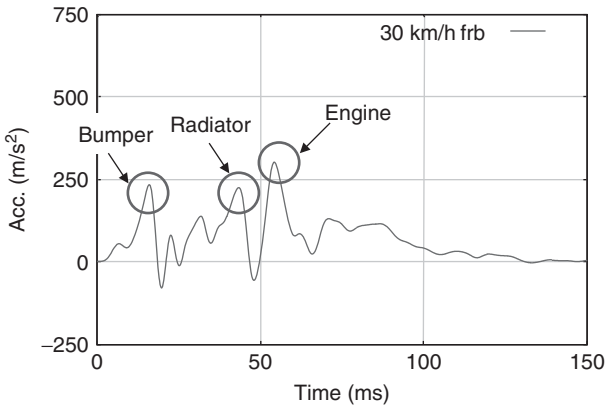


Figure 20. Collision waveform (30 km/h frontal collision). (Reproduced by permission of Denso Corporation.)

Figure 21 shows a block diagram of the collision determination algorithm.

6.8.3 *Preprocessing*

The signals from both the satellite sensors and the G sensor within the ECU are first subjected to preprocessing. Abnormality determination processing using a guard value is performed to carry out the malfunction determination for the G sensor. A high-pass filter that possesses time constants of several seconds to tens of seconds is used to correct any drift from the zero point due to the temperature and changes due to aging of the G sensor. A low-pass filter with a specified time constant is then applied for preprocessing of the signals before they move on to the subsequent stages. These filters are composed of digital filters and all of the processing is done by software.

6.8.4 *Signal analysis*

In the signal analysis area, the collision waveform is broken into pulses using a method called *wavelet conversion*. This method is used because it can detect pulses of various time constants to grasp the characteristics of various structural members and parts. In the case of a large vehicle that suffers a large amount of deformation to absorb a large amount of collision energy, the time constants of the pulses from the structural members will also become larger. However, a smaller vehicle may not be able to obtain a sufficient amount of deformation, and so the pulse time constants will become smaller. A sharp pulse is produced if the deformation reaches the engine or other large auxiliary equipment, so this can also be used as a characteristic parameter for that vehicle. The set of collision waveforms for tuning is first passed through this analyzer and the characteristic pulses of the time constants for that vehicle are investigated. The key pulses are determined based on these results and the calibration parameters for the subsequent-stage determination area are also determined.

6.8.5 *Collision determination*

The collision determination is performed in this block based on the sizes of the pulses, which are a characteristic parameter of that vehicle, and when the pulses appear. In general terms, the combination logic is assembled in accordance with each of the different types of collision: a frontal collision that strikes both side members, which are the main load-bearing structural members of the vehicle, an offset collision that strikes only one structural member, a diagonal collision that strikes structural members diagonally, a collision with a pole that strikes no structural members, and the like. Tuning can then be

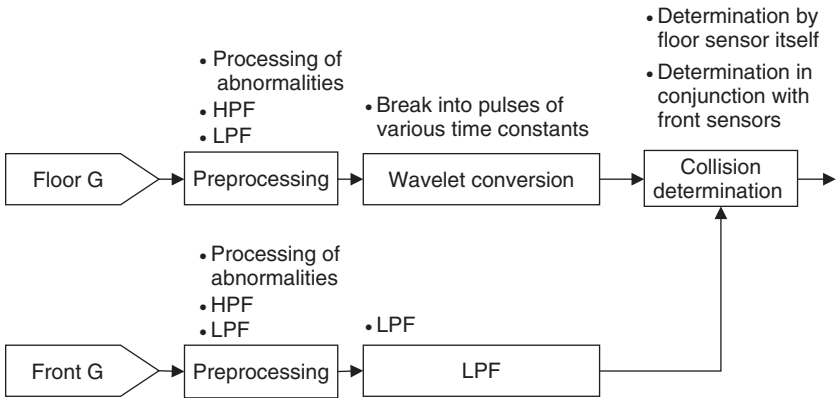


Figure 21. Block diagram of collision determination algorithm. (Reproduced by permission of Denso Corporation.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

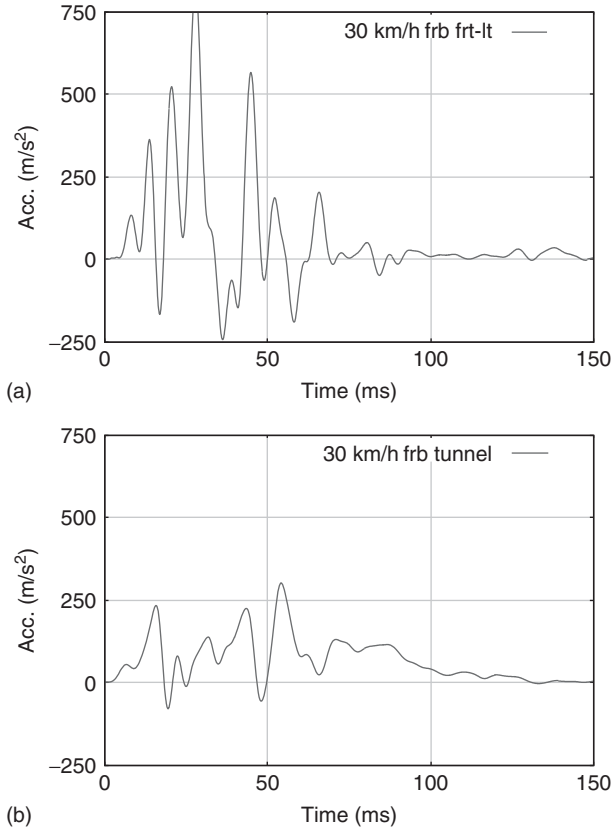


Figure 22. (a,b) Difference in waveforms at the front and floor. (Reproduced by permission of Denso Corporation.)

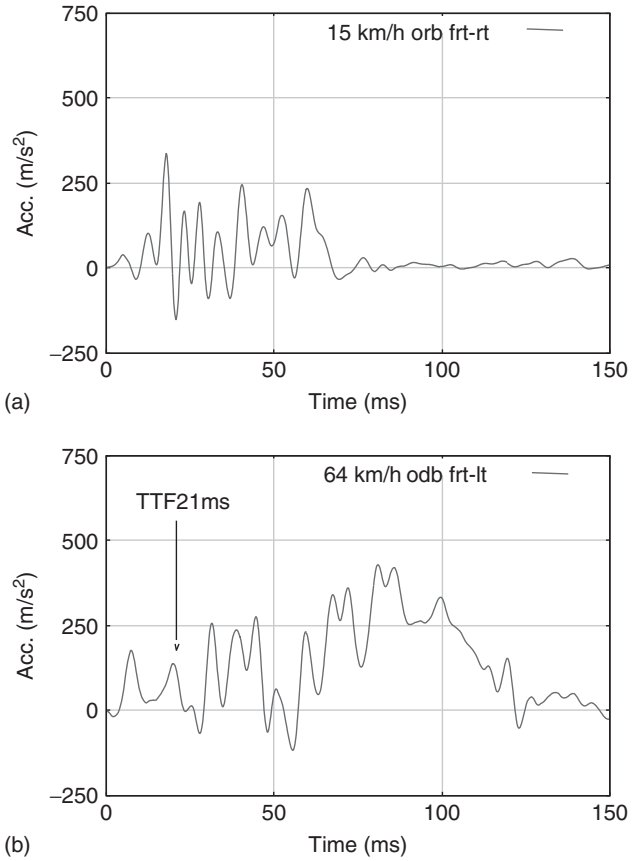


Figure 23. (a,b) Difficulty of determining collision with front sensors only. (Reproduced by permission of Denso Corporation.)

performed by simply setting the calibration parameters appropriately.

6.8.6 *Determination in conjunction with satellite sensors*

Figure 22 shows acceleration waveforms from the front satellite sensor and the floor ECU in the same frontal collision at 25 km/h. The graph shows that there is an extremely large acceleration produced in the crushable zone where the front satellite sensor is located because of the large deformation during the initial stage of the collision. In comparison, the increase in the acceleration is delayed in the floor ECU, which is located in the noncrushable zone. This collision has a must fire requirement, so it is necessary for the airbags to deploy.

Figure 23a shows the acceleration waveform from an offset frontal collision at 15 km/h using the same vehicle model as in Figure 22. Only one half of the vehicle receives the impact, so the acceleration in the area where the front satellite sensor is located is very intense. This collision has a no fire requirement, so it is not necessary for the airbags to deploy. Figure 23b shows the collision waveform of a 64 km/h ODB crash where the required TTF is 21 ms. The rise of the acceleration up to the required TTF is slow, even though the collision velocity is over 4 times faster. Consequently, the collision determination is performed by combining the waveform from the upper graph with the acceleration of the floor ECU.

The principle of an integrated collision determination is the same as for the individual determinations. The acceleration pulse produced at the time of the collision is detected. The size of the pulse and the time difference between the front and floor sensors are grasped as the characteristic parameters of the vehicle. These are then set as the calibration parameters. Figure 24 shows the acceleration waveform of an ODB collision at 64 km/h. The pulse produced at the front sensor is indicated by A, and this same pulse is then transmitted to the floor sensor, as indicated by B. Collision determination is then performed using this information.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

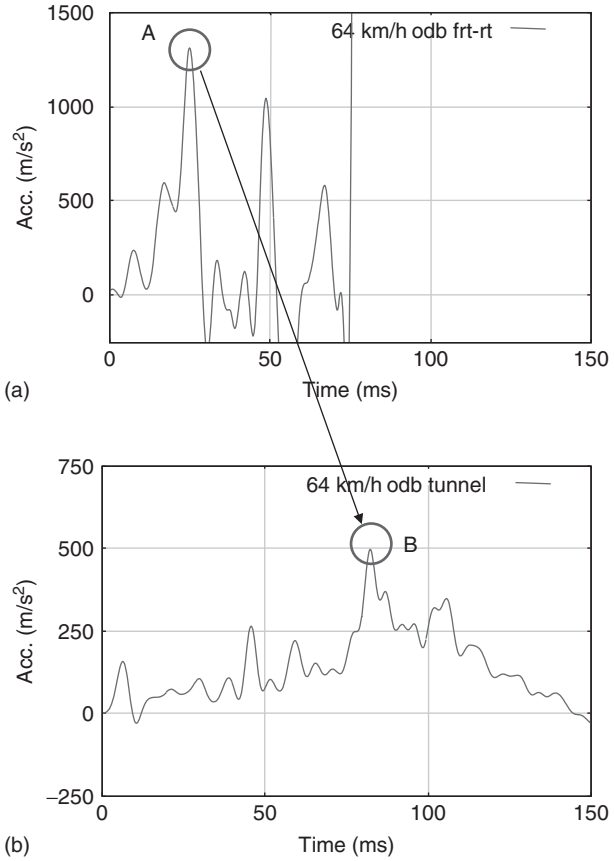


Figure 24. (a,b) Transmission of collision pulse. (Reproduced by permission of Denso Corporation.)

7 THE FUTURE OF AIRBAG

A collision phenomenon, sensing technologies, the sensors for airbags, ECU, and the collision determination algorithm were described as passive safety technologies of a vehicle collision. In recent years, other safety technologies have been placed on the market, such as crash compatibility technology (which protects the occupants of the other vehicle in the event of a collision with a compact car) and airbags and active hoods (which protect pedestrians). These kinds of technologies, which protect the other party in the collision, have been advancing and evolving. There are also other safety technologies such as fuel shut-offs, electric power shut-offs, and mayday systems in addition to those that protect vehicle occupants directly. Moreover passive safety technologies will be combined with driving assist technologies, active safety technologies, and precrash safety technologies. A safety system with few traffic accidents will be developed.

FURTHER READING

Kato, M. (2010a) *Automotive Electronics: Systems*, Nikkei Business Publications, Inc., Tokyo.

Kato, M. (2010b) *Automotive Electronics: Basic Technologies*, Nikkei Business Publications, Inc., Tokyo.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto227
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Chassis ECU (ACC and Sensor)

Takayuki Nagai[1] and Masayuki Furuhashi[2]

[1]*DENSO Corporation, Kariya, Japan*
[2]*DENSO Corporation, Inabe, Japan*

1 Introduction 1
2 ACC Operation 1
3 System Configuration 3
4 Future Development 6
References 6
Further Reading 6

1 INTRODUCTION

1.1 Adaptive cruise control (ACC)

Adaptive cruise control (ACC) (Figure 1) is an enhanced cruise control (CC) system with drastically improved functions. CC basically enables the vehicle to drive at a constant speed. It has become widespread, for example, in the United States where long-time driving on roads with a low traffic volume is the norm, as it saves the driver the trouble of constantly keeping the accelerator depressed. However, the adoption of CC is limited, in countries such as Japan where even highways often have heavy traffic, because of frequent resetting of CC after braking. Compared with CC, ACC is easier to use for the driver. It recognizes objects ahead of the vehicle and, if it judges that the vehicle is a relevant ACC target, the system controls the vehicle speed in accordance with the speed of the preceding vehicle. In addition, a new type of ACC system called *full speed range adaptive cruise control* (*FSRA*) has also been developed (Figure 2). It is capable of following the preceding vehicle in the entire speed range from start to stop.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto228
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 ACC OPERATION

2.1 ACC in driving assistance systems

Figure 3 shows a block diagram that illustrates the relationship among the driver, vehicle, and driving assistance systems. Figure 3a indicates a vehicle without assistance control. The driver operates the vehicle according to the surrounding conditions. Figure 3b shows a vehicle with chassis control. The control detects locking of the tires on a frozen road through the ABS (antilock brake system) ECU (electric control unit). In this case, it lowers the oil pressure of the brakes and helps the vehicle recover enough grip to enable the driver to perform avoidance operation through steering. The role of the chassis control is to maintain the vehicle state in accordance with the driver's intentions. Figure 3c adds driving assistance systems. Unlike the chassis control ECU, which is positioned in parallel with the vehicle, driving assistance systems are positioned in parallel with the driver. These systems provide the driver with information by processing the surrounding conditions. Specific examples of these systems include the navigation system, night vision, adaptive front-lighting system (AFS), and the like. Figure 3d incorporates vehicle speed and in-lane driving controls. ACC is included in this final type of systems.

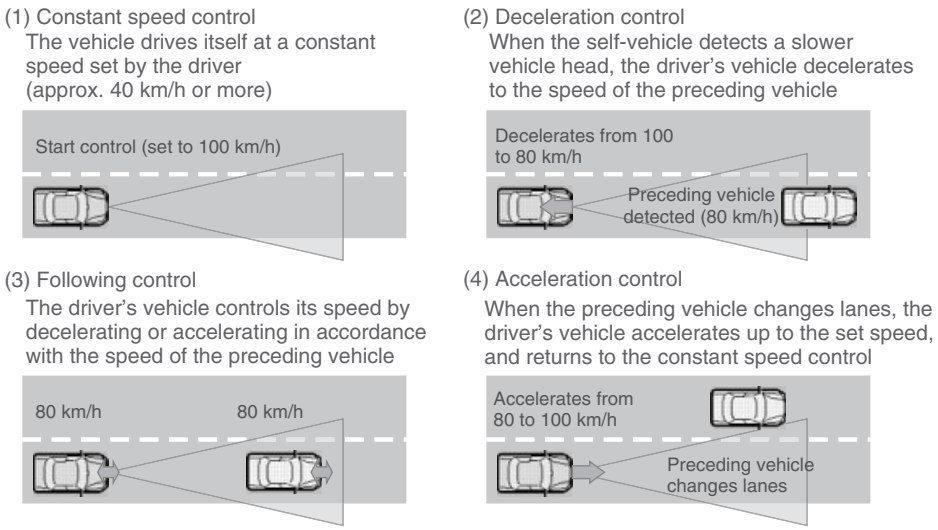


Figure 1. ACC control. (Reproduced with permission from Denso.)

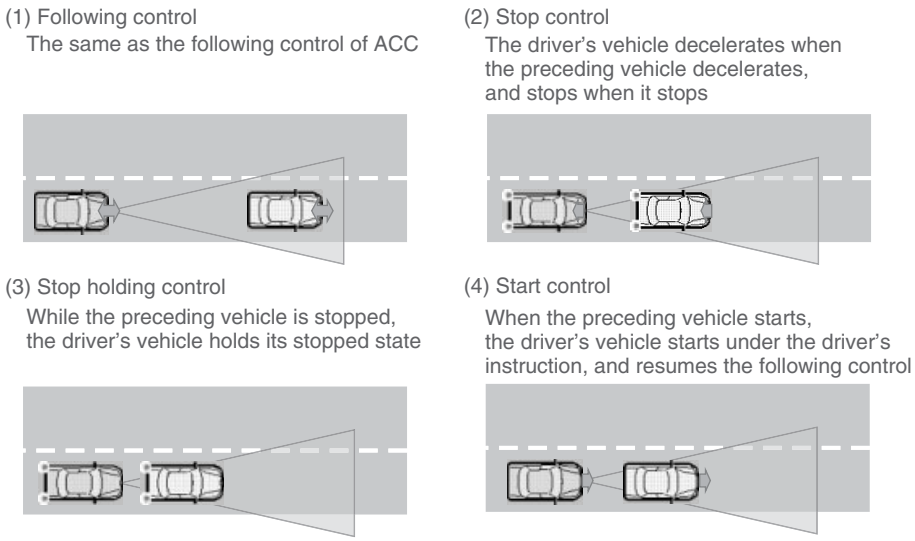


Figure 2. FSRA control. (Reproduced with permission from Denso.)

2.2 Distance control

Conventional CC achieves constant speed driving by controlling only the engine to hold the speed set by the driver. ACC has radar that functions as the eyes of the system. When the preceding vehicle is driving slower than the set speed, it can control the brake to decelerate the self-vehicle and follow the preceding vehicle at a certain following distance. When there is no longer a preceding vehicle, the system accelerates again and resumes driving at the set speed.

Furthermore, FSRA covers the entire speed range from start to stop. Even when the self-vehicle stops because the preceding vehicle has also slowed down and stopped, it resumes following the preceding vehicle when the driver turns on a switch or depresses the accelerator.

2.3 Drive override

The essential concept is that driver operation always takes precedence over ACC. When the driver depresses the brake pedal, ACC is cancelled, and deceleration is carried

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto228
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

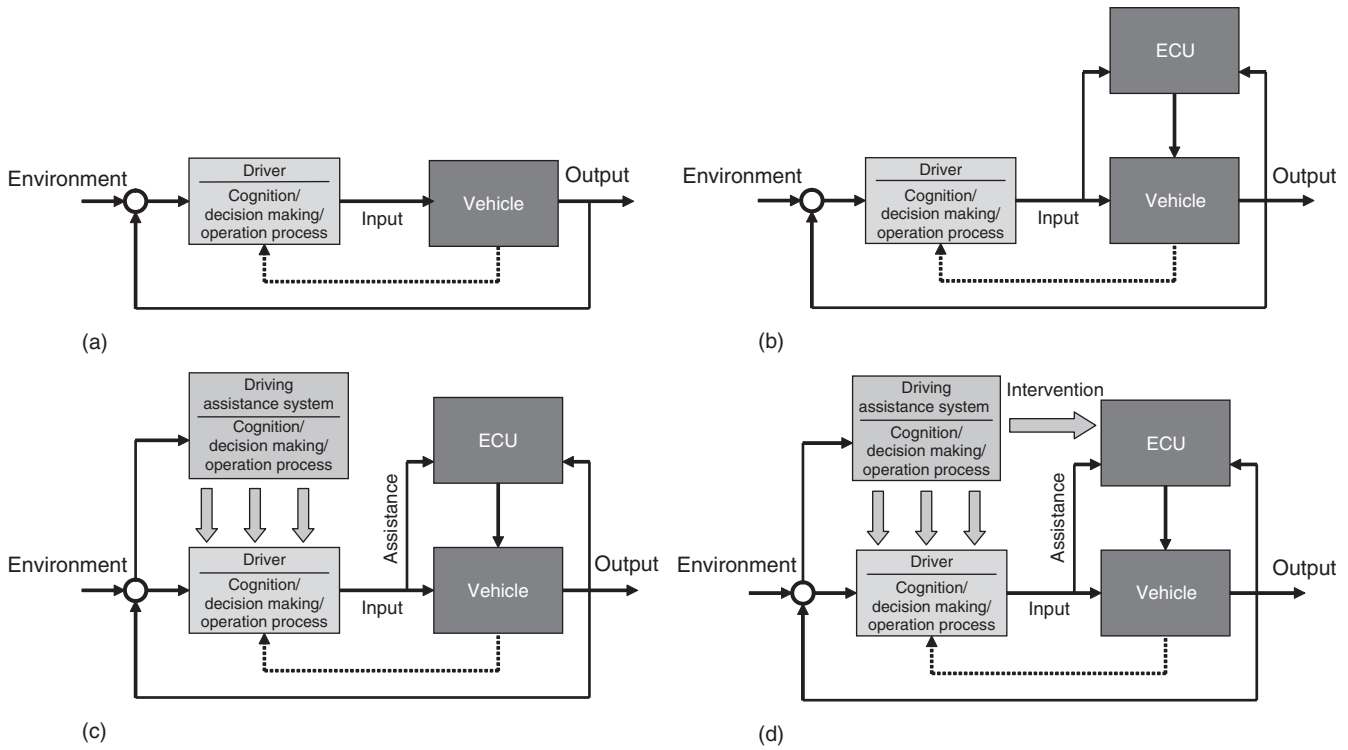


Figure 3. (a–d) Relationship among driver, vehicle, and driving assistance systems. (Reproduced with permission from Denso.)

out by normal braking based on the driver's intention. Furthermore, when the driver depresses the accelerator to increase the speed, the vehicle accelerates accordingly. When the driver releases the accelerator, the system returns to ACC mode.

3 SYSTEM CONFIGURATION

3.1 Overall configuration

The example in Figure 4 shows a configuration that incorporates AEBS (advanced emergency braking system) and LKAS (lane keeping assist system) in addition to ACC. ACC uses a millimeter-wave radar for detecting vehicles ahead. AEBS also uses the millimeter-wave radar for detecting potential collision objects. LKAS employs a video sensor for recognizing lane markings. The information collected by these sensors (such as the distance and the relative speed to the preceding vehicle, and the azimuth) are processed by driving assistance system ECU together with vehicle information from other ECUs to identify the target operative amount necessary for safe driving. ACC gives instructions to the engine and brake control ECUs to operate the engine and electronic stability control (ESC) actuator. The ACC operation state and alerts are shown to the driver in the display.

3.2 Range sensor (millimeter-wave radar)

Range sensors to perform ACC include radar, LIDAR (light detection and ranging), and stereo cameras. Figure 5 shows an example of a millimeter-wave radar. Automotive millimeter-wave radars employ various detection methods. The radar shown employs the frequency modulation-continuous wave (FM-CW) method to detect distance and relative speed and digital beam forming (DBF) to detect the azimuth. Figures 6 and 7 show the principle of each method (2004).

In Figure 6, the horizontal axes show the time, and the vertical axes show the frequency. The radar launches pulses modulated to form a triangle shape as shown in the figure. The time lag until it receives the reflective pulse is proportionate to the distance. Here, as it is difficult to measure the time lag accurately, the distance is obtained by calculating the frequency difference of $fs - fr$ instead. The bottom graph in Figure 6 shows a case when a relative speed is present between the self-vehicle and the preceding vehicle. In this case, the entire reflective pulse shifts because of the Doppler effect. Thus, the $fs - fr$ difference

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto228
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

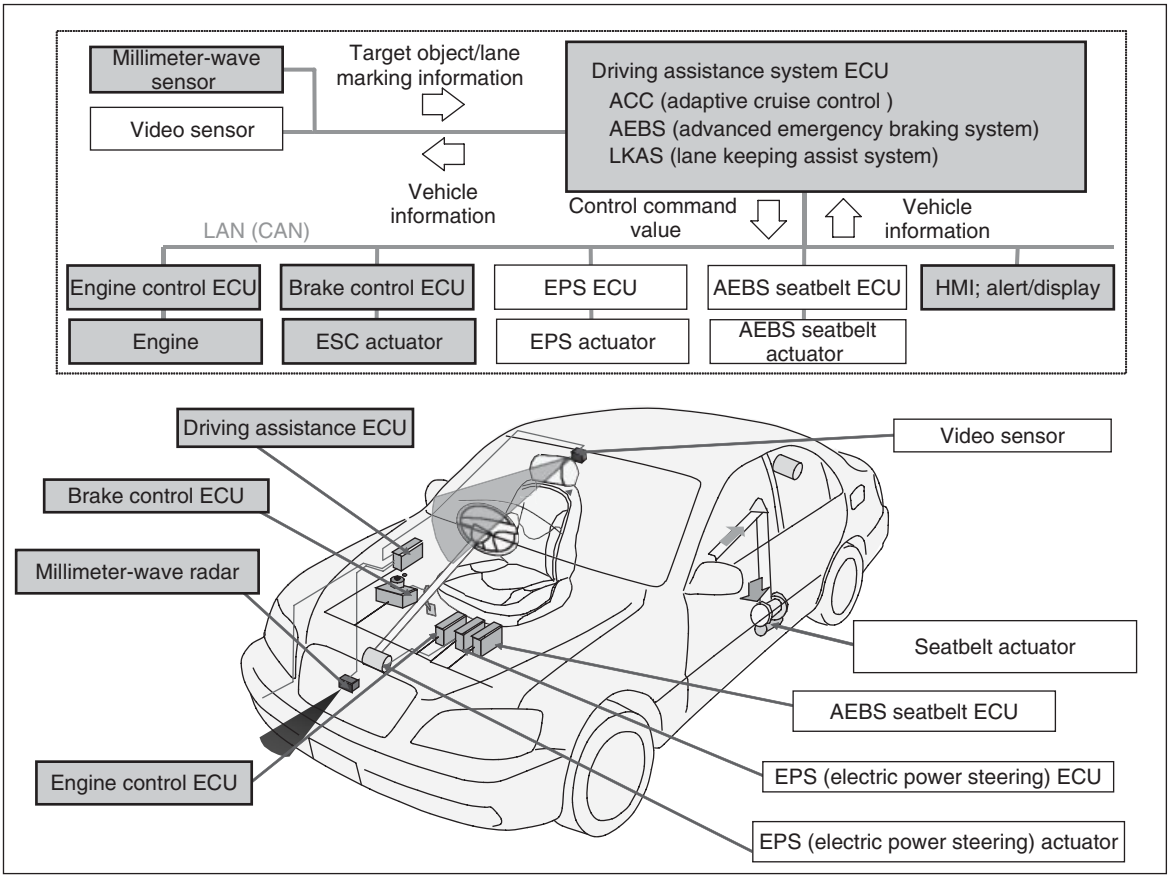


Figure 4. Example of driving assistance/safety system configuration. (Reproduced with permission from Denso.)



Figure 5. Millimeter-wave radar. (Reproduced with permission from Denso.)

becomes smaller by the amount of the Doppler effect at the rising portion, and the difference becomes larger at the descending portion. The relative speed can be obtained by subtracting the differences in the rising and the descending frequencies, and the distance by adding up the differences.

Figure 7 shows the azimuth detection principle. An outgoing pulse is transmitted from one channel antenna, and the reflective pulse is received by multiple antennas (five channels in this example). As the path length of the reflective waves that reach the channels varies depending on the azimuth it comes from, the azimuth is obtained by detecting the phase difference between the channels.

3.3 Controller (driving assistance system ECU)

ECUs used for driving assistance systems including ACC obtain information from sensors via the network and send requests to other ECUs to control the brakes, engine, and steering. Furthermore, they judge diagnosis information from sensors and actuators comprehensively and shift the system operation to the fail-safe mode as needed. The control system has a hybrid configuration that consists of a state transition model and a continuity model, manages ACC operation, and achieves safe and smooth transition between the normal driving and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto228
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

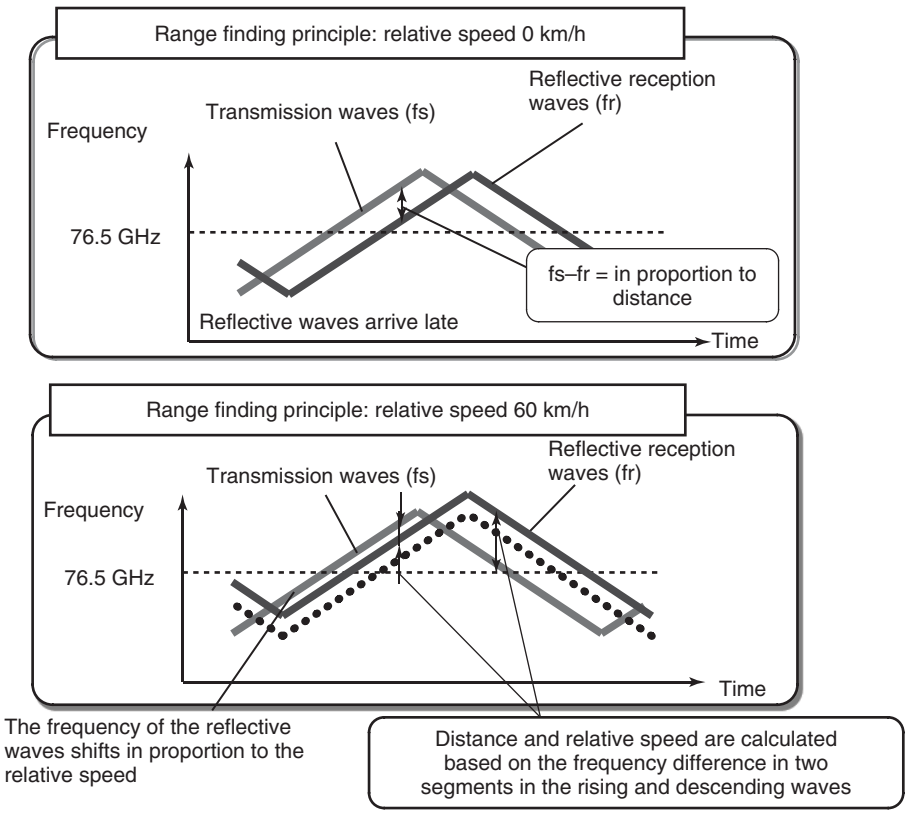


Figure 6. Range finding principle based on FM-CW. (Reproduced with permission from Denso.)

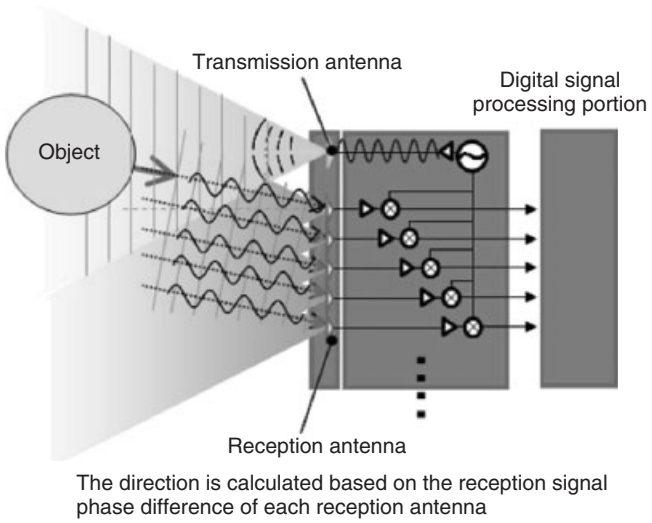


Figure 7. Direction measurement based on DBF. (Reproduced with permission from Denso.)

the ACC operation states in accordance with the driver operation (brake pedal, accelerator pedal, and input to ACC system). In addition, the driving assistance system ECU often includes applications such as AEBS and LKAS. When the value to execute each control is parallel and the control command value issued to the vehicle system from various applications differs, it arbitrates the control command value based on the priority among applications and outputs the command values to the engine/brake ECUs. It includes the operation to shift ACC mode to the AEBS mode if the frontal collision is inevitable because of the sudden deceleration of the preceding vehicle. System configuration is shown in Figure 8.

The CAN (controller area network) communication is a main communication approach employed nowadays. Adoption of high speed and large-volume communication technology such as FlexRay achieves a high level and high speed recognition and control.

Control logic development is based on models, which are fitted directly into the rapid prototyping ECU of the actual vehicle. The control constants directly impact the driver's comfort, and these are important calibration elements in addition to the calibration of the sensors and actuators. Automatic code generation method from the model is also utilized (Figure 9).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto228
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

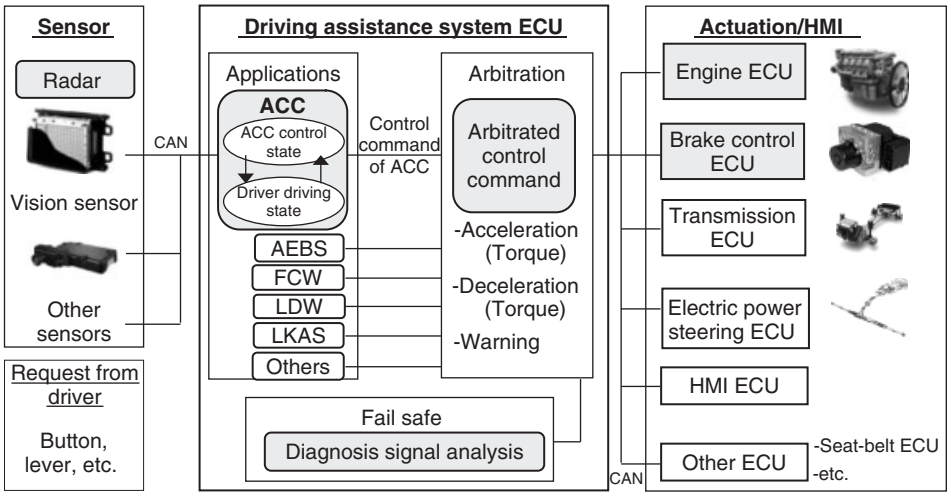


Figure 8. Roll and process flow in the driving assistance system. (Reproduced with permission from Denso.)



Figure 9. ACC ECU (driving assistance system ECU). (Reproduced with permission from Denso.)

4 FUTURE DEVELOPMENT

Although ACC presently achieves driving assistance through longitudinal speed and distance controls, it is likely to be fused with lateral vehicle control in the future. Integrated longitudinal and lateral vehicle control will require more sophisticated recognition and decision-making process in the system by utilizing obstacle and maneuverable space identification, road profile detection, digital mapping, and vehicle-to-infrastructure and/or vehicle-to-vehicle communications as a maximum set of sensing elements. As the recognition confidence level increases, autonomously drivable scenario will expand, including autonomous lane change as one of the challenging maneuvers. The technology also would be usable to realize fuel-efficient driving assistance, route guidance assist, and convoy driving.

The technical evolution as described earlier will require the driving assistance system ECU to deal with much larger inputs and outputs. Calculation amount inside ECU will also increase to handle comprehensive surrounding recognition, increasing applications, and busier communication with other ECUs.

REFERENCES

Hiroshi, M., Noriyuki, T., Atsushi, K., and Tomoya, K. (2004) A forward-looking sensing millimeter-wave radar. *Denso technical review*, **9** (2).

FURTHER READING

Kato, M. (2010) *Automotive Electronics: Systems*, Nikkei Business Publications, Inc.(In Japanese), Tokyo.

Tokoro, S. (2003) Electronically scanned millimeter wave radar for pre-crash safety and adaptive cruise control system, *IEEE International conference in IV*.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto228
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Body and Lighting ECU (Key-less Entry, Sonar, HID, LED Usage for Lamps)

Masafumi Yamaura
DENSO Corporation, Kariya, Japan

1 Outline of Body Controls 1
2 Body Control ECUs 1
3 Headlamp Control 4
Further Reading 7

1 OUTLINE OF BODY CONTROLS

The purpose of body controls is to improve aspects of vehicle performance, such as comfort, usability, safety, theft prevention, serviceability, and the like. This is accomplished by controlling the output of motors, lamps, and the like, based on inputs mainly from switches and sensors. Table 1 lists some examples of products and functions that aim to improve these aspects of performance. Many of these body control products are designed to enhance the perceived value of the vehicle and to create an appealing impression to the user. For this reason, these controls were conventionally installed mainly in luxury vehicles. In recent years, however, technological innovations have reduced costs to the level where more and more have become available in mass-market vehicles.

2 BODY CONTROL ECUS

Each vehicle contains multiple body control engine control units (ECUs). This section describes the integrated body ECU, clearance sonar, and headlamp control (functions such as the air conditioning, cluster, remote door locking, and the like are covered elsewhere).

2.1 Integrated body ECU

This refers to an ECU that incorporates multiple body system ECUs for functions that originally existed independently. The in-built functions of this integrated ECU differ depending on the vehicle, based on the installed functions and electronic systems in the vehicle as a whole. One factor in the emergence of integrated ECUs is the dramatic increase in wire harnesses for electronically connecting ECUs with inputs (switches and sensors), outputs (motors and lamps), and other ECUs as the scope of body control functions has spread. The integration of ECUs is one method of minimizing the space required for installation and costs. The feasibility of integration is generally determined based on the commonality of inputs and outputs (i.e., whether the same inputs and outputs are connected), the similarity of control (body functions often operate with the ignition turned off), the functions adopted on the vehicle (i.e., what functions are standard or optional), and the division of responsibilities with other ECUs. Since the 1980s, the number of wire harnesses has been reduced even further by the adoption of multiplex communication, which is capable of transmitting data over a single wire harness in a time-sharing format. This has also led to the development of new functions using data obtained through multiplex communication. Figure 1 illustrates the merits of integrating body system ECUs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto229
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Table 1. Examples of body control products and functions.

Performance aspect	Main products and functions
Comfort and usability	Automatic air conditioning, electronic cluster, (remote) door locking, electronic key systems, power windows, power seats, sliding roofs, power sliding doors, power back door, power luggage compartment opening, power fuel lid opening, automatic lights, interior illumination, electrically retracting and adjustable outer mirrors, electrical sun shades
Safety	Automatic wipers, clearance sonar, tire pressure monitoring systems, HID, AFS
Theft-prevention	Vehicle theft-prevention systems, remote security systems, user customization
Serviceability	Fault diagnostics, user customization

Some products and functions cover multiple performance aspects.

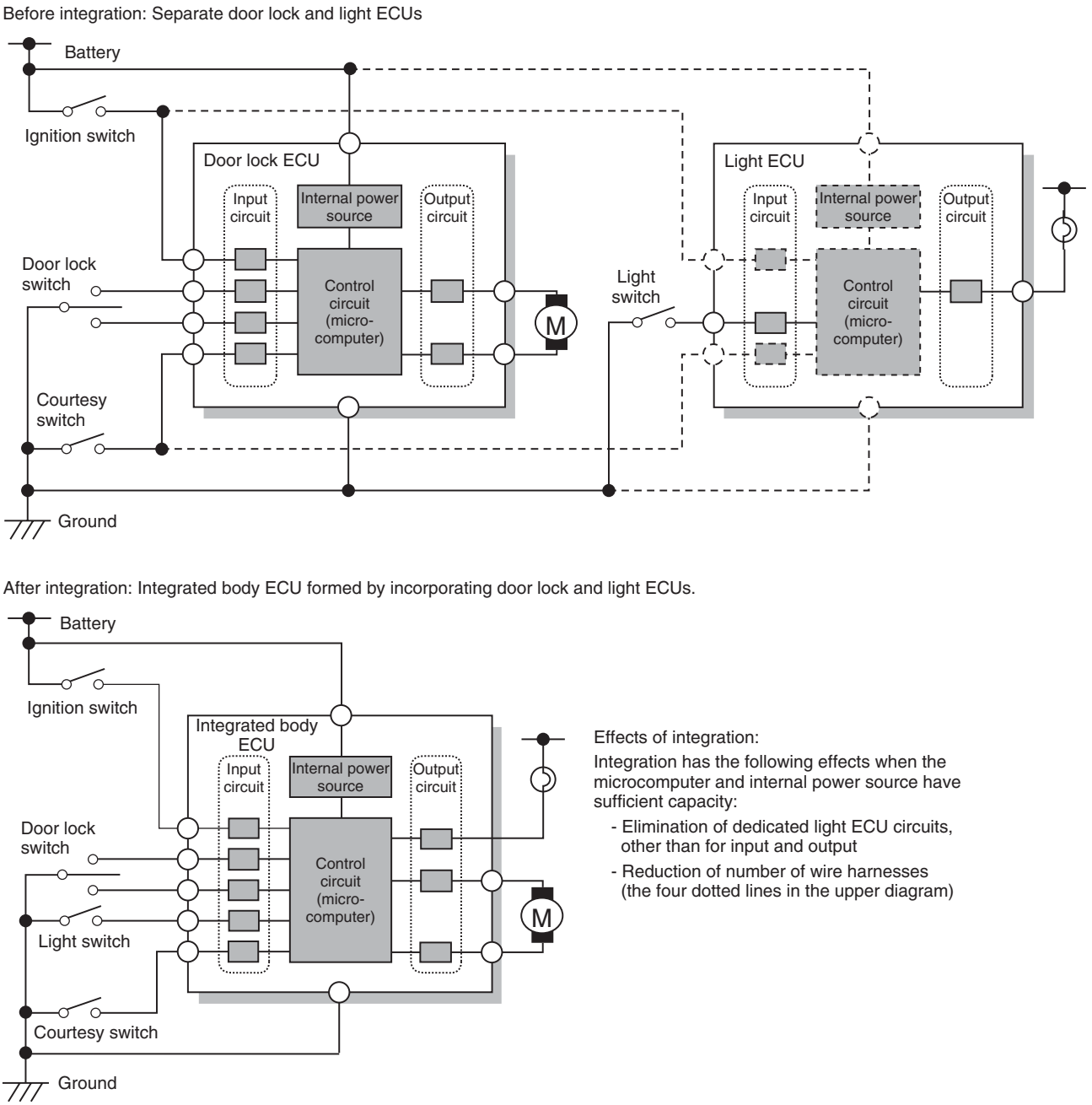


Figure 1. Merits of integrated body ECUs.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto229
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.2 Clearance sonar

This is a system that detects objects close to the vehicle and notifies the driver of their location and distance via a buzzer or display. It operates when the vehicle is being driven at low speeds such as during parking maneuvers, using ultrasound sensors with an integrated reception/transmission function built into the front and rear bumpers.

2.2.1 Example system configuration

Figure 2 shows the configuration and detection range of a typical system with eight sensors. In the system shown in the figure, the vehicle information required to control the sonar (i.e., the vehicle speed and whether it is moving forward or backward) is obtained through the controller area network (CAN) or another type of vehicle local area network (LAN). This information is used to control the sensors. Detected objects are notified to the driver by a buzzer and displayed on the cluster in front of the driver by the LAN. The detection range of the system is 50 to 60 cm from the sensors at the vehicle corners and 100 to 150 cm in front and behind the vehicle. The buzzer tone changes from an intermittent to a continuous sound as the vehicle approaches the object to help the driver identify the distance aurally.

2.2.2 Principle of object detection

Figure 3 shows the principle of object detection. Ultrasonic waves are transmitted from the sensors, which are reflected back by objects. The return time t of the ultrasonic waves is measured to calculate the distance to the object D. Furthermore, the control of each ultrasonic sensor changes the transmission timing and interval of the sensors dynamically (i.e., randomly) to prevent false detection due to ultrasonic waves from the same system or from other sources.

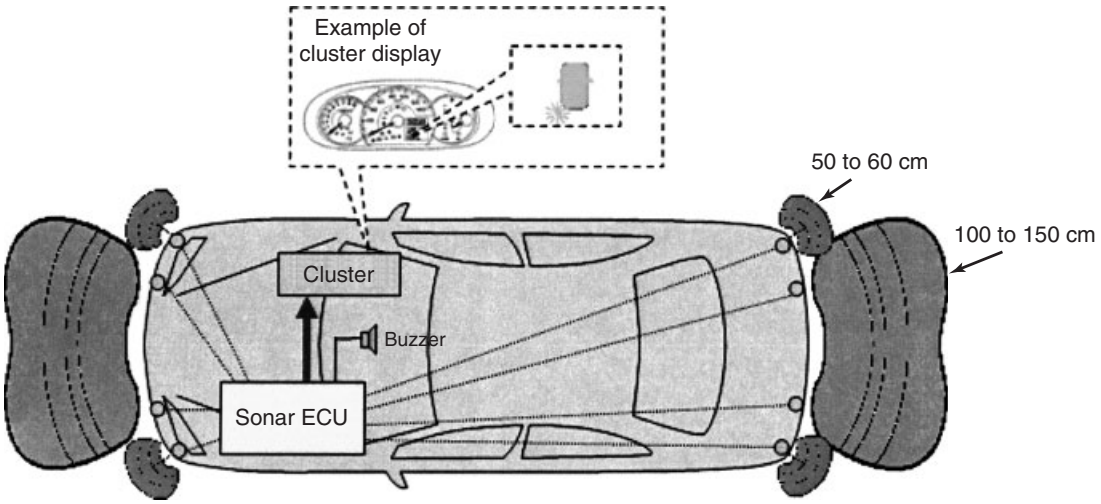
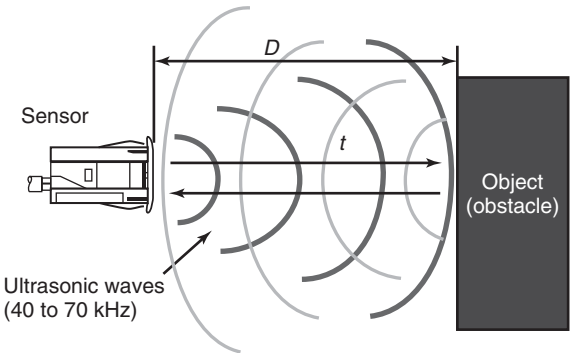


Figure 2. Example configuration and detection range of sonar system.



Method of calculating distance D between sensor and object:
- Transmission rate of ultrasonic waves $v \approx 331.5 + 0.6\,T$
(T : temperature)
- t = Time for ultrasonic waves to return from object
- Distance to object $D = v\,t / 2$

Figure 3. Principle of object detection.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto229
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

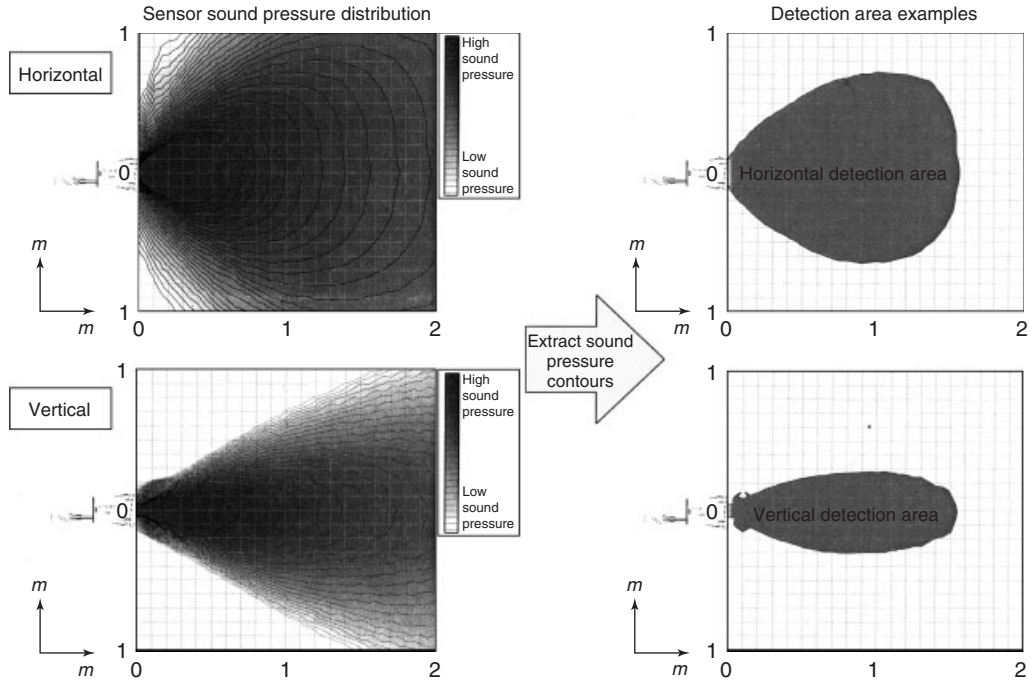


Figure 4. Detection area control.

2.2.3 *Detection area control*

As Figure 4 shows, the detection area can be divided by sound pressure contours for the ultrasonic waves. These are formed by applying voltage transformation to the sound pressure reflected from the object within the sensor. Actual sensor control is performed so that certain items are not detected. These include small objects and tire-stops on the floor, which should not be considered as collision risks, as well as side walls when reversing into a garage.

2.2.4 *Ultrasonic sensors*

Figure 5 shows the structure and appearance of an ultrasonic sensor. It consists of an ultrasonic microphone portion that transmits and receives ultrasonic waves, which is composed of a piezoelectric device (lead zirconate titanate (PZT)) and an aluminum case, a voltage boosting circuit that drives the piezoelectric device, and an ultrasonic wave reception circuit. The circuit inside the circuit uses a dedicated IC to reduce size. As the sensor is installed inside the bumpers, it has a waterproof structure that integrates the circuit with the connector using a filler material. The surface of the microphone is painted the same color as the bumper to improve its appearance. The materials used are selected to prevent damage from stone chipping during driving and the structure is designed to quickly damp fluctuations in the ultrasonic waves.

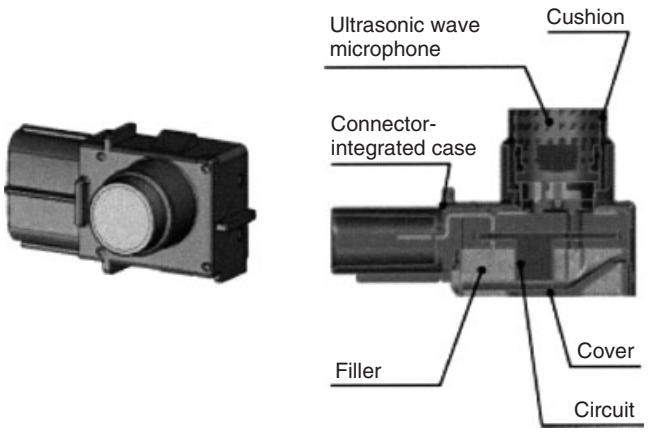


Figure 5. Ultrasonic sensor.

3 HEADLAMP CONTROL

3.1 Evolution of headlamps

Since the 1960s, automotive headlamps have used halogen lamps as a light source. In the 1990s, brighter high intensity discharge (HID) lamps were introduced and entered widespread use. Sophisticated adaptive front-lighting systems (AFSs) were developed in the 2000s to enhance forward visibility at intersections and on curves at nighttime by optimizing the distribution of light in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto229
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Evolution of headlamps.

Date	1960	2000	2005	2010	2015
Social trends	**Reduction of glare to oncoming vehicle**		**Environmental protection**	**Increasing demand for safety and usability**	
Headlamp evolution				**LED headlamps**	(○ Introduction of AFS for LED headlamps)
		HID headlamps			
		• Auto-leveling	• Mercury-free HID • Introduction of AFS for HID headlamps		(○ Introduction of energy-savin AFS)
	Halogen headlamps				

accordance with the driving conditions of the vehicle (such as the vehicle speed, steering angle, and the like). Furthermore, LED headlamps, which were introduced in 2007, have the merits of increasing bulb life, saving energy, and achieving stable brightness levels immediately after being switched on. LED headlamps are also more compact, which increases the design freedom of the headlamps and surrounding areas. As improving safety and usability are constant priorities, it is likely that further improvements will be made in the future to light sources (in terms of brightness and energy saving) and visibility enhancement control. This section describes the controls for HID and AFS (Table 2).

3.2 HID headlamps

3.2.1 Outline of HID headlamps

According to statistics, the rate of fatal traffic accidents is roughly three times higher at night. Poor visibility is recognized as one factor contributing to this issue. This led to the introduction of discharge bulbs as light sources. Discharge headlamps emit white light similar to the sun and are three times brighter than conventional halogen headlamps, while consuming 30% less power.

3.2.2 Configuration and control of HID headlamps

An extremely high voltage of approximately 20,000 V is required for discharge bulbs to operate. The power must also be controlled to maintain brightness at a constant level. These controls are performed by an electronic circuit called the *ballast*, which is located outside the headlamp housing. The ballast consists of a DC–DC converter that boosts the battery voltage, full bridge inverters that generate square waves for stable illumination of the bulbs, and a control circuit that controls these items. It also includes a fail-safe circuit, which shuts off operation of the headlamps when an abnormality occurs, and a starter circuit that generates

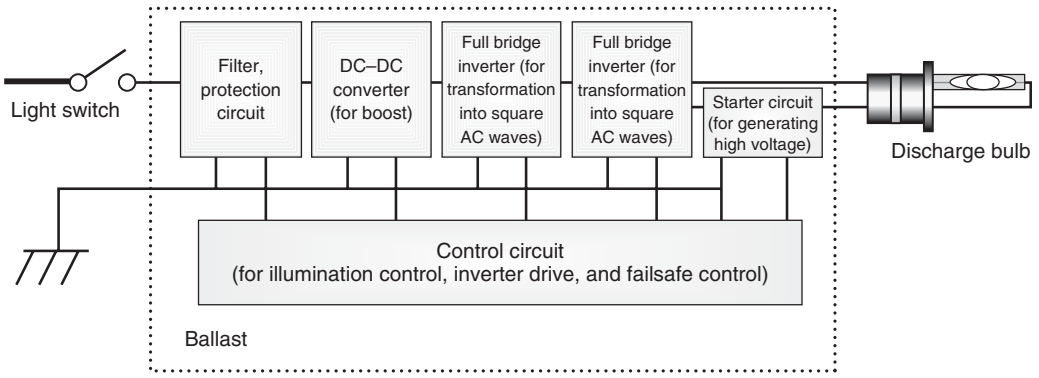


Figure 6. Configuration of ballast.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto229
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

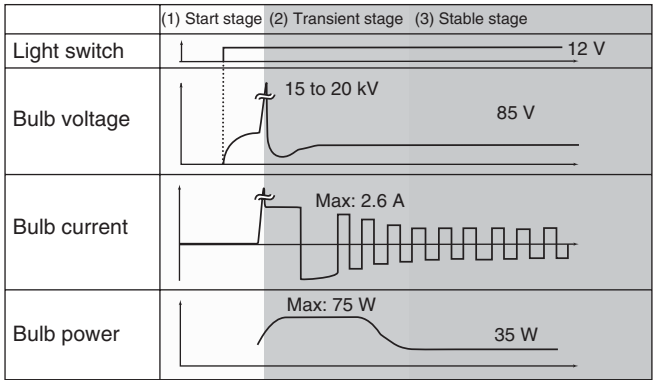


Figure 7. Ballast control operation.

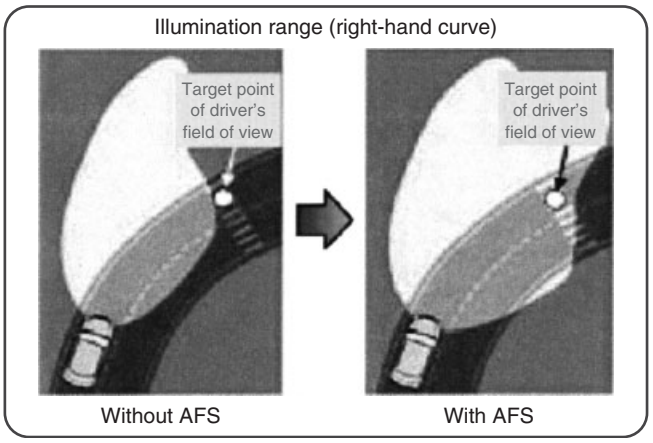


Figure 8. Effect of AFS.

a high voltage pulse to break down the insulation between the bulb electrodes when the headlamps are switched on (Figure 6).

Ballast control operation can be categorized into the following three stages when the discharge bulb is switched on (Figure 7).

3.2.2.1 Start stage. The control applied a high voltage pulse between the electrodes of the discharge bulb to break down the insulation. Although the required voltage changes depending on the bulb state, it is generally in the range of 15,000 to 20,000 V.

3.2.2.2 Transient stage. This is the stage between insulation breakdown and stable illumination of the bulb. This stage requires an extremely large amount of power to increase the brightness in a short period of time. However, as applying unrestricted power to the bulb will dramatically reduce the lifetime of the bulb, the power must be controlled within a maximum bulb output power and current. Subsequently, the control reduces the power in accordance with the bulb impedance.

3.2.2.3 Stable stage. In this stage, the voltage and power of the bulb are controlled to stable values of 85 V and 35 W, respectively.

Consequently, ballast design requires control technology capable of handling large voltages and power levels.

3.3 Adaptive front-lighting system

3.3.1 Outline of AFS

AFS is a headlamp system that automatically changes the light distribution pattern in accordance with the conditions after detecting the driving environment (i.e., whether the vehicle is being driven around a curve, through city streets, at high speed, or while it is raining) using sensors installed in the vehicle. In a curve, headlamps with conventional fixed optical axes will illuminate a different location to the intended vehicle travel direction (i.e., the outside of the curve). In contrast, AFS allows the headlamps to illuminate the intended vehicle travel direction (i.e., the inside of the curve). In addition, AFS reduces the glare of the headlamps visible by oncoming vehicles and pedestrians. Therefore, AFS plays an important role in danger avoidance by making approaching vehicles clearer (Figure 8).

3.3.2 AFS configuration and control

AFS is based on a conventional auto-leveling system, which consists of height control and vehicle speed sensors. Additional components in AFS include a steering sensor to detect the angle of the steering wheel and headlamp

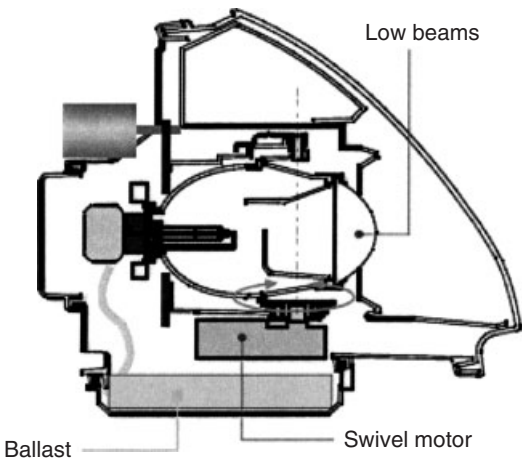


Figure 9. AFS lamp structure.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto229
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

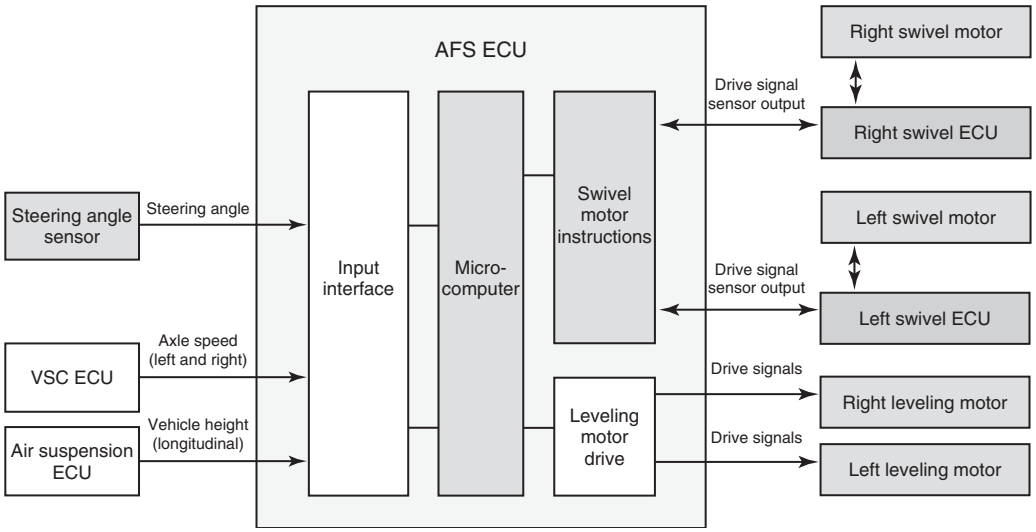


Figure 10. AFS configuration.

swivel actuators that rotate the headlamps in the horizontal direction. The AFS ECU determines the control (swivel) angle of the low beam headlamps based on the information from each sensor (i.e., the vehicle speed and steering angle) and operates the swivel motors in the headlamps. In addition to the swivel motor control, the AFS ECU has an auto-leveling control function that moves the headlamps up and down in accordance with the longitudinal angle of the vehicle. These two controls are combined to optimize the headlamp light distribution for all driving conditions. Figure 9 shows the AFS lamp structure and Figure 10 shows the AFS configuration.

FURTHER READING

Kato, M. (2010) *Automotive Electronics: Systems*, Nikkei Business Publications, Inc., Tokyo.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto229
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Telecommunications

Hiroaki Kuraoka
DENSO Corporation, Kariya, Japan

1 Outline of Communication 1
2 Bluetooth 1
3 Dedicated Short-Range Communication 3
4 Vehicle-Infrastructure Cooperative Systems 6
5 Mobile Telephone Networks 8
6 Conclusion 15
Further Reading 15

1 OUTLINE OF COMMUNICATION

In addition to the fundamental driving, turning, and stopping functions, modern vehicles are also capable of utilizing various services through wireless-based communication with objects outside the vehicle. Although this is used to simply refer to radio reception, it now encompasses radio waves from the Global Positioning System (GPS), analog and digital television broadcasts, and external wireless communication with electronic toll collection (ETC) systems, mobile telephone networks, and the like. It has even come to include hands-free use of telephones inside the vehicle through Bluetooth.

There are three categories of wireless communication used in vehicles. The first category is used inside the vehicle (near field). A typical example of this technology is Bluetooth. The second category is short-range wireless communication, which is used for ETC and other functions that are operated from fixed locations. The third is broadband wireless communication, which is used for mobile telephone networks. This chapter describes the characteristics of each category of wireless communication technology, as well as examples of the available services (Figure 1).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 BLUETOOTH

2.1 What is Bluetooth?

Bluetooth is a wireless communication technology for mobile devices. The main companies behind this technology include Ericsson from Sweden, IBM from the United States, and Toshiba from Japan. As Bluetooth is a form of short-range communication that uses radio waves in the 2.4 GHz band, it does not require licensing or registration to use. Bluetooth is used for relatively low speed communication, such as for personal computers and mobile telephones. Its characteristics are as follows.

1. It divides the 2.4 GHz band (ISM band) into 79 frequency channels and hops between use frequencies at random.
2. Low power consumption (during reception: 10 mA, during standby: several μA).
3. Compact size (active chip area: max. 1 cm^2).
4. Low cost (chip price: max. 5 US dollars).
5. Joint transmission of audio and data (calls can be made with up to three different people).
6. Security assured by authentication on connection and data encryption.
7. Profiles can be established for each usage scenario (Table 1).

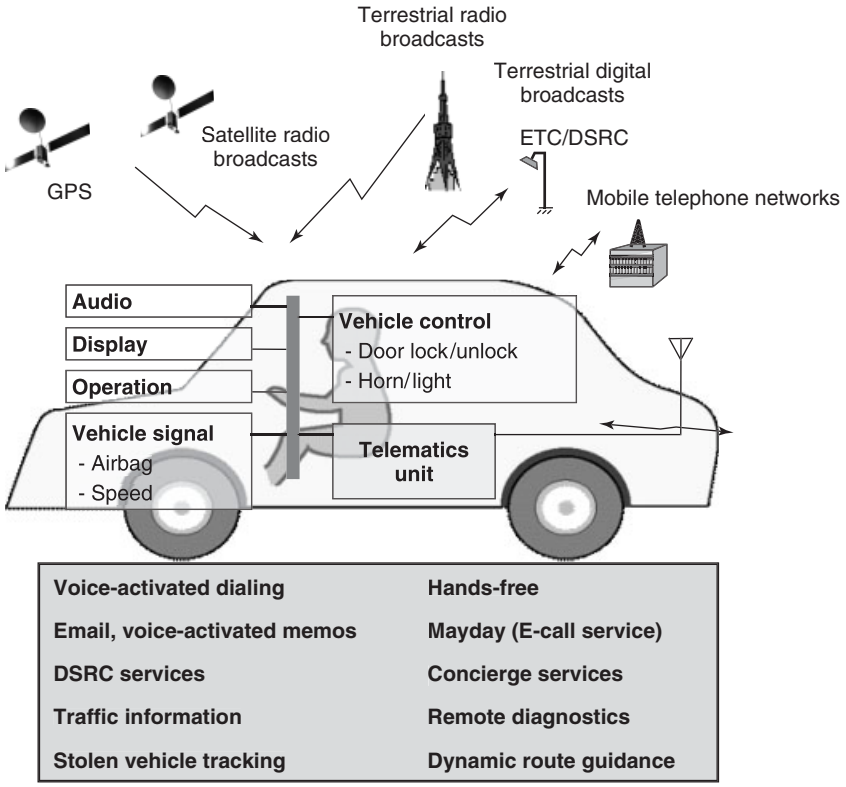


Figure 1. External communication functions.

Table 1. Bluetooth specifications.

Item		Specifications	
Wireless	Wireless band used	Industrial, scientific, and medical (ISM) bands (2.400 GHz to 2.4835 GHz)	
	Transmission distance	Class 1: Approx. 100 m (100 mW) Class 2: Approx. 25 m (2.5 mW) Class 3: Approx. 10 m (1 mW)	
	Bit rates	1, 2, 3 Mbps[a]	
	Primary modulation protocols	Gaussian frequency shift keying (GFSK) (frequency modulation protocol), differential quadrature phase shift keying (π/4DQPSK), differential phase shift keying (8DPSK)[a]	
	Secondary modulation (expansion) protocol	Frequency hopping spectrum expansion protocol 79ch ($f_{(k)} = 2402 + k$ [MHz], $k = 0, 1, 2, \cdots 78$)	
	Frequency hopping rate	1600 hops/s (625 μs)	
Transmission speed	Synchronous mode (synchronous connection-oriented (SCO) link)	Voice: 64 kbps	
	Asynchronous mode (asynchronous connectionless link (ACL))	Symmetrical	433.9 kbps
		Asymmetrical	Maximum download: 723.2 kbps, upload 57.6 kbps

[a]Protocol with Bluetooth 2.0 + enhanced data rate (EDR).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

2.2 Services achieved using Bluetooth

Figure 2 shows an outline of the services accomplished using Bluetooth. In addition to the typical example of hands-free communication, Bluetooth can be used to forward and play music data from iPods or other mobile music players on the vehicle's audio system, to perform wireless communication between mobile telephones for data communication, to forward an address book from a mobile telephone to the vehicle's navigation system, and to forward e-mails from a mobile telephone to an on-board device.

Protocols are determined for each service a communication device can use. These protocols are called *profiles*. If two devices are compatible with the same profile, then that service can be used. Examples of profiles include the hands-free profile (HPF), dial-up network (DUN) for data communication, the object push profile (OPP) for data exchange, the basic imaging profile (BIP) for forwarding still images, and so on. Figure 3 shows an example of a device configuration for hands-free telephone calls. This consists of a mobile telephone and vehicle navigation system with Bluetooth compatibility, and switches on the steering wheel.

First, initialization and profile confirmation (profile check) is performed between the mobile telephone and the vehicle navigation system. Once pairing is complete, the driver presses a button on the steering wheel to speak into the telephone. The telephone call is carried out through the microphone and speakers in the vehicle. The driver can also use the system to respond to calls from outside the vehicle.

3 DEDICATED SHORT-RANGE COMMUNICATION

3.1 What is DSRC?

Dedicated short-range communication (DSRC) is a short-range wireless technology that uses 5.8 GHz band radio waves. It is used in ETC systems, which are nonstop automatic toll collection systems for highways recommended by the Japanese Ministry of Land, Infrastructure, Transport and Tourism (MLIT). DSRC is also used for the vehicle information and communication system (VICS), which collects and sends out real-time traffic information on highways, and various on-board intelligent transport system (ITS) devices that expand the functions of ETC.

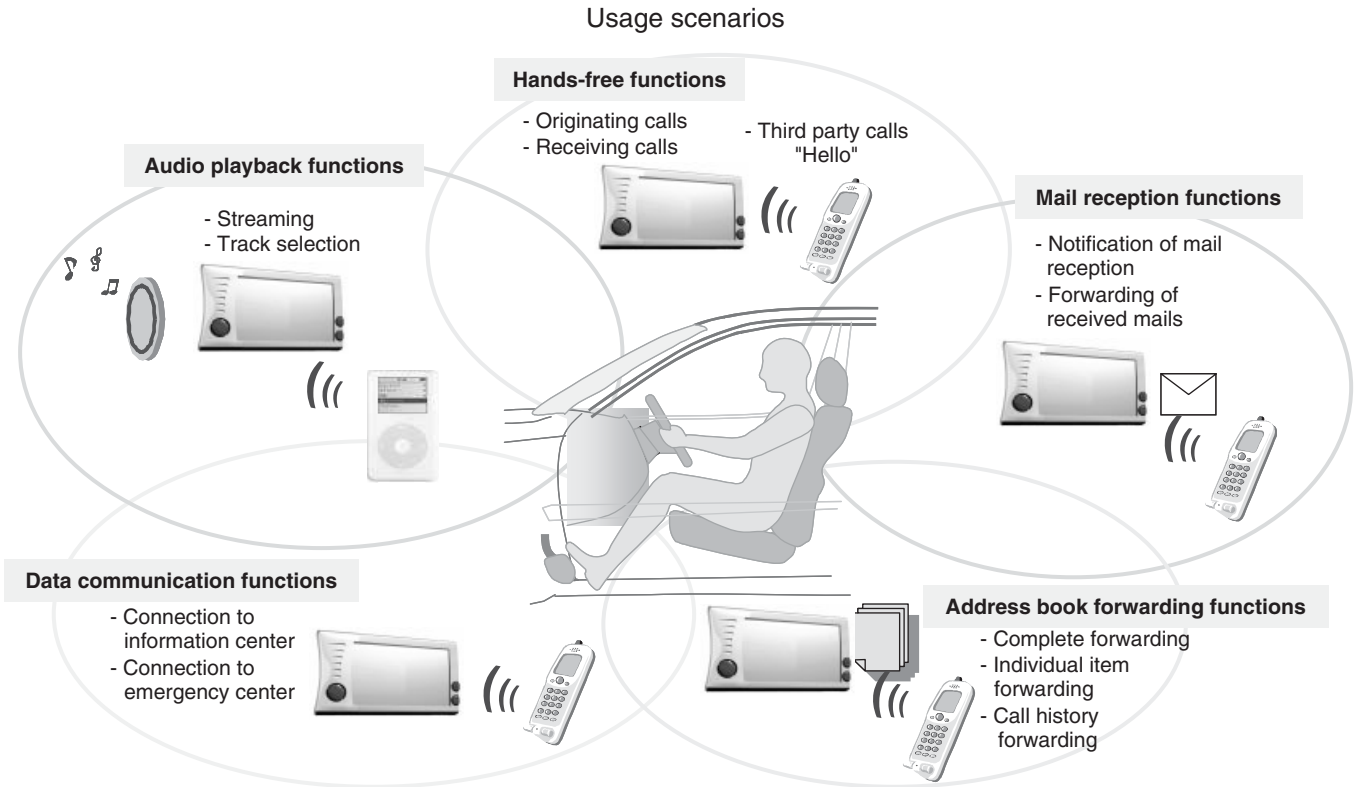


Figure 2. Bluetooth services.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

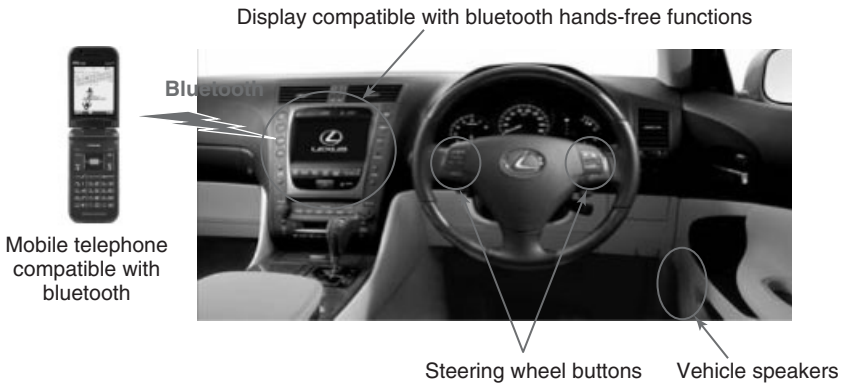


Figure 3. Device configuration for hands-free telephone calls.

3.2 Outline of ETC

ETC uses antennas located at highway toll booths in combination with an on-board device and wireless communication to enable the automatic payment of tolls without the driver having to stop the vehicle (Figure 4).

One of the merits of ETC is the alleviation of congestion. In Japan, most vehicles are now equipped with ETC devices. In the past, 35% of congestion occurred at toll booths, but as a result of ETC, major congestion no longer occurs at the on/off ramps of highways (Figure 5). The vehicle capacity of tool booths has increased dramatically as the drivers do not have to stop to use ETC. Capacity has jumped from 230 vehicles/hour to 800 vehicles/hour.

Another merit is the fact that vehicles no longer have to stop or accelerate away from toll booths after paying. As a result, the noise and emissions levels around toll booths have decreased, and fuel efficiency has been improved. Cashless payment is more convenient and it has enabled the introduction of various discount systems for commuters and travelers at night. ETC systems have spread to other countries outside Japan. It was introduced in Italy and the United States in 1990, and Singapore in around 2000. Many other countries are also thinking about introducing the system (Figure 6).

3.3 ETC technology

3.3.1 Active protocol

ETC in Japan has adopted an active protocol that uses 5.8 GHz band radio waves. With this protocol, the on-board

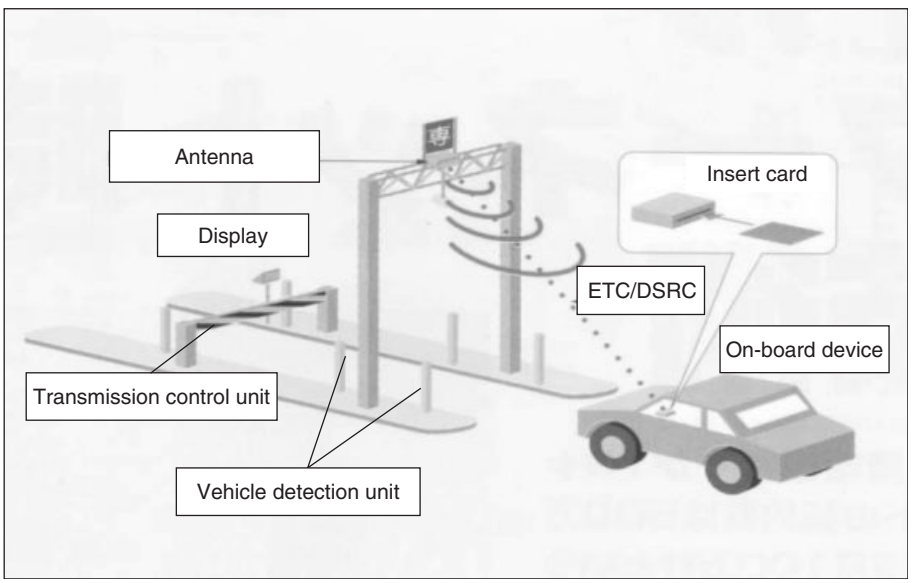


Figure 4. ETC system configuration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

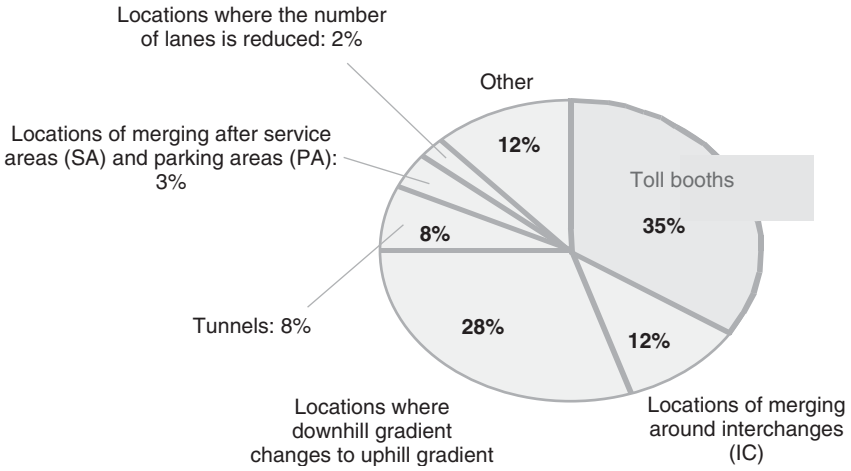


Figure 5. Congestion locations before ETC introduction.

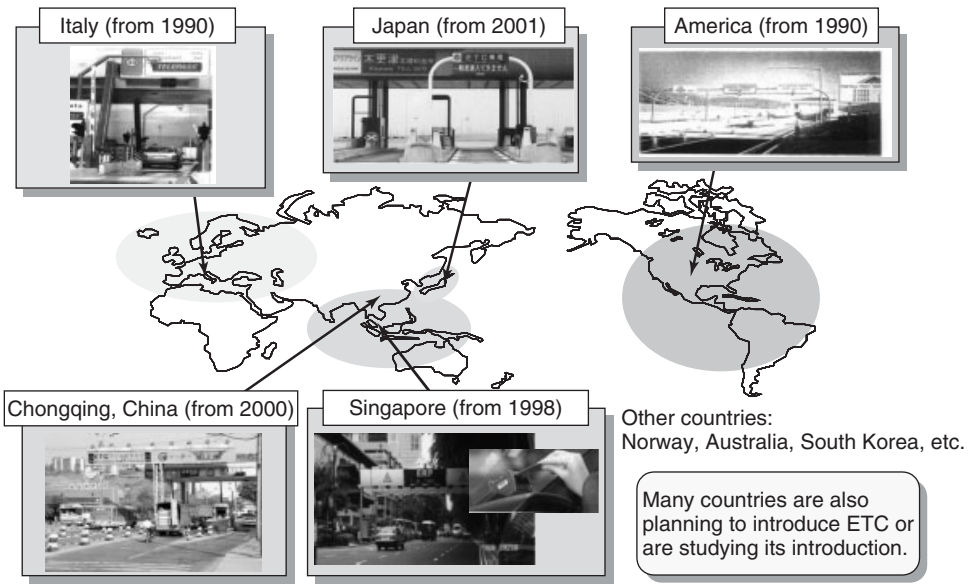


Figure 6. Examples of countries that have introduced ETC.

device has a power supply and is capable of transmitting its own radio waves. This has the following merits (Figure 7).

1. Strong reception and high noise resistance
2. High speed and large-capacity communication (1024 kbps)
3. Wide communication area (max. 30 m)
4. Capable of simultaneous communication with multiple vehicles (max. eight vehicles)
5. Future expandability (i.e., the potential of use with applications other than ETC)

In contrast, some ETC systems outside Japan use a passive protocol. This protocol reflects transmitted radio waves and performs data communication by changing the rate of reflection.

3.3.2 Security

ETC in Japan uses card authorization similar to that of a credit card. Improper use of the card is prevented by providing an integrated circuit (IC) chip in the card. The roles of the on-board device and IC card are separated. The on-board device memorizes only its device ID and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

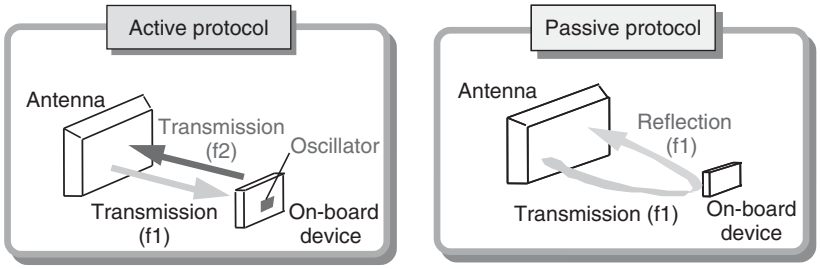


Figure 7. ETC communication protocols.

the vehicle information. The IC card stores the personal information and use history of the user. Communication between the on-board device and roadside antenna is also encrypted to help secure a high level of security (Figure 8).

3.4 Systems compatible with ITS spot service

ITS Spot Service is used by systems that carry out two-way communication between roadside and on-board devices over short distances and small zones. The application of ETC in vehicles is not only capable of paying tolls but it can also be expanded to the reception of VICS information, the payment of parking fees in car parks or the like, the utilization of services that provide information about tourist attractions, and so on (Figure 9). Some ITS on-board devices compatible with ITS Spot Service that have a selection of these functions are already on the market.

3.5 ITS spot service technology

Table 2 shows the specifications of an ITS Spot Service-compatible system. Unlike the ETC system, it carries out modulation by quadrature phase shift keying (QPSK). This achieves a fast communication speed of 4 Mbps. The characteristics of an ITS Spot Service-compatible system are as follows.

1. Use of 5.8 GHz-wave radio waves
2. Signal speed: 1 Mbps and 4 Mbps
3. Modulation protocol: amplitude shift keying (ASK) or QPSK
4. Communication over short distances
5. Large capacity and two-way mobile communication within a small zone mobile communication in small zones

4 VEHICLE-INFRASTRUCTURE COOPERATIVE SYSTEMS

These are new types of systems that aim to enhance safety through infrastructure-to-vehicle or vehicle-to-vehicle communication using 700 MHz or 5.9 GHz band radio waves. These systems use the IEEE 802.11p communication protocol, which is a standard for wireless LAN communication, to allow communication between multiple vehicles.

The requirements of this technology are described below. Vehicle-to-vehicle and infrastructure-to-vehicle communications have been defined, and various services are currently being examined. The services being considered for use with vehicle-to-infrastructure cooperative systems are shown in Figure 10.

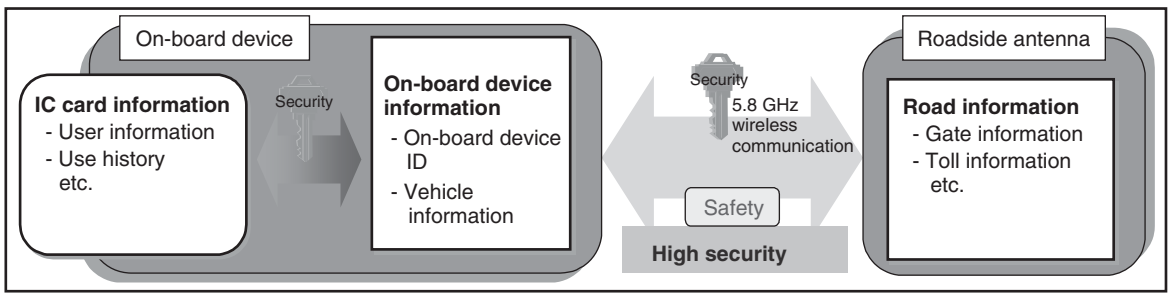


Figure 8. Security of ETC.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

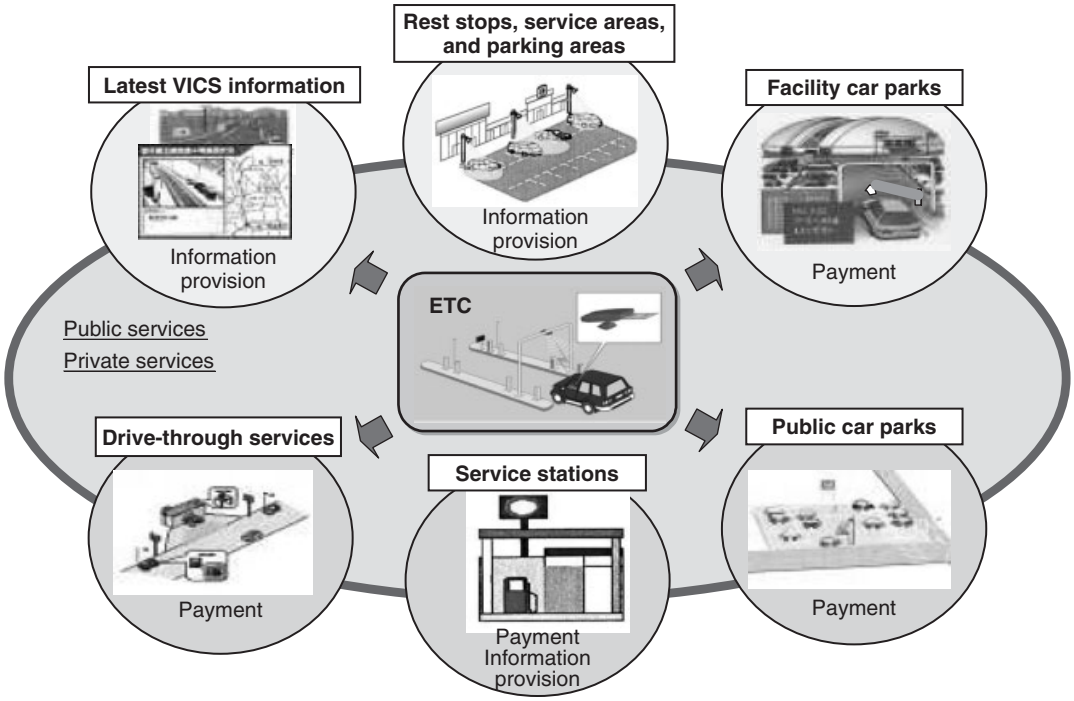


Figure 9. Example ITS Spot Service applications.

Table 2. ITS Spot Service-compatible system specifications.

		ITS Spot Service communication systems	Automatic payment systems for toll roads (ETC)
Wireless frequency band		14-wave 5.8 GHz	4-wave 5.8 GHz
Modulation protocol		ASK or QPSK	ASK frequency modulation Information to be transmitted $D(t)$ 1 0 1 t ASK modulation waveform t
Single transmission speed	ASK modulation protocol	1024 kbps	1024 kbps
	QPSK modulation protocol	4096 kbps	—
Transmitter power output	Base station	Max 300 mW	
	Land mobile station	Max 10 mW	
Permitted occupied bandwidth		Max 4.4 MHz	Max 8 MHz
Antenna gain	Base station	Max 20 dBi	
	Land mobile station	Max 10 dBi	

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

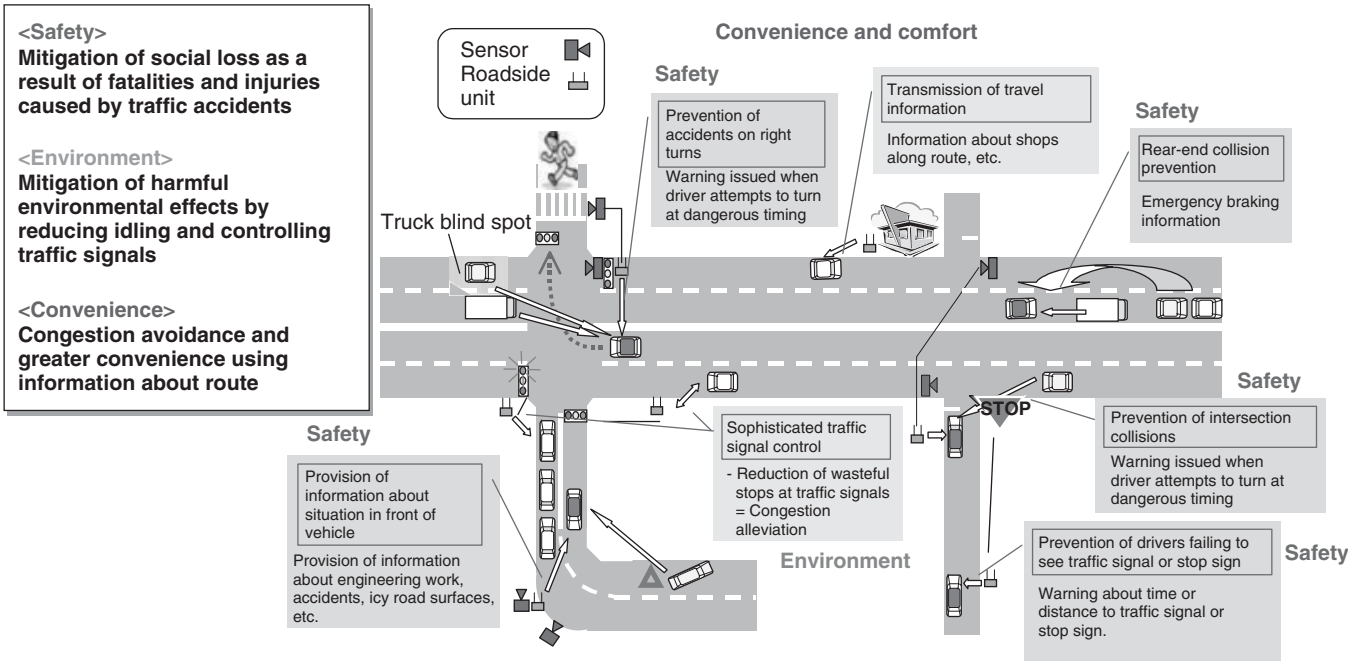


Figure 10. Vehicle-infrastructure cooperative systems.

1. Systems to help prevent intersection collisions
2. Systems to help prevent drivers failing to see a traffic signal or stop sign
3. Systems to help prevent rear-end collisions
4. Systems to help prevent collisions when turning right (particularly in countries such as Japan that drive on the left of the road)
5. Systems that provide information about objects in front of the vehicle

As the radio wave environment changes depending on the country, there are differences in frequencies, wavebands, and so on (Table 3).

5 MOBILE TELEPHONE NETWORKS

5.1 Evolution of broadband wireless

With broadband wireless, the mobile telephone wireless communication network can also be used for data communication. Therefore, the vehicle is installed with a communication module compatible with mobile telephone networks also for communicating with objects outside the vehicle. This communication module can be used to obtain real-time traffic information for the vehicle navigation system, tourist information around a destination, hands-free telephone calls, and so on.

First-generation mobile telephones were bulky and used analog communication. These telephones were developed exclusively as audio-based car phones. Digitalization was achieved in the second generation, and advances in semiconductor technology led to successive reductions in size and weight. Second-generation mobile telephones were also capable of low speed data communication. Third-generation devices can be used for applications requiring high speed data communication, such as music downloading and online gaming (Figure 11). The growing popularity of smart phones and increasing network speeds (currently in generation 3.9) are encouraging the construction of high speed infrastructure systems.

However, the rapidly increasing number of smart phones is consuming large amounts of wireless resources, giving rise to concerns about communication problems. Owing to the depletion of radio wave resources, the next trend in broadband wireless systems may be toward micro cells.

5.2 Communication protocols for mobile telephones

Although vehicles are used throughout the world, telematics services that use mobile telephone networks depend on the wireless environment in each region. Therefore, wireless devices for each region have to be prepared before a service can be established.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Comparison of radio wave media for vehicle-infrastructure cooperative systems.

Communication system / Characteristics		700 MHz	5.9 GHz	5.8 GHz
		WAVE (RC006)	WAVE (802.11p)	DSRC (T-75)
		Japan	Europe and the U.S.	Japan
Communication control		Autonomous distribution (asynchronous access, base station not required) Many-to-many communication (connection authentication not required)		Base station control (synchronous access, base station required) One-to-one communication (connection authentication required)
	Number of terminals allowed	Approx. 100 vehicles		7 vehicles
	Possibility of infrastructure-road or vehicle-vehicle communication	- Infrastructure-road: Possible - Vehicle-vehicle: Possible		- Infrastructure-road: Possible depending on the application - Vehicle-vehicle: Not possible
Radio wave wraparound performance		Good wraparound performance	Poor wraparound performance	Poor wraparound performance
Communication distance		Several hundred meters (100 mW): Medium band	Several hundred meters (100 mW): Medium band	30 m (10 mW): Short range
Modulation protocol		Orthogonal frequency division multiplexing (OFDM)		ASK/QPSK
	Usage environment	- Compatible for use on streets containing many reflective surfaces - Compatible with vehicles travelling at high speed		-Poor compatibility for use on urban streets - Not compatible with vehicles travelling at high speed

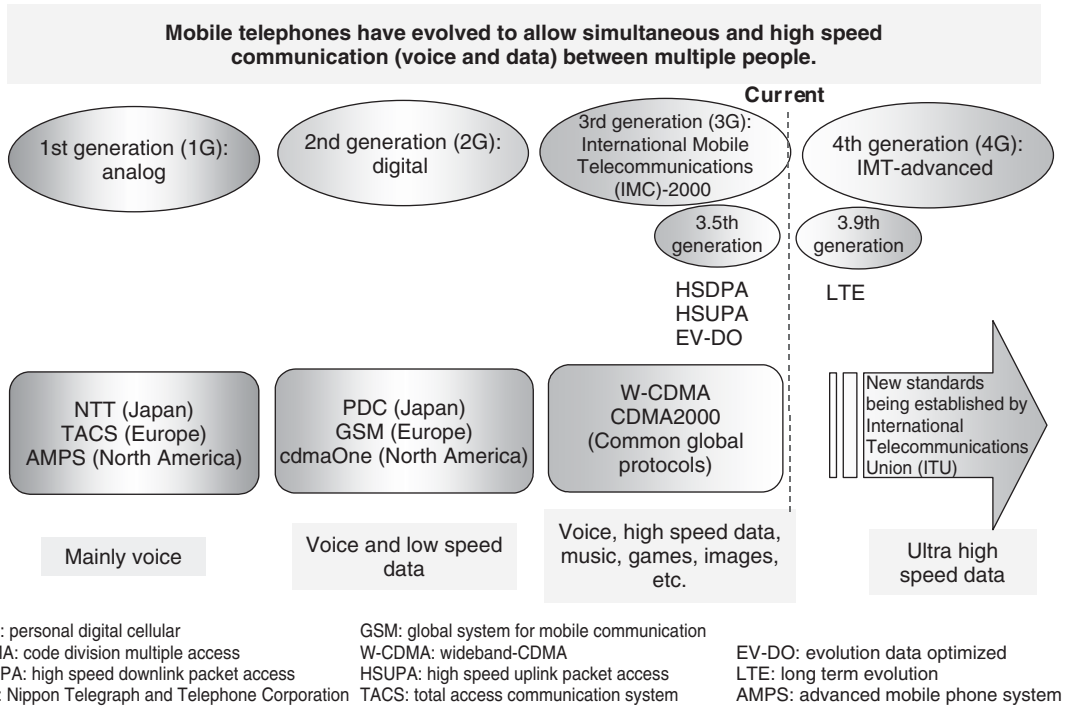


Figure 11. Evolution of mobile telephones.

As shown in Figure 12, the mainstream communication protocol around the world is the second-generation Global System for Mobile Communications (GSMs). The third-generation CDMA 2000 protocol is also prevalent in Japan and the United States. Japan uses the W-CDMA (high speed packet access, HSPA) and CDMA 2000 protocols.

Telematics services in Japan frequently use mobile telephones in place of modems. Dial-up connections using Bluetooth are also common. Some dedicated on-board devices have been developed. These include Toyota's G-Book service, which uses a CDMA 2000 communication module, and Honda's Internavi Premium Club, which uses

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

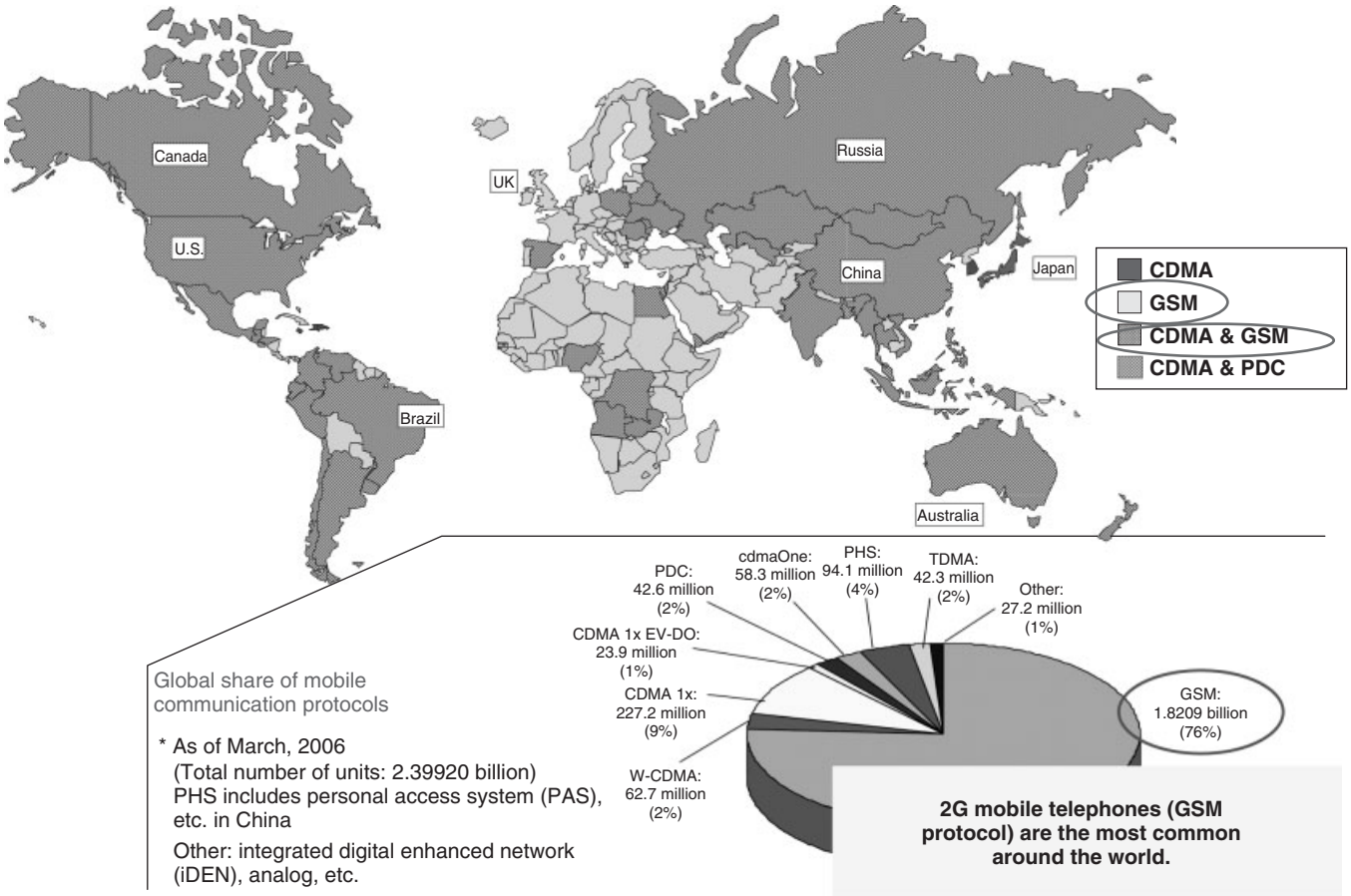


Figure 12. Mobile telephone communication protocols.

a PHS communication module. In North America, General Motors (GM) has developed the OnSTAR service, which also uses a CDMA 2000 communication module.

In China, the government has ordered the reorganization of the communication industry and the introduction of third-generation mobile telephones is progressing. China Mobile Ltd. uses TD-SCDMA, which was developed independently in China. The state-owned telecommunications operators China Telecom and Unicom are making progress in rolling out the HSPA and CDMA 2000 protocols.

In contrast, GSM (GPRS) remains the mainstream protocol in Europe. The 2.5th-generation GSM protocol with an upward compatible EDGE (384 kbps) was partially adopted, but HSPA third-generation systems are now being introduced.

5.3 Wireless frequencies

The frequency bands used by mobile telephones are determined by law. Devices used inside the vehicle must also conform to the Radio Law. Similar laws are in place in each country.

Different communication formats are used in each country, such as GSM, CDMA 2000, W-CDMA, HSPA, and LTE (see subsequent text for a definition of these terms). Communication must also conform to the technical standard determined by each communication company.

5.4 Evolution of communication protocols

5.4.1 Frequency division multiple access

First-generation mobile telephones used frequency division multiple access (FDMA). This protocol divides and allots the wireless channel frequency bands between multiple participants. A guard band was set between the channels to prevent interference between users (Figure 13).

5.4.2 Time division multiple access

Time division multiple access (TDMA) divides and allots the frequency bands on the same wireless channel

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

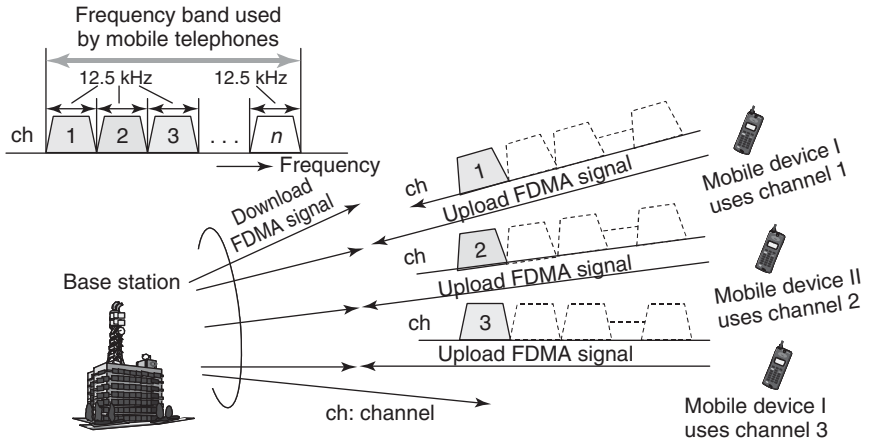


Figure 13. FDMA protocol.

between multiple participants on a time-basis. A guard time is provided to prevent signals from users overlapping (Figure 14).

5.4.3 Code division multiple access

Code division multiple access (CDMA) divides and allots the frequency bands on the same wireless channel between multiple participants using a spread code. Transmission data is modulated by the transmitting system. Transmission occurs after the data expanded by the spread code to the broadband region. The received broadband signal is then returned to the short-range region and demodulated by a reverse spread code (Figure 15).

The CDMA protocol is superior for the following reasons.

1. Large capacity: Data capacity can be increased by utilizing the same frequency in the neighboring channel.
2. Broadband: The high speed spread code enables broadband transmission for high speed multimedia services.
3. High quality: CDMA is resistant against interference or obstruction from existing systems. It is also less likely to interfere with or obstruct these systems itself. CDMA can also be used with rake receivers to prevent transmission quality deterioration due to multi-path propagation.
4. Mobile communication processing function: CDMA is capable of soft handovers with little line disconnection.

5.4.4 3.5th-generation technology

When communication technology is updated from the third to the 3.5th generation, audio data that does not require high

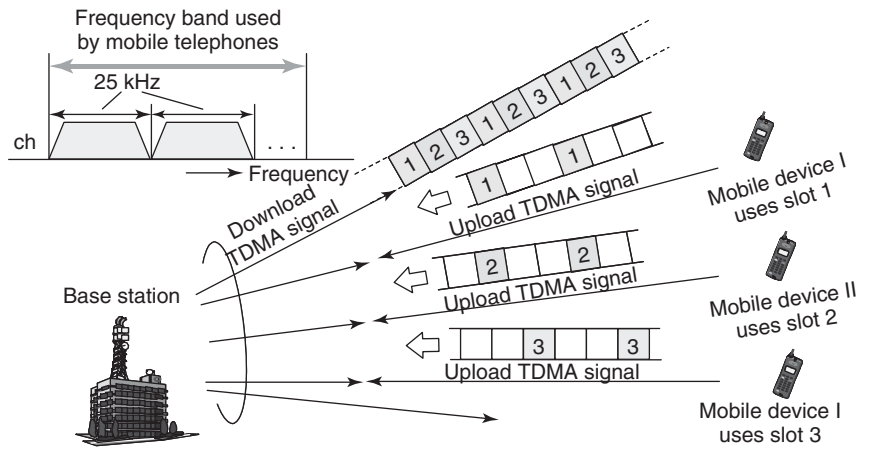


Figure 14. TDMA protocol.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

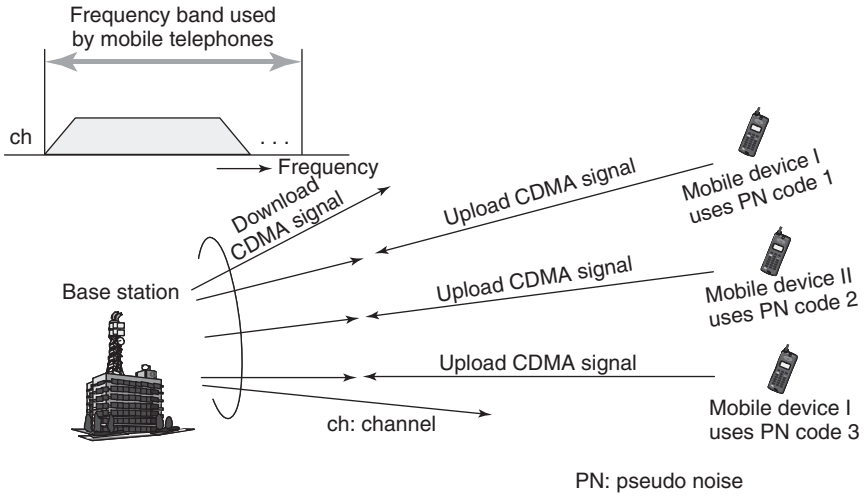


Figure 15. CDMA protocol.

speed communication will be transmitted at approximately 8 kbps. A high speed data communication function will be prepared separately from the audio channel to boost speed by changing the modulation protocol when required by the wireless communication state (Figure 16).

5.4.5 3.9th-generation communication

This technology is referred to as *long-term evolution* (*LTE*), and is capable of achieving communication speeds of 100 Mbps or higher. Multiple-input and multiple-output (MIMO) technology has been introduced to increase antenna speed, and orthogonal frequency-division multiple access (OFDMA) is used as the modulation protocol. OFDMA combines an orthogonally modulated subcarrier and technology such as 64-quadrature amplitude modulation (QAM) to achieve high throughput (Figures 17 and 18).

5.5 Services using mobile telephone networks

Various automakers are providing telematics services in vehicle navigation systems that use mobile telephone

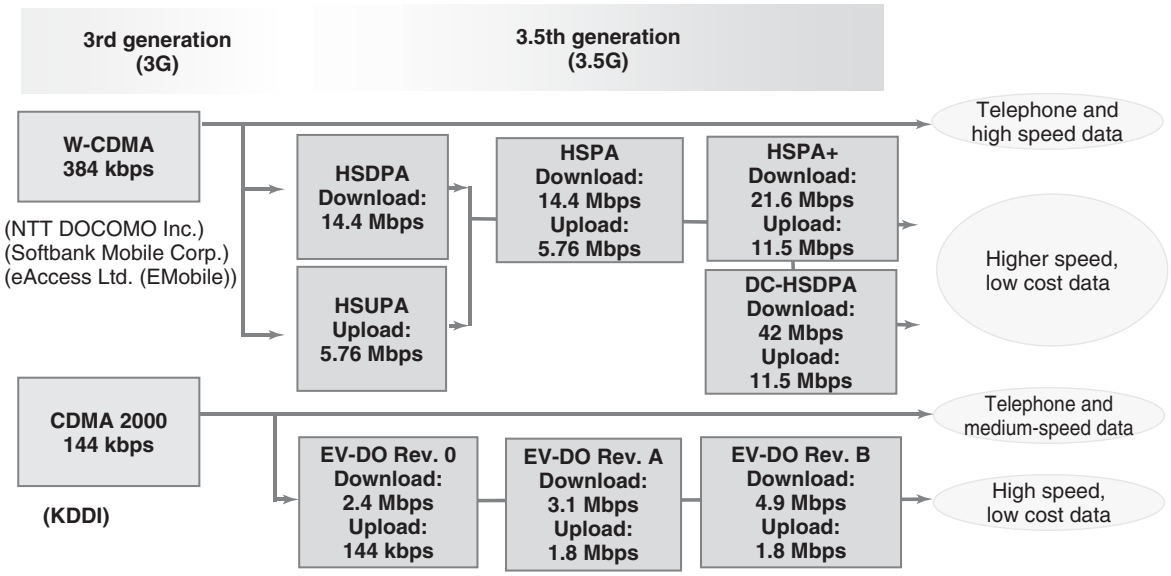


Figure 16. 3.5th-generation protocol.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

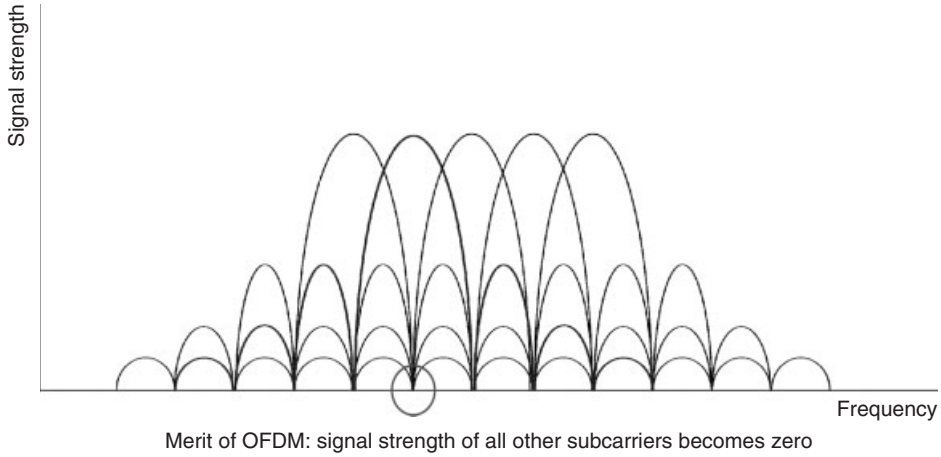


Figure 17. 3.9th-generation protocol.

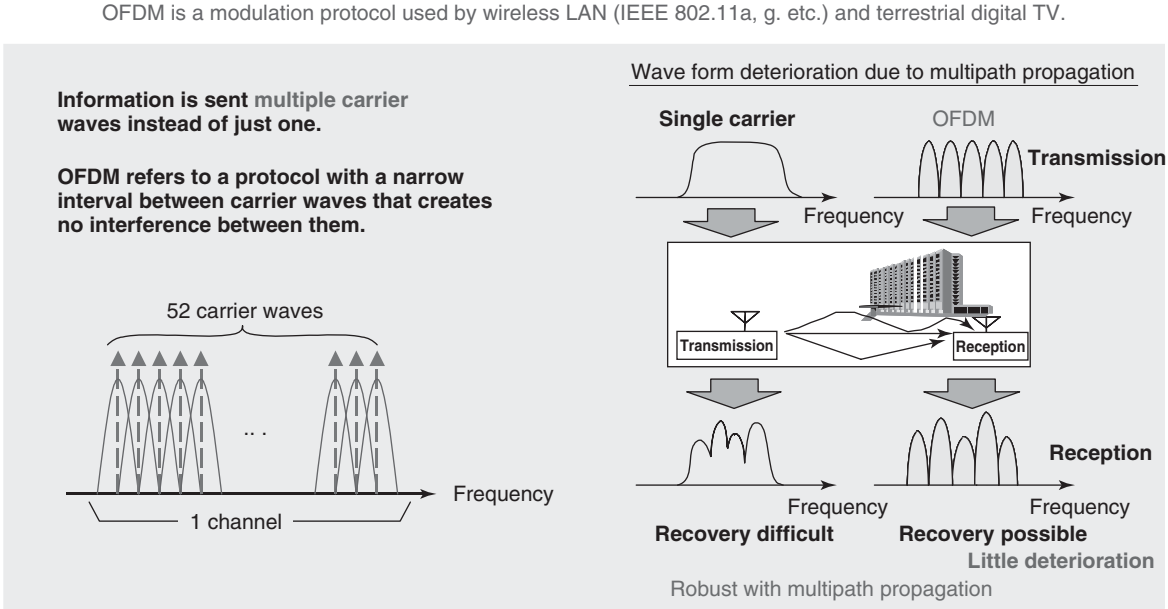


Figure 18. OFDM.

networks in combination with communication modules (Figure 19).

Automakers have introduced two broad categories of services, dedicated vehicle services and general-purpose services. Dedicated vehicle services include those related to (i) the vehicle navigation system and traffic information, (ii) safety and emergency services, (iii) vehicle services, and (iv) intelligent systems. In contrast, general-purpose services include (i) the provision of information such as news and weather forecasts, (ii) mobile commerce, (iii) multimedia entertainment, and (iv) mobile communication functions that display e-mails and the like through on-board devices. The following section gives a brief description of some of the general services offered by each automaker.

5.5.1 Emergency notification

System configuration: The system connects airbag deployment signals to an on-board wireless device. This device is connected to a microphone and speakers, and optionally to a backup battery. The backup battery ensures functionality even if the vehicle battery is damaged in a collision or the battery lines have been cut. Because the controller

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

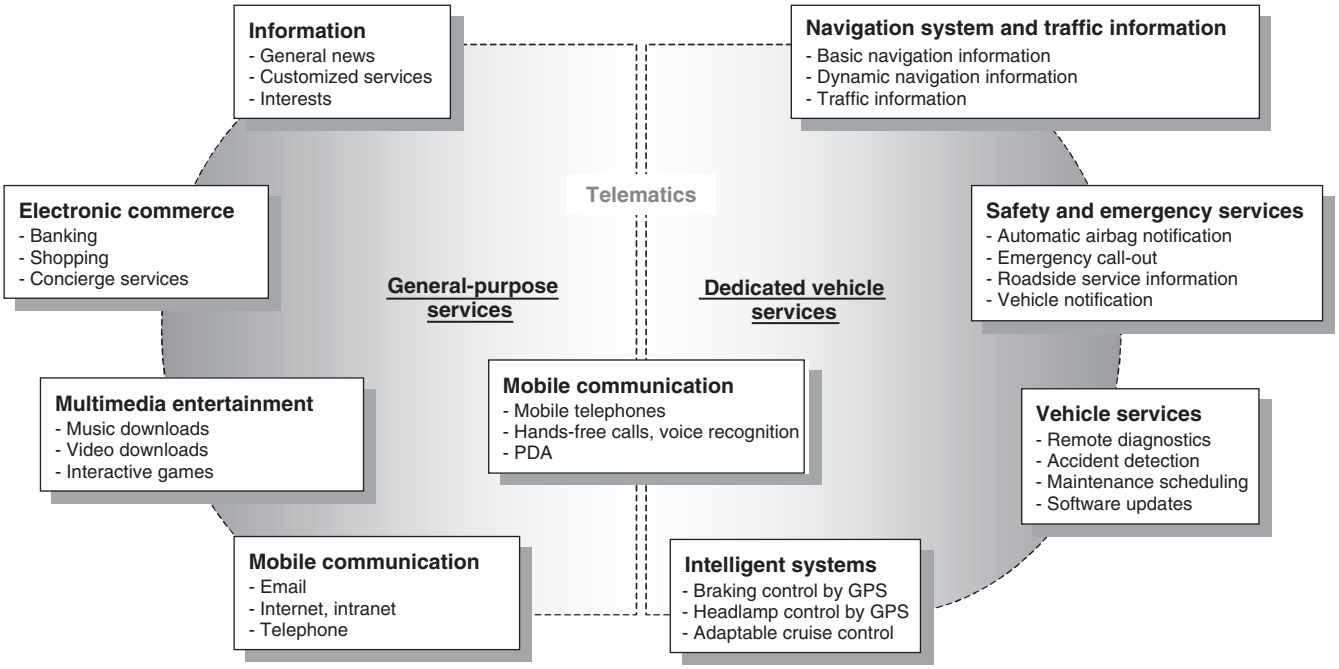


Figure 19. Example of telematics services.

area network (CAN) bus is also connected with each electronic control unit (ECU) in the engine compartment, it may suffer a short circuit or communication interruption after a collision. For this reason, a direct line is also used for communication.

1. If an accident occurs, the airbag deployment signal is communicated to the data communications module (DCM).
2. The DCM automatically notifies the support center of the accident location and other information.
3. The support center that received the emergency notification calls the driver, confirms the urgency of the accident, and summons emergency vehicles to the scene as required.
4. If the accident results in a situation in which the user of the system cannot respond to the operator, the operator calls emergency vehicles to the scene.

5.5.2 *Manual notification*

The on-board wireless device includes a switch that is pressed by the user to activate this service. Pressing the switch connects the user to an operator at a service center, who then provides the necessary road service by speaking to the user.

5.5.3 *Theft notification and tracking*

The on-board device is connected to an intrusion center and the smart key system. An alarm is then communicated by an on-board wireless device if abnormal vehicle entry is detected. Triggered by this alarm, the on-board wireless device communicates the vehicle abnormality to the emergency center, which sends a mail explaining the situation to the user's registered e-mail account. In addition, the user may request the center to track the vehicle. The center is capable of identifying the vehicle's current position by sending inquiry signals to the vehicle. Depending on the vehicle, the center can also send commands to lock the doors and shut the thief inside, or to reduce the vehicle speed.

5.5.4 *Operator services*

Users of vehicles with a wireless device-compatible navigation system can audibly notify the operator of their destination. The operator can then make complex navigation system settings by remote center command. Other operator services include remote door unlocking, access to medical networks, and so on.

5.5.5 *Remote services*

As the on-board wireless device is connected to the CAN bus, it can store various types of information at the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

center. That information can be viewed through the center homepage using a smart phone to check if the doors have been locked, or the like. The doors can then be locked or unlocked remotely. Other functions that can be activated remotely include the engine, automatic air-conditioning, and so on.

6 CONCLUSION

For vehicle users, the mobile phone is a necessary item. Telematics devices are becoming available in all vehicles as an accessory equipment.

The adoption of the LTE communication protocol provides a rough indication of the future of telecommunications. One example currently evolving is 4G communications. As it is a broadband protocol with a delay of less than 50 ms, LTE has the potential to enable various applications that have been difficult to achieve in the past. As an example, LTE may allow real-time remote operation of certain devices installed in the vehicle or of the whole vehicle itself. It might facilitate prevention of accident through active diagnostics that measure more detailed vehicle information in real time. Furthermore, LTE may also accelerate the introduction of autonomous driving technology through background communication.

FURTHER READING

Kato, M. (2010) *Automotive Electronics: Systems*, Nikkei Business Publications, Inc., Tokyo.

Kato, M. (2010) *Automotive Electronics: Basic Technologies*, Nikkei Business Publications, Inc., Tokyo.

Details of OnSTAR. http://www.onstar.com/web/portal/landing (accessed 10 December 2013).

eSafety Initiative. http://www.esafetysupport.org/en/esafety_activities/index.html (accessed 10 December 2013).

Japan Mayday Service. http://www.helpnet.co.jp/ (accessed 10 December 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto231
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Active Safety, Pre-Collision Safety, and Other Safety Products (Millimeter-Wave Radar, Image Recognition Sensors)

Yoshihiko Teguri

DENSO Corporation, Kariya, Japan

1 Active Safety Systems 1
References 9

1 ACTIVE SAFETY SYSTEMS

1.1 Relationship among the driver, vehicle, and driver assistance systems

Figure 1 shows block diagrams that illustrate the relationship among the driver, vehicle, and driver assistance systems. Figure 1a shows a vehicle without any controls, which is driven in accordance with the surrounding environment. Figure 1b shows a vehicle with a chassis control system. For example, the driver of this vehicle can perform avoidance maneuvers by steering as the anti-lock brake system (ABS) electronic control unit (ECU) detects tire lockup on frozen roads and reduces the brake pressure to recover grip. The role of chassis control systems is to enable the vehicle to be driven in accordance with the driver's intentions.

Figure 1c adds driver assistance control. As a result, the driver is located in parallel to the driver assistance system. This system processes the information from the surrounding environment and provides information to the driver. Specific examples include car navigation systems, night vision, adaptive front-lighting systems (AFS), and so on. Figure 1d includes systems such as precrash safety systems (PCS), adaptive cruise control (ACC), and lane keeping assistance systems (LKAS) that also function to control vehicle speed or the driving lane (Kato, 2010a,b).

Figure 2 summarizes the driver assistance and safety systems based on the relationship with a human model. This figure categorizes driver assistance systems into three categories based on Rasmussen's reflex skill, rule, and knowledge (SRK) model of decision-making. Chassis control is an example of skill-based assistance because it functions to supplement an insufficient action by the driver. ACC is rule based and it replaces the driver's function to maintain a set distance to the preceding vehicle based on vehicle speed. Car navigation systems are knowledge based and function to support the driving plan (Kato, 2010a,b).

This way of describing driver assistance systems suggests that supporting the driver and the vehicle through control that simulates the driver's brain is the future of the automobile, within the feasible range of autonomous driving. The sections below describe these driver assistance systems in more detail. These systems include those that simply provide information to the driver, and those that control the vehicle and assist driver operations (Masegi *et al*., 2007).

1.2 Driver assistance systems that only provide information

1.2.1 Night vision

Night vision is a device that enhances the field of view of the driver at night by displaying infrared images on

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto232
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

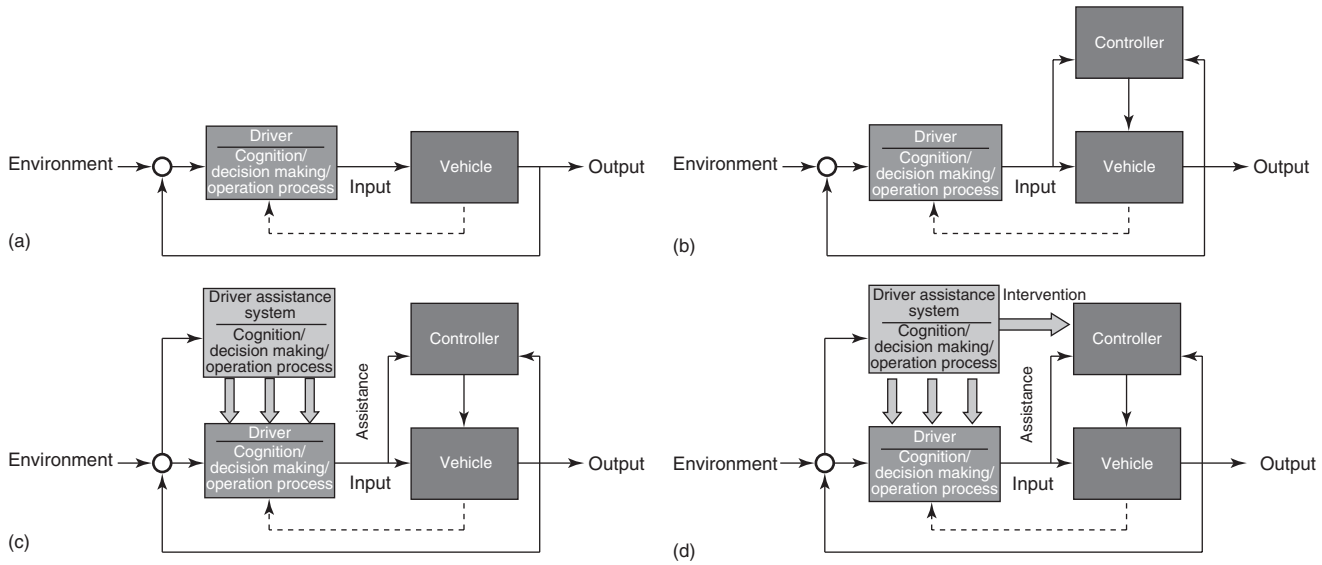


Figure 1. Relationship among the driver, vehicle, and driver assistance systems.

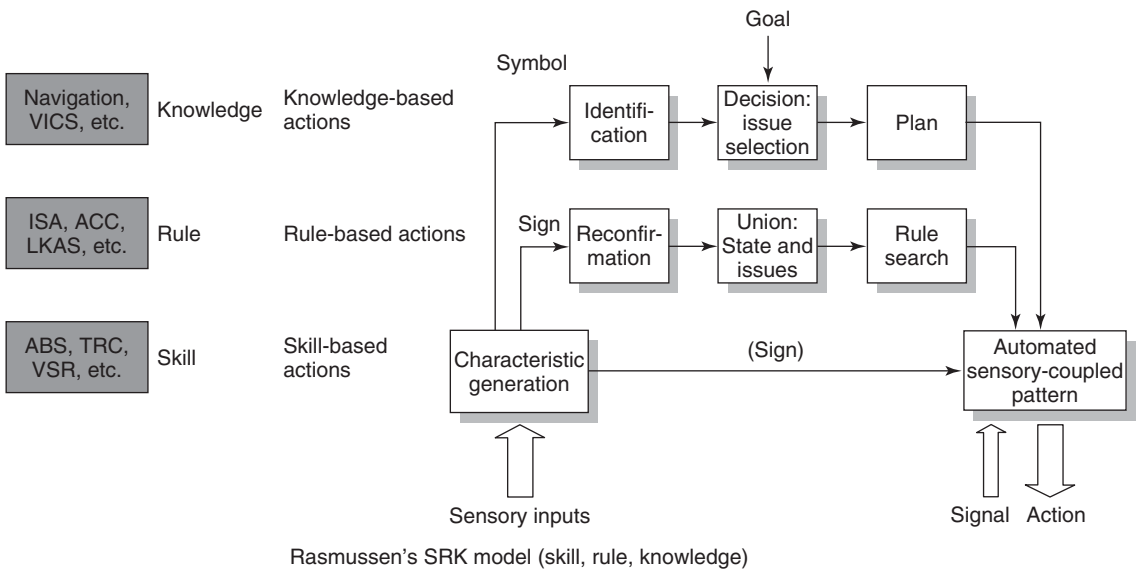


Figure 2. Human model and driver assistance systems.

a screen. There are two types of system, far-infrared and near-infrared. Far-infrared systems display images obtained from heat sources. The camera and lenses of this type of system are costly and the image quality is poor because of the long wavelengths of far-infrared rays. These include two-camera stereo systems that use heat sources to detect and emphasize the location of pedestrians at night.

Near-infrared systems illuminate the front of the vehicle with near-infrared light to obtain and display images using a camera sensitive to near-infrared rays. Some systems use a dedicated lamp or pass the light through a visible light cut filter placed in front of the high beam lights. Near-infrared rays can be used to obtain images that are close to those obtained using normal visible light. As a result, these systems are used for pattern matching recognition of pedestrians at night, the results of which are then emphasized and notified to the driver (Figure 3). The quality of these images is close to that of images obtained with visible light because near-infrared rays have short wavelengths.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto232
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

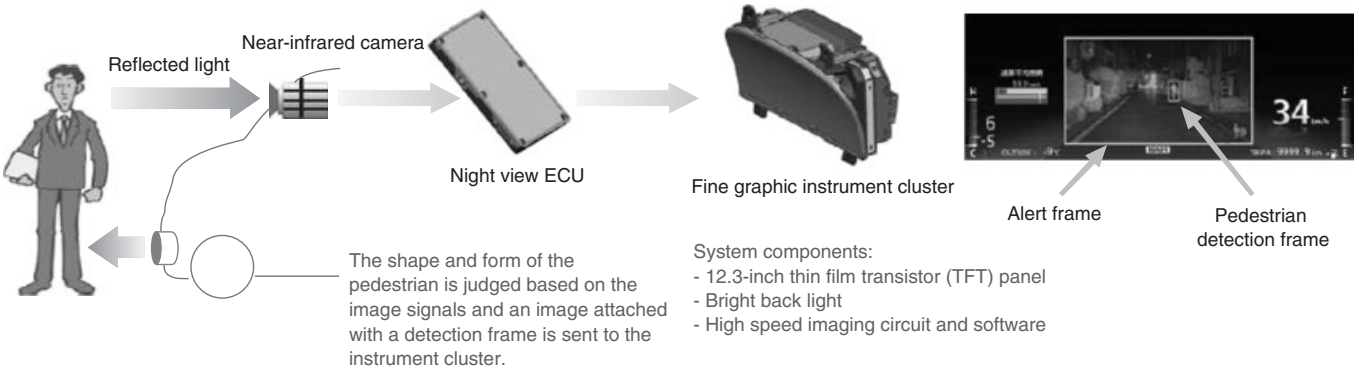


Figure 3. Night vision.

1.2.2 Adaptive front-lighting systems

AFS controls the optical axis of the vehicle headlamps to illuminate curves (Figure 4). For example, when the driver turns the steering wheel to the right to negotiate a right-hand curve, the steering angle and yaw rate sensors detect that the vehicle has entered a curve and move the optical axis of the headlamps to the right using motors. This helps the driver see the driving lane more clearly.

1.2.3 Other systems

Other examples of driver assistance systems that only provide information are proximity sensors such as rear-view cameras and sonar (Kato, 2010a,b).

1.3 Vehicle control systems

1.3.1 Adaptive cruise control

A conventional cruise control system maintains a set constant vehicle speed by controlling the engine only. In contrast, ACC uses radar as eyes to decelerate the vehicle using the brakes if the preceding vehicle is driving slower than the set speed. In this case, ACC maintains a constant vehicle-to-vehicle distance and then accelerates the vehicle to the set speed once the preceding vehicle has disappeared. Full speed range adaptive cruise control (FSRA) is capable of performing the functions of ACC at all vehicle speeds. Although a conventional ACC system will not function below around 40 km/h, FSRA can be used in all speed

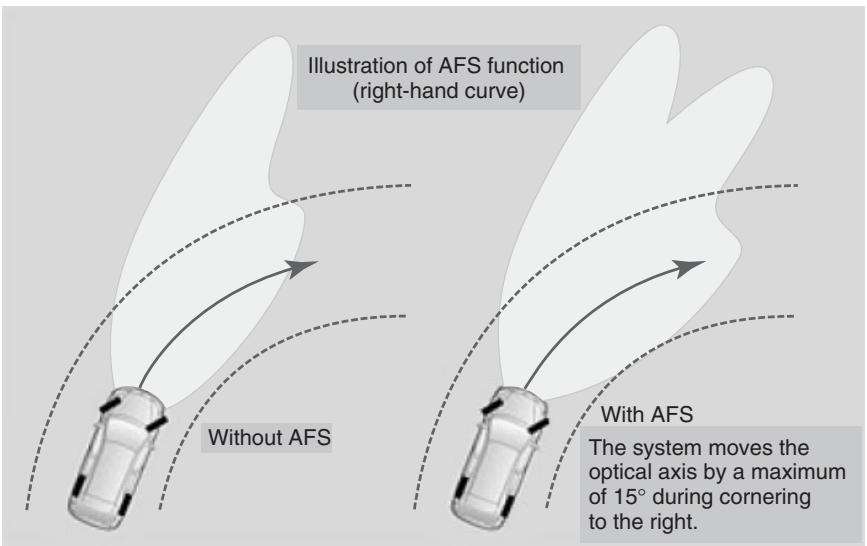


Figure 4. Adaptive front-lighting systems.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto232
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

ranges from starts to stops. After the preceding vehicle has decelerated to a stop, the driver can perform a simple operation to follow that vehicle once it moves off (Kato, 2010a,b).

1.3.2 LKAS and lane departure warning (LDW)

LKAS recognizes the driving lane using a camera to detect lane markings. If the vehicle is about to leave the lane, LKAS applies assistance torque to the steering wheel to bring the vehicle back to the center of the lane. LKAS is designed to stop operating if the driver is not holding the steering wheel. Lane departure warning (LDW) warns the driver when the vehicle has left its lane. This alarm may use sounds or vibrations. Both of these systems monitor the operation of the turn signal switch and do not perform control or issue a warning if the driver intentionally turns the steering wheel by a large amount (Kato, 2010a,b).

1.3.3 Precrash safety systems

PCS systems may also be called *automatic emergency braking systems* (*AEBSs*) in Europe. Such systems use radar to measure parameters such as the distance to forward objects and relative speeds. If there is a risk of a collision, PCS systems warn the driver, tighten the electrically powered seatbelts, and increase the pressure of the brake fluid. If the driver depresses the brake pedal, PCS systems assist the braking operation. If the driver does not depress the brake pedal, the system brakes automatically. In Japan, the functions of PCS systems are regulated by technical guidelines. In passenger vehicles, the brakes are applied at an acceleration of 5 m/s^2 or more, which is more than twice the speed of ACC systems (maximum 2.45 m/s^2).

Although the most common type of radar is the 76 GHz band millimeter-wave radar, laser radars are also used. Systems have also been adopted that combine radar with cameras to increase the accuracy of collision judgment. Other systems monitor the state of the driver using a camera inside the vehicle and image processing to detect the orientation of the driver's face or the motion of the driver's eyelids. These systems then apply warning braking or early warning notification if the attention level of the driver is judged to be low.

Currently, PCS systems for head-on collisions are most common. However, PCS systems for collisions from the rear have also been developed and it is likely that the scope of these systems will be expanded to include side-on collisions, rollovers, and the like in the future.

Europe is leading the way to legislate the installation of PCS systems in trucks, which have the potential to cause major damage in a collision. In the United States, it has been proposed to expand the New Car Assessment Program (NCAP) to include a forward collision warning (FCW) system, which notifies drivers when a collision cannot be avoided, along with LDW. In the future, it is possible that these advanced safety systems will become standard equipment on all vehicles in the same way as airbags and ABS. The main issue is the cost of these systems, but this may come down with mass production after the introduction of regulations and standards (Kato, 2010a,b).

1.3.4 Closing vehicle warning (CVW) and blind spot monitor (BSM)

Systems such as closing vehicle warning (CVW) and blind spot monitor (BSM) monitor the blind spots behind the driver and issue warnings as required when changing lanes. The aim of CVW systems is to encourage safe lane changing by warning the driver when a vehicle is approaching from behind at high speed in an adjacent lane. BSM systems have the same aim and warn the driver of vehicles in the blind spots at the side. The range covered by CVW and BSM sensors is different. The range of CVW is approximately 30–50 m, whereas BSM covers only a few meters close to the driver's vehicle. Although these systems mainly use 24 GHz millimeter-wave radar, some also use cameras.

1.4 The eyes of driver assistance systems

This section describes the sensors used by driver assistance sensors. These include conventional acceleration (*G*) sensors, gyroscopes, global positioning system (GPS) information, and the like. However, the characteristic feature of these systems is the capability to identify target objects by a process of recognition from measured physical quantities. Millimeter-wave radar and cameras are the most typical examples of such sensors.

1.4.1 Millimeter-wave radar

Figure 5 shows a millimeter-wave radar manufactured by Denso Corporation, which is used in ACC and PCS systems. Figure 6 shows the antenna and circuit boards used in the radar. Automotive millimeter-wave radars employ various detection methods. The radar shown below employs the frequency modulation-continuous wave (FM-CW) method to detect distance and relative speed, and digital beam forming (DBF) to detect the direction. Figure 7 illustrates the principles of each method.

In Figure 7, the horizontal axes show the time, and the vertical axes show the frequency. The radar launches

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto232
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 5. Millimeter-wave radar.

pulses modulated to form a mound shape as shown in the figure. The time lag until it receives the reflective pulse is proportionate to the distance. Here, as it is difficult to measure the time lag accurately, the distance is obtained by calculating the frequency difference of fs−fr instead. The bottom graph in Figure 7 shows a case when a relative speed is present between the driver's vehicle and the preceding vehicle. In this case, the entire reflective pulse shifts because of the Doppler effect. Thus, the fs−fr difference becomes smaller by the amount of the Doppler effect at the rising portion, and the difference becomes larger at the descending portion. The relative speed can be obtained by subtracting the differences in the rising and descending frequencies, and the distance by adding up the differences.

Figure 8 shows the direction detection principle. An outgoing pulse is transmitted from a one-channel antenna, and the reflective pulse is received by a five-channel antenna. As the path length of the reflective waves that reach the channels varies depending on the direction from which it comes, the direction is obtained by detecting the phase difference between the channels. The latest sensors add a new signal processing technology called *multiple signal classification* (*MUSIC*) to improve the separation performance of targets positioned closely (Miyake Y *et al*., 2007).

When driven on actual roads, radio wave interference due to the reflection of various objects at the side of the road generates ghost images and objects that are irrelevant to the detection targets, such as manhole covers that are almost the same height as the road surface. The post-processing technology that recognizes, categorizes, and discards these objects is extremely important. One typical issue is merging, in which two adjacent objects are detected as one. If the preceding vehicle merges with a stationary object at the side of the road, the system may judge that the vehicle has instantly braked. In response, tracking with

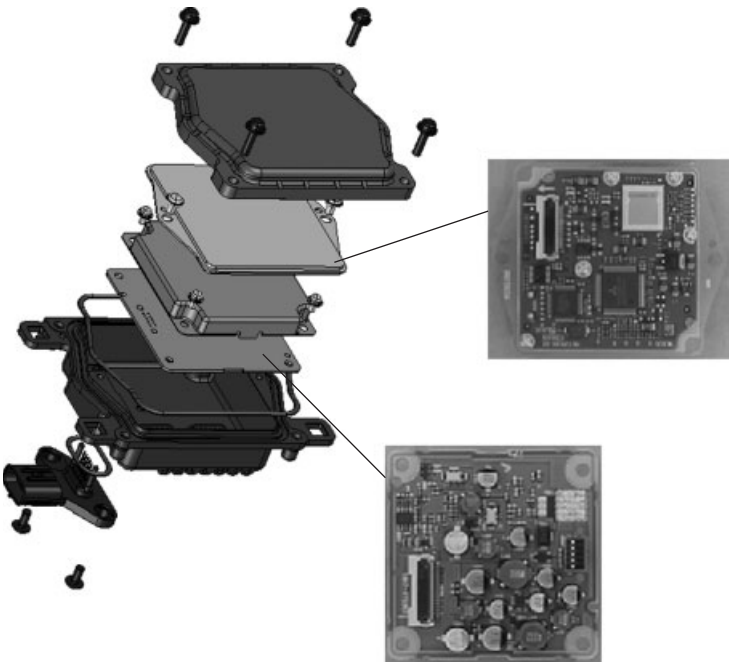


Figure 6. Structure and circuit board of millimeter-wave radar.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto232
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

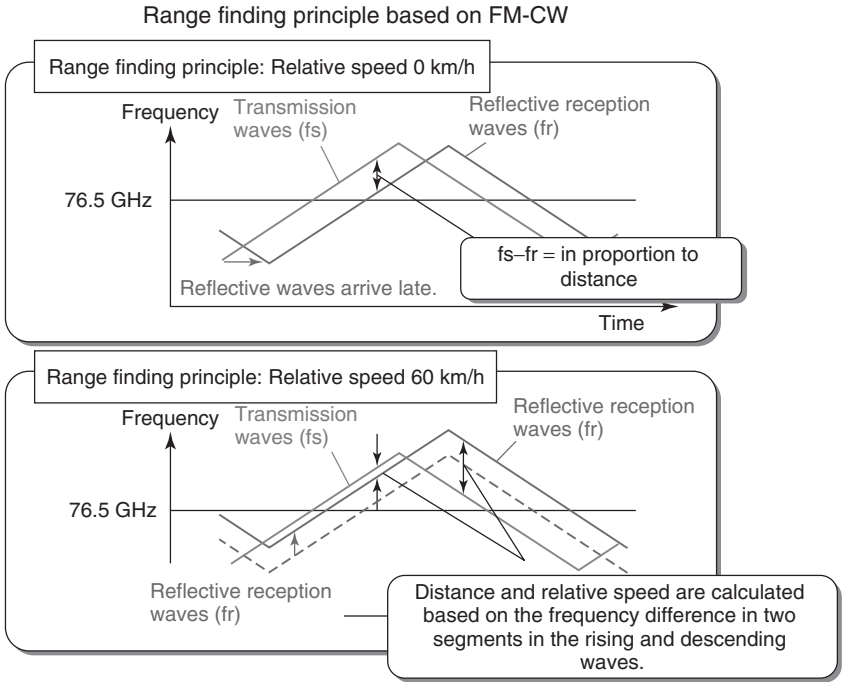


Figure 7. Principle of millimeter-wave radar range finding.

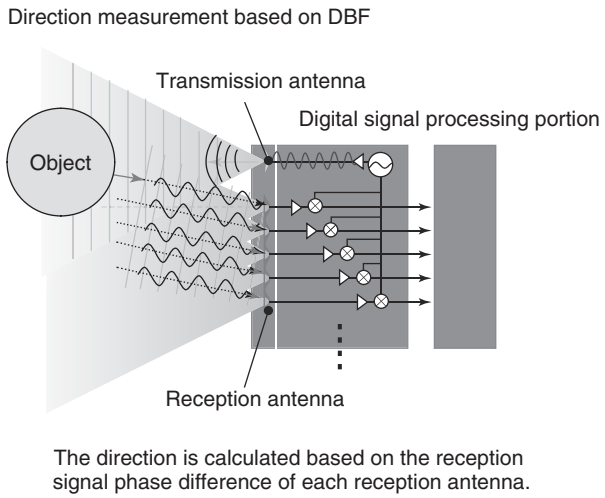


Figure 8. Principle of millimeter-wave direction measurement.

Kalman filters, processing with target selection algorithms, and methods to improve resolution are employed to enable appropriate recognition.

FM-CW and DBF are means of analyzing and processing reflective waves using calculations based on fast Fourier transform (FFT). The growing sophistication of automotive microcomputers and advances in digital signal processing technology are supporting the adoption of systems that use millimeter-wave radar.



Figure 9. Forward monitoring automotive camera.

1.4.2 Cameras

Figure 9 shows a camera manufactured by Denso Corporation and used for detecting lane markings. Figure 10 shows the camera module and Figure 11 shows the circuit board. The charge-coupled device (CCD) image sensor is black and white, and the video graphics array (VGA) resolution and signal processing use a general-purpose 32-bit automotive central processing unit (CPU).

Figure 12 shows the lane marking detection algorithm. Pairs of edges (i.e., points at which light and dark areas meet) obtained by applying a smoothing and differentiation process to a single scanning line are used as hints to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto232
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

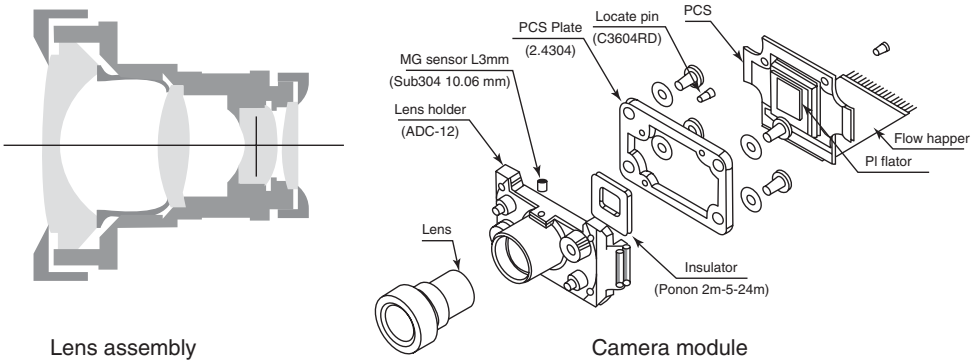


Figure 10. Camera module.

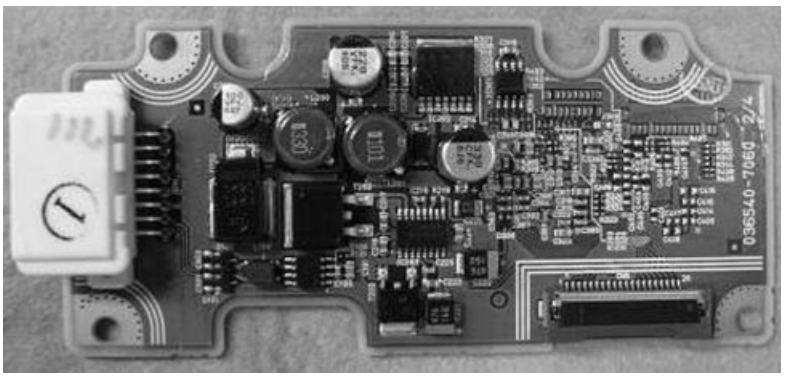


Figure 11. Camera circuit board.

detect straight lines using an algorithm called the *Hough transform*. Finally, tracking is performed using a Kalman filter to improve the precision frame by frame. As roads are not always straight, a method called *model matching* is used to handle curves.

Cameras are more susceptible to disturbances than millimeter-wave radar. The key for a camera-based system is achieving stable recognition with respect to disturbances such as day/night differences, the entrance and exits of tunnels, sunlight and shade, the afternoon sun, rain, wet road surfaces, faded or interrupted lane markings, glare, and halation. As a result, optical technology such as the image sensor, exposure control, and optical design of the lens assembly is extremely important.

Recently, software that recognizes vehicle bodies has been integrated with millimeter-wave radar for use in precrash brake systems. This so-called sensor fusion technology that incorporates millimeter-wave radar systems is capable of accurately calculating the potential of collisions, leading to more appropriate brake application (Masegi *et al*., 2007).

Recognition targets have also begun to expand from lane markings to include systems that recognize traffic signs such as speed limits, and warn the driver if the set speed is exceeded, systems that warn the driver of pedestrians, and so on. Pedestrian recognition uses an image processing technology based on pattern matching. The recognition targets are extremely complex and use dedicated high-performance processors such as the EyeQ from MobilEye and IMAPCAR from Renesas Electronics. EyeQ is hardware that contains a recognition image processor (IP) and uses an innovative means of accessing an image memory bus. IMAPCAR is a unique high speed architecture that enables parallel processing by 128 processors, and the like.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto232
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

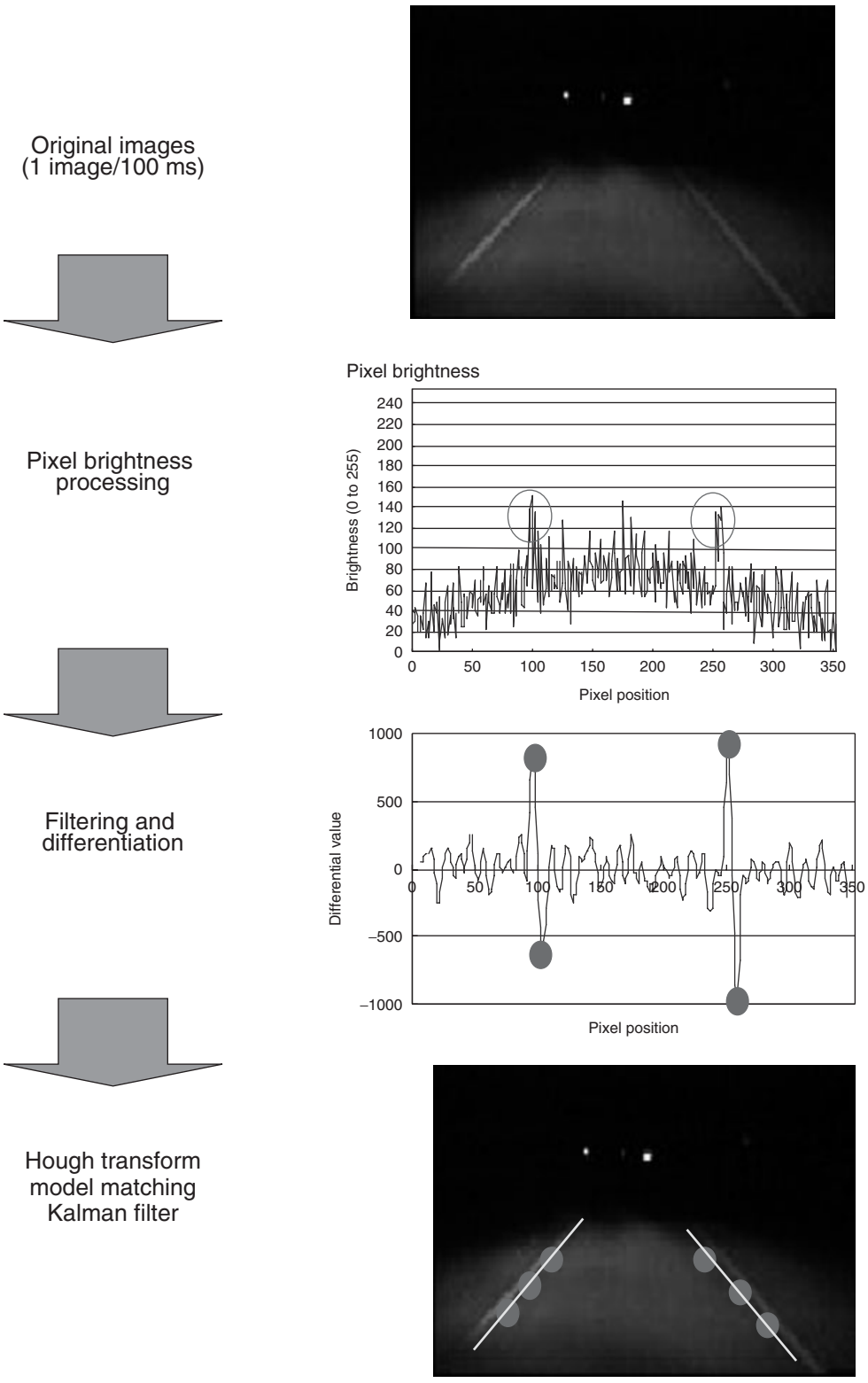


Figure 12. Principle of road marking detection.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto232
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

1.5 Future driver assistance systems

Sensors have already been developed that monitor the front, sides, and rear of the vehicle, as well as the driver. Infrastructure-coordinated systems can be combined with these autonomous type vehicle sensors to recognize the traffic flow situation. In addition to safer driving, such systems can be used to help reduce fuel consumption and provide route information. Furthermore, it may also be possible to use vehicle-to-vehicle communication to achieve platoon driving and various other services (Gallagher B *et al*., 2007).

REFERENCES

Kato, M. (2010a) *Automotive Electronics: Systems,* Nikkei Business Publications, Inc., Tokyo.

Kato, M. (2010b) *Automotive Electronics: Basic Technologies*, Nikkei Business Publications, Inc., Tokyo.

Masegi, M., *et al*. (2007) *Denso Technical Review,* **12** (1), pp 12.

Miyake, Y., *et al*. (2007) *Denso Technical Review,* **12** (1), pp 23.

Gallagher, B., *et al*. (2007) *Denso Technical Review,* **12** (1), pp 81.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto232
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Junction Blocks Simplify and Decrease Networks When Matched to ECU and Wire Harness

Takezo Sugimura, Kaoru Sugimoto, and Mamichi Tsuyuki

The Furukawa Electric Co., Ltd, Hiratsuka, Japan

1 Introduction 1
2 Car Electronics 1
3 Increase of ECUs 3
4 History of Wire Harnesses 4
5 Networks Between ECUs 5
6 Conclusion 7
References 7

1 INTRODUCTION

In this chapter, there is a description of the growth in the number of electronic control units (ECUs) and wire harnesses with the change in vehicle systems over time and the introduction of vehicle installation local area network (LAN). The number of ECUs on a vehicle has increased with new introductions and extension of vehicle systems year by year. Integration has been applied to control this ECU increase.

Along with this, the number of wire harnesses and their weight have increased, too. In addition, with the introduction of multiplex communication control, the computerization of the system and in-vehicle LAN has been introduced. Reduction of the amount of wire harnesses has been achieved by the introduction of this in-vehicle LAN.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto233
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 CAR ELECTRONICS

2.1 Transition of car electronics

In the 1950s, motorized body electric equipment has progressed, with the introduction of, for example, starters, wipers, power windows, and air-conditioning.

In the 1960s, automotive electronics still had single-control-function devices, such as voltage/current regulators and igniters. Later, in the 1970s, it became possible to control complex functions such as anti-lock brakes (ABS), air-conditioning, and fuel supply control by the advent of the microcomputer.

Since the 1980s, with the improved performance of the semiconductor, the reduction of cost and size in automotive electronics has progressed rapidly. After that, the installation of versatile electronic control for various individual systems made the wide expansion of applications possible (Figure 1).

In the late 1990s, engine, brakes, and power-integrated control systems were mounted on to HEVs (hybrid electric vehicles), which support an energy regeneration function.

Moreover, in recent luxury cars, vehicle lane-keeping functions, with integrated engine, brakes, steering control, and cooperative driver assistance systems that utilize infrastructure and information technology, have appeared with the aim of improved safety and comfort.

In addition, electric vehicles (EVs) and fuel cell vehicles that have no internal combustion engines have been developed with full-by-wire systems that have no mechanical backups at all. The aim is to introduce a system of full-by-wire (The Institute of Electrical Engineers of Japan, 2006, pp. 8–9).

		1970	1975	1980	1985	1990	1995	2000	2005	2010	2015
Powertrain	Power						Methanol		Hydrogen		
						Electric vehicle		Fuel cell vehicle			
		ICE (carburetor)		ICE (injection)			Hybrid				
	AT						CVT				
		3 speed mechanical AT			4 speed electronic controlled AT		5 speed electronic controlled AT	6 speed electronic controlled AT	7 speed electronic controlled AT		
	Throttle	Cable				Electronic controlled throttle					
	AWD	Mechanical			Viscous		Electric torque distribution mechanism				
Chassis	Steering				Speed sensitive power steering			Electric power steering (large car)			
							Electric power steering (medium car)				
				Hydraulic power steering		Electric power steering (small car)					
	Supension	Mechanical		Electronic controlled air suspension			Electronic controlled hydraulic	Electronic controlled magnetic fluid		(linear motor)	
	Brake					Traction control			Integrated chassis control system		
						Brake assist		Collision mitigation brake			
		Hydraulic		ABS			ESC				
	Cruise assist							Night vision monitoring / parking assist			
							Adaptive cruise control	Lane keeping assist			
				Cruise control (vacuum)			Electronic controlled cruise control				
	Occupant protection				Driver's seat airbags	Passenger seat airbags	Side airbags	Occupant detection sensor			
BODY	Air conditioning	Air conditioning	Cooler	Air conditioning	Auto air conditioning			Occupant detection air conditioning			
		Radiator cooling fan	Mechanical fan		Electric fan		Variable electric fan				
		Rear seat air conditioning				MPV (cooling)		MPV (heating and cooling)			
		Compressor	Mechanical compressor				Electric compressor	Hybrid compressor		CO_2 electric compressor	
	Entertainment & information	AUDIO	Cassette tape			CD		MD			
			AM Radio		AM/FM Radio				Digital radio		
		NAVI					Telematics	Rear seat entertainment			
						CD NAVI	DVD NAVI	HDD NAVI			
		Instrument panel	Mechanical Mater			Electronic controlled mater		Multi information display mater			
	Light & sight	Head light					Adaptive head light				
			Tungsten		Halogen		HID			LED	
		Defogger		Rear defogger				Front deicer		EHW	
		Side window	Mechanical	Power window		Key-off operation	Jam prevention				
	Other	Door lock	Mechanical		Electric door lock		Keyless entry	Smart entry			
		Mirror	Mechanical		Electric mirror	Retractable	Heated mirrors	Auto retractable			
		Seat	Mechanical			Power seat/heated seat	Memory seat	Memory seat with tilt /telescopic			
		Trunk	Mechanical				Power sliding door	Power tail gate & power trunk lid			

Figure 1. Transition in car electronics. (Reproduced from Institute of Electrical Engineers of Japan.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto233
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Currently, by-wire systems such as throttle-by-wire, brake-by-wire, and shift-by-wire have been put to practical use. However, the complete control only using the by-wire technologies has not been achieved. For example, brake-by-wire keeps a safety mechanism in combination with a mechanical transmission. Steering-by-wire has still not been put to practical use, but progress of motorizing the steering is moving forward with electric power steering (EPS) and active-steering. The installation of the steering-by-wire system has started on many vehicles. For complete steering-by-wire, the most important developmental challenges are the establishment of reliability, development of legislation, and a dedicated communication protocol for practical use. These are progressing, and it is expected that market introduction will occur in the near future. By the introduction of by-wire systems, the mechanical transmission mechanism can be eliminated from the operation of the vehicle, and thus the flexibility of automotive design will be significantly increased. Many benefits such as space, power consumption, lighter weight, fuel consumption, and cost reduction can be expected. For this reason, the next generation of cars will be built with a system only using full-by-wire.

3 INCREASE OF ECUs

3.1 History

3.1.1 Outline

Undergoing this transition of car electronics, the number of ECUs and circuit wire harnesses has undergone rapid changes. An example of the growth of the number of ECUs mounted on a luxury car is shown in Figure 2.

By increasing the systems installed in a given model of car, at every full model cycle, the number of vehicle ECUs tends to increase. The number of ECUs is increasing in luxury vehicles more than smaller cars. In the example of the luxury car shown in Figure 2, the number of ECUs exceeded 70 in 2005, whereas in 2010, 100 or more ECUs are installed. The reason for the ECU increase is due to the addition of comfort and safety features at an acceptable price and environmental protection functions since 2000 (Figure 2).

In classifying ECUs by their functions, such as "HEV/EV system," "safety system," "information systems," "body electronics system," and "power train system," every function tends to increase in the number of ECUs.

3.1.2 Power train

Power train systems have one or two ECUs for the transmission and the engine. The number of ECUs depends on whether the function is integrated or independent. In recent years, for the realization of high engine efficiency, various systems have become electric-electronic. Processing by the engine ECU only is possible if the electronic control is simple, such as adjusting value timing, but if the aim is to control a complex motion, such as electric controlled variable valve timing and lift, to avoid strain on the engine ECU, a dedicated valve train ECU has begun to be installed.

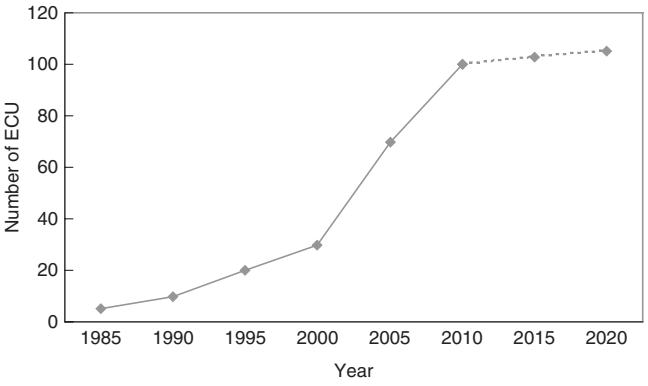


Figure 2. Transition of the number of ECUs (luxury car).

An automatic stop and start control system, which is basically controlled by the engine ECU, may be performed by an increasing number of dedicated ECUs to accommodate factors such as supply voltage fluctuation.

3.1.3 Body system

Body system ECUs include air-conditioning, meters, smart key, latches, body control module (BCM), and so on. Basically, each ECU is installed individually. In some cases such as BCM, some control functions are integrated, for example, lighting system such as head lamps, door locks, wipers, and interior lamps. On the other hand, by functional integration to the BCM, there are increased problems in the number of circuit wire harnesses and need for increased installation space because of the larger ECU. For this reason, the case for mounting a dedicated ECU depends on how many functions for control are needed. For example, ECUs mounted on the doors, the engine, or around the steering. It is considered that there will be an increase in separated BCM functions controlled by ECUs in distributed areas.

3.1.4 Information system

ECUs of information systems, such as navigation, audio, telematics, and digital TV, have dedicated ECUs for each system, because there are many system configuration options. In the future, the addition of various information

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto233
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and services, such as communication between car and car and car and road transport infrastructure, will increase the number of ECUs further.

3.1.5 Safety system

The ECUs for safety systems, such as airbag, electronic stability control (ESC), vehicle camera systems, radar systems, and peripheral monitoring, have, because there are many system configuration options, dedicated ECUs mounted for each system. In the future, the addition of advanced safety functions will require an increase of ECUs.

3.1.6 Hybrid electric vehicle/electric vehicle

In the ECU systems of HEVs/EVs, the number of ECUs differs from that of an EV to a mild-HEV or a strong-HEV. Typically, dedicated ECUs are mounted for motor control, HEV/EV management, inverter control, and battery management.

3.2 Problems of increased ECUs

While increased ECUs can deliver benefits, a significant problem is that the space where the ECU can be mounted in the vehicle is limited. Usually, the place is in the vehicle cabin, such as the foot space under the driver's seat or passenger's seat, as a good temperature and/or vibration condition is required. On the other hand, expanding the cabin living space in the quest to comfort leaves less space for the ECU. Thus, it is difficult to keep the space for the ECU and this has become a serious problem.

In order to solve these problems, reducing the number of ECUs by integration is underway. For example, in the case of BCM, some ECUs belonging to the body electronics function are integrated into one ECU. In the case of safety systems such as the driving support ECU, the system is integrated with the face-direction-detecting ECU, the millimeter-wave radar ECU, and the stereo-camera ECU. In the case of the information system, an ECU is mounted in the center display module that is integrated with the audio and navigation ECU.

This ECU integration trend is expected to expand more and more in future generations of vehicle. In the future, the integration of ECUs will proceed beyond the function field, such as information systems and safety systems with the system detecting the field around the car.

3.3 Predictions for the future

Beyond the integration of ECUs, integrating power lines and signal lines that were necessary for the individual ECU is becoming possible; the effect is a reduction in overlapping circuits on wire harnesses.

Thus, as was indicated in Figure 2, the number of ECUs on a vehicle will tend to decrease as existing systems integration proceeds. However, this reduction effect is offset by the increase of ECUs with new systems. Therefore, the trend in the number of ECUs may remain flat until around 2020, or it may only increase slightly (Tanokura, Shindou, and Okubo, 2006, p. 115).

4 HISTORY OF WIRE HARNESSES

4.1 Historical trends in number and weight

An example of the trend in the number of circuits mounted on a luxury car is shown in Figure 3. The trend in wire harness mass for the same luxury car is shown in Figure 4.

For environmental control, safety, and comfort, electrical systems in cars have been increasing in functionality and circuits per vehicle wire harness every year.

Moreover, the number of wire harnesses will also increase as new electrical systems are added, and ECUs and multiple types of electrical equipment such as switches, sensors, valves, and motors are connected in a one-to-one connection.

In the past, when new electrical systems have been added, the number of circuits increased to a dozen at most, but the electrical systems advanced in recent years, which mean that 30–60 extra circuits have been added in one system.

4.2 Problems of increased wire harness

Although the space for mounting the vehicle wire harness is limited, owing to an increase in the number of these circuits, the problem of finding a suitable location for the

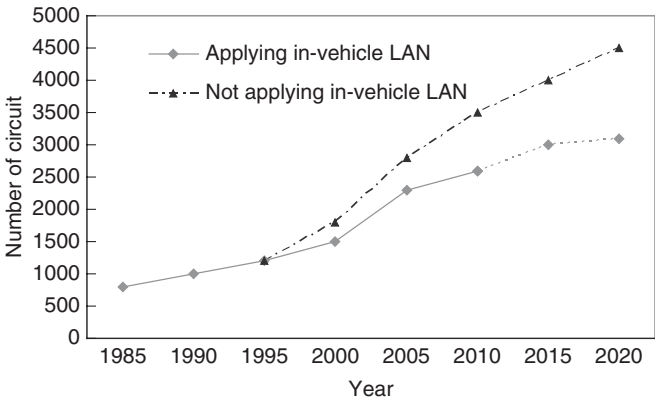


Figure 3. Transition of the circuits (luxury car).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto233
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

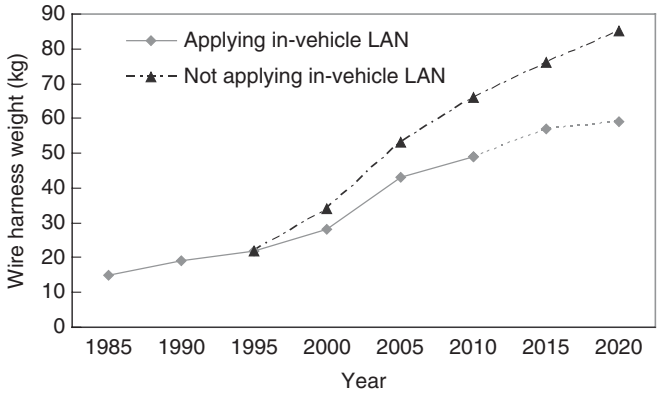


Figure 4. Transition of the wire harness weight (luxury car).

wire harness has arisen. In addition, the increase in the number of circuit wire harnesses has led to an increase in the vehicle weight and a deterioration in the ease of assembly work of the vehicle. As is required to improve the fuel efficiency of automobiles because of the rise of environmental concerns in recent years, the reduction in the weight of the entire car is being promoted and therefore lighter wire harnesses are also required.

4.3 Introduction of in-vehicle LAN

In order to solve these problems, the technology of in-vehicle LAN had begun to be introduced from the 1980s. Figure 3 shows the number of circuit wire harness introduced with the presence or absence of in-vehicle LAN. Comparing the 1990s to the present, the circuits increased from 1000 to 2600 circuits (applying in-vehicle LAN), but the increase of the number of circuits is estimated to be up to about 3500, without the introduction of in-vehicle LAN.

Figure 4 shows the mass growth of the wire harnesses in the presence or absence of in-vehicle LAN. Comparing the 1990s to the present, wire harnesses have increased from 15 to 49 kg (applying in-vehicle LAN), and the increase in weight of wire harnesses is estimated to up to about 66 kg if there is no application of in-vehicle LAN.

In this way, even though new electrical systems have been added each year, by introducing the vehicle LAN mode, it is possible to suppress the increase in the number of circuits and wire harness mass.

4.4 Rearranging of power supply distribution and the wiring

Generally, there are two roles for the junction block. The first is the storing/fixing/protecting the safety equipment such as a fuse, the relay, or the control of the power supply ON/OFF by the relay or rearranging the power supply distribution according to load and the joint circuit of the power supply. Secondly, there is a role for an organizing BOX wire harness to split the wire harness, to make for easier assembly in the vehicle. Duplication of circuit power lines and signal lines can be reduced by the aggregation of components for safe distribution and function, including power in the junction block, thus realizing a reduction of wire harnesses.

4.5 Decentralization of the power supply

In recent years, with the increase in the number of electrical systems for environmental, safety, and comfort purposes, there has been a rise in the number of circuit wire harnesses. Architecture has been considered to be a solution to this problem, by the distribution of the power source to the junction block for each area (i.e., a decentralization of the power supply). Specifically, a junction block placed on the engine, where the electrical equipment is concentrated, in the driver's seat and passenger's seat sections, the instrument panel section, or the center cluster section. Such decentralization of power reduces the wire harness power delivery (e.g., circuit lighting or a ground circuit). In addition, further lightweight wire harnesses can be used by minimizing the total wire length of power lines and electrical equipment cable leading to further weight reduction. Moreover, the power distribution architecture with in-vehicle LAN connections between the junction blocks for each area makes it possible to share power control information. The purpose of this power distribution method has the effect of reducing power by efficient distribution and management of ordered power, avoiding enlargement or concentration of the wire harness. In addition, ordered power distribution architecture contributes to the standardization of power supply arrangements: the standardization of power distribution (The Institute of Electrical Engineers of Japan, 2006, p. 18). It follows that junction blocks will have the role of power management and distribution control functions.

5 NETWORKS BETWEEN ECUs

5.1 Histories

Figure 5 shows an example of network deployment and the time of application LAN technology started to be introduced into cars from the 1980s. Many automobile manufacturers adopted in-vehicle LAN interfaces using their own original

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto233
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

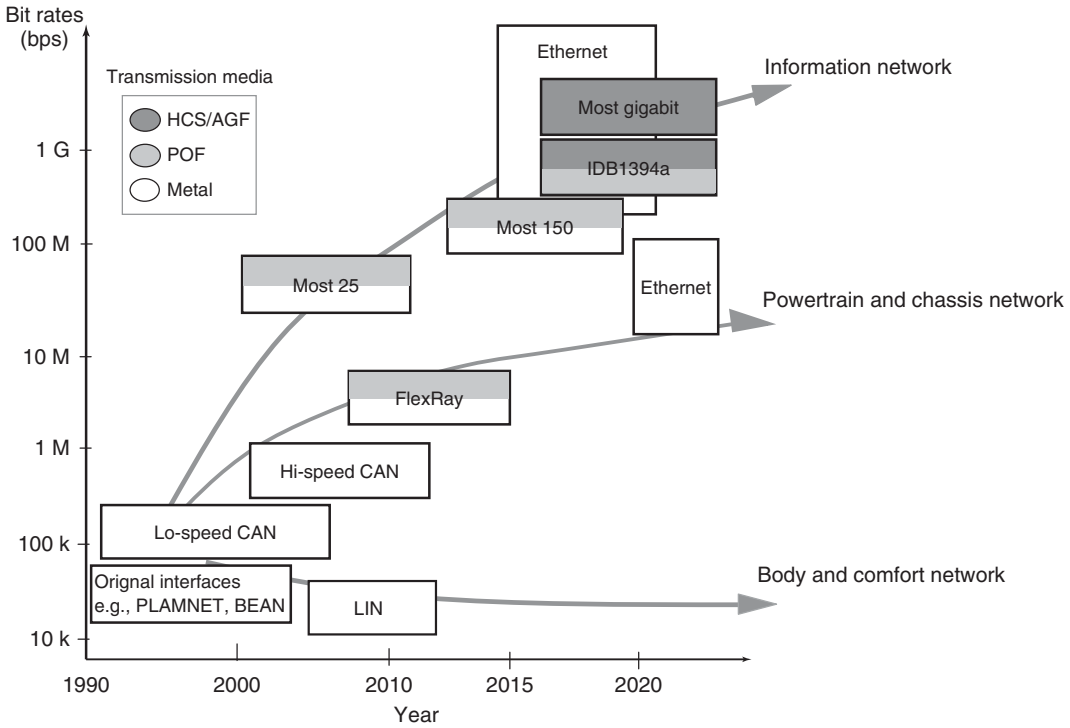


Figure 5. Network deployment and transition period.

standards developed independently of each other. Introduction of the automotive in-vehicle LAN began with the body control system. Examples are Chrysler "C2D", GM "J1850VPW," in the year 1990, Daimler "CAN," BMW "I-BUS" and "K-BUS," Toyota "BEAN," Mazda "PLAMNET," Honda "MPCS," and Nissan "IVMS." In many of these original interfaces, the data transmission speed was about 10 kbps.

5.2 Standardization of in-vehicle LAN

Not only does the wire harness continue to increase in number of systems installed in vehicles, but also there is a problem with the lack of data transmission speed in the standard interfaces previously adopted. From around 2000, world automobile manufacturers started to adopt the controller area network (CAN), which is a single interface for standard in-vehicle LAN.

In this way, CAN has become the mainstream in-vehicle LAN, although other LAN interface standards have also appeared since 2000. Those are not developed for each automobile manufacturer, as was the case in the 1980s, but standardized organizations such as Tier1 suppliers, automobile manufacturers, and semiconductor manufacturers participate and formulate optimal in-vehicle LAN standards for each vehicle system.

5.3 LAN structure

As shown in Figure 5, several kinds of standard LAN interface are used according to the difference of data transmission speed, purpose, or reliability.

Classification according to the in-vehicle LAN domain can be divided into three systems: control systems, body systems, and information systems. CAN and FlexRay are control systems, CAN and LIN are for body electronics systems, and MOST is for the information system; all are used widely as the typical in-vehicle LAN.

In this way, communication goes ahead through multichannel LAN in the car, which is made up of plural subnetworks and used properly according to the domain. Currently, a typical luxury car may have multichannel networks composed of a LIN local network of multiple channels, MOST subnetworks, and nine channels of CAN subnetworks. Transmission of information between these subnetworks has been realized by a dedicated gateway ECU; its gateway function is set up to connect with the ECU's multiple plural subnetworks.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto233
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5.4 Problems of high functionality of electronic control

So far, with the high function of cars, LAN has been applied to cooperation or integrated controlling functions, by sharing actuator and sensor information between ECUs. In future, equipment related to the precrash safety advanced safety features developed in order to prevent a car accident, which will have wide applicability in the next generation of cars, and if cooperative control of vehicle infrastructure and road traffic by intelligent transport systems (ITSs) becomes widely used, the importance of in-vehicle LAN systems connecting each ECU is expected to progress further.

On the other hand, MOST and CAN have limits with this transmission speed; the maximum transmission speed of CAN is 1 Mbps and MOST25 is 25 Mbps. It follows that there can become a difficult situation transmitting by MOST25 or CAN, with an increased number of ECUs or growth in complexity of control, when the amount of data increases dramatically.

For example, control systems for the basic function of vehicle running required the improvement in motor control, to enhance vehicle safety, for which LAN interfaces need higher reliability and a faster real-time response. In addition, in information systems, LAN interface standards require emphasis on communication speed and a large amount of data transmission, with increased application of environmental monitoring cameras and simultaneous streaming of audio and video that this will mandate (Tanokura, 2011, p. 11).

5.5 Next-generation standard LAN interface

To compensate for these weaknesses and new requirements, various next-generation standard LAN car interfaces are being developed. In control systems, the standard interface equivalent to the next generation of "CAN" is "FlexRay," which has better speed and reliability, and guarantees a response at a given time, with a maximum transmission speed of up to 10 Mbps, and has begun to be adopted in some luxury cars. Other standardizations, such as the "automotive Ethernet" as a long-term measure, are being developed.

In information systems, the equivalent to next-generation "MOST" are "MOST150" with up to 150 Mbps transmission speed, "IDB1394" with a maximum transmission speed of 800 Mbps, and the "Ethernet" with a maximum transmission speed of 1 Gbps, as the standards for transmitting voice or video data.

The "Ethernet" has been developed as the standard to encompass both control and information systems in the long term.

In this way, the transmission speed of the next-generation standard LAN interface is moving in the direction of large capacity/high speed communications. In addition, using a standard LAN interface of the next generation, it will be possible to interact faster with much more information. This will be needed unless the number of ECUs does not decrease dramatically, using multichannels with multiple subnetworks (Noumi *et al*., 2007, p. 20). However, it is thought that the future LAN constitution in the car will be multichannel and layered according to domain, with CAN, LIN, MOST, and Ethernet.

6 CONCLUSION

The introduction to the function of the junction box, the in-vehicle LAN, and integration of multiple ECUs discussed in this chapter has become indispensable in the solution of issues such as the increase in the number and weight of the wire harness circuits to transmit power and signals with the increase in number and ECUs as new electrical systems that are mounted on a vehicle. To build the overall architecture of the automobile electrical system by combining these solutions, to be developed further are the methods that deliver lightweight wire harnesses and reduced numbers of ECUs and the number of electrical systems for an increasingly complex future vehicles.

REFERENCES

Institute of Electrical Engineers of Japan (2008) Management technology of automotive electric power supply systems. Technical Report No. 1121. The Institute of Electrical Engineers of Japan, Japan.

Investigation Committee on Next Generation Automotive Electric Power Systems (2006) Roadmap of next generation automotive electric power systems. Technical Report No.1049. The Institute of Electrical Engineers of Japan, Japan.

Noumi, K., Nishihashi, S., Ishikawa, Y., and Takahashi, J. (2007) Approaches to the development of the in-vehicle LAN system *FUJITSU TEN Technical Journal*, **49**. Fujitsu Ten, Japan.

Tanokura, Y. (2011). Toyota Eyes Adoption of Ethernet for On-Board LAN. *Nikkei Electronics* (June 27).

Tanokura, Y., Shindou, T., Okubo, S. (2006). Improving the Lexus. *Nikkei Electronics* (December 18).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto233
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The Wire Harness

Takafumi Kawauchi, Masaharu Onda, Kenji Matsuo, Yukihiro Kawamura, Yoshihiro Shiotani, and Nobuaki Sakai

Furukawa Automotive Systems, Shiga-ken, Japan

1 Introduction 1
2 The Wire Harness 1
3 Wire 4
4 Connector 4
5 Power Supply Block 6
6 Wire Harness Exterior Parts 7
7 The Future of Wire Harness 8
References 9
Further Reading 9

1 INTRODUCTION

A power supply system in a vehicle electrical system has three basic functions: power generation represented by an alternator, power supply-storage function represented by a battery, and power distribution represented by a wire harness. In this chapter, we discuss the power distribution function of the wire harness.

2 THE WIRE HARNESS

The "wire harness" is a collective term for the composition of a set of wires used for supplying power and transmitting signals linked to electrical equipment by means of connectors, clamps to install these wires in the units, and protectors to protect the wires from damage. Functional parts such as the junction block (J/B) and the relay block (R/B) are also included in some cases. The wire harness that is used for modern automobiles with an internal combustion engine has drastically changed in the latter half of the twentieth century with the introduction of advanced electronics in vehicles.

An electronic engine-management system, with which controlled ignition and injection timing was introduced by Bosch in 1979, achieved fuel efficiency improvement and exhaust emission reduction. Today, this system is widely used because of tight regulations on exhaust-gas emissions, as part of global environmental consciousness. The introduction of the engine-management system has largely influenced the wire harness structure. For instance, the R/B has been added to the wire harness because this system composed of many relays. Owing to the addition of R/Bs, the wire harness divided into parts, as the wire installation was unable to pass through the hole between the engine room and the indoor area with the normal through the hole. For a switch system utilizing the R/B in the indoor area, direct switch types replaced relay types, so this replacement was an important factor in altering wiring routes.

Wire harnesses have increased their presence in terms of role and volume in the introduction of advanced electronics. The total length of wire per vehicle has extended from 35 m (16 connectors) in 1949 to 150 m (200 connectors) in 1965 and then surged to 2000 m (1500 connectors) in 1990 (Eckermann, 2001, p. 204). For vehicle reliability, weight reduction, and space saving, the introduction of local area network (LAN: multiplex telecommunication technology) systems started in the 1980s, and controller area network (CAN) systems were announced by Bosch in 1986 and practically used in 1992 in order to reduce wire volumes (see Increase of ECU and Wire Harness,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto234
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

JB Simplify and Decrease Networks). Nevertheless, the introduction of sophisticated electronic systems such as the by-wire system, already introduced in the aviation industry, has induced much higher dependence on electronics. The cost of electronics was about 15% of a vehicle in 1992; this rose to reach 25% in the early 2000s and is now expected to surge to up to 50% after 2010 (Eckermann, 2001, p. 205; Gomes and Fernandes, 2010, p. 378). As a result, wire length now extends to 1000 m even for small cars, and is no less than 4000 m for luxury models.

Owing to the wire volume increase in recent years, the wire harness was required to be much smaller and lighter, so conductor downsizing and insulation thickness thinning have been developed. As for conductors, core size was reduced from 0.5 mm^2 of annealed twisted copper wire to 0.3 mm^2, and 0.3 mm^2 of annealed core wire was replaced by 0.13 mm^2 of alloy copper after 2000. Meanwhile, the insulation thickness was reduced to 0.2 mm. As a result, about 10–40% core size reduction and about 30–60% weight saving was achieved (in comparison, 0.5 mm^2 of annealed twisted copper wire with 0.3 mm^2 of alloy copper and 0.3 mm^2 of annealed core wire with 0.13 mm^2 of alloy copper). Further, wire harness weight reduction (about 30% reduction compared to conventional wire harness) was achieved by the introduction of a flexible, flat cable with rectangular-shaped conductors. There has also been the movement to introduce aluminum for the conductor instead of the now dominant copper, to reduce the risk in fluctuations of material prices.

Moreover, investigation of environmental technology was activated during late 1990s. After the introduction of lead-free and high recyclability insulation material, halogen-free polyethylene (PE) was developed around 2000 as a replacement for polyvinyl chloride (PVC) and has since been applied to the automobile wire harness. For PVC-insulated wires, much thinner insulation reduced its usage.

The wire harness in electric vehicles (EVs) such as hybrid electric vehicles (HEVs), EVs, and fuel cell vehicles (FCEVs), introduced with the objectives of fuel efficiency improvement and carbon dioxide reduction of exhaust gases that should withstand over 200 V and be required to resist vibration and electromagnetic noise (IRC, 2010, p. 795).

The wire harness volume tends to enlarge in high demands for comfort, convenience, and safety such as intelligent transport systems (ITSs) (see Intelligent Transport Systems: Overview and Structure (History, Applications, and Architectures), and precrush safety systems. In addition, the wire harness has begun to play an important role in the requirements of communications with social infrastructure, households, and other automobiles, thus meaning that continuous improvements of weight reduction, downsizing, and cost reduction of wire harness are indispensable.

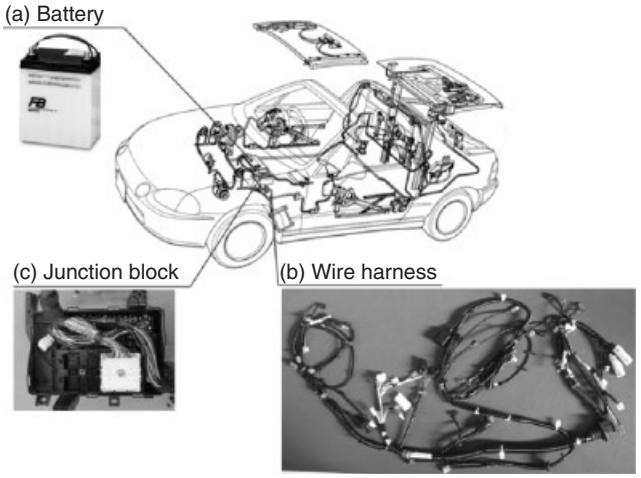


Figure 1. (a–c) Example of wire harness and related parts in a vehicle.

2.1 Role of the wire harness

The battery (Figure 1a) supplies electric power and the engine is started by activating the starter motor. After the engine starts, the power-generating alternator supplies electric power to the power supply block. The power supply block is equipped with power-related parts such as wire-protecting fuses and switching relays. Once the power supply block is activated, power is supplied to each unit.

The wire harness (Figure 1b) connects these units and supplies power and transmitting signals. The total wire length depends on the vehicle and has a range of approximately 1000 m for a small vehicle to 4000 m or more for a higher grade vehicle. Although the wire harness is installed out of sight, it serves an important function in the vehicle.

Power supply blocks such as the R/B with relays, fuse block with fuses, and J/B (Figure 1c) with relays and fuses and circuit connections are used in a vehicle. These power blocks are located in the engine room (under the bonnet/hood) or the instrument panel. The location of the power supply block depends on the individual car maker or model.

2.2 Wire harness types and examples of grouping

Wire harness grouping depends on the installation condition, power supply distribution, and operating environment.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto234
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The following are the wire harness types and features in a typical sedan-type vehicle.

2.2.1 Battery harness

The battery harness supplies power from the battery to the starter, power block, and other systems. Owing to the high current, large-size wires are used.

2.2.2 Engine harness

The engine harness is assembled on the engine itself. Wires are routed along the engine, which generates high heat and therefore heat-resistant wires and tubes are used on this wire harness. Also, designing an engine harness requires engine vibration to be considered.

2.2.3 Engine room harness

This type of harness is assembled on the body panel in the engine room. Waterproofing is important as well as the engine harness. Wire protection is also required when it is assembled around metal brackets, brake pipes, and so on.

2.2.4 Instrument harness

Instrument harnesses are installed inside the instrument panel. This harness is mainly connected to various instruments and audio switches. Since this harness is located inside the vehicle, chattering noise suppression is needed.

2.2.5 Floor harness

The floor harnesses are installed on the indoor floor of the vehicle, and located widely from the front to the rear of the vehicle. These harnesses tend to be longer as they cover the floor area.

2.2.6 Other small harnesses

Small harnesses are divided in the process of installation process and operating environment. Examples of small harness types are found in the door, trunk room, and so on. Door harnesses require a high durability design that will withstand door opening/closing.

Other than the examples listed above, there also are hybrid harnesses combining the engine room harness, instrument harness, and floor harness or combining engine harness and engine room harness.

2.3 Functions of wire harness components

This section introduces the main components of wire harnesses and their functions. Examples of the wire harness components are shown in Figure 2 and their functions are listed in Table 1.

(a) Wire
(b) Connector
(c) Clamp
(d) Grommet
(e) Tape
(f) Protection part

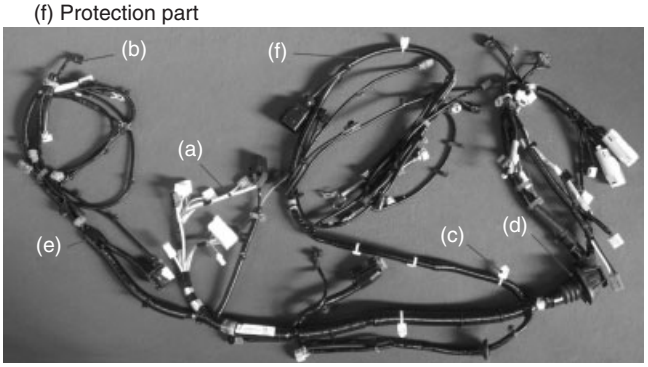


Figure 2. Wire harness components.

Table 1. Wire harness components, a summary.

Item	Component	Summary
a	Wire	The wire transmits the power and signal and it consists of a metal conductor, such as copper, and insulation materials made of plastics such as PVC.
b	Connector	The connector connects the wire harness to equipment or wire harness to wire harness. It consists of the copper terminal attached to the wire end and resin housing.
c	Clamps	Clamps are used to fix the wire harness on to the vehicle body.
d	Grommets	Grommets are used when wire harness passes from the inside to the outside of the vehicle body and it protects the wire from physical damage and prevents water penetration.
e	Tapes	Tapes are used to bundle wires and fix components.
f	Protection parts	The protection parts protect electric wires from physical interference from neighboring parts. Corrugated tubes or protectors are used.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto234
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

A detailed description of the components is given in the following.

3 WIRE

3.1 Standard for automotive wire

Automotive wires are used in a variety of harsh conditions and are exposed to vibration, heat, cold, machine oil, and the elements. For this reason, wire specification standards have been established in each country for standard performance, for example, the SAE standards of America, DIN standards of Germany, and the JASO standards of Japan. Automotive manufacturers are obliged to use these wires for wire harnesses or lead wires. These standards for automotive wires define the temperature rating for the environment, conductor size for the allowable current, insulating materials for each wire type, and wire performances for each wire type that must be met through the test methods for wires.

In part because of globalization, the establishment of a unified standard (ISO standard) has been promoted more recently. However, the situation is still a time-consuming process, because each of the automotive manufacturers and markets are still using different standards and specifications of wire suitable to them.

3.2 Conductor material

For automotive wire conductors, annealed copper, bunched with a few or hundreds of drawing wires, is commonly used. Flexible conductors with much finer drawing wires are used the areas such as those experiencing engine vibration or the opening and closing of the doors. Tin-plated conductors are also commonly used in order to obtain high reliability of the connection with terminals or to withstand high temperatures in the engine bay. In this way, various automotive wires corresponding to different requirements have been developed.

3.3 Insulation material

PVC has been commonly used as an insulation material. For wires used in high temperature environments such as the engine bay, heat-resistant insulation materials such as cross-linked PVC and cross-linked PE have been used. In a time of high global environmental consciousness, lead-free insulating material has been promoted by the End-of-Life Vehicles Directive in Europe. Although more advanced halogen-free insulation material has been introduced, PVC is still the major insulation material for automobile wires today.

3.4 Twisted pair wire, shielded wire

Wires emit weak electromagnetic waves during electric current flow. Conversely, noise due to electromagnetic interference is generated by equipment. For this reason, twisted pair wires or shielded wires are used in some cases. Twisted pair wires are twisted at a constant pitch, so the current flows in opposite directions in each wire. As a result, the electromagnetic waves emitted from the wire are canceled. Shielded wires include a shielding layer that shields the electromagnetic wave. Copper wire braid or/and metal foil (for high density shields) are used as shielding layers.

3.5 High voltage cable

Conventionally, the electrical voltage in passenger cars is about 12 V, while it is 24 V for heavy-duty vehicles such as buses and trucks. In recent years, voltage and current in vehicles has tended to rise with the introduction of electric vehicles powered by motor only and hybrid vehicles powered by a combination of engine and electric motor. As a result, special automotive wire is needed for high voltage and high current circuits. The insulation material for such cables is required to withstand high voltage and high heat resistance due to heat generated by the current. In addition, a shielding layer such as braid shield is required for shielding the electromagnetic waves emitted by high current energizing. From a safety standpoint, the insulation color of this cable is orange for easy recognition from other wires in case of an urgent check-up or maintenance.

3.6 Future vision for wires

With high global environmental consciousness, a need for lighter vehicles has emerged in the market. With this trend, smaller and lighter automobile wires are required. For this purpose, ultrathin wall wires with compressed conductors and thinly coated insulation have been introduced. In addition, the development of much smaller size wires than those available at present is in progress. In addition, aluminum wire whose conductor is aluminum alloy, which is lighter than copper wire, has been developed.

4 CONNECTOR

The connector, composed of a terminal and housing, is a component for transmitting electrical current between

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto234
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

devices or wire harnesses. The "terminal" is electrically connected to the wires and equipment. The "housing" has the mechanical connectivity features of insulation between terminals, terminal retention, and locking ability.

The male terminal structure is shown in Figure 3 and female terminal structure in Figures 4 and 5. A terminal is composed of a contact section (male: tab, female: box) for mating, a wire barrel section for connection with wires, and an insulation barrel section for holding wire insulation.

A spring blade is placed in the female terminal and contact pressure is created between the male tabs and the spring blade. This type of connection is widely used. Another type of connection, the Faston type, is composed of a male terminal with a hole and female terminal with a dimple corresponding to the hole in the male terminal. For high current applications, multicontact terminals have been also used.

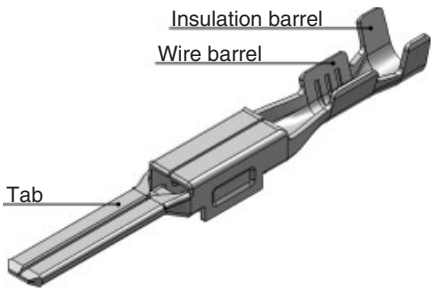


Figure 3. Structure of the male terminal.

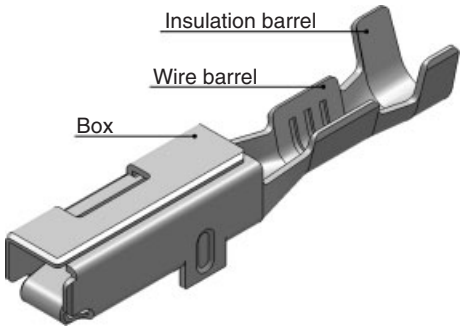


Figure 4. Structure of the female terminal.

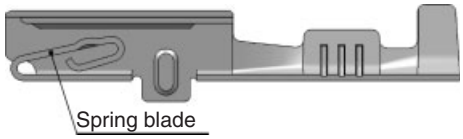


Figure 5. Internal structure of the female terminal.

Terminal material is copper alloy such as brass, with Sn, Ag, and Au plating for stable electrical contact. Crimped connections with the wire have been widely used, while other connections such as insulation displacement or ultrasonic welding have also been used.

The housing is composed of a cavity to accommodate the terminals, a lance to hold the terminals, and a retainer to strengthen the hold of the terminals and to prevent half mating. The structure of the cavity is shown in Figure 6 and the structure of the retainer is shown in Figure 7. When the terminals are inserted, the retainer is in its released position. After all the terminals are inserted, the retainer is fitted in the locked position. In a defective insertion state, the retainer doe not get fitted in the locked position. Owing

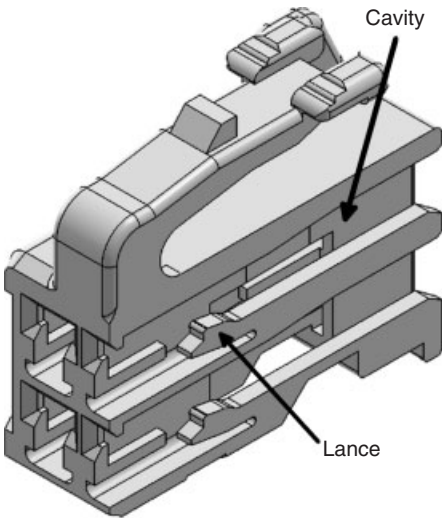


Figure 6. Structure of the cavity.

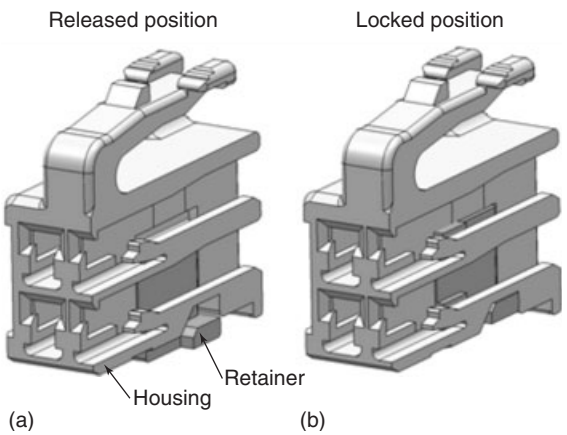


Figure 7. Structure of the retainer. (a) Released position. (b) Locked position.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto234
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

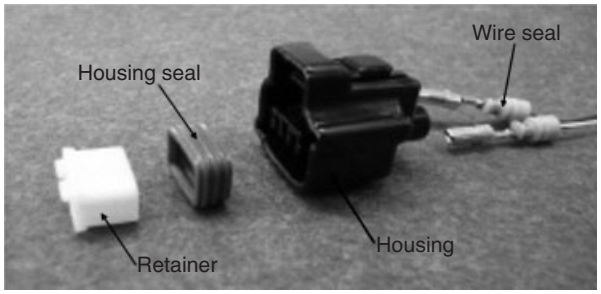


Figure 8. Structure of a waterproof connector.

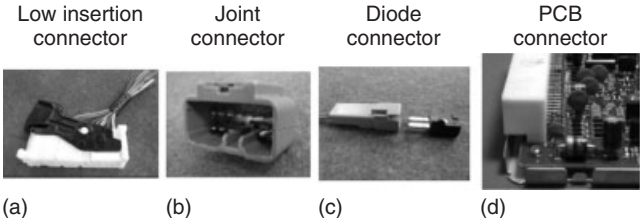


Figure 9. (a–d) Types of automotive connectors.

to the retainer holding the terminals in the locked position, the retention force is stronger than when a lance alone is used. The structure of a waterproof connector is shown in Figure 8. The waterproof connector is composed of a housing seal for waterproofing between housings and a wire seal for waterproofing of the wires.

Polybutylene terephthalate (PBT) has been commonly used as housing material, while PBT filled with fiber glass or polyamide (PA) has been used for heat-resistant applications. Silicon rubber has been commonly used as a seal material, while nitrile rubber or acrylic rubber has been used for oil-resistant applications.

Automotive connectors must be able to maintain stable connections under harsh automotive environments, so they need to have heat resistance, vibration resistance, and water resistance. Owing to enlarged wire harnesses corresponding to the recent introduction of advanced electronics into vehicles, the multipolarization and miniaturization of connectors are required. In addition to the conventional low voltage connectors for HEV and EV, connectors that withstand high current and high voltage are required. For this application, connectors with high dielectric strength insulation and shield for noise reduction are needed.

The following are features of automotive connectors:

1. ease of connection and disconnection in a single operation of the locking mechanism;
2. waterproof type for harsh environment applications;
3. various types of terminals for high or low current circuit, including the hybrid type (using two or more terminal types);
4. having a retainer for strengthening terminal retention and preventing half mating (particularly for small terminals).

Four types of automotive connectors are shown in Figure 9, including the nonwaterproof and the waterproof types mentioned above. For mating, connectors are categorized into the following types: the wire-to-wire connecting type (male and female type), the direct connecting to the equipment type (females only), and the type mounted on a printed circuit board (PCB connectors). Other features of these types are the following: (i) the type that has a lever and slide for reduction of high insertion forces induced by multiple terminals, (ii) the joint connector (J/C) type for connecting multiple circuits with the bus bar type male terminal, and (iii) the type with built-in diodes for preventing circuit bypassing. In addition to the examples shown in Figure 9, round terminals, fastened with nuts and bolts, are used for starter motor, alternator, and ground.

5 POWER SUPPLY BLOCK

The J/B connected with electronic equipment through the wire harness, is an electric box that makes intensive connections of circuits, and has incorporated circuit components such as fuses and relays. J/Bs are mainly installed in the engine bay or in interior near the instrument panel. In many applications, a J/B is incorporated with an ECU, especially large-scale ECUs corresponding to the introduction of advanced electronics.

Among electric connection boxes, a box with built-in relays is called a *R/B* and the box with built-in fuses is called the *fuse block* (*F/B*). In principle, these boxes usually do not have the intensive connection functions of circuits. Parts with no power supply function and with only connecting functions are called *J/C*. The J/B has functions such as joint integration and circuit reduction, power distribution, and circuit component integration. The J/B has been effective in downsizing, weight saving, cost reduction, and installation improvement of large wire harnesses induced by the adoption of advanced electronics.

Examples of internal wiring of J/Bs for various applications are shown in Figure 10 and the suitable type is selected for any particular application.

5.1 Wire connection type

In this type, wires are inserted with the terminals into a case, and, owing to easy drainage, the case is usually installed

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto234
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

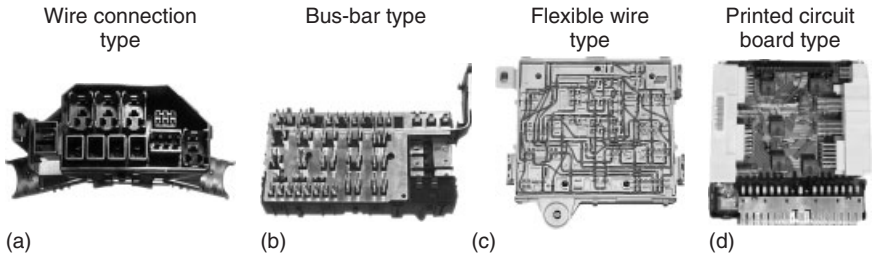


Figure 10. (a–d) Types of internal wiring system of junction blocks.

in the engine bay. Although production equipment costs are low and part costs are also low, the installation costs of wire harnesses are high. Moreover, it is unsuitable for connection with ECUs.

5.2 Bus bar type

For internal circuits of this type, copper strips are pressed and bent into the required form. This has been one of the main types since the introduction of J/B. Among this type of circuits, one consists of laminating the bus bar with plastic plates for complex circuit patterns. The plastic molded bus bars may be used for vibration resistance or waterproofing applications and bus bar integration with a divided bus bar may be used for improvement of bus bar flexibility and reduced tool charges. Although this is a simple structure with low machining costs and ease of high current handling, the costs of the die and press equipment for bus bars are high and their flexibility for circuit changes is low. So, this type is most suitable for mass production.

5.3 Flexible wire type

This type uses electric wires for internal circuit connectors instead of bus bars. In many applications, a hybrid type with wires for low current and bus bars for high current are used. There is also a case where a flexible flat cable is used in addition to an electric wire. This type has higher flexibility for circuit changes than the conventional bus bar type and weight saving is possible. However, it is best suited to small-lot production owing to high wiring equipment costs and machining costs.

5.4 Printed circuit board (PCB) type

The PCB type is one of the main circuit systems. Patterns are printed on the circuit board using an etching process instead of using a bus bar or wires. Compared to the two types mentioned in the preceding sections, the smaller size and weight of the PCB make savings possible and allow the flexibility of circuit change, so design change is comparatively easy.

Owing to the high temperatures induced by high circuit component integration, heat dissipation may pose a problem in many cases. Therefore, countermeasures such as circuit pattern layout design in consideration of heat rejection and cooling may be needed. In the case of the metal-core PCB, which has a thin copper plate embedded in an inner layer, heat generated from circuit parts can be distributed, so this type is suitable for highly integrated circuits and high temperature applications.

6 WIRE HARNESS EXTERIOR PARTS

In order to avoid damage caused by obstacles (such as the edges of the body panel) and harsh environments (such as high temperature or vibration) for wire harnesses, the wire-protecting parts described below are used (Figure 11). For exterior parts, the material is required to be flame retardant, flexible, heat resistant, and abrasion resistant. In some cases, oil resistance and chemical resistance are also required.

6.1 Tapes

Tapes are used to bind wires and fix components. Vinyl tapes are commonly used. In some cases, heat-resistant and flame-retardant vinyl tape or butyl tape with water-proofing and cloth for noise suppression is used. Environmentally friendly halogen-free tapes are also used in other cases.

6.2 Tubes

Tubes are used to protect the wires. Although they offer inferior protection compared to corrugated tubes, plane tubes incur a lower cost than corrugated tubes (the corrugated tube is described in the following.) Vinyl is generally

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto234
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

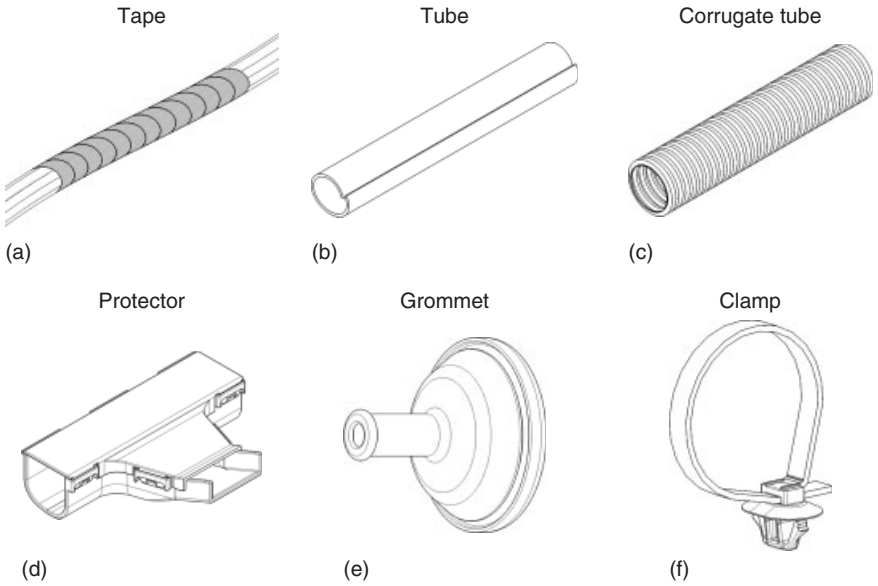


Figure 11. Types of wire harness exterior parts.

used as tube material. In addition, heat-resistant vinyl tubes, slit tubes for easy handling, spiral tubes for high flexibility, aramid cloth for heat-resistance, and twisted tubes for noise reduction are used.

6.3 Corrugated tubes

Corrugated tubes are used to protect the wire harnesses in areas with severe mechanical interference. In addition to polypropylene (PP) corrugated tubes, flame-retardant and heat-resistant corrugated tubes are also used. Slit corrugated tubes are generally used.

6.4 Protectors

Protectors are used to protect the wires and to regulate the routing of the wires. Different shapes are needed for each vehicle and the cost is higher than that of a corrugated tube. PP protectors and heat-resistant protectors are also used.

6.5 Grommets

Grommets are used at pass-through holes to protect wires and to prevent water penetration into the interior.

Ethylene-propylene-diene monomer (EPDM) is commonly used as grommet material. Chloroprene rubber and silicon rubber are used for oil-resistant and heat-resistant applications, respectively. In some cases, a resin inner type grommet is used for easy harness assembly.

6.6 Clamps

Clamps are used to bind and fix wires to the body panel or structure parts for instrument panel reinforcement. Generally, wires are bound with bands and tapes; corrugated tubes are also used. For fixing clamps onto the vehicle body, usually, an anchor-shaped portion is inserted into a hole drilled in the body. In some cases, a stud-shaped portion or plugs are inserted into the supporting bracket.

7 THE FUTURE OF WIRE HARNESS

The wire harness, as part of the infrastructure of automobiles, has achieved comfort, safety, and convenience required in automobiles, with reliable power supply and signal transmission. It has also contributed to the development of the automotive industry An automotive power train shifts electricity (such as on EV and HEV) from an internal-combustion engine; and automotive electrical systems are now evolving at an accelerated rate. The wire harness is necessary for the corresponding high current, large-voltage, and high speed communications. Furthermore, the wire harness is expected the communication and the electric power relation with not only inside of the automobile but outside of it, such as other automobiles, a social infrastructure, and households. The continuous improvements in weight reduction, downsizing, and cost reduction of the wire harness will make it indispensable for continuous evolution of the automotive industry.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto234
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

REFERENCES

Eckermann, E. (2001) *World History of the Automobile*, SAE, Pennsylvania.

Gomes, L. and Fernandes, M.J. (2010) *Behavioral Modeling for Embedded Systems and Technologies: Applications for Design and Implementation*, Idea Group Inc., Pennsylvania.

IRC (2010) *Production and Circulation Research of 200 Items of Automobile parts '10*, IRC, Nagoya.

FURTHER READING

Asaoka, T., *et al*. (2008) *Environmental Measure and Electronization of Automobile*, Techno-Associate, Tokyo.

Iida, H. (2004) *Modern Automotive Terms Explained: Dai-Sharin*, Sanei-Shobo, Tokyo.

IRC (1996) *Production and Circulation Research of 200 Items of Automobile parts 96*, IRC, Nagoya.

Momoda, T. (2012) "16th ITS Strategy of USA." Automotive Technology. January.

Toyota Motor Corporation (n.d.) FCV (Fuel cell vehicle), <http://www.toyota.co.jp/jpn/tech/environment/fcv/> (accessed 20 December 2011).

Toyota Motor Corporation (2011) Sustainability Report 2010 – Energy/Global Warming. Toyota Motor Corporation, <http://www.toyota.co.jp/jp/csr/report/10/download/pdf/sr10_p24_p31.pdf> (accessed 20 December 2011).

Voss, W. (2005) *A Comprehensible Guide to Controller Area Network*, Copperhill Media, New York.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto234
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The Fascination of Car Body Manufacture: Requirements for Car Body Manufacture from Viewpoint of Production

Andreas Kropf

Volkswagen AG, Wolfsburg, Germany

1 Introduction 1
2 A Challenge for the Automobile Industry 1
3 Cross-Brand Standardization in Body Assembly 2
4 Modular Transverse Construction (MTC) 2
5 MTC Modular Strategy 2
6 Interaction of MTC and MPK in Body Assembly 2
7 Challenge of Lightweight Design 3
8 Think Blue. Factory. 3
9 Use of Laser Technology in Body Assembly 4
10 Demographics in Production—Volkswagen Integrated Ergonomic Strategy 4
11 The Volkswagen Way to a Learning Organization 4
12 Summary: Requirements for Body Assembly 4
Further Reading 5

1 INTRODUCTION

Being a globally operating car manufacturer, today Volkswagen Group has over 94 production sites across the world with more than 30 bodywork plants for the Volkswagen brand. In 2011, these bodywork plants produced around 62% (5 million vehicles) of all group vehicles (around 2 million Volkswagens in Europe). Over 800,000 vehicles were built in Wolfsburg alone. Our production turntables and a multibrand strategy at individual sites allow us to react to regional market developments. This requires a high level of flexibility in the body designs and also far-reaching standardization of the production systems and the production processes across the world.

2 A CHALLENGE FOR THE AUTOMOBILE INDUSTRY

The following chapter deals with four aspects from the viewpoint of a globally operating car manufacturer.

Market: A differentiated perception of mobility on the automobile markets sets new challenges for all volume manufacturers not only in the development of their products but also in the design of their production processes and their production systems. Today our customers can choose from over 200 different group models. Modular transverse construction (MTC) and the modular production kit (MPK) help to handle the increasing variant diversity in car production in the economic sense. Building high quality cars and, at the same time, saving resources requires an ideal green factory. For bodywork assembly, this means systematically and consistently using all options to save energy in joining processes and with production equipment.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Volkswagen AG
DOI: 10.1002/9781118354179.auto237
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

CO_2/E-mobility: Reducing the weight of bodies with new lightweight concepts improves the carbon footprint and allows greater ranges for electric vehicle concepts. New material combinations are leading to further development of today's joining processes and joining methods.

Demography: The demographic developments in Europe will change the way we organize our production processes in car factories to a great extent. For example, an integrated ergonomic strategy and supplementary training concepts are also required in body assembly in view of the aging workforce. Our being a learning organization ensures that the expertise of our staff is further developed to the same extent as our products and manufacturing technology.

3 CROSS-BRAND STANDARDIZATION IN BODY ASSEMBLY

The answer to the increased complexity is cross-brand standardization of products, processes, and production equipment. MTC for products, the Volkswagen production system (VPS) for manufacturing processes, and the MPK for production equipment provide the framework for this standardization (Figure 1).

4 MODULAR TRANSVERSE CONSTRUCTION (MTC)

Combining fixed and variable dimensions in the vehicle architecture reduces the complexity considerably. In addition to conventional combustion engines, MTC allows all common alternative drives from natural gas and hybrid versions to purely electrical drives to be installed in identical positions (Figure 2).

5 MTC MODULAR STRATEGY

MTC is the logical development of our platform strategy and is complemented further by modular systems, for example, modular longitudinal construction (MLC). MTC consists of seven modular kits: front axles, front ends, front seats, front floors, rear seats, rear ends, and rear axles.

6 INTERACTION OF MTC AND MPK IN BODY ASSEMBLY

The interaction between MTC and MPK is illustrated by the example of side panel manufacture and the group framer. Standardized components such as the two-shell side panel will allow standardized joining sequences and joining processes. In turn, these product or production standards enable use of standardized production equipment such as the group framer. The implementation of the modular strategy therefore not only leads to considerable unit cost reduction and to a reduction of the run-up times for model changes but also ensures the same quality standards across the world.

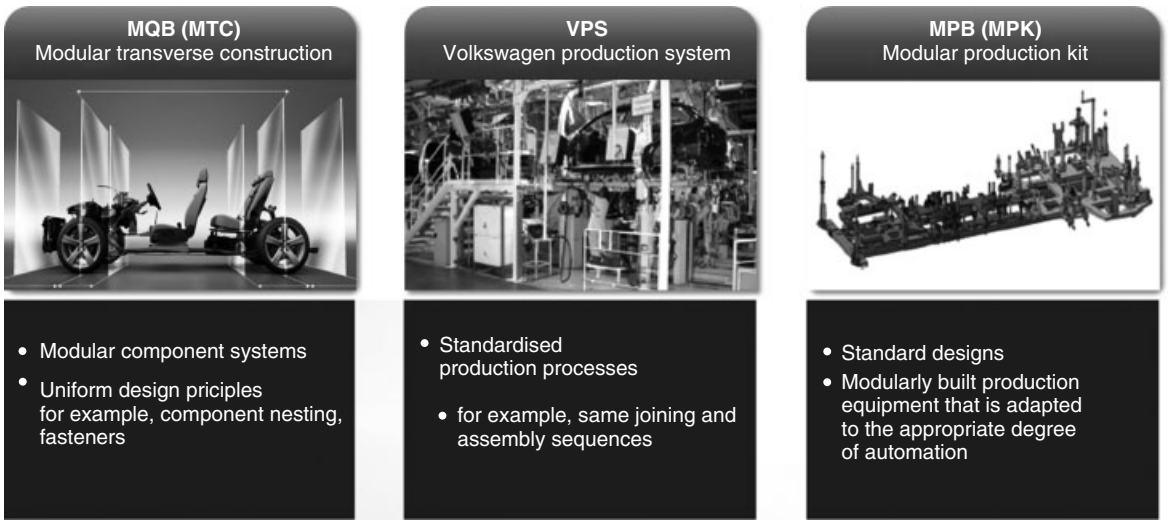


Figure 1. Group-wide standardization in body production. (Reproduced by permission of Volkswagen AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Volkswagen AG
DOI: 10.1002/9781118354179.auto237
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

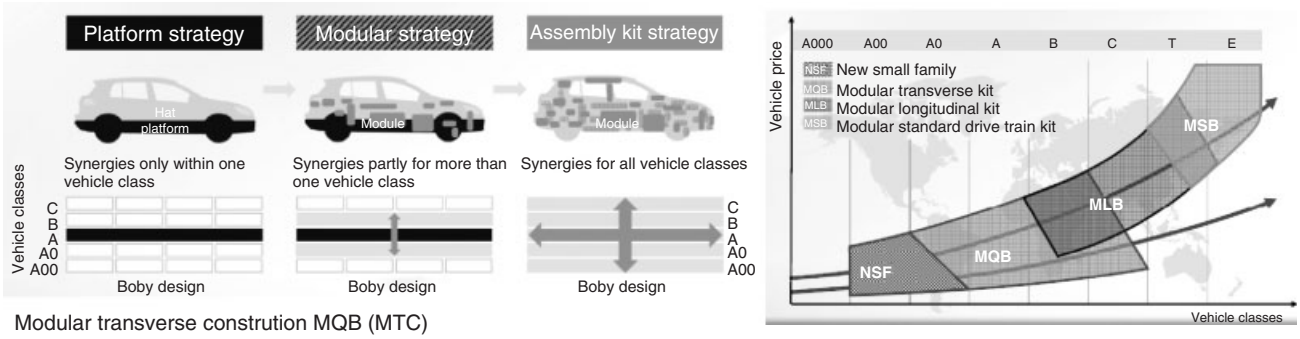


Figure 2. Modular transverse construction MQB (MTC). (Reproduced by permission of Volkswagen AG.)

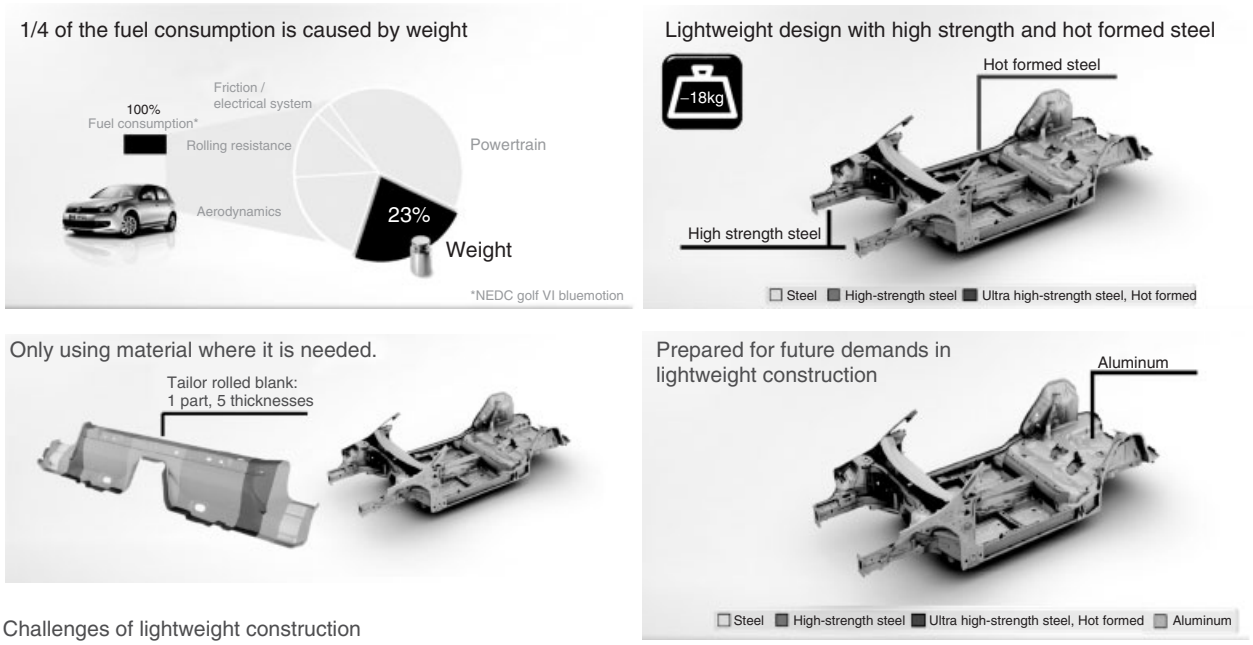


Figure 3. Challenges of lightweight construction. (Reproduced by permission of Volkswagen AG.)

7 CHALLENGE OF LIGHTWEIGHT DESIGN

By combining different materials, lightweight concepts in body assembly make a considerable contribution to reducing the weight of our vehicles. These lightweight concepts lead among other things to further reductions in sheet metal thickness, to an increasing proportion of high strength and ultra-high strength thermoformed steels, or to completely new material combinations, for example, steel and aluminum (Figure 3).

8 THINK BLUE. FACTORY.

The "Think Blue. Factory." scheme sees Volkswagen take on ecological responsibility not just with its particularly low consumption vehicles featuring BlueMotion technology, but also in the production of vehicles. Our factory in Chattanooga represents a milestone on the road to the green factory. It has received LEED® certification in platinum for sustained and environmentally friendly construction. With the Energy Way in Wolfsburg, we are using best-practice examples to show what savings potential is already

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Volkswagen AG
DOI: 10.1002/9781118354179.auto237
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

possible today in different areas of traditional factories. Our aim: Volkswagen production should become 25% greener by 2018. "Think Blue. Factory." in Body Assembly means putting all processes to the test also in terms of a complete energy approach in order to secure further savings potential. Furthermore, a complete approach concerning the question of energy generation has begun in Wolfsburg. Our production halls are equipped with a photovoltaic system covering 40,000 m^2 of roof area including the roof of the Touran body assembly plant in hall 10. The latest generation of robots with energy-optimized drives are used with the new vehicles. There is also potential for optimization in conveyor technology, joining technology, and joining processes. The use of the latest diode laser technology can, for example, improve the energy balance of laser welding by a factor of 10. Remote laser technology allows greater joining speeds. This reduces the required number of joining stations.

9 USE OF LASER TECHNOLOGY IN BODY ASSEMBLY

The laser-welding clamp allows laser welding without a laser cabin. As a result, investments and energy costs resulting from operation of the laser cabin extraction system are eliminated and space requirement is reduced considerably. The laser-welding clamp can be integrated at a later stage into existing spot-welding lines and thus increase flexibility in body assembly. In the future, the use of laser technology will be concentrated even more on the areas with direct benefits for customers. This is why, for example, the laser-soldered roof seam is a characteristic design element of our vehicles. The laser-welded joins in the area of the door entries increase ease of entry for our customers as a result of narrower flanges and improve the crash behavior.

10 DEMOGRAPHICS IN PRODUCTION—VOLKSWAGEN INTEGRATED ERGONOMIC STRATEGY

Demographic developments in Europe will change the way we organize our production processes in automobile factories for at least as long as the aforementioned technical and ecological changes are in place. The workforce that will build the Golf in 2018 will be in the work process almost 10 years longer than the team for the Golf IV. For body assembly, as for all other areas of production, this calls for a further reduction in the burden of industrial work by means of an integrated ergonomic strategy. This integrated ergonomic strategy sees Volkswagen tackle three points: people, technology, and organization. In body assembly, this means ergonomic design of all workstations starting with provision of materials and insertion windows to the use of assembly aids for fitting doors and lids and the use of man movers in the finish areas. A further point is the decoupling of man and machine by using insertion concepts that are independent of cycles. This relieves staff, enables the integration of indirect activities into the work procedure, and generally increases productivity.

11 THE VOLKSWAGEN WAY TO A LEARNING ORGANIZATION

As a result of demographic developments, new training strategies are also necessary in body assembly. As a learning organization we ensure that the expertise of our staff is further developed to the same extent as our products and the manufacturing technology. For staff in body assembly this is done, for example, on-site in our learning workshops. In addition to imparting basic skills in real work situations, professional training courses are held. The learning workshop concept is part of the Volkswagen Way. The Volkswagen Way is an extensive organization development process and a permanent feature of company strategy. In addition to the VPS, it covers the basic elements of demographics, learning organization, communication, and management and thus contributes to the success cycle of the 2018 strategy. Only an attractive employer can take good players on board and turn them into a top team. Only a top team can give an excellent performance.

12 SUMMARY: REQUIREMENTS FOR BODY ASSEMBLY

Volume manufacturers such as Volkswagen face the challenge of simultaneously building different body structures with the greatest possible flexibility in existing and future body designs, thus creating additional growth potential across the world and even in niche markets. Volkswagen is responding to this challenge with a cross-brand modular strategy—MTC for products and the MPK for production equipment (Figure 4). New innovative joining methods are the key to future lightweight design concepts in body assembly. A second key is the consistent implementation of sustained, that is, resource-saving production in all areas of vehicle production. "Think Blue. Factory." combines economic and ecological goals in the factory. Despite all

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Volkswagen AG
DOI: 10.1002/9781118354179.auto237
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

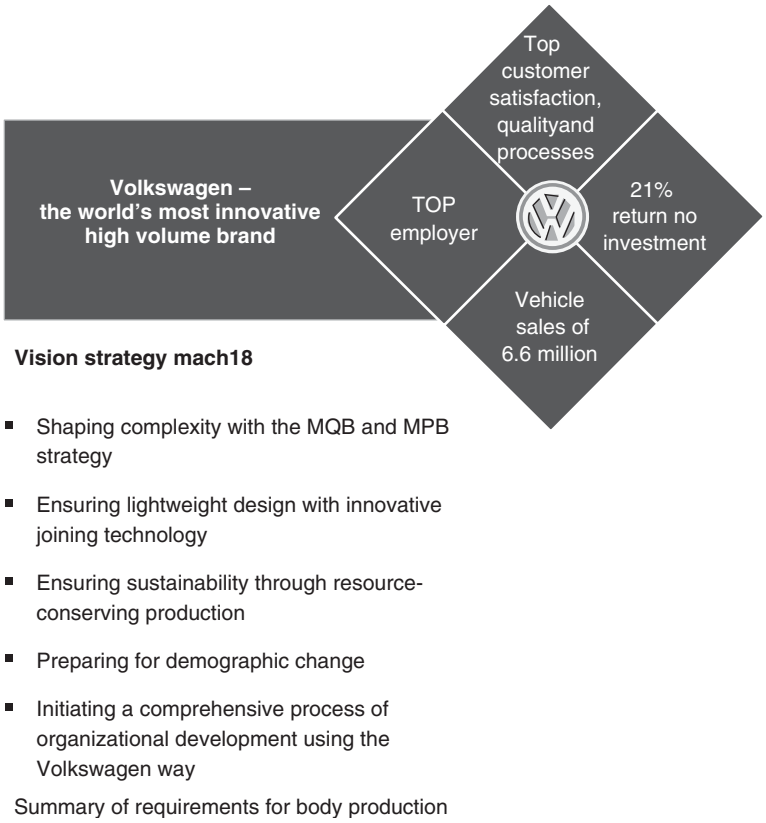


Figure 4. Summary of requirements for body production. (Reproduced by permission of Volkswagen AG.)

this, the complexity of future production procedures in body assembly will also remain manageable for staff as regards demographic aspects. For Volkswagen as a volume producer, the Volkswagen Way will provide the necessary method, building blocks, and elements for this extensive change process. Together with the aims of the 2018 strategy, the areas of action for body assembly on the way to its becoming the most innovative volume manufacturer are clearly described.

FURTHER READING

J. Hillmann, H. de Boer, B. Lohmann, Kostenoptimierter Stahlleichtbau für Karosserien in Großserienfertigung, 13. Dresdner Leichtbausymposium 18. - 19. Juni 2009.

A. Stalmann, Herausforderungen des Leichtbaus für Fertigungsprozesse der Großserie, 5. Symposium Faszination Karosserie und Fahrzeugkonzepte, 13./14.03.2012, Wolfsburg.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 Volkswagen AG
DOI: 10.1002/9781118354179.auto237
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Styling of Cars, from Sketch to Realization, Main Trends and Milestones

David R. Brisbourne and Brian A. Clough

Coventry University, Coventry, UK

1 Introduction 1
2 Design Process Overview 2
3 Design Definition Stage 2
4 Design Creation Stage 3
5 Design Delivery Stage 9
6 Delivery to Market (Manufacturing Engineering) 12
7 Summary 12
Appendix: A—Graphical Representation of Design Process and Gateways 12
Related Articles 13
References 13
Further Reading 13
Online Resources 13

1 INTRODUCTION

Today, all motor cars perform reliably and efficiently so that the economy and performance metrics across a particular class of vehicle are now very similar from one manufacturer to another. Large global car manufacturers who control a number of brands even use a platform-sharing strategy to reduce development costs. The VW Group remains one of the greatest exponents of platform sharing and its stable includes brands that are considered conservative, sporty, mainstream, and premium. It has several platforms for different sizes of vehicle. Versions of the VW "Group A" platform (PQ34/PQ35) for compact and midsize cars have underpinned 15 very distinctive products based on the same C-segment platform (Škoda Octavia, Skoda Yeti, Skoda Superb, SEAT León, SEAT Toledo, SEAT Altea, VW New Beetle, VW Golf, VW Jetta, VW Eos, VW Tiguan, VW Touran, VW Scirocco, Audi A3, and the Audi TT). In this situation, it is even more important that the models are different from each other in the minds of the consumer, particularly if there is considerable price differentiation between badges.

Brand image and visual design have become critical product differentiators. Buyers may be fully aware that the current Skoda Fabia is underpinned by the same platform as the VW Polo and Audi A1 but the cars will be seen in a very distinct brand pecking order. Some badges can command a higher price and it is vital that the design reflects this. To sell in high enough numbers to recover the investment and return a profit, the car must be desirable and meet the needs and expectations of its intended buyers (who are also free to shop elsewhere) and so the role of the automotive designer in the success of a new car should not be underestimated. Advertisers are able tap into the passions and desires of car buyers with greater emotional effect than for any other product.

Image alone is not enough without substance, and as inferred above, all of the various parties involved in developing a new car must work closely to ensure that all the important criteria are met in each specialist discipline. Visual surfaces and functional requirements are inextricably linked and the styled surfaces of car exterior and interior have structural or operational tasks to perform. Although design is constrained by many ergonomic and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

engineering "hard points," the top 20–25 mm of the visible surfaces belong to the automotive designer during the design creation stage up until design freeze. Thereafter, styling-generated surface data drives many of the body engineering and trim design functions working inward from the surfaces. During the design delivery stage, after design freeze, further changes in the styled surfaces can only be initiated by engineering requirements.

However, designers stay involved in the development of a new car from the first sketches to the first car produced, and may even be involved in marketing activities extending to point of sale and merchandise design. This is to ensure consistency of brand image and to ensure that the design intent is maintained through to final production.

Different manufacturers operate on shorter or longer timescales and have their own terminology for the deliverables and gateways. This chapter investigates the detailed process and key milestones in designing a car from first styling sketches through to realization in the context of the process employed by many mainstream manufacturers (Appendix 1).

2 DESIGN PROCESS OVERVIEW

The process of bringing a new vehicle to market is complex and extremely costly and involves many stakeholders, including separate design and engineering departments within a company as well as external suppliers (described as original equipment manufacturers or "OEMs") not to mention the very buyers whose needs will have led to the product being earmarked for production in the first place. For a car company to remain competitive, the process must not only be efficient and timely but also the product itself must be desirable and targeted carefully at the intended buyers.

Typically 36 to 40 months in duration, a new vehicle programme has become more synchronised due to technological developments in Computer Aided Design and Computer Aided Manufacture and in the visualisation of 3D virtual styling and engineering models. Data from packaging, drivetrain, and chassis engineers can be used by styling to create extremely precise armatures to support physical (clay) and digital (CAD) models. In turn, surface data captured from styling models can then be quickly utilized by body, trim, and drivetrain engineers in a carefully choreographed sequence of concurrent activity.

Different companies have their own particular approaches to vehicle development programs but common to all is the use of "gateways." These gateways are effectively checkpoints at the end of key stages of activity where specific "deliverables" must be completed before progression to the next stage of development.

Design/styling departments may have involvement throughout the development of a new vehicle from the earliest exploration of new ideas and concepts carried out in the, usually separate, advanced design studio through every stage of design for production even up to point of sale. At all stages, there is pressure to maintain the original design intent. Typically, the production styling studio will operate from program start through to final production approval when it becomes the main responsibility of the engineering and manufacturing departments to deliver the car to market. In reality, designers remain engaged with the program right up to launch to deal with design-related issues and they often also take part in presenting the vehicle to dealer principals, press, and customers via dealer networks or launch events.

The remainder of this chapter explains the stages of the design process in detail and indicates the gateways that must be passed at the end of each stage.

3 DESIGN DEFINITION STAGE

3.1 Advanced design and concept definition

The advanced design and concept definition stage allows a company to explore new potential product concepts, emerging trends, and potential applications of new technology outside its main business activity and without the pressure of clearly defined timescales. This stage includes inputs from other parts of the business, including transformation and strategic projects (which looks beyond just the product to strategic issues such as the company's integrated environmental policies or manufacturing strategy, new processes, and perhaps new manufacturing or market locations), business plan (to determine if the long-term aim is to build on an existing strategy or to lead with a new riskier product), and cycle plan (which takes an overview of the current availability of research and development, design, engineering, and manufacturing resources—not all programs can be accommodated at once and to the same scale, so the overall plan of products needs to be considered), and turns it into content for programme teams.

Few designs originate directly from the design studio. Within a company, there are many factors that influence the decision to proceed with a given project. These can include the state of the market or economy generally, a company's resources, or even emerging technologies and trends. Sometimes, a company may wish to explore

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

design concepts that are unrelated to its current products or business strategy. A group of executives who support a particular strategic direction may call upon the services of design to create the most convincing artwork, scale models, and prototypes to sell a concept to the main board of directors or even to the buying public. Concept cars may be produced which, although costing millions, are a low cost way to test the market response to new ideas before risking the huge investment of a production programme.

Usually, however, the advanced design team will be engaged in a planned program of strategic design activity, working with a multidisciplinary team that may include product planning, marketing (evaluating customer requirements), engineering, purchasing, and external suppliers. For the program to be approved for production design, a business case will have to be formed that predicts a high level of success. Sometimes, a project may not get approval because it is "ahead of its time," or the company may simply not have the production capacity, or it may not be considered viable owing to predictions of low demand or high build costs. Very few, if any, companies are now willing to take risks on a product that cannot pay for itself and in order to proceed further, a concept must meet the criteria of all the stakeholders. Some products may take a number of years for the conditions to be right, while others may never see the light of day; and there are many concepts that spend time in the press spotlight at motor shows but go on to be stillborn.

Concept exploration is a carefully managed process that involves all the various disciplines up until a gateway known as *final check* (*FC*). This is where approval to proceed with a new concept is sought at a management review. If approval is denied, the team may be sent away at this point to go around the concept exploration process loop again. Between final check and program start, a number of management reviews will take place to review the program parameters and to assess its viability. Once all parameters are met and management reviews are completed, the scope of the program and its scalability (a measure of the complexity and resource demands of a programme) will be agreed ready to progress to programme start. The two key gateways at the end of concept definition stage are

FC—final check
Project teams working on concept development prepare concept presentations.
This may take the form of 2D artwork and 3D models (may even include driveable concept vehicle).
Approval to proceed is sought at management review.
Program parameters are assessed for viability.
The scope of the programme and "scalability" are agreed on.

PS—programme start
If all areas agree the business case is viable, approval will be granted for programme start.

4 DESIGN CREATION STAGE

4.1 Program approval/programme start

For an advanced concept to gain approval for a production, programme start, the business case criteria must met for all the stakeholders of marketing (from customer research), product planning, design, purchase, engineering, manufacturing, and external OEM suppliers. If all areas agree the business case is viable, the project starts.

A brief is generated, which defines the new vehicle architecture and the place of the new product within the business strategy and future plans. The type of vehicle, target customer group, manufacturing materials and processes, drivetrain, platform, production volumes, and pricing are established. The new product's function is planned and confirmed and the scopes of themes and concepts that will be explored are put into place. Approval to proceed is then granted and the design process begins

4.2 Theme development—interior and exterior design

Design is not simply a question of sketching ideas randomly. The design team will initially gather information about competitors in a process known as *benchmarking*. There are many ways to present this information beginning with "product positioning" diagrams which, in graphical form, place vehicles broadly against two or more axes representing specific design criteria (Figure 1). The advantage of this method is the ability to quickly identify "white space" indicating there may be a market niche for a new product. The features and specifications of vehicles closest to the design target can then be compared in more detailed tables (Figure 2).

Mood boards and scenario boards may also be generated to inspire and guide concept development by identifying a theme or topic relevant to future customers (Figure 3).

Visual mood boards may include words and images that evoke the identity a brand wishes to project or provide the guiding theme of a new product (e.g., "sporting luxury" or "comfort and connectivity"). Future scenario boards may include content relating to considerations such as green issues, economy, recycling, premium quality, technology, communication, information technology, luxury, comfort, youth culture, targeted generation groups (x and y, baby boomers, silver surfers), and so on.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

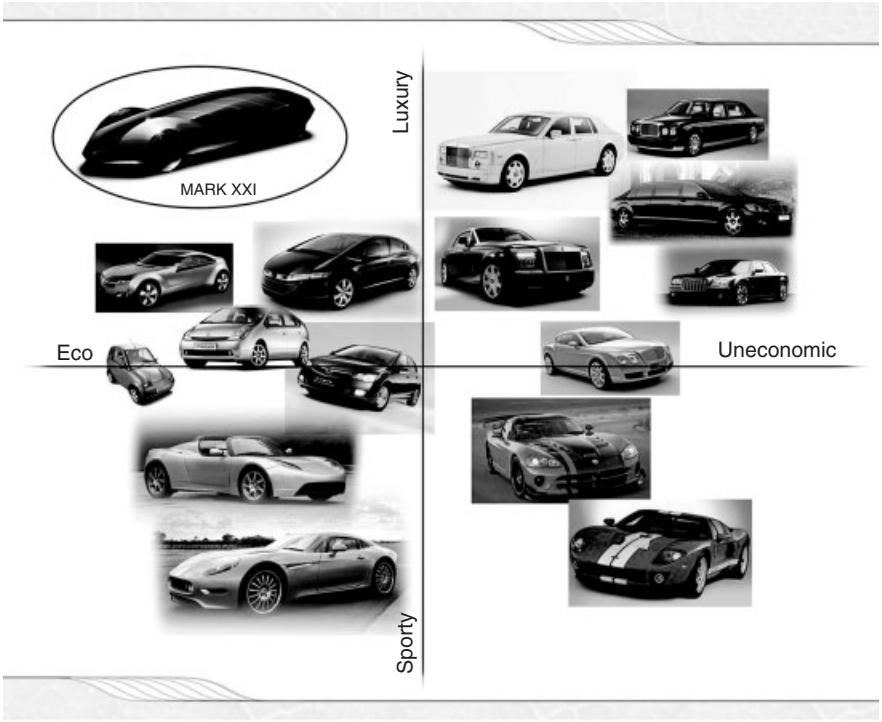


Figure 1. Product positioning diagram. (Reproduced with permission from Christopher Pollard. © Christopher Pollard.)

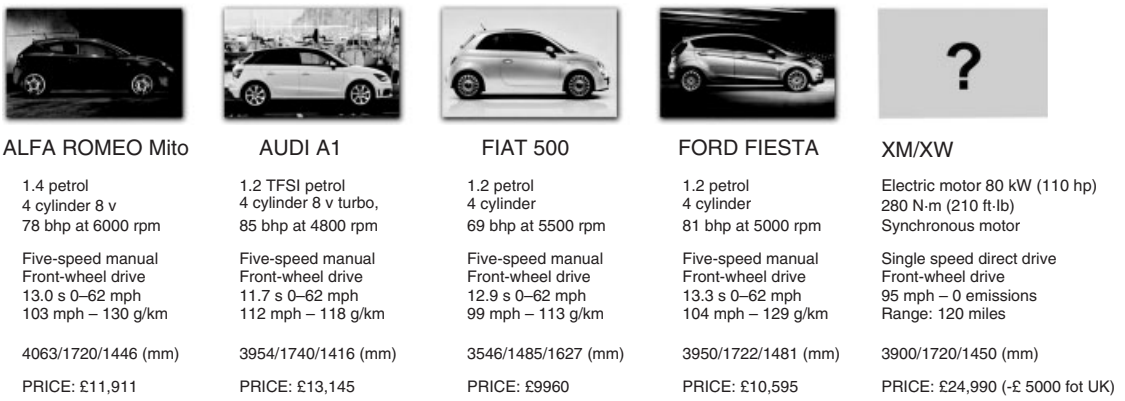


Figure 2. Detailed benchmarking comparison table. (Reproduced with permission from Daniel Chindris. © Daniel Chindris.)

Increasingly, design teams will create detailed "personas" to represent the target users of a new vehicle. On the basis of user research gathered from visits, interviews, and surveys with real potential users, these personas amalgamate various physical and cognitive traits and tastes of the collected behavioral groups into archetypes. Personas boards (Figure 4) then show the persona in pictures and words along with evidence of their lifestyle, places where they live and work, their leisure activities, personal image, preferred brands, and so on. Personas effectively become "real" consumers and allow designers to understand with some clarity who they are designing for. Well-defined personas become a reference tool during the design work.

Two-dimensional sketching, which, of course, starts straight away, is a quick and efficient way to generate a number of ideas. The process is also motivated by competition because selected themes have the potential to go forward to future stages led by the designers who create them. This encourages creativity and productivity. As designers start to explore potential themes and endeavor

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 3. Visual mood board combining keywords and images. (Reproduced with permission from Iain McShane. © Iain McShane.)

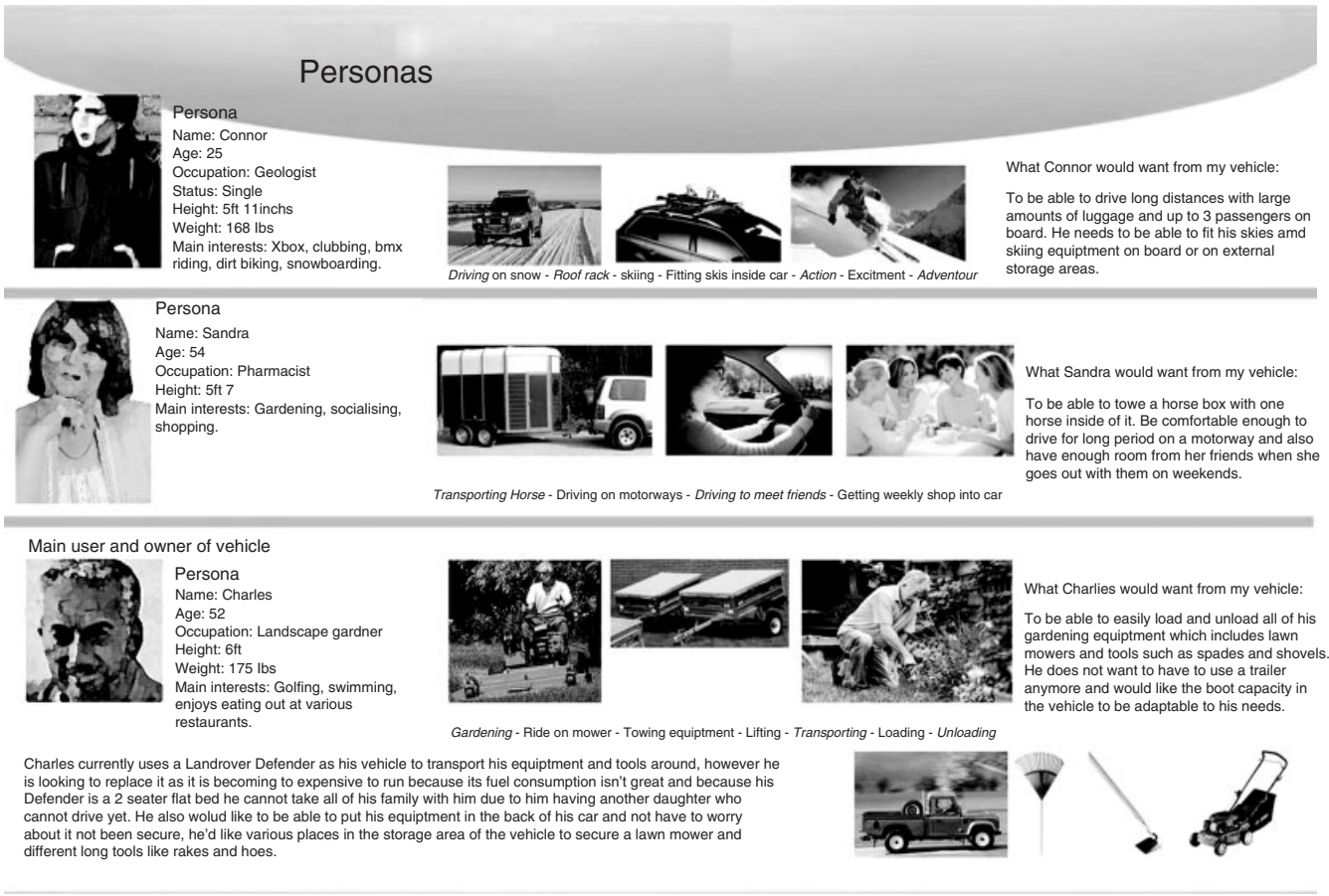


Figure 4. An example of a persona board. (Reproduced with permission from Alix Dobson. © Alix Dobson.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to communicate those thoughts to the team and senior management, their 2D sketches are loosely drawn and not too tightly constrained by the package. Their purpose is to capture the emotion and excitement of the new concept. Hence the early sketches and renderings will not have the evidence supporting them to direct or influence the vehicle "hard point" negotiations that take place at this time. At this stage, the research carried out by the designers, marketing teams, and chief program engineers is concerned with appropriateness of the design, its potential for innovation, and general feasibility rather than detailed engineering.

Increasingly 3D modeling software (Autodesk ALIAS) is being used as a design visualization tool during these early stages (most design graduates now have experience of ALIAS model making). Designers who are familiar with ALIAS can "sketch-models" their proposals (Figure 5). Images of the model can then be imported into 2D graphics software (e.g., Adobe Photoshop, Illustrator, or Autodesk Sketchbook Pro) and rendered over. Particularly when used for interiors, this technique provides a strong, and dimensionally accurate, composite sketch capable of describing their thoughts without ambiguity and including ideas about color, materials, and textures.

Initial engineering package hard points may well be in place at this point, especially if the project is a "facelift" of an existing vehicle or so-called "new top hat." The former is a midlife refresh of an existing design that may include new front- and rear-end treatments affecting bumpers and lights. The latter is a program where the chassis and drivetrain are carried over from a previous program/production vehicle but with a completely new body and interior. For a completely new "ground-up" design involving new chassis and drivetrain, an initial concept package will exist from the advanced design and product definition phases. Sometimes, an existing product may be used as the starting point with the designers challenging what they wish to change.

4.3 Multiple theme selection (design shortlist)

Two-dimensional theme sketches are the quickest way to generate ideas and direction for a new design program and often a single "key sketch" is identified, which provides an identity. However, sketches have their limitations as automotive design eventually leads to a three-dimensional outcome. To truly gain confidence in the evolving design directions, 3D physical and digital models are created to properly realize the designer's thoughts. Physical exterior clay models may be produced at scales of 25%, 30%, or 40%. Virtual alias models can be produced for both exterior and interior and viewed at close to full size for evaluation.

Typically, in the large studios of major car companies (e.g., Ford or General Motors), around eight initial exterior themes are developed as scale models (Figure 6), some hand-modeled in clay and some machined from digital data. At the end of this phase, the models are presented painted or "dressed" for selection. The models are developed with skilled physical or digital modelers over 10–12 weeks, the work being done in close conjunction with the studio feasibility engineers and senior design managers.

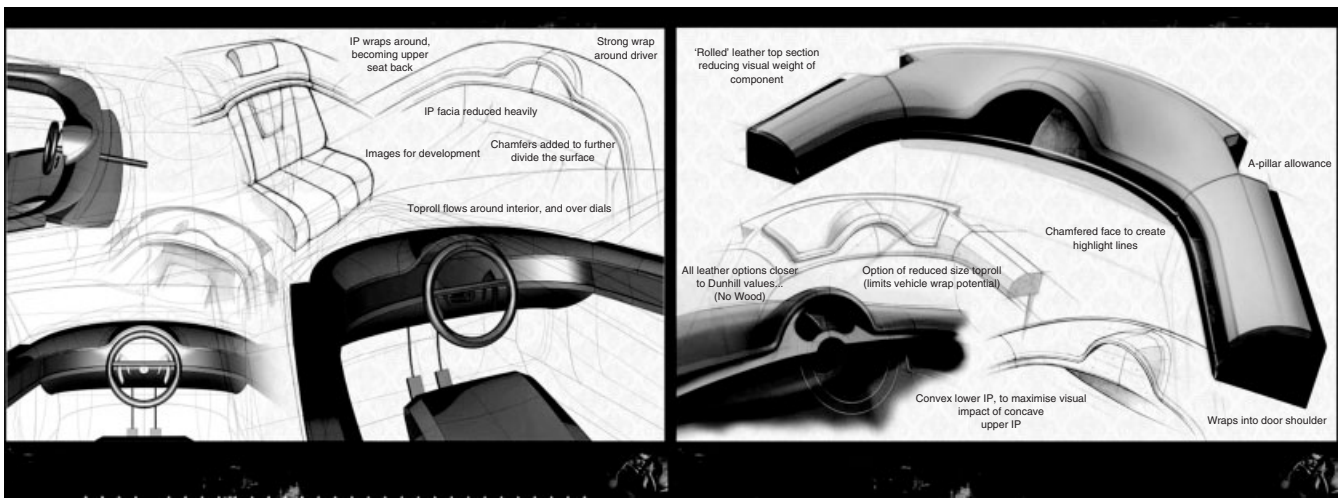


Figure 5. Three-dimensional CAD "sketch" models (Autodesk Alias) rendered in 2D graphics software (Adobe Photoshop). (Reproduced with permission from Ben Quaintance. © Ben Quaintance.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

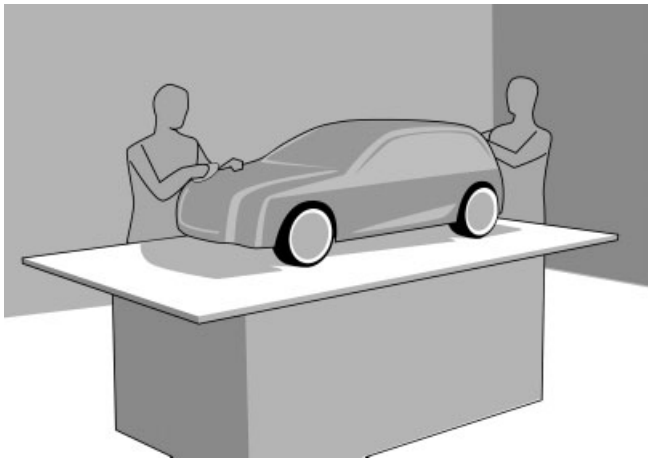


Figure 6. Three-dimensional development using a scale clay model. (Reproduced with permission from Brian Clough. © Brian Clough.)

The eight exterior themes may well be split across four actual "half models" (Figure 7) where each side of the same model will explore a different theme. A designer may develop more than one theme in the program and each will have a modeler to translate the design into 3D.

All hand-worked models will at some point be digitally scanned to capture the surface form, allowing data to be compared with the other themes and help support feasibility engineers in their parallel work streams. Computer visualization software such as "Autodesk Showcase" and "Bunkspeed SHOT" can be used to show material changes and lighting effects on photorealistic digital models in convincingly real environments quickly and with considerable sophistication (Figure 8).

Some digital models can even be animated or "driven" in real time to see how they may look on the road long before a physical prototype exists.

Interior design and model making may begin some weeks after the exterior is underway. Interior models are developed virtually in Alias and reviewed on large screens so that they can be seen close to full size. The data can be exported to ergonomics simulation software such as RAMSIS and evaluated even before a physical seating buck is created. Comfort, ergonomics, and packaging are usually then explored through engineered full-size seating bucks and rigs (Figure 9).

Interior design is more complex because of the number of surfaces and parts that need to be generated in different materials and interior designers must work closely with ergonomists and color and trim designers. Interiors are modeled precisely over the engineering hard point data to give a high level of feasibility so that they will be readily deliverable if selected. A couple of important gateways occur at the end of the design shortlist stage:

PSC—programme strategy confirmed
- Project strategic targets and guidance are put in place.
- The business case is confirmed across marketing, finance, quality, functional targets, hardware selection and programme timing.

FDC—first development competition for design
- Themes are evaluated in line with programme brief—sketches, physical scale models, and virtual models are presented to management.
- A decision is then made to go forward on a reduced number of themes.
- Management decides on the direction for continued development

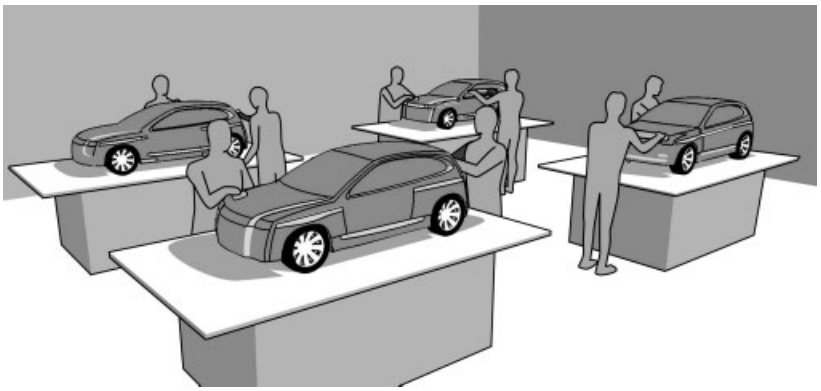


Figure 7. Developing multiple themes using scale-split models. (Reproduced with permission from Brian Clough. © Brian Clough.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

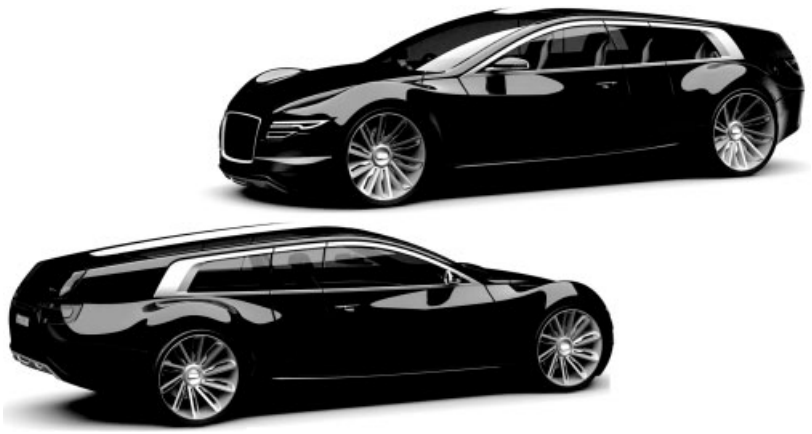


Figure 8. Digital models presented using visualization software. (Reproduced with permission from Phillip Dean. © Phillip Dean.)



Figure 9. Interior seating buck clay model shown with seats partly dressed. (Image used with the kind permission of Jaguar Land Rover 2013.)

4.4 Theme reduction

Following a senior design and programme management review that is supported by color and materials proposals, marketing information, and engineering assumptions, the number of themes is reduced down to four exterior and two or three interior proposals. The four exterior proposals are developed using two more split clay models (different themes on each side), but this time modeled at full size (Figure 10).

Interior design moves forward with full-size clay interior bucks, which over time will be dressed to be a full facsimile of the design intent (Figure 11).

Exterior clay models can be dressed using painted film called *Dynoc* and rolled outside to be viewed in natural light. The theme reduction phase concludes with another important gateway:

PTCC—programme target compatibility checkpoint
Target attributes for the design are agreed and approved.
Systems selection (electrical, powertrain, chassis) is completed and confirmed.

4.5 Design selection (final design selection)

To enter the final phase of creative design development, a single exterior and interior theme is selected at the most important gateway of the design creation stage:

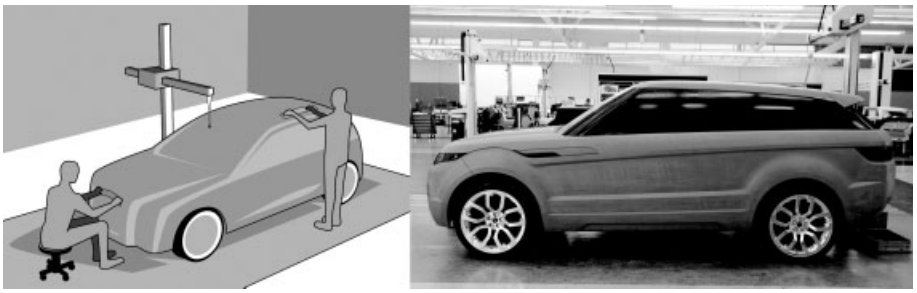
Final design selection "Go for One"
The final exterior and interior design themes are selected to go forward into production, hence "Go for One."

4.6 Single theme

All the concept and design development now comes down to a single full-size clay model of the exterior and interior, which represents all the design development up to this point and confirms that there is now one clear direction approved by the business.

This phase of development focuses on fine-tuning the design details of the full-size models in conjunction with suppliers and engineers, so it is a fully balanced interpretation of what the design team imagined. The end of the single-theme phase also concludes the design creation stage with design freeze or "pencils down." From now on, any detail interior or exterior design changes can only be initiated by "engineering change" documents released in a rigorously controlled manner by the program team. This

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) (b)

Figure 10. (a) Three-dimensional development, full-size clay model. ((a): Reproduced with permission from Brian Clough. © Brian Clough. (b): Image used with the kind permission of Jaguar Land Rover 2013.)



Figure 11. Interior buck fully dressed with representative materials. (Image used with the kind permission of Jaguar Land Rover 2013.)



Figure 12. Fully see-through exterior and interior design Realization model. (Image used with the kind permission of Jaguar Land Rover 2013.)

gateway signals the end of any design changes being initiated by the design (styling) team:

Design freeze (design creation stage "Pencils Down")
Acknowledges the hand over from design creation to design delivery.
The point at which further design changes can only be initiated by engineering

5 DESIGN DELIVERY STAGE

5.1 Microfeasibility

A new car contains thousands of component parts that require detailed design and engineering solutions during the design delivery phase. Component engineers work with design and with external suppliers to confirm specification, quality, material selection, fit and finish, functionality, and durability. During this tightly managed intense period, production dates are on the horizon and decisions critical to success.

"Design Digital Data" which is the digital representation of 3D design surfaces sourced from clay, physical, or virtual models is confirmed and handed over to engineering in preparation for production surfacing. By this phase, the surfaces and features will have design engineering feasibility due to close, ongoing collaboration with both a company's own engineering teams and those of its suppliers. At this stage, the design (clay and digital model) may still be subject to some levels of engineering initiated change before going into full production.

Designers are responsible for continuing the relationship with engineers to ensure the design is developed

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

as intended and to avoid the focus of the theme being diluted. They will add value with every opportunity that presents itself during perceived quality and "optical quality" discussions. Three gateways conclude the microfeasibility phase:

PTC—program target compatibility
  Ongoing program targets are set.
  The business case is updated in line with latest program assumptions.
M1DJ—first drivable prototype (M1) data judgment
  Data readiness for M1 (the first drivable prototype) tooling, test, and build verification
  All underbody components are now produced to the first production build (Job #1) specification intent
DDJ—design data judgment
  Clay model freeze

5.2 Surfacing

The surfacing process is where the creative design surfaces meet engineering feasibility and tolerance of associated component parts. Surfacing as a department takes design digital data and engineering data from a company's own engineering departments and those of external OEM suppliers and develops working surfaces for tooling, including all the relevant requirements from material selection like grain, draught angles, and fit and function.

The forum for this phase is described in many companies as "optical quality" and provides the opportunity for making judgments optically using the digital surface data (typically, ICEM Surf), where parts are "stacked" in a virtual form in a highly precise format. Owing to the accuracy of the data, minute irregularities that may be invisible to the naked eye on the clay can be smoothed out in the software. Joint conditions, material differences, grain finishes, and molding split lines are all areas that are explored in conjunction with their individual tolerances, so designers can see what a customer may experience when the vehicle is built. The aim of this phase is to confirm the design intent virtually so the data is ready to release for initial tooling.

5.3 Feasibility cube

The feasibility cube takes the "surface transfer data" to build a functioning representation of the design data. Surface transfer data come from the styling models developed by design, supported by studio engineering who ensure feasibility, and also with some supplier input. Data are sourced directly from digital styling models or by scanning exterior and interior physical clay models to capture the design intent before obtaining final approval to pass to engineering who will go on to create production surfaces. Detail can change beyond this point but usually for technical reasons. Feasibility cubes are precise physical representations of the exterior and interior styling surfaces milled in sections in hard resins and foam such as "Ureol®," and assembled on aluminum jigs. They are accurate to fractions of a millimeter enabling all interfaces and gaps to be critiqued to ensure they match the styling intent. Assembled feasibility cubes look like real cars with opening doors but there are no telltale welds or clinched joints around door edges (Figure 13). Seats, dashboards, door cards, and other trim items are not padded but machined from hard materials that represent their surfaces as the designers modeled them.

Feasibility cube data are also the basis for the complete exterior/interior see-through models known as *design realization models*, *styling reference models*, or "*prove out*" *models* (Figure 12). Some manufacturers call them "*design experience cars*." To all intents and purposes, these are real cars built very accurately in a variety of appropriate materials and processes (fabricated metal and composites, rapid prototyping using "additive" 3D printing or "subtractive" machining from solid) all fully painted and trimmed in representative finishes. They may sometimes, but not always, be fully drivable. The design realization model is a design property and is used to confirm design intent and to help start to sell a design throughout the business, including the sales councils and dealer networks. They are so convincing that they may even be used in early press release photographs many months before a real production car is available. The feasibility cube phase ends with the first of three appearance approval gateways:

AA1—appearance approval 1 (of 3)
  First of three appearance approvals complete
  Feasibility of interior and exterior themes approved
  Production intent surfaces released
  Upper body mechanical package and software activity started

5.4 Modify surface and engineering

This is the final phase of design "fine-tuning" intended to catch all the snagging issues identified and tracked by the engineering change documents. All surfacing and engineering issues should be resolved and released for

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

final approval at this point. This stage ends with two more gateways including the second appearance approval.

PA—programme approval
- Business case, quality targets, and market equation confirmed
- Ongoing programme objectives approved
- Mass production upfront activities launched
- Upper body mechanical package and software activity completed

AA2—appearance approval 2 (of 3)
- Final exterior and interior design approval
- Design data released

5.5 Design and engineering freeze (design delivery "pencils down")

Design and design engineering now stop.

5.6 Function cube (environmental cube)

The function cube (or tooling reference model) is a master physical representation of a vehicle body machined in aluminum with fixing points to allow body trim and hardware to be fitted (Figure 13). This reference model will be maintained throughout the life cycle of a complete vehicle (to incorporate future design updates/ facelifts etc.). Built to show every aesthetic design surface as it will be released for tooling, the function cube is usually machined from solid aluminum billets, thinly painted and assembled to look like a real if extraordinarily heavy version of the vehicle. Doors open and every component interface will be as described by the "data tree." Since a function cube represents the surfaces exactly as they will look when the real car is assembled from the pressed parts, even the shape of welds and clinched joints will be milled into the surfaces.

Since new cars contain thousands of separate parts that may be changed during the design development, it is essential that a record is maintained of changes to ensure that only the latest version of a component is in use. The data tree records and holds all program components in a virtual 3D state. The complete work history of each individual component is contained here so that all components are the most current.

The data tree also allows virtual 3D models to be built by the various module teams to illustrate how their areas come together and function in isolation or in conjunction with the rest of the vehicle. Virtual reality systems can even employ component data to assess tasks such as serviceability and access to engine components.

The only components that are not created to machining precision tolerances are the glazed areas, typically because on an all-new vehicle they will initially be made from Perspex not glass. The "glass" will include the "blackout"

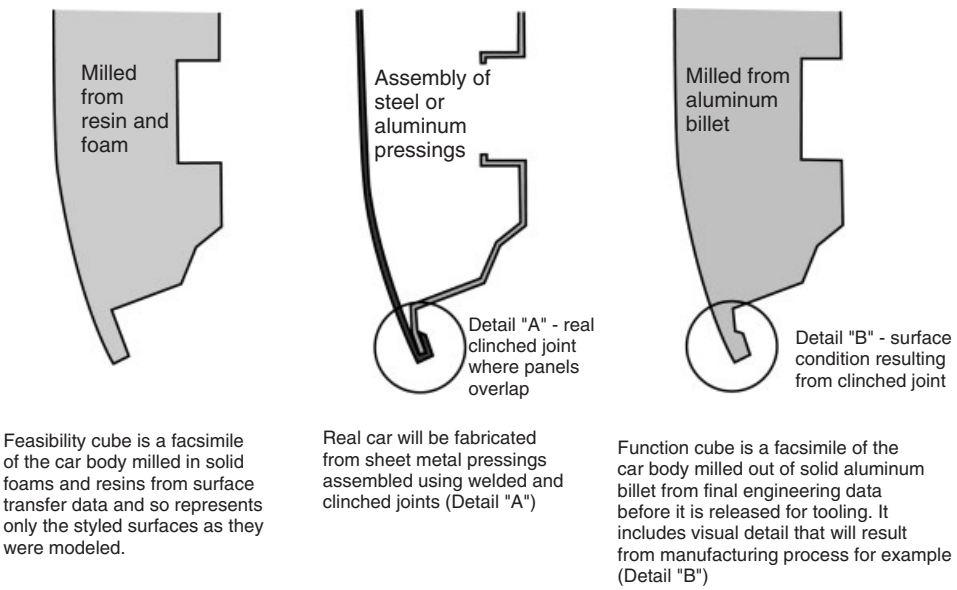


Figure 13. The difference between feasibility cube and function cube explained using the example of a car door bottom edge. (Reproduced with permission from Brian Clough. © Brian Clough.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

obscuration areas where paint or decals hide pillars, adhesive joints, and items mounted to the glass such as interior mirrors. Even the areas that are left unswept by the windscreen wipers will be "frosted" for visibility evaluation purposes.

The purpose of the function cube is for every component area to evaluate critically all the surfaces and how they relate to each other across the whole vehicle. If something is deemed unsatisfactory in a particular area, the module teams responsible often initiate change themselves before tooling. Some items such as bumpers and any metal pressings for the "body in white" are time critical at this point as they are the longest lead items, so any change that is initiated has to proceed through the change and correction process swiftly and efficiently to prevent delay to the programme.

Completion of the function cube signifies the end of the design delivery stage with three final gateways remaining:

M1DC—(M1 = first drivable prototype) development completion
  Verification of upper body systems and subsystems
  Engineering change notices issued
  M1 development completion authorized by engineering management
FAA—final appearance approval (3 of 3)
  Appearance approval completed
  Interior and exterior surface final refinements/highlights approved
  Design data updated
FDJ—final data judgment
  Engineering designs and associated confirmation complete
  Confirmation across all business functions that designs satisfy business requirements
  Outstanding issues have solutions in place meeting measured company standards, and are judged deliverable for future phases

6 DELIVERY TO MARKET (MANUFACTURING ENGINEERING)

6.1 Beyond final data judgment

In theory, designers could let a program go at this point as it now becomes the core responsibility of the engineering and manufacturing departments to deliver the car to market. However, it is rare that the designers cease to be involved through the manufacturing engineering stage.

As an intrinsic part of the program team, the designers, especially color and material designers, are still engaged in making sure that design intent is delivered, supporting program activity until launch where they will often also take part in presenting the vehicle to dealer principals, press, and customers via the dealer network or launch events. Their ongoing responsibilities can include

Generating and illustrating the various trim levels
Working with marketing agencies on artwork and images for communications and launch material
Preparing properties (models, displays, and actual vehicles)
Preparing artwork for press releases detailing the design story
Interviews
Filming for development stories, attending motor shows, vehicle launch events, customer events, to support launch
Being part of snagging teams, problem-solving last minute issues

In effect, designers will be involved from the earliest concept sketches until the first customers take delivery of their new car.

7 SUMMARY

A designer's role is vast and varied and many may only ever work in a specific area during a program; early sketching, or detailed delivery, but those experienced and rounded enough to work from start to finish hold a unique insight into the project they are inextricably linked to. It becomes part of their life and it is like bringing a living thing into the world, complex but wondrous. Designers require a strong empathy with the wider team around them and they are required to lead the design language debate but they also have to reach acceptable compromises in order to deliver a desirable design solution in a very tight time frame and to cost targets.

For a designer there is no greater pride than viewing a piece of design they have played a large or small part in being used by the customer they have designed it for. The lucky ones may even get to own and drive the car that they have helped to design.

APPENDIX: A—GRAPHICAL REPRESENTATION OF DESIGN PROCESS AND GATEWAYS

Figure A.1

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

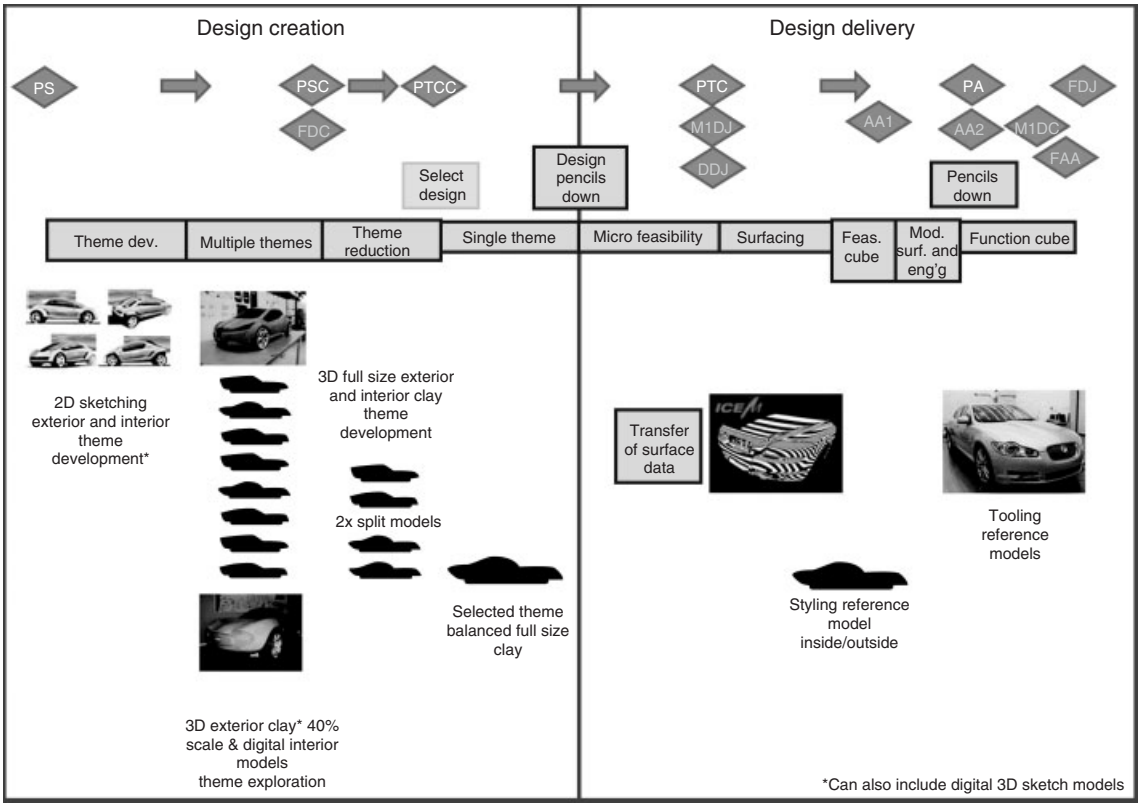


Figure A.1. Design process diagram with gateways. (Reproduced with permission from David Brisbourne. © David Brisbourne.)

RELATED ARTICLES

Body Design, Overview, Targeting a Good Balance Between all Vehicle Functionalities
Innovative Structural Design
The Fascination of Car Body Manufacture: Requirements for Car Body Manufacture from Viewpoint of Production
Fundamentals, Basic Principals in Road Vehicle Aerodynamics & Design
Vehicle Architecture for Meeting Anthropometric, Posture, Comfort, Health Requirements of Passengers
Human Machine Interface Design in Modern Vehicles
Vehicle Seat Design, Development and Manufacturing

REFERENCES

Horbury, P."Foreword" to (2003) *How to Design Cars like a Pro*, 1st edn, Tony Lewin pub, Minneapolis, USA. Autobooks.

Tovey, M. (2012) Designers Role in The Automobile Industry in *Design for Transport – a User Centred Approach to Vehicle Design and Travel* (ed. M. Tovey), Gower, Farnham, England, p. 271.

FURTHER READING

Tovey, M. (2012) *Design for Transport – A User Centred Approach to Vehicle Design and Travel*, Gower.

Macey, S. and Wardle, G. (2009) *H-Point – The Fundamentals of Car Design & Packaging*, Design Studio Press, Culver City CA.

Lewin, T. (2003) *How to Design Cars Like a Pro*, 1st edition, Motorbooks International.

Lewin, T. and Boroff, R. (2010) *How to Design Cars Like a Pro*, 2nd edn, Motorbooks International.

ONLINE RESOURCES

CarDesignNews, http://www.cardesignnews.com/site/home

Car Body Design, http://www.carbodydesign.com

Car Design Online, http://www.cardesignonline.com/

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto238
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Fundamentals, Basic Principles in Road Vehicle Aerodynamics and Design

Simone Sebben, Tim Walker, and Christoffer Landström

Volvo Car Corporation, Gothenburg, Sweden

1 Introduction 1
2 Some Fundamentals of Fluid Mechanics Applied to Vehicles 1
3 Forces and Moments 4
4 Importance of Aerodynamics on Road Vehicles 8
5 Development Process 11
References 16
Further Reading 16

1 INTRODUCTION

Road vehicle aerodynamics considers the interaction between the vehicle, the road, and the surrounding airflow, and the resulting effect on fuel consumption, handling, and cooling performance. Drag and lift are the best-known aspects of aerodynamics. Drag for its contribution to fuel consumption and lift for its influence on vehicle handling. However, aerodynamics influences a large number of other vehicle attributes that contribute to customer satisfaction.

Aerodynamics is an important aspect of today's automotive product development process, and most car manufacturers utilize both wind tunnels and computational fluid dynamics (CFD) for optimizing the aerodynamic properties of their products. This chapter gives an introduction to basic principles of road vehicle aerodynamics including some basic fluid mechanics, the importance of aerodynamics in automotive engineering, and an overview of the aerodynamic design process with focus on passenger cars.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 SOME FUNDAMENTALS OF FLUID MECHANICS APPLIED TO VEHICLES

The aim of this section is to briefly describe some basic fluid dynamic principles and their importance to road vehicle aerodynamics. A more thorough explanation of all the basic definitions can be found in several fundamental books in fluid mechanics, for example, White, 1986; White, 2006; and Panton, 2005; and applied to road vehicle aerodynamics books such as Allen, 1982; Hucho, 1998; Hummel, 1998; and Barnard, 2001.

2.1 Air properties

2.1.1 Density

The density of any material is defined as its mass per unit volume and it is commonly denoted by the symbol ρ. For liquids and gases, this physical property is a function of both pressure, p, and temperature, T. Air, as most other gases, follows the ideal-gas law, stated in Equation 1.

$$\rho = \frac{p}{RT} \tag{1}$$

where R is the gas constant.

Air density decreases with increasing altitude, as air pressure does, and it is inversely proportional to temperature. At sea level ($p = 1$ atm and $T = 288$ K), air has a density of

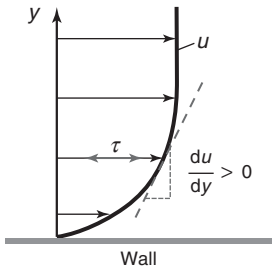


Figure 1. Velocity profile near a wall and the shear stress proportional to the gradient of velocity.

1.225 kg/m^3. Aerodynamic forces acting on ground vehicles are directly proportional to air density. At speeds for which passenger cars operate, variations of air density can be considered negligible and the air can be considered as incompressible.

2.1.2 Viscosity

Viscosity is the mechanical property of a fluid that represents a measure of its resistance to deformation by shear. When a fluid is sheared, it starts to move at a strain that is inversely proportional to its dynamic viscosity μ. The applied shear is also proportional to the velocity gradient. According to Newton, for a shear layer near a wall (Figure 1), the shear force acting on a surface is defined as Equation 2.

$$\tau = \mu \frac{\mathrm{d}u}{\mathrm{d}y} \tag{2}$$

For air at sea level, the value for dynamic viscosity is 1.789×10^{-5} N·s/m^2. As we discuss later, viscosity is the reason for the existence of drag.

2.2 The Bernoulli equation

A fundamental principle to the study of fluid flow past bluff bodies was derived by Daniel Bernoulli based on Newton's second law, and it relates pressure and air speed. Bernoulli's principle states that for an inviscid flow (a flow that has no resistance to shear stress), an increase in the speed of the fluid occurs simultaneously with a decrease in pressure or the fluid's potential energy.

A common form of the Bernoulli equation used by aerodynamicists is seen in Equation 3.

$$p + \frac{1}{2}\rho V^2 = \text{constant} \tag{3}$$

where p is the static pressure and $\frac{1}{2}\rho V^2$ is called the *dynamic pressure*.

The sum of the static pressure and the dynamic pressure is the total pressure, $p_{\text{tot}} = p + \frac{1}{2}\rho V^2$. Equation 3 states that for inviscid flow, the total pressure is constant along a streamline. A streamline is a curve that shows the path that a fluid element will travel in any point in time. Streamlines are always tangent to the velocity vector of the flow, as shown in Figure 2.

For external flow around a vehicle, p and V are taken as the free-stream pressure and the free-stream velocity, respectively. On the road, this means that p is the atmospheric pressure and V is the vehicle speed.

2.3 Stagnation pressure and pressure coefficient

According to the Bernoulli equation, if the flow comes to a rest, the static pressure assumes its highest possible value. Because this maximum value of the pressure is associated with zero-flow velocity, it is called *stagnation pressure* and the position where it occurs is known as the *stagnation region*.

A stagnation region is always present on the nose of the vehicle, where the flow is split over the top and under the bottom of the car. A typical pressure distribution at the center line of a passenger car is seen in Figure 3.

It is convenient in aerodynamics to work with values of pressure, which are nondimensionalized. The advantage

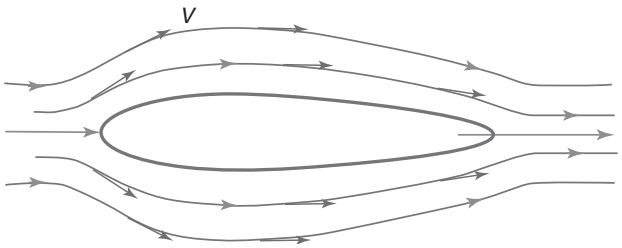


Figure 2. Streamline pattern around an arbitrary body shape immersed on an inviscid flow stream.

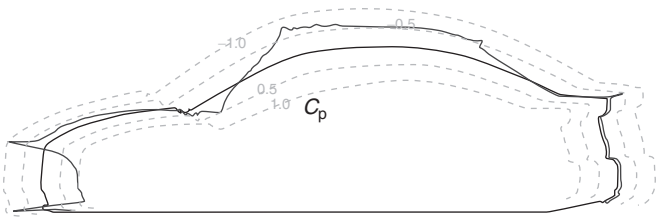


Figure 3. Typical pressure distribution at the symmetry line of a passenger car.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

of working with the nondimensional pressure coefficient rather than with actual pressure values is the fact that the pressure coefficient will not vary with the vehicle speed. A dimensionless pressure coefficient, denoted as Cp, is defined in Equation 4 as the ratio

$$C\mathrm{p} = \frac{p - p_\infty}{\frac{1}{2}\rho V_\infty^2} \tag{4}$$

or

$$\frac{\text{local static pressure} - \text{free stream static pressure}}{\text{free stream dynamic pressure}}$$

where p is the local static pressure and V_∞ and p_∞ the free-stream velocity and free-stream static pressure.

If the Bernoulli equation is applied into Equation 4, we obtain

$$p + \frac{1}{2}\rho u^2 = p_\infty + \frac{1}{2}\rho V_\infty^2$$

or

$$C\mathrm{p} = \frac{p - p_\infty}{\frac{1}{2}\rho V_\infty^2} = 1 - \left(\frac{u}{V_\infty}\right)^2 \tag{5}$$

Equation 5 states that in stagnation regions where the flow field comes to a rest, that is $u = 0$, Cp reaches its maximum value of 1. Equation 5 also yields $C\mathrm{p} = 0$ at the free stream where $u = V_\infty$. Negative values of Cp are also possible and are encountered, for example, in regions where the velocity is higher than the free-stream velocity. On the example given in Figure 3, regions of negative Cp are seen at the leading edge of the bonnet, roof, and base.

2.4 The boundary layer

The boundary layer is defined as the layer of fluid in the immediate vicinity of a bounding surface where the effects of viscosity are significant. The aerodynamic boundary layer was first defined by Ludwig Prandtl. It simplifies the equations of fluid flow by dividing the flow field into two areas: one inside the boundary layer, dominated by viscosity and creating the majority of drag experienced by the body, and the other outside, where viscosity can be neglected without significant effects on the solution. The thickness of the velocity boundary layer, $\delta(x)$, is usually defined as the distance from the body at which the flow velocity is 99% of the free-stream velocity. With increasing distance from the leading edge, the boundary layer gets thicker and less stable and eventually becomes turbulent. Velocity fluctuations are superimposed to the mean velocity leading to more mixing between fast- and slow-moving airs. The result is an increase in the thickness of the boundary layer and a velocity profile that is more flat with higher velocity values close to the walls. This is schematically shown in Figure 4.

Similarly to the representation of the flow on a flat plate, as the air passes a vehicle, the creation of a boundary layer close to the surface of the car also takes place. Within this boundary layer, which is a few centimeters thick at the most, the velocity of the air decreases from its value at the free stream (at the outer edge of the boundary layer) to zero at the surface of the vehicle owing to nonslip conditions. This tiny region close to the vehicle walls has an important influence on the development of the flow around it. At the front part of the vehicle, the airflow is practically steady and with no major perturbations. Downstream the vehicle front end, the flow develops to a turbulent type. In the turbulent region, the flow is unsteady and can contain a few areas with small recirculation zones; nevertheless the flow is mostly attached and streamlined with the shape of the body. At the rear, the flow separates, the boundary layer gets dispersed, and the flow is entirely governed by viscous effects. A representation of the flow around a vehicle can be seen in Figure 5.

2.5 Flow separation and reattachment

As explained previously, all solid objects traveling through a fluid (or alternatively a stationary object exposed to a moving fluid) acquire a boundary layer of fluid around

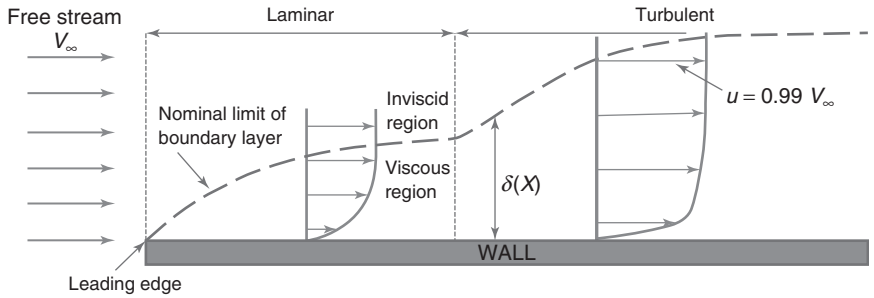


Figure 4. Representation of the boundary layer over a flat plate.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

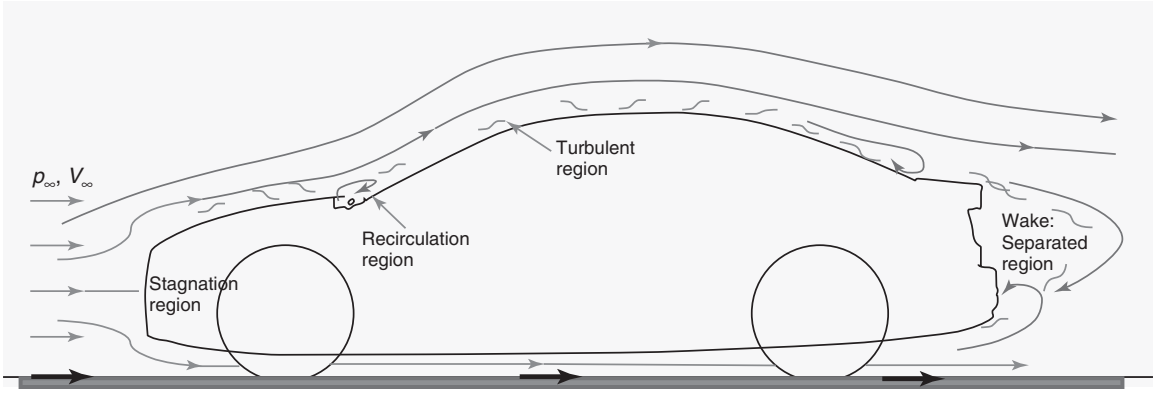


Figure 5. Schematic representation of the flow around a vehicle.

them where the effect of viscous forces is significant. When the boundary layer travels far enough against an adverse pressure gradient (when the static pressure increases in the direction of the flow, $\frac{dp}{dx} > 0$), flow separation occurs. The fluid flow becomes detached from the surface of the object, and takes the form of eddies and vortices. The separation point S is defined as the point between the forward and the backward flows, where the shear stress is zero, $\left(\frac{du}{dn}\right) = 0$ where n is the vector normal to the surface at each point. The overall boundary layer thickens at the separation point and is forced off the surface by the reversed flow. This can be visualized in Figure 6.

Compared to laminar boundary layers, turbulent boundary layers can tolerate much stronger flow deceleration, or adverse pressure gradient, before the flow separates. For a given adverse distribution, the separation resistance of a turbulent boundary layer increases slightly with increasing Reynolds number. This is because turbulent mixing increases momentum transport from the outer flow toward the wall.

In aerodynamics, flow separation often results in increased drag, particularly pressure drag (see definition on Section 3.2.3). Depending on the flow conditions, the separated flow can terminate and may again become reattached to the body. Between the point of separation and reattachment, a separation bubble is formed. A variety

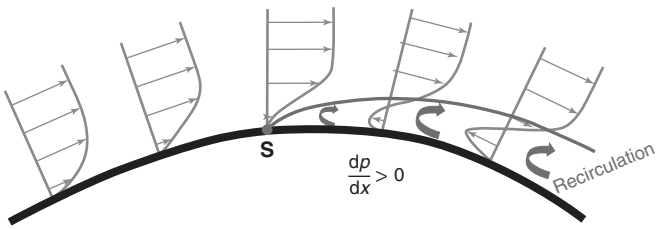


Figure 6. Flow separation over a curved surface.

of factors can influence reattachment, such as: the pressure gradient may become favorable owing to body geometry; the flow initially laminar may undergo transition within the bubble and may become turbulent. In aerodynamics, a short bubble may not be of much consequence.

2.6 The Reynolds number

The Reynolds number, Re, is a dimensionless number that gives a measure of the ratio of inertial forces to viscous forces and quantifies the relative importance of these two types of forces for given flow conditions. It is defined as in Equation 6

$$\mathrm{Re} = \frac{\rho V_\infty L}{\mu} \tag{6}$$

where L is a characteristic length of the vehicle.

Different flow patterns around a same vehicle occur for different Reynolds numbers. The Reynolds number can be used to determine dynamic similitude between different experimental cases. It is also used to characterize different flow regimes, such as laminar or turbulent flow: laminar flow occurs at low Reynolds numbers, where viscous forces are dominant and is characterized by smooth, constant fluid motion; turbulent flow occurs at high Reynolds numbers and is dominated by inertial forces, which tend to produce chaotic eddies, vortices, and other flow instabilities.

For passenger cars, typical values of the Reynolds number are of the order of 10^7.

3 FORCES AND MOMENTS

3.1 Forces and moments of immersed bodies

Any body of any shape will experience forces and moments when immersed in a fluid stream. For a vehicle or any

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

arbitrarily shaped body, the flow will exert forces and moments about all three coordinate axes. Figure 7 represents the forces and moments acting on a vehicle submitted to a cross-wind load.

The force on the body along the *X*-axis is called *drag*, F_D, and the moment around this axis is called the *rolling moment*, *R*. It is essentially the force that needs to be overcome in order for a body to move against the stream. The force on the body acting normal to the ground plane is called *lift*, F_L, and the moment around its axis is the yaw moment, *N*. The third component is called the *side force*, F_S, and the moment around its axis is the pitching moment, *M*. For a symmetric body about its drag–lift axis traveling directly into a stream (V_∞ parallel to *X*-axis), the side force, yaw, and rolling moments disappear and only the drag, lift, and pitch moments are left.

3.1.1 Bluff bodies and streamline bodies

A body is said to be a bluff body when the source of air resistance is dominated by pressure forces. When the drag is dominated by viscous forces, the body is called a *streamlined body*. Whether the flow is viscous drag dominated or pressure drag dominated depends solely on the shape of the body. A streamlined body looks like a fish or an airfoil at small angles of attack (Figure 8a). A bluff body looks like a cylinder or a block (Figure 8b). Road vehicles are bluff bodies. For a given frontal area and velocity, a streamlined body will likely have a lower resistance than a bluff body. See references White, 1986; Hummel, 1998; and Barnard, 2001 for drag values of different shaped objects.

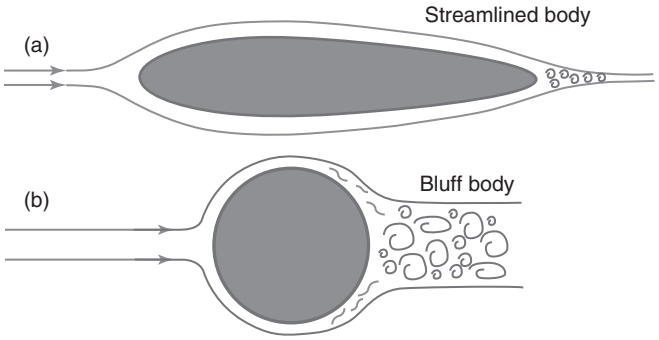


Figure 8. Airflow around a streamlined body and a bluff body.

3.2 Aerodynamic drag

Drag is the force that opposes motion. In aerodynamics, aerodynamic drag is the fluid drag force that acts on any moving solid body in the direction of the fluid free-stream flow. In vehicle aerodynamics, drag is conventionally taken as the resulting force acting along the *X*-coordinate (Figure 7). Drag comes from forces due to pressure distribution over the body surface and forces due to skin friction, which is a result of the fluid viscosity. These sources of drag are denoted as pressure drag and frictional drag, respectively. For passenger cars, drag is the most important aerodynamic force. It is the dominant source of resistance to motion at speeds above 50–70 kmph. There is a lot to gain from drag reduction as it directly influences fuel consumption, top speed and acceleration as we shall see in Section 4.

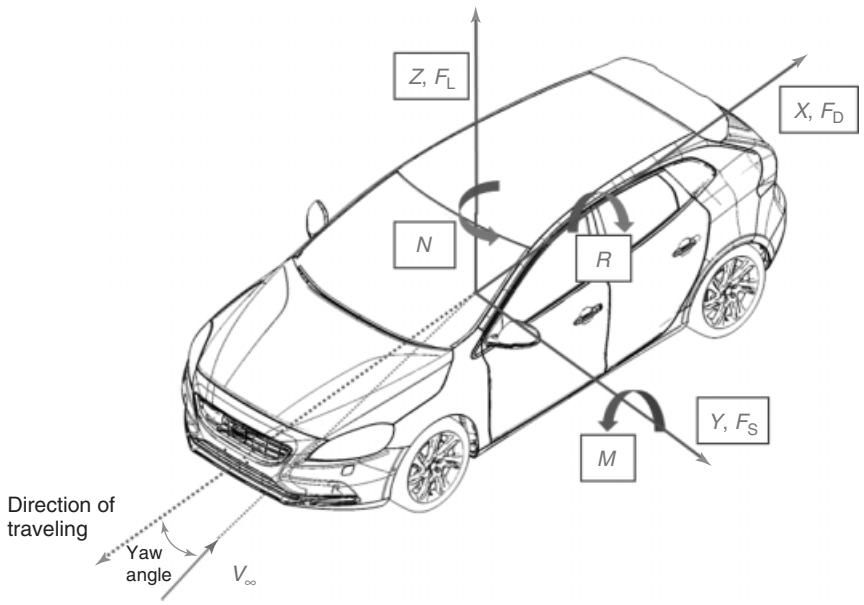


Figure 7. Forces and moments acting on a vehicle.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

3.2.1 Drag coefficient

Drag coefficient is a dimensionless quantity used to express the resistance of an object in a fluid stream. It is a common variable in automotive design. Drag depends on the properties of the fluid and on the size, shape, and speed of the object in motion. This relationship is expressed in Equation 7.

$$C_{\mathrm{D}} = \frac{F_{\mathrm{D}}}{\frac{1}{2}\rho V_{\infty}^{2} A} \tag{7}$$

where C_{D} is the drag coefficient, F_{D} the drag force, $\frac{1}{2}\rho V_{\infty}^{2}$ the free-stream dynamic pressure, and A the cross-sectional area as shown in Figure 9.

The drag coefficient can also be denoted as C_{X} or C_{W}. Equation 7 implies that for vehicles with approximately the same frontal area and traveling at the same speed, the vehicle with the lowest drag coefficient value will produce less drag.

3.2.2 Friction drag or viscous drag

Friction drag comes from friction between the fluid and the body surface. Frictional forces act tangentially to the body surface. It is this friction which is associated with the development of boundary layers. For passenger cars, which as described are bluff bodies, the contribution of viscous drag to the total drag is relative small, of the order of 7% to 10%.

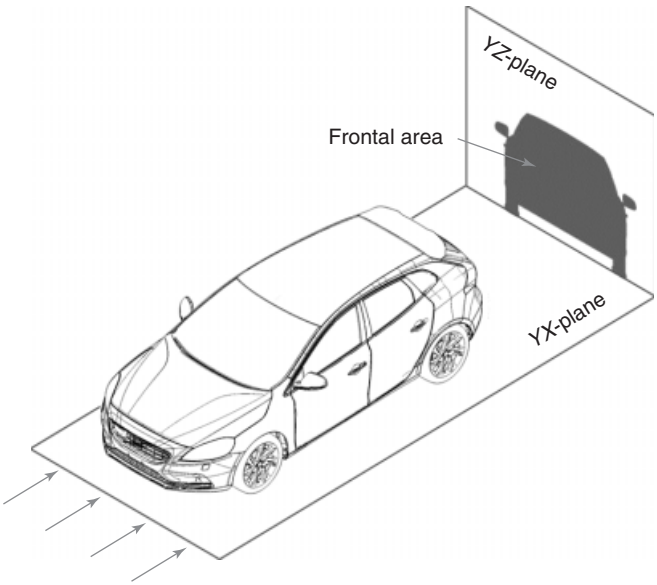


Figure 9. Definition of frontal area of a vehicle.

3.2.3 Pressure drag or form drag

Pressure drag or form drag arises because of the shape or form of the body. The size and shape of the body are the most important factors in form drag. Generally, bodies with large cross-sectional area have higher drag than slick or streamlined bodies. Form drag is associated with flow separation and the formation of a wake, as shown in Figure 8. Formally, both viscous drag and pressure drag appear because of the effects of viscosity that is responsible for the creation of a boundary layer. If the body was moving through an inviscid fluid, the viscous drag and the pressure drag would be zero. It is, however, useful to make a distinction between these two types of drag because they are due to different flow phenomena. Frictional drag is important for attached flows (flows with no separation), and it is related to the surface area exposed to the flow. Pressure drag is important for separated flows, and it is related to the cross-sectional or frontal area of the body.

3.2.4 Local drag

It is common in vehicle aerodynamics to analyze the areas of the flow field where drag is generated using a variable called the *local drag* (sometimes also called the *micro drag*). The local drag is derived from the integral form of the momentum equation and uses simplified approaches to exclude terms that are less relevant to the formation of drag. The resulting expression, Equation 8, relates the aerodynamic drag to the pressure and velocities at a given plane. Its derivation can be found in the works of Onorato, Costelli, and Garonne (1984); Cogotti (1989); and Ivanic and Guillieron (2005).

$$C_{\mathrm{D_{local}}}.A = \int_{S} (1 - C\,\mathrm{p_{tot}})\mathrm{d}S - \int_{S} \left(1 - \frac{V_X}{V_\infty}\right)^2 \mathrm{d}S + \int_{S} \left(\frac{V_Y}{V_\infty}\right)^2 \mathrm{d}S + \int_{S} \left(\frac{V_Z}{V_\infty}\right)^2 \mathrm{d}S \tag{8}$$

where S is the wake section at the plane of interest.

The local drag can be subdivided into wake drag, which includes the total pressure energy loss term and the streamwise velocity energy loss term. The two last terms in Equation 8 are known as the *vortex drag* or *induced drag*. That is,

$$\text{wake drag} = (1 - C\,\mathrm{p_{tot}}) - \left(1 - \frac{V_X}{V_\infty}\right)^2$$

$$\text{vortex drag} = \left(\frac{V_Y}{V_\infty}\right)^2 + \left(\frac{V_Z}{V_\infty}\right)^2$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For bluff bodies, wake drag is always larger than vortex drag, as it is related to form drag. Wake drag can be reduced by body streamlining. At the free stream, $C_{D_{local}} = 0$, because $V_X = V_\infty$ and $Cp_{tot}, V_Y, V_Z = 0$.

Figure 10 shows a typical example of the wake behind the rear end of a vehicle plotted with the local drag variable.

3.3 Aerodynamic lift

As defined in Section 3.1, for a body immersed on a fluid stream, lift is the component of the force that is perpendicular to the oncoming flow. Lift might be desired to be upward, in the case of an aircraft, or downward, in the case of road vehicles, especially racing cars. In racing car aerodynamics, the term *down force* is usually used to describe negative lift.

3.3.1 Lift coefficient

The lift coefficient is a dimensionless quantity that relates the lift generated by a body, the free-stream dynamic pressure of the fluid flow around the body, and a reference area associated with the body. Similarly to the drag coefficient, the lift coefficient is defined as in Equation 9

$$C_L = \frac{F_L}{\frac{1}{2}\rho V_\infty^2 A} \tag{9}$$

where C_L is the lift coefficient and F_L the lift force. A and $\frac{1}{2}\rho V_\infty^2$ have already been defined.

In case of lift, the use of A as the frontal area is more a matter of convenience, because lift is actually more related to the area parallel to the flow direction. The lift coefficient can also be denoted as C_Z or C_A. Modern passenger cars have values of C_L that are considerably low, whereas racing cars are designed to have negative values of lift (down force). In vehicle aerodynamics, both the magnitude and the balance between the front and the rear axle lifts are of importance. The front and the rear axle lifts are denoted by C_{LF} and C_{LR}, respectively.

3.4 Moments

As described schematically in Section 3.1, in addition to drag and lift forces, there are other forces and moments acting on a vehicle. As for drag and lift, it is also convenient to use dimensionless coefficients to refer to the side force and the three torques, Equations 10–13, respectively. Thus,

$$C_S = \frac{F_S}{\frac{1}{2}\rho V_\infty^2 A} \quad \text{Side force} \tag{10}$$

$$C_M = \frac{M}{\frac{1}{2}\rho V_\infty^2 A l} \quad \text{Pitching moment} \tag{11}$$

$$C_N = \frac{N}{\frac{1}{2}\rho V_\infty^2 A l} \quad \text{Yawing moment} \tag{12}$$

$$C_R = \frac{R}{\frac{1}{2}\rho V_\infty^2 A l} \quad \text{Rolling moment} \tag{13}$$

Note that as the moments have dimensions force × length, an additional characteristic dimension, l, is introduced in the definitions. This characteristic length is usually taken as the wheelbase, which is the distance between the front and the rear axles.

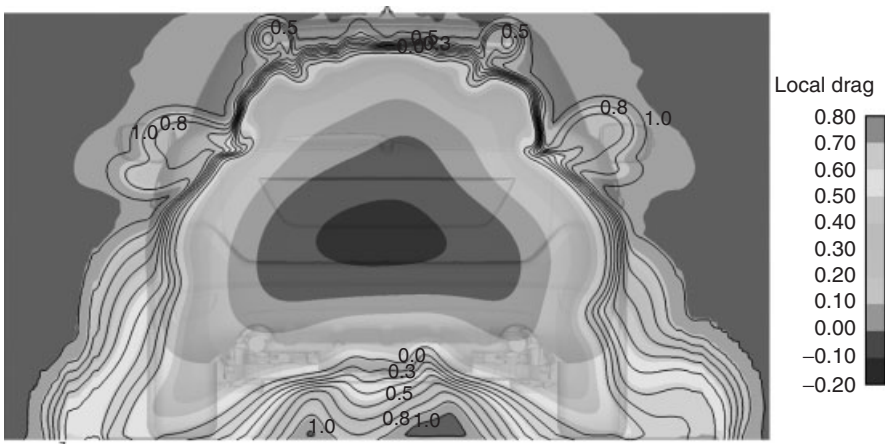


Figure 10. Local drag distribution at a plane 100 mm behind the rear end of a vehicle. (From Sterken, Sebben, and Löfdahl (2013). Copyright © 2013 SAE International. Reprinted with permission.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4 IMPORTANCE OF AERODYNAMICS ON ROAD VEHICLES

The airflow around a vehicle has a significant effect on several different attributes of a road vehicle. Table 1 presents a summary of such key areas, and as can be seen, there are several critical aspects related to aerodynamics that needs to be considered when developing a vehicle. The first section of Table 1 is labeled performance and includes aspects such as fuel consumption, acceleration, top speed, and emissions. As all of these are related to engine power, reducing the aerodynamic drag will reduce the power requirement for a given driving condition and may thus reduce fuel consumption and emissions. For a given engine, reducing aerodynamic drag will allow both improved acceleration and in particular higher top speed.

Handling is an important area not only for racecars but also for passenger cars. The handling properties of a passenger car are of course affected by several factors, aerodynamic lift forces being only one of them. At high speeds, the front-axle and, especially, the rear-axle lift forces have a significant influence on handling. Achieving a stable vehicle during strong winds gusts and high speed braking are two examples where aerodynamics is important. On high performance vehicles, it is not uncommon to introduce variant specific add-on devises such as a custom spoiler to reduce rear lift to be able to meet high speed requirements, often at the cost of slightly increased aerodynamic drag.

Contamination, number three in Table 1, focuses on keeping critical areas of the car clean from water, snow, and dirt. This can, for example, be to avoid significant water flow over the a-pillar onto the side window, which may obscure driver visibility, prevent water sprays from the wheels from entering the engine bay, or making sure the vehicle operates without issues in very snowy conditions.

Table 1. Overview of areas of importance for road vehicle aerodynamics (Hucho, 1998).

1. Performance • Fuel consumption • Acceleration • Top speed • Emissions	2. Handling • High speed stability • Braking stability • Cross-wind stability
3. Contamination • Visibility • Dirt/snow packaging • Splash and spray	4. Thermodynamics • Engine cooling • Transmission cooling • Brake cooling • Exhaust system cooling
5. Comfort • Climate control • Heating • Wind noise	

The fourth key area is summarized in the thermodynamics section of Table 1. In this case, cooling performance is the main consideration as the airflow around and through a vehicle will dictate the cooling performance of critical components such as the engine components, transmission, and brakes. Ensuring that sufficient amounts of air passes through the engine bays is an important requirement as the engine bays are getting increasingly occupied. Together with increased need for closing off the engine bay area for reasons such as NVH (noise, vibration, and harshness), contamination, and aerodynamic drag, getting sufficient cooling air through the engine bay poses a challenge to the engineer. Consequently, there is a need to optimize these kind of interacting attributes together.

Of course, the demand for cooling is not constant but depends on the specific driving condition. On modern cars, it is therefore becoming common with different kinds of mechanical systems to close off the cooling inlets when cooling is not required. Such systems are sometimes referred to as shutters. By closing off the cooling inlets in the front of the vehicle, drag reductions in the range of 5% can easily be achieved.

The last focus area in Table 1 reflects comfort aspects, which largely relates to aero acoustics but also internal climate control. Ensuring a comfortable climate in the driver compartment of the vehicle depends on several aerodynamics-related issues. Fresh air needs to be extracted from the airflow around the car and then conditioned to the desired settings. Another example is defrosting of the front screen, which is subject to legal requirements.

Aerodynamically generated noise relates to both external and internal noise levels. There are several design features on a passenger car that generate noise; a lot of work is, for example, often spent on optimizing the a-pillar and external rear view mirrors.

Of course, it is relatively easy to optimize each and every one of these attributes one by one, but as already indicated, in every vehicle project, there will always be a need for compromises between not only the areas in Table 1 but also with other aspects such as cost, weight, legal requirements, and manufacturing. The great challenge in developing a vehicle with good aerodynamic properties therefore not only lies in the purely aerodynamic field but also to be able to find solutions that do not only meet the aerodynamic targets but can also be managed by all other project members. Sections 4.1 and 4.2 will give more detailed presentations of the influence of aerodynamics on the performance and handling attributes separately.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.1 Aerodynamic influence on performance

As described in Section 3.1, the aerodynamic forces and moments acting on a road vehicle are usually divided into components, three forces and three moments. Aerodynamic drag, lift, and side force together with pitch, yaw, and roll moments are the common ones that are usually measured or calculated for a given passenger car configuration. Lift and side force are divided into front and rear components at the front and the rear wheel centers.

All parts of the exterior, wheel housings and wheels, as well as the engine bay and underbody contribute to the aerodynamic drag. Figure 11 shows typical numbers for how the drag generation is distributed on a modern passenger car. Approximately half of the aerodynamic drag is generated from the exterior. The wheels and wheel housings generate somewhere between 25% and 30% of the aerodynamic drag (Landström, 2011), and the remaining 20–25% is generated by the underbody and engine bay region. These numbers are estimates and should be treated as such; of course, there will always be specific models where these numbers do not apply. It is also important to understand that the net aerodynamic force on a specific region of a car is not only an effect of the design of one specific part, but may often be a result of the design of several surrounding parts influencing the flow simultaneously. This is referred to as interference effects and is a very common phenomenon in aerodynamics because of the complex geometries and strong nonlinear physics of fluid flow.

To be able to understand such interactions between different parts of the flow around a road vehicle, it is often necessary to investigate the local flow field, in addition to the net forces. Flow field investigations can be done both experimentally in wind tunnels and numerically using CFD.

Equation 14 gives the total propulsion force necessary to operate a vehicle at any driving condition. The first term on the right represents inertia effects from accelerating the vehicle. For simplicity, this has been summarized as the product of vehicle mass and acceleration, but should also include terms for rotating parts. The second term is called *climbing resistance* and represents the gravitational effect if driving up or down a slope. The third term is the rolling resistance and the fourth term is the aerodynamic drag. As three of the four contributors to the necessary propulsion force are dependent on mass and aerodynamic drag depends on the size and shape of the vehicle and the velocity, it is reasonable to assume that their relative contributions to the necessary propulsion force will depend on driving conditions.

$$F = m\frac{\mathrm{d}V}{\mathrm{d}t} + mg\sin\alpha + mgf_{\mathrm{R}}\cos\alpha + \frac{1}{2}\rho V^2 C_{\mathrm{D}}A \quad (14)$$

Figure 12 shows the required power necessary to operate an average passenger car as a function of different steady state velocities on flat ground, hence eliminating the first two terms in Equation 14. The remaining force has been split into aerodynamic drag and rolling resistance. For a

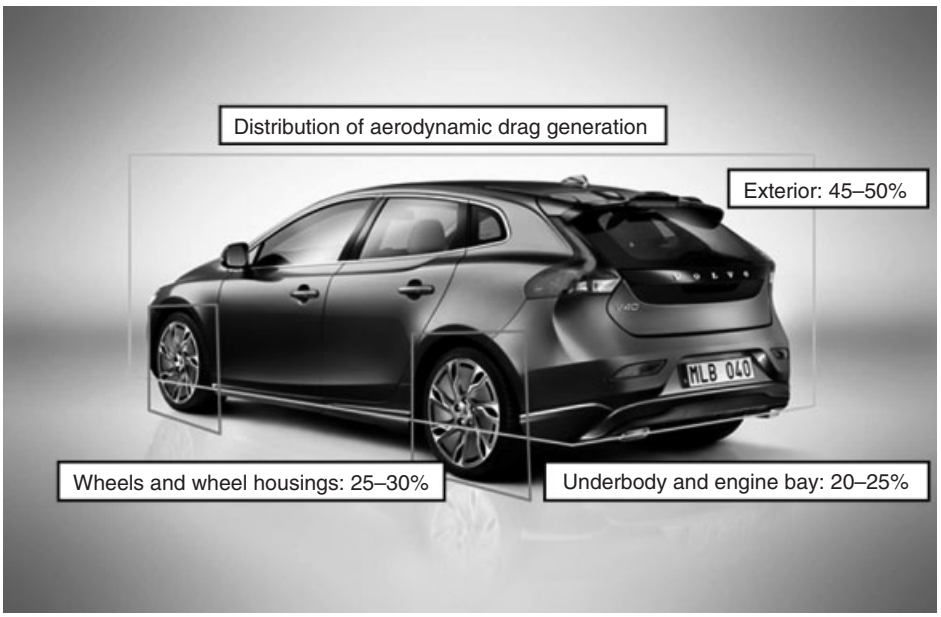


Figure 11. Aerodynamic drag distribution on a modern passenger car. Approximately half of the drag originates from the exterior, 25–30% from the wheels and wheel housings, and the remainder from the underbody and engine bay area.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

passenger car, the curves typically intersect in the range 50–70 kmph, and above these velocities, the contribution from aerodynamic drag grows significantly larger than rolling resistance.

From Figure 12, it is obvious that the contribution to fuel consumption from aerodynamic drag is strongly dependent on velocity. As shown by Woll (2005), reducing mass by 10% gives approximately twice the reduction in fuel consumption as a 10% reduction in aerodynamic drag when considering the New European Driving Cycle (NEDC), keeping all other factors constant. If instead considering driving at a constant speed of 150 kmph, a 10% reduction in aerodynamic gives four times larger reduction in fuel consumption than the corresponding reduction in weight. The reason for this difference lies in the shape of the NEDC with a relatively low average velocity of 33 kmph, and several sections of acceleration and braking that increase the relative importance of inertia in Equation 14.

Figure 13 shows an example of how the total amount of energy used by a car during NEDC is divided among different areas. In this case, the aerodynamic drag represents approximately a quarter of the overall energy loss during the NEDC. Rolling resistance and inertia losses makes up approximately half of the energy consumption, both depending on vehicle mass. If considering driving at a constant velocity of 90 kmph (Figure 14), distribution changes significantly. The aerodynamic drag now generates slightly more than half of the total energy loss. At even higher velocities, this part should increase even more according to Figure 12. This is of course an effect of inertia losses not being present in the constant velocity case.

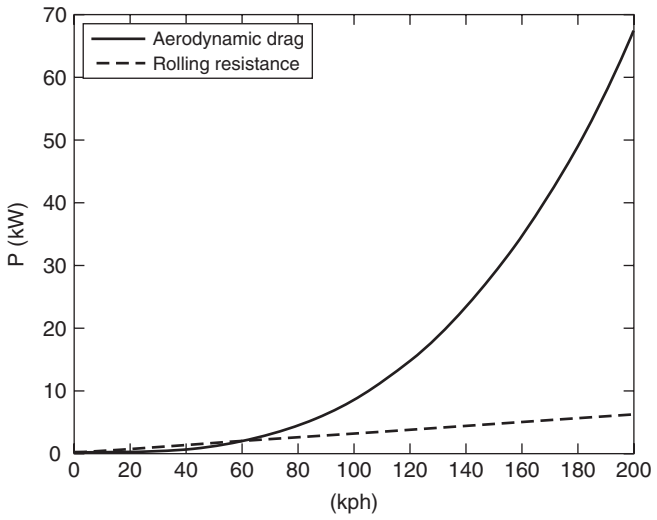


Figure 12. Power requirement for an average passenger car divided into aerodynamic drag and rolling resistance at different steady state velocities.

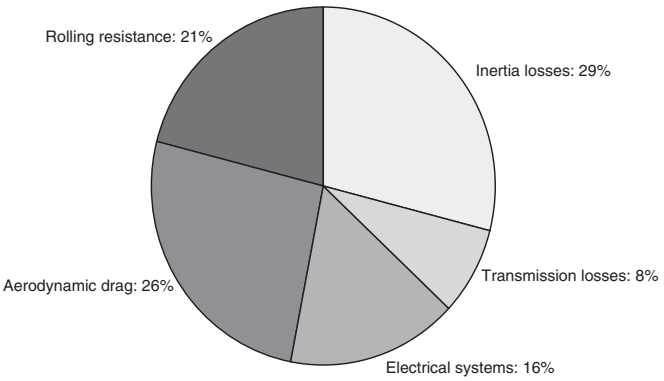


Figure 13. Energy losses for a passenger car during NEDC.

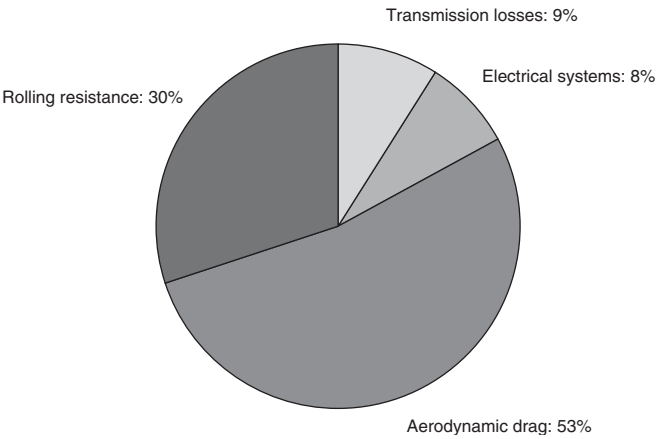


Figure 14. Energy losses for a passenger car during 90 kmph constant velocity.

Even though reducing weight is more efficient in reducing fuel consumption in the NEDC, achieving a weight reduction may be significantly more challenging—and expensive—compared with a reduction in aerodynamic drag.

The aerodynamic drag, however, may be significantly easier to reduce at a low cost. If the general shape of the exterior design of the vehicle can be improved during the design phase, the aerodynamic drag can be reduced at practically no additional cost. Improving the flatness of the underbody is common on many modern cars in order to reduce aerodynamic drag even further. This will most likely require underbody panels, but still at a relatively low cost. Other small features such as spoilers and deflectors can achieve significant reduction in aerodynamic drag and lift. Consequently, if the basic principles of aerodynamic design such as well-defined separation regions and a smooth underbody can be introduced and accepted early in the vehicle development process, the potential for low aerodynamic drag is good.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.2 Aerodynamics influence on handling

Apart from fuel economy and top speed, which for road vehicles is fundamentally decided by the aerodynamic drag of the vehicle, the other major influence of aerodynamics is the handling of the car. This is normally a result of the aerodynamic lift acting on the car, or transient or unsteady changes to the lift or side forces acting on the car during nonsteady conditions.

The high speed stability of a car is dependent on many factors, primarily weight distribution, suspension design, tires, and aerodynamics. Normally road cars have positive lift forces acting on both the front and the rear axles. Depending on the type of car and how the manufacturer wants the car to behave, the magnitude and difference between the front and the rear lift forces will need to be analyzed and controlled. Too much lift on the front axle will create a car with less steering response, too little lift may result in the steering becoming too responsive and nervous for some drivers. Too much lift on the rear axle will reduce the vertical loading on the rear tires leading to less lateral grip and consequently greater yaw sensitivity. This behavior will be more noticeable when braking at high speed.

Reducing the rear lift will often improve stability although if the rear lift is too low or even negative, that is, down force, the front lift will increase, perhaps leading to a lack of steering sensitivity. In general, a modification to the front of a car that results in a reduction to the front lift will almost certainly result in an increase in rear lift. Similarly, changes that increase the front lift will reduce the rear lift. Typically, the magnitude of the changes on the front lift will be reflected by an opposite change of between 50% and 75% on the rear lift. Although modifications at the rear of the car that effect the rear lift can also affect the front lift, the degree of influence is only 0%–30% of the change of the rear axle lift change.

Because of these interactions and also the influence that modifying the car will have on drag, it is important to consider the complete aerodynamic package that would be suitable for a particular type of car.

It is also important to consider the change to both drag and lift on the car as the angle of the wind relative to the cars direction changes as a result of natural side winds or the passing of other vehicles or buildings will be significant.

Figure 15 is a diagram for an SUV-type vehicle for drag, front lift and rear lift as a function of the relative, or yaw, angle of the wind.

It is important to identify if there are any sudden changes to drag and or lift at a particular yaw angle. This typically arises when the airflow around the rear of the car, near areas of flow separation, changes from a separated flow to an attached flow, or vice versa. If this occurs the rapid change of forces acting on the car may be felt by the driver as an unwanted nervousness in the behavior of the car. To avoid this, small changes to surfaces or radii in the problem area can create a better airflow and avoid rapid flow pattern changes.

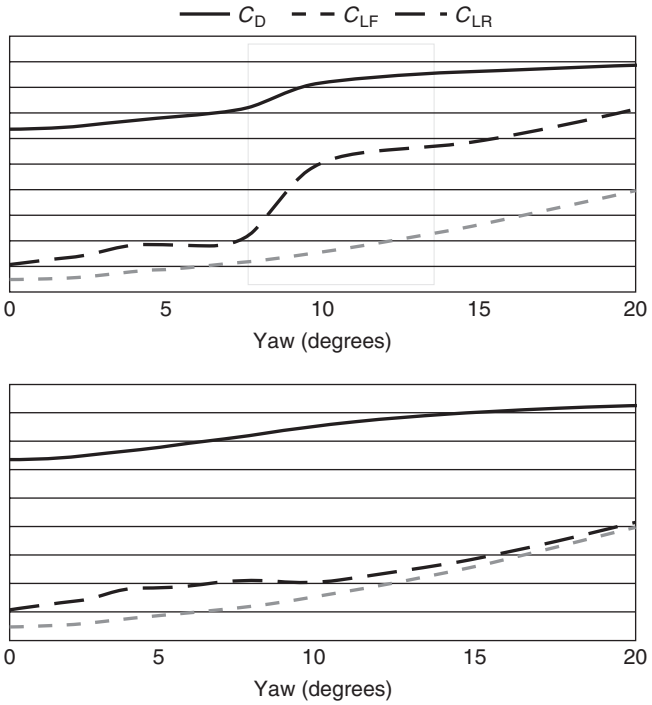


Figure 15. C_D, C_{LF}, and C_{LR} versus yaw angle for two roof wing configurations.

The upper graph in Figure 15 shows such a case where between 8° and 10° yaw, there is a significant change to the drag and rear lift forces. By means of a small modification to the upper rear corner of the roof wing, the sudden step is eradicated as shown in the lower graph.

As the airflow over the car is intrinsically time dependent, this change of force on the surface of the car can even be apparent when there is no particular change of yaw angle. Normally this is not noticeable but in some designs, it can create repetitive nervousness with a frequency of between 1 and 2 Hz.

5 DEVELOPMENT PROCESS

The typical aerodynamic development process for high volume production cars is, in general, largely generic. This process has evolved in parallel with the development process for the entire car with increasing complexity

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and demands leading to an increased need for aerodynamic research and development. Although including aerodynamic features and concepts into the car's design was present already since the early 1900s, this chapter relates mostly to the vehicles developed since the mid-1960s. A more detailed chapter relating to the historical development of automotive aerodynamics can be found in the work by Le Good *et al*. (2011). This time period reflects the beginning of the era when manufacturers started to be aware of the benefits of aerodynamics even for normal production cars. Figure 16 shows the evolution of Cd for a large number of different vehicles from the 1950s until the present day.

A central part in the development process of production cars is that the design of the vehicle has to be attractive in order to appeal to potential customers. The degree to which the external design of the vehicle can or may be influenced by aerodynamics will vary enormously among different model types, manufacturers, and market demands. There is never, therefore, a single solution to how the design process of a particular automotive project will be. Obviously, if the project in question has particular aims of offering significantly lower fuel consumption through low aerodynamic drag, more emphasis of the aerodynamic characteristics will have to be considered. Conversely, if a project is focused on offering a stunning looking form with little consideration for fuel consumption, design will be prioritized before low drag. Even in this case, however, the aerodynamicist will still need to ensure that other areas that are influenced by the airflow around the vehicle are satisfactory.

5.1 Design process—from multiple models to single model

A typical vehicle project will normally start with several different designers making sketches or models of a vehicle. These may sometimes be completely free of engineering restrictions, conceptual designs, but will normally have a number of design limitations already set. This would typically include major dimensions (width, height, length, etc.), but will probably include some basic aerodynamic guidelines. These guidelines, or aerodynamic hard-points, will be a result of previous experience with similar vehicle concepts and development, as well as preparatory research and development made in advance of the project start. The project targets will also have been investigated with regard to the fuel consumption levels, handling and stability needs, cooling air requirements, and other areas where aerodynamic performance is of significance.

Already from the dimension criteria in a project, one vital characteristic of the aerodynamic drag has been decided, that of the frontal area of the car. Since the late 1950s, the frontal area of passenger cars has steadily increased as a result of design trends, safety aspects, and customer demands and needs for bigger interior space. Subsequently, the frontal area has, market segment for market segment, increased by approximately 10% with each generation change. As the aerodynamic drag is a product of the C_D and the frontal area, the increase to area obviously has a negative effect. Restricting or controlling the size of the frontal area is therefore an important factor when defining the layout and dimensions of a new project.

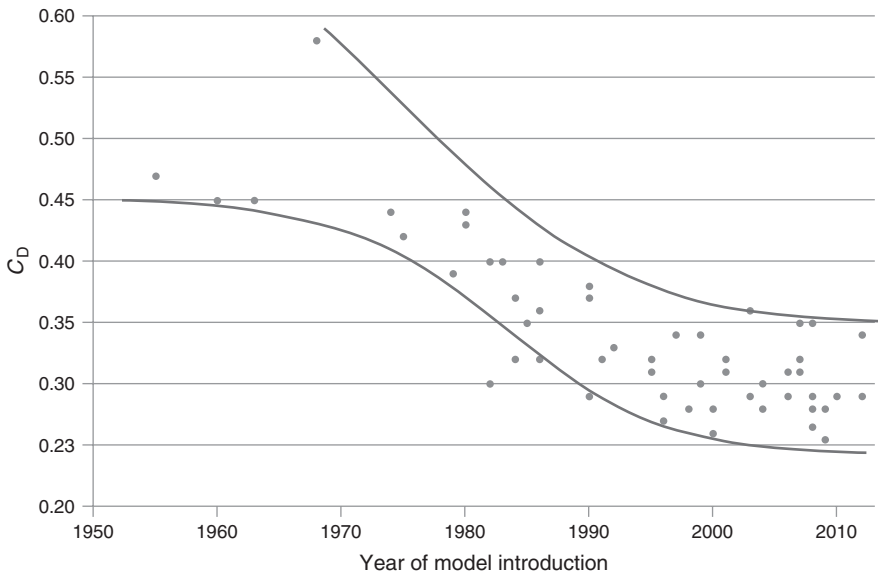


Figure 16. Evolution of C_D versus year of model introduction.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

An overview of the frontal area of C and D/E segment cars from 1993 to 2013 is shown in Figure 17.

The range of typical frontal areas, expressed in square meters, for different vehicle segments is given in Table 2.

During this period of many design models being studied, the aerodynamics group will try and analyze as many as practicable, in order to give feedback to each designer. This analysis will often be based again on previous experience, simple geometrical analysis of some or all the models, or even the results of testing the design. This may be done with physical testing of scale- or full-scale models and, since the mid-1990s, the increasing use of CFD. The degree to which method is used varies considerably between manufacturers, depending on the facilities and techniques available.

The purpose of aerodynamic analysis at this stage of the project is not to develop each model to its full potential but rather to identify the possibility of that model being able to reach the aerodynamic targets of the project. Fundamental analysis of the shape and design features can easily highlight areas where the design needs to be aerodynamically improved, and indeed areas where the aerodynamics works well. Typically, this analysis follows the basic guidelines for good aerodynamic design as outlined in Chapter 7.c.ii. This analysis, together with suggestions for improvement, is then given back to the designer and further changes to improve the situation may be implemented in the next iteration of the model. In some circumstances, however, this may be in conflict with the design theme that the designer is aiming for and may not be accepted.

It is not unusual at this stage of the design processes that the geometrical differences between different design proposals can be large. Similarly, the magnitude of suggested changes to the design of the car in order to improve the aerodynamic performance will also be large. In practice, this could be in the order of moving surfaces several centimeters. Even at this stage of the project, it is important to consider other attributes. For example, certain

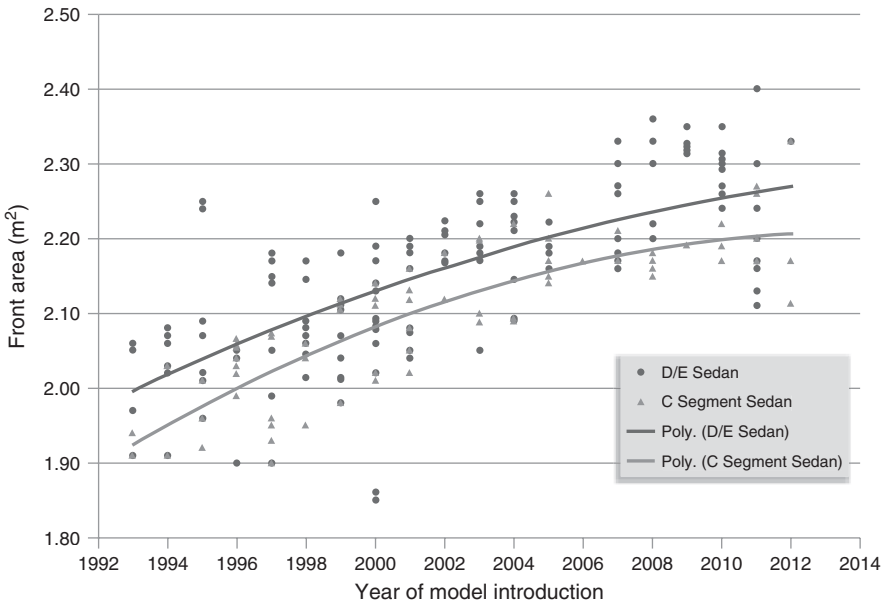


Figure 17. Frontal area for C and D/E sedans between 1993 and 2013.

Table 2. Frontal area range for different size vehicles.

Segment	Description	From (m^2)	To (m^2)
A	Micro cars	1.60	2.55
B	Subcompact	1.60	2.60
C	Compact	1.80	2.45
D/E	Medium	1.90	2.85
F	Full size	2.05	2.85
J/M	Compact SUV	2.50	2.75
S/V	Medium SUV	2.35	2.95

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

design proposals can be identified as having difficulties of supplying sufficient cooling air without making significant changes to the design language or packaging. It is important that these conflicts are identified during the concept phase. Failure to do so will lead to difficulty of meeting the project targets during the production phase of the project, resulting in increased costs or reduced performance. Figure 18 summarizes the number of models and magnitude of possible design change during the course of the project.

As the project continues, the number of design proposals normally becomes less, perhaps being reduced from 10 or more to 6 or less within half a year. The reasoning behind this reduction is often styling based although other attribute conflicts, including aerodynamics, could be an influencing factor.

The reduction in the total number of models in the project may offer possibilities of the aerodynamics team working and analyzing all the models more thoroughly. Depending on the company development facilities and resources, more full-scale models may be built and tested in the wind tunnel. Alternatively, more CFD analysis, probably using more detailed models will also be made. By this stage of the project, the possibility to modify the design of the car in order to influence the aerodynamic performance of the car becomes more and more restricted. In practice, the surfaces of the car will not be modified during the development loops more than a few centimeters.

At this stage in the development process, it is also important to consider the interaction between the upper body design and the underside of the car. As described in Section 4, the mass flow of air over and below the car will sometimes determine how the optimum shape of the rear of the car will be. Furthermore, the degree to which the underside of the car is aerodynamically shaped will also influence the upper body shape optimization. Therefore, it is necessary to coordinate the development of both the upper and the lower surfaces.

Eventually, typically between 2 and 3 years before the start of production, a single model will be selected by the project from the models that have been developed. The choice of model will often be based on the design esthetics, together with an analysis of many different attributes that the project will need to consider.

Once the project has moved into this stage, the emphasis of aerodynamic development is more focused on the detail development and optimization. The development focus is normally shifted toward full-scale physical wind tunnel testing, although detailed CFD models may still be used for particular analysis. By this stage in the project program, the full-scale model should have a fully detailed underbody, representative cooling airflow and sufficient exterior detailing to give a correct aerodynamic behavior. This will be necessary in order to make correct suggestions for modifications to the design of the car if more work is required to meet the project targets. Typically, sheet metal changes of more than 5 mm will not be common at this stage as the risk of influencing other characteristics of the car, and thereby delaying the project, become greater the nearer the final "frozen" design of the car is made. It may still be possible to make design changes to the areas of the car where the lead-time for the part tooling is shorter. This could include the design of the rear lamps, or other exterior parts where small changes may be needed in order to obtain a clean flow separation, for example. It could also include small modifications to the underbody of the car, spoiler lips, and deflectors or under body panel design, which can still make a significant change to the drag and lift behavior of the complete vehicle. It is the aerodynamicist's responsibility to ensure that all the aerodynamic fine tuning is complete before the design is finalized.

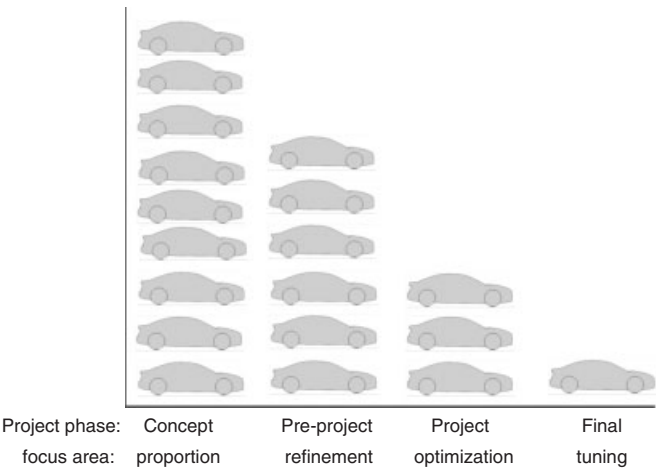


Figure 18. Design model and aerodynamic development process.

5.2 Attribute balancing

The drag targets will also consider those project prerequisites that will affect the aerodynamic performance of the car. This will include the basic underbody layout of the car, the packaging needs of the occupants, powertrain installation, visibility and safety requirements, manufacturing possibilities and limitations, and of course the cost limitations.

As a road-going vehicle fundamentally behaves as a wing creating both drag and lift forces, an acceptable balance between these two has to be found. This will depend on the project in question. If the vehicle has a relatively low top

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

speed (perhaps <190 kmph) perhaps it is sensible to prioritize lower drag than ensuring low lift. On the other hand, if the high speed handling properties of the car are considered more important, the lift levels on the front and the rear axles may be prioritized before drag. In practice, this balance will almost always be an engineering compromise.

In general, it is easier to modify details on the body of the car to adjust the lift forces and distribution front and rear than modify the drag levels. Often this is managed by modifying the height and shape of the car in the area where the airflow separates from the body. For sedans, this is typically on the upper surface of the boot/trunk lid and side surfaces of the rear fender.

5.3 Interaction effects

A fundamental part of the aerodynamic development process that has to be considered is the interaction effects of different areas of the body. Furthermore, these effects may well change the degree to which a design modification will work on the car. Therefore, for example, it is very important to analyze how the upper body design of a car can be optimized for a given under body design and vice versa. An example of this is shown in Figure 19. This summarizes a number of different aerodynamic changes when tested in isolation, compared to their respective effect on drag when tested as the complete optimized aerodynamic package.

As can be seen the effect adding the individual floor panels of the car is $\Delta C_D = -0.005$, whereas the combined effects is -0.008. Even more striking is the individual effect of the rear diffusor panel and the roof wing. The individual effects are only -0.004 but the combined effect, with the smoother underbody in position, is three times greater.

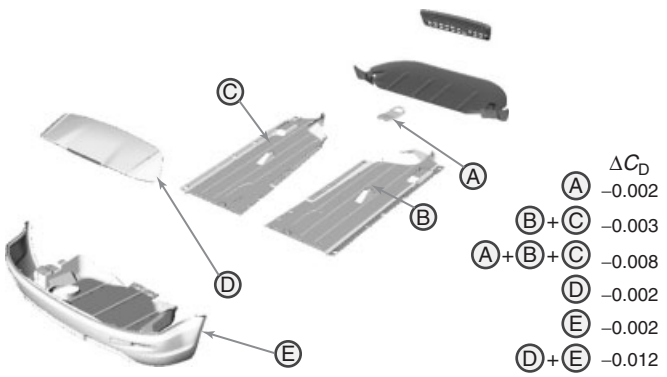


Figure 19. Interaction effects on drag for under and upper body parts.

5.4 Future trends

From the early 1980s, some volume production cars have offered designs with C_D of 0.30 or less. Since then, cars with C_D's of 0.30 have not been unusual, and from 2000 to 2010, there have been vehicles marketed with C_D as low as 0.25. Unfortunately, the reduction to C_D has in many cases been neutralized by an increase in frontal area, although a combination of low C_D and a small frontal area has been shown to be possible without creating a design that might only appeal to a limited customer base.

The need to reduce the fuel consumption in order to meet current and future legislation has forced manufacturers to include a low aerodynamic drag on more and more variants. Apart from keeping the frontal area to a minimum, for a given size of car, more effort will be required to achieve C_D values of 0.20 or less. As mentioned in Section 4.1, more than half the total drag on a typical car is from the underside, cooling, and wheels. Subsequently, more effort is being made to improve the flow underneath the car, generally achieved by designing the underside of the floor to be flatter or to have larger and flatter paneling. Initially, this concept was used by premium manufacturers but is becoming more normal even in high volume production cars.

A consequence of the improved under floor airflow is that the drag contribution from the cooling air and the tires has become more significant. By reducing the cooling air through the radiator and engine bay when not needed, this penalty is reduced. Furthermore, restricting the cooling air can also contribute to reduced fuel consumption by helping the engine reach its operating temperature quicker. Currently, systems that restrict ambient air reaching or passing through the cooling package are the most common. By 2020, there will probably be more variants that can control when and which cooling system should receive cooling air. The increased use of hybrids will create a need to cool different temperature cooling circuits in different manners, and will be suitable for new methods of flow control.

Optimized cooling air drag, a smoother under body, combined with a good basic aerodynamic design and attention to details has already shown that C_D values of under 0.24 are possible with a conventional design. The challenge to reach C_D values of 0.20 or less with road legal cars has been shown to be possible with the Saturn EV1 and the Volkswagen XL1 (Figure 20).

Both these cars (two seater, coupé-like) however have design features and packaging that would not be acceptable for high volume production cars. The challenge facing the automotive aerodynamicist and the design department is to create a five-seater vehicle with a C_D of under 0.20 that

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 20. 2013 Volkswagen XL1. (Reproduced by permission of Volvo Car Corporation.)

is not considered radical or unattractive. Covered wheel arch openings, for example, offer lower drag but may not be acceptable by the customer. The task is therefore to find aerodynamic solutions that offer a similar influence on the airflow around the car without the visual impact of fully covered wheel arches. New wheel concepts may be one solution, more sophisticated under body aerodynamic control may be another. One other possibility that may be introduced is the use of devices that actively influence the local airflow over certain critical areas of the car. These devices could be either moveable panels or even systems that attempt to influence the local boundary layer or flow separation behavior. These devices will undoubtedly add cost to the car but if the benefits are considered to outweigh the investments and product cost they may become a standard feature on higher priced vehicles by 2020.

REFERENCES

Allen, J.E. (1982) *Aerodynamics: The Science of Air in Motion*, 2nd edn, Allen Brothers & Father, Suffolk, UK.

Barnard, R.H. (2001) *Road Vehicle Aerodynamic Design*, 2nd edn, Mechaero Publishing, Hertfordshire, England.

Cogotti, A. (1989) A strategy for optimum surveys of passenger-car flow fields. SAE Paper No. 890374.

Hucho, W.H. (1998) Introduction to automobile aerodynamics in *Aerodynamics of Road Vehicles* (ed. W.H. Hucho). ISBN: 0-7680-0029-7.

Hummel, D. (1998) Some fundamentals of fluid mechanics in *Aerodynamics of Road Vehicles* (ed. W.H. Hucho). ISBN: 0-7680-0029-7.

Ivanic, T. and Guillieron, P. (2005) Aerodynamic drag and ways to reduce it. In Road Vehicle aerodynamics Lecture Series, von Karman Institute for Fluid Dynamics, Lecture Series 2005–05.

Landström, C. (2011) Passenger Car Wheel Aerodynamics. Ph.D. Thesis, ISSN: 0346-718X, Chalmers University of Technology.

Le Good, G., Johnson, C., Clough, B., and Lewis, R. (2011) The aesthetics of low drag vehicles. SAE Paper 2011-37-0016.

Onorato, M., Costelli, A.F., and Garonne, A. (1984) Drag measurement through wake analysis. SAE, SP-569.

Panton, R.L. (2005) *Incompressible Flow*, 3rd edn, John Wiley & Sons, New Jersey, USA.

Sterken, L., Sebben, S., and Löfdahl, L. (2013) Experimental and numerical investigations of the base wake on an SUV. SAE Paper No. 2013-01-0464.

White, F.M. (1986) *Fluid Mechanics*, 2nd edn, McGraw-Hill Book Co, New York, USA.

White, F.M. (2006) *Viscous Fluid Flow*, 3rd edn, McGraw-Hill Book Co, New York, USA.

Woll, T. (2005) Verbrauch und Fahrleistungen in *Aerodynamik des Automobiles*, 5. Auflage edn (ed. W.H. Hucho). ISBN: 0-7680-0029-7

FURTHER READING

Dieselnet, Emission Test Cycles, http://www.dieselnet.com/standards/cycles/ece_eudc.php (accessed 7 December 2013).

Wikipidia, http://en.wikipedia.org/wiki/Drag_(physics) (accessed 7 December 2013).

Wikipidia, http://en.wikipedia.org/wiki/Viscosity (accessed 7 December 2013).

Wikipidia, http://en.wikipedia.org/wiki/Boundary_layer (accessed 7 December 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto241
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Exterior Vehicle Noise Development: Assessment and Control

Ismael Fernández, Juan J. García-Bonito, and Javier Iturbe

Applus + IDIADA, Tarragona, Spain

1 Introduction 1
2 Sound and Noise 1
3 The Nature of Sound 1
4 Sound Pressure: the Decibel 2
5 Acoustic Intensity 3
6 Addition and Subtraction of Decibels 4
7 Frequency Weighting Scales 5
8 Noise Control 5
9 Exterior Noise Measurement 6
10 Automotive Applications 11
References 16
Further Reading 16

1 INTRODUCTION

All vibrating bodies surrounded by a fluid, gas, or liquid compress the layers of fluid adjacent to their surface. This compression is transmitted to the mass of surrounding fluid and is conveyed beyond the body. In these circumstances, the vibrating surface acts as a source of noise. Independently of whether the source of noise radiates in an open space (free field) or in an enclosed space (reverberate field), the acoustic field at a particular point is determined by the fluctuating pressure due to the wave propagation. The measurement of this acoustic pressure is of vital importance in any test to quantify the level of noise produced by a component or vehicle. Furthermore, the analysis of the variation in this pressure in the time and frequency domain gives us fundamental information about the origin of the noise studied.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 SOUND AND NOISE

Sound is a disturbance in the balanced state of the air molecules that vibrate due to the propagation of a compression wave created by a vibrating object. This disturbance is associated with an alternative displacement of the layers of air and a fluctuation of pressure (acoustic pressure) that, on reaching our ears, produces the auditory sensation. Sound may generate pleasant or unpleasant sensations depending on the spectral and temporal characteristics of the fluctuating pressure affecting our eardrums.

A clearly unpleasant or annoying sound is called *noise*. Nevertheless, the level of nuisance does not only depend on the type of sound but also on our attitude toward it; for example, a type of music that some people may like may annoy others, especially if it is very loud. Another example of the subjectivity of the possible nuisance of a sound is the sensation that the exhaust system of a sports car produces: it is pleasant for fans of this type of vehicles, but extremely annoying for others.

3 THE NATURE OF SOUND

Sound is defined as a pressure variation that the ears can detect, ranging from very low variations to levels that could

cause damage. The study of sound is called *Acoustics* and covers all fields of generation, propagation, and reception of sound. Noise is an unavoidable part of daily life and technological development, which has produced a notable increase in the noise level coming from machines, factories, traffic, and the like.

Figure 1 shows a piston mounted at the end of a cylindrical tube producing an alternative forward and backward movement. When the piston moves forward, the particles near to its surface accumulate creating a compression zone; when it moves backward, the particles separate creating a rarefaction zone or dilation. In compression, the air pressure will be a little higher than its value of balance and in dilation it will be a little lower. Therefore, a series of compressions and dilations will be created along the tube. Thus, a sound wave will be generated in the tube, which can be heard at the other end. The speed of sound propagation along the tube is a function of the elasticity and air density, which depend on the static pressure and temperature. At atmospheric pressure and at 20°C, the speed of sound in air is 344 m/s.

Figure 2 shows the variation in pressure around the static value P_o (atmospheric pressure) as a function of time for a generic point inside the tube. The maximum fluctuation is called *sound amplitude* and the number of oscillations per second that coincides with the rhythm of vibration of the piston is called the *sound frequency*. We will call this sound a *pure tone* because it has been created by a sound source that oscillates at a single frequency.

Very few man-made noises or natural sounds are pure tones. Even the sounds of a musical instrument, which seem to be made up of a clear single note, contain more than one frequency. Normally, the sensation of frequency that we perceive (tonality) corresponds to the dominant frequency of the sound heard.

The range of audible frequencies, for a young and healthy person, goes from approximately 20 Hz until 20,000 Hz. Frequencies below 20 Hz give rise to infrasound and frequencies higher than 20 kHz to ultrasound. On the other hand, as we will see later on, not all audible frequencies are perceived equally by the human ear.

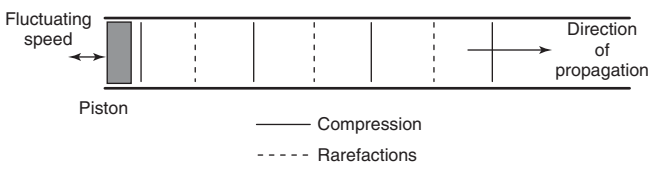


Figure 1. Vibrating piston at the end of the tube generates a compression wave that propagates along the tube, producing a compression and rarefaction (wave), which propagates at the speed of sound (345 m/s). (Reproduced by permission of IDIADA.)

4 SOUND PRESSURE: THE DECIBEL

As described above, sound is made up of a series of oscillations of pressure in relation to atmospheric pressure. Therefore, it is logical to think that this pressure is the most obvious parameter to express the magnitude of the sound field. Strictly speaking, the amount to measure is the difference between the fluctuating value of pressure and its value of balance P_o (atmospheric pressure). However, as the frequency of the wave increases, it is more difficult to measure these fluctuations as they become faster. It is obvious that, even for the lowest vibration able to produce audible sound, it is necessary to make a time average that serves as an indicator of the amplitude of the signal. The way this average is calculated is as follows:

- Square all the values of pressure difference during any cycle. Hence, negative values become positive.
- Average
- Calculate the square root.

The final result is known as the *root mean square* (r.m.s.) *value* represented as $p_{\text{r.m.s}}$. From now on, we will refer to the value of acoustic pressure that characterizes an acoustic field at the measuring point as the r.m.s. value of the pressure difference.

Mathematically, this is expressed by

$$p_{\text{r.m.s.}} = \sqrt{\frac{1}{T}\int_0^T p^2(t)\mathrm{d}t} \tag{1}$$

The measurement unit of pressure is the Newton/m^2, called *Pascal* (Pa). Typical acoustic pressures have values of fractions of a Pascal, whereas atmospheric pressure is approximately 100,000 Pa (=1 atmosphere). Therefore, sound waves are associated with tiny variations in air pressure.

Unfortunately, the measurement of sound pressure directly in Pascal gives rise to a series of difficulties, whose origin comes from the characteristics of the human ear, and which are explained below:

- The pressure range that the human ear can perceive is wide. The weakest sound pressure that can be perceived by a normal, healthy person is 2×10^{-5} Pa at a frequency of 1000 Hz (auditory threshold), while the sound pressure that begins to be painful is 100 Pa (pain threshold). That is, the scale of audible pressures covers a dynamic range of 1 to 5,000,000, which leads to the use of unmanageable numbers.
- Our auditory system does not respond lineally to the stimuli that we receive but rather has a logarithmic form.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

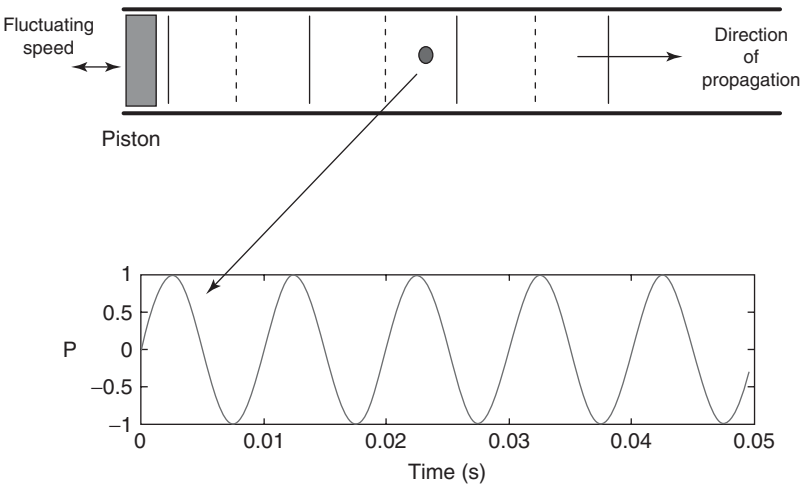


Figure 2. Variation in acoustic pressure with time at a particular point in the tube, which is induced by the vibration of the piston. If that vibration is sinusoidal, pressure will follow the same trend. (Reproduced by permission of IDIADA.)

For example, if the pressure of a pure tone of 1 kHz is doubled, the sensation produced by it will not be double. This is a clear demonstration of the nonlinearity of hearing.

For these two reasons, it seems reasonable to use a logarithmic scale in order to quantify the sound pressure. However, as the logarithm of a number lower than 1 is negative, this would mean that any sound pressure whose value were a fraction of 1 Pa would be expressed by a negative number. In order to avoid this problem, a reference amount for which it will always be necessary to divide the corresponding pressure before taking logarithms is introduced. This amount is in fact the threshold of perception, that is, 2×10^{-5} Pa. On the basis of this, the level of sound pressure is defined by:

$$\mathrm{SPL} = 10 \cdot \log \frac{p^2}{p_{\mathrm{ref}}^2} = 20 \cdot \log \frac{p}{p_{\mathrm{ref}}} \tag{2}$$

This expression is called the *sound pressure level* and is measured in decibels relative to 2×10^{-5} Pa. The word *level* is added to indicate that the amount has a certain level above a preset reference value. The acronym *SPL* or the symbol L_{p}. is used.

5 ACOUSTIC INTENSITY

Another parameter of interest in the definition of an acoustic field is the intensity. The *intensity* of a sound wave is the amount of acoustic energy crossing the unit of area normal to the direction of propagation of the wave per unit of time. Therefore, it is expressed in units of power (energy per unit of time) per surface unit, that is, in watt/m^2. In the case of a flat wave, as seen in the tube in Figure 1, the acoustic intensity is given by:

$$I = \frac{p^2}{\rho c} \tag{3}$$

where p is the acoustic pressure (r.m.s), ρ is air density (kg/m^2), and c denotes the sound speed (m/s). Figure 3 illustrates the concepts associated with the measurement of intensity. The vector crossing the plane that denotes the unit of area indicates acoustic power and has information about its associated direction.

As with the sound pressure, one can talk about the sound intensity level defined as

$$\mathrm{SIL} = 10 \cdot \log \frac{I}{I_{\mathrm{ref}}} \tag{4}$$

where *SIL* is the sound intensity level, I is the intensity in W/m^2, and I_{ref} is the intensity of reference that, in this case, is taken as 10^{-12} Watt/m^2. The measurement of sound intensity offers the opportunity to detect *in situ* the area of a vibrating surface that emits more noise. In this context, the expression *emits more noise* suggests that the surface injects acoustic energy into the environment and acts like a loudspeaker. Intensimetry may also reveal those parts of a surface acting as energy drains and subsequently absorbing noise.

It is of vital importance to keep in mind the vectorial nature of the measurement of intensity, that is, all measurements of intensity have amplitude and direction. Therefore,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

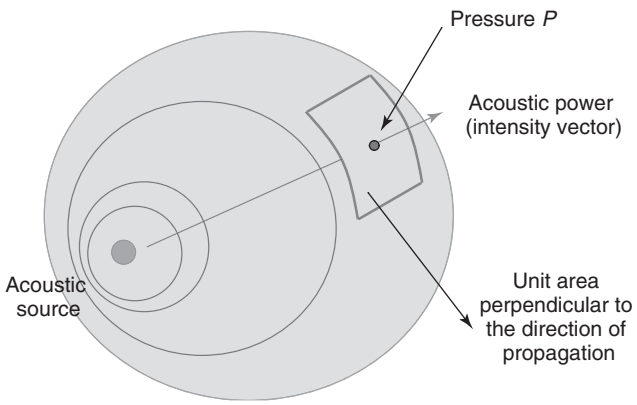


Figure 3. Acoustic field created by a source expands when going away from the vibrating surface that generates it. The measure of acoustic intensity in a particular point of the acoustic field indicates the power going through the perpendicular section into the direction of propagation. (Reproduced by permission of IDIADA.)

it is very useful for the location of sources in acoustically complex spaces to be known.

6 ADDITION AND SUBTRACTION OF DECIBELS

6.1 Addition of decibels

When two sound sources radiate sound, they both contribute to the existing sound pressure level at a point far from both sources. Let us assume that dB_1 and dB_2 are the sound pressure levels due to two sources that are radiating simultaneously. The acoustic pressure at a point in the space associated with each source is given by

$$dB_1 = 10 \cdot \log \frac{p_1^2}{p_{ref}^2} \Rightarrow p_1^2 = p_{ref}^2 \cdot 10^{\frac{dB_1}{10}}$$
$$dB_2 = 10 \cdot \log \frac{p_2^2}{p_{ref}^2} \Rightarrow p_2^2 = p_{ref}^2 \cdot 10^{\frac{dB_2}{10}} \quad (5)$$

In order to know what is the total sound pressure produced by the acoustic surfaces, we will refer to the concept of intensity established previously: intensity is the energy per unit of surface and per unit of time. Therefore, it seems logical to think that the intensity due to both sources will be the sum of the intensities of each source. We have also seen that intensity is proportional to the r.m.s. pressure squared (Equation 3), and so, we can deduce that the r.m.s. pressure squared due to the two sources will be the sum of the r.m.s. pressures squared for each of the sources; therefore, it will be

$$p^2 = p_1^2 + p_2^2 = p_{ref}^2 10^{\frac{dB_1}{10}} + p_{ref}^2 10^{\frac{dB_2}{10}} \quad (6)$$

And the total sound pressure level will be

$$dB_{1+2} = 10 \cdot \log \frac{p^2}{p_{ref}^2} = 10 \cdot \log \frac{p_{ref}^2 \left(10^{\frac{dB_1}{10}} + 10^{\frac{dB_2}{10}}\right)}{p_{ref}^2}$$
$$= 10 \cdot \log \left(10^{\frac{dB_1}{10}} + 10^{\frac{dB_2}{10}}\right) \quad (7)$$

This result could be extended to n sources, that is,

$$dB_n = 10 \cdot \log \left(\sum_{i=1}^{n} 10^{\frac{dB_i}{10}}\right) \quad (8)$$

According to Equation 8, the sum of two equal levels of SPL produces an increase of 3 dB. Therefore, two sources producing individually a sound level of 80 dB at a certain point will generate a total pressure of 83 dB when acting simultaneously.

6.2 Subtraction of decibels

In some cases, it is necessary to subtract levels of noise. By means of a mathematical development similar to that explained in the previous section, it is deduced that the difference in two levels of noise given by dB_1 and dB_2 is

$$dB_{1-2} = 10 \cdot \log \frac{p^2}{p_{ref}^2} = 10 \cdot \log \left(10^{\frac{dB_1}{10}} - 10^{\frac{dB_2}{10}}\right) \quad (9)$$

The most usual case of the subtraction of decibels arises when we need to measure the noise of a component in the presence of background noise. In these cases, it is important to know whether the measured noise is dominated by the noise of the machine, the background noise, or by a combination of both. The procedure is as follows:

(a) The total existing noise with the component in L_{s+n} operation is measured.
(b) The component is stopped or encapsulated and the background noise measured L_n.
(c) The ΔL difference is calculated as $L_{s+n} - L_n$.

If ΔL is <3 dB, the background noise is too high for an accurate measurement and, therefore, the level of noise produced by the component cannot be measured accurately as long as the background noise does not decrease. On the other hand, if the difference is above 10 dB, the background noise can be ignored. If the difference is between 3 dB

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and 10 dB, the correct level of noise can be found using Equation 9.

7 FREQUENCY WEIGHTING SCALES

The human ear is not consistent in its sensitivity to different frequencies. Therefore, the ear will perceive a sound at 1000 Hz with greater intensity than one of the same intensity at 200 Hz. In order to take into account this phenomenon, weighting scales of different frequencies have been standardized. These are shown as A, B, C, and D in Figure 4. Initially, these weighting scales were developed for different levels of noise: scale A was for low noise level (40 dB), scale B for medium sounds (70 dB), and scale C for loud sounds (100 dB). However, in the end, scale A has been applied universally to all levels. Scale D is used exclusively for aircraft noise.

It is important to observe that scale A noticeably attenuates frequencies below 300 Hz. This aspect should be borne in mind in the analysis of vehicle interior noise at low speeds as this is dominated by the structural excitation that the engine introduces into the passenger compartment, predominately at low frequencies (<200 Hz).

8 NOISE CONTROL

The term *noise control* refers to all those techniques that reduce the level of acoustic pressure at a point in free space or inside a given enclosed space. The method used to get this noise reduction (or control) basically depends on the level of understanding of the noise source and of the difficulty and cost associated with implementing this method. The techniques of noise control most commonly used at present are based on the following principles:

- Acoustic isolation
- Acoustic absorption
- Reduction of the excitation force that acts on the radiant surface
- Reduction of the level of vibration on the surface of the source
- Reduction of the transmissibility of the medium of propagation (air/structure)
- Cancellation of noise through destructive interference (active noise control)

Of the above techniques, the first two stand out for the simplicity of their application in a great number of situations. Acoustic isolation and absorption use the idea of inserting an acoustic barrier combined with noise-absorbing materials. The combination of both principles could be used in those cases in which the noise problem is found at high frequencies (>1 kHz). However, these methods are ineffective in reducing noise at low frequencies (<400 Hz) as; in this case, the wavelength of the sound to be controlled is much greater than the thickness of the materials and the absorbents used in the construction of the barriers (or encapsulation).

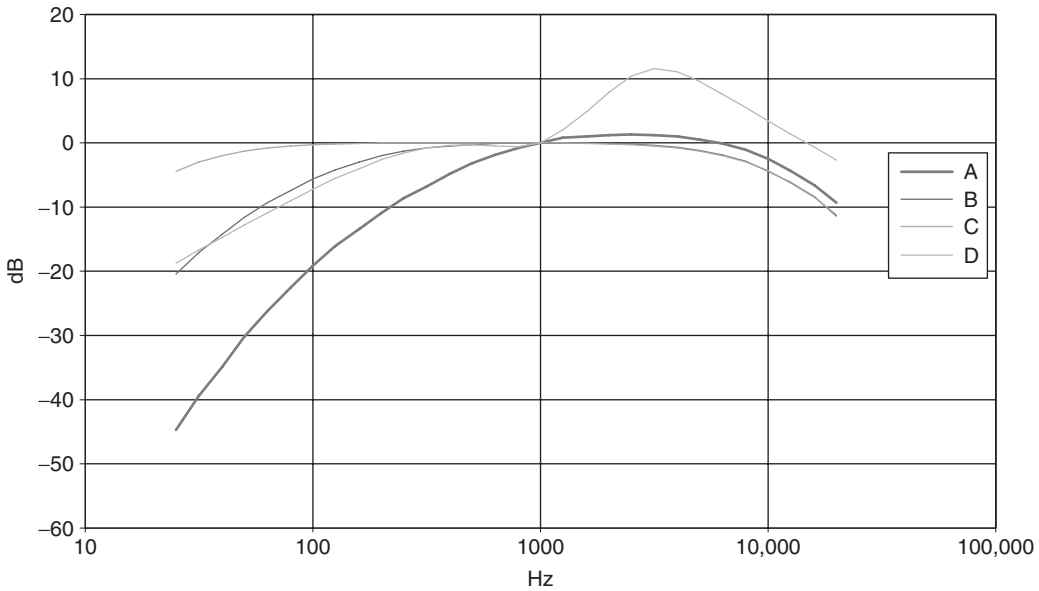


Figure 4. Weighing frequency scales A, B, C, and D. (Reproduced by permission of IDIADA.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

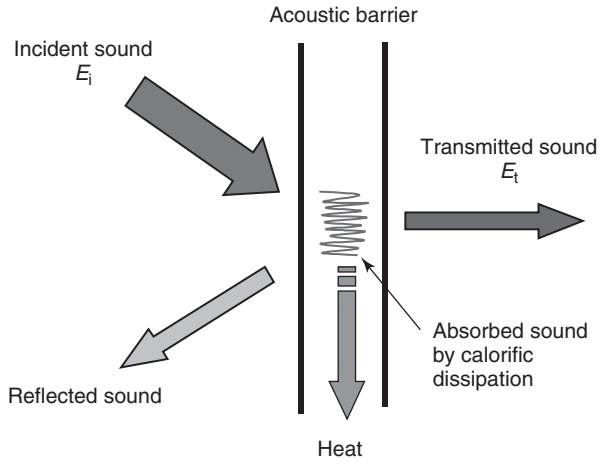


Figure 5. Sound reflected, absorbed, and transmitted. The energy of the incident wave is not destroyed; it is only reflected and transformed into heat. The rest of the energy generates a transmitted acoustic wave of lower amplitude than the incident wave. (Reproduced by permission of IDIADA.)

Noise does not always spread freely in space (propagation in free field). On many occasions, the sound reaches obstacles such as roofs, walls, and structural panels which cause reflections. Vehicle engine noise, for example, finds an obstacle in the wall that separates the engine compartment from that of the passengers (the firewall). A car with the bonnet open is noisier (external noise) than with bonnet closed as the bonnet obstructs the noise escaping to the exterior. When an acoustic wave reaches a solid surface, such as the firewall of a vehicle, part of the acoustic energy is transmitted to the other side of the barrier, another part is reflected, and the barrier absorbs the rest. Figure 5 shows the distribution of the acoustic energy of a wave that hits a barrier of this type.

Isolation and absorption are the two characteristics that make a material interesting from an acoustic point of view. They are two different characteristics, but the combination of both is very desirable from the point of view of noise control.

9 EXTERIOR NOISE MEASUREMENT

The main exterior noise measurement for homologation purposes is performed nowadays in wide open throttle acceleration condition, but in a short time, a new testing procedure will enter into force. This procedure is the result of wide studies of driver behavior in the urban traffic and it intends to simulate the real urban driving conditions. On the contrary to the current method, it takes into account the vehicle noise at full throttle acceleration and constant speed conditions.

The current method is explained in standards and regulations (Regulation No 51–02 – Rev.1/Add.50/Rev.2/Amend.1, 2012), (EU Directive 92/97/EEC amending Directive 70/157/EEC, 1992); (ISO 362, 1998) and it will be explained here. In order to explain the procedure, we need to introduce the concept of vehicle category.

9.1 Vehicle category

Vehicles can be classified into the following categories:

- Category M_1. Vehicles used for the carriage of passengers and comprising not more than eight seats in addition to the driver's seat.
- Category M_2. Vehicles used for the carriage of passengers, comprising more than eight seats in addition to the driver's seat, and having a maximum mass not exceeding 5 tons.
- Category M_3. Vehicles used for the carriage of passengers, comprising more than eight seats in addition to the driver's seat, and having a maximum mass exceeding 5 tons.
- Category N_1. Vehicles used for the carriage of goods and having a maximum mass not exceeding 3.5 tons.
- Category N_2. Vehicles used for the carriage of goods and having a maximum mass exceeding 3.5 t but not exceeding 12 tons.
- Category N_3. Vehicles used for the carriage of goods and having a maximum mass exceeding 12 tons.

9.2 Test track

The exterior noise tests are performed in a track that consists of two areas (Figure 6):

- A drive line with a length l_a and a width of 3 m minimum. The minimum value of l_a is 20 m for long vehicles and 10 m for the other ones.
- A propagation area that extends at least 10 m from the center of the drive line CC′ and at least 10 m at both sides of the line PP′.

The microphones must be located at a distance of 7.5 ± 0.2 m from the reference line CC′ (Figure 6) of the track and 1.2 ± 0.1 m above the ground. Their axes of maximum sensitivity must be horizontal and perpendicular to line CC′.

Two lines AA′ and BB′, parallel to line PP′, situated, respectively, 10 m forward and 10 m rearward of PP′ are marked out on the test runaway.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

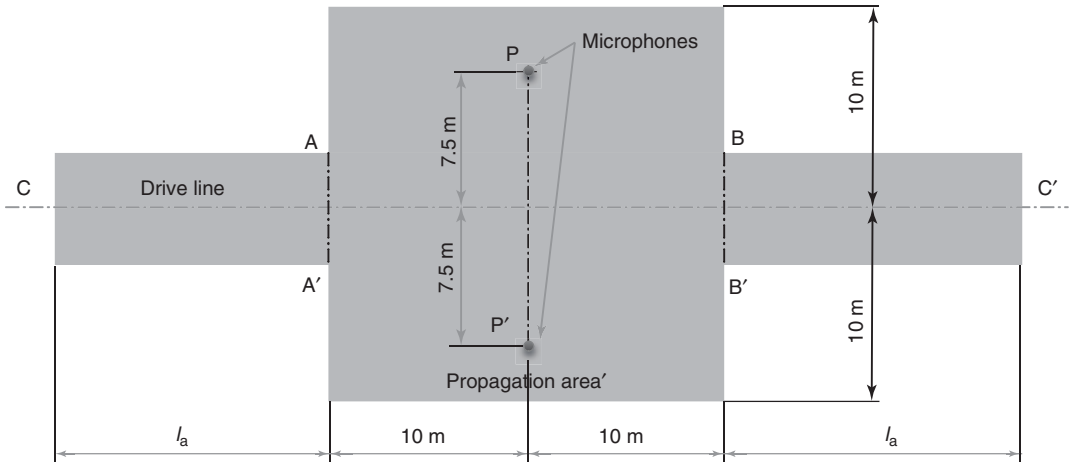


Figure 6. Exterior noise test track. (Reproduced by permission of IDIADA.)

9.3 Test procedure

9.3.1 Test with vehicle in motion (pass-by test)

The vehicle is driven at constant speed as specified below on the drive line until it reaches the line AA′. Its longitudinal median plane will be as close as possible to the line CC′. Just when the vehicle front end reaches line AA′, the driver has to depress the pedal as rapidly as practicable and will keep in this position (wide open throttle) until the vehicle rear end reaches line BB′, when the pedal also will be released as rapidly as practicable (Figure 7).

The maximum sound level, expressed in A-weighted decibels, shall be measured, at both vehicle sides, as the vehicle is driven between lines AA′and BB′. Such values shall constitute the results of the measurement (Figure 8). The time weighting has to be "fast".

9.3.1.1 Choice of gear ratios

9.3.1.1.1 Vehicles with manual transmission

9.3.1.1.1.1. Vehicles of categories M_1 and N_1. When these vehicles are fitted with a gearbox with four forward gears or less, they shall be tested in second gear.

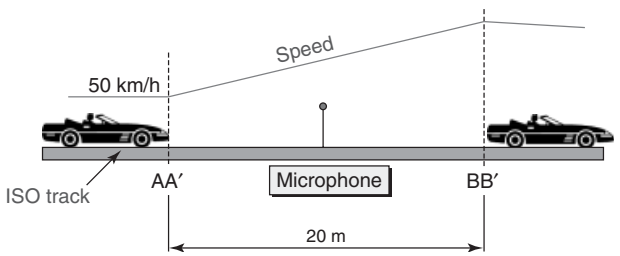


Figure 7. Layout of the exterior noise test. (Reproduced by permission of IDIADA.)

When they are fitted with a gearbox with five forward gears or more, they shall be tested in second and third gear.

9.3.1.1.1.2. Vehicles of categories other than M_1 and N_1. These vehicles shall be tested sequentially at different gear ratios. Initial testing will be carried out using the gear ratio which gear is (x/n) or the next higher gear ratio if (x/n) is not an integer.

where
x is the number of forward gears
$n = 2$ when the engine power is not >225 kW
$n = 3$ when the engine power is >225 kW

The testing shall continue from the gear (x/n) to the next higher gear. Shifting up ratios from (x/n) shall be terminated when in the gear x in which the rated rotating engine speed S is reached just before the rear of the vehicle has passed the line BB′.

9.3.1.1.2 Vehicles with automatic transmission. The test will be conducted with the selector in the position recommended by the manufacturer for normal driving.

9.3.1.2 Approach speed

9.3.1.2.1 Vehicles with manual transmission

9.3.1.2.1.1. Vehicles of categories M_1 and categories other than M_1 whose engine power is not $>$ 225 kW The approach speed is the lowest of the following: $V_A = 50$ km/h or V_A corresponding to $N_A = \frac{3}{4}S$

where
V_A: Uniform vehicle speed at the approach of line AA′

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

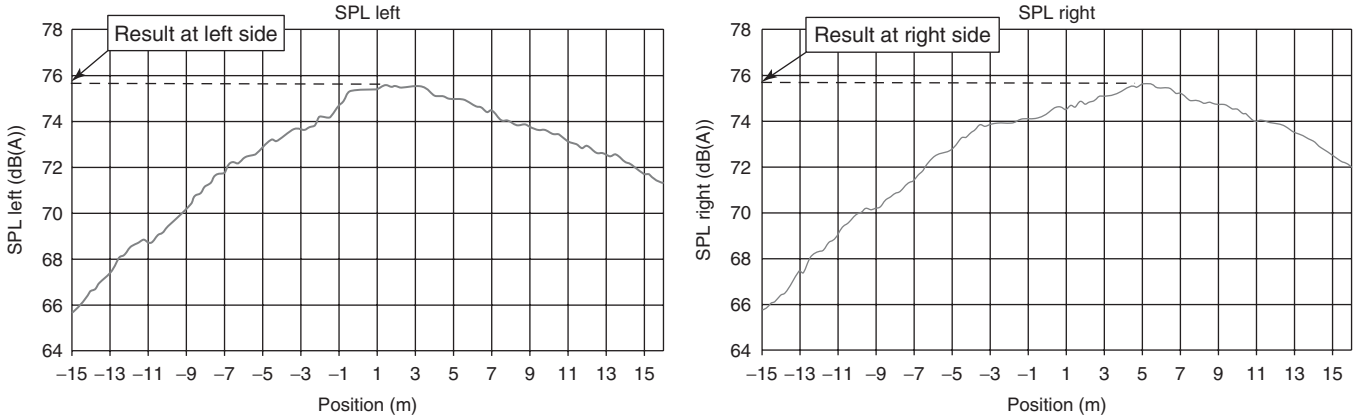


Figure 8. Results of the exterior noise test. (Reproduced by permission of IDIADA.)

S: Rated engine speed—engine rotation speed at which the engine develops its maximum power, according to Regulation 85 (2013).

9.3.1.2.1.2. Vehicles of categories other than M_1 whose engine power is >225 kW The approach speed is the lowest of the following: $V_A = 50$ km/h or V_A corresponding to $N_A = \frac{1}{2}S$

where
V_A: Uniform vehicle speed at the approach of line AA′
S: Rated engine speed—engine rotation speed at which the engine develops its maximum power, according to Regulation 85 (2013)

9.3.1.2.1.3. Vehicles powered by an electric motor The approach speed is the lowest of the following: $V_A = 50$ km/h or $V_A = \frac{3}{4}V_{max}$

where
V_A: Uniform vehicle speed at the approach of line AA′
V_{max}: Maximum vehicle speed declared by the vehicle manufacturer

9.3.1.3 Interpretation of results.

9.3.1.3.1 Vehicles of categories M_1 and N_1. Two runs are performed at each one of the tested gears. When the vehicle has more than four forward gears, we have to test in second and third gear. Then, we will have four results for the second gear and four results for the third gear, according to Figure 8 (two runs and two sides per run). We have to take the maximum of the four results in the second gear and the maximum of the four results in the third gear, and we calculate the average value between them.

$$dB_{avg} = \frac{dB_{max\,2} + dB_{max\,3}}{2} \quad (10)$$

where,
$dB_{max\,2}$ is the maximum of the four results in second gear
$dB_{max\,3}$ is the maximum of the four results in third geard
B_{avg} is the average

When the vehicle is tested in one gear only, we have to take the maximum of the four results. Finally, we round the value and reduce by 1 dB and the obtained value is the final result of the vehicle in motion.

9.3.1.3.2 Vehicles of categories other than M_1 and N_1. Vehicles other than categories M_1 and N_1 are tested at different gears (Section 9.3.1.1.1.2). The final result of the vehicle in motion test is the maximum of all results corresponding to all the tested gears, rounded and reduced by 1 dB.

9.3.2 Test with stationary vehicle

In order to facilitate the checks of the vehicles in-use, the sound level must be measured close to the exhaust outlet.

9.3.2.1 Target engine speed. The target engine speed is defined as follows:

- When $S \leq 5000$ rpm, target speed $N_T = \frac{3}{4}S$
- When $5000 < S < 7500$ rpm, target speed $N_T = 3750$ rpm
- When $S \geq 7500$ rpm, target speed $N_T = \frac{1}{2}S$

where
N_T: target engine speed. Engine rotation speed at which we perform the test

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

S: Rated engine speed—engine rotation speed at which the engine develops its maximum power, according to Regulation 85 (2013).

9.3.2.2 Test procedure. The vehicle will be located in the center of the test area with the gear lever in neutral position and the clutch engaged.

The microphone will be located at a distance of 0.5 ± 0.01 m from the reference point of the exhaust pipe as defined in Figure 9.

The engine speed will be gradually increased from idle to the target engine and held constant when we reach it. Then, the pedal will be rapidly released and the engine returned to idle. The sound pressure will be measured during a time period consisting of a constant engine speed of at least one second and throughout the entire deceleration period. The maximum sound level will be taken as the test value. The result has to be given in A-weighted decibels (dB(A)) and the time weighting will be "fast".

9.3.3 Compressed air noise test

Vehicles having a maximum permissible mass exceeding 2800 kg are subjected to an additional measurement of the compressed air noise with the vehicle stationary, if the corresponding brake equipment is part of the vehicle. This test is described in Regulation 51 (2012).

9.4 Evolution of the exterior noise measurement procedure

The first noise limits for motor vehicles having at least four wheels, adopted by the European Economic Community, entered into force in 1970. The test procedure used at that time was almost similar to the one described above,

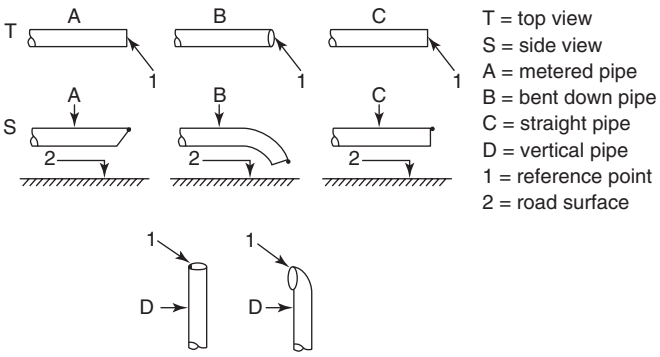


Figure 9. Reference point of the exhaust pipe. (Reproduced from Regulation No 51-02-Rev.1/Add.50/Rev.2/Amend.1.)

Table 1. Exterior noise limits from 1970 to 1992 (in dB(A)).

	Year of Entry into Force			
	1970	1981	1984	1992
Cars	82	80	77	74
Buses	91	85	83	80
Trucks	91	88	84	80

and it was intended to evaluate the noise emitted by the powertrain.

Since 1970, engines and the rest of the powertrain have become much quieter as a result of the big development work done during the last 40 years. This is reflected by the noise limit reduction from 1970 to 1992 shown in Table 1.

These reductions range between 8 dB(A) for the cars and 11 dB(A) for trucks and buses. It is interesting to remark that a reduction of 3 dB means that the acoustic power emission is reduced by 50%. This means that the acoustic power emission of a car is now 16% of the one of 1970 and the acoustic power emission of a truck or bus is now 8% of the one of 1970.

Nowadays, the vehicle noise emission is not any more dominated by the powertrain. The noise generated by the rolling of tires on the road surface (rolling noise) is also significant for the overall noise. This noise is not well evaluated in the actual procedure that measures vehicles in wide open throttle condition only.

Consequently, a new procedure that takes into account the actual level of noise due to vehicle emission in urban traffic has been developed recently although it has not entered into force yet. This procedure is described in references (Regulation No 51–02 – Rev.1/Add.50/Rev.2/Amend.1, 2012) and (ISO 362–1, 2007). This new procedure is described in the following paragraphs.

9.4.1 Vehicles of categories M_1, N_1, and M_2 with test mass not >3500 kg

The test of vehicles of categories M_1, N_1, and M_2 with test mass not >3500 kg according to the new procedure shows the following characteristics.

We perform acceleration and constant speed (steady) tests instead of only acceleration tests. (Figure 10b)

In the acceleration tests, the vehicle reaches the speed of 50 km/h at the center of the track instead of at the beginning (Figure 10a and b).

In the new method, the concepts of reference acceleration and urban acceleration are defined.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

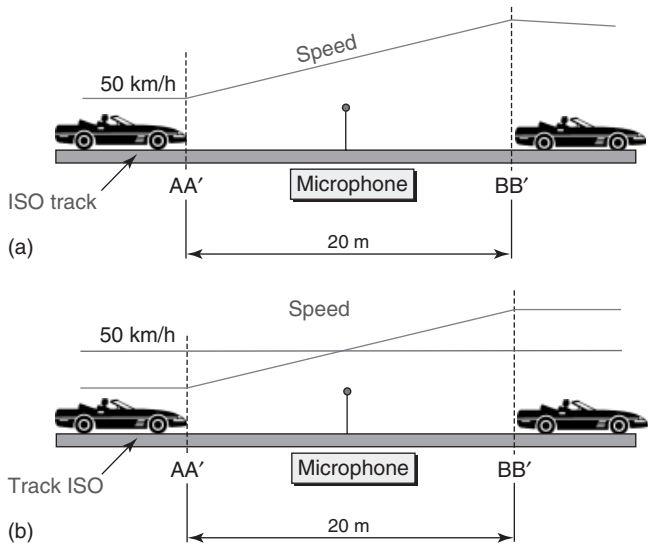


Figure 10. Exterior noise test. (a) Current method with wide open throttle acceleration test only. (b) New method with wide open throttle acceleration and constant speed tests. (Reproduced by permission of IDIADA.)

9.4.1.1 Reference acceleration. Reference acceleration is defined according to the following expression:

$$a_{\text{wot ref}} = 1.59 \cdot \log(\text{PMR}) - 1.41 \tag{11}$$

where

$a_{\text{wot ref}}$ is numerical value of the reference acceleration in m/s^2

PMR is the power to mass ratio, a dimensionless quantity according to the equation

$$\text{PMR} = \frac{P_{\text{n}}}{m_{\text{t}}} \cdot 1000 \tag{12}$$

where

P_{n} is the numerical value of the engine power measured according to Regulation 85 (4), expressed in kW

m_{t} is the numerical value of the vehicle mass, expressed in kg

9.4.1.2 Urban acceleration. Urban acceleration is defined according to the following expression:

$$a_{\text{urban}} = 0.63 \cdot \log(\text{PMR}) - 0.09 \tag{13}$$

where

a_{urban} is the numerical value of the urban acceleration in m/s^2

PMR is the power to mass ratio, as defined in 12

Both accelerations are characteristic parameters of the vehicle.

In the new method, reference acceleration is the required acceleration in the test track. We can find two scenarios.

First scenario: the vehicle has a gear position i whose corresponding wide open throttle acceleration is in a tolerance band of $\pm 5\%$ of the reference acceleration.

$$\frac{|a_{\text{wot } i} - a_{\text{wot ref}}|}{a_{\text{wot ref}}} \leq 0.05 \tag{14}$$

where

$a_{\text{wot ref}}$ is the reference acceleration as defined in Equation 12

$a_{\text{wot } i}$ is the wide open throttle acceleration corresponding to the gear position i

In this case, we perform the acceleration measurements in this gear.

Second scenario: none of the gear ratio positions gives a wide open throttle acceleration that fulfills the condition in Equation 14. In this case, we choose a gear ratio i whose wide open throttle acceleration $a_{\text{wot } i}$ is higher than the reference acceleration $a_{\text{wot ref}}$ and another gear ratio $i+1$ whose wide open throttle acceleration $a_{\text{wot } (i+1)}$ is lower than the reference acceleration $a_{\text{wot ref}}$.

Then, we test both gears and the final acceleration noise level $L_{\text{wot rep}}$ is the weighted average of sound levels of both gears, and the weighting factor depends on the differences between the achieved accelerations in the different gears.

In the constant speed tests, the vehicle passes the test track at 50 km/h in the same gear or gears as used in the acceleration tests. When we perform the test in two gears, the final constant speed noise level $L_{\text{crs rep}}$ is the weighted average of sound levels of both gears, and the weighting factors are the same as the ones used in the acceleration tests.

The urban acceleration is defined in Equation 13. This expression is the result of statistical studies of the behavior of different vehicles, driven in different driving conditions (economical, normal, and aggressive) in different cities.

The urban noise level is a weighted summation of sound levels from acceleration and constant speed tests, where the weighting factor depends on the ratio between the urban acceleration a_{urban} and the reference acceleration $a_{\text{wot ref}}$. This weighting factor is called *part power factor* for urban driving k_{p} and it is defined as:

$$k_{\text{p}} = 1 - \frac{a_{\text{urban}}}{a_{\text{wot ref}}} \tag{15}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

where
k_p is the part power factor
a_{urban} is the urban acceleration as defined in 13
$a_{wot\ ref}$ is the reference acceleration as defined in 11

The part power factor ranges between 0 and 1. When a car has a part power factor close to 1, this means that the reference acceleration is much higher than the urban acceleration ($a_{urban} \ll a_{wor\ ref}$). This vehicle uses a small part of its power in urban acceleration condition. This is the situation of high powered cars.

When a car has a part power factor close to 0, it has the reference acceleration close to the urban acceleration ($a_{urban} \approx a_{wor\ ref}$). This vehicle uses almost full power in urban acceleration condition. This is the situation of low powered cars.

The final test result is calculated by combining the levels from acceleration and constant speed tests using the following equation:

$$L_{urban} = k_p \cdot L_{crs\ rep} + (1 - k_p) \cdot L_{wot\ rep} \qquad (16)$$

where
L_{urban} is the urban noise
k_p is the part power factor as defined in 15
$L_{crs\ rep}$ is the noise level in constant speed
$L_{wot\ rep}$ is the noise level in acceleration

High powered vehicles have a part power factor close to 1; consequently, according to expression 16, its urban noise is dominated by the noise level measured in constant speed condition. On the contrary, low powered cars have a part power factor close to 0 and, according to Equation 16, its urban noise is dominated by the noise level measured in acceleration condition.

9.4.2 *Vehicles of categories M_2 with test mass >3500 kg, N_2, M_3, and N_3*

For other categories of vehicles, the test procedure is less complex. The wide open acceleration test has to be performed within a target speed range and a target vehicle speed range instead of a target acceleration. Testing can be done in one gear if target ranges are reached in this gear. If no gear fulfills the target conditions for the vehicle speed, we have to test two gears, one above and one below the prescribed speed range. The constant speed test is not included for these vehicle categories.

10 AUTOMOTIVE APPLICATIONS

10.1 Evaluation of ambient urban noise

Most of the investigations on acoustic pollution are based on the measurement of the levels of ambient noise produced by different sound sources and especially by transport vehicles. It has been demonstrated that traffic is the most important and widespread source of noise in industrial countries.

In order to express the existing relationship between exposure to noise and the subjective responses of people affected by this environmental factor, all the studies have used various noise indicators. However, in fact, it has been shown that all these indicators are closely related to each other. One of these is the equivalent continuous sound level measured during a 24-h period, (L_{eq} (24 h)). It can be used, with certain limitations, to predict the general response of a community (subjective nuisance for the residents) in the face of the impact produced by a wide variety of different sound sources. The L_{eq} (24 h) evaluates the average value of the noise level to which a predetermined zone is exposed during a period of time. In general, the L_{eq} measured in an interval of T time is calculated by means of the expression:

$$L_{eq}(T) = 10 \cdot \log\left(\frac{1}{T}\int_0^T \frac{p^2(t)}{p_0^2}\mathrm{d}t\right) \qquad (17)$$

That is to say, its value is the average constant amplitude of noise with the same acoustic energy as that of the noise measured. Figure 11 illustrates the time variation of an arbitrary acoustic signal during an interval of time T, its energy, and the value of L_{eq} associated with this interval, that is, L_{eq} (T). In general, urban noise measurements show a frequency range between 125 and 2 kHz, which is typical of the noise generated by traffic. Above these frequencies, the spectral level decays progressively. These bands of relative low frequency tend to be easily transmitted through building structures (walls, windows, foundations, etc.), making the impact on the population still greater than expected.

More recent investigations have shown that the characterization of ambient noise in a given urban location with the L_{eq} parameter alone could be incomplete. This is particularly true for those environments that experience big variations in noise level during the time interval studied. In these circumstances, we also need to describe the statistic and time variability of the noise level as a random variable.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

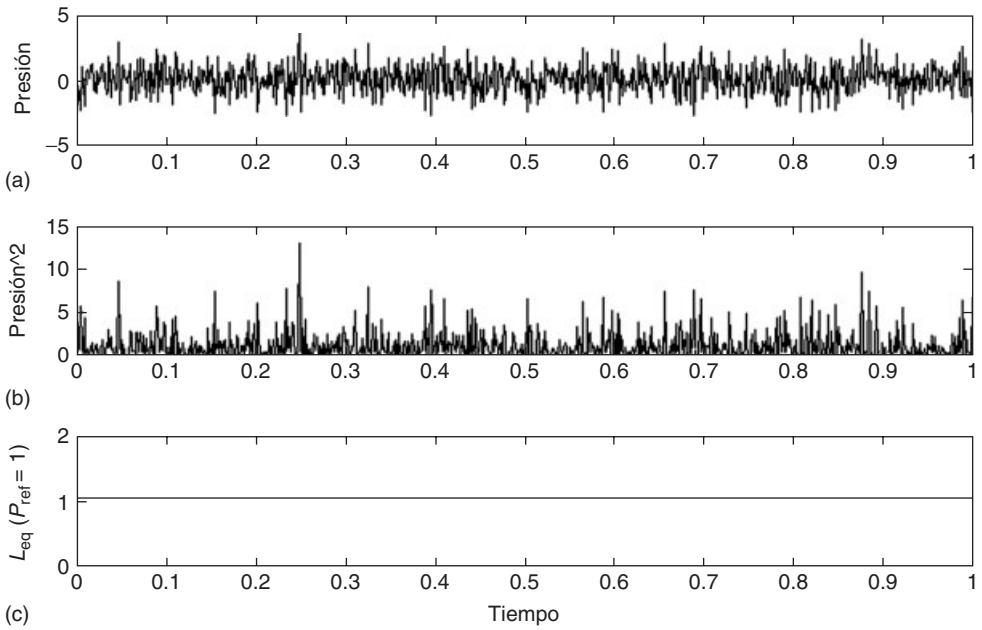


Figure 11. Example of an acoustic time signal (a), its energy as a function of time (b), and the equivalent average level L_{eq} *(T)* for $T = 1$ (c). (Reproduced by permission of IDIADA.)

10.2 Analysis of contribution of external noise sources

The above-described procedures (current and new) do not give information about the structure or the main components of the total noise emitted by a vehicle. In order to understand this structure, we should bear in mind that the noise produced by vehicles is composed of the contribution of the following sources:

- Engine
- Intake system
- Exhaust system
- Rolling
- Aerodynamic

The contribution of each one of these acoustic sources to the overall noise level changes, both in level and spectral content, depending on the vehicle speed and/or the engine speed. The reduction in vehicle noise emissions achieved in the last years is linked to the increasing knowledge of the contribution associated with each noise source present in a vehicle. This section describes a method for analyzing the contribution that each acoustic source in a vehicle makes to the total noise. The results obtained in IDIADA for a group of vehicles show that rolling noise has acquired a notable importance as a source that contributes to the external measured noise both in the pass-by test and in normal driving conditions. In the case of the pass-by test described, we can consider that the contributing acoustic noise sources are:

- Engine
- Intake system
- Exhaust system
- Rolling

Depending on the type of vehicle and the problem, we can consider others such as gearbox, transmission system, radiation from the walls of the exhaust and intake mufflers, radiation of the catalytic converter, and the like.



Figure 12. Vehicle with minimum or residual noise by means of the use of intake and exhaust jumbo mufflers and engine encapsulation. (Reproduced by permission of IDIADA.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Obtaining a vehicle that complies with the levels marked in the Directive (EU Directive 92/97/EEC amending Directive 70/157/EEC, 1992) and the Regulation (Regulation No 51–02 – Rev.1/Add.50/Rev.2/Amend.1, 2012) obviously happens by achieving partial reductions in the levels of noise radiated by some or all of the sources that contribute to the level of exterior noise of the vehicle. In order to get these partial reductions, we need to know which sources radiate and their contribution to the global levels of emitted noise. To do this, *source decomposition studies* are carried out. These studies allow us to quantify the contribution of each source to the global level of exterior noise and also to study it as a function of the position of the vehicle on the track, the vehicle side, and the vehicle speed and/or engine speed. The contribution could be expressed as a percentage of the total or as a contribution in dB(A). The source decomposition method consists of reducing as much as possible the different noise sources until a vehicle with minimum or residual noise has been obtained. The vehicle is equipped with additional intake and exhaust super-silencers and the engine is encapsulated as shown in the example in Figure 12.

Next, the source under study is uncovered and the levels of noise measured during the pass-by test in both states are compared. The difference between the level of this last test and that of minimum or residual noise is the contribution of that source. The contribution of each source F_i in dB(A) along the position of the vehicle on the track is given by:

$$F_i = 10 \cdot \log\left(10^{\mathrm{NF}_i/10} - 10^{\mathrm{NR}/10}\right) \tag{18}$$

where
F_i: contribution of the i-th source.
NF_i: level of noise with the i-th source unshielded.
NR: level of minimum or residual noise.

The percentage contribution of an acoustic source to the level of global noise is calculated using the following equation:

$$F_i(\%) = 100 \cdot \left(10^{\mathrm{NF}_i/10}/10^{\mathrm{NT}/10}\right) \tag{19}$$

where

F_i: contribution of the i-th source.
NF_i: level of noise with the i-th source unshielded.
NT: level of total noise.

Figure 13 shows the result obtained for the contribution of the exhaust system of the vehicle (Figure 12) to the total noise of the vehicle (second gear).

If this process is repeated for each source of interest, the contribution of each one of these sources to the overall noise level is obtained. Figures 14–17 show the noise contributions of the exhaust system, the intake system, the engine, and the tire rolling in second and third gear, all as a function of the position of the vehicle shown in Figure 12. Values are expressed as a percentage of the total level in Figures 14 and 15 (see Formula 19) and in dB(A) in Figures 16 and 17 (see Formula 18).

We can see that the most important noise contributions come from the engine and the tires both in second and in third gear. This result is general for current passenger cars, which is illustrated by the comparative graphs (Figures 18 and 19) obtained in IDIADA between six vehicles, four diesel (vehicles A, B, D, E), and two petrol (vehicles C and F), of the same category and currently on the market. Figures 18 and 19 show the measured contribution of the different noise sources of the above vehicles for the position in which the maximum noise level is obtained during the pass-by test.

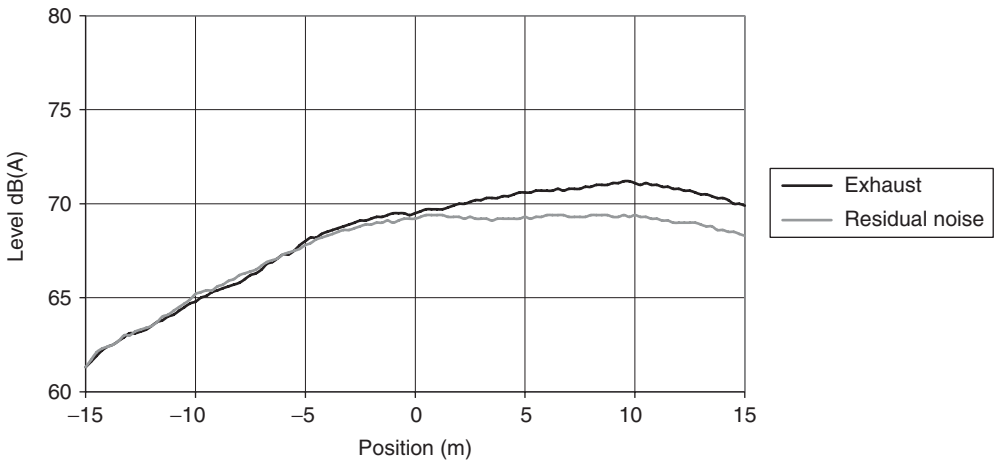


Figure 13. Contribution of the exhaust system to the global noise (second gear, left side). (Reproduced by permission of IDIADA.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

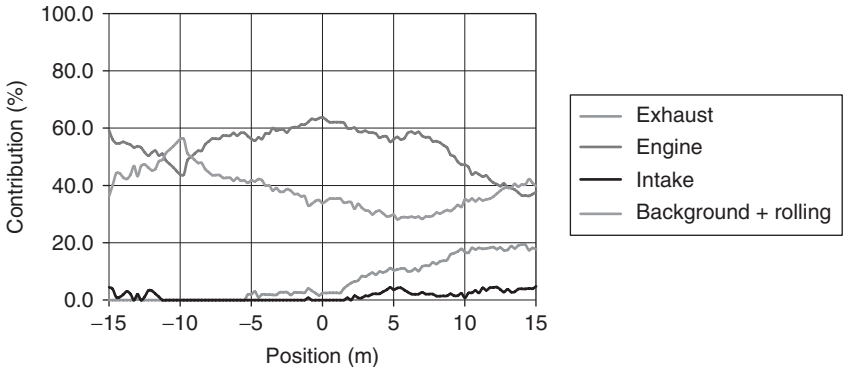


Figure 14. Contribution of acoustic sources to the global noise level as a percentage of the total noise measured (second gear). (Reproduced by permission of IDIADA.)

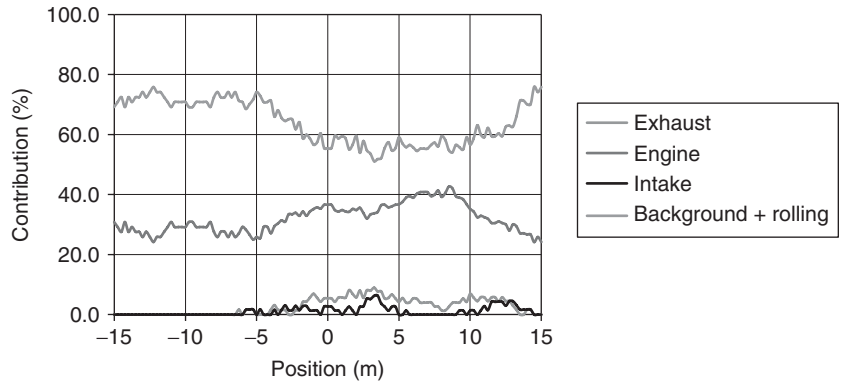


Figure 15. Contribution of acoustic sources to the global noise level as a percentage of the total noise measured (third gear). (Reproduced by permission of IDIADA.)

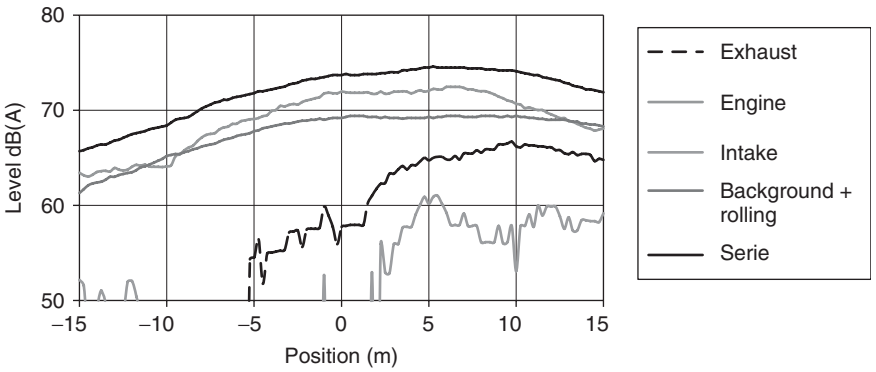


Figure 16. Contribution of acoustic sources to the global noise level in dB(A) (second gear). (Reproduced by permission of IDIADA.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

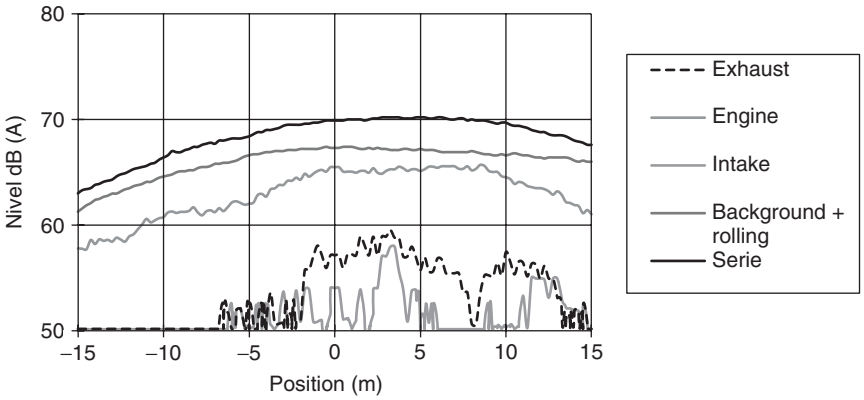


Figure 17. Contribution of acoustic sources to the global noise level in dB(A) (third gear). (Reproduced by permission of IDIADA.)

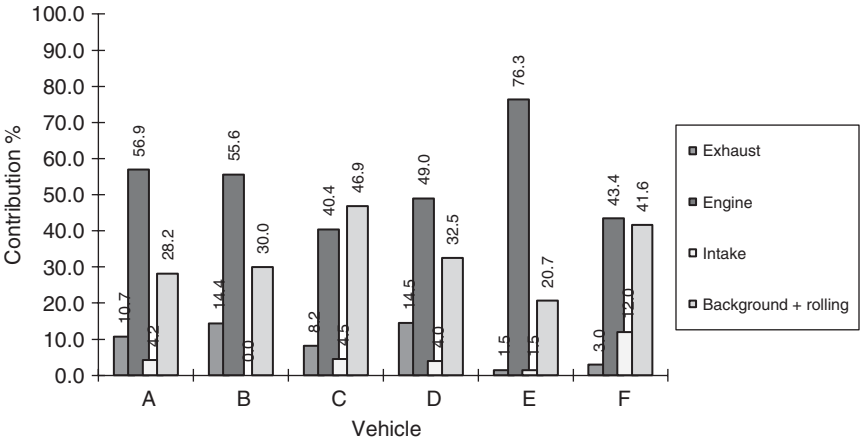


Figure 18. Comparison of noise sources for diesel (A, B, D, and E) and petrol (C and F) vehicles (second gear). (Reproduced by permission of IDIADA.)

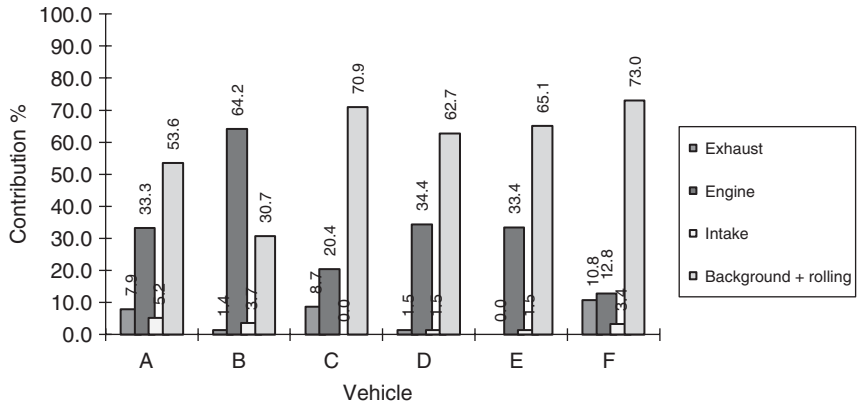


Figure 19. Comparison of noise sources for diesel (A, B, D, and E) and petrol (C and F) vehicles (third gear). (Reproduced by permission of IDIADA.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Figures 18 and 19 demonstrate the small contribution that intake and exhaust systems have on the global noise emitted by current vehicles and that the greatest contribution is from the engine and tires.

REFERENCES

EU Directive 92/97/EEC amending Directive 70/157/EEC on the approximation of the laws to the Member States relating to the permissible sound level and the exhaust system of motor vehicles. (1992) Council Directive of 10th November 1992. Brussels. Official Journal of the European Commission L371, December 19.

ISO 362. (1998) Acoustics – Measurement of noise emitted by accelerating road vehicles – Engineering method, Geneva, Switzerland.

ISO 362–1. (2007) Acoustics – Measurement of noise emitted by accelerating road vehicles – Engineering method Part 1: M and N categories. Geneva, Switzerland.

Regulation No 51–02 – Rev.1/Add.50/Rev.2/Amend.1. (2012) Uniform provisions concerning the approval of motor vehicles having at least four wheels with regard to their noise emissions. Regulation UN/ECE, Geneva, Switzerland, May 23.

Regulation No 85–00 – Rev.1/Add.84/Rev.1. (2013) Uniform provisions concerning the approval of internal combustion engines intended for the propulsion of motor vehicles of categories M and N with regard to the measurement of the net power and the maximum 30 minutes power of electric drive trains. Regulation UN/ECE, Geneva, Switzerland, August 21.

FURTHER READING

Beranek, L.L. (1980) *Noise Reduction*, Krieger.

Kinsler, l.E. (1982) *Fundamentals of Acoustics*, John Wiley & Sons.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto245
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Human Machine Interface Design in Modern Vehicles

Stefan Becker[1], Parrish Hanna[2], and Verena Wagner[3]

[1]*Ford Werke GmbH, Cologne, Germany*
[2]*Ford Motor Company, Dearborn, MI, USA*
[3]*University of Graz, Graz, Austria*

1 Introduction: What it Means—Definition and Scope of HMI 1
2 Who Will Use it: psychology and physiology of the user 3
3 How it has Developed Over Time: a Brief History of Automotive HMI 5
4 How to Make a Good HMI: Development Process 5
5 What it is: Interface Design 10
6 What it is: Interaction Design 13
7 What it Will be: Future Trends 13
Related Articles 15
References 15

1 INTRODUCTION: WHAT IT MEANS—DEFINITION AND SCOPE OF HMI

The human machine interface (HMI) in a vehicle is subject to constant change. The growth in number and complexity of technical systems in the vehicle cockpit confronts the driver with new displays, components, and interaction logic. It is estimated that the majority of all automotive innovations is in the area of electronics. The majority of these electronic innovations communicate with the driver and/or need to be monitored and/or controlled by the driver. Therefore, an appropriate HMI is of fundamental relevance for the easy, safe, and pleasant control of a vehicle. Nowadays, the human machine interface can be seen as the "Face" of the electronic system. Customer research has demonstrated that positive and negative experiences with these systems can massively influence the overall acceptance of a vehicle (Ford internal investigations).

On the background of the fundamental relevance for business success, this chapter aims to provide an overview to the HMI:

– *Definition and scope*: What HMI is and has to cover (Section 1)
– *User*: What we know about the user and his or her needs (Section 2)
– *History*: How HMI concepts, displays, and controls are developed over time (Section 3)
– *Process*: What the HMI development process is (Section 4)
– *Interfaces*: How displays and controls can be classified (Section 5)
– *Interaction*: What the fundamentals of interaction design are (Section 6)
– *Trends*: What future innovations in HMI can be envisaged (Section 7).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

1.1 From human machine interface to human machine interaction

In the beginning, automotive HMI focused on the interface. Regions of interest have been displays and controls located in the cluster or the center stack of a vehicle as well as in the steering wheel. In addition to mechanical components, HMI was also interested in nonmechanical interfaces such as voice control. The fact that system functionality becomes more complex leads to a stronger dialog orientation. So the "I" in HMI has changed its meaning from interface to interaction, which meant thats nowadays HMI stands for human machine interaction, which also includes states and transitions and not only displays and controls (Figure 1).

The term *interaction* describes the circumstances that "(. . .) we no longer simply use machines, we interact with them" (Suchman, 1990, p. 25), which also includes that "each action by the user affects an immediate machine response" (Suchman, 1990, p. 25). Interfaces can be seen as one important part of the interaction between humans and machines.

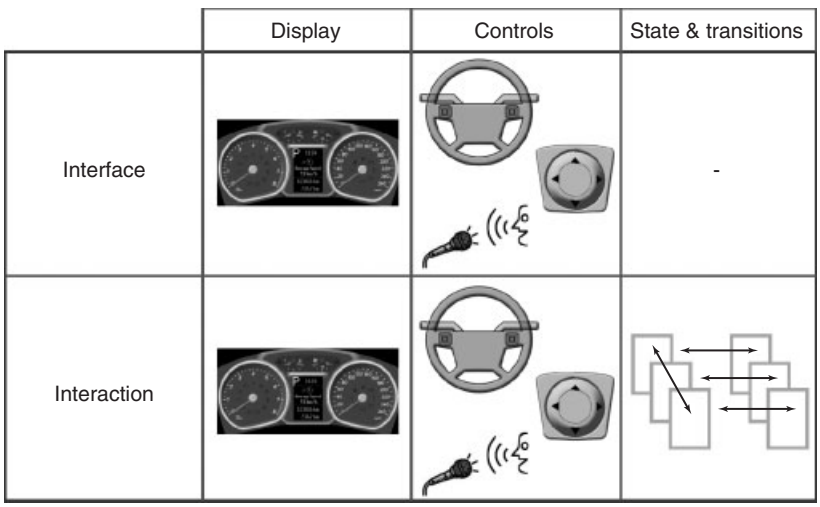
1.2 Scope: infotainment, comfort features, and assistance

Automotive HMI deals with three different scopes: infotainment, comfort features, and driver assistance features (see Table 1 for more details).

For these three scopes, both generic and special issues result. Issues that will be addressed in the automotive HMI development process include the following:

- General issues
  - o How should the system be switched "on" or "off"?
  - o How should all relevant information be configured and presented?
  - o How is the status of the system displayed?
  - o Does the system fulfill user needs?
  - o What is/are the basic function(s) of the system?
  - o Are the functions and the dialogs easily comprehended by the user?
- Special issues
  - o Can the system be controlled easily?
  - o What are the appropriate controls to accomplish an accurate, safe, and comfortable interaction with the system?
  - o Is the system distracting while driving?
  - o Do the dialogs and graphics of the system create a positive "experience/user experience"?

The primary components is of an appropriate automotive HMI is the consistent logic of operation throughout the entire cockpit, which includes the operation of the cluster display as well as the display(s) of the center stack unit. Complimentary use of well-known operation paradigms enables the easy, intuitive, fast, and safe operation of the system and leads to minimum driver distraction. The HMI of a vehicle can also be seen as an important contributor to strengthen brand identity and customer restraints.



Note. □ … *states,* <-> … *transitions*

Figure 1. From human machine interface to human machine interaction.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Scopes of automotive HMI.

Scope	Examples
Infotainment	• Trip computer • Radio and media • Navigation
Convenience	• Power sliding door; power tailgate • Keyless vehicle • Assisted parking • Vehicle personalization and seat adjust
Assistance	• Longitudinal control (cruise control; speed limiting, distance control; Stop&Go) • Lateral support and control (blind spot monitoring; lane departure warning and intervention) • Traffic sign recognition; night vision

2 WHO WILL USE IT: PSYCHOLOGY AND PHYSIOLOGY OF THE USER

For the definition of system requirements, the understanding of the user from a psychological (see, e.g., Gerrig, 2012) and physiological perspective (see, e.g., Bear, Paradiso, and Connors, 2007) is fundamental for a safe system operation (see, e.g., Salvendy, 2012) and product acceptance (Becker, 2004).

2.1 Perception

In terms of driving a car and operating any kind of on-board HMI, visual perception is the prime source of information. Visual perception reflects the activity of the peripheral sensor ("eye") and its signal analysis on the so-called visual cortex. Beyond readability (see Section 2.5), four typical characteristics of that system have to be considered: contrast perception, color blindness, movement perception, and geometrical feature extraction.

Contrast perception is fundamental for vision and any identification of meaningful objects. This "job" is performed by specific neurons in the retina as well as visual cortex. The better the contrast is, the faster the objects can be detected and identified. Therefore, a good contrast ratio can be seen as the first requirement based on human factors (Figure 2).



Figure 2. Contrast requirements.

Color blindness is the next perceptual item to be considered. It occurs in about 8% of the population, making the people unable to see one or more specific colors or even no color at all. Here, good graphic design should compensate for that in using a strategy of "multicoding" (redundant use of color as well as other graphical attributes). This could, for example, mean coding a street on a navigation map by color and thickness of line.

Movement perception strongly directs attention. This effect can be used for guiding the user's eyes to the position of interest on a screen. The realization could, for example, be made by any kind of screen transition animation.

Geometrical feature extraction means a process of detecting the "key" information in complex visual stimuli. This means detection of simple geometrical elements such as lines, angles, and quadrangular structures. The more the screen design reflects in a "minimalistic design" approach such simple structures, the faster the signal detection and the appropriate the reaction could be.

2.2 Cognition

This topic deals with attention, information reduction, and semantic interpretation.

Attention is essential for correct information selection, processing, and reaction. In the automotive environment in particular, the so-called divided attention needs special consideration. This means that any operation of an HMI has to be regulated against the primary task of driving a car. Knowing that a real "timesharing" is not possible for the human brain HMI systems should be designed to keep the driver always in the loop and in control. An intelligent definition of controls in the steering wheel, intelligent dialogs, and use of voice control is of substantial help.

Information reduction and semantic interpretation means that every (overwhelming) sensory input has to be extracted down to the essential information and its fundamental

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

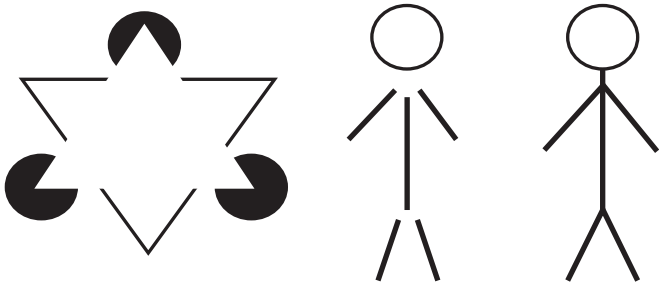


Figure 3. Information reduction by form recognition ("Gestalt").

semantic. The so-called Laws of Good Gestalt (Figure 3) are demonstrating that effect. Therefore, good graphics design should use simple geometrical structures supporting any semantic categorization and building of classes of identical elements (e.g., clustering frequency bands and radio stations graphically).

This information reduction and extraction process is in addition accelerated by graphical elements creating "*mental models*" of a function. Such mental models describe user's assumptions on how a function works. A good HMI design ensures that these assumptions are in line with the *functional models* describing how the system really works.

2.3 Learning and memory

Comprehensibility is not sufficient for an effective HMI operation. Particularly with elderly users, it is well known that memory plays a fundamental role for easy and correct operation. In particular, functions that are used seldom may require a well-working long-term memory. It is obvious that memory load can be reduced by consistency in a dialog, always using identical logical rules, terminology, and symbols. Beyond that, the use of widely accepted informal standards of operation is very helpful. Such so-called stereotypes are telling the story of operation by its *affordance* (Figure 4), guiding the user in operation. This is very often the basis of what is called an *intuitive system*.

In addition, such a process of recognition of essentials can be supported by a picture language using the opportunity of human cognition to think on the basis of the picture. Operational logic is could therefore be translated into the so-called metaphors, describing graphically the way of operation (e.g., the "jukebox" metaphor for operating music folders).

2.4 Emotion and motivation

We study human behavior with the intend of delivering rewarding and pleasant experiences to our consumers in



Figure 4. Examples of stereotypes with high affordance showing the "five-way toggle" stereotype of TV/Sat remote controls (Samsung, Technisat; Comaq). (Reproduced with permission of Stefan Becker. © S. Becker.)

life. "*User experience*" (*UX*) is meanwhile one of the key terms within the HMI-expert language. A pleasant HMI experience can be triggered by comprehensible dialogs and emotionally appealing visualizations. Fulfilling such requirements is a key to success and product acceptance (Hanna, 2012). On the other hand, incomprehensible menus can lead to *disappointments and frustrations*. On a long-term perspective and based on several "bad" experiences, feelings of "learned helplessness" might raise. Such customers might not be willing to invest in the future in "innovative" products but would look back more to "good old days" when everything was pretended easy. This first requires a dialog structure that supports exploring without losing the orientation (see, e.g., the "undo" button on the upper left corner in Microsoft products). Beyond that, special strategies reflecting the needs of more "conservative" versus "innovative" users have to be defined. This could, for example, lead to "beginners" versus "expert" modes in an interaction system.

2.5 Psychomotor performance and age influence

It is a well-known effect that at least in Western hemisphere, the structure of the society is changing toward elderly people. The average buyer of a car in Germany

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is currently around 50 years old with a constant trend to increase. And people within their 80s are still driving a car. HMI designers need to pay special attention to the *presbyopia effect*. This means that the distance for clear vision is moving from 40 cm for 25-year-old to 4 m for 70-year-old people (Goldstein, 2010). Even this impairment can be (partially) compensated by spectacles, it is accompanied by a slower refocusing from near to distance view (slower "*accommodation*"). This effect requires the legibility of "big-fonts views" and may be as well a reason to think about head-up displays having the projected virtual display in front of the car.

2.6 Intercultural differences

Intercultural differences play an important role in the design process of an automotive HMI for a global marketed brand (Becker *et al*., 2011). Different understandings of the meaning of the representations of icons and/or colors can give rise to usability problems of products (Bourges-Waldegg and Scrivener, 1998). Therefore, some issues such as the acceptance and understanding of icons, color preferences of different markets, and preferences of commonly used logic of operation of different culture groups has to be kept in mind during the automotive HMI development process. For example, icons designed to mark points of interest such as graveyards, zoos, and restaurants in the navigation settings of a navigation system have to be designed in a way that users of different cultural areas are able to easily understand them. Sometimes, these cultural differences do not allow the use of unitary/global icons and market-specific icons and symbols have to be implemented. The same is valid for color coding: in Western countries such as Europe or North America, the color "red" is connected to "alarm" and/or a warning. In Eastern countries such as China or Taiwan, the color red has a contrary meaning and stands for "luck" (Zißler-Gürtler, 2002). Also, the way of writing (from left to right or from right to left) can have an impact on the logic of operation of a system, especially when a hierarchical logic of operation is used. On the basis of these examples, it can be advised to investigate cultural characteristics as one important step to better understand customers' needs (Choong and Salvendy, 1998).

3 HOW IT HAS DEVELOPED OVER TIME: A BRIEF HISTORY OF AUTOMOTIVE HMI

In the automotive context, a brief description of the history of the human machine interaction can be illustrated as "from a *classic* radio to head-up display and gesture control."

Figure 5 describes the history of HMI innovations in the consumer electronics (CE) as well as in the automotive environment.

Key milestones in CE were defined by the introduction of the mouse concept (1981), followed by first touch-screen applications (1983), functional integration on one touch screen (in 1990), and smart phones with new interaction paradigms up to high performance voice systems and gesture control (Figure 5). It is obvious that the automotive innovations always followed these CE developments with a certain time gap.

4 HOW TO MAKE A GOOD HMI: DEVELOPMENT PROCESS

To ensure that an automotive HMI is accepted and preferred by costumers, a development process has to be followed that is informed by user needs that are then translated into technical recommendations (for a deeper understanding of the process see: Norman, 1988; Nielsen, 1993; Cooper, Reimann, and Cronin, 2007; Moggridge, 2007; Rogers, Sharp, and Preece, 2011; Nielsen and Budiu, 2012).

The so-called HMI development process (see also Figure 6) usually starts with research that concentrates on different segmentations and a global marketing perspective. On the basis of the research results, goals for the new HMI will be identified and finally defined. These goals help derive recommendations for action. The next step is to describe relevant use cases, which lead to a description of functions that have to be implemented into the new system. Then, the HMI framework definition has to be crafted—which includes the interface framework that defines the controls and displays of the system, the interaction framework (logic of operation), and also the visual framework (graphics). To succeed, the different parts of the development process as well as iterative validation need to occur. This iterative process happens throughout the whole HMI development process. Here, it is described in detail.

4.1 Research: from market segmentation to personas

The first step of the automotive HMI development process can be described as the development of customer and marketing studies that help understand the segmentation and global user needs for the new product. Questions that should be answered with the help of these studies include the following:

- Who are the customers that will use the new system?

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

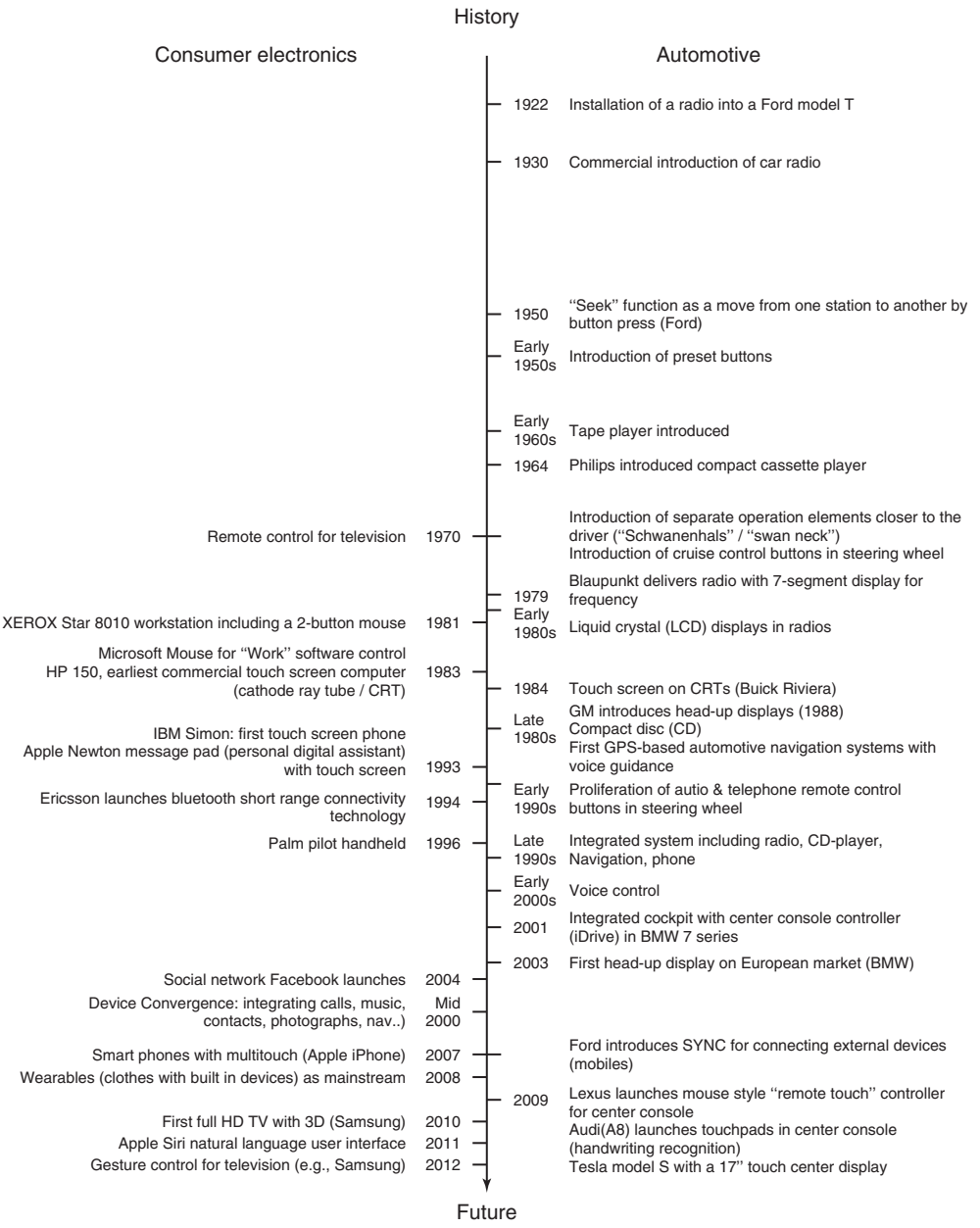


Figure 5. Historical milestones in development of automotive (and consumer) HMI (see also: Bhise, 2012; Wikipedia.org/vehicle audio; Wikipedia.org/Autoradio; Wikipedia.org/Steering wheel; Wikipedia.org/Touchscreen).

- How can they be described both demographically and psychographically?
- Is the product normally used by males or females? One or many?
- What is their cultural background?
- What are their mindsets and motivations for using the system?
- How are they anticipated to use the system?
- Concerning the use of technology, are they innovators or laggards? What are the needs, habits, and expectations of the future users of this product?

The questions that should be answered at this stage explore the dimensions in which possible users or user groups are hypothesized to be different.

A commonly used strategy to define different user groups is called *personas*. Personas can be defined as models that

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

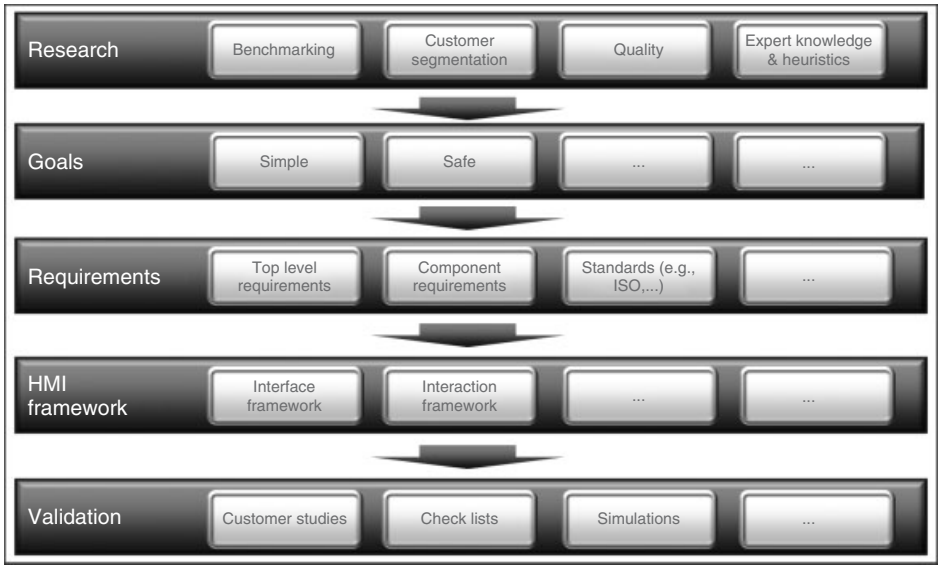


Figure 6. HMI development process: from research to validation.

describe a prototype of the user of the system and his or her needs (Cooper, Reimann, and Cronin, 2007). The representation of personas as "real" people shall help the designers and engineers to develop a product that focuses on real customer needs, habits, and expectations. Normally, personas are represented within 1–2 pages. One important part in the creation of personas is to decide whether a global HMI strategy should be informed by one specific user or if different users are needed to define the personas of a global HMI perspective. In other words, should the strategy of the HMI development process follow market-specific HMI needs or should the goal be to develop one global HMI that has the strength to inspire users all over the world?

4.2 Goals, use cases, and functions

On the basis of the research results described in Section 4.1, goals for the new HMI had to be derived, identified, and finally defined. Some top-level goals can be described as follows:

- comprehensibility,
- controllability, and
- user experience.

Goals can be seen as a first step that describes at a high level what the newly designed system shall provide in order to satisfy user requirements. The next step is "(...) breaking these goals down into the more specific physical or cognitive actions necessary to obtain the goal" (Wickens and Hollands, 1999, p. 6). This can be done with the help of use-case definitions and task analysis. The description of use cases of a new system can be seen as the basis of the requirements definition (see Section 4.3 for more details). In the product development process, use cases describe all tasks that a user can perform with the system (Wickens and Hollands, 1999). The integration of use cases in the automotive HMI development process helps design an appropriate system and avoid errors. One significant benefit of the integration of use case analyses in the early phase of the automotive HMI development process is that potential error sources of the system can be eliminated before the first prototype of the system is built. As an example, one use case of a radio in a vehicle can be that the user not only wants to hear his or her favorite radio station but also prefers to store it as a preset to enable quick access when he wants to listen to the same radio station during his or her next trip. This use case example for a radio task can also lead to the requirement that there has to be a function included in the radio system where radio presets can be stored (e.g., a physical button and a dedicated space on a touch-screen display).

4.3 Requirements

On the basis of the goals, recommendations for action are derived. These result in requirements that the newly developed system has to fulfill. Besides requirements that each organization individually defines for a product, there are ISO standards that a system has to fulfill. Part 10 of the ISO standard 9241–110 (2006) deals with general ergonomic principles that apply to the design of dialogs

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

between humans and information systems. The different principles are as follows:

- suitability for the task (which stands for an appropriate functionality and minimization of unnecessary interactions),
- suitability for learning (guiding the users to minimize the time for learning),
- suitability for individualization (adaptability of system to user and his or her context),
- conformity with user expectations (consistency to user's model of operation),
- self-descriptiveness (comprehensibility by feedback and help functions),
- controllability (controllability of dialog by the user), and
- error tolerance, which means that the systems keep functioning even in case of failure.

Besides the requirements of the ISO standards, Nielsen (1994) introduces 10 usability heuristics for user interface design. These heuristics can be seen in Table 2.

A comparable heuristic system was developed by Shneiderman and Plaisant (2009).

4.4 HMI framework definition

When questions of segmentation and the global perspective are answered, goals of the new system are defined and use cases and functions are described. The next step in the automotive HMI development process is to define the HMI framework of the system. The HMI framework definition includes the interface framework (see Section 5 for more details) that defines the displays (technology, size, orientation, and resolution) and controls (location, type, and functional and logical behaviors) of the system. It also includes the interaction framework (see also Section 6), which specifies the implemented logic of operation including the use of metaphors and stereotypes, the specification of the underlying menu structure as well as the distribution of information on a screen (screen layout), and the information flow between screens. Finally, it also includes the visual framework that defines the graphics that the user will see while interacting with the system (style guide).

4.5 Validation

The validation process of a human machine interface can be described as an iterative one (Wickens and Hollands, 1999). On the basis of the results of customer studies, a refinement of the design of the interface will be done. Further, the refined system will be evaluated by customers and the results can lead to new refinements and so on. Also, the results of the validation process give the developer of a system an answer as to whether the defined system requirements are implemented in an appropriate way. Furthermore, validation results demonstrate whether the purpose is accomplished or not (e.g., the system is reaching the top-level goals and defined tasks can be executed accurately and in a user-friendly way). Therefore, the so-called RESPONSE check list (Becker, 2005), which is shown in Figure 7, can be used especially for driver assistance features.

Concerning validation, the choice of research method is a very important step. The researcher has the choice between a plentitude of different research methods—such as reports of accidents/incidents, press feedback, surveys, field studies, laboratory experiments, task simulations (mock-up or static cockpit simulator, moving-based simulator, and vehicle-based simulator—investigations under real road conditions), or models based on human simulations. They differ in terms of cost, their appropriateness concerning the circumstances in which the system will be used, and the ease of product change because of their results. They also differ in their ease of control and inference and regarding their design relevance of conclusions (Wickens and Hollands, 1999). Figure 8 shows examples for simulators that are normally used to validate automotive HMIs.

It is particularly important to involve users in the evaluation process of an interface! So, during the HMI development process, customers usually assess the actual system that is already in the market against a new reference model. During these evaluations, the customers have to accomplish the most important and typical tasks that are defined during the use case analyses (see Section 4.2). Following Wickens and Hollands (1999), different measures of performance and behavior can be used to measure the performance of a user with a system:

- measures of speed or time,
- measures of accuracy or error,
- measures of workload or capacity demands, and
- measures of preference.

The choice of the measures is just as important as choosing the appropriate research method.

Designing an automotive HMI has to be seen as a challenge that deals with the circumstances that driving effectively requires the safe control of the vehicle to always be the first task. Therefore, the HMI of an in-vehicle system has to be designed by keeping in mind that using the system should "only" be the secondary task, besides driving, which can be demanding by itself. It is also important that the human–machine interface is designed in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Ten usability heuristics for user interface design.

Heuristic	Description
1. Visibility of system status	The system should always keep users informed about what is going on, through appropriate feedback within reasonable time
2. Match between the system and the real world	The system should speak the users' language, with words, phrases, and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order
3. User control and freedom	Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialog. Support undo and redo
4. Consistency and standards	Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions
5. Error prevention	Even better than good error messages is a careful design that prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action
6. Recognition rather than recall	Minimize the user's memory load by making objects, actions, and options visible. The user need not remember information from one part of the dialog to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate
7. Flexibility and efficiency of use	Accelerators—unseen by the novice user—may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions
8. Aesthetic and minimalist design	Dialogs should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialog competes with the relevant units of information and diminishes their relative visibility
9. Help users recognize, diagnose, and recover from errors	Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution
10. Help and documentation	Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large

Reproduced from Nielsen (1994). © John Wiley & Sons, Inc. See also http://www.nngroup.com/articles/ten-usability-heuristics/

a way so that the driver always has available to him or her the necessary resources to respond to unexpected events. As a consequence of this, the level of workload can be seen as important safety-related factor in the HMI development (Lumsden, 2011). So, the designer of a system has to know how busy the driver will be using the system during driving. Will the driver be able to respond to unexpected events, how many resources will be required from the driver to fulfill a task with the system, and how does the driver feel using the system? These questions can be summed up in the concept of mental workload and help the developer to design an efficient and easily usable system. Wickens and Hollands (1999) propose three different workload assessment techniques for the analyses of mental workload:

- secondary-task technique,
- psychophysiological measures, and
- subjective measures.

With the help of the secondary task technique, resources that will be demanded for the primary task can be made visible. Recording psychophysiological measures such as cardiovascular measures (heart-rate variability and heart rate), electromyogram from facial muscles, pupil diameter, eye movements (saccades and fixations), electrodermal measures (e.g., amplitude, frequency of nonspecific electrodermal responses, and skin conductance level), blood pressure, pulse volume amplitude, and changes in the electroencephalogram (EEG) can show the autonomic or central nervous system activity while operating the system (Boucsein, 2006; Boucsein and Backs, 2009; Wickens and Hollands, 1999). The third workload assessment technique is subjective measures. Two commonly used subjective assessments are the NASA Task Load Index (TLX) scale (Hart and Staveland, 1988) and the subjective workload assessment technique (SWAT) (Reid and Nygren, 1988).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Question	Yes / No / not suitable
Are system reactions understood by other road users? If not can they still control the situation (e.g., system based deceleration without activation of brake lights)?	
Is the driver's attention necessary to keep him in the **physical control loop** while the system is running?	
Is the vehicle **controllable** in the case of a system **malfunction by overruling or switching off** the system?	
Can **system parameters be changed while driving** without causing unexpected behaviour?	
Is the system function **self - explanatory** (i.e., without user manual)?	
Can the operation or observation of the system (e.g., possible driver distraction by displays) be achieved **without a major change in attention distribution** relating to the driving task so that potentially hazardous situations may not occur?	

Figure 7. Excerpt of RESPONSE check list. (Reproduced from Becker *et al*. (2005). © European Commission, RESPONSE project.)

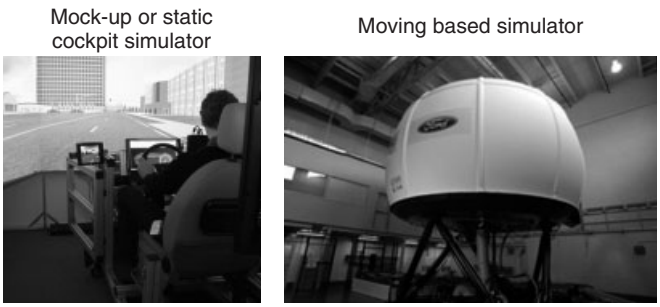


Figure 8. Examples of simulators used for validation of automotive HMIs. (Reproduced with permission from Ford. © Ford.)

5 WHAT IT IS: INTERFACE DESIGN

Interface design is concerned with the classification of displays, controls, and other input possibilities such as voice control or other modalities (e.g., gestures). The *classification of displays* for an automotive system can be seen as the decision which display technologies should be applied. For example, the development engineer has the choice between segmented and full graphics capabilities, nontouch versus touch-screen displays. And if he or she decides to implement a touch-screen display, then it should be between monotouch and resistive touch displays or multitouch and capacity touch displays (see also Table 3).

In the automotive environment, there is a clear trend toward high value displays realized by LCDs with active light emission, color, and high resolution dot matrix. They are designed as the so-called thin film transistor (TFT) displays.

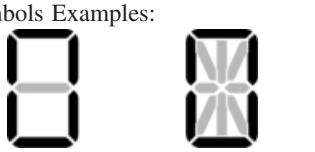
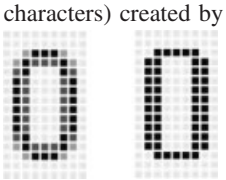
In terms of the *classification of controls* (Table 4), the development engineer needs to choose, for example, between discrete (e.g., on/off) versus analog (e.g., volume control) control modalities, touch versus controller-based operation of the system or a combination of both (touch screen with supplementary list controller mostly implemented as additional rotary), the usage of a keypad versus a touch pad for handwriting input, and more.

5.1 Voice control

Voice control is meanwhile a well-known and accepted technology of interaction. It is used for all menu operations as well as selection of specific content such as addresses for navigation or artists within an audio database. If the system is to be controlled via voice or other modalities such as gestures, then the interface design has to ensure that the commands are embedded into the overall dialog structure. Also, there should be implemented some support from the system for situations in which the system is not able to understand the input of the driver in an appropriate way. This support can, for example, be a list displayed on the multifunctional display where the driver has the choice to select between different possibilities. The main advantage is the reduction of driver distraction by avoiding manual control (e.g., on a navigation entry keyboard) and reading of a display. While, at the beginning, the voice dialog was very much driven by keywords that the user had to learn, voice

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Classification of displays from a HMI point of view perspective (see also Wikipedia.org/wiki/Electronic visual display; www.pcmag.com/encyclopedia)

Light Emission Effect	Active Displays	Passive Displays
	Liquid crystal display (LCD) + backlight (e.g., LCD TV screen) *Electroluminescence* Organic light emitting diode (OLED) *Photoluminescence* Plasma display panel (PDP)	*LCD* *Electrophoresis* Electronic paper (e.g., e-book reader)
Display mode of observation	Direct view displays Picture directly seen on display Example: navigation screen	Projection displays Picture projected Examples: head-up display, laser projection displays
Layout of picture elements	Segmented displays Fixed shape of characters, numbers, and symbols Examples:  7-segment display 14-segment display (source: Ford)	Dot-matrix displays (graphic displays) Any graphical element (including characters) created by a pixel matrix  (source: Ford)
Coloring	Monochrome displays	Color displays
Special view effects	Dual view displays	3-Dimensional (3D)
Control capability	Nontouch Pure display	Touch screen Resistive technology (gaps of two layers are electrically connected by mechanical force of finger on surface layer) Capacitive touch (change in electrostatic field by electrical conductivity of human body)

dialogs (such as Ford SYNC System; Figure 9) are meanwhile flexible in understanding. Dialogs between system and user can be carried out nearly "naturally" based on the opportunity of "natural language understanding (NLU)."

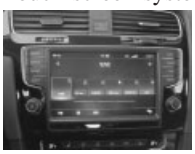
Concerning *gesture communication*, a distinction must be drawn between display-based gestures and camera-based gestures. Display-based gestures are already known from use of touch screens of a smart phone or tablet PC. Nowadays, these well-known gestures can help the user to easily operate with a touch system in a vehicle. There are currently only limited applications on the market such as proximity sensing. Here, more or less hidden buttons are enlarged when the finger of the user is moving toward the screen (e.g., Volkswagen Golf Infotainment system). In the future, another typical and intuitive application might be a wiping gesture for horizontal moving of content.

Camera-based gestures, on the other hand, allow the user to control functions with the help of hand gestures in the space. These operation concepts are well known from video game consoles. However, before a successful introduction into the automotive environment, several open issues have to be solved:

- *Dialog initiation*: How to start a dialog (comparable to Push to Talk button).
- *Semantic bandwidth*: Compared to speech (unlimited bandwidth) gesture, dialogs can only transmit a limited number of dedicated information. Full information bandwidth would require learning of a complex gesture language such as gestures for hearing impaired people.
- *Ambiguity handling*: If the system cannot clearly identify the message, a modality change to written speech on screen would be required (in the case of voice control, such a modality change is not required).
- *Intercultural comprehensibility*: Sometimes specific gestures signalizing cultural-specific messages.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Classification of controls from a HMI point of view.

Type of Control	Generic Function	Typical Application
Push	1-way logic	On/off function
Push/pull	2-way logic, up and down	Opening/closing power windows
Rocker	2-way logic	Volume control
Toggle	4-way	Control of outside mirrors
Toggle	5-way	Dialog control List up/down Menu deeper/higher Ok for select 
Rotary	Analog	Set temperature
Rotary/push	Analog + 1-way logic	Volume + on/off 
Rotary/shift	Analog + 4-way	Multimedia control in center stack (e.g., BMW) Also providing haptic feedback 
Touch pad	Analog (2D)	Handwriting input in center console
Rotary/touch pad	Analog (1D and 2D)	Multimedia controller in center stack with integrated touch pad (e.g., Audi A3, 2013) 
Touch screen	Direct addressing	Multimedia system
"Hybrid" touch screen	Direct addressing on screen + rotary controller	Touch screen system (e.g., Golf 2013) 

(Reproduced with permission from Ford. © Ford.)

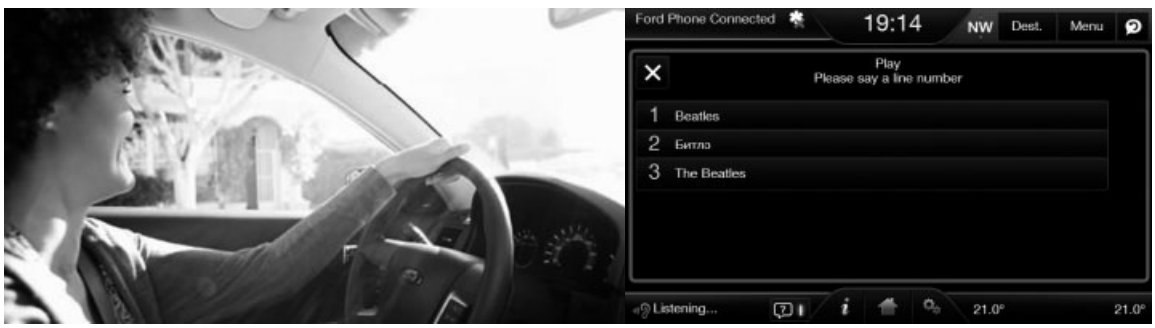


Figure 9. Ford SYNC with voice control system. (Reproduced with permission from Ford. © Ford.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6 WHAT IT IS: INTERACTION DESIGN

In a first step, decisions that have to be made on the overall screen layout include the choice of which information is presented where on the display(s). An important design goal for the overall screen layout is that a consistent grid of information enables a fast orientation of the user and results in a faster finding of content. As an example, a variant of a multifunctional display of Ford is shown in Figure 10. The overall screen layout of this Ford's multifunctional display is divided into three sections: a bottom line that includes information such as the actual time, settings of the air conditioning system, and the charge condition of the bonded mobile phone. On the left side of the display, different information of the above-mentioned classes are shown (e.g., different wave spectrums of radio signals as shown in Figure 10). On the middle-right side of the screen, more detailed information of the selected device (in Figure 10: radio) is presented.

The second step of interaction design is the implemented logic of operations and menus. This means that the organization of screen flows has to be designed. For example, the presented content of the cluster can be organized based on a hierarchal logic. Using a hierarchal logic allows the user to operate with well-known stereotypes like those from other devices such as a mobile phone. A system that uses a hierarchal logic of operation supports the user to easily understand the geometric logic in which the menus and menu entries are arranged within the system. The use of a hierarchal logic for the cluster display can, therefore, be seen as geometrical representation of the system.

And in the third step, interaction design deals with the visualization, which metaphors and stereotypes should be used to facilitate the users' operation, and the logic of operations and menus.

Interface metaphors are the central component of the conceptual model of a system. The metaphors used should help the user to easily understand the functionality without additional help. One example of a well-designed HMI using metaphors is the translation of hardware calculators into software. These software calculators look and work like the hardware the users are used to since years, which makes them easy to use (Rogers, Sharp, and Preece, 2011). Examples for automotive HMI metaphors can be seen in visual displays of driver assistance features. As an example (Figure 11), the radar beam used in the adaptive cruise control driver assistance feature shows the driver the distance to the vehicle in front of their own vehicle. The radar beam is used as a metaphor, which can be intuitively understood by the driver. Another metaphor also used in the automotive HMI is the so-called jukebox metaphor. This metaphor enables the user not only to choose by list browsing (text based search) between different albums but also to search between different album covers that are shown in a graphical way. The user can scroll between the different album covers by vertically moving the covers, like the jukebox had worked in formerly.



Figure 10. Example of a multifunctional display of Ford. (Reproduced with permission from Ford. © Ford.)

7 WHAT IT WILL BE: FUTURE TRENDS

7.1 Goals: combination of safety and experience

A trend can be seen toward a combination of the safe usage of a system and user experience. So, it becomes more and more important to develop in-vehicle infotainment systems that do not distract the driver during driving. Also, because of the increasing number of driver assistance systems implemented in vehicles, the controllability of them can be seen as another important part in future vehicles. It will be important that the driver is able to understand the system behavior in every traffic situation and in every use case in which the driver assistance features may use or intervene automatically. Also, considerations will be important so as to ensure that the driver always has the ability to intervene and override or switch off these systems without a safety-related risk. The objective is a user experience that delivers on the so-called fascination of simplicity and is designed in such a way that results in extreme ease of operation and use (Hanna, 2005, 2012).

7.2 Interfaces

7.2.1 Universal interfaces and connectivity

The tracking of trends in the industry of consumer products shows a trend toward a universal usage of smart phones and/or tablet PCs. At this stage, smart phones are used to control machines, home environments, offices, and, maybe

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

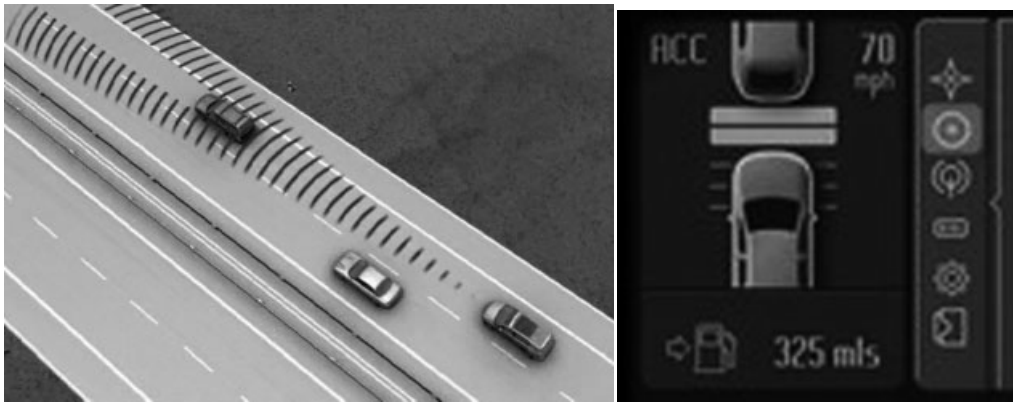


Figure 11. Example for automotive HMI metaphor: the radar beam used in the adaptive cruise control driver assistance feature as part of the driver information (cluster). (Reproduced with permission from Ford. © Ford.)

in the nearer future, vehicles. Also, the usage of smart phones allows a fast reaction on changing use cases, because every new use case that comes up can be covered by offering a new app for mobile devices. The integration of mobile devices can deliver new interfaces in a vehicle, which enable new ways of interaction in the automotive context. In this context, issues regarding driver distraction become important as well.

7.2.2 Augmented reality

Augmented reality describes a technology that adds virtually generated objects and elements into the real-world scene. To create a virtual overlay, which is added to the real-world scene, users normally have to wear special goggles, but in the future context of the automobile, these accessories may not be required. Numerous concepts and technologies are evolving that deliver augmented reality elements via an existing technology: the head-up display. Also, some automotive manufacturers have presented their ideas to implement augmented reality into vehicles at motor shows of the past years [see, e.g., the "DICE"-concept (dynamic and intuitive control experience) of Mercedes-Benz—2012 presented at the CE Show in Las Vegas]. Other conceptual demos include the enhanced vision system of GM, an augmented reality layer placed on the rear seat windows such as Toyota presented with its Window to the World interactive vehicle concept, and finally the "Window of Opportunity" presented by GM.

7.3 Graphic-based interaction

Nowadays, interaction between a driver and a vehicle happens through text. This causes two different issues for the application of a new HMI of a vehicle: First, because of a global selling of vehicles of one brand, the wording used has to be translated into a lot of different languages. Second, complex functions are often very hard to understand when they are "only" described by words (without pictures). One example which demonstrates that shortfalls of a text-only system are user manuals and specifically the descriptions of how driver assistance features will work and intervene. One in-vehicle system that solves this problem is the onboard user manual "your BMW in brief" offered by BMW. This system allows the driver to get more information on selected functions of the vehicle without knowing the correct wording of the topic. The functions are represented as pictures and the user can select the different pictures to get more information on the underlying system.

Whatever the future will be, designing HMI solutions is one of the most challenging activities in the automotive environment. HMI is meanwhile at all OEMs in the focus of cross-functional teams incorporating people from

- electric (function and technology),
- design (seamless User Experience of HMI and Interior),
- marketing and planning (Global user requirements and product definition), and
- quality department (short- and long-term learning from market).

It needs the collective intelligence of engineers and computer scientists, interaction and visual designers, as well as psychologists and physiologists. At least one thing is clear: the working life of HMI people will not be boring.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

RELATED ARTICLES

Intelligent Transport Systems: Overview and Structure (History, Applications, and Architectures)
Evolution and Future Trends
Applications—Intelligent Vehicles: Driver Information
Driver Assistance
Applications—Intelligent Vehicles: Autonomous Vehicles
Driver Distraction
Historical Overview of Electronics and Automobiles: Breakthroughs and Innovation by Electronics and Electrical Technology
Telecommunications
Active Safety, Pre-collision Safety and Other Safety Products (millimeter wave, image recognition, laser)

REFERENCES

Anderson, J.R. (2009) *Cognitive Psychology and its Implications*, 7th edn, Palgrave Macmillan, New York.

Bear, M.F., Connors, B.W., and Paradiso, M.A. (2007) *Neuroscience – Exploring the Brain*, Lippincott Williams & Wilkins, Baltimore, USA.

Becker, S. (1996) Mental Models, Expectable Consumer Behaviour and Consequences for System Design and Testing. In *Proceedings of the 1996 IEEE Intelligent Vehicles Symposium*. Tokyo, Piscataway: IEEE Service Center, pp. 313–318.

Becker, S. (2004) Usability und produktakzeptanz in *Automotive Management - Strategien und Marketing in der Automobilwirtschaft* (eds B. Ebel, M. Hofer, J. Al-Sibai), Springer, Berlin, pp. 250–270.

Becker, S., *et al*. (2005) RESPONSE 2: Final report. ADAS – From market introduction scenarios towards a Code of Practice for Development & Evaluation, Deliverable D4. RESPONSE Consortium. European Commission, Cologne/Brussels.

Becker, S., Mariet, R., Menrath, I., *et al*. (2011) One Ford – one HMI? Human machine interface in between global brand identity and regional customer requirements in *VDI-Berichte 2132, Elektronik im Kraftfahrzeug*, VDI Verlag GmbH, Düsseldorf, pp. 625–638.

Behnke, B. (2009) *Betriebliche Gesundheitsförderung älterer Arbeitnehmer*, GRIN, Norderstedt.

Bhise, V.D. (2012) *Ergonomics in the Automotive Design Process*, CRC Press, Taylor & Francis Group, Boca Raton, FL.

Boucsein, W. (2006) Psychophysiologische methoden in der ingenieurspsychologie in *Sonderdruck aus Enzyklopädie der Psychologie: Themenbereich D Praxisgebiete: Serie III Wirtschafts-, Organisations- und Arbeitspsychologie. Band 2: Ingenieurspsychologie* (eds B. Zimolong and U. Konradt), Hogrefe, Göttingen, pp. 317–358.

Boucsein, W. and Backs, R.W. (2009) The psychophysiology of emotion, arousal, and personality: methods and models in *Handbook of Digital Human Modeling. Research for Applied Ergonomics and Human Factors Engineering* (ed. V.G. Duffy), CRC Press, Boca Raton, pp. 35-1–35-18.

Bourges-Waldegg, P. and Scrivener, S.A.R. (1998) Meaning, the central issue in cross-cultural HCI design *Interacting with Computers*, **9**, 287–309.

Choong, Y.-Y. and Salvendy, G. (1998) Design of icons for use by Chinese in mainland China *Interacting with Computers*, **9**, 417–430.

Cooper, A., Reimann, R., and Cronin, D. (2007) *About Face 3. The Essentials of Interaction Design*, Wiley Publishing, Inc., Indianapolis.

Gerrig, R.J. (2012) *Psychology and Life*, 20th edn, Pearson Education Inc., New Jersey.

Goldstein, E.B. (2010) *Sensation and Perception*, 8th edn, Wadsworth, Belmont, CA.

Hanna, P. (2005) Customer Storytelling at the Heart of Business Success. Boxes and Arrows, July 16th 2005.

Hanna, P. (2012) The evolution of simplicity and meaning *Journal of Product Innovation Management.*, **29**, 352–354.

Hart, S.G. and Staveland, L.E. (1988) Development of NASA-TLX (task load index): results of empirical and theoretical research in *Human Mental Workload* (eds P.A. Hancock and N. Meshkati), North Holland Press, Amsterdam, pp. 139–183.

ISO 9241–110 (2006) Ergonomics of Human System Interaction. Dialog Principles, International Organization for Standardisation, Geneva, Switzerland.

Lumsden, J. (2011) *Human-Computer Interaction and Innovation in Handheld, Mobile and Wearable Technologies*, IGI Global, Hershey.

Moggridge, B. (2007) *Designing Interactions*, MIT Press, Cambridge, Massachusetts.

Nielsen, J. (1993) *Usability Engineering*, Academic Press, San Diego.

Nielsen, J. (1994) Heuristic evaluation in *Usability Inspection Methods* (eds J. Nielsen and R.L. Mack), John Wiley & Sons, New York.

Nielsen, J. and Budiu, R. (2012) *Mobile Usability*, New Riders, Berkeley, CA.

Norman, D.A. (1988) *Psychology of Everyday Things*, Basic Books, New York.

Reid, G.B. and Nygren, T.E. (1988) The subjective workload assessment technique: a scaling procedure for measuring mental workload in *Human Mental Workload* (eds P.A. Hancock and N. Meshkati), North Holland Press, Amsterdam, pp. 185–213.

Rogers, Y., Sharp, H., and Preece, J. (2011) *Interaction Design: Beyond Human-Computer Interaction*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Salvendy, G. (2012) *Handbook of Human Factors and Ergonomics*, 4th edn, John Wiley & Sons Inc., Hoboken, New Jersey.

Shneiderman, B. and Plaisant, C. (2009) *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 5th revised edn edn, Addison-Wesley Longman, Amsterdam.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Suchman, L.A. (1990) What is human-machine interaction? in *Cognition, Computing, and Cooperation* (eds S.P. Robertson, W. Zachary, J.B. Black), Ablex Publishing Corporation, Norwood, New Jersey, pp. 25–55.

Weinschenk, S. (2011) *100 Things every Designer Needs to Know About People*, New Riders, Berkeley, CA.

Wickens, C.D. and Hollands, J.G. (1999) *Engineering psychology and human performance*, 3rd edn, Prentice-Hall, New Jersey.

Zißler-Gürtler, D. (2002) Wenn Rotsehen Glück bedeutet *markenführung marketing/kommunikation*, **3**, 120–124.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto248
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Physics of Car Crashes: Design Concepts for Safer Cars

David C. Viano[1] and Priya Prasad[2]

[1] *ProBiomechanics LLC, Bloomfield Hills, MI, USA*
[2] *Prasad Engg LLC, Plymouth, MI, USA*

1 Introduction 1
2 Field Accident Data 1
3 Crash Dynamics 4
4 Intrusion 5
5 Injury Criteria and Tolerances 6
6 Compatibility 7
7 Restraint Systems 9
Related Articles 12
References 12

1 INTRODUCTION

The investigation of real world crashes in the 1950s spawned research and improvements in vehicle crashworthiness, occupant restraints, and friendly interiors (Schwimmer and Wolf, 1961). By the early 1980s, the concept of HARM was used to develop priorities for safety improvements (Malliaris, Hitchcock, Hedlund, 1982; Malliaris, Hitchcock, Hansen, 1985). HARM combined the incidence and severity of crash injuries into priority rankings of crash types and injury sources. Today, field accident data is available online to study the most significant sources of injury. It is also possible to investigate electronic files with individual crashes for in-depth understanding (http://www.nhtsa.gov/NASS).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

A vehicle in motion has kinetic energy. In a crash, energy is absorbed by deformation of vehicle structures, causing rapid decelerations that slow the vehicle. Primary measures of crash severity are the change in velocity (ΔV) and amount of intrusion of the occupant compartment. These measures describe the physics of a car crash and risks for occupant injury.

2 FIELD ACCIDENT DATA

Priorities for automotive safety are based on an understanding of field accident injuries. In-depth analysis of real-world crashes provides information on the sources for injury by crash type and occupant position in the vehicle. The field data describe the type of crash and deformation to vehicle structures that involve acceleration of the vehicle's cg (center of gravity) and intrusion of the occupant compartment. Field data also address the use of occupant restraints in the crash. This gives information on the effectiveness of seat belts, airbags, and other safety features.

Some of the most comprehensively investigated field accidents are collected by the NHTSA (National Highway Traffic Safety Administration). The National Automotive Sampling System (NASS) was designed to help NHTSA identify the most common safety problems in motor vehicle crashes, prioritize methods to reduce injury rates, and help enable regulations. NASS is made of NASS-GES and NASS-CDS. NASS-GES is based on police reports and NASS-CDS data provides comprehensive information on police reports, NASS investigator reports, and medical records.

NASS-CDS is a nationally representative sample that is publicly available. It contains detailed crash information

representative of a random sample of thousands of minor, serious, and fatal crashes. The data are collected by trained investigators. The crashes investigated in NASS-CDS are based on a probability sample of all US police-reported crashes. A NASS-CDS crash must be police reported and involve at least one towed vehicle.

Field research teams study about 5000 crashes a year, involving light vehicles (passenger cars, light trucks, vans, and utility vehicles). Approximately 400 variables are collected. Trained crash investigators obtain data from crash sites. They study evidence such as skid marks and broken glass. They locate the vehicles involved, photograph them, measure the crash damage, and identify occupant contact marks on the vehicle interior. The researchers then follow up with on-site investigations. They interview crash victims and review medical records to determine the nature and severity of injuries.

Because of its quality and nationally representative attributes, NASS-CDS data has been used by industry, universities, and other research organizations to assess the significance of various crash scenarios, identify ways to improve vehicle crashworthiness, to improve restraint system performance, and to identify needs and assess the safety benefit of various countermeasures. It provides important means of understanding injury mechanisms and is the basis for cost–benefit analyses used to support rule-making and report to Congress on the traffic accidents in the United States.

Figure 1 shows the risk for severe injury by crash severity (ΔV), crash type, and belt use based on 1994–2009 NASS-CDS. The risk for severe injury increases with crash severity and is different for unbelted compared to belted occupants. Risks are higher for unbelted occupants. For belted occupants, the highest risks are in side impacts followed by frontal crashes and then rear impacts at any particular crash delta V. There is an increase in risk for side impacts above 15–20 mph, for frontal crashes above 25–30 mph, and for rear impacts above 40–45 mph. For

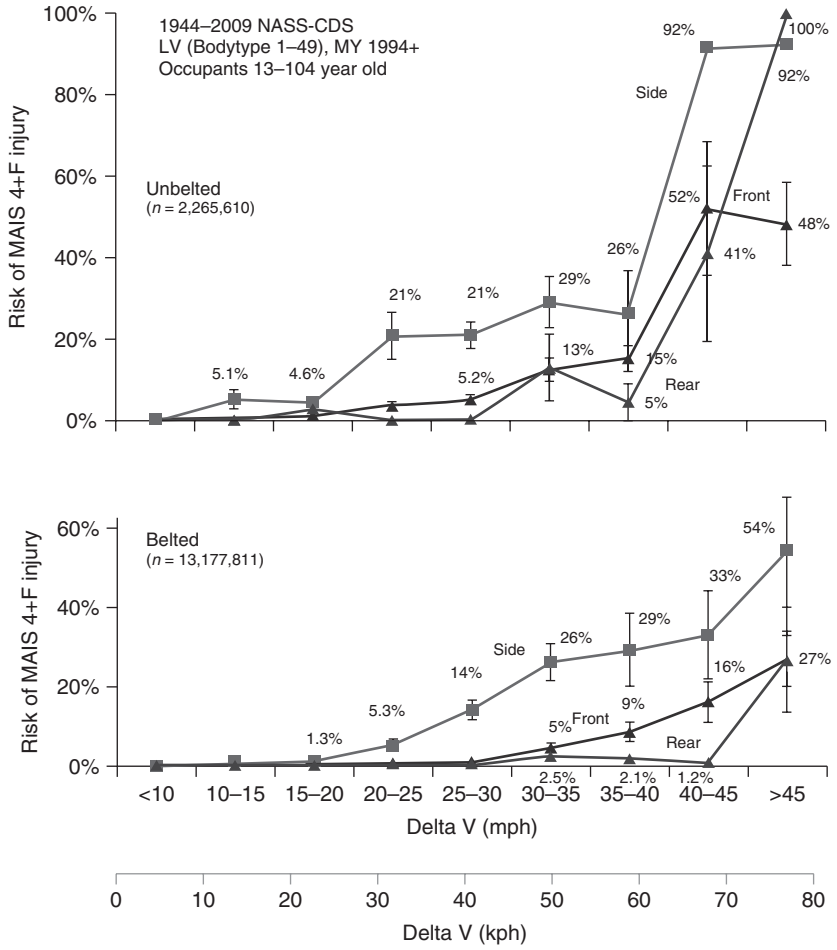


Figure 1. Risk for severe injury by delta V, crash type, and seat belt use.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

unbelted occupants, the increase in risk starts at lower severity crashes.

The trend in risk with delta V can be fit by a power function. For unbelted occupants in frontal crashes, the risk (R) in percent is $R = 0.000659\Delta V^{2.829}$ indicating risk increases by more than the square of crash ΔV, which is in miles per hour. For belted occupants in frontal crashes, $R = 0.00000874\Delta V^{3.728}$, indicating a stronger influence of crash severity at higher delta V but with lower risks than for unbelted occupants. For this sample of crashes, essentially all frontal crashes involve airbag deployment, so the risk is for the combination of belt use and airbags for crashes above 15 mph. For unbelted occupants in frontal crashes, the airbag is also part of the restraint.

For unbelted occupants in side impacts, the trend in risk with delta V shown in Figure 1 is given by $R = 0.000932\Delta V^{3.012}$. For belted occupants in side impacts, $R = 0.0000339\Delta V^{3.779}$. For unbelted occupants in rear impacts, $R = 0.0000636\Delta V^{3.288}$ and for belted occupants, $R = 0.000000165\Delta V^{4.744}$. The data show a strong correlation between injury risk and crash severity.

Table 1 shows the occupancy of vehicles of different types and the risk for severe injury (MAIS 4+F). Occupants of cars constitute 59.6% of occupants exposed to towaway crashes. In 71.4% of the crashes the driver is the only occupant. In 18.8% of the crashes, there is a driver and right-front passenger. Only 9.8% of the crashes involve three or more occupants in the vehicle. For vans, 18.1% of crashes involve three or more occupants. The lowest occupancy is in trucks with only 5.9% of crashes involving three or more occupants.

When severe injury is considered, the importance shifts to vehicles with multiple occupants. For vans, 46.3% of severely injured occupants are in vehicles with three or more occupants. The fraction drops to 12.1% in trucks. The highest risks are in sport utility vehicles (SUVs) and trucks with six or more occupants.

Table 2 shows the occupancy by crash type and the risk for severe injury (MAIS 4+F). In terms of exposure, 36.4% of occupants are in frontal impacts, 16.3% in side crashes, and only 5.8% in rollovers. In contrast, the largest fraction of severely injured occupants are in rollovers

Table 1. Occupancy, severe injury, and injury risk by vehicle type.

	Number of occupants							
Vehicle type	1	2	3	4	5	6	7+	All
Number of vehicles ($n = 2{,}406{,}964$)								
Cars (%)	71.4	18.8	6.3	2.7	0.6	0.1	0.0	59.6
SUV (%)	67.8	20.3	6.9	3.1	1.4	0.3	0.1	18.3
Vans (%)	60.4	21.5	8.1	4.5	2.5	1.7	1.2	7.4
Trucks (%)	75.4	18.7	3.9	1.4	0.5	0.0	0.1	14.7
All (%)	70.5	19.3	6.2	2.7	0.9	0.2	0.1	100.0
Number of exposed occupants ($n = 3{,}239{,}781$)								
Cars (%)	52.7	25.4	12.6	6.9	2.0	0.2	0.2	60.0
SUV (%)	50.5	25.7	10.7	7.1	4.5	1.0	0.4	18.3
Vans (%)	38.8	23.1	12.8	9.6	6.0	5.4	4.4	8.5
Trucks (%)	62.6	24.7	7.8	3.1	1.0	0.2	0.6	13.2
All (%)	52.4	25.2	11.6	6.7	2.6	0.8	0.7	100.0
Number of occupants with severe-to-fatal injury (MAIS 4+F, $n = 21{,}869$)								
Cars (%)	51.6	27.1	9.9	7.0	3.5	0.4	0.4	62.7
SUV (%)	46.7	25.3	12.4	6.7	4.2	3.2	1.5	15.9
Vans (%)	34.0	19.8	9.6	8.6	6.8	8.3	13.0	6.1
Trucks (%)	56.6	31.3	5.7	3.7	1.5	0.7	0.4	15.3
All (%)	50.5	27.0	9.6	6.5	3.5	1.4	1.4	100.0
Risk MAIS 4+F/vehicles								
Cars (%)	0.69	1.38	1.49	2.48	5.34	6.80	10.34	0.96
SUV (%)	0.54	0.98	1.42	1.71	2.32	7.37	10.97	0.79
Vans (%)	0.43	0.69	0.89	1.44	2.02	3.67	8.03	0.76
Trucks (%)	0.71	1.59	1.38	2.58	3.07	13.80	3.83	0.95
All (%)	0.65	1.27	1.41	2.20	3.59	5.45	8.36	0.91
Risk MAIS 4+F/MAIS 0+F								
Cars (%)	0.96	0.92	0.66	0.80	1.44	1.44	1.54	0.91
SUV (%)	0.94	0.75	0.93	0.69	0.64	2.61	2.23	0.87
Vans (%)	0.76	0.55	0.44	0.53	0.63	0.83	1.73	0.68
Trucks (%)	1.29	1.44	0.72	1.14	1.73	2.35	0.67	1.27
All (%)	0.99	0.92	0.69	0.77	1.04	1.34	1.64	0.92

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Occupancy, severe injury, and injury risk by crash type.

	Number of occupants							
Crash type	1	2	3	4	5	6	7+	All
Number of vehicles ($n = 2{,}406{,}964$)								
Rear (%)	6.6	1.6	0.7	0.3	0.1	0.0	0.0	9.4
Side (%)	11.2	3.5	0.9	0.4	0.1	0.0	0.0	16.3
Front (%)	26.2	6.9	2.0	0.9	0.3	0.1	0.0	36.4
Rollover (%)	4.0	1.1	0.4	0.2	0.1	0.0	0.0	5.8
All (%)	70.5	19.3	6.2	2.7	0.9	0.2	0.1	100.0
Number of exposed occupants ($n = 3{,}239{,}781$)								
Rear (%)	4.9	1.8	1.3	0.6	0.1	0.0	0.1	8.9
Side (%)	8.3	4.9	1.9	1.1	0.5	0.2	0.1	17.1
Front (%)	19.4	9.8	4.2	2.5	1.1	0.3	0.2	37.4
Rollover (%)	2.9	1.6	0.8	0.7	0.2	0.1	0.1	6.5
All (%)	52.4	25.2	11.6	6.7	2.6	0.8	0.7	100.0
Number of occupants with severe-to-fatal injury (MAIS 4+F, $n = 21{,}869$)								
Rear (%)	0.7	0.7	0.2	0.0	0.1	0.1	0.0	1.9
Side (%)	14.2	7.8	3.4	1.7	0.8	0.1	0.5	28.6
Front (%)	15.0	6.4	2.2	1.5	0.4	0.1	0.0	25.7
Rollover (%)	15.0	7.6	2.7	1.9	1.5	0.7	0.7	30.1
All (%)	50.5	27.0	9.6	6.5	3.5	1.4	1.4	100.0
Risk MAIS 4+F/veh								
Rear (%)	0.10	0.39	0.27	0.06	1.42	4.42	0.49	0.19
Side (%)	1.15	2.01	3.31	3.97	4.87	2.43	15.93	1.59
Front (%)	0.52	0.85	0.97	1.57	1.22	1.41	0.88	0.64
Rollover (%)	3.44	6.31	6.54	7.49	24.22	21.46	36.67	4.74
All (%)	0.65	1.27	1.41	2.20	3.59	5.45	8.36	0.91
Risk MAIS 4+F/MAIS 0+F								
Rear (%)	0.26	0.43	0.14	0.02	1.28	2.96	0.08	0.27
Side (%)	1.52	1.23	1.38	1.17	1.09	0.48	3.26	1.37
Front (%)	0.65	0.51	0.40	0.46	0.30	0.26	0.18	0.55
Rollover (%)	3.91	3.42	2.41	2.02	5.70	3.91	5.22	3.47
All (%)	0.99	0.92	0.69	0.77	1.04	1.34	1.64	0.92

(30.1%, followed by side impacts (28.6%) and frontal crashes (25.7%). The highest risk for severe injury is in rollovers at 3.47%, followed by side impacts at 1.37%. The lowest risk is in rear impacts at 0.27%.

Table 3 shows exposure and risk for injury by ejection status. For nonejected occupants, the risk for injury AIS 3–6 is greatest in rollovers at 9.38% followed by side impacts at 6.59%. The risk is shown for each body region by crash type. The most common AIS 3–6 injuries are to the head (2.82%) and chest (2.52%) for nonejected occupants. With complete ejection, the risks increase to 63.3% for the head and 40.1% for the chest in rollovers. When nonejected occupants are compared to completely ejected occupants, the highest relative risk is 479 times for AIS 3–6 neck injuries in frontal impacts (Viano and Parenteau, 2010).

Table 4 summarizes the effectiveness of seat belts in preventing injury to the head, chest, spine, and lower extremities (LX) by crash type. The data show high effectiveness in preventing AIS 4+ injuries to the head and torso and AIS 3+ injuries to the LX.

3 CRASH DYNAMICS

For frontal impact, the vehicle has a traveling speed when it collides with another vehicle or roadside object. The collision deforms vehicle structures and decelerates the occupant compartment. The deceleration of the occupant compartment depends on the amount of crush and is often approximated by a Haversine function. Integration of the acceleration pulse gives the delta V of the impact and double integration gives the stopping distance of the vehicle's cg.

Figure 2 shows the peak acceleration and distance traveled for a Haversine of 80, 120, and 160 ms duration crashes of different severity in terms of delta V. For example, a 35 mph frontal NCAP crash test of 120 ms duration involves 30.4g peak acceleration and 1.07 m (42.2″) stopping distance for a 40 mph delta V. The overall severity of the crash is greater than the impact speed because of restitution that involves rebound of the vehicle away from the barrier. The vehicle acceleration increases as the duration of the collision decreases and the stopping distance decreases.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Body region injury risk by ejection status and crash type.

	Occupants	AIS 3–6 injuries by body region								
	MAIS 0–6 F	Head (%)	Face (%)	Neck (%)	Thorax (%)	Abdomen (%)	Spine (%)	UX (%)	LX (%)	All (%)
Nonejected										
All crashes	28,679,211	1.15	0.084	0.0046	1.09	0.21	0.26	0.47	0.90	4.20
Front	12,423,550	0.71	0.082	0.0052	0.89	0.20	0.16	0.56	1.18	3.81
Side	5,607,800	2.16	0.101	0.0042	2.05	0.33	0.30	0.46	1.17	6.59
Rear	1,824,887	0.66	0.008	0.0032	0.18	0.015	0.103	0.018	0.100	1.11
Rollover	2,164,305	2.82	0.216	0.0122	2.52	0.44	1.24	1.00	1.00	9.38
Partial ejection										
All crashes	108,744	45.6	1.5	0.1259	22.9	4.2	8.3	13.6	9.4	106
Front	9,057	66.2	3.0	0.0000	33.5	5.6	4.8	18.9	18.3	151
Side	22,493	99.6	2.9	0.4210	42.9	7.6	4.5	23.7	17.1	199
Rear	6,563	1.9	0.00	0.0000	1.6	0.42	0.87	0.00	1.13	5.9
Rollover	57,830	32.3	1.2	0.0729	17.4	3.8	8.8	11.5	6.2	81.3
Complete ejection										
All crashes	157,558	58.2	2.1	0.4694	39.9	9.1	11.1	7.3	19.4	148
Front	18,242	52.3	0.79	2.5090	55.0	7.9	10.5	20.2	46.5	196
Side	28,912	47.6	1.6	0.1244	31.2	5.3	6.1	3.0	13.0	108
Rear	4,170	8.8	0.00	0.0000	14.1	2.9	2.9	2.5	7.2	38.5
Rollover	86,645	63.3	1.6	0.2272	40.1	7.6	14.3	6.6	16.5	151

Table 4. Seat belt effectiveness by crash type and body region.

Crash	AIS 4+ injury			AIS 3+
Type	Head (%)	Thorax (%)	Spine (%)	LX (%)
Front	83	89	84	82
Side	88	79	90	77
Rear	95	86	43	89
Rollover	86	83	65	74
All	78	76	69	70

Figure 2 also summarizes the severity of a number of consumer and Government crash tests to the front, side, and rear of vehicles. The impacts involve rigid moving or fixed barriers, deformable moving or fixed barriers, and are full-width or offset.

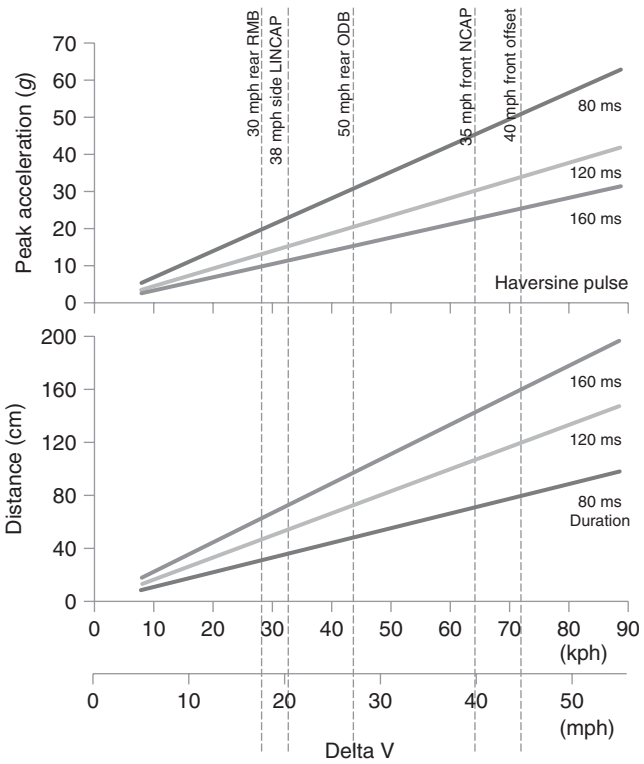


Figure 2. Peak acceleration and distance for Haversine crash pulses of 80, 120, and 160 ms duration with nominal values for Government and consumer crash tests.

4 INTRUSION

As the severity of a crash increases in offset impacts, the risk for deformation of the occupant compartment increases. Intrusion or deformation of the occupant compartment increases the risk for injury, particularly to the occupant adjacent to the intrusion. This occurs because of the velocity of the intrusion.

Evans (1991) showed that deformation of the occupant compartment or intrusion adjacent to the seated position of an occupant increased fatality risks by up to four times that of occupants seated at a distance from the intrusion. Viano and Parenteau (2008) found the same effects on the risks for severe injury. When the occupant compartment was deformed in front, side, or rear impacts, the severity of the collision increased and there were higher risks of injury.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

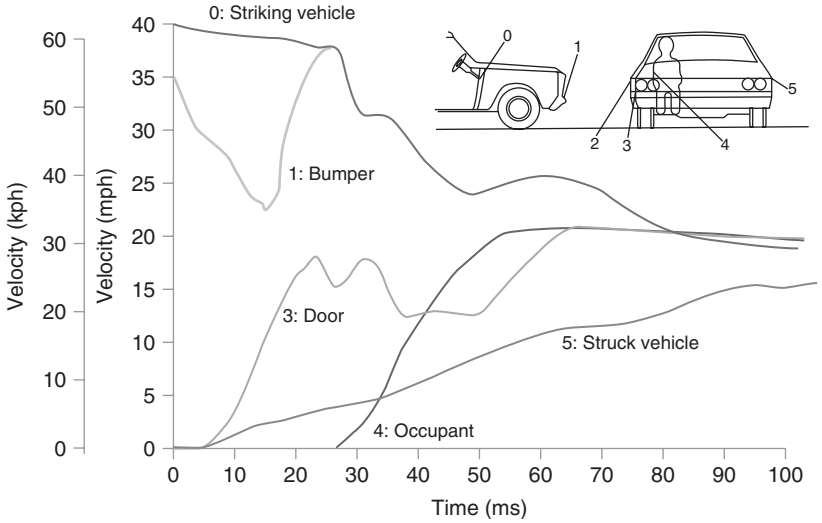


Figure 3. Side impact showing the velocity of the striking vehicle, struck vehicle, door intrusion, and occupant. (Based on Strother *et al*. 1984. Reprinted with permission from the Association for the Advancement of Automotive Medicine.)

The relationship between injury and intrusion has been studied in side impact crashes. Strother *et al*. (1984) described the injury mechanism for a near-side occupant in a side impact. Figure 3 shows the velocity in a car-to-car side impact. The striking vehicle is traveling at 40 mph when its bumper impacts the side of another vehicle. Since the bumper is in the crush zone of the striking vehicle, its velocity drops as it deforms the door of the struck vehicle. At the same time, the door of the struck vehicle rapidly accelerates up to more than 15 mph in 20 ms in this example. The occupant is adjacent to the door and is struck at high velocity. The occupant is accelerated in the direction of the door intrusion. Later, the struck vehicle accelerates through the delta V of the crash.

The higher injury risks in side impacts shown in Figure 1 are related to intrusion of the side interior and high velocity loading on the occupant in the struck vehicle. There have been a number of improvements in side impact protection since this example was presented. Vehicle side structures have been reinforced to reduce the intrusion velocity and amount. Padding and inflatable restraints have been added to door and windows providing protection for the head, neck, and torso (Lau, Capp, Obermeyer, 1991).

5 INJURY CRITERIA AND TOLERANCES

Injury criteria are mechanical parameters such as force, acceleration, or deflection that are related to injury of a body region (Eiband, 1959; and Melvin, 1992). For example, the acceleration of the head has been used to describe the risk for skull fracture and brain injury (Mertz, Prasad, Nusholtz, 1996). Head injury is related to the duration of the impact and acceleration raised to the power 2.5. This led to HIC (Head injury Criterion), which is specified in government and consumer testing with dummies. Today, a family of dummies is used in crash testing. They represent adults, children, and infants, and instrumentation is used in the dummies to measure mechanical responses of the head, neck, chest, abdomen, pelvis, and extremities in the collision. The responses are related to injury risks based on biomechanical analysis and consensed injury criteria and tolerance levels.

Table 5 summarizes some information from a comprehensive collection of human tolerances developed by Mertz, Irwin, and Prasad (2003). It covers injury criteria for various body regions and tolerances for the family of dummies. There are three adult hybrid III dummies, three hybrid III child dummies, and three infant dummies. The tolerances include the HIC_{15} and peak head acceleration, upper neck force and bending tolerances, as well as the *Nij* intercepts, chest injury tolerances for acceleration, compression, and viscous response, and the femur compression tolerance. Other tolerance criteria and values included in Mertz, Irwin, and Prasad (2003) are not summarized here. In addition, there is a new family of dummies for testing in side and rear impacts with specific response measurements and injury criteria (www.humaneticsatd.com, www.carhs.com). For completeness, Table 5 also includes some basic anthropometry information on the nine crash test dummies. This includes standing and seated height, weight, head mass, and other dimensions of the dummy.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 5. Anthropometry and injury reference values by occupant size/age (based on Mertz, Irwin, and Prasad 2003)).

	Large male	Mid-size male	Small female	10 yo child	6 yo child	3 yo child	18 mo infant	12 mo infant	6 mo infant
Anthropometry									
Body mass (lb)	225.5	172.0	102.8	71.3	45.9	31.9	24.6	21.3	17.2
Head mass (lb)	10.9	10.0	8.1	8.1	7.7	6.7	6.0	5.5	4.6
Standing height (in)	73.4	68.9	59.6	54.1	46.0	37.5	32.0	29.4	26.4
Sitting height (in)	38.2	35.7	32.0	28.3	25.0	21.5	19.9	18.9	17.3
Forehead height (in)					22.8	19.3		16.9	15.4
Occipital condyle height (in)		29.7			20.0	16.7	15.4	14.6	13.3
Vertex to occipital condyles (in)		6.0	5.4	5.3	5.0	4.8	4.5	4.3	4.0
Injury tolerances (IARV)									
Head									
HIC (36 ms)		1000							
HIC (15 ms)	670	700	779	741	723	568	440	389	377
Peak acceleration (g)	175	180	193	189	189	175	160	154	156
Neck									
Tension (lb)	1130	937	589	515	425	321	243	222	209
Compression (lb)	1085	899	566	494	409	310	234	216	200
Flexion (in lb)	2230	1682	841	690	531	372	257	239	221
Extension (in lb)	1133	850	434	354	266	186	133	124	115
Nij	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Intercept values									
Tension (lb)	1838	1524	957	834	692	524	396	362	339
Compression (lb)	1681	1393	876	762	634	479	362	330	310
Flexion (in lb)	3584	2699	1354	1,106	850	593	414	376	350
Extension (in lb)	1566	1177	592	485	372	259	181	165	153
Chest									
Acceleration (g)	54	60	73	82	93	92	89	87	88
Deflection (in)	2.2	2.0	1.6	1.4	1.2	1.1	1.0	0.9	0.9
Viscous (mph)	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2
Peak velocity (mph)	18.3	18.3	18.3	18.8	19.0	17.9	17.7	17.0	17.4
Femur									
Load (lb)	2584	2038	1384	1,007	607	339	265	200	157

6 COMPATIBILITY

Compatibility between two vehicles generally refers to the response of vehicles and its occupants in two-vehicle collisions. Although the subject has been of interest since the beginning of crashworthiness studies, no consensus has been reached as to the definition of compatibility. Some studies have tried to define compatibility in terms of an aggressivity metric, which essentially had a good correlation with vehicle mass (Gabler and Hollowell, 1998, 2000), and others have used experimental data (Digges and Eigen, 2001; Barbat, Li, Prasad, 2001a, 2001b). However, it is commonly agreed that in two-vehicle collisions, the vehicle responses in terms of delta V, crush, and intrusion are strongly affected by the mass ratio, relative stiffness, and the relative geometry of the two vehicles. The above three factors also affect the injury responses of occupants in collisions.

For example, in a head-on collision between two vehicles of unequal masses, all else being the same, the likelihood of fatality of occupants in the lighter vehicle will be higher than that in the heavier vehicle. Similarly, even if the two vehicles are equal in mass, the likelihood of fatality in the softer vehicle will be higher than that in the stiffer vehicle. The geometrical alignment of the crash interface between the two vehicles can result in under-ride/override between the two vehicles. In general, the injury outcome of occupants in the vehicle being overridden is greater than that of occupants in the vehicle that is under-riding, except for large truck under-ride. The vehicle and occupant responses are a direct consequence of the physics of the collision.

If we assume that the masses of the two vehicles are m_1 and m_2 and they are in a head-on collision while moving at velocities v_1 and v_2 (closing velocity $= V_c = v_1 + v_2$), conservation of momentum dictates that $m_1{*}v_1 - m_2{*}v_2 = (m_1 + m_2){*}v_c$, where $v_c =$ common velocity and $m_1 > m_2$.

As a result of conservation of momentum, the delta V of the two masses will be inversely proportional to their masses, that is, (delta v_1 of m_1)/(delta v_2 of m_2) $= m_2/m_1$.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For example, if m_1 is twice m_2, the mass m_1 will undergo a velocity change one-half that of the mass m_2. Since the velocity changes of the two masses take place in the same time duration of impact, the average acceleration of the lighter mass will be two times that experienced by the heavier mass. The occupants in the lighter vehicle will have to contend with two times the vehicle accelerations than the occupants in the heavier vehicles. The injury responses of the occupants will be proportional to the vehicle accelerations, but not in a linear manner. Eiband's (1959) curves indicate that it should be in the order of $(m_2/m_1)^{2.5}$. Accident data also indicate that this relationship is nonlinear and is represented by a power function of delta V as indicated in the section on accident data. In other words, if R_{21} is the ratio of severe-to-fatal injury risk of occupants in vehicle 2 versus that in vehicle 1, and the occupants are restrained, $R_{21} = (m_1/m_2)^{3.728}$, where, the exponent 3.728 is derived by a power fit of field-observed severe-to-fatal injury risk as a function of delta V for belted occupants in frontal crashes as mentioned in an earlier section of the chapter.

If the mass ratio is 2, an occupant in the lighter vehicle is nearly 13.25 times more likely to be severely injured than an occupant in the heavier vehicle. On the basis of acceleration effects only, it should have been $2^{2.5} = 5.7$ approximately. The difference in observed and predicted risk of occupants in the lighter car arises from other factors like relative stiffness and geometry of the front ends of the colliding vehicles. Evans (2004) analyzed FARS data for two vehicle head-on crashes in which at least one occupant was killed and derived the relative risk R_{21} as a function of mass ratio (m_1/m_2) as $R_{21} = (m_1/m_2)^{3.58}$ where R_{21} is the ratio of fatal risk in vehicle 2 colliding against vehicle 1. The absolute risk of fatality as a function of delta V could not be determined from the available data because delta V is not available in FARS. Once again, if the mass ratio is 2, the likelihood of an occupant being killed in the lighter vehicle is nearly 12 times that of an occupant in the heavier vehicle. This predicted risk is not too different from 13.25, and perhaps within the errors of estimation, derived for risk of severe-to-fatal injuries based on momentum principles applied to field data discussed earlier in this chapter.

In two-vehicle collisions, energy principles dictate that the kinetic energy absorbed in the collision $= (1/2)*((m_1*m_2)/(m_1 + m_2))*V_c{}^2$. The energy is absorbed by deformation of structures of the two vehicles. The amount of crush experienced by the individual vehicles depends on the relative stiffness of the interacting structures and determines the stopping distance of the collision. For example, if we assume that the two vehicles are equal in mass and front-end stiffness, both vehicles will absorb half the energy and will have equal amount of crush. However, if one vehicle is substantially stiffer than the other, the softer vehicle will have to absorb the major portion of the energy and will crush substantially more than the stiffer vehicle. An extreme manifestation of this phenomenon is in front-to-side collisions in which the side structure is substantially softer than the front end of a vehicle, resulting in substantially higher intrusions of the side structure of the struck vehicle than in the front end of the striking vehicle.

When vehicles of substantially different masses are involved in frontal collisions, there is disparity in mass as well as stiffness. This results in generally greater crush and structural intrusions in the lighter vehicle. This phenomenon is generally observed in crashes between LTVs (light truck and van) and cars. The LTV fleet in the United States is heavier than the car fleet and the LTV is generally stiffer than cars because of their functionalities. Figure 4 shows the acceleration of a midsize car and an SUV of nearly the same mass. In a collision between the two vehicles, the car will have more crush than the SUV. For a more "compatible" crash, changing the front-end stiffness of the car or the SUV may be in order so that both vehicles participate in energy absorption and reduce the crush of the car front end. This could be accomplished by increasing the front-end length of the SUV. However, it may not be practical as the increased length of the SUV will affect some of its functions such as off-road capability. Also, changing the front-end stiffness of the SUV or the car would not be effective if the structural interactions between the SUV and the car front ends were not adequate, for example, if the bumpers and the energy-absorbing structures in the two vehicles did not match in height, resulting in the car being overridden by the SUV. Car bumper heights above the ground are standardized as they have to be in the Part 581 zone specified in a Federal Motor Vehicle Safety Standard. However, the bumper heights of LTVs are not regulated. These observations have led to the development of two principles for "compatible" crashes: "stiffness matching and geometry matching." Obviously, stiffness matching cannot be accomplished if the geometries are not matched.

Geometry matching between the LTV and the car fleet was the subject of research by a working group composed of vehicle manufacturers in the United States. In December 2003, the WG recommended that to improve "compatibility" between LTVs and cars, the light trucks primary frontal energy absorbing structure shall overlap at least 50% of the Part 581 Zone. If the above geometrical requirement could not be met, there must be a secondary energy-absorbing structure (SEAS), connected to the primary energy-absorbing structure, whose lower edge shall be no higher than bottom edge of the Part 581 zone. This

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

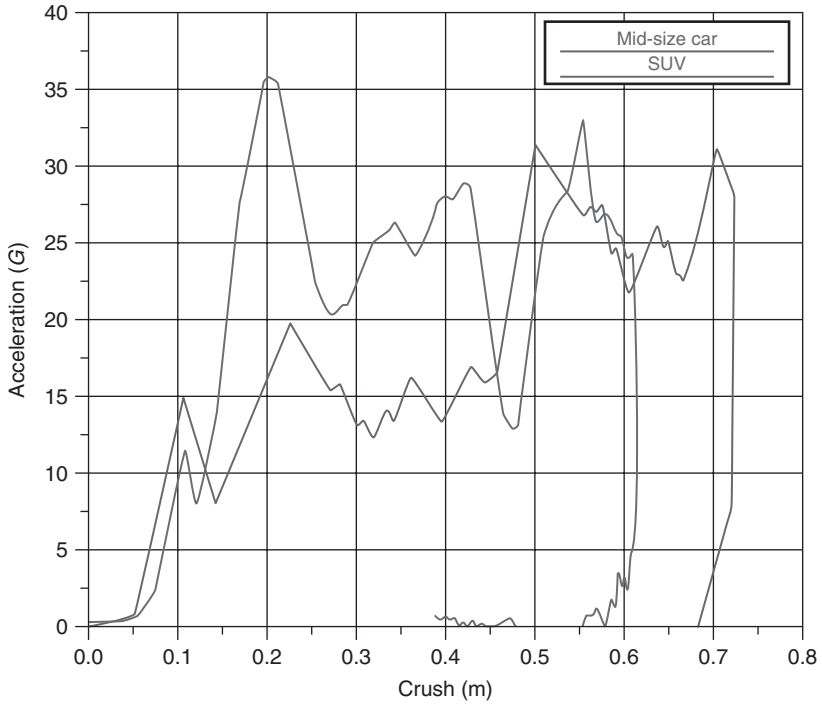


Figure 4. Crash pulses of two vehicles at 56 kph against a rigid barrier (NCAP).

secondary structure shall be designed to reduce the structural override of a passenger car during a frontal crash. The SEAS is commonly known as the *BlockerBeam*™ was introduced by the Ford Motor Company between 1999 and 2001 in heavy SUV and super duty pickup trucks. The details of the research performed by the WG were reported by Barbat (2005).

The recommendation of the WG was accepted by the domestic and foreign automotive manufacturers in the United States for all vehicles (<10,000 lb GVWR) manufactured and sold in the United States by MY 2009. The industry also agreed to fit all vehicles <10,000 lb GVWR with side airbags capable of providing head protection when impacted by an LTV, noting that practical structural changes in LTVs to improve compatibility in front-to-side crashes were not possible (see Insurance Institute of Highway Safety (IIHS) side impact test procedure, www.iihs.org).

7 RESTRAINT SYSTEMS

This section discusses some of the basic principles of restraint systems currently used in vehicles to provide occupant protection in front, side, and rear impacts and rollovers (Table 4). In case of a collision, two impacts occur. The first impact is when the vehicle collides against a fixed object or a moving object. Just before impact, the occupant and the vehicle are traveling at the same velocity. Subsequent to the initiation of impact of the vehicle with an external object, the vehicle structures deform causing the vehicle to slow down. This creates a differential velocity between the occupant and the vehicle, causing the occupant to move toward the point of impact. Without any restraint system, for example, seat belts, the occupant will contact the interior of the vehicle at a velocity dictated by the initial velocity of the occupant and the velocity of the contacted part of the vehicle. The occupant decelerates until they come to rest relative to the vehicle. The occupant–vehicle contact phase is generally referred to as the "*second collision*". In order to minimize injurious loadings on the occupant, energy-absorbing foams or energy-absorbing deforming structures are designed into the vehicle. The goal of the traditionally recognized restraint systems, for example, seat belts and airbags, is to minimize or avoid, if practicable, injurious contacts with the vehicle interior by minimizing occupant excursion within the occupant compartment during impact and retaining the occupant within the vehicle.

Since the early 1970s, three-point seat belts have become common in all vehicles as the primary restraint system. Since the mid-1980s, inflatable restraints have been introduced in various phases in all vehicles in the United States and more recently in vehicles around the world. The first application of inflatable restraints was for providing

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

incremental protection to belted front outboard occupants in frontal crashes, hence the term *supplemental restraint system* (SRS) for airbags. The early development of airbags can be found in Viano (1987), which is a compendium of published safety research through 1986 and more recently in Kent (2003). Initially, the frontal airbags were introduced in the United States, by law, with the goal of providing protection to unbelted occupants in frontal impacts. The test procedure to demonstrate such a level of protection was specified in the FMVSS 208. The test procedure involved a crash of the vehicle against a rigid barrier at 30 mph with unbelted driver and passenger dummies. The levels of injury criteria required by the law could be met. However, field data indicated that airbags designed to meet the required injury criteria could seriously injure occupants who were in the path of the deploying airbag, typically in low-speed crashes and with small occupants. FMVSS208 crash test velocity was reduced to 25 mph for the unbelted dummy tests to reduce the severity of airbag-induced injuries. This law became fully effective as of MY 2006. In an interim period 1998–2006, the unbelted crash could be substituted with a sled test driven by a generic half-sine pulse. Some background information of the development and changes of the FMVSS 208 are described by O'Neill (2009). The effectiveness of the sled test-certified airbags has been reported by Kahane (2006).

Inflatable restraints have undergone major developments since their first introduction for frontal protection. Even knee airbags have been introduced in fairly high volumes. More recently, airbags deploying from the shoulder belt have been introduced in the rear seats of several vehicles manufactured by Ford Motor Co. (Sundararajan *et al*., 2011). Inflatable restraints to provide head protection in side impacts were introduced in the mid-1990s. This was in the shape of a tube originally stowed in the side rails of the vehicle. Approximately, in the same time frame, inflatable restraints to provide chest protection in side impacts were also introduced. These side airbags were mounted on the door. The head airbag was integrated with the chest airbag as one system and is commonly referred to as the "*combo side airbag*". This combo bag was mounted on the outboard side of the front seats. The "combo bag" has evolved further to provide chest and pelvis protection, and the head protection is now provided by an inflatable curtain. The initial application of the inflatable curtain was for providing head protection in side impacts, an event lasting milliseconds. The curtain technology has further evolved to provide protection in rollovers, an event lasting for seconds, by the use of coated bags that retain pressure for 5–6 s. Future curtains will have to comply with the requirements of FMVSS 226, which specifies a series of tests and acceptance criteria, including containment in the vehicle to avoid ejection.

Over the years, substantial improvements in the belt system has taken place, some would not be possible without standard frontal airbags. Today's belts have pretensioners that apply forces on the occupant early in the crash to create an early coupling of the occupant with the vehicle, increasing ride-down of structural crush. Most shoulder belts in the front outboard seating positions have load limiting. The level of load limiting varies from vehicle to vehicle, but has proven to be effective in reducing chest injuries (Foret-Bruno *et al*., 2001; Mertz and Dalmotas, 2007). The level of shoulder belt load limiting used in today's vehicles could not be used without the airbag as lower belt loads lead to greater occupant excursion within the vehicle and partial and full ejection in some circumstances. Since the seat belt is the earliest form of restraint system, its installation in vehicles in United States is also regulated. For example, the anchor locations are controlled by the FMVSS 210, and comfort and convenience specifications have been developed by NHTSA. The seat belt retractors have to meet the requirements of the FMVSS 209, which specifies the deceleration levels ($0.7g$) at which the retractor will lock and the maximum amount of belt payout from the retractor (most retractors use a "no lock" at $<0.3g$ and "must lock" at $>0.45g$ within 25 mm of webbing payout). Most retractors also include webbing sensitive locking at $2g$ acceleration within 50 mm.

Early designs of the seat belt were influenced substantially by studying accident data, identifying injuries that could have been caused by the belt system, and injuries caused in spite of the belt system. During the early development of the seat belt, lack of biofidelic crash test dummies, led to kinematic design criteria to judge the quality of the restraint system. On the basis of analysis of accidents in Germany, Adomeit and Heger (1975) proposed seat belt anchor locations and seat cushion designs that would reduce the likelihood of submarining under the belt and reduce the likelihood of unfavorable kinematics of the occupant. Although the guidelines developed by Adomeit and Heger (1975) are still valid and generally followed by designers, the quality of restraint systems is now judged in crash tests with sophisticated dummies.

Crash tests to evaluate restraint systems: Regulatory and new car assessment programs (NCAPs) are in place in the United States, Europe, Japan, Australia, Korea, China, and Latin America, and more are expected in future. A common thread in all the NCAPs is a frontal crash against a crushable barrier, engaging 40% of the front end of the vehicle at 56 to 64 mph initial velocity. In Europe, this test is also a regulatory test; however, the velocity is 56 kph as opposed to 64 kph in the NCAP. In the United States, FMVSS 208

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

has required a test against a rigid barrier engaging the full front end of the vehicle. The impact velocity used to be 30 mph, but has been raised to 35 mph since MY 2006. The United States also has an NCAP frontal test at 35 mph against a full-width rigid barrier and the rating is based on dummy occupant responses. However, the vehicles sold in the United States are also designed for a test similar to that conducted in European NCAP.

These tests are conducted by the IIHS and are generally referred to as the *offset deformable barrier* (ODB) test. The rating system utilizes dummy occupant measures and structural responses of the vehicle. The structural responses considered in the rating system include post-crash intrusions of the toeboard and the dash. As a result, the overall stiffness of the front end is generally dictated by this test and has become a *de facto* standard for structural strength of the front end of vehicles. Since an ODB is used in the test, in general, the occupant compartment accelerations are lower and the crash pulse is of longer duration than those observed in the rigid barrier tests. The common understanding among safety engineers is that the ODB test is a test of the vehicle structure and the rigid barrier test, generating higher vehicle accelerations, is a test of the restraint system. We now discuss the physics of vehicle and occupant responses in rigid barrier impacts with a simple spring mass model.

Figure 5 shows a vehicle occupant represented by a point mass M_D. The mass is connected to the vehicle by a linear spring representing the stiffness of a restraint system that develops a force when a certain amount of prescribed slack is taken out when the occupant moves to the left. The vehicle and the occupant are initially traveling at the same velocity. When the vehicle impacts an external object it begins to slow down depending on the stiffness of the vehicle, causing the occupant to move to the left. After the slack is taken out, restraining forces are developed on the occupant. The restraint force in this case is the highest when the occupant stops moving relative to the vehicle. The maximum force can be easily estimated by solving the equation of motion of the mass M_D. For simplicity, we will assume that the mass is 1 kg, the spring stiffness is 2.5 kN/m and the initial velocity of the vehicle and the occupant is 56 kph. Note that the natural frequency of the spring–mass system is approximately 8 Hz and the time period of one oscillation is approximately 125 ms. In addition, the time period of one-half oscillation is 62.5 ms. The natural frequency selected in this example is an approximation of modern airbags plus belt restraint systems. The solution of the differential equation governing the motion of the occupant can be obtained in closed form, for illustrative purposes. We show the results of several different conditions of vehicle structural characteristics in terms of peak deceleration of the occupant mass, M_D, expressed as multiples of the acceleration due to gravity (g).

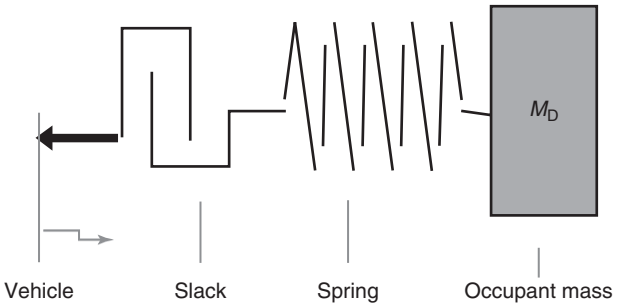


Figure 5. Spring–mass model of occupant in vehicle crash with slack.

Several scenarios are studied with and without the slack in the restraint system (Table 6). In the first scenario, we assume that the vehicle stops suddenly, in which case the occupant will move to the left, load up the restraint system and stop. In this case, the peak deceleration of the occupant will be at the instant when the energy absorbed by the restraint system is exactly equal to the initial kinetic energy of the occupant. Simple calculations show that the peak deceleration of the occupant is 79.3g. In the other scenarios, we will stop the vehicle with a prescribed deceleration–time history that is Haversine in shape. The peak acceleration of the pulse and the duration were adjusted to yield approximately 17.8 m/s delta V of the vehicle. The 17.8 m/s corresponds to an initial velocity of 15.56 m/s plus a rebound velocity of 2.2 m/s (commonly observed rebound velocity of vehicles tested against rigid barriers at 15.56 m/s). As the vehicle deceleration decreases, the stopping distance increases for the assumed impact. The results show the importance of increasing stopping distance in reducing the forces on the occupant. It is also shown that even with 30 mm of slack in the restraint system, forces on the occupant increase when compared to those without slack. Hence the deployment of pretensioners in modern vehicles that minimize slack in the belt system.

The preceding analysis is a simplified analysis of the dynamics of an occupant involved in frontal crash and how the front-end design and restraint parameters can affect the forces on and decelerations of the occupant. Since the start of crash tests against rigid barriers, the topic of the optimum shape of the vehicle deceleration time history (the "crash pulse") has been studied. The earliest study was reported by Egli (1967) and a more recent one has been reported by Kral (2006). Kral's (2006) study followed those by Takahashi *et al*. (1993) and two studies by Motozawa and Kamei (2000), Motozawa *et al*. (2003). All studies

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 6. Effect of belt slack and stopping distance on occupant acceleration.

Stopping||distance (mm)	Haversine||acceleration (g)	*Duration* (*ms*)	Occupant acceleration (*g*) Slack in belt (mm)	
			0	30
0	Infinite	0	79.3	79.3
430	60.0	60	77.6	82.0
517	45.3	80	69.4	75.3
713	36.3	100	60.0	66.8
855	30.2	120	51.4	58.5

indicate that the optimum pulse should have an early rise followed by a dwell period that is followed by a square wave pulse. This is generally known as a *front-loaded pulse*. Kral (2006) goes on to mention that a front-end structure capable of delivering such a front-loaded crash pulse may not be ideal in lower velocity collisions. Designing a front-end structure to deliver the optimum pulse has proven to be elusive as most vehicles in the field do not exhibit the desired characteristics of the theoretically optimum pulse. One aspect not considered in the above studies is the aspect of compatibility with other vehicles. A fast rising, front-loaded pulse would be aggressive to other vehicles, especially in front-to-side crashes.

RELATED ARTICLES

Vehicle Seat Design, Development and Manufacturing
Adaptive Restraint Systems: Towards Integral Safety
Body Design, Overview, Targeting a Good Balance Between all Vehicle Functionalities

REFERENCES

Adomeit, D., Heger, A. (1975) Motion Sequence Criteria and Design Proposals for Restraint Devices in Order to Avoid Unfavorable Biomechanic Conditions and Submarining. *19th Stapp Car Crash Conference*, SAE 751146, SAE Warrendale, PA.

Barbat, S., Li, X., Prasad, P. (2001a) A Comparative Analysis of Vehicle-to-Vehicle and Vehicle-to-Rigid Fixed Barrier Frontal Impacts. *17th ESV Conference, Paper No. 01-S7-O-01*, Amsterdam, Netherlands.

Barbat, S., Li, X., Prasad, P. (2001b) Evaluation of Vehicle Compatibility in Various Frontal Impact Configurations. *17th ESV Conference, Paper No. 01-S7-O-09*, Amsterdam, Netherlands.

Barbat, S.D. (2005) Status of Enhanced Front-to-Front Vehicle Compatibility Technical Working Group Research and Commitments. *ESV Conference, Paper No. 05–463*, US DOT, Washington, D.C.

Digges, K., Eigen, A. (2001) Measurement of Stiffness and Geometric Compatibility in Front to Side Crashes. *ESV Conference, Paper No. 349*, US DOT, Amsterdam, Holland.

Egli, A. (1967) Stopping the occupant of a crashing vehicle - a fundamental study. SAE 670038, SAE, Warrendale, PA.

Eiband, A.M. (1959) Human tolerance to rapidly applied accelerations: a summary of the literature. NASA Memorandum 5-19-59E, National Aeronautics and Space Administration, Washington.

Evans, L. (1991) *Traffic Safety and the Driver*, Van Norstand Reinhold, New York.

Evans, L. (2004) *Traffic Safety*, Science Serving Society, Bloomfield Hills, MI.

Foret-Bruno, J.Y., Trossille, X., Le Coz, J.Y., Bendjellal, F., Steyer, C., Phalempin, T., Villeforceix, D., Dandres, P., Got, C. (2001) Thoracic Injury Risk in Frontal Car Crashes with Occupant Restrained with Belt Load Limiter. SAE 2001-22-0009, *42nd Stapp Car Crash Conference*, SAE, Warrendale, PA.

Gabler, H.C., Hollowell, W.T. (1998) The aggressivity of light trucks and vans in traffic crashes. SAE 980908, SAE, Warrendale PA.

Gabler, H.C. and Hollowell, W.T. (2000) The Crash Compatibility of Cars and Light Trucks. *Journal of Crash Prevention and Injury Control*, **2** (1), 19–32.

Kahane, C.J. (2006) An Evaluation of the 1998–1999 redesign of frontal airbags. NHTSA Technical Report DOT HS 810085, Washington, D.C.

Kent, R. (2003) Air bag development and performance: new perspectives from industry, government and academia. SAE PT-88, Society of Automotive Engineers, Warrendale, PA.

Kral, J. (2006) Yet another look at crash pulse analysis. SAE 2006-01-0958, SAE, Warrendale, PA.

Lau, I.V., Capp, J.P., Obermeyer, J.A. (1991) A comparison of frontal and side impact: crash dynamics, countermeasures and subsystem tests. SAE 912896, Society of Automotive Engineers, Warrendale, PA.

Malliaris, A., Hitchcock, R., Hansen, M. (1985) Harm causation and ranking in car crashes. SAE 850090, Society of Automotive Engineers, Warrendale, PA.

Malliaris, A., Hitchcock, R., Hedlund, J. (1982) A search for priorities in crash protection. SAE 820242, Society of Automotive Engineers, Warrendale, PA.

Mertz, H.J., Prasad, P., Nusholtz, G. (1996) Head injury risk assessment for forehead impacts. SAE 960099, Society of Automotive Engineers, Warrendale PA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Mertz, H.J., Dalmotas, D.J. (2007) Effects of Shoulder Belt Limit Forces on Adult Thoracic Protection in Frontal Collisions. SAE 2007-22-0015, *51st Stapp Car Crash Conference*, SAE, Warrendale, PA.

Mertz, H.J., Irwin, A.L., and Prasad, P. (2003) Biomechanical and scaling bases for frontal and side impact injury assessment reference values, SAE 2003-22-0009, *Stapp Car Crash Journal*, **47**, 155–188.

Motozawa, Y., Kamei, T. (2000) A new concept for occupant deceleration control in a crash. SAE 2000-01-0881, SAE, Warrendale PA.

Motozawa, Y., Tsuruta, M., Kawamura, Y., Noguchi, J. (2003) A new concept for occupant deceleration control in a crash-part 2. SAE 2003-01-1228, SAE, Warrendale PA.

Nahum, A.M. and Melvin, J.W. (1992) *Accidental Injury: Biomechanics and Prevention*, 2nd edn, Springer-Verlag, New York. ISBN: 0-387-97881-3-540-97881

O'Neill, B. (2009) Preventing Passenger vehicle occupant injuries by vehicle design - a historical perspective from IIHS. *Traffic Injury Prevention*, **10** (2), 113–26.

Schwimmer, S., Wolf, R.A. (1961) Preliminary Ranking of Injury Causes in Automobile Accidents. *5th Stapp Car Crash Conference SAE 1961-12-0001*, Society of Automotive Engineers, Warrendale, PA.

Strother, C.E., Smith, G.C., James, M.B., and Warner, C.Y. (1984) Injury and intrusion in side impacts and rollovers. SAE 840403, Society of Automotive Engineers, Warrendale PA.

Sundararajan, S., Rouhana, S.W., Board, D., *et al*. (2011) Biomechanical assessment of a rear-seat inflatable seatbelt in frontal impacts. *Stapp Car Crash Journal*, **55**, SAE, Warrendale, PA

Takahashi, K., Suzuki, N., Sonoda, Y., Komamura, T., Suzuki, T., Tawarayama, T., Dokko, Y. (1993) Optimization of vehicle deceleration curves for occupant injury. SAE 9307515, SAE, Warrendale PA.

Viano, D.C. and Parenteau, C.S. (2010) Ejection and severe injury risks by crash type and belt use with focus on rear impacts. *Traffic Injury Prevention*, **11** (1), 79–86.

Viano, D.C., Parenteau, C.S. (2008) Fatalities by seating position and principal direction of force for 1st, 2nd and 3rd row occupants. SAE 2008-01-1850, Society of Automotive Engineers, Warrendale PA.

Viano, D.C. (1987) Passenger car inflatable restraint systems: a compendium of published safety research, SAE PT-31, SAE, Warrendale, PA.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto252
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Reparability and Insurance Ratings in the Development of Cars

Frank Leimbach and Helge Kiebach

KTI GmbH & Co. KG, Lohfelden, Germany

1 Introduction 1
2 Historical Background 1
3 Insurance Scheme 3
4 Terminology 4
5 Important Test Procedure 4
6 Measures for the Conformance to Low Speed Crash Tests 8
7 Summary 11
References 12

1 INTRODUCTION

Decades ago, the analysis of real world accidents determined that most accidents happen at low speeds. German statistics (Anselm 1997) revealed that 85% of all frontal impacts occur at velocities below 9 mph (15 km/h), resulting in the primary crash safety requirement for the front structure of cars to fully absorb the energy of an offset impact at 9 mph (15 km/h).

From an insurance cost view, the requirements are to limit the deformation or damage to structural parts and cost-intensive components.

The crash safety performance could be implemented relatively easily, that is, in the design of rigid bumpers and beams. However, the design of car front structures is a very complex process, including marketing and visual design criteria, as well as pedestrian safety. Along with the insurance cost aspects, other regulatory and consumerism low speed crash tests are often employed to reduce damage and improve the repairability of vehicles.

In addition to frontal testing, low speed tests for the rear and side areas of the vehicle have also been introduced for similar purposes.

This chapter addresses the history of low speed repair crash tests and the relevance of crash repair tests in the calculation of insurance premiums for fully comprehensive cover. Various standards and crash tests for consumer information and insurance tests have been introduced in many markets. RCAR (Research Council for Automobile Repairs) provides an overview of the variety of low speed test procedures.

2 HISTORICAL BACKGROUND

In the United States, the Insurance Institute for Highway Safety (IIHS) started to conduct low speed crash tests at 5 mph into a flat barrier in 1969. These tests led to the first federal bumper test rules for cars, requiring the bumpers to resist damage at impacts up to 5 mph (8 km/h). In April 1971, the National Highway Traffic Safety Administration (NHTSA) issued its first regulation on passenger car bumpers.

Federal Motor Vehicle Safety Standard (FMVSS) 215 was initially enacted in September 1972, imposing requirements to prohibit damage to specified safety-related components such as headlamps and fuel systems in a series of perpendicular barrier impacts at 5 mph (8 km/h) for front bumpers and 2.5 mph (4 km/h) for rear bumper systems.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.

This article is © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

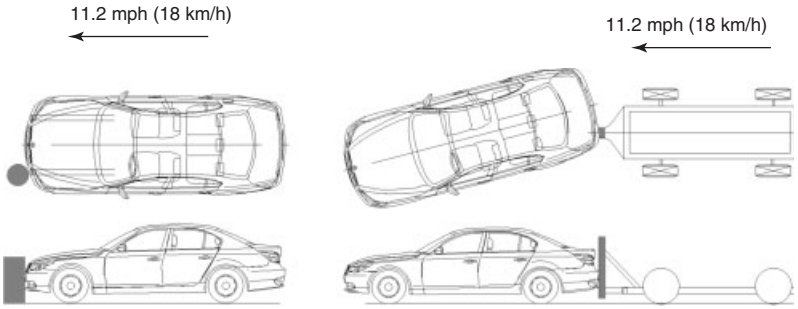


Figure 1. Test setup of the first crash repair tests in Germany.

These standards were further enhanced for MY 1974 passenger cars, specifying standardized height front and rear bumpers that could resist angled impacts at 5 mph (8 km/h) with no damage to the car's lights, safety equipment, or the engine.

Performance requirements have been improved several times. For MY 1979 cars, the standard required no damage to safety-related parts and exterior surfaces not involving the bumper system. This was known as *Phase 1* and required an impact test speed of 5 mph. More stringent tests were introduced for MY 1980–1982, applying 5 mph longitudinal front and rear impacts and 3 mph pendulum (corner) impacts resulting in no damage to the bumper itself beyond a 3/8 inch "dent" and 3/4 inch "set," or displacement from original position. These requirements, limiting damage to the bumper itself, are referred to as *Phase II*.

The last change in the bumper standard took place on 14 May 1982 when the tests were modified, reducing the test impact speeds from 5 to 2.5 mph for longitudinal impacts and from 3 to 1.5 mph for pendulum impacts. The Phase II bumper damage requirement was dropped and replaced with the previous Phase I requirements. After considering the available data and public comments, the NHTSA completed a Final Regulatory Impact Analysis (FRIA) in support of the latest rule amending the bumper standard to the 2.5 mph impact with Phase I damage requirements. These standards became effective on 6 July 1982, affecting 1983 and subsequent model year cars as American standard (SAE J980a) and have been adopted by the European Community, enacted as part of ECE Regulation No. 42 in June 1980 (UN, 1980). This regulation applies to the behavior of certain parts of the front and rear structures of passenger cars when involved in a collision at low speed and says: "Exterior protection is assured by protective devices, which are essentially elements located at the front and rear ends of vehicles and designed in such a way as to allow contacts and small shocks to occur without causing any serious damage" (United Nations, 1980).

During the late 1970s, Germany introduced a car insurance classification system (Gustafsson, 2001). These frontal-impact tests required the left chassis leg/side member to crash centrally against a pole of 40 cm in diameter (Figure 1). For the rear test, a 20 cm wide, 500 kg ram was driven into the left rear corner of the car at an angle of 15°. The impact speed applied in both tests was 11 mph (18 km/h). The damage caused by these tests was calculated and the repair procedure was documented.

In 1982, the frontal and rear test procedures were revised, along with the addition of side crash tests and the average repair cost per vehicle type was determined.

These procedures involved front and rear offset crash tests with a 40% width overlap (excluding the exterior mirrors) against a rigid barrier (0° angle) at a collision speed of 9.3 mph (15 km/h [+1/−0 km/h]). The rigid, mobile barrier weighed 1000 kg and was 70 cm high with a ground clearance of 20 cm. The mobile barrier moved parallel to the test car's longitudinal axis. For the side crash test, a rigid barrier weighing 1000 kg was rammed into the left side front door of the test car at a collision speed of 6.2 mph (10 km/h) at 45°.

When determining the result, in order to determine the final average cost per vehicle, each test was weighted based on the actual incidences of each type of impact (frontal 54%, rear 30%, and side crash test 16%).

The German system and its application to insurance rating had a great influence on the car manufacturers' interest in lowering crash repair costs. They realized that they were partly responsible for setting the insurance premium level of the cars they were producing and selling, affecting sales, so focusing on the reduction of crash repair costs would help in their aims.

This test method has been further developed in cooperation with other RCAR members and is documented in an RCAR document. The method was later incorporated in an RCAR Standard (low speed 9.3 mph offset insurance crash test). The test was included in the type class calculation of the German Insurance Association in 1991.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In 2006, the low speed offset crash test was revised. The angle of front and rear impacts was changed from 0° to 10°, and the rear impact moving barrier weight was increased from 1000 to 1400 kg. This test is referred to as the *RCAR Structural Test*. Speed and overlap have been not modified.

Car manufacturers design their vehicles to perform well in the RCAR structural test, but some vehicles do not exhibit a good crash behavior in "real world" crashes. In some cases, manufacturers have eliminated the bumper beam (a strong part) as a backstop and replaced it with localized measures such as crush cans in order to perform well in the tests. Such sub-optimized designs are in most cases not robust or are damaged too easily, often leading to expensive damage in car-to-car crashes.

Insurance claims data also indicates that rear bumpers are often underridden in low speed impacts by a striking vehicle because of bumper system movement or vertical "dive" of vehicles during braking, resulting in misalignment of the "strong" areas and excessive damage. In these cases, it is desirable to have bumper systems that have sufficient vertical overlap to ensure proper engagement. To this end, bumpers should ideally be mounted at slightly different heights at front and rear, but should also have sufficient height to maintain engagement over a wide range of circumstances. Insurance data also shows that rear bumpers are often overridden when struck by high ride-height vehicles (SUVs, pickup trucks, and so on). Vehicle damage would be reduced in both these situations with taller front and rear bumper beams.

With this in mind, an international RCAR working group has developed test procedures to assess how well a vehicle's bumper system protects the vehicle from damage in low speed impacts. The damage in these tests closely replicates the damage patterns observed in real world low speed crashes.

This test represents a typical city accident, with interaction of the bumpers of the involved vehicles. The test vehicle crashes with a collision speed of 6.2 mph (10.0 ± 0.5 km/h) against a deformable, energy-absorbing element with a solid rear backing, which replicates real world damage where an underride situation occurs. The test encourages designs that absorb crash energy while limiting intrusion into the vehicle structure. Vehicles with qualifying bumper beam heights of 100 mm or more shall be subject to the dynamic test even if the front-to-rear bumper engagement is less than 75 mm if a reasonable test result can be anticipated. The vehicle's bumper system must sufficiently engage with the bumper barrier to be deemed acceptable. Systems that use the vehicle main structure as a backstop for energy management will be regarded as unacceptable.

3 INSURANCE SCHEME

In countries where it is legally permitted, insurance rating systems have been introduced. The cost of repairs as a result of low speed crash test criteria (mostly, the RCAR standard) plays a major role in deciding the insurance group rating and, therefore, the whole life insurance costs of the vehicle.

In Germany, motor vehicle insurance premiums (motor liability, semi-comprehensive cover, and fully comprehensive covers) are calculated by separating the individual car types into categories. Until 1996, the main criterion for the classification was the engine power. As of 1996, the classification depends on two criteria; the frequency of claims as a percentage of total number of claims and the average loss across the total vehicle stock. Fully comprehensive cover consists of 25 categories.

While the frequency of claims depends primarily on driver behavior, the average loss can be affected by the vehicle's construction. This system regulates itself because the frequency of claims and the average loss of any car type must be a substandard of all cars in the market. At the first classification of a new car type, the frequency of claims is as far as possible assumed from the comparable predecessor model. The claim amount is derived directly from the RCAR test, which is then projected on the total of average losses. Since 2010, the RCAR bumper test has served as test procedure by the German Insurance Association in addition to the RCAR low speed test.

The car insurance scheme in the United Kingdom is like the German model. Since 1992, cars have been condensed to car groups and then classified by car type. Testing is based on the RCAR test. After a crash, the damage is determined and the car is repaired. The original price of the car, the engine power, the availability of body in white as a spare part, and the protection against theft (as standard equipment) are further influences on the classification. The frequency of claims is not considered. France, Spain, Italy, Sweden, Norway, and Finland also use tests that are in line with the RCAR test.

In the United States, there is no uniform insurance scheme. In the majority of the States, motor liability insurance is obligatory. In some States, simply proof that there is enough capital available for adjustment of a loss is a sufficient precondition to get the vehicle registered. The calculation of the insurance premium depends on driver experience, car type, place of residence, family status, age, sex, and capital assets of the owners.

In the United States, the repair costs are determined in tests that are carried out by the IIHS. The Institute's series of four tests (front and rear full-width impacts at 6 mph and front and rear corner impacts at 3 mph) produce the types and amounts of damage that commonly occur in low speed

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

collisions. The results of this test are made public, allowing the consumer to choose a car with low repair costs.

The Insurance Australia Group (IAG) is the operator of the Technical Research Centre in Australia and a member of RCAR. The IAG performs research into repair costs and gives recommendations regarding classification of cars. Furthermore, the repair costs determined from tests are published on the Internet. As in the United States, the Australian consumer can use this information in choosing a car with low repair costs.

4 TERMINOLOGY

The terms *damageability* and *repairability*, as used in the following sections, are according to the guidelines of RCAR (2008), which says:

"Damageability is the capacity of a vehicle to withstand the force of a collision and embraces the ability of the vehicle to absorb crash energy and in so doing limit the physical displacement, deformation and damage to structures and high cost components."

"Repairability means the possibility and ease of repair, firstly in the physical sense and secondly in terms of cost. Good repairability will mean that the vehicle can be restored to its pre-accident condition either by repair or by economic component replacement."

5 IMPORTANT TEST PROCEDURE

5.1 RCAR tests

5.1.1 RCAR low speed offset crash test (until 2006)

The RCAR low speed offset crash test was established in 1983 in Germany and was developed at the suggestion of the insurance industry. The RCAR test specification describes a front crash and a rear crash. The damage, as a consequence of a side crash, which is needed in the German first classification of cars, is considered as a theoretical damage within the RCAR procedure. For this, a partial replacement of the sill and a replacement of driver's door will be calculated.

In the front and rear impact tests, the vehicle overlaps the face of the barrier by 40% ±25 mm (Figure 2) at 0° (parallel) to the front/rear of the vehicle. The 1000 kg mobile barrier is propelled into the test vehicle with an impact speed of 9.3 mph (15 km/h [+1/−0 km/h]).

An alternative rear impact test method is acceptable, whereby the test vehicle is propelled rearward into a fixed barrier that is dimensionally compliant with the fixed dimensions of the mobile barrier. The speed of the vehicle on impact must be adjusted according to the mass of the test vehicle and is determined by Equation 1.

$$v = 15\,\text{km/h}\sqrt{\frac{1000\,\text{kg}}{1000\,\text{kg} + \text{m}}} + 1\,\text{km/h} \qquad (1)$$

where v = test speed in kilometer per hour and m = mass of test vehicle.

5.1.2 New RCAR low speed offset crash test (since 2006)

In 2006 (RCAR, 2006), the low speed offset crash test was revised. The impact angle was changed from 0° to 10° and the rear impact moving barrier weight was increased from 1000 to 1400 kg (Figure 3). This test is referred to as the *RCAR Structural Test*. Speed and overlap were not modified.

5.1.3 RCAR bumper test

The RCAR bumper test uses a standardized bumper beam that is 100 mm tall with a flexible, energy-absorbing cover that replicates a real bumper on a car. A 200 mm tall backplate is fitted to the top surface of the bumper barrier, 25 mm behind the front face, which represents the upper

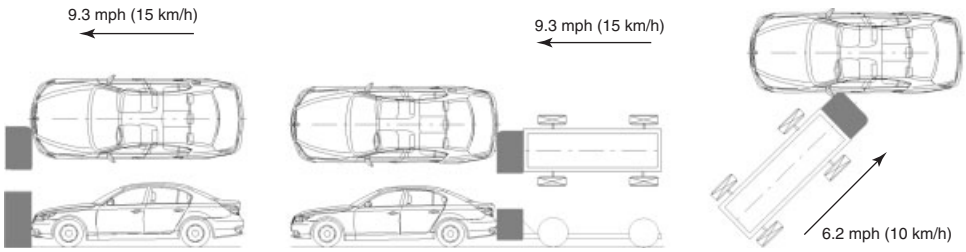


Figure 2. Test setup of RCAR low speed offset crash test (until 2006).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

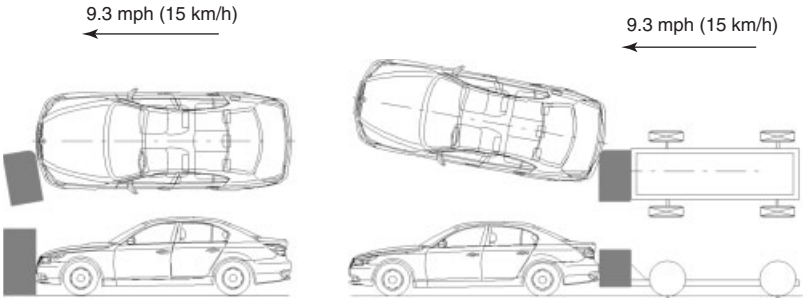


Figure 3. Test setup of new RCAR low speed offset crash test (since 2006).

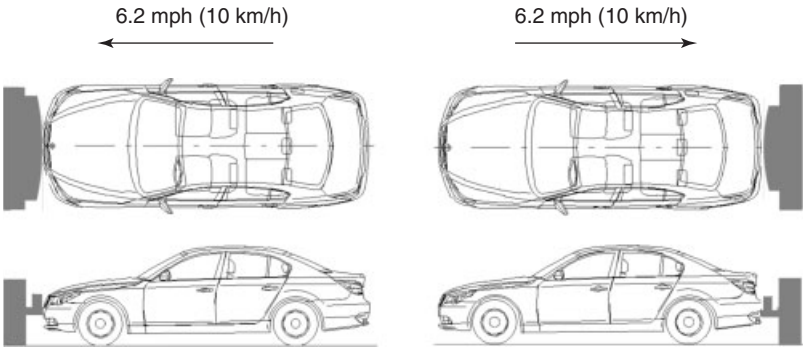


Figure 4. Test setup of RCAR bumper test (RCAR 2010). (Reproduced from RCAR, 2010. © Research Council for Automobile Repairs.)

structure of the impacting vehicle (see Figure 4, RCAR, 2010). The tests are run at 6 mph (10 km/h) with the bumper barrier height set to 455 mm from the ground to the lower edge of the barrier for the front test and 405 mm for the rear test. These dimensions recreate the underride phenomenon seen in real world crashes and encourage the car manufacturers to produce tall bumper beams that are at least 100 mm tall, and that overlap between front and rear. The RCAR test only recommends testing beams that are within the 75 mm overlap test zone.

5.2 IIHS tests

The IIHS's low speed crash test protocol was revised in 1997, 2001, and 2002 (IIHS, 1996, 1997, 2001a, b, 2002, 2012).

5.2.1 IIHS low speed test 1996 until 2007

This older crash test protocol (Version 1) described four different low speed crash tests (IIHS, 1996) at a nominal 5 mph (8 km/h) impact speed. The following tests were performed on each vehicle model:

- front into full-width flat barrier;
- rear into full-width flat barrier;
- right front into 30° angle barrier;
- rear center into pole.

The first two tests followed the patterns of the perpendicular barrier impacts required by federal regulations for passenger cars (49 CFR, Part 581). The latter two tests were added to simulate a broader range of impacts occurring in actual on-the-road crashes.

For the front- and rear-into-flat-barrier tests, the impact barrier is an unyielding (rigid) block of reinforced concrete weighing 145,150 kg positioned perpendicular to both the crash hall floor and the longitudinal centerline of the test vehicles (Figure 5). The barrier is augmented with a solid steel face plate measuring 366 cm wide, 184 cm high, and 8 cm thick. The impact area of the face plate is covered with 2 cm-thick plywood.

For the front-into-angle-barrier test, a rigid steel fixture is bolted to the impact barrier face plate (Figure 6). The fixture includes a solid steel face plate measuring 214 cm wide, 92 cm high, and 4.5 cm thick. The entire face plate is covered with 2 cm-thick plywood. The angle barrier face plate is perpendicular to the floor, and the angle between the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

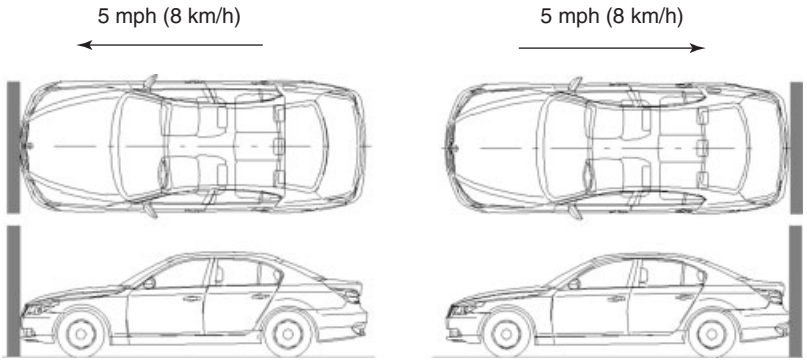


Figure 5. Test setup of IIHS low speed test front and rear into full-width flat barrier (1996 until 2007).

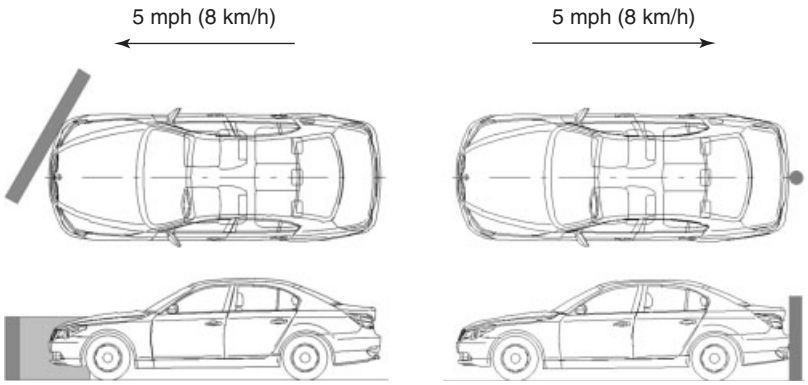


Figure 6. Test setup of IIHS low speed test front into inclined barrier and rear into pole (1996 until 2007).

longitudinal centerline of the test vehicles and the plane of the face plate is 60° (90° minus 30°). If any vehicles have hoods or front fenders whose top leading edges are more than 92 cm above the floor, the entire angle barrier fixture is raised, so that the bottom edge of the angle barrier face plate is 18 cm above the floor. Otherwise, the fixture is not raised.

For the rear-into-pole test, the test pole is 152 cm long and 18 cm in diameter (Figure 6). The pole extends 92 cm above the floor surface and 60 cm below.

For the flat barrier impacts, this target point is located midway between the vertical edges of the impact barrier face plate. For the angle barrier impacts, the target point is offset 53 cm to the left of the vertical centerline of the angle barrier face plate. For pole impacts, the target point is the pole surface located in the tangent plane perpendicular to the track centerline. The crash test speed range is 4.95 ± 0.15 mph (7.96 ± 0.24 km/h).

Following the completion of the tests, the damage will be estimated. The appraisers indicate "no damage" if there is damage only to the external bumper surfaces and no other damage to the vehicle body or bumper or localized dents are no more than 0.95 cm deep, and overall bumper distortion or displacement is no more than 1.9 cm from the original contour and there is no breakage of fasteners. These criteria were part of the federal requirements in effect from 1 September 1979 to 6July 1982 for passenger car bumpers, under which minor cosmetic damage to exterior bumper surfaces was permitted after specified 5 mph (8 km/h) barrier and pendulum impacts and 3 mph (5 km/h) corner pendulum impacts (49 CFR, Part 581). For part replacement indicated in the estimates, new original equipment replacement parts at full list prices are specified. If vehicles have clear coat (two-stage clear over color) paint, all estimates requiring refinishing include the appropriate additional labor time.

5.2.2 IIHS low speed test (since 2007)

The baseline for these tests is the crash test procedure according to RCAR. Four different bumper crash tests into a contoured, bumper-like barrier are conducted on each vehicle model. Front and rear full-overlap tests are

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

conducted at 6 mph (10 km/h, see Figure 4), and front and rear corner tests (Figure 6) are conducted at 3 mph (5 km/h). The four tests were developed to promote compatible, stable, and energy-absorbing interfaces among vehicles in the fleet. Each vehicle is run into a steel barrier designed to mimic the design of a car bumper, with the barrier's plastic absorber and flexible cover simulating typical cars' energy absorbers and plastic bumper covers. The barrier heights for full-overlap and corner tests differ to simulate a broader range of impacts occurring in actual on-the-road crashes.

For front and rear corner tests, the impact barrier is mounted with the forward-most portion of the bottom edge of the barrier that is 406 mm from the floor. At impact, the vehicle overlaps the lateral edge of the barrier by 15% of the vehicle's width, as measured at the wheel wells (including moldings and sheet metal protrusions) at the corresponding axle—front axle for front corner tests and rear axle for rear corner tests (Figure 7).

The damage-estimating process is conducted as it would be done in a repair shop; each bumper assembly is generally removed and dismantled to check for possible hidden damage. Damage repair estimates are conducted using industry standard appraisal techniques and documented in a computerized system developed by Audatex, a Solera company.

For hourly labor rates indicated in the estimates, Audatex supplies an average of labor rates for body repair and refinishing used in actual estimates by its clients across the country as of the most recent calendar year quarter. This average rate is rounded to the nearest dollar and used in calculating labor costs. Similarly, the cost for paint and related materials are based on the average rate used by Audatex clients during the most recent quarter (rounded to the nearest dollar) and is directly proportional to the total refinishing time for each estimate.

For part replacement indicated in the estimates, new original equipment replacement parts at full list prices are specified (based primarily on the most recent Audatex information, but secondarily on data from the appropriate Mitchell Collision Estimating Guide, Motor Crash Estimating Guide, or vehicle manufacturers or dealers). No discounts, betterments, appearance allowances, insurance deductibles, taxes, or vehicle storage fees are applied. If a vehicle has clear coat (two-stage clear over color) paint, all estimates requiring refinishing include the appropriate additional labor time (in most cases, automatically computed by the Audatex system, otherwise manually calculated by the appraisers).

The front and rear full test damages are multiplied by two—because in the real world, full-width impacts occur approximately twice as often as corner impacts; that total is added to the front and rear corner test damages. The sum is then divided by 6 to get the weighted average. This number determines the overall rating. The good/acceptable boundary is $500, the acceptable/marginal boundary is $1000, and the marginal/poor boundary is $1500. However, no vehicle can earn a rating of good or acceptable if it is deemed undrivable or unsafe after a test because of headlamp or taillamp damage, hood buckling, coolant loss, or the like. The IIHS's bumper test protocol was revised in 2009 (IIHS, 2007, 2009, 2010, 2012).

5.3 Pedestrian protection tests

There is always a conflict between the requirements of the front of a car for the pedestrian protection test and the requirements for the car front repair crash tests. The original EC Directive on pedestrian protection, 2003/102/EC (published on 6 December 2003), introduced pedestrian protection requirements in two stages. Both stages utilized the same test procedures, but the injury limits for Stage 2

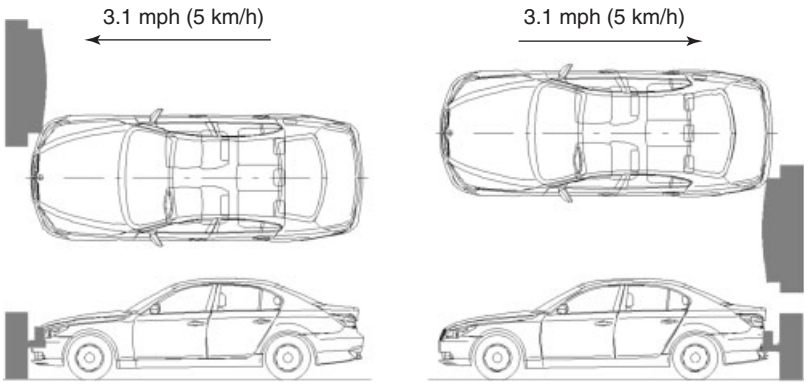


Figure 7. Test setup of IIHS low speed corner test (since 2007; according to the test of RCAR).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

were more stringent than those applied in Stage 1 (valid from 1 October 2005).

Additional energy-absorbing material is often used to enhance crash protection and insurance ratings. The thickness, width, height, and stiffness of the energy-absorbing material must now be optimized to meet the pedestrian legform impactor acceleration criteria. Therefore, the distance between the bumper face and the cross member will be limited. A method of managing the energy resulting from an impact to the bumper, as used by some manufacturers, involves a combination of deforming energy-absorbing element and the crush cans attached to the side member. This arrangement forms a two-stage energy management system.

The pedestrian safety tests are performed at 25 mph (40 km/h) of the 13.4 kg legform impactor in "free flight." In the process, the energy-absorbing element (e.g., a foam pad) behind the bumper has to absorb an energy of 830 J. This is the result of Equations 2 and 3, respectively

$$E = \frac{m}{2} v^2 \tag{2}$$

$$E = \frac{13.4\,\text{kg}}{2}\ (11.1\,\text{m/s})^2 = 830\,\text{J} \tag{3}$$

where v = test speed in meter per second and m = mass of test device.

At an impact of the vehicle against a rigid barrier, such as the RCAR test at 9.3 mph (15 km/h), initially, the barrier penetrates deep into the vehicle's front structure. Assuming a car test weight (m) of 1500 kg and an impact speed of 9.3 mph (15 km/h), the bumper structure must absorb a maximum energy of around 13,000 J (Figure 8).

Owing to the required space necessary for the fulfillment of the pedestrian protection tests using the lower legform, there is a little space for an effective energy absorption in a vehicle impact against an obstacle.

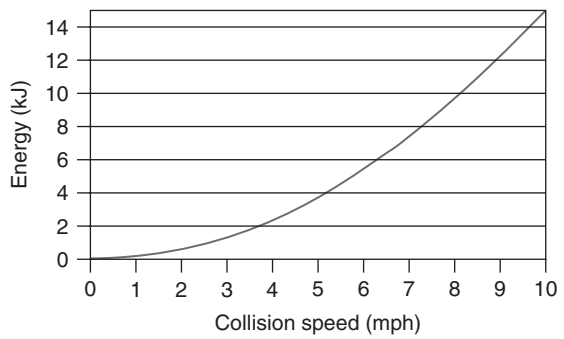


Figure 8. Increase of the collision energy with rising speed.

This means that impact forces are transmitted to the main vehicle structure more quickly over shorter distances in low speed crashes. This conflicts with the requirement to damage as few parts as possible during the crash repair test. Headlights, fenders, and engine cover cannot always be set back for design reasons, which also increase the risk of damage. Therefore, the requirements on the bumper systems increase to afford pedestrian protection (soft deformation at the front) as well as to absorb the energy of a low speed crash in shorter distances, without large amounts of total deformation.

6 MEASURES FOR THE CONFORMANCE TO LOW SPEED CRASH TESTS

The "RCAR Design Guide—A manufacturers' guide to ensure good design practice for repairability and limitation of damage" can serve as a tool to assist for the automobile manufacturer in optimizing damageability and repairability. Implementation of the guidance in this document can ensure a competitive vehicle with strong economic and practical selling points; the insurer and the policy holder pay less, and the vehicle repair industry is able to carry out the repairs needed easily.

It is important to ensure, however, that both damageability and repairability enhancements must be achieved without compromising the safety of vehicle occupants or other road users.

The most important test procedures are tests at speeds up to 9.3 mph (15 km/h), as performed by the insurance industry (e.g., testing of RCAR and the American Institute IIHS). In addition, the bumper systems must also conform to performance testing in relation to the overall vehicle concept. The primary standards are UN/ECE-R42, CMVSS 215, and FMVSS 215, which provide the ability to drive the car after a crash at low speed, that is, between 2.5 and 5 mph (4 and 8 km/h). Other mandatory and legally prescribed test standards include those on pedestrian safety (2003/102/EC) and occupant protection in crashes at speeds of up to 40 mph (64 km/h; e.g., UN/ECE-R94, FMVSS 208, and CMVSS 208).

An optimization of the bumper systems to meet these requirements is sometimes so different that the systems used can vary depending on the market (e.g., USA, Europe, and Japan). The design in terms of crash behavior is often complicated by the test specifications being subjected to changes in procedures, as new influences come into play.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6.1 General crash behavior

Figure 9 shows the stages of damage to the vehicle structure at different collision speeds. To minimize the repair costs, the components should, as far as possible, be consecutively involved from the outside inward, as the energy is absorbed.

In a frontal crash up to 2.5 mph (4 km/h), no damage to the main structure should be evident. The energy should be absorbed only by the external bumper skin and absorbent components inside the bumper, that is, foam elements. In accordance with regulation ECE-R42, little damage should occur; for example, marking on the outer bumper skin and punctures are acceptable.

At speeds between 2.5 mph (4 km/h) and 9.3 mph (15 km/h), the repair costs should be at an acceptable level, that is, cross members and elements such as impact absorbers or crushable crash boxes (Figure 10). Only the extreme front sections of side member should be damaged.

At speeds above 9.3 mph (15 km/h), the longitudinal beams will probably be damaged, in addition to the deformation of the wheel arches and other load-bearing parts in the front structure. In frontal impacts with even higher speeds, the deformation often extends to the bulkhead, regularly leading to an economic total loss.

To ensure appropriate crash impact absorption in front of the passenger compartment, the structure should be designed from the outset to accommodate high crash loads. This design should consider the increasing requirements of the crash test procedures, such as the EuroNCAP, 40% overlap at 40 mph (64 km/h). This means that, in accordance with differing crash repair tests, parts will already be appropriately stiff by design, combining to absorb energy in the early phase of a high speed crash as well as individually in low speed impacts.

During the RCAR test, it is most important that no airbag or belt tensioner is triggered; otherwise, the repair costs will rise sharply. The threshold for triggering the front airbags must, therefore, be above the delay profile caused by the RCAR test. Neither should the threshold be too high because in a crash test against a soft, deformable barrier the airbag should not trigger too late; for example, crash test with 40% coverage at 25 mph (40 km/h) in accordance with FMVSS 208 (Rüter *et al*., 2007).



Figure 9. Stages of damage.

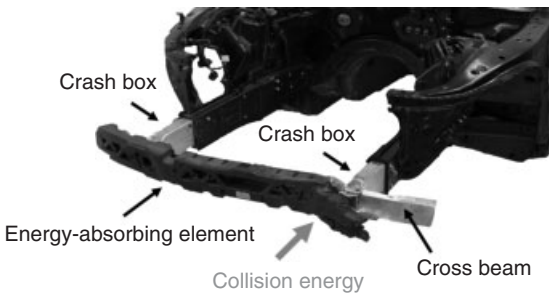


Figure 10. Bumper system after crash (BMW 5 series, type F10, European model).

6.2 Bumper skin

Bumper skins are design elements and often form supporting element for other components; for example, license plate, parking sensors, nozzles of the headlight washers, and distance sensors.

In addition, the skins are contributing more and more to conform to regulatory requirements such as in pedestrian protection tests. For this to be achieved, some of the kinetic energy of the impactor is converted to rotational energy. In addition, soft foams are integrated behind the bumper skin and in front of the cross member. Therefore, even more space is required to allow for the cross members as well as energy-absorbing elements. These requirements may lead the manufacturers to make the energy-absorption elements stiffer in order to ensure effective energy reduction over shorter distances. As the thickness of the skins is only approximately 2 mm, they are very flexible, and in an impact like that tested in the RCAR crash repair test, they absorb very little energy.

The outer panels are bolted to the front module. In the region of the fender, there are very often connectors, so that a slight collision can flex the outer skin of the body, and no lasting damage to the body occurs. At the 9.3 mph (15 km/h) impact speed tests according to RCAR procedure, the most likely damaged part is the bumper skin, so that a replacement is normally necessary. In the low speed crash tests by the IIHS of approximately 5 mph (8 km/h) impact speed, replacing the bumper skin after the crash test is not always required.

As with the crash-absorbing elements (impact damper and crash boxes), the bumper skins of the European version of a particular vehicle may be different to versions of the same model sold in the United States in order to comply with the different requirements (USA: FMVSS 215 "bumper standard" and Europe and worldwide: RCAR).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

6.3 Cross member

In a frontal crash, the cross member has to absorb as much energy as possible. This requires a force versus deflection curve, which remains as consistent as possible. Furthermore, the beams must have a defined bending rigidity, so that, for example, in a central pole rear impact (IIHS Test Specification), the transverse beam buckles immediately and forces are introduced into the longitudinal beams. In a crash with an overlap of 40%, in which only one side member is exposed directly to the load, the load should be guided to the opposite side longitudinal beam as well.

If the outer contact surface area of a cross member should fall below a minimum level, its resistance during crash tests with a deformable barrier will be reduced; for example, EuroNCAP with 40 mph (64 km/h) impact speed and 40% overlap. It may be necessary to design quite a large surface in order to prevent the barrier (representing the structure of another car) penetrating too deeply.

In relation to the "packaging," meaning the integration of the beam system into the overall concept of the front or rear module, there are very limited margins to the requirements the vehicle manufacturer has to fulfill (Figure 11).

Outer design specifications of the vehicle often determine the shape of the cross member. A highly curved bumper skin requires the inner parts to be adapted to the shape of the bumper. Especially, when designing compact cars, usually the case for the European and Japanese markets, very little deformation space is available and the deformation distance must be very short. The choice of locations for the connection points between the cross member and the side members may also be limited.

The material used for steel and aluminum cross beams is critical and the requirements for the systems has led to a noticeable tendency toward high strength materials. Hot-formed steel cross beams can achieve a tensile strength of about 1300 MPa, and the weight of the systems will be reduced.



Figure 11. Energy-absorbing element front, cross beam bolted to the side member (BMW 5 series, type E60, European model).

6.4 Crash-absorbing elements

Owing to the demand for in-expensive repair costs, especially in view of the RCAR test, energy-absorbing elements, for example, impact absorbers and crash boxes, have proven to be very effective. In the event of a crash at speeds up to 9.3 mph (15 km/h) by these measures should ensure that damage to the side member structure and other expensive parts is avoided in most cases.

Almost every car manufacturer uses energy-absorbing elements at the front and rear of the car. These specific elements absorb the energy resulting from crash loads at low impact velocities [collisions to about 9.3 mph (15 km/h) and pedestrian accidents]. Various energy-absorption elements can be used, including

- foam plastic elements;
- structured plastic elements;
- crush boxes;
- dampers.

Plastic elements, usually hollow body with reinforcing ribs are used as alternative solution for foam elements in order to prevent minor damage and for pedestrian protection. These elements are either integrated in the bumper skin or separately attached to the vehicle structure, for example, cross member.

For higher energy absorption designed to prevent structural damage, for example, in the RCAR test, about 5 mph (8 km/h) impact speed, foam elements and structured plastic parts are not alone sufficient because of their low strength and stiffness. Elements placed between the cross beam (bumper support) and the ends of side members should have a higher energy absorption capacity. For this purpose, there are various systems for the purpose of absorbing the energy, most of which use the common principle of the deformation of metals (usually, aluminum or steel).

Within the actual side member structure, an energy-absorption element can be formed by including regular folds in the member. While effective, this method still requires the replacement or repair of the actual side member. By incorporating a separate "crash box" of folded metal, which is bolted to the end of the side member, in case of damage, the cross beam and energy-absorbed elements are simply unbolted from the longitudinal member and replaced. The type and strength of the materials used for the crash boxes, side members and the cross beam (usually, aluminum and steel), and the design of the folds can be combined to produce the required amount of controlled deformation when an impact occurs, reducing the risk of damage to the main side members at the tested speeds.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Crash energy into the main structure can also be reduced by the compression of a round tube inside a tightly fitting or tapered outer sleeve. The inner tube is pushed into the outer sleeve in a crash, either expanding or seizing, as it does so. Other types absorb the energy by shearing, and in all these cases, nearly constant force-displacement characteristics can be generated. The advantages of these types of crash box are low weight and costs. All of these "crash box" solutions enable energy absorption by means of irreversible deformation.

Reversible crash absorbers, however, can absorb the energy from a crash load up to a certain limit and without plastic deformation get the original shape back. These elements can be installed as crash boxes between the crossbeam and the side rails as well.

In these systems, silicone, gases, and/or liquids act as shock absorbers and are usually reusable up to 9.3 mph (15 km/h), suiting the requirements of RCAR tests. It should be noted that the impact damper is designed for a particular stress direction. In the later RCAR low speed offset crash test (Section 5.1.2), the systems have to be adapted to the new test conditions; therefore, for example, the damper can be installed at an angle of 10° to the vehicle's longitudinal axis. Disadvantages of these reversible systems are the relatively high mass and the expensive production.

For simplified replacement, the advantage of bolting the components together is that the crash boxes, cross members, and longitudinal members may be made of different materials in order to perform correctly. Compared to the often-welded cross members in the past, there are advantages in production; for example, preassembly of the front module, inserting the engine from the front. In order to save more costs, a combination of common crossbeams and crash boxes can be used for various models.

6.5 Side member

The longitudinal members should be constructed with an increasing modulus of resistance from front to rear. This could be achieved by ensuring that the cross-sectional area or material thickness is larger toward the rear of the side member, for example. Buckle initiators can be used in order to restrict unavoidable partial repairs to the front section of the side beam. This can be achieved by means of varying stiffness or "weak" areas toward the front of the side members.

If a towing eyelet is permanently connected, for example, welded, to the longitudinal beams, an impact on the towing eye in various directions, for example, "snatching" when towing, can initiate torsional moments to the longitudinal beam, leading to failure of the member. In order not to adversely affect the deformation behavior of the longitudinal beams in case of collisions, the towing eyelet is provided with a thread so it can be easily fitted, but only when required.

7 SUMMARY

Good deformation behavior, ease of repair, and low cost of spare parts have a favorable impact on the running costs of a car. Often, these properties are subject to crash repair tests and are included in the calculation of insurance premiums. Relevant tests for evaluating the ease of repair of a car are the crash repair test by RCAR and the low speed crash tests by the US IIHS. In the RCAR test procedures, one front crash test and one rear crash test are described, in which the vehicle crashes at 9.3 mph (15 km/h [+1/−0 km/h]) with a 40% overlap against a rigid barrier, which is inclined by 10°. In the rear test, the vehicle is crashed at 9.3 mph (15 km/h [+1/−0 km/h]) with an overlap of 40% against a mobile barrier at an angle of 10°. The IIHS test procedure describes four tests for front and rear of the car.

Tests by the insurance industry have shown that the cross member systems have been optimized to such an extent, due in part to this test method, that in case of slight modification of the test procedure, the energy-absorption elements can fail and may result in considerably higher repair costs. This was taken as an opportunity to test and modify the RCAR and IIHS low speed tests.

Partly owing to the differing requirements of RCAR tests and the IIHS tests and partly owing to different markets, different systems are used. Energy-absorbing elements for vehicles that are intended for the American market may be different to those fitted to the same model offered on the European market.

In addition to the requirements in terms of ease of repair, there are many factors that influence the design of the crossbeam systems. Thus, only in terms of crash behavior in the initial impact phase can a variety of test procedures be considered. These include low speed repair crash tests conducted worldwide. Regarding the fulfillment of statutory safety standards, crash tests in the lower speed range (from 2 mph) are only considered in order to demonstrate the drivability of the car.

The cross beam systems also contribute to the occupant protection in high speed crash tests [up to 40 mph (64 km/h)]. Within the European Directive on pedestrian protection (2003/102/EC), which has been mandatory since 2005 for all new models, under the bumper skin and in front of the cross beam, soft, energy-absorbing elements are housed that are thick enough to satisfy the test type with the leg impactor. The majority of the energy absorption

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

is done behind the soft cross beam system. Therefore, the crossbeam system must absorb energy within a very short distance; otherwise, the impact forces are transmitted to the main structure, increasing the average deceleration of the entire vehicle as well as causing the total deformation to be too large and more expensive to repair.

Therefore, manufacturers are increasingly using absorption elements such as the impact-absorbing crash box placed between cross beams and side beams. These elements can absorb energy at minor crash loads [i.e., between 5 and 9.3 mph (8 and 15 km/h)] and prevent damage to the longitudinal beam to a large extent. Bolting the crash-absorbing elements on the longitudinal beams allows for a quick and cost-effective repair.

REFERENCES

Anselm, D. (1997) *Die Pkw-Karosserie*, Vogel Fachbuchverlag, Würzburg.

Gustafsson, H. (2001) RCAR'S History, the story of RCAR's birth and progress 1960–2000, http://www.rcar.org/About/RCAR%20History.pdf (accessed 3 September).

IIHS (1996) Low-speed crash test protocol, http://www.iihs.org/ratings/protocols/pdf/archived_versions/test_protocol_bumper_vI_1196.pdf (accessed 3 September).

IIHS (1997) Low-speed crash test protocol (Version II), http://www.iihs.org/ratings/protocols/pdf/archived_versions/test_protocol_bumper_vII_0897.pdf (accessed 3 September).

IIHS (2001a) Low-speed crash test protocol (Version III), http://www.iihs.org/ratings/protocols/pdf/archived_versions/test_protocol_bumper_vIII_0401.pdf (accessed 3 September).

IIHS (2001b) Low-speed crash test protocol (Version IV), http://www.iihs.org/ratings/protocols/pdf/archived_versions/test_protocol_bumper_vIV_1001.pdf (accessed 3 September).

IIHS (2002) Low-speed crash test protocol (Version V), http://www.iihs.org/ratings/protocols/pdf/archived_versions/test_protocol_bumper_vV_0502.pdf (accessed 3 September).

IIHS (2007) Bumper test protocol (Version VI), http://www.iihs.org/ratings/protocols/pdf/archived_versions/test_protocol_bumper_vVI_0407.pdf (accessed 3 September).

IIHS (2009) Bumper test protocol (Version VII), http://www.iihs.org/ratings/protocols/pdf/archived_versions/test_protocol_bumper_vVII_0609.pdf (accessed 3 September).

IIHS (2010) Bumper test and rating protocol (Version VIII), http://www.iihs.org/ratings/protocols/pdf/test_protocol_bumper.pdf (accessed 3 September).

IIHS (2012) Bumper test protocol previous versions, http://www.iihs.org/ratings/protocols/archive.html (accessed 3 September).

RCAR (2006) The procedure for conducting a low speed 15 km/h offset insurance crash test to determine the damageability and repairability features of motor vehicles, http://www.rcar.org/Papers/Procedures/rcar_test_protocol_angled_barrier.pdf (accessed 3 September).

RCAR (2008) Design guide - a manufacturers' guide to ensure good design practice for repairability and limitation of damage, http://www.rcar.org/Papers/Design%20Guides/DesignGuide_v1_1.pdf (accessed 3 September).

RCAR (2010) Bumper test procedures, http://www.rcar.org/Papers/Procedures/BumperTestProcedure.pdf (accessed 3 September).

Rüter, G., Zoppke, H., Bach, P., *et al*. (2007) Einfluss des Versicherungs-Einstufungstests auf die Belange der passiven Sicherheit, http://www.bast.de/cln_030/nn_42640/DE/Publikationen/Berichte/unterreihe-f/2007-2000/f62.html (accessed 3 September).

UN (1980) Regulation No. 42, http://www.unece.org/fileadmin/DAM/trans/main/wp29/wp29regs/r042e.pdf (accessed 3 September).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto254
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Adaptive Restraint Systems: Toward Integral Safety

Klaus Kompass and Manfred Schweigert

BMW Group, Munich, Germany

Christian Gruber and Christian Domsch

BMW AG, Munich, Germany

Ronald Kates

Ronald Kates Consulting, Munich, Germany

1 Introduction 1
2 Potential of Adaptive Restraint Systems 2
3 State-of-the-Art and Future Restraint Systems 3
4 Sensor Technology 5
5 Effectiveness Analysis 6
6 Effects of Unneeded Activation 9
7 Conclusions 10
References 12

1 INTRODUCTION

Vehicle crash sensors record the accelerations of car components and pressure increases in door cavities. If defined thresholds are exceeded, several restraint systems are triggered in stages to protect the vehicle occupants. The airbags and seat belts deploy their protective functions like clockwork and react adaptively to variable occupant characteristics and accident severities.

Inertial sensors measure translational and rotational vehicle accelerations during vehicle motion. These kinematic data are continually compared to the control input given by the driver. If dynamical thresholds are exceeded and critical driving situations are predicted, for example, stability control systems work to stabilize the vehicle and thus support controllability.

The increasing variety of available driver assistance systems has led to stronger utilization of forward-looking sensors. These sensors monitor the car's surroundings. They also monitor the driver's current state and attention level. In traffic situations requiring higher attention, driver assistance systems support situational awareness by targeted information and/or warnings. In addition, suspension control and brakes are preconditioned in order to provide maximum response at the first sign of a driver reaction. If the driver fails to react independently in time to avoid an impending collision with a pedestrian or a leading vehicle, the car can trigger an automatic braking sequence. By reducing the collision speed, the accident consequences for all involved persons would be reduced in a considerable percentage of conflicts.

Information concerning the vehicle and driver states and the traffic environment enable targeted preparation of the vehicle and the occupants for an impending collision. To this end, all available data from forward-looking, inertial, and crash sensors can be fused to predict crash criticality. In accident scenarios compatible with measurement of the so-called precrash data, reversible actuators can be deployed to precondition the occupants and the vehicle by moving them into optimal positions and orientations. Irreversible restraint systems can also be pretensioned to deploy adaptively,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.

This article is © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

according to the situation, and provide protection at the earliest possible moment following the collision. In addition, protective strategies are deployed in a controlled manner and adapted to the constitution of the occupants, the accident scenario, and the collision speed. This integrated safety approach links active and passive safety systems and aims to reduce the risk of injury even beyond what can be provided by passive safety systems alone.

This summary has described the potential and the limits of adaptive safety systems for preconditioning the vehicle and the occupants before an accident. In the following, the achievable protective performance will be illustrated using several examples; methods for reliable and comprehensive assessments of the protective sequence of integral safety systems under laboratory conditions will be elucidated.

2 POTENTIAL OF ADAPTIVE RESTRAINT SYSTEMS

Vehicle safety has already achieved a very high level of performance. The risk of serious injury or mortality in traffic accidents has been steadily decreasing in recent years despite increasing exposure (mileage driven), primarily because of the development of passive safety systems.

On the other hand, it is evident that further development of passive safety is reaching a point of diminishing returns, although some weak points still can be identified based on our comprehensive knowledge of passive safety development. However, classical methods may not suffice to further improve passive safety.

For one thing, collision scenarios in tests can never precisely match the spectrum of accidents under realistic traffic conditions. Moreover, the true positions of occupants within the vehicle can deviate markedly from an optimal configuration—not to mention unbelted occupants. Presumably for an increased sense of comfort, seat backs are often quite reclined, seatbelts are fastened with too much slack, or tightening of the seatbelt is hindered by thick winter clothing.

In addition, the variance among occupants is quite large: systems must protect people who are short, tall, thin, corpulent, and all combinations thereof. Demographic trends result in increasing proportions of older drivers and passengers. They have stricter requirements on maximum stresses because of their more fragile physiology, including vulnerability to skeletal, muscular, and internal injuries. Indeed, it is certainly beneficial even for average occupants to minimize the forces acting on the body in an accident.

The key mechanisms that can be addressed by adaptive restraint systems are matching of the occupant's forward displacement in a crash to the accident severity, reduction of seat belt tension acting on the occupants, and/or adaptation of this tension to the forces acting in the crash.

In recent years, research on occupant restraint systems has focused on the development of adaptive restraint systems. Adaptive restraint systems are able to adapt their behavior to individual occupant characteristics and crash conditions using information from sensors that gather information during the precrash and crash phases. They can provide the best compromise between occupant load and forward displacement in given situations. Nowadays, adaptive load limiters and airbags are already available in cars. However, they currently only distinguish two classes: occupants smaller than a "5th percentile woman" versus all larger occupants. Scientific research is currently focusing on more advanced adaptive safety measures, for example, the so-called continuous restraint control systems, which can continuously optimize their configuration during impact considering individual occupant characteristics and crash severity.

The potential of adaptive restraint systems has been shown in a parameter study in the PRISM project of the European Union, where simulations were performed with MAthematical DYnamic MOdels (MADYMO) based on late-model vehicle models and various dummy sizes. In the frontal Euro NCAP crash test, the ISS1 value of the occupants could be significantly reduced (up to 60%) by adapting various restraint system parameters to the specific occupant (Lemmen *et al*., 2005). Extreme cases included a small driver close to the airbag, a very large driver, or a driver missing the airbag entirely. The restraint system was optimized for each case by, for example, varying the load limiting level of the seat belt between 2000 and 7000 N and varying the firing time of the driver airbag between 10 and 40 ms.

Hesseling (2004) has investigated the possibility of controlling airbag and seat belt behaviors to mitigate the consequences of a crash for the occupants. By controlling the airbag mass flow and vent size and the belt force, a significant reduction in chest and head accelerations could be achieved. A MADYMO simulation model of a mid-size passenger car was used with a dummy, representing a 50-percentile (median) male. Simulation of a US NCAP frontal crash test2 with this simulation model and the controlled restraint system showed a reduction of chest and head injuries by 60% and 50%, respectively, compared to the same test with a conventional restraint system. Similar results were found in simulations with a 5-percentile female dummy. However, conclusions drawn from this research are subject to several limitations because of idealized assumptions: it was assumed that the crash pulse (the deceleration profile of the vehicle during the crash) was known *a priori*;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

that ideal restraint actuators were present; and that all required measurement data, such as biomechanical occupant responses, were available in real time.

The aforementioned research approaches show that adaptive restraint systems may offer considerable potential in terms of injury reduction in frontal vehicle crashes. The challenge in the implementation of these systems in production vehicles is to ensure that the adaptive system will always perform as well or better than the corresponding nonadaptive system in terms of injury reduction. When an adaptive restraint system receives incorrect information on, for example, occupant size or impact conditions, the occupant's injuries could theoretically be more severe than in the case of nonadaptive systems. Therefore, a study will be performed on the malfunctioning of adaptive restraint systems and the consequences thereof. To be able to make a comparison between the two systems, a system definition is needed, for both a current car with nonadaptive restraint systems and a future car with adaptive restraint systems.

3 STATE-OF-THE-ART AND FUTURE RESTRAINT SYSTEMS

3.1 Modern restraint systems

The protective performance of modern restraint systems begins a few milliseconds after collision detection. The data required to trigger the appropriate "choreography" of an adaptive restraint system are measured by vehicle sensors.

Occupants are classified at the beginning of the trip, for example, by pressure-sensitive seat cushion mats and by detection of the longitudinal seat position. In most cases, two occupant classes are currently distinguished: occupants smaller than a "5th percentile woman" (height less than 152 cm and weight less than 54 kg) and all others.

Numerous sensors measure accelerations occurring at different points of the chassis and pressure increases in the door cavities. In a collision, the severity and impulse direction are derived from sensor signal data, particularly the rise in signal amplitude. Front, side, and rear collisions can be clearly distinguished. The crash severities are divided into different stages. Once the signals exceed defined thresholds and have been verified for self-consistency, in the majority of cases, pyrotechnic propellant charges are ignited to activate belt pretensioners and airbags.

The use of adaptive restraint systems enables the selection of trigger times for multistage airbags and force characteristics for adjustable safety belts. Thus, the protective effect of the restraint system can be optimally adapted to different accident scenarios, their resulting crash severities, and classes of occupants. These adaptive capabilities are implemented by several different techniques. Airbags can be fitted with two pyrotechnic propellant charges, whereby the stages can be triggered simultaneously, separately, or with a time offset. Additional airbag volume can be released or ventilation holes can be opened according to time or pressure actuation by means of controllable air bag tethers. The diverse range of possibilities extends to self-adaptive solutions, in which ventilation openings are closed or remain open after physical contact with the occupant depending on his or her size. Belt forces are adapted by the so-called "force limiters" using linear, digressive, or switched controllers. The required belt forces are generated and modulated using friction, straining, or torsion effects.

By coordinated action of all these adaptive components in an accident, the system aims to provide a restraint capability with the lowest possible burden and optimized to the accident severity and the occupant classification.

3.2 Controlled restraint systems

Until now, progress in implementation of restraint systems has included self-adaptive and multistage devices. A further stage of development could involve protection systems with feedback controllers that can adapt their restraining action within milliseconds during the entire course of the crash. This development would of course require rapid and reliable sensing of all relevant feedback parameters, real-time processing of these data in the controllers, and a mechatronic actuator concept for adaption of the restraining system components. The sensor, restraint system, and controller requirements would need to be improved by substantial factors compared to the state of the art. While some progress has been made, there remain fundamental technical challenges.

One possible solution could utilize a precrash model: for example, just before an impending accident involving two cars, all data from both cars relevant for prediction of accident severity and estimation of occupant injury risks would be measured by forward-looking sensors. Here, it is also conceivable to send the relevant data shortly before the crash via car-to-car communication systems. On the basis of this sensor data, the optimal triggering strategy for the controllable restraint system could then be determined either by real-time prediction of the collision dynamics or by a knowledge-based approach.

3.3 Preconditioning before T_0

Using current technologies, it is possible to precondition the vehicle and its occupants for an impending collision. Once the predictive sensors have detected an unavoidable

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

collision, the occupants and vehicle components could be moved to optimal positions using reversible actuators. In this way, the full potential of both the vehicle structure for energy absorption and restraint systems for the reduction of injury risk can be optimally utilized. If the driver manages to avoid the accident or if a collision does not occur, all actuators would be restored to their original positions. Reversible electromotive belt pretensioners could play a particularly important role in accident scenarios with automatic brake intervention. Before emergency braking, they link the occupant to the seat structure at an early stage and reduce the slack in the belt, which may be present because of an effect similar to slack in a film reel. The deceleration caused by forward displacement of the upper torso is reduced, and the occupants remain in a favorable position for the seat airbag. Occupants sitting with highly reclined seat backs just before the collision are rapidly shifted to an optimal, upright seating position. In addition, by seat angle adjustment, a seat ramp is formed, which helps prevent "submarining," the tendency to slide out under a seatbelt in an accident. An adaptive shock absorber control on the front axle counteracts brake pitch by setting a more rigid damping characteristic just before automatic braking. Thus, the crash-absorbing structures of the front end maintain the best possible overlap with the structures of the collision partner. If a sensor system becomes available that can detect an unavoidable accident situation with a very high reliability, it would be feasible to design an irreversible precrash system.

Various factors play a role in the design of a restraint system. Use cases demanded by lawmakers and by several consumer groups must be fulfilled by one single restraint system on crash test dummies differing in construction and size percentiles. Car manufacturers are succeeding in achieving these rather high occupant protection goals increasingly often, as evident from a glance at EuroNCAP evaluation results of current vehicles. On the other hand, more and more effort is required for small safety gains. This raises the question of how occupant safety can be enhanced by intelligent concepts in an efficient and customer-oriented way.

The fact that vehicles have been designed for human occupants rather than dummies suggests an approach. A look at customer-driven vehicles reveals that occupants do take a large variety of seating positions. In the case of the driver, seating positions in the real world closely correspond to those in crash tests, because these positions are geometrically constrained within narrow limits by the requirement that the driver must reach the pedals and the steering wheel as well as maintain a good view of the instruments and the vehicle exterior. Hence, for a given driver size percentile, seating positions remain within narrow limits. In the case of the passenger seat, these limitations do not apply, so that predominantly comfort aspects have an influence on the seat position chosen. Another factor regarding passengers is that they often simply keep whatever position was previously set, only to avoid the effort of positioning the seat for only a "few minutes." In either case, the result is a higher probability that the passenger side occupant sits in a less favorable position for a crash.

While satisfying the comfort needs of some passengers, highly reclined seat back settings are not recommended for safety reasons. The belt then does not make proper contact around the chest and pelvis and can slip into the rib cage area (Figure 1). In a crash, submarining can occur, where the occupant slides out under a seatbelt and, in the worst case, can suffer severe internal injuries. A suitable approach toward reconciling the conflicting demands of safety and comfort could be to use the power seat motors automatically to bring the seat back into an upright position before an impending crash.

The seatbelt itself offers further potential for improvement. Despite the installation of pyrotechnic belt

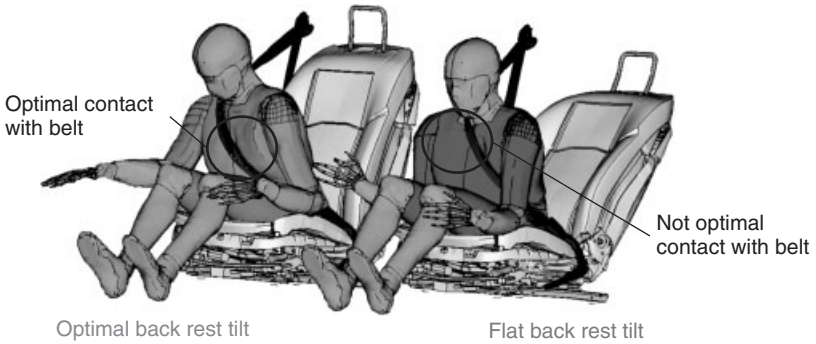


Figure 1. Crash simulation. light gray: upright seat back positioning with optimal seat belt contact. dark gray: highly reclined (flat) back rest with suboptimal seat belt contact during the crash.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

pretensioners as standard equipment, loose fitting belts cannot always be entirely avoided. Customers rarely tighten the seat belt to remove slack after buckling the belt. Here, a reversible electromotive belt retractor (REMA) can realize its safety potential by gently tightening the belt after it is buckled and more strongly pretensioning the belt in critical driving situations. REMA has a high relevance in practice, whereas crash tests assume that the vehicle is not braking; in reality, braking often occurs just before the crash. In these cases, REMA provides the additional advantage that timely pretensioning counteracts the forward displacement of the occupants resulting from their forward momentum in the braking vehicle. If the crash does occur, more distance remains available to spread out and thus further reduce the forces on the occupants.

Nowadays, in some cars, the already high safety level with standard equipment is extended by the features presented here. If the vehicle detects a precrash situation, it corrects an excessive seat back reclining (in certain seat configurations) to the extent possible. Electromotive belt pretensioners are activated on the driver and the passenger sides, resulting in a reduction of possible belt slack and prevention of excessive forward displacement of the occupants in case of strong braking. In accordance with the situation, open windows and the sunroof are closed. These strategies serve to reduce the risk that extremities—which flail about because of violent motions of the vehicle—might hang out of the vehicle, or that external objects might penetrate the vehicle through the openings.

Basically, two classes of precrash situations can currently be detected: critical driving states and unavoidable rear-end collisions. Critical driving states, for example, skidding or driver-initiated emergency braking, can be detected by sensors relatively easily. In these cases, the occurrence of a crash cannot be predicted with certainty, and in case of a crash, whether it will occur within seconds or even milliseconds. However, if the system is triggered in such a critical situation without a subsequent crash, customers still do not perceive the triggering as a false alarm. On the contrary, the driver is sensitized to the ability of the system to correctly trigger in critical situations and tends to see this triggering as a positive confirmation of the decision to purchase the system. The second class of situations consists of unavoidable rear-end collisions detected using environmental sensors installed for driver assistance systems, for example, cameras or radar sensors. The distinction between truly critical precrash situations from "sporty" but uncritical driving maneuvers is not that simple, so that a reliable crash prediction is possible only a few hundred milliseconds before the crash. Unjustified triggering is more intrusive for the customer in these cases, which implies that the classification as an unavoidable crash must be made late enough to avoid false alarms/triggers.

Future improvements will optimize these features. Faster seat motors are required to improve the chances of achieving the optimal seat position within the precrash in situations detected by environmental sensors. In addition, the range of reliably detectible precrash events will be extended to collisions with a vehicle located to the rear as well as side crash constellations, which are even more difficult to detect in a timely manner.

The currently available and future functions described here illustrate how adaptive capability mentioned earlier can be achieved in practice. Even if the customer ignores the recommended settings for seats and seatbelts for whatever reason, these functions aim to realize the full occupant protection potential of the conventional restraint system.

Another reason for the improvement of safety functions is the demographic change (aging society). Older people usually withstand only lower restraint forces than younger people do. Owing to the fact that the correlation between displacement and force is physically limited, it is reasonable to use a compromise between displacement and restraint force in all kinds of accidents. Adaptive restraint systems can help to optimize this compromise.

At this point, it should be mentioned that the age of the occupants is not a relevant parameter for designing the safety systems. The aim should always be to offer the lowest possible occupant loads for all kind of passengers. Owing to this, a decision between older or younger passengers is not required.

4 SENSOR TECHNOLOGY

To accommodate the use of the various restraint systems, several sensors are present in the vehicle to measure relevant parameters. In this case, the sensors that can be used in the BMW 5 series of the model year 2012 will be described as typical for current premium vehicles. The following sensors are present in the restraint system and relevant for a frontal crash:

- Acceleration sensors measure whether the vehicle's acceleration level exceeds a certain value, meaning the vehicle is in a crash situation.
- Buckle sensors measure whether an occupant is belted or not.
- A sensor in the seat foam of the passenger seat determines whether an occupant is present. The sensor gives three possible outputs: no occupant seated, child seat mounted, or occupant seated.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

- A seat position sensor recognizes whether the seating position is in the front. Therefore, the occupant may be a 5th percentile female or smaller.

Integrated safety will play an important role in the further improvement of vehicle safety (Dannenberg, 2004; Heudorfer and Meissner, 2008; Bogenrieder *et al*., 2007). A future car is, therefore, likely to contain sensors to measure crash conditions and occupant characteristics, such that optimal protection is possible for each occupant.

Currently, research focuses on vision systems, occupant classification, and tracking sensors and wireless communication between cars and infrastructure.

4.1 Restraint system sensors

4.1.1 Object detection

To be able to accurately detect an oncoming crash, automotive manufacturers have been working on various solutions for object detection such as radar- and camera-based systems. For example, the Mercedes S-class (model year 2012) currently can be equipped with short- and long-range radars (two 24 GHz radar devices and one 77 GHz radar device, respectively) to detect upcoming vehicles and obstacles (Schittenhelm, 2009). The BMW 7 series of the model year 2013 uses a sensor system of one combined mid- and long-range 77 GHz radar sensors and a mono-video camera. Infrared sensors for obstacle detection at night (Jerzembek, 2008) are available for some premium vehicles.

It is expected that the costs for sensors for object detection, particularly for video- and radar systems, will decrease significantly in the near future, so that it is safe to assume that these systems will be present in future nonpremium cars as well (Whydell, 2010).

4.1.2 C2X communication

Next to the development of vision systems, wireless car-to-car and car-to-infrastructure communications (in short, C2X-communication) could also contribute to a safer vehicle. These systems would be able to warn the driver about a possible oncoming hazard; for example, when a vehicle a few cars ahead suddenly performs emergency braking. With car-to-infrastructure communication, a traffic light could provide data predicting when the red phase will begin, so that a driver who fails to stop could receive a timely warning (Kompass, 2008). Currently, several tests are being conducted to develop a system that functions reliably between cars from different manufacturers and different kinds of infrastructure (Badstübner, 2008; Herrtwich, 2008). Car manufacturers and scientists agree on the potential of C2X communication for improving traffic safety, and it is, therefore, likely that this technology will be part of future vehicles (Hartenstein *et al*., 2004). In the sim^{TD} research project, car-to-x communication and its applications are being tested in a large-scale test field around the German city of Frankfurt am Main (www.simtd.org, 2012).

4.1.3 Occupant sensing

Finally, an occupant-sensing system is needed to trigger adaptive restraint systems. This system needs to be able to classify an occupant in terms of weight, size, and seating position, measure the motion of the occupant, and provide information on the possible type and severity of injuries of the occupant. With this information, the adaptive restraint system can provide optimal occupant protection in case of a crash. However, unlike the vision and communication systems mentioned earlier, automobile manufacturers and scientists have not reached consensus on a reliable occupant-sensing system concept.

It is very likely that future cars will be able to predict an oncoming crash and that this information can be used to adapt the restraint systems to this specific crash scenario. It is expected that in future, vehicles information on crash severity will be available before the crash occurs.

5 EFFECTIVENESS ANALYSIS

In the development of adaptive safety systems, attention should be focused on the true effectiveness in real accident situations with respect to avoided collisions and reduced severity. However, this assessment is difficult to extract from statistics, as completely avoided accidents ("near misses") are not recorded in accident databases, and injury accident statistics do not provide data on complete avoidance of injuries. Moreover, it is important to assess whether a proposed safety measure will have a positive effect on accidents before introduction of the measure. For this reason, virtual methods of safety assessment are key elements in the entire development process for integral safety.

Large variances are inherent in the spectrum of relevant accident and risk situations and in the influence of driver reactions and responses. These variances considerably complicate the assessment of integral safety system effectiveness. No definitive standard test cases have yet been established, and system design needs to be oriented to effectiveness in the field—rather than optimized for testing. In order to achieve reliable assessment, a linked framework including driving dynamics, system operation, and crash

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

simulation is needed. Standards for determining the quality of these simulations have not yet been established and are also needed.

Stochastic simulation (also known as *Monte Carlo simulation*) offers a method to assess the effectiveness of safety systems rapidly—while considering the spectrum of traffic situations in which the systems might be triggered, the variability of possible driver responses, and other variable factors. A target scenario is simulated a large number of times with stochastically varying characteristics (on the order of one million runs). All relevant quantities (e.g., the braking force of a driver after a warning) are drawn anew in each run from appropriate, calibrated frequency distribution models. Owing to the large number of computational runs, a large variety of possible combinations of influence factors are drawn, including unusual but critical combinations. Simulation results vary correspondingly: most runs involve enough favorable factors to avoid an accident; a small proportion of runs—generally involving several unfavorable parameters—result in virtual accidents of varying severity or in near misses. If the distributional model parameters and the process model are close to the real scenario, then the distribution of accidents and severities should resemble the corresponding data in reality. This similarity can be verified by studying appropriate accident databases. Now, a novel safety system can be modeled and introduced into the virtual simulation environment in order to assess the effectiveness in a controlled virtual experiment by comparing the changes in accident frequency and severity with and without the system.

In order to compute a useful metric from stochastic simulation, injury severity distributions need to be generated from the state variables describing the virtual occupants and the virtual collision. Currently, these distributions are usually derived from statistical injury risk curves (usually obtained from logistic regression models) based on accident databases. However, the level of modeling detail is severely limited because of the statistical power and possible sampling biases of existing databases.

If a more precise or differentiated (i.e., by body part) assessment of injury probabilities is desired, a crash simulation with its corresponding computational effort could be utilized. However, even this method does not generally consider the motion of the vehicle and the occupants before the crash; for example, if the vehicle goes into a skid or braking is very strong. Appropriate simulation tools can of course be implemented to include vehicle dynamics and occupant motions resulting from the vehicle dynamics as well. The challenge lies in combining the various simulation domains, which typically have different time and spatial resolutions as well as modeling depth. By appropriate tools for cosimulation of different models, a comprehensive toolkit can be generated. A comprehensive simulation sequence enables assessment of integrated safety systems with active and passive elements on the basis of injury severity as a common metric. There is still a need for standardization of requirements on simulation models for effectiveness analyses.

Figure 2 illustrates a computation with a simulation sequence as described earlier. Simulation of vehicle dynamics provides a representation of the vehicle's

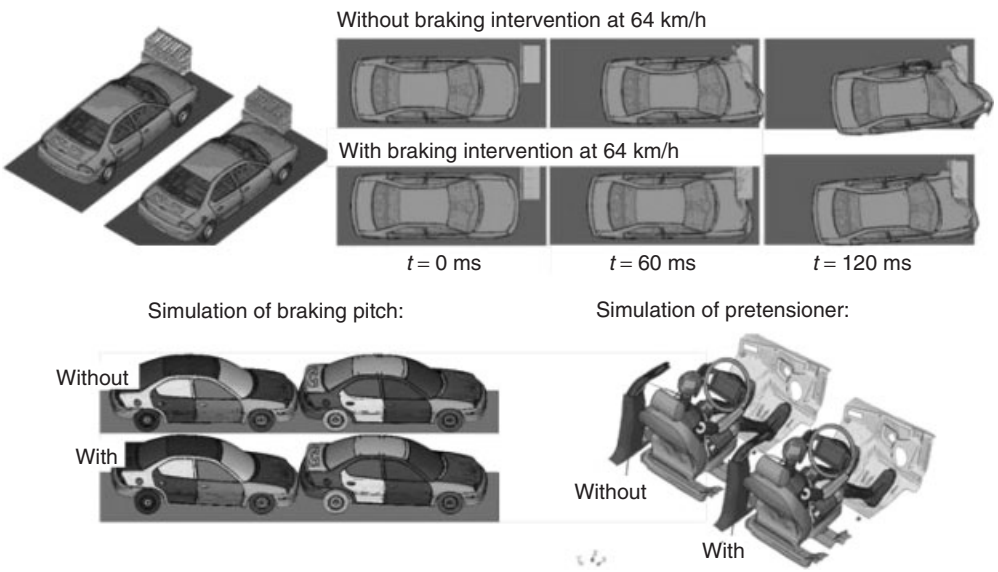


Figure 2. Simulation of offset frontal crash with braking intervention and precrash seatbelt pretensioning.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

shift in orientation during braking. At the same time, forward displacement of the occupants with and without pretensioners can be shown. The starting configuration of the crash simulation varies correspondingly; the subsequent crash simulation then provides the computed occupant forces.

Any simulation requires validation, however. Validation is implemented using testing in a real environment. However, there are as yet no standardized methods established for testing adaptive and integral safety systems.

Currently, vehicle safety is dominated by passive safety measures. Correspondingly, test procedures are designed to assess the effectiveness of passive safety measures, which by definition begin with the crash itself. Thus, the benefits of any safety measures that work by preconditioning are not assessed by laboratory crash tests, although these measures could exhibit a high effectiveness in real accidents.

An example is reversible electromotive belt pretensioners. Their effectiveness is essentially attributable to the pretensioning of seat belts before the impending collision. They allow pretensioning before the collision has even taken place because they are reversible if the collision in fact does not occur. In real accidents, a large number of drivers will brake strongly before the collision (Figure 3), leading to forward displacement of occupants.

A conventional crash test (constant speed) without forward displacement thus underestimates the system's effectiveness. This assessment deficit causes an unnecessary disincentive for introducing such systems.

In the future, vehicles will be increasingly equipped with systems that warn the driver of an impending rear-end collision and support braking. These systems may institute automatic emergency braking if there is an unavoidable collision and the driver fails to react. Owing to the sharp braking of the vehicle, the occupants will normally experience forward displacement. This can lead to an occupant position closer to the airbag, so that the space available to move forward during the crash becomes smaller.

A reversible electromotive pretensioner can be activated as soon as an environmental sensor detects that the distance to the preceding vehicle is so small that an immediate crash seems imminent. Strong braking by the driver or an unstable driving state can also activate a reversible electromotive pretensioner. This activation couples the occupant tightly to the restraint system. In the event that the collision does not in fact occur, the pretensioning can be reversed. The consequences were shown in a crash test considering the effects of reversible systems.

It has not yet been possible to test the effectiveness of forward-looking occupant protection systems by crash tests with test rigs used up to now. However, innovations in vehicle safety will be introduced in increasing numbers of vehicles in coming years. For this reason, we need test methods capable of reproducibly testing the effectiveness of precrash functions. To this end, it would be advisable to develop a policy for implementing novel technical functionalities of traffic safety within currently valid standards.

A demonstration test was developed in a cooperative effort between DEKRA and BMW that showed how the effect of forward-looking protective systems can be evaluated in a crash test (Berg, Rücker, and Domsch, 2011).

A BMW 530d prepared especially for this test was able to determine the distance to the relevant target (in this case, the crash block) as well as the relative velocity, on the basis of radar sensor data from the ACC (active speed/distance control device). In this way, it was possible to record the

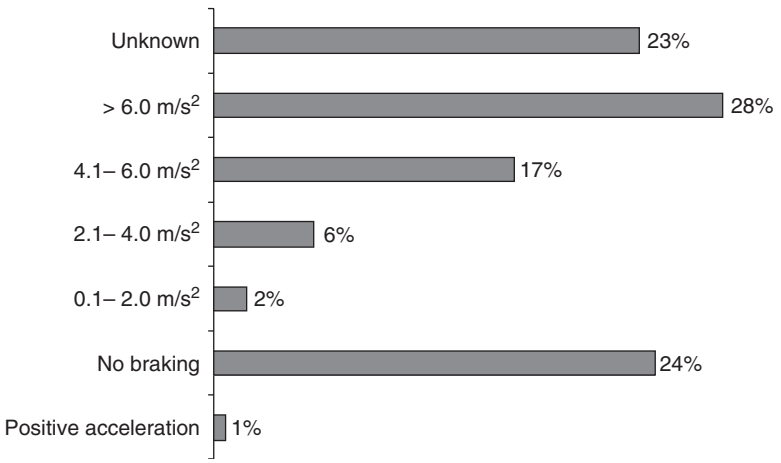


Figure 3. Frequency distribution for braking decelerations during the precrash phase ($N = 1492$ accidents, (GIDAS).)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

entire signal sequence starting from the sensor right up to the safety system reaction (i.e., triggering of automatic emergency braking). The response of the safety system in the test thus corresponded precisely to the corresponding response in a comparable real accident scenario.

While the vehicle was approaching the crash block, various safety functions—in some cases, prototypes—were activated. In addition to the ACC radar sensor with special object detection, identification, and selection capabilities, a stability control system (electronic stability program, ESP/dynamic stability control, DSC) with prototypical functions was required to implement an automatic emergency stop under full braking. In addition, the vehicle was equipped with reversible electromotive belt retractors for the driver and the passenger. The strategy for warning the driver and initiation of emergency braking was also implemented as a prototype. Finally, precrash deactivation of the fuel pump was implemented. The automatic emergency calling function corresponded to the standard BMW assist production version ("advanced emergency call") and was also utilized.

In the course of the test, the "point of no return" was reached, where a collision was no longer physically avoidable by any driver reaction (neither braking nor swerving). The automatic emergency braking function was initiated at this moment, producing the maximum vehicle deceleration possible on the basis of the traction between the tires and the road.

An initial speed of 64 km/h was selected for the crash test run, corresponding to the initial speed of the Euro NCAP or IIHS unbraked frontal offset tests. As in the Euro NCAP and IIHS frontal offset tests, there was a 40% overlap of the vehicle front with the deformable barrier. The driver and the passenger were represented by belted, instrumented dummies (Hybrid III 50th percentile male). Child dummies were not used.

In contrast to normal test procedures, in which precrash systems are not permitted to be activated, they were intentionally activated here. After the vehicle had been accelerated to the test speed, it approached the crash block with constant velocity. The sequence of safety measures was defined as follows: At a TTC of 2.1 s (TTC = time to collision, defined assuming the vehicle maintains its current velocity), the driver is warned optically of the impending rear-end collision. This warning is displayed as a red warning symbol on the instrument panel and as a warning symbol in the head-up display. Thus, the driver sees the symbol directly in his or her field of view. At the same time, the brakes are prefilled, and the triggering threshold of the brake assistant is lowered. In the system configuration presented here, an acute warning is given to the driver at TTC of 1.7 s, in which the optical warning is accompanied by an acoustic alarm. The reversible belt pretensioners are activated at TTC = 1.1 s, in order to avoid forward displacement of the occupants during braking. Automatic emergency braking is initiated at TTC = 0.9 s. In the test, the speed was reduced from 64.8 to 40.4 km/h (−38%).

The test rig controller was designed to detect the automatic precrash vehicle deceleration and reduce the speed of the towing cable correspondingly.

Even in the unbraked crash test, the production BMW 5 series vehicle had exemplary safety behavior. These results are highlighted by outstanding assessments in US-NCAP, EURO-NCAP, and IIHS test procedures. In the test with precrash braking, because of the reduced speed, the measured forces on the occupant dummies were further reduced by considerable amounts compared to the 64 km/h unbraked crash. The relative changes of several characteristic force values for the driver and the passenger dummies are displayed in Figure 4

6 EFFECTS OF UNNEEDED ACTIVATION

Adaptive restrain systems can contribute to a minimization of the risk of occupant injuries. However, it is important to understand what effects can occur in case of an unneeded activation. The following discussion illustrates the influence of a one-sided optimization.

To this end, consider an adaptive restraint system optimized for a 50 km/h crash into a rigid barrier. Suppose that the restraint system can be adapted using available data from the precrash phase of the situation. Now, the precrash phase data could be subject to errors for various reasons. This consideration raises the question of what influences uncertain or faulty data could have on the effectiveness of the system.

One example of an unneeded activation will be investigated: the velocity of a full overlap rigid barrier crash is wrongly detected and the settings of the restraint system are, therefore, not optimal for the situation. Two simulations are performed: one with a velocity of 40 km/h and the other with a velocity of 60 km/h. The restraint system settings are selected to be optimal for the case of

Table 1. P values in (%) for standard and unneeded activated restraint systems.

	P_{Standard} (%)	P_{Unneeded} (%)	Relative difference (%)
40 km/h	12.7	8.5	−33
60 km/h	14.9	45.2	+203

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

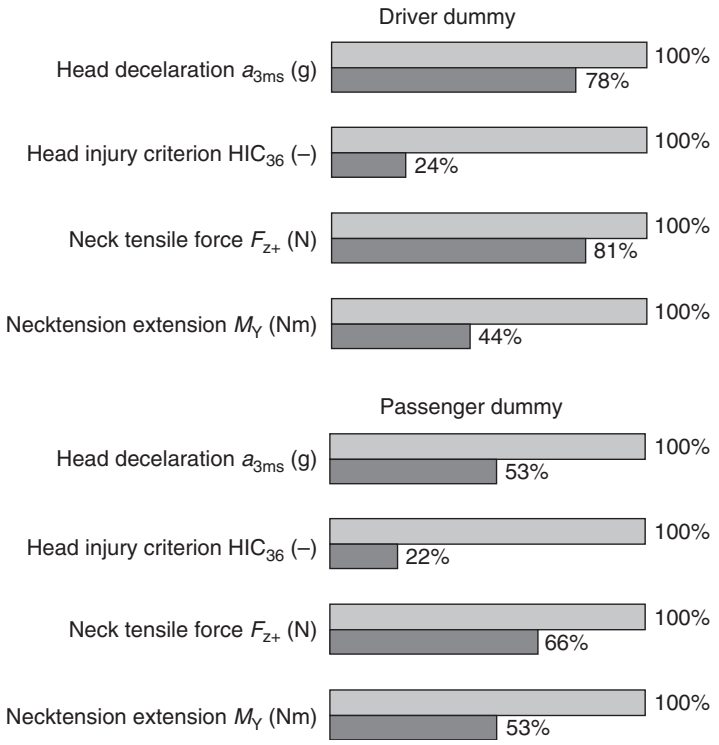


Figure 4. Comparison of relative injury measures to driver and passenger dummies in a braked crash test at 40 km/h collision speed compared to an unbraked crash test at collision speed of 64 km/h (normalized to 100%).

50 km/h. The dummy's injury level is compared with that of simulations with the standard restraint system at the same velocity.

The resulting *P*-values are shown in Table 1 for the standard restraint system ("standard") and the adaptive restraint system ("unneeded"). Each component of the *P*-value is displayed in Figure 5, together with the values of the various crash simulations.

If the relative difference is considered, it is clear that the *P*-value increases significantly for a velocity of 60 km/h. This is caused by the fact that the dummy hits the steering wheel with its head, which leads to a high probability of severe head injury ($P_{\text{Head}} = 35\%$) as can be seen in Figure 5a.

However, in the case of 40 km/h, the *P*-value is improved by 33% relative to the standard case. This is caused particularly by decreased values of chest deflection (Figure 5b) and Nij (Figure 5d).

Summarizing, measurement errors in adaptive restraint systems can lead to more serious injuries, particularly if the accident severity is considerably higher than forecasted. This conclusion implies that further development is required both to reduce errors to a minimum and to make sure that errors do not have negative effects on occupant injuries.

7 CONCLUSIONS

Traffic safety has made great strides in recent years, but the development of conventional, strictly passive safety devices is approaching a point of diminishing returns.

Adaptive restraint systems can contribute substantially to reduction of forces on the vehicle occupants during crashes. By combining these systems with integral safety systems that precondition the vehicle and its occupants or trigger additional appropriate vehicle responses, one can further reduce injury risks in serious accidents. This reduction is due to avoidance of some accidents entirely, decreased relative collision speeds, and individualized and accident-specific optimization of structures and restraint systems; the protective effect is adapted both to the occupants characteristics and to the details of the collision dynamics.

In system development, the problem of false positives is an important consideration. The collision is 100% certain according to vehicle sensors only at the moment of contact. As a prediction of collisions and their dynamics become increasingly uncertain with increased time to collision, in many cases, reversible protective measures are the only viable options. Even then, a careful coordination is

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

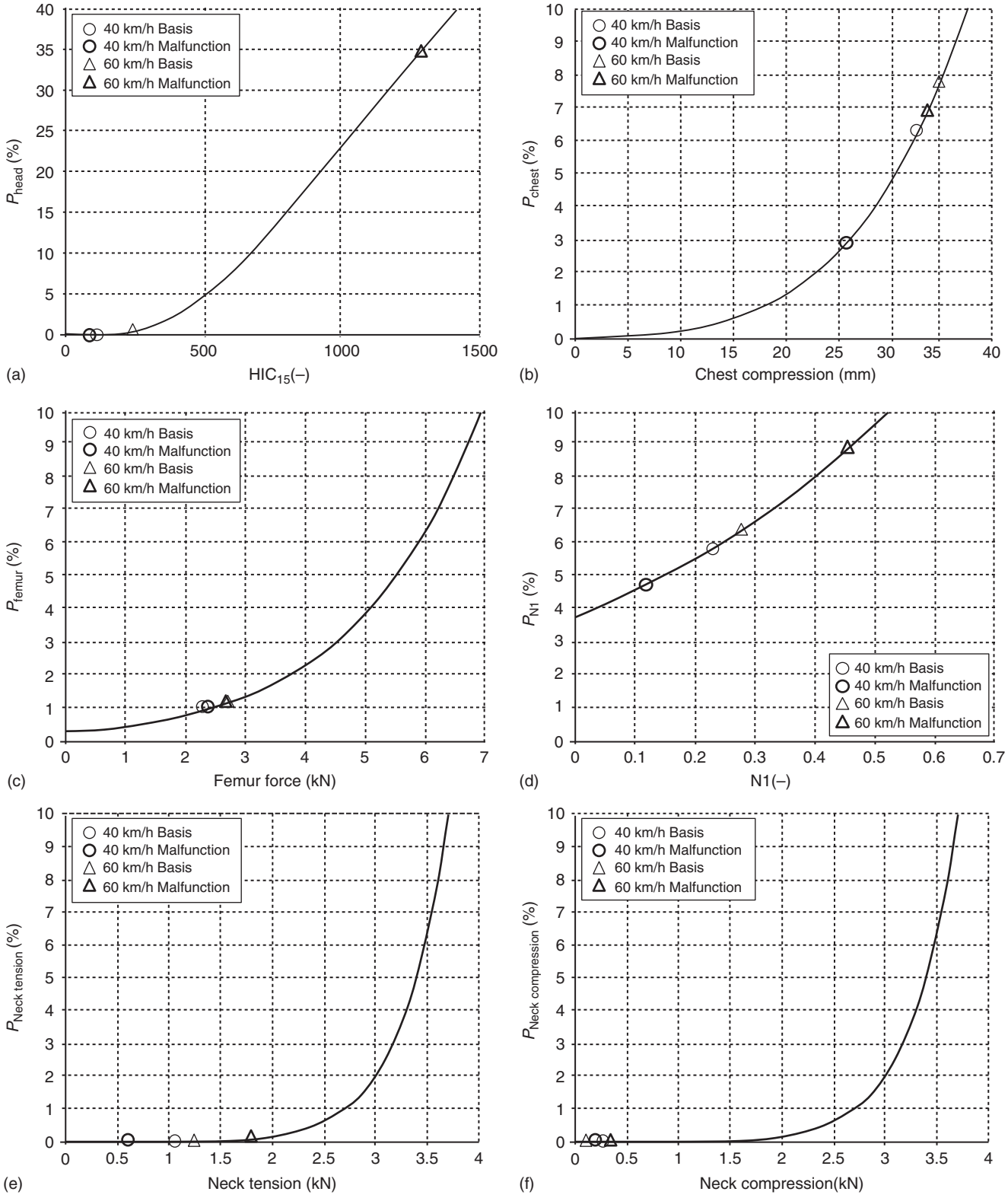


Figure 5. (a–f) Probability curves of US NCAP test in case of unneeded activation.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

required, because the driver should under no circumstances be hindered in driving.

Active dynamic control of restraint systems during a crash promises high effectiveness, although further research will be needed in the areas of actuators and sensors. In particular, faulty measurements need to be intensively considered; a falsely interpreted sensor signal must not lead to increases in occupant stresses under any circumstances.

The design of safety measures should be optimized with respect to accidents and traffic scenarios as they occur in the field. To this end, simulative methods should be further developed and validated, in order to achieve rapid and reliable assessment of effectiveness.

REFERENCES

Badstübner, J. (2008) Planspiel zur Sicherheit. Automobilindustrie, März.

Berg, A., Rücker, P., and Domsch, C. (2011) *Crashtest mit einem Fahrzeug mit Pre-Crash-Funktionen und automatischer Pre-Crash-Bremsung*, Springer Fachmedien, Wiesbaden, p. 10.

Bogenrieder, R., Merz, U., Bachmann, R., *et al*. (2007) Pre-sense. Technical Report, DaimlerChrysler AG.

Dannenberg, J. (2004) Das Jahrzehnt der Assistenzsysteme. Automobilentwicklung, Juli.

Hartenstein, H., Sarma, A., Festag, A., *et al*. (2004) Fleetnet: Bringing cartocar communication into the real world. Technical Report, NEC Europe, University of Mannheim, University of Karlsruhe.

Heudorfer, B. and Meissner, D. (2008) *Aktiver Eingriff in passive Systeme: Von passiver Sicherheit zu sicherem Fahren*, Forschung für das Auto von Morgen, Springer, Berlin 2008, ISBN: 978-3-540-74150-3, pp. 215–238.

Jerzembek, M. (2008) Siebter Sinn. Automobilindustrie, Juli.

Kompass, K. (2008) *Fahrerassistenzsysteme der Zukunft - auf dem Weg zum autonomen Pkw?*, Forschung für das Auto von Morgen, Springer, Berlin 2008, ISBN: 978-3-540-74150-3, pp. 261–284.

Lemmen, P., Couper, G., Neale, M., *et al*. (2005) Development and evaluation of smart restraints: Ec prism. Technical Report, PRISM.

Herrtwich, R.G. (2008) Daimler setzt auf car2car-Vernetzung. Automobilwirtschaft, März.

Hesseling, R.J. (2004) Active restraint systems - feedback control of occupant motion. PhD Thesis.

Schittenhelm, H. (2009) The Vision of Accident Free Driving - How Efficient are We Actually in Avoiding or Mitigating Longitudinal Real World Accidents. *In Proceedings of International Technical Conference on Enhanced safety vehicles*.

Whydell, A. (2010) Assistance for all. Vision Zero International, January.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto259
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Pedestrian Protection Overview

Carlos Arregui-Dalmases[1,2], Rikard Fredriksson[3], Jason R. Kerrigan[2], Joan Velazquez-Ameijide[1], and David Sanchez-Molina[1]

[1] *Universitat Politècnica de Catalunya - Barcelona Tech, Barcelona, Spain*
[2] *University of Virginia, Charlottesville, VA, USA*
[3] *Autoliv Research, Vårgårda, Sweden*

1 Introduction to Traffic Injuries 1
2 Epidemiology and Characterization of Pedestrian Accidents 1
3 The Pedestrian, the Environment, and the Vehicle 2
4 The vehicle–Pedestrian Collision, Influencing Factors 2
5 Distribution of Pedestrians' Injuries 8
6 Vehicle Structures Responsible for Pedestrian Injuries 9
7 The Consequences of the Second Impact 9
8 Injury Biomechanics 10
9 Pedestrian Tests and Research Tools 13
10 Current Strategies used by the Industry to Protect the Pedestrian 14
11 Requirements Incompatible with Pedestrian Protection in New Design Vehicles 16
12 Active Safety Systems 17
References 20
Further Reading 23

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

1 INTRODUCTION TO TRAFFIC INJURIES

In recent years, society has gradually become aware of this epidemic problem. Administrations have reacted to this by incorporating requirements for the certification of vehicles, demanding the use of restraint systems, indicating the way vehicles should move on their highways, and educating their citizens. Other entities such as automobile and suppliers companies have prioritized safety as the main requirement in every new model design. All these developments have led to greater protection inside the vehicle during the accident, leaving the behavior of the vehicle for the rest of entities that coexist with it. This will become, among others, the bone of contention in the near future:

- Vehicle compatibility: the way the vehicle behaves compared to other vehicles during the event of impact, without pursuing the protection of the occupants whatever the price.
- Pedestrian protection: There is a need to study the problem of pedestrian casualties and fatalities and how can they be mitigated.

2 EPIDEMIOLOGY AND CHARACTERIZATION OF PEDESTRIAN ACCIDENTS

The World Bank estimates that between 41% and 75% of all road traffic fatalities worldwide are pedestrians (World

Bank Group, 2006) and that nearly 35% of all pedestrian fatalities are children (World Bank Group, 2002). According to Japanese data, pedestrians are four times more likely to be killed in a traffic accident than vehicle occupants (ITARDA, 2004). Like most of the approximately 1.2 million road traffic fatalities occurring each year, the majority of all pedestrian fatalities occur in developing nations where traffic volume exceeds the infrastructural capacity and the latest safety countermeasures are not yet available (World Bank Group, 2002). Pedestrian fatalities occupy a lower, yet still substantial, percentage of road traffic fatalities in industrialized nations: 11% in the United States, 12% in France, 14% in Australia, 21% in the United Kingdom, 32% in Japan, and nearly 50% in South Korea (NHTSA, 2006; ATSB, 2007; CARE, 2007; NPA, 2006; Youn *et al*., 2005). While pedestrian fatalities are a considerable problem, the number of pedestrians surviving vehicle impacts with injuries far exceeds the number of fatalities. In the United States (US), for instance, there are more than 13 injured pedestrians that survive vehicle impacts for every pedestrian fatality (NHTSA, 2006).



Figure 1. Example of scenario where a dangerous situation can emerge. (Photo reproduced by permission of Carlos Arregui.)



Figure 2. Example of vehicle designed without pedestrian protection requirements. (Photo reproduced by permission of Carlos Arregui.)

3 THE PEDESTRIAN, THE ENVIRONMENT, AND THE VEHICLE

Three factors are important to fully understand a vehicle–pedestrian collision. The environment, the vehicle, and the pedestrian contribute to accidents and determine their outcome (Haddon, 1980). Further, the accident can be divided into three parts on a time-scale; pre-, in-, and post-crashes. In each of these time events, the environment, vehicle, and road user are more or less influential factors important to consider.

The pre-crash phase describes the sequence leading up to the accident. In this phase, the road design, rain, lighting conditions, vehicle condition, the pedestrian, and driver behavior all interact, and if one or more of these parameters are faulty or unfavorable, a dangerous situation can emerge. If the risk parameters are not minimized, the dangerous situation can lead to an accident. Examples of poor road design could be sight obstructions or lack of safe pedestrian crossings (Figure 1). Poor vehicle brakes and lack of stability control (ESC; electronic stability control) are examples of vehicle factors. Distraction and alcohol intoxication of the driver or pedestrian are road user factors that can contribute to risky situations. When the accident is unavoidable, the vehicle impacts the pedestrian and the accident proceeds to the in-crash phase. During this phase, vehicle design and speed are examples of vehicle-influencing factors (Figure 2), whereas pedestrian vulnerability is an influencing factor for the road user and surface rigidity for the road. Finally, in the post-crash phase, rapid, emergency care can influence the outcome of pedestrian injuries.

4 THE VEHICLE–PEDESTRIAN COLLISION, INFLUENCING FACTORS

4.1 Pedestrians' age distribution

Two differentiated groups monopolize most of the leading role in pedestrian accidents: those younger than 12 years old and those older than 65.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

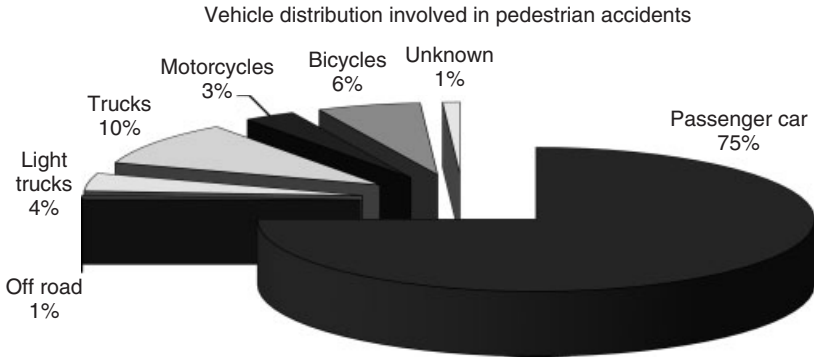


Figure 3. Vehicle type distribution, based on GIDAS database.

Some of the causes that make children a group of risk are (Montoro *et al*., 2000)

- limited visual field due to their low stature
- problems of outlying vision, reduced field of vision
- audition deficiencies (shown up to the age of 7)
- lack of attention
- ignorance of rules and risks.

Some of the causes that make elderly people a group of risk are (Montoro *et al*., 2000)

- sensory problems
- smaller attention to the environment
- muscular and bony degeneration.

4.2 Vehicles involved in pedestrian accidents

Passenger cars are responsible for 70–80% of all pedestrian accidents in developed countries, varying this percentage depending on the world area studied, followed by trucks, slight trucks, and pick-ups, being responsible for 85–90% of all pedestrian with injuries. Figure 3 presents the distribution of types of vehicles with an injury of AIS 1+.

Currently, the automotive industry and the administration focus their efforts on passenger cars, with little or no research carried out by other vehicle types.

4.3 Speed of collision

The speed between vehicle and pedestrian is one of the main causes responsible for the seriousness of the pedestrian injuries. Figure 4 presents the impact speed

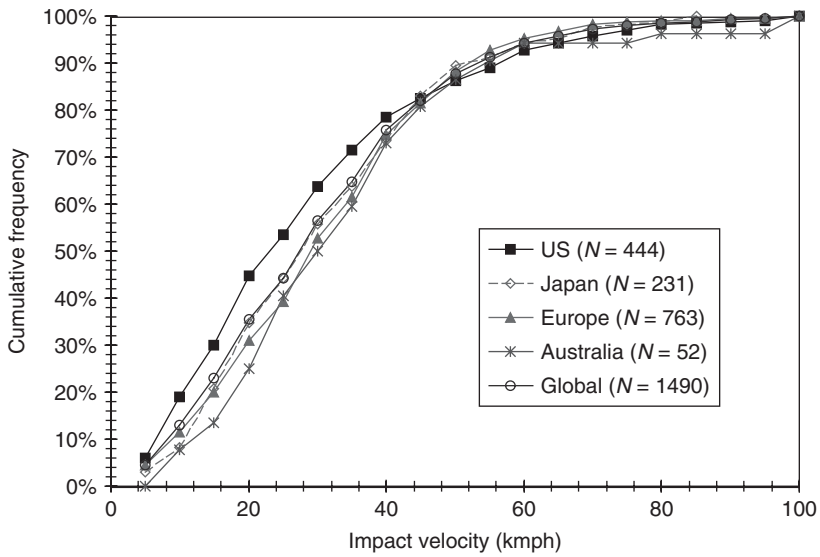


Figure 4. Impact velocity distribution. (Based on Mizuno, 2002.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

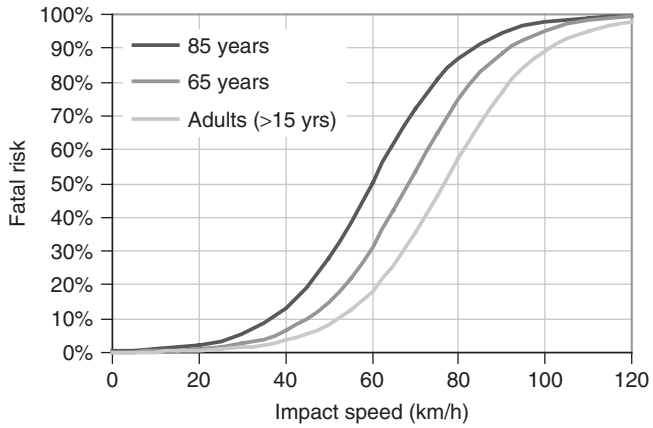


Figure 5. Fatality risk as a function of impact speed for different ages of pedestrians. (Based on data from Rosén and Sander (2009).)

and the accumulative frequency, with data collected from USA, the European Union, Japan, and Australia. For all of them (maybe with the exception of USA, where the frequency is slightly higher), it can be noticed that approximately 70–75% of pedestrian accidents take place up to 40 km/h.

Rosén and Sander (2009) presented an injury risk curve describing the relationship between vehicle impact speed and risk for fatal outcome for the pedestrian in a vehicle frontal impact. This was based on 490 accidents representative to Germany. The study showed that the risk for fatal outcome in a 50 km/h impact was twice as high as an impact at 40 km/h and four times higher than an impact at 30 km/h, indicating the importance of impact speed to determine outcome (Figure 5).

Older pedestrians are over-represented in severe and fatal pedestrian crashes with a higher injury and fatality risk (Henary, Ivarsson, and Crandall, 2006; Loo and Tsui, 2009; Rosén and Sander, 2009). Henary *et al*. found in 552 US vehicle-to-pedestrian accidents that pedestrians 60 years or older had an almost threefold higher mortality rate compared to adults 19–50 years old. Loo and Tsui found, in a study of 4290 accidents in Hong Kong, a 3.6 times higher mortality rate for pedestrians 65 years or older compared to 15–64-year-old pedestrians. Rosén and Sander (2009) presented a pedestrian injury risk function where they concluded that age and speed were the two most important parameters for risk of fatal outcome. The risk function was used to extract risk functions for different ages, compared to average adults, in figure. Males are reported as more frequently involved in pedestrian crashes, but no gender difference has been found for the fatality risk (Rosén and Sander, 2009; Zhang *et al*., 2008).

4.4 Kinematics of pedestrians: the injury pattern

Before analyzing the distribution of injuries suffered by a pedestrian, it is important to point out what can be understood by the term *pedestrian accident*. Obviously, all accidents are different and all external elements add more entropy to understand an already turbulent scenario. Nevertheless, there are common elements or patterns.

Figure 6 shows the occurrence of the impact direction on a pedestrian. It can be concluded from them that the most probable direction is the lateral. This fact is extremely important in understanding what impactors are used as a tools for developing cars with pedestrian protection countermeasures in a crash laboratory.

4.4.1 The sequence of a pedestrian accident (adult case)

In the event of a pedestrian–vehicle collision, the first contact takes place between the bumper and the lower extremity. Later on and as a consequence of the impact produced below the pedestrian's center of gravity because of the speed of the vehicle, a wrap around the body on the vehicle takes place. The second contact occurs between the front part of the vehicle and the area of the thigh and

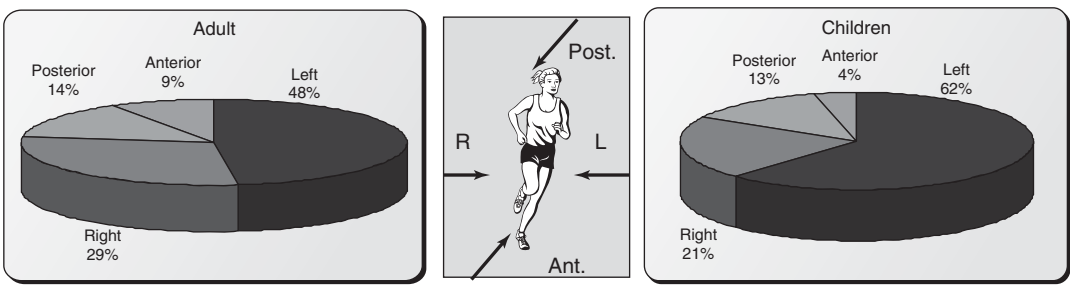


Figure 6. Impact direction occurrence. (Based on data in Sakurai *et al*., 1994.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

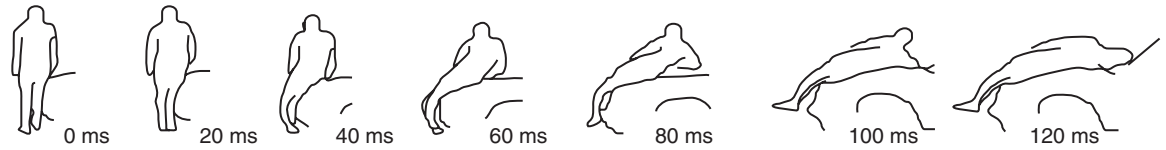


Figure 7. Adult pedestrian kinematics.

hip. The third impact corresponds to the pedestrian's torso and the vehicle's hood, and finally, the head hits against the hood or windshield, depending on several variables (Figure 7).

Once the collision against the automobile is over, the possibility of increasing the pedestrian's injuries does not disappear, which is known as the *second impact*. This second impact involves the pedestrian's collision against the road or the elements found in the scenario, and it is interesting to point out its great randomness and the unforeseeable nature of the injuries suffered. Generally, it is not possible to dissociate absolutely the lesions suffered at the first impact from those sustained at the second impact.

Likewise, pedestrian kinematics will be affected by the geometry of the vehicle, the vehicle speed, the pedestrian's anthropometric measures, the possibility of braking, the direction of the impact, the internal structure of the front part of the vehicle, and so on.

For instance, in Figure 8, different types of kinematics are represented according to the vehicle's geometry.

If a deeper investigation needs to be performed, accident reconstruction or finite element model analysis is not the right research tool; in that case, pedestrian test

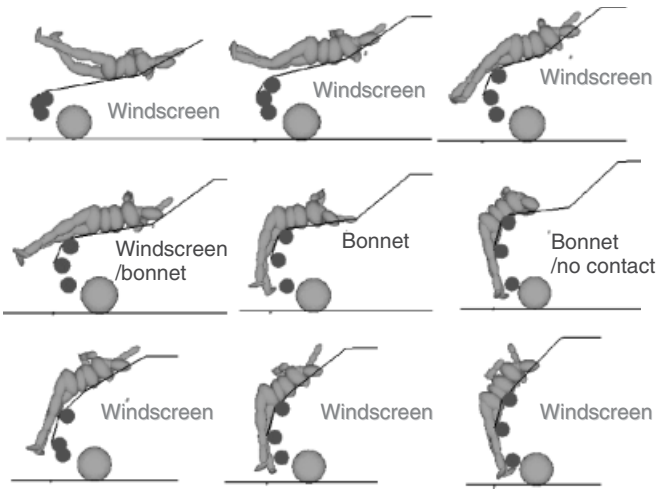


Figure 8. Pedestrian–vehicle different kinematics depending on vehicle profile. (From Mizuno, 2002. Reproduced by permission of Y. Mizuno.)

should be conducted in suitable laboratories. Kerrigan *et al*. performed 12 PMHS (postmortem human subject) tests, with three sedan-type cars and small to tall pedestrians (154–187 cm), the wrap around distance (WAD) to head impact was between 60 and 540 mm greater than the pedestrian stature in each test (Kerrigan, Arregui, and Crandall, 2009; Kerrigan, Crandall, and Deng, 2007; Subit *et al*., 2008).

A pedestrian body versus hood-sliding effect is visible in pedestrian tests with sedan-type cars. The pelvis slides up onto the hood surface after the impact of the thigh to the hood leading edge. This is related to sliding motions of similar distances. In similar tests with sports utility vehicles (SUVs), with a higher hood leading edge, the sliding effect was less pronounced, with an 85–90 mm difference in WAD to stature (Kerrigan, Arregui, and Crandall, 2009). This indicates a higher WAD-to-head impact in collisions with sedan-type cars compared to collisions at same impact speed with vehicles with higher front ends such as SUVs, or in vehicles with a more vertically inclined hood surface as in small compact cars.

High speed video images from a PMHS are given in Figure 9 for sedan-type vehicle and Figure 10 for SUV vehicle.

Accident data was used to investigate this sliding effect. Fredriksson and Rosén (2010) used German (GIDAS) accident data to derive a head impact WAD equation, where WAD depended on pedestrian stature and car impact speed. Using their equation to calculate the sliding effect (head impact WAD—body height) for three body heights, the following dependence on impact speed could be derived (Figure 11).

Head impact speed relative to the car can be both higher and lower than the initial car impact speed. Kerrigan *et al*. reported, in 10 PMHS tests with sedan-type cars, head impact velocities ranging from 68% to 130% of the car impact speed (Kerrigan, Arregui, and Crandall, 2009; Kerrigan, Crandall, and Deng, 2008). There seems to be a trend that a higher impact velocity ratio is recorded when the head impact is to the windshield compared to the hood.

Head impact times are dependent on car type, pedestrian stature, and impact speed. Kerrigan *et al*. and Subit

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

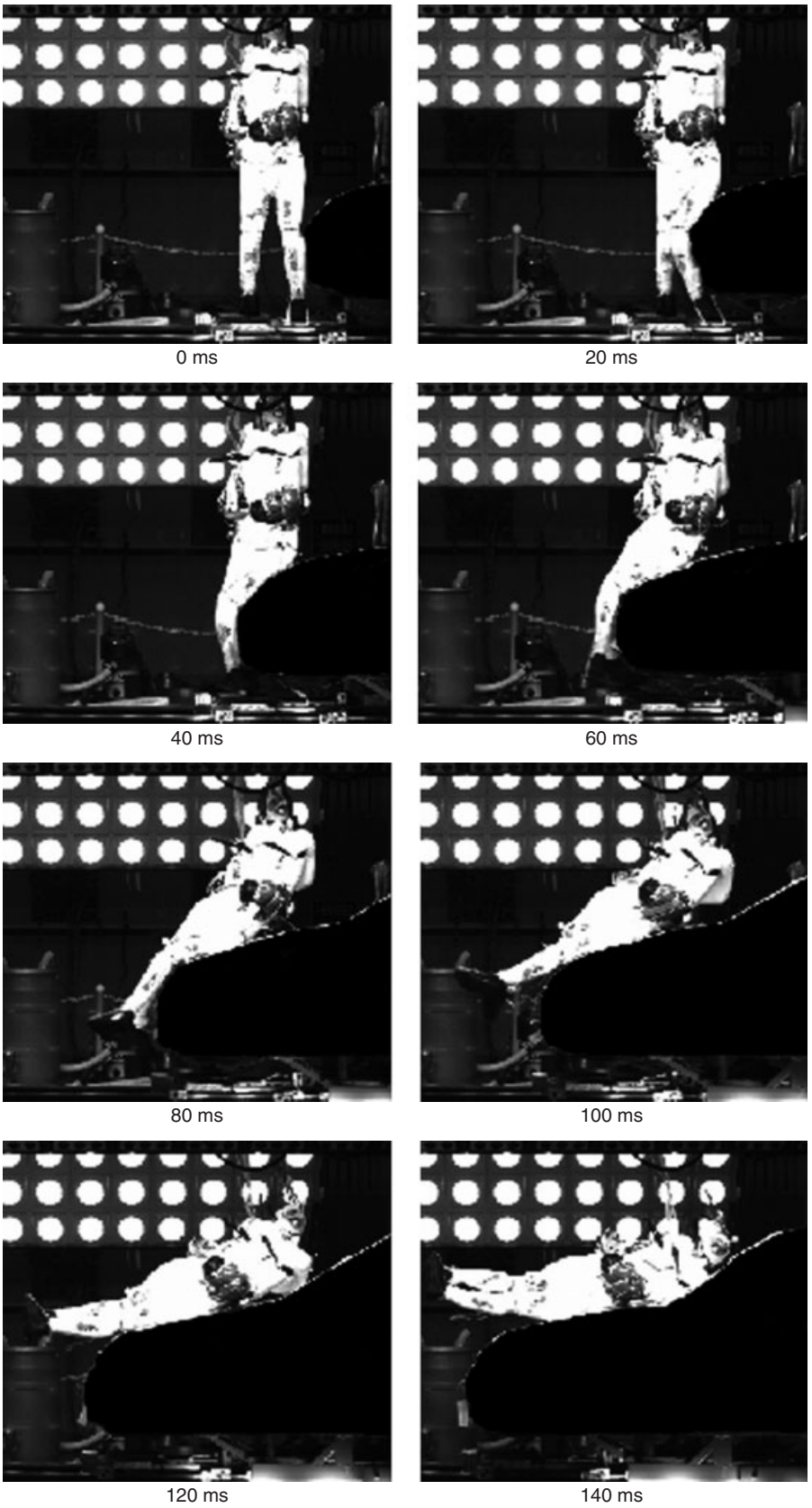


Figure 9. High speed video images for sedan test. (Reproduced with permission from Kerrigan *et al*., 2005a,b. © J. Kerrigan.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

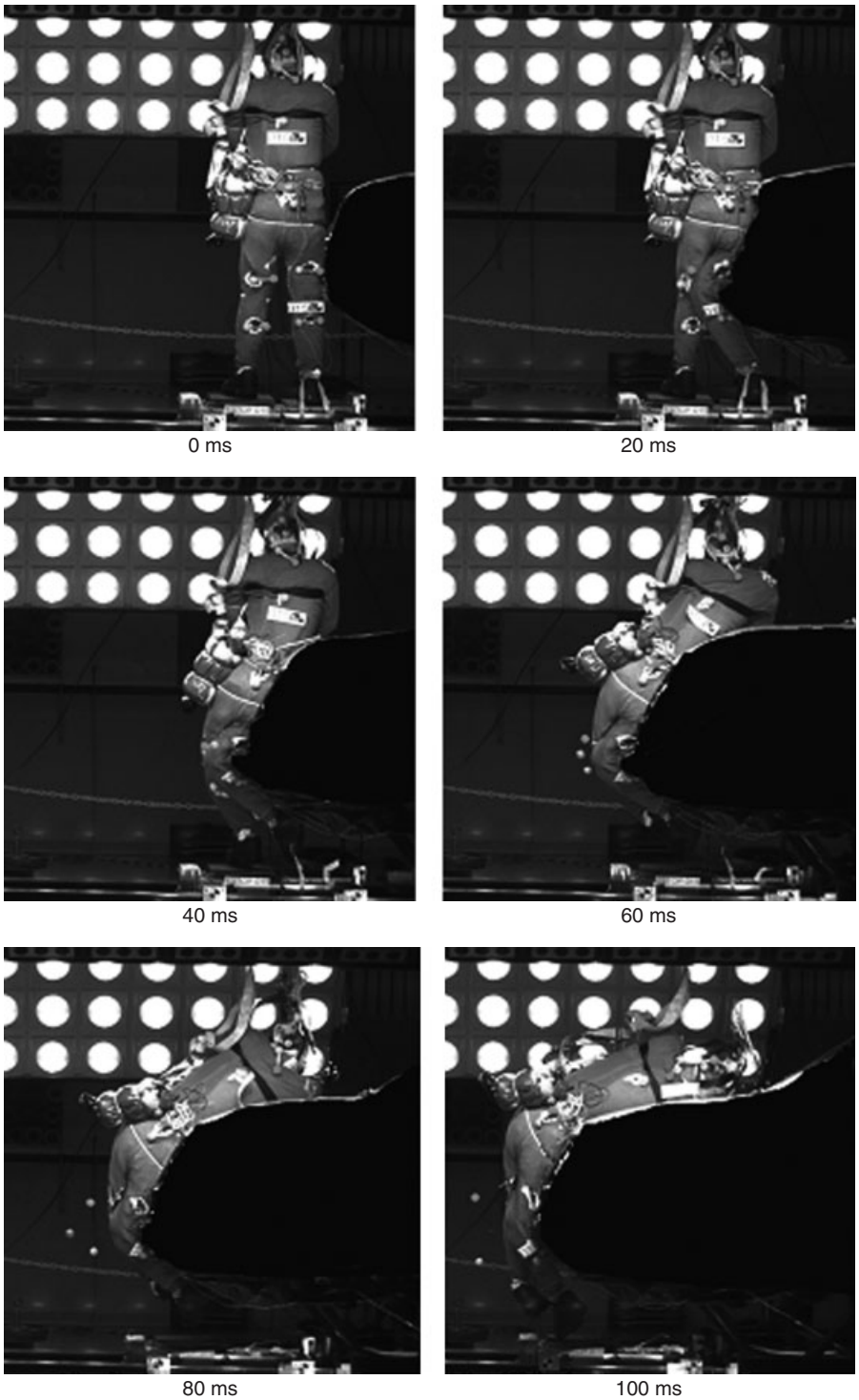


Figure 10. High speed video images for SUV test. (Reproduced with permission from Kerrigan *et al*., 2005. © J. Kerrigan.)

et al. reported, in their 12 tests with sedan-type cars at 40 km/h, head impact times ranging from 107 to 151 ms from first car-to-pedestrian impact, with the shorter times for shorter pedestrians. Four SUV tests have been performed with differing stature adult PMHS to measure head impact time (Kerrigan, Arregui, and Crandall, 2009; Schroeder, Fukuyama, and Yamazaki, 2008). The head impact times ranged from 90 to 116 ms. Subit *et al*. (2008)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

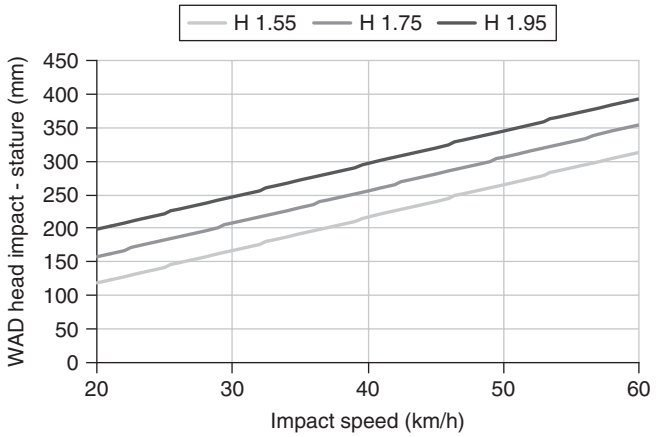


Figure 11. Difference (in mm) between head impact WAD and body height (stature) as a function of car impact speed ("sliding") for three pedestrian body heights. (Based on accident data from Fredriksson and Rosén (2010).)

performed two tests with a small compact car and reported impact times of 91–94 ms for one short and one tall adult PMHS.

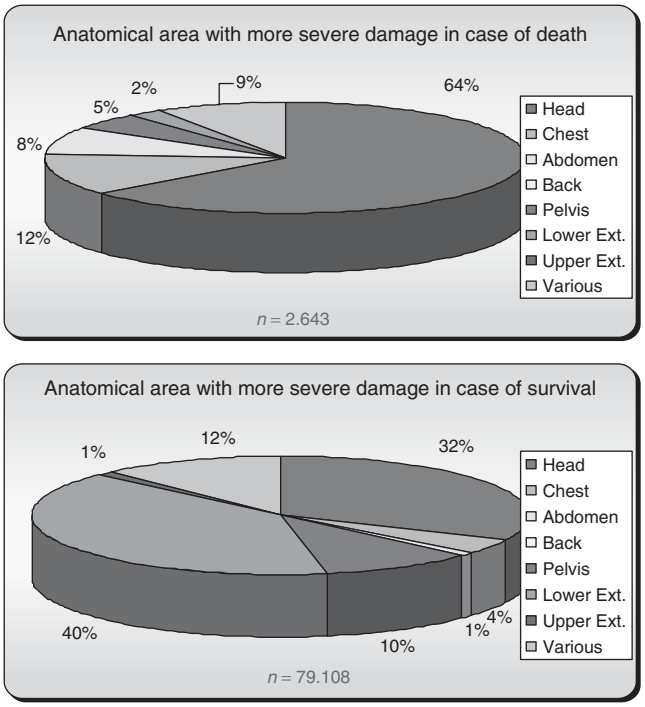


Figure 12. Pedestrian accidents divided into fatalities and survivals. (Based on Matsui, 1998.)

5 DISTRIBUTION OF PEDESTRIANS' INJURIES

In case of a pedestrian–vehicle collision, almost every anatomical area is exposed to sustain injuries. Table 1 shows the injury distribution according to anatomical areas. It can be observed that head and lower extremities are the most frequently injured anatomical areas. Another remarkable aspect is the similarity of data collected in USA, the European Union, and Japan.

Nevertheless, more precise information can be obtained by analyzing this type of data in detail. For instance, if pedestrian accidents are divided into fatalities and survivals and represented graphically (Figure 12), it can be appreciated that in the event of fatality, the head is the most injured structure, followed far away by chest and abdomen, and legs have a minor representation. In case of survival, the scenario is completely different; legs are mainly represented as the most injured area followed by the head. It is important to point out that in spite of the little influence of the lower extremities for survival, injury to this region can have serious and lasting effects on the quality of life.

Table 1. Injury distribution related to anatomical areas.

(For AIS 2–6) Injury Location	$n = 1342$ US (%)	$n = 1140$ Europe (%)	$n = 360$ Japan (%)	$n = 163$ Australia (%)	$n = 3305$ Global (%)
Head	32.7	29.8	28.6	39.0	31.3
Face	3.7	5.3	2.4	4.4	4.3
Neck	0.0	1.8	4.5	0.5	1.3
Chest	9.5	11.6	8.5	9.3	10.2
Abdomen	7.7	3.8	4.8	6.0	5.6
Pelvis	5.3	7.9	4.5	4.4	6.3
Arms	7.9	8.1	9.0	8.8	8.1
Legs	33.3	31.3	35.7	27.5	32.4
Unknown	0.0	0.5	2.1	0.0	0.5
TOTAL	100	100	100	100	100

Based on of IHRA/PS accident data 2002.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

The fact that the most damaged structures are head and lower extremities explains the policy dictated by the WG 10–17 (groups of work that study pedestrian protection) of EEVC (European Enhanced Vehicle-Safety Committee), and consequently focused on head, femur, and lower extremities impactors.

6 VEHICLE STRUCTURES RESPONSIBLE FOR PEDESTRIAN INJURIES

One of the goals of accident research consists of being able to discern what elements of the automobile have caused certain lesions to the pedestrian. From Table 2 for pedestrians older than 15, with a damage of AIS 2+, it can be seen that the area of the vehicle that most frequently injures the head is the windshield, the thorax is damaged by the rear part of the hood, the abdomen and the pelvis by the line of the hood, the femur practically in the same way by the line of the hood and the bumper, and finally the knee and the low leg are damaged mostly by the bumper.

Considering the previous table for an age group up to 15 years old for a damage equal to AIS2+ (Table 3) can be concluded that the hood is the structure that most frequently damages the head, chest, and pelvis, the line of the hood damages the abdomen, and the bumper is responsible for most of the damages in the femur, knee, and lower extremities (tibia, fibula).

The explanation of these results lies in the difference of height between children and adults, giving a different impacted area. Children are more frequently damaged by the front part and the hood in its frontal area, as for adults the front area injures the lower extremities and pelvis, the head being impacted against more rearward structures by the simple fact of the body wrap around. Figure 13 reproduces the main car structures involved in a pedestrian collision.

7 THE CONSEQUENCES OF THE SECOND IMPACT

The second impact can be defined as the impact suffered by a pedestrian once he or she leaves contact with the vehicle. This definition brings some problems, especially one found in literature and known as *run over* (pedestrian accident in which a large vehicle, such as SUV, or pick-up is involved). In this scenario, the pedestrian's contact surface contains his or her center of gravity, so that he or she is projected frontward. Later on or at the same time of the second impact, a run over can take place, producing visceral squashing and pneumatic tattoos.

D. Otte studied in 2001 a total of 293 pedestrians throughout more than 20 years of accident research

Table 2. Vehicle structures responsible for adult pedestrian injuries.

	Anatomical Area ($n = 2773$)						
Aggressor element	Head	Chest	Abdomen	Pelvis	Femur	Knee	Inferior Ext.
Bumper	20	2	3	3	29	69	429
Frontal	—	8	13	6	9	10	32
Bonnet line/wing	7	36	65	80	33	5	24
Upper bonnet/wing	140	122	39	35	3	1	1
Windshield	303	28	3	10	—	—	—
Roof beam/A pillar	159	34	7	14	1	—	—
2nd impact	125	21	2	8	4	3	5

Based on IHRA/PS Accident Data 2002.

Table 3. Vehicle structures responsible for child pedestrian injuries.

	Anatomical Area ($n = 532$)						
Aggressor element	Head	Chest	Abdomen	Pelvis	Femur	Knee	Inferior Ext.
Bumper	4	1	2	—	30	7	47
Frontal	5	1	—	1	5	1	3
Bonnet line/wing	8	7	13	5	7	1	6
Upper bonnet/wing	83	17	5	8	—	—	—
Windshield	41	2	2	—	—	—	—
Roof beam/A Pillar	9	1	—	—	—	—	—
2nd impact	46	1	—	1	—	—	—

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

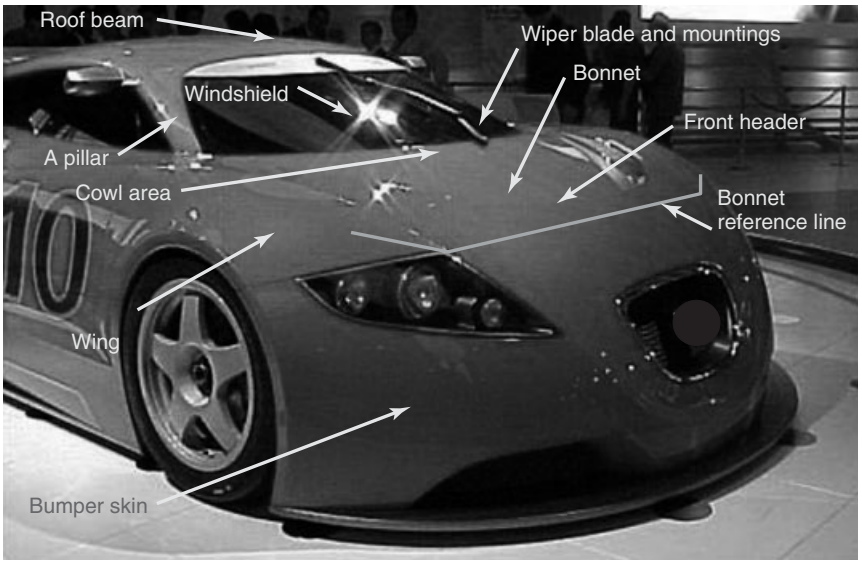


Figure 13. Automobile main structures involved in the pedestrian–vehicle collisions.

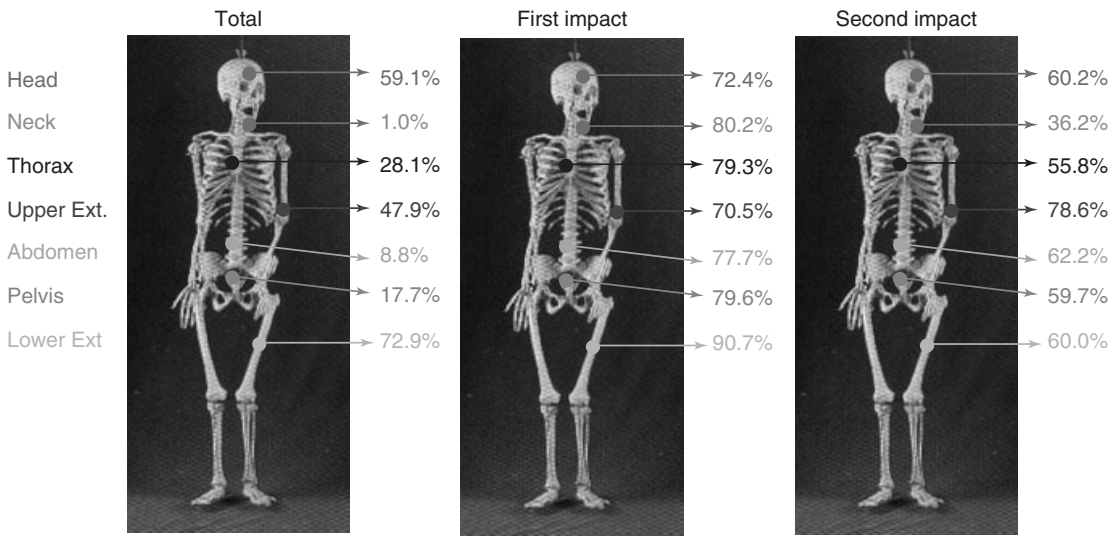


Figure 14. Injury frequency for pedestrians distinguishing between primary and secondary impacts. (Based on data in Otte and Tohlemann, 2001.)

undertaken at the Medical University of Hanover in Germany. Figure 14 reproduces the frequencies for $n = 293$ pedestrians distinguishing between primary and secondary impacts and indicating 100% of each injured region.

This figure shows that most of the frequencies are superior in the first impact, with the only exception of upper extremities. Furthermore, the existence of a first impact does not avoid the second one from happening. Similarly, the study determines injury severity, as shown in Figure 15, and shows that injuries are more serious during the first impact.

The conclusion is that it is essential and necessary to pay attention to the design of the frontal part of the vehicle, and also performing more research on the second impact.

8 INJURY BIOMECHANICS

8.1 Pedestrian head injury

Many factors should be considered in order to understand properly how a mechanical input to the head can result in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

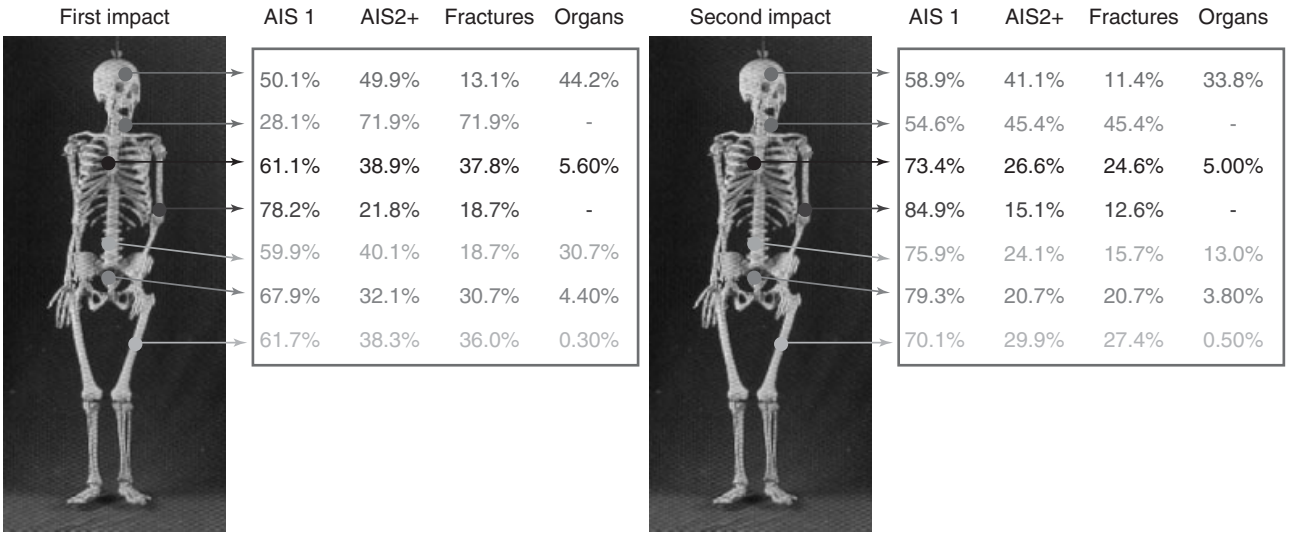


Figure 15. Injury categorization for pedestrians distinguishing between primary and secondary impacts. (Based on data in Otte and Tohlemann, 2001.)

a determined type of head injury: the severity, the nature of the mechanical input, the impact location, the direction of this input, the age of the patient, his or her gender, anthropometrics, and previous state, and also the treatment and recovery of the patient.

Head injuries are either the most or second most commonly reported injuries to pedestrians struck by vehicles (Mizuno, 2003; Zhao *et al*., 2010). Furthermore, among serious or life-threatening injuries, head and brain injuries far outnumber injuries to all other body regions (Fildes *et al*., 2004). Previous studies have shown that head and neck injuries sustained by pedestrians account for almost 60% of all harm to pedestrians (Fildes *et al*., 2004).

As explained previously, in an effort to mitigate the risk of head (and other) injuries to pedestrians, researchers have developed tools, such as pedestrian dummies and computational models, to further understand the dynamics of vehicle–pedestrian impact (Untaroiu *et al*., 2008). While the local stiffness of the individual vehicle structures involved in head-to-vehicle impact is a primary concern in decreasing the risk of head injury, impact simulations with pedestrian dummies and computational models allow for examination of other factors that affect head injury risks. For instance, the magnitude of the accelerations sustained by the head in head-to-vehicle impacts is dictated not only by the vehicle stiffness but also by the impact velocity and impact angle, which dictate the magnitude and duration of the impact forces applied to the head. In addition, Okamoto and Kikuchi (2006), in a study that involved vehicle-pedestrian impacts with the Polar II

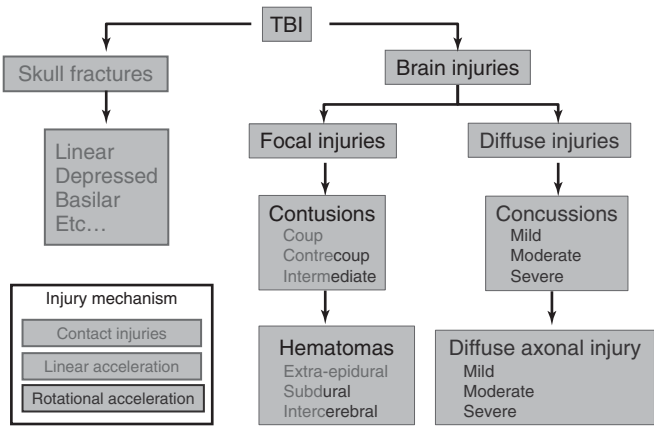


Figure 16. Head injuries related to injury mechanisms. (Based on Gennarelli, Spielman, and Langfitt, 1982, 2002.)

pedestrian dummy, used the dummy's neck instrumentation to explore the magnitudes of forces applied to the head through the neck during impact. As their goal was to compare pedestrian dummy impacts to those of headform impactors, Okamoto and Kikuchi used neck forces to examine similarities and differences between the dummy and the impactor, without examining how neck forces directly affect impact kinematics and estimates of injury risk. Figure 16 summarizes the main head injuries and their injury mechanisms.

Contact injuries are related to a direct impact between the head and the vehicle, linear and rotational accelerations are due to the head impact and/or the head kinematics during all the event of a pedestrian–vehicle collision.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Twenty-three percent of all head injuries analyzed have the translation acceleration as a single injury mechanism; in 40% of all cases, rotational acceleration is the only injury mechanism and in 37% of the cases, rotational or translational acceleration can be the injury mechanism for the sustained head injury (Arregui-Dalmases, 2006).

8.2 Pedestrian lower extremity injury

By combining both the frequency and the severity of pedestrian lower extremity injuries, pedestrian lower extremity injury mitigation priorities could be described as

- diaphysis fractures of the tibia and fibula,
- soft tissue injuries and fractures of the knee joint,
- femoral diaphysis fractures,
- malleolar (ankle) fractures.

As injury tolerance is a function of loading mode, the mechanisms associated with these lower extremity injuries must be considered if injury countermeasures are to be designed specifically to prevent these injuries.

Previously it was showed that the majority of pedestrians were struck laterally by the vehicle and the majority (65%) of pedestrians were in mid-stance gait (one foot in front of the other) when struck (Kam *et al*., 2005). Accordingly, most pedestrian lower extremity injury mechanisms are derived from direct loading to the lateral or medial side of the lower extremity (Figure 17).

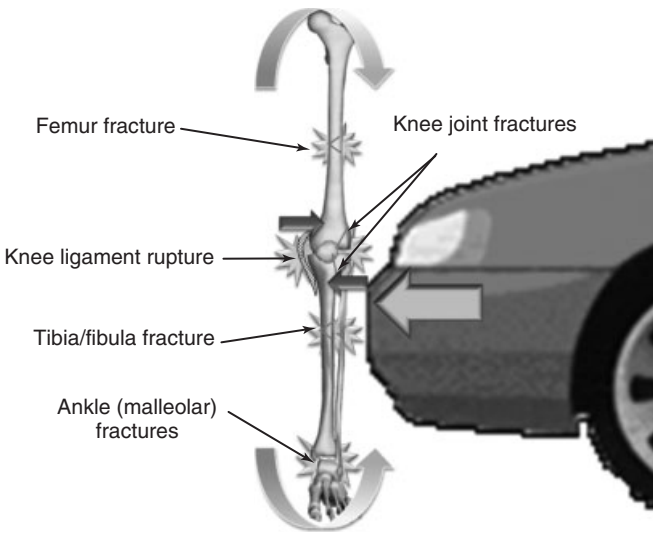


Figure 17. Schematic diagram predicting lower extremity injuries as a result of vehicle front-end loading causing bending and shear in the lower extremity components.

Direct loading of the vehicle bumper in the area of the knee joint can cause the knee to experience bending loads and shearing loads if the vehicle's geometry is such that the vehicle loads the proximal leg without loading the distal thigh, or vice versa (Yang, 2005). In addition, even when the knee joint is subjected to pure (four-point) bending, shear displacements can still occur as a result of the complex geometry of the tibial and femoral epiphyses (Bose *et al*., 2008). Thus, lateral shearing and bending have been recognized as the two most important knee injury mechanisms (Yang, 2005). Yang (2005) and many others have explained that lower extremity long bone (diaphyseal) fractures, as well as both fracture and soft tissue injuries to the knee joint, are the result of lateral bending of the entire lower extremity owing to vehicle bumper and hood leading edge loading of the lateral aspect of the lower extremity (Figure 17). Biomechanics experiments using PMHS lower extremities loaded laterally by vehicles or vehicle-like structures have produced

- long bone fractures (Pritz *et al*., 1975; Cesari *et al*., 1980; Bunketorp *et al*., 1981; Ashton, Cesari, and Van Wijk, 1983; Bunketorp *et al*., 1983; Kallieris and Schmidt, 1988; Schroeder *et al*., 2000; Snedeker *et al*., 2005; Kerrigan, Crandall, and Deng, 2008),
- knee joint fractures (Pritz *et al*., 1975; Cesari *et al*., 1980; Bunketorp *et al*., 1981; Ashton, Cesari, and Van Wijk, 1983; Bunketorp *et al*., 1983; Kallieris and Schmidt, 1988; Untaroiu *et al*., 2007; Kerrigan, Crandall, and Deng, 2008)
- soft tissue injuries (Cesari *et al*., 1980; Bunketorp *et al*., 1981, 1983; Schroeder *et al*., 2000; Untaroiu *et al*., 2007; Kerrigan, Crandall, and Deng, 2008), and
- ankle injuries (Cesari *et al*., 1980; Bunketorp *et al*., 1981, 1983; Untaroiu *et al*., 2007; Kerrigan, Crandall, and Deng, 2008).

The majority of pedestrians are struck laterally, whereas their lower extremities are positioned in gait-like stance, which dictates the mechanisms of the most common injuries. Diaphyseal fractures of the long bones are expected to occur as the result of lateral bending loads applied to the lower extremity. Knee joint injuries, including soft tissue injuries and epiphyseal fractures, are expected to occur as a result of either pure bending (valgus on the struck-side limb or varus on the contralateral limb) or a combination of bending and lateral-medial shear loading. Finally, ankle fractures are expected to occur as a result of either eversion (struck-side limb) or inversion (contralateral limb).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

9 PEDESTRIAN TESTS AND RESEARCH TOOLS

9.1 Regulation and consumer test, pedestrian impactors

The striving of car manufacturers to perform well in car safety consumer tests has led to a rapid development of safety systems and improved safety for car occupants during the 2000s.

In 2005, the first legal requirements for pedestrian protection were introduced in both Europe and Japan. The EEVC WG 17 impactors and test methods were adopted by the European directive (EC, 2003). The lower legform is launched horizontally toward the vehicle bumper at 40 km/h (Figure 18). Requirements were set for tibia acceleration as well as knee bending and shearing. For higher front-end vehicles, the upper legform is used in a vertical orientation to assess bumper performance, using requirements of force and bending moment. The upper legform for the front hood edge is used only for monitoring purposes where the impactor mass, impact angle, and speed are dependent on vehicle geometry. The child and adult headform tests are performed at the front and rear sections of the hood area, respectively, determined by the WAD but limited to the hood area. Injury is assessed by the head injury criterion HIC_{15}. The headform impact speed in the directive was reduced to 35 km/h, which was argued to reflect a car impact speed of 40 km/h. The WG17 based its conclusions on AIS2+ injuries and, to a limited extent, AIS3+ injuries.

The Japanese directive includes the head-to-hood tests only, using the headforms developed by ISO. The impact angles are different from the European directive and depend additionally on vehicle geometry, divided into three vehicle categories. The headform impact speed is 32 km/h, lower than the European directive.

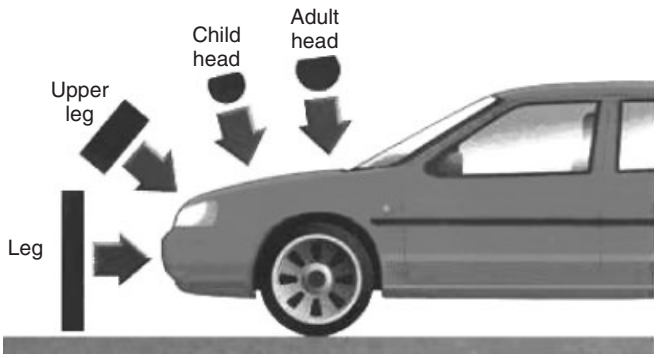


Figure 18. Pedestrian impactors, based on EEVC WG 17. (Reproduced from Euro NCAP. © Euro NCAP.)

A second phase of the European directive was introduced in 2009 (EC, 2009). It was basically harmonized with the global technical regulation (GTR). The two headforms, of 2.5 and 4.8 kg, were replaced by the ISO child headform of 3.5 kg. The injury criteria were raised to a slightly higher level. Further, a requirement of equipping the vehicle with a brake assist system, to assist the driver to brake optimally, was introduced.

In 2009, a GTR to harmonize pedestrian protection was introduced (UN, 2009). It was almost identical to the second phase of the EC regulation, with legform-to-bumper and headform-to-hood tests based on a crash speed of 40 km/h, but did not require brake assist and included the possibility of raising the upper vehicle mass limit to 4.5 tons if so decided by the individual country. The EU regulation is limited to 2.5 tons. Further, the GTR plans to adopt the FlexPLI legform to replace the WG17 legform used by the EC regulation and as an intermediate solution in the GTR. In the early phase of the GTR development, the intention was to include the windshield area, but this part of the test was removed owing to feasibility issues. It was considered unfeasible to design a pedestrian-friendly windshield frame while meeting other vehicle stability requirements. Further, it was concluded that the glass impact caused a spread in the test results for identical windshields but that this was not yet fully understood (UN, 2009). In addition, the GTR includes a test method for deployable hood systems. Part of this test method uses numerical pedestrian models, a new method to assess pedestrian protection in a regulation. The GTR was based on AIS2+ pedestrian injuries.

In 1997, the European consumer organization Euro NCAP introduced pedestrian protection assessment of the most sold vehicles in Europe. They adopted the EEVC WG17 impactors and test methods, including the lower legform, upper legform, and child and adult headforms. In contrast to legal tests, they did not limit the headform tests to the hood area, but included the windshield area to a WAD of 2100 mm. They also retained the headform test speed of 40 km/h (Euro NCAP, 2011). The pedestrian rating of the car was initially excluded from the overall rating of the car. In 2009, Euro NCAP changed their assessment protocol to include pedestrian protection in the overall rating (Euro NCAP, 2009), which has led to a rapid development of the secondary (passive) pedestrian protection of cars.

The regulations and consumer pedestrian tests have been driving forces in the introduction of pedestrian safety measures, such as pedestrian bumpers and hoods in production cars.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

9.2 Full-body dummies and models

As early as the 1980s, pedestrian-specific test devices were developed. Aldman *et al*. (1985b) developed a rotationally symmetrical pedestrian dummy (Figure 5). In the early 2000s, Autoliv and Chalmers University developed pedestrian dummies in adult and child sizes. The adult dummy, a 50th percentile adult male, was based on existing frontal and side impact dummy parts with new parts designed for the lumbar spine and knee joints (Björklund and Zheng, 2001). The child dummy, equivalent to a 6-year-old child in size and weight, was based on a Hybrid III dummy with a redesigned neck, lumbar spine, and knees (Renaud and Tapia, 2004; Renaud *et al*., 2005). Both dummies were tested at three different impact speeds and two car types and were compared to the Chalmers Madymo pedestrian model (Yang and Lövsund, 1987). The intention of these dummies was limited to study kinematics, not injury assessment. Honda and Gesac developed the Polar dummy, based on the Thor dummy (Akiyama *et al*., 1999, 2001). The Polar dummy was a more advanced pedestrian dummy, designed for both kinematic and injury assessments. The most important features were a flexible lower spine, deformable knee structures including ligaments, and a deformable tibia with properties including fracture, and the Polar II version was validated against PMHS tests (Figure 18) (Kerrigan *et al*., 2005a, 2005b). The Society of Automotive Engineers (SAE) pedestrian dummy task group developed a performance specification for an adult pedestrian dummy (SAE, 2009). The performance specification was based on PMHS tests using a mid-sized sedan and compared to the existing Polar II dummy in a report by SAE (2008). The Polar dummy is still under development to Polar III, where new properties and injury assessment are under consideration (Figure 19) (Akiyama *et al*., 2009; Okamoto, Akiyama, and Takahashi, 2009; Takahashi *et al*., 2009).

The main advantages of the pedestrian dummy are:

1. The assessment of dummy kinematics interaction with the vehicle in a global way.
2. Assessment of the EEVC subsystems.
3. Previous evaluation and comparisons with human cadavers.
4. Avoiding problems of excessive focus on the subsystems.
5. Testing with another type of vehicles.
6. Study the second impact, etc.
7. Collision identification systems for pedestrians to activate all the passive safety mechanisms incorporated in the vehicle.
8. Increase knowledge on injury biomechanics and injury causes.

10 CURRENT STRATEGIES USED BY THE INDUSTRY TO PROTECT THE PEDESTRIAN

10.1 Passive safety system, thorax and head protection

Secondary or passive safety systems have been developed for the vehicle front, focusing on the bumper, hood edge, hood, and windshield areas.

To mitigate thorax injuries to the hood edge in impacts to vehicles with higher front ends, such as SUVs, an airbag was proposed for the front hood edge (Fredriksson *et al*.,



The main advantages of the pedestrian dummy are:

1. The assessment of dummy kinematics interaction with the vehicle in a global way.
2. Assessment of the EEVC subsystems.
3. Previous evaluation and comparisons with human cadavers.
4. Avoiding problems of excessive focus on the subsystems.
5. Testing with another type of vehicles.
6. Study the second impact, etc.
7. Collision identification systems for pedestrians to activate all the passive safety mechanisms incorporated in the vehicle.
8. Increase knowledge on injury biomechanics and injury

Figure 19. Polar-II Dummy. (Reproduced from Honda. © Honda.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 20. Pop-up bonnet. (Reproduced with permission from Autoliv. © Autoliv.)

2007). The hood, wings, and wiper engines have also been passively redesigned to improve energy absorption (Belingardi, Scattina, and Gobetto, 2009; Han and Lee, 2003).

Even if the hood surface design is optimized for energy absorption, there may not be a sufficient deformation distance available to underlying parts in the engine compartment. It has been theoretically and experimentally proved that deformation distances of 60–70 mm can be sufficient to achieve HIC values below 1000 (Okamoto *et al*., 1994; Zellmer and Glaeser, 1994). A solution for this is to lift the hood in case of pedestrian impact. Active hoods, pop-up hoods, or deployable hoods are different names for the concept of lifting the hood surface, usually by actuators in the rear corners of the hood (Fredriksson, Håland, and Yang, 2001; Nagatomi *et al*., 2005). These systems are currently (2012) in production in 20 car models from 14 different car brands. They lift the rear hood part between 50 and 120 mm to enable energy absorption of the head impact preventing a second "bottoming out" impact to structures underneath the hood in the engine compartment. Fredriksson *et al*. 2009 showed, in a combined experimental and finite element study, that an under-hood distance of 100 mm reduced both skull fracture-related and brain-related injury criteria to acceptable levels in 40 km/h headform impacts. The same study with dummy tests using Polar II and a real vehicle showed a large reduction in head loading by a deployable hood system compared to a standard hood. For deployable hoods to be activated in accidents, they are connected to a sensor and an actuator, which must make the decision and perform the lifting motion within a short time period. Dummy tests and simulations have shown that the deployed hood for a standard sedan-type passenger car must be in position within less than 60 ms after the first leg impact to the front of the car at a crash speed of 40 km/h. For the lower part of the windshield and the a-pillars, airbags have been proposed to enhance head protection (Autoliv annual report, 2002, 2010; Crandall, Bhalla, and Madeley, 2002; Maki, Asai, and Kajzer, 2003). A windshield airbag typically needs 90–100 ms for the same type of car, whereas a car design with a more vertical hood and front (e.g., multi-purpose vehicles MPVs or vans) requires positioning significantly earlier of both the deployable hood and airbag. A pedestrian airbag needs more energy absorption distance than a deployable hood, owing to the limited energy absorption capability in the first phase of the airbag impact. A typical airbag thickness is 200 mm and the volume can vary between 80 and 140 L depending on the car size, roughly similar to a passenger side occupant airbag. The first airbag of this kind, called *Pedestrian Airbag Technology* (*PAT*), was entered into production in 2012 in the Volvo V40 model (Figures 20 and 21).

The following figures reproduce main current strategies used for protecting the thorax-head (Figures 22–31).

10.2 Passive Safety System, Lower Extremity Protection

In recent years, a rapid development in car bumper design has been seen. While the average car scored low in the legform-to-bumper test in 2004, most cars were rated "green" (full score) in Euro NCAP tests in 2009. The distance between the bumper and the bumper beam has been increased, the bumpers were redesigned with solutions such as thicker foam and an extra lower stiffener below the bumper to reduce loading of the knee, which typically impacts at bumper height for an average adult. Airbag solutions have also been proposed to distribute and reduce the load on the lower extremities (Pipkorn, Fredriksson, and Olsson, 2007) and headlights have been redesigned to be more energy absorbent (Lucas, 2000). Modern cars have a more aerodynamic design leading to a lower, less protruding, front hood edge with a lower risk of pelvis and thigh injuries (Figures 32–34).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) (b)

Figure 21. (a, b) Pedestrian airbag; protecting the pedestrian from the rigid elements. ((a) Reproduced with permission from Autoliv. © Autoliv and (b) Reproduced with permission from Volvo. © Volvo Car Group.)

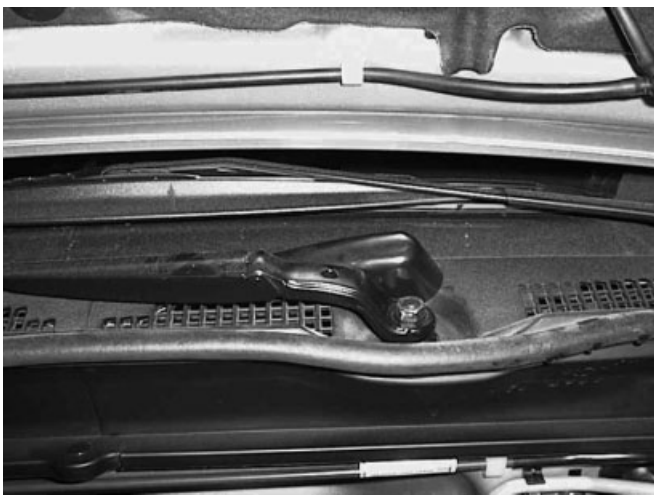


Figure 22. Avoid direct impact against rigid structures, deformable elements.



Figure 23. Deformable hinges in case of direct impact.

11 REQUIREMENTS INCOMPATIBLE WITH PEDESTRIAN PROTECTION IN NEW DESIGN VEHICLES

Unfortunately, the desire to include a new requirement in an existing product comes into conflict with of existing requirements and regulations. Among these requirements opposed to pedestrian protection the following stand out:

- Decreasing the vision field; increasing the distance between hood and engine hinders the driver's visibility, and conflicts with EC 77/649–90/630. This also increases the coefficient of friction C_x, which contributes to vehicle consumption.
- Modifying lock stiffness conflicts with FMVSS 113, 401. The lock must withstand the aerodynamic force of the vehicle and always work properly. On the other hand, in the vehicle design, there are also internal requirements such as misuse (using the vehicle in an inappropriate manner), for instance the bonnet slam test it is performed, consisting of opening and closing the bonnet 5000 times without resulting in large deformations or cracks.
- Modifying windshield stiffness conflicts with EC 92/22–2001/92. It must have enough resistance against the objects penetration. It should also break without projecting parts.
- A-pillar stiffness (EC 96/79–99/98). It must have enough rigidity for a good behavior in a frontal crash, oblique crash, and rollover. This is one of the worst places to impact against in case of pedestrian accident, and one of the most challenging to modify.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 24. Deformable elements in case of wing impact.



Figure 26. Increase distance from bonnet to still structures.



Figure 25. Deformable windshield wiper axis.



Figure 27. Engine at the back (trunk) would help to increase deformation distances.

- Internal structures weakness, hinges, and so on have to fulfill all their functions and be able to collapse during a pedestrian impact.
- Difficulty for processes, these new elements tend to have less stiffness than those of the previous vehicles. This results in added difficulties for production, handling, keeping the geometry, and so on.
- Bumper and bumper beam weakness to absorb low speed structural crash test. RCAR (Research Council for Automobile Repairs).

One of the most important points is client's satisfaction, because excessive loss of stiffness in the hood or wings will seem to imply low quality. An example of this restriction can be the sensation of the client; if when the client opens the bonnet, he or she feels that the hood is weak, which could create dissatisfaction.

12 ACTIVE SAFETY SYSTEMS

Primary or active safety systems have been introduced to either aid the driver in reducing speed or automatically reduce the speed of the impacting car in a pedestrian crash. The "brake assist" system in the brake pedal senses the braking intention of the driver and automatically optimizes braking performance. The brake assist systems were

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 28. Devices with fracture under control.



Figure 30. Energy management in bonnet structure.



Figure 29. Move away the rigid elements from the impact areas.



Figure 31. Rain trough without rigid structures.

mandated in new vehicles in Europe in 2008. Infrared systems detecting living creatures such as animals or pedestrians and displaying the image on a screen to the driver were introduced in the early 2000s (Cadillac, Lexus) and were later followed by systems that additionally warned the driver (BMW, Audi, Honda, Mercedes, and Toyota). As brake assist systems are dependent on driver action, they were estimated to be activated only in 50% of accidents (Hannawald and Kauer, 2004). It is then natural to develop this system into an automatic system without driver intervention. A system was introduced in 2009 that detected pedestrians and gently applied the brakes if no driver action was noticed after a warning (Lexus, 2011). Recently, an auto-brake system was introduced that detects pedestrians and automatically applies full braking before an imminent impact (VolvoCars, 2010). This system has been claimed to be able to brake to a full stop from 25 km/h and thereby completely avoid low speed pedestrian crashes. At higher speeds, crash energy can be substantially reduced (Figure 35).

The pre-crash, or primary, safety measures and the in-crash, or secondary, safety measures can be combined into integrated systems. Integrated pedestrian systems have not been introduced in production cars. It is unclear whether an integrated system would be more effective than a single primary system such as autonomous braking. When

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 32. Increase distance between the bumper and the bumper beam.



Figure 33. Plastic frontal elements, increase distance to stiff elements.

developing an integrated system, it is also important to study how the two parts of the system interact. This can be performed using full-body impacts, and introducing both primary and secondary countermeasures.

12.1 Effectiveness/potential of countermeasures

Studies have tried to estimate the effectiveness of pedestrian protection systems. Lawrence *et al*. (2006) estimated the effectiveness of reducing fatally and seriously injured pedestrians, by introducing brake assist systems, to 10%.



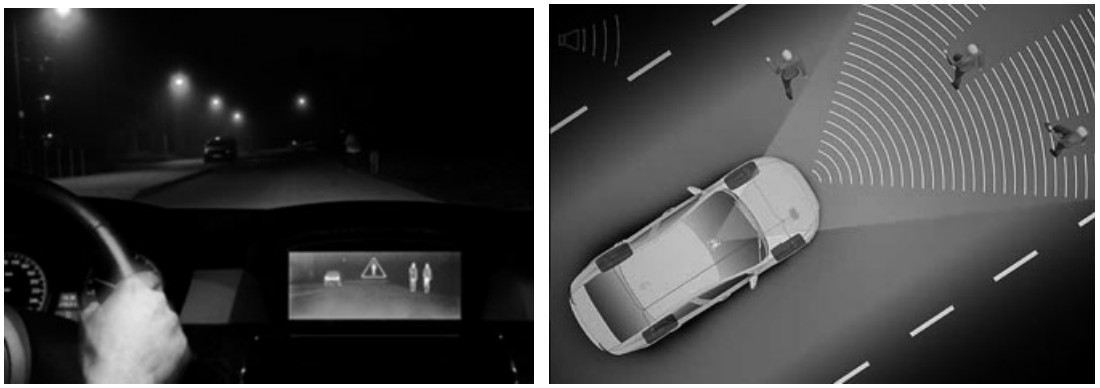
Figure 34. Aerodynamic new designs.

If the vehicle could brake autonomously, the effectiveness of the system would be increased. Rosén *et al*. (2010) estimated that an auto-brake system, activated for all visible pedestrians within a forward-looking angle of 40° one second before impact, would reduce fatalities (when struck by car fronts) by 40% and seriously injured by 27%.

It is unclear whether primary safety measures, such as automatic braking, can be enhanced by secondary safety measures. Fredriksson and Rosén (2010) studied 54 representative, severely head injured (AIS3+) pedestrians in detail to estimate the potential of theoretical primary and secondary systems and the potential of combining them into an integrated system. The primary safety system was assumed to brake (up to 0.6 g, depending on road friction) for all visible pedestrians one second before crash. The secondary system consisted of a deployable hood system and a lower windshield/A-pillar airbag covering up to 2.1 m WAD, estimated to be fully effective (when impacted) in avoiding AIS3+ injury up to 40 km/h and then have a linearly decreasing effectiveness. The research concluded that the passive (secondary) system could protect 34% of the severely head injured (AIS3+) and the active (primary) system 44%. If combining the systems into an integrated system, it protected a significantly higher number, 64% of the pedestrians, from severe (AIS3+) head injury (Figure 36).

Although the Fredriksson and Rosén study showed theoretically that primary and secondary systems complement each other to increase the protection potential, there is a need to further study the potential of integrated systems including information from real tests or simulations with the countermeasures.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



(a) (b)

Figure 35. (a) Primary pedestrian safety systems; (driver display of) pedestrian warning system. (Reproduced with permission from Autoliv. © Autoliv). (b) Auto-brake system detecting pedestrians at danger. (Reproduced with permission from Volvo. © Volvo Car Group.)

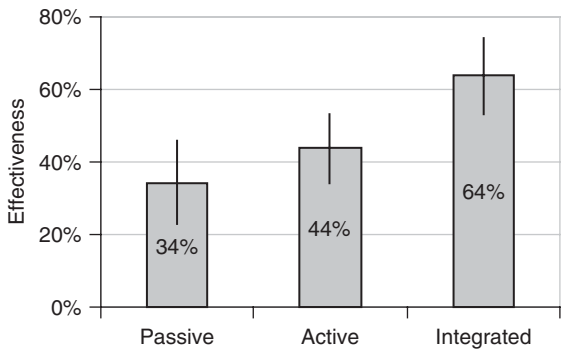


Figure 36. Effectiveness of passive (secondary), active (primary), and integrated systems. (Fredriksson and Rosén (2010). Reproduced with permission from R. Fredriksson.)

REFERENCES

Akiyama A., Okamoto M., Ito O., Takahashi Y. (2009) Development of omni-directional injury criteria for a pedestrian dummy for evaluating rib fracture. SAE (Society of Automotive Engineers) World Congress; Detroit, USA.

Akiyama A., Okamoto M., Rangarajan N. (2001) Development and Application of the New Pedestrian Dummy. *International Technical Conference of Enhanced Safety of Vehicles (ESV)*, Amsterdam, Netherlands.

Akiyama A., Yoshida S., Matsuhashi T., Moss S., Salloum M., Ishikawa H., Konosu A. (1999) Development of human-like pedestrian dummy. Japanese Society of Automotive Engineers, Chiyoda-Ku, Tokyo, Japan. Paper No. 9934546.

Aldman B., Kajzer J., Cesari D., Bouquet R., Zac R. (1985) A New Dummy for Pedestrian Test. *10th International Technical Conference on Experimental Safety Vehicles (ESV)*, Oxford, England.

Arregui-Dalmases, C. (2006) Rotational acceleration as a traumatic brain injury mechanism in pedestrian–vehicle collisions. PhD Thesis. Universidad Politécnica de Cataluña, Barcelona, Spain.

Ashton, S.J., Cesari, D., Van Wijk, J. (1983) Experimental reconstruction and mathematical modeling of real world pedestrian accidents. SAE 830189. Society of Automotive Engineers, Warrendale, PA., SAE, Feb, pp. 205–223.

Australian Transport Safety Bureau (ATSB) (2007) *Road Deaths Australia 2006 Statistical Summary*. Australian Government, ATSB Research and Analysis Report, Road Safety, DOTARS 50249, June, (accessed 4 June 2007).

Autoliv (2002) Autoliv annual report 2001. Stockholm, Sweden.

Autoliv (2010) Autoliv annual report 2009. Stockholm, Sweden.

Belingardi G., Scattina A., Gobetto E. (2009) Development of an Hybrid Hood to Improve Pedestrian Safety in Case of Vehicle Impact. *21st International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, Stuttgart, Germany. Paper No. 09–0026.

Björklund M., Zheng Q. (2001) Development and evaluation of a pedestrian anthropomorphic test device. MSc Thesis. Dept. of Machine and Vehicle Design, Chalmers University of Technology, Göteborg, Sweden.

Bose, D., Bhalla, K.S., Untaroiu, C.D., *et al*. (2008) Injury tolerance and moment response of the knee joint to combined valgus bending and shear loading *Journal of Biomechanical Engineering*, **130**, 10.1115/1.2907767.

Bunketorp O., Aldman B., Jonson R., Romanus B., Roos B., Thorngren L. (1981) Experimental Studies on Leg Injuries in Car-Pedestrian Accidents. *Proceedings of the 6th International IRCOBI Conference on the Biomechanics of Impacts*, 243–255.

Bunketorp O., Romans B., Hansson T., Aldman B., Thorngren L., Eppinger R. (1983) Experimental Study of a Compliant Bumper System. *Proceedings 27th STAPP Car Crash Conference*. Society of Automotive Engineers, Warrendale PA. SAE-831623 Oct, pp. 287–297.

Cesari D., Ramet M., Cavallero C. *et al*. (1980) Experimental study of pedestrian kinematics and injuries. International IRCOBI Conference on the Biomechanics of Impacts (5th) Proceedings, Amsterdam, 1980, pp 273–285.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Crandall, J.R., Bhalla, K.S., and Madeley, N.J. (2002) Designing road vehicles for pedestrian protection *British Medical Journal*, **324** (7346), 1145–1148.

Community database on accidents on the roads in Europe (CARE-European Road Accident Database) (2007) European Commission, Transport, Road Safety, Updated April 2007 (accessed 4 June 2007).

EC (2003) Directive 2003/102/EC of the European Parliament and of the Council of 17 November 2003 relating to the protection of pedestrians and other vulnerable road users before and in the event of a collision with a motor vehicle and amending council directive 70/156/eec. In: UNION, T. E. P. A. T. C. O. T. E., ed. 2003/102/EC. Brussels: Official Journal of the European Union.

EC (2009) Regulation (EC) no 78/2009 of the European Parliament and of the Council of 14 January 2009 on the type-approval of motor vehicles with regard to the protection of pedestrians and other vulnerable road users, amending directive 2007/46/ec and repealing directives 2003/102/EC and 2005/66/EC. In: UNION, T. E. P. A. T. C. O. T. E., ed. Vol 78/2009. Strasbourg: Official Journal of the European Union; 2009.

Euro NCAP (2009) Assessment Protocol - Overall Rating Version 5.0. European New Car Assessment Programme (Euro NCAP).

Euro NCAP (2011) Pedestrian Testing Protocol Version 5.2.1. European New Car Assessment Programme (Euro NCAP).

Fildes, B., Gabler, H.C., Otte, D., Linder, A., Sparke, L. (2004) Pedestrian Impact Priorities Using Real-World Crash Data and Harm. *2004 International Conference on the Biomechanics of Impacts (IRCOBI)*, Graz, Austria.

Fredriksson R., Flink E., Boström O., Backman K. (2007) Injury Mitigation in SUV-to-Pedestrian Impacts. *20th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, Lyon, France. Paper No. 07–0380.

Fredriksson R., Håland Y., Yang J. (2001) Evaluation of a New Pedestrian Head Injury Protection System with a Sensor in the Bumper and Lifting of the Bonnet's Rear Part. *17th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, Amsterdam, Netherlands. Paper No. 131.

Fredriksson R., Rosén E. (2010) Integrated pedestrian countermeasures - potential of head injury reduction combining passive and active countermeasures. IRCOBI (International Research Council on the Biomechanics of Impact) Conference; Hannover, Germany.

Fredriksson R., Zhang L., and Boström O. (2009) Influence of deployable hood systems on finite element modelled brain response for vulnerable road users. *International Journal of Vehicle Safety*, **4** (1), 29–44.

Gennarelli, T., Spielman, G.M., and Langfitt, T.W. (1982) Influence of the type of intracranial lesion on outcome from severe head injury *Journal of Neurosurgery*, **56**, 26–32.

Haddon, W., Jr, (1980) Advances in the epidemiology of injuries as a basis for public policy *Public Health Reports*, **95** (5), 411–421.

Han Y.H., Lee Y.W. (2003) Development of a vehicle structure with enhanced pedestrian safety. SAE (Society of Automotive Engineers) World Congress; Detroit, USA.

Hannawald L., Kauer F. (2004) Equal effectiveness study on pedestrian protection. Technische Universität Dresden.

Henary, B.Y., Ivarsson, J., and Crandall, J.R. (2006) The influence of age on the morbidity and mortality of pedestrian victims *Traffic Injury Prevention*, **7** (2), 182–190.

Institute for Traffic Accident Research and Data Analysis (ITARDA) (2004) Information No. 50: Pedestrian Traffic Accidents, http://www.itarda.or.jp/english/info50/50top.html (accessed 4 June 2007).

Kallieris, D., Schmidt, G. (1988) New Aspects of Pedestrian Protection Loading and Injury Pattern in Simulated Pedestrian Accidents. Paper 881725, *Proceedings 32nd Stapp Car Crash Conference*, pp. 185–196.

Kam C., Kerrigan J., Meissner M., Drinkwater C., Murphy D., Bolton J., Arregui C., Kendall R., Ivarsson J., Crandall J., Deng B., Wang J.T., Kerkeling C., Hahn W. (2005) Design of a full-scale impact system for analysis of vehicle pedestrian collisions. Paper 2005-01-1875, Society of Automotive Engineers, Warrendale, PA.

Kerrigan J.R., Arregui C., Crandall J.R. (2009) Pedestrian Head Impact Dynamics: Comparison of Dummy and PMHS in Small Sedan and Large SUV Impacts. *21st International Conference on the Enhanced Safety of Vehicles (ESV)*, Stuttgart, Germany. Paper No. 09–0127.

Kerrigan, J.R., Crandall, J.R., and Deng, B. (2007) Pedestrian kinematic response to mid-sized vehicle impact *International Journal of Vehicle Safety*, **2** (3), 221–240.

Kerrigan, J.R., Crandall, J.R., and Deng, B. (2008) A comparative analysis of the pedestrian injury risk predicted by mechanical impactors and post mortem human surrogates *Stapp Car Crash Journal*, **52**, 527–567.

Kerrigan J.R., Murphy D.B., Drinkwater D.C., Kam C.Y., Bose D., Crandall J.R. (2005a) Kinematic Corridors for PMHS Tested in Full-Scale Pedestrian Impact Tests. *19th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*.

Kerrigan J.R., Kam C.Y., Drinkwater D.C., Murphy D.B., Bose D., Ivarsson B.J., and Crandall J.R. (2005b) Kinematic comparison of the Polar-II and PMHS in pedestrian impact tests with a sport-utility vehicle. IRCOBI Conference on the Biomechanics of Impact.

Lawrence G.J.L., Hardy B.J., Carroll J.A., Donaldson W.M.S., Visvikis C., Peel D.A. (2006) A study on the feasibility of measures relating to the protection of pedestrians and other vulnerable road users-Final 2006. EC Report UPR/VE/045/06, Contract ENTR/05/17.01. TRL (Transport Research Laboratory) Limited, UK.

Lexus (2011) Advanced pre-collision system (APCS) with driver attention monitor, (accessed 13 January 2011).

Loo, B.P. and Tsui, K.L. (2009) Pedestrian injuries in an ageing society: insights from hospital trauma registry *Journal of Trauma*, **66** (4), 1196–1201.

Lucas (2000) Flexible lamp mounting. USA patent application US6190030. 1999-02-18.

Maki T., Asai T., Kajzer J. (2003) Development of Future Pedestrian Protection Technologies. *18th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, Nagoya, Japan. Paper No. 03–0165.

Matsui Y., Ishikawa H. (1998) Validation of pedestrian upper legform impact test, reconstruction of pedestrian accidents Paper 98-S10-O-05.

Mizuno Y. (2003) Summary of IHRA pedestrian safety WG activities-proposed test methods to evaluate pedestrian protection afforded by passenger cars. NHTSA Paper 580, *Proceedings 18th*

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Conference on the Enhanced Safety of Vehicles (ESV), Nagoya, Japan.

Mizuno Y. (2002) Summary of IHRA Pedestrian Safety WG Activities.

Montoro L., Alonso F., Esteban C., Toledo F. (2000) Manual de seguridad vial: El factor humano, Ed. Intras.

Nagatomi K., Hanayama K., Ishizaki T., Sasaki S. (2005) Development and Full-scale Dummy Tests of a Pop-up Hood System for Pedestrian Protection. *19th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, Washington DC, USA.

National Highway Traffic Safety Administration (NHTSA) (2006) Traffic Safety Facts 2005 Data-Pedestrians. DOT HS 810 624.

National Police Agency (Japan) (2006) Traffic Accidents Situation, 2006 Data, (accessed 1 June 2007).

Okamoto M., Akiyama A., Takahashi Y. (2009) Pedestrian dummy pelvis impact responses. SAE (Society of Automotive Engineers) World Congress; Detroit, USA.

Okamoto Y., Akiyama A., Nagatomi K., Tsuruga T. (1994) Concept of Hood Design for Possible Reduction in Pedestrian Head Injury. *14th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, Munich, Germany. Paper No. 94-S7-W-14.

Okamoto Y., Kikuchi Y. (2006) A Study of Pedestrian Head Injury Evaluation Method. *Proceedings 2006 International IRCOBI Conference on the Biomechanics of Impact*, pp. 265–276.

Otte D., Tohlemann T. (2001) Analysis and load assessment of secondary impact to adult pedestrians after car collisions on roads. IRCOBI, Isle of Man.

Pipkorn B., Fredriksson R., Olsson J. (2007) Bumper Bag for SUV to Passenger Vehicle Compatibility and Pedestrian Protection. *20th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, Lyon, France. Paper No. 07–0056.

Pritz, H.B., Hassler, C.R., Herridge, J.T., Weis, E.B., Jr., (1975) Experimental Study of Pedestrian Injury Minimization Through Vehicle Design. *Stapp Car Crash Conference. Nineteenth. Proceedings*. Warrendale, Society of Automotive Engineers, 1975, pp. 725–751. SAE 751166.

Renaud F., Tapia F. (2004) Development of a mechanical neck for a 6-year-old pedestrian dummy. M.Sc. Dept. of Machine and Vehicle Systems, Chalmers University of Technology, Göteborg, Sweden.

Renaud F., Tapia F., Fredriksson R., Yang J. (2005) Development of a Mechanical Neck for a Six-year-old Pedestrian Dummy. *IBS Conference*, Ohio, USA.

Rosén, E., Källhammer, J.-E., Eriksson, D., *et al*. (2010) Pedestrian injury mitigation by autonomous braking *Accident Analysis & Prevention*, **42** (6), 1949–1957.

Rosén, E. and Sander, U. (2009) Pedestrian fatality risk as a function of car impact speed *Accident Analysis & Prevention*, **41** (3), 536–542.

SAE (2008) Pedestrian dummy full scale test results and resource materials. Surface Vehicle Information Report J2868, SAE International.

SAE (2009) Performance specifications for a midsize male pedestrian research dummy. Surface Vehicle Information Report J2782, SAE International.

Sakurai M., Kobayashi K., Ono, K., A. Sasaki (1994) Evaluation of Pedestrian Protection Test Procedure in Japan. *The fourteenth International Technical Conference on Enhanced Safety of Vehicles*, 94-S7-O-01, pp. 1114–1130.

Schroeder G., Fukuyama K., Yamazaki K. (2008) Injury Mechanism of Pedestrians Impact Test with a Sport-utility Vehicle and Mini-van. *IRCOBI (International Research Council On the Biomechanics of Impact) Conference*, Bern, Switzerland.

Schroeder, G., Konosu, A., Ishikawa, H., Kajzer, J. 2000 Injury mechanism of pedestrians during a front-end collision with a late model car. JSAE Spring.

Snedeker J., Walz F., Muser M., Lanz C., Schroeder G. (2005) Assessing femur and pelvis injury risk in car–pedestrian collisions comparison of full body PMTO impacts, and a human body finite element model. NHTSA, Paper 05–0103, *Proceedings of the 19th International Technical Conference on Enhanced Safety of Vehicles*, Washington, DC.

Subit D., Kerrigan J., Crandall J., Fukuyama K., Yamazaki K., Kamiji K., Yasuki T. (2008) Pedestrian-vehicle Interaction: Kinematics and Injury Analysis of Four Full-scale Tests. *IRCOBI (International Research Council On the Biomechanics of Impact) Conference*, Bern, Switzerland.

Takahashi Y., Okamoto M., Akiyama A., Kikuchi Y. (2009) Estimation of knee ligament injury measures for a pedestrian dummy. SAE (Society of Automotive Engineers) World Congress; Detroit, USA.

UN (2009) Global Technical Regulation no. 9 - pedestrian safety. ECE/TRANS/180/Add.9. United Nations, Geneva, Switzerland.

Untaroiu, C., Kerrigan, J., Kam, C., *et al*. (2007) Correlation of strain and loads measured in the long bones with observed kinematics of the lower limb during vehicle-pedestrian impacts *Stapp Car Crash Journal*, **51**, 433–466.

Untaroiu, C.D., Shin, J., Ivarsson, B.J., *et al*. (2008) A study of the pedestrian impact kinematics using finite element dummy models: the corridors and dimensional analysis scaling of upper-body trajectories *International Journal of Crashworthiness*, **13** (5), 469–478.

VolvoCars (2010) A revolution in pedestrian safety - Volvo's automatic braking system now reacts to people as well as vehicles, www.volvocars.com/za/top/about/news-events/pages/default.aspx?itemid=24 (accessed 22 August 2010).

World Bank Group (2002) The "*Road Crash Problem*" *Road Safety*, (accessed 4 June 2007).

World Bank Group (2006) The "*How to Improve Road Safety*" Road Safety, (accessed 19 March 2006).

Yang, J. (2005) Review of injury biomechanics in car-pedestrian collisions *International Journal of Vehicle Safety*, **1** (1/2/3), 100–117.

Yang J.K., Lövsund P. (1987) Development and Validation of a Human-body Mathematical Model for Simulation of Car-pedestrian Collisions. *IRCOBI (International Research Council On the Biomechanics of Impact) Conference*, Hannover, Germany.

Youn Y., Kim S., Oh C., Shin M., Lee C. (2005) Research and Rule-Making Activities on Pedestrian Protection in Korea. NHTSA Paper 05–0117. *Proceedings of the 19th International Technical Conference on Enhanced Safety of Vehicles*, Washington DC, US.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Zellmer H., Glaeser K.-P. (1994) The EEVC-WG 10 Head Impact Test Procedure in Practical Use. *14th International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, Munich, Germany. Paper No. 94-S7-O-03.

Zhang, G., Cao, L., Hu, J., and Yang, K.H. (2008) A field data analysis of risk factors affecting the injury risks in vehicle-to-pedestrian crashes *Annals of Advances in Automotive Medicine (AAAM)*, **52**, 199–213.

Zhao, H., Yin, Z., Chen, R., *et al*. (2010) Investigation of 184 passenger car-pedestrian accidents *International Journal of Crashworthiness*, **15** (3), 313–320.

FURTHER READING

Gennarelli T., Pintar F., Yoganandan N. (2003) Biomechanical Tolerances for Diffuse Brain Injury and a Hypothesis for Genotypic Variability in Response to Trauma. *Annual proceedings AAAM*.

Lopez, A.D., Mathers, C.D., Ezzati, M., *et al*. (2006) Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data *Lancet*, **367** (9524), 1747–1757.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto260
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Automotive Applications for Magnesium

Roger Grimes and Vit Janik

University of Warwick, Coventry, UK

1	Introduction	1
2	History	2
3	Production of Magnesium Alloys	2
4	Properties of Magnesium	6
5	Applications of Cast Magnesium	7
6	Applications of Wrought Magnesium	10
7	The Future	13
References		15
Further Reading		17

1 INTRODUCTION

Magnesium is the lightest of all structural metals having a density of 1.74 g/cm^3 compared with 2.7 g/cm^3 for aluminium and 7.8 g/cm^3 for steel. Thus, it might be expected to find major use in the construction of motor vehicles, particularly in an era where there are major concerns over fossil fuel consumption and environmental damage. Probably, the best-known automotive application of magnesium alloys was their introduction, in 1936, into the cylinder block and gearbox housing of the Volkswagen Beetle. This was very successful and, over the life of the model, consumed large quantities [42,000 tonnes per annum at its peak in 1971 (Friedrich and Schumann, 2001)] of magnesium alloy. Despite this success, there has, until relatively recently, probably been a perception that magnesium alloys are only appropriate for use in exotic vehicles.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

The fuel crises of the mid-1970s led in the United States to major efforts to reduce their dependency on imported oil (35% in 1973) and, particularly, to reduce the consumption by automobiles. In 1978, legislation was introduced in the form of the Corporate Average Fuel Economy (CAFE) regulations to force the vehicle builders to construct more fuel efficient vehicles and, in consequence, smaller vehicles were introduced together with new construction techniques and greater use of lightweight materials. This resulted in a dramatic reduction in the average weight of a North American car between 1977 and 1980, followed by a slower reduction between 1980 and 1986. However, from 1986 onward, and despite the growing use of low density materials, average weight has increased by about 1% per annum (Figure 1). In Europe, a similar pattern of weight creep has been observed. Here, the major driver for mass reduction has been the concern for the environment, and targets for reduced CO_2 emissions were reached by voluntary agreements between the vehicle builders. But, as in the United States, despite the greater employment of low density materials and the introduction of improved construction methods, average car weight has continued to rise.

Between 1960 and the fuel crises of the mid-1970s, the aluminium content and the plastics content of passenger cars remained roughly constant and averaged about 2%. With the search for lighter vehicles, these contents had risen to about 10% and 15%, respectively, by the year 2000. The magnesium content of the Lincoln LS, a midsize vehicle launched in 2000, was only 0.3% (aluminium 12% and plastic 10%). Thus, in the early efforts to reduce vehicle mass, magnesium hardly figured at all, probably because of a variety of unfavorable perceptions such as corrosion, flammability, cold workability, and, particularly, cost. However, from about this time onward, there has been a steady growth in the range of vehicle applications

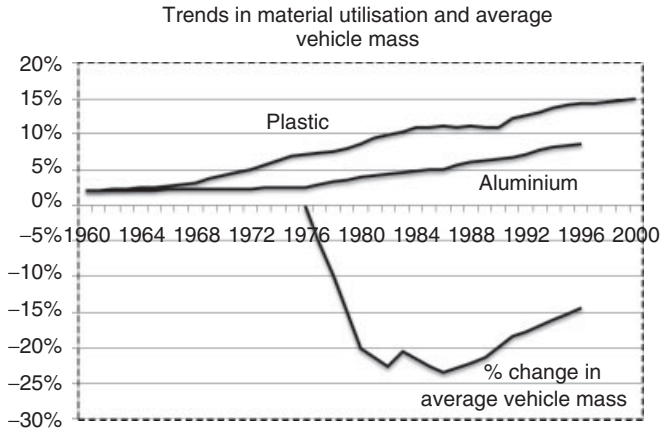


Figure 1. Trends in material utilization and average vehicle mass.

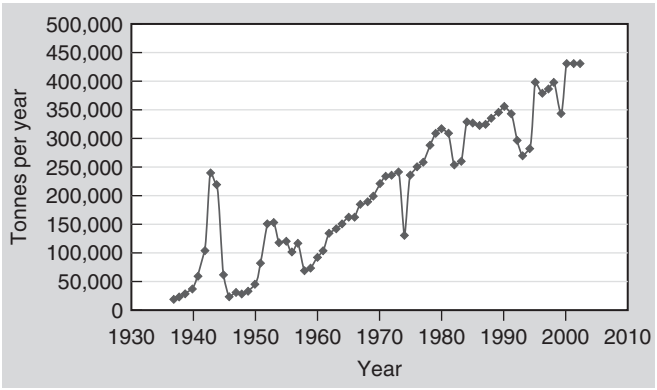


Figure 2. Virtually linear growth in primary magnesium production. (Reproduced from JOM, 59(2), 2007, 30–38, The evolution of technology for light metals over the last 50 Years: Al, Mg, and Li, J.W. Evans, Figure 1. With kind permission from Springer Science and Business Media.)

that has been taken over by magnesium and a dramatic increase in R&D activities investigating means by which the weight-saving potential of magnesium alloys may best be practically exploited. At present, the growth in actual applications has, virtually completely, been in castings and, particularly, high pressure die castings (HPDCs) and the actual average magnesium content remains at a very low percentage of the vehicle mass. Nevertheless, it seems likely that the combination of the mass reducing potential that magnesium alloys provide with the environmental pressures on vehicle manufacturers will lead to significantly greater use in the coming years.

The following sections of this chapter outline the history of magnesium, its manufacture, its availability, the properties that make it attractive for vehicle construction, and likely developments in cast and wrought applications and attempt to forecast future progress.

2 HISTORY

The existence of magnesium was proved by Davy in 1808, and the separation of elemental magnesium was achieved by Busy in 1828 (Polmear, 1999). He achieved this by reducing magnesium chloride with potassium. Subsequently, two industrial processes were developed, one based on electrolysis of molten magnesium chloride and one on reduction of the magnesium salt with ferrosilicon and derivatives of these processes form the basis of all the primary magnesium production systems in use today. Magnesium is the eighth most abundant element in the Earth's crust and is found in numerous well-distributed minerals as well as in sea water at about 0.2%. It is most frequently extracted from dolomite ($MgCO_3 \cdot CaCO_3$) or from $MgCl_2$ derived from brine. However, several other minerals could be employed and are currently under consideration as the magnesium source in some new, possible, extraction operations (see the following sections).

Magnesium alloys found some application in early motor vehicles but, for many years, the largest use for magnesium was as an alloying addition to aluminium and the growth in magnesium production was largely driven by the growth in the consumption of aluminium alloys containing magnesium. Significant quantities are consumed in desulfurizing steel but, for the past 15 years, or so, there has been a steady growth in the quantity of magnesium being used for the manufacture of die castings for motor vehicles such that roughly equal quantities of magnesium are now consumed in alloying aluminium and in automotive castings. While the growth in production of primary magnesium has fluctuated considerably, particularly peaking during wars and dropping back thereafter, over the past 60 years, growth has been virtually linear (Figure 2) and now stands at close to 600,000 tonnes per annum. The use of wrought magnesium alloys has consumed a very small fraction of this total and although, from time to time, there has been some automotive consumption of sheet and extrusions, current actual applications are negligible.

3 PRODUCTION OF MAGNESIUM ALLOYS

3.1 Extraction of primary magnesium

The separation of magnesium from its ores inevitably consumes large quantities of energy. While several extraction variants exist, essentially, two processes

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

are currently employed: either electrolysis of molten magnesium chloride, in a process that is basically similar to that used to separate aluminium from alumina, or thermal reduction of magnesium oxide. The former is more environmentally friendly but relatively capital intensive and, compared with aluminium extraction, has the disadvantage that the molten magnesium is less dense than the electrolyte from which it has been separated and so floats on the surface of the cell and must be protected from the atmosphere. Aluminium is more dense than the electrolyte and so collects on the bottom of the cell from where it can be periodically tapped off. The energy involved in the electrolytic part of the extraction of the two elements is fairly similar and can be about 13 kWh/kg (Evans, 2007) for either, albeit older designed cells may have significantly higher power consumption. Unfortunately, the most commonly used protective gas in the magnesium primary industry is sulfur hexafluoride (SF_6) with global warming potential 23,000 times higher than CO_2 (Ramakrishnan and Koltun, 2004). Even though the magnesium industry uses relatively small amounts of SF_6, the overall current contribution of SF_6 to global warming is estimated to be 0.2% annually, with a further potential rise because of the increased utilization of magnesium in the automotive industry. Several alternative cover gases with lower warming potential than SF_6 have been introduced recently (United States Environmental Protection Agency, EPA, 2008). The leading candidates for further application in magnesium primary production and recycling include sulfur dioxide, 1,1,1,2-tetrafluoroethane cover gas known as *AM-cover*® developed by Advanced Magnesium Technologies, and fluorinated ketone fluid available under the commercial designation Novec™ 612 supplied by 3M™.

The thermal reduction of magnesium oxide was developed by Pidgeon and Alexander (1944) and is a batch process in which the magnesium oxide is reduced by ferrosilicon:

$$2(CaO \cdot MgO) + FeSi \rightarrow Mg_{(g)} + 2CaO \cdot SiO_2 + Fe$$

Calcined dolomite, or magnesium oxide, is charged together with ferrosilicon into retorts and heated to about 1200°C under vacuum. Magnesium vapor is formed, as indicated in the preceding reaction, which condenses in the cold part of the retort. Not only is the process energy intensive during the reduction phase but the production of the ferrosilicon employed in the reduction process is also a high temperature, energy-intensive operation.

Ramakrishnan and Koltun (2004) compared the energy consumption involved in electrolytic extraction with that required for thermal reduction when all the steps of the extraction processes were included. They concluded that in energy terms the electrolytic process was significantly better than the thermal process requiring 174 and 284 MJ/kg Mg, respectively.

The geographic distribution of primary magnesium production has undergone dramatic changes since the year 2000 and these are considered in Section 7.1. These changes have, however, led to new producers considering entering primary production and to the consideration of employing new, or evolved, processes for the production. It seems appropriate to consider the new processes in this section.

Probably, the nearest to actual production is the SilMag operation, a joint venture between the Norwegian company Hydro and the German Advanced Metallurgical Group (AMG). The actual extraction of magnesium is to be fairly conventional electrolysis of magnesium chloride using the cells at Porsgrunn from the mothballed Norsk Hydro plant. The intention is to use olivine, a magnesium iron silicate with a magnesium content up to about 33% and found in abundance on the west coast of Norway, as the starting mineral. The ore would be treated with hydrochloric acid to produce magnesium chloride and silica. Annual capacity is predicted to be about 34,000 tonnes of magnesium with a similar quantity of silica. SilMag believe that energy consumption would be about 25 kWh/kg Mg and there would be virtually no CO_2 produced. Production should commence in 2014 (Willekens, 2012).

The Zuliani process, being developed by Gossan Resources, is a variant of the thermal reduction of calcined dolomite with ferrosilicon but operating continuously at atmospheric pressure to produce a high purity, liquid product. The major power source would be hydroelectricity from Manitoba where the prices are both stable and very low. Modeling and bench-scale operations have indicated a potential green house gas (GHG) emission of only 9.1 kg CO_2/kg Mg. However, while there are plans for a 5000 tonnes per annum pilot plant, at present, the process has only been demonstrated on a bench scale (Zuliani and Reeson, 2012).

Another variant of thermal reduction has been under investigation since 2003 by CSIRO in Australia (Prentice, Poi, and Haque, 2010). They have improved on previous attempts to use carbothermic reduction of the magnesite by employing extremely rapid quenching of the reaction products and, thus, suppressing the reversion reaction and allowing the magnesium to condense and solidify.

$$MgO_{(s)} + C_{(s)} \leftrightarrow Mg_{(v)} + CO_{(g)}$$

GHG emissions would lie between 6.4 and 21.3 kg CO_2/kg Mg and, thus, at worst, be roughly equivalent to

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

the best operations of the Pidgeon process in China at 25 kg CO_2/kg Mg (Ehrenberger *et al*., 2008). The CSIRO version of carbothermic reduction is called *MagSonic*™ and detailed analysis of the costs of running the new process compared with either electrolytic extraction or the Pidgeon process has been conducted (Prentice and Haque, 2012). Their paper concluded that, potentially, of the three processes, MagSonic™ would be the least labor intensive, have the lowest global warming potential, and be only marginally more capital intensive than Pidgeon. However, thus far, the MagSonic™ process has only been operated as a laboratory batch process, whereas continuous industrial operation is envisaged.

3.2 Production of magnesium alloy sheet

At present, commercial magnesium alloy sheet is manufactured by the traditional route of employing vertical, semi-continuous, direct chill (DC) casting to produce a rectangular cross-section ingot that, after appropriate scalping and heat treatment, is hot rolled to an intermediate gauge on a reversing mill before reheating and warm rolling to final gauge and temper. While facilities exist that are capable of producing large quantities of sheet by this route, the demand for magnesium sheet is very small and thus the price is, inevitably, high. If magnesium sheet is to find significant application in volume motor vehicles, it is likely to be in competition with aluminium alloy sheet and must be able to compete with aluminium on a price basis. Given that any sheet structure fabricated from magnesium will be lighter than the comparable aluminium structure, it has been estimated that, to be competitive, magnesium sheet should be no more than 1.3 times the price of aluminium (Zuliani and Reeson, 2012). Currently, it is suggested that the price of magnesium sheet is five times that of aluminium (US AMP Magnesium Vision 2020, 2006).

The suggestion is that a sheet production system based on twin-roll strip casting (TRC) could reduce the cost and simplify manufacture to the extent that the cost of magnesium alloy sheet would be not more than 20% higher than that for aluminium alloy sheet (US AMP Magnesium Vision 2020, 2006). Detailed modeling by Herling and Carpenter (2005) of the likely cost of magnesium alloy sheet produced via TRC suggested a reduction to US$4.34/kg from US$9.92/kg for conventionally cast and rolled magnesium sheet. This was for a 10,000 tonnes per annum operation producing 1.5 mm gauge sheet at a width of 1 m.

The first of the current investigations into the potential of TRC for the production of magnesium alloy sheet was that of CSIRO in Australia (Liang and Cowley, 2004). Starting in about 2000, they developed a casting system capable of producing 300 mm wide strip by 3 mm gauge and estimated that full-scale production by this route would save 60% of total production costs and reduce associated GHG emissions by 50%. They also suggested that sheet made via TRC should, because of the rapid solidification, have superior properties together with reduced segregation and preferred crystallographic orientation compared with that made via conventional casting. The technology developed is available but has not led to a sheet manufacturing operation. More major developments of TRC sheet production systems are taking place in Korea and Germany.

In Korea, in 2002, the Pohang Iron and Steel Company (POSCO), in conjunction with the Research Institute for Science and Technology (RIST), initiated planning for diversification into a magnesium business (The Korea Herald, 2011). By 2008, they were able to produce 600 mm wide by 3 mm strip in coils up to 1220 mm diameter and by 2011 had increased the cast strip width to 2000 mm (Park *et al*., 2011). POSCO plans to have an integrated magnesium sheet production line, initially supplying thin-gauge, narrow strips for application in electronic devices on a pilot plant basis but providing wide sheets for automobile applications in due course. While the TRC equipment for producing 2 m wide coils is being moved from RIST to the POSCO plant at Suncheon, the warm rolling mill that will be necessary to complete the integrated line has not yet been installed but some 1500 mm wide coils cast at RIST have been successfully rolled to strip at Magnesium Elektron North America (MENA) (Choo *et al*., 2012). In June 2010, POSCO entered a collaboration agreement to exchange technology with the Thysen Krupp Stahl subsidiary Magnesium Flachprodukte GmbH (MgF).

In Germany, research into the potential for greater application of magnesium alloys in motor vehicles had been underway since the 1990s, and in 2001, Thysen Krupp Stahl, in conjunction with the University of Freiberg, formed the organization Magnesium Flachprodukte GmbH to accelerate the project. In 2002, MgF acquired a TRC machine capable of casting 700 mm wide strip of magnesium alloys in gauges between 3.5 and 8 mm (Figure 3) (Kawalla *et al*., 2008). Sheets cut from the TRC coils were individually hot/warm rolled but, in late 2009, the plant was augmented with a hot strip rolling mill, thus making possible accurate simulation of a full-scale production operation.

In Turkey, the Materials Institute TUBITAK MRC has developed a TRC capable of producing strip up to 1.5 m wide and in the gauge range 4.5–6.5 mm (Kaya *et al*., 2008; Duygulu *et al*., 2009) and has found the resulting warm-rolled strip to be appreciably less anisotropic than comparable strip produced via the ingot route (Duygulu *et al*., 2009).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 3. Twin-roll casting plant of the MgF Magnesium Flachprodukte GmbH in Freiberg. (From Kawalla *et al*. (2008). Reproduced by permission of Croatian Metallurgical Society.)

In China, Taiwan Mach has made significant magnesium investment in Shandong Province including a cast-rolling line said to be operational in late 2012. Its capacity is expected to be 2000 tonnes per annum, but strip dimensions are not disclosed (Metals Week, 2012a).

As the overwhelming majority of magnesium alloys have a close packed hexagonal crystallographic structure, they have very limited slip systems on which to deform at ambient temperature and, hence, very limited ductility. Rolling has to be carried out either hot or warm and this tends to produce strong basal plane crystallographic textures that add to the difficulties in subsequent forming operations. While strip casting reduces the overall strain involved in strip production, textural problems can remain. Various techniques, such as equal channel angular extrusion (ECAE), have been employed to modify and refine the structure of magnesium alloys, but the majority of such techniques could not be employed on an industrial scale. However, it has been shown that the redundant work introduced by asymmetric (or shear) rolling can greatly modify the resultant strip microstructure (Cui and Ohori, 2000) and this technique might be applicable on a large scale. Kim, Kim, and Wang (2009) applied this technique to ZK60 and found that the grain size could be refined to $1-2\,\mu m$ by a single pass, although the experiment was performed on a relatively small scale with material that had been extruded before rolling. An alternative route leading to significant reduction of undesirable basal texture after rolling of magnesium alloys can be achieved by modification of their microstructure with rare earth (RE) elements, the most powerful options being cerium, neodymium, or yttrium (Hantzsche *et al*., 2010).

A joint project between Oak Ridge National Laboratory and MENA was duly set up to investigate the extent to which shear rolling could be employed to produce magnesium alloy sheet with improved cold formability (Randman *et al*., 2011). The five tasks in the project scope included design of a shear mill, alloy selection, and demonstration of component fabrication. Shear rolling did not dramatically refine the grain structures that were generally fairly uniform apart from thin ($\sim 20\,\mu m$) surface layers of refined structure, while the basal texture was tilted away from the sheet normal and toward the rolling direction. In the course of the development, an experimental rolling mill was designed and built at Oak Ridge National Laboratory with cooperation from Hunter Fata Inc., enabling detailed investigation of the shear rolling process (Oak Ridge National Laboratory, 2012). This has now enabled Hunter Fata to design a new magnesium reversing warm rolling mill (Passoni *et al*., 2012) for producing large wide coils from a twin-roll caster. The mill incorporates features designed to overcome the limitations of previous magnesium warm mills so that hot coilers are placed on either side of the rolling stand and are able to carry out intermediate annealing as well as allowing reheating. The mill drive system is capable of either conventional rolling or shear rolling.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Thus, while at present there is no magnesium producer capable of manufacturing wide, thin magnesium sheets in the volume (and at the price) that would be required for significant use in volume vehicles, the technology for casting and rolling such a material does exist.

4 PROPERTIES OF MAGNESIUM

It is obviously not appropriate to attempt comprehensive listing of the properties of magnesium alloys in a chapter like this and readers seeking such information are referred to publications such as the ASM Speciality Handbook *Magnesium and Magnesium Alloys* (Avedesian and Bake, 1999) or manufacturers', such as Magnesium Elektron, web sites (http://www.magnesium-elektron.com). This section will simply give a brief summary of the alloy designation system and some of the more widely used alloys. Although not formally recognized internationally, the American Society for Testing Materials (ASTM) alloy descriptions are widely used. In this system, the first two letters of the designation indicate the major alloying elements, while the numbers indicate the alloy content, rounded to the nearest whole number. Table 1 indicates the element code letters used in the ASTM system. Table 2 gives a brief summary of the properties of some of the main sheet materials competing for structural vehicle applications. The listed magnesium alloy, AZ31B, seems at present to be the sheet alloy of choice, although it was developed long ago and there are current development activities to produce improved alloys. It is likely that for superplastically formed applications, AZ31B would be in competition with the aluminium alloy AA5083, whereas if the aluminium alloy had been conventionally formed, it would be likely to be a stronger alloy, such as AA6022. Magnesium and its alloys have a significantly lower modulus than either aluminium or steel, which give them a considerable advantage in reducing noise, vibration, and harshness (NVH) problems and in improved dent resistance but give the disadvantage that if magnesium sheet is substituted for aluminium sheet in stiffness critical applications, there has to be an increase in the gauge of the magnesium alloy to compensate. Magnesium has excellent machining behavior and can be readily recycled if one is dealing with clean segregated scrap. However, when dealing with end-of-life vehicles and postconsumer scrap recycling, the situation becomes much more difficult and more work is needed (Fechner, Hort, and Kainer, 2009).

The cold forming capability of magnesium is severely restricted, so that the manufacture of the body panels would have to be either by warm conventional forming or

Table 1. Element code letters used in the ASTM designation system for magnesium alloys.

Letter	Alloying Element
A	Aluminium
C	Copper
E	Rare earth metals
H	Thorium
K	Zirconium
L	Lithium
M	Manganese
Q	Silver
S	Silicon
W	Yttrium
Z	Zinc

Table 2. Properties of a variety of competing sheet materials.

Property	AZ31B	AA5083	DDQ steel	AA6022 T43[a]
Density (g/cm^3)	1.78	2.65	7.86	2.69
Modulus (GPa)	45	71	207	71
0.2% proof stress (MPa)	140	145	165	272
Tensile strength (MPa)	240	290	310	325
Specific strength (MPa)	135	109	40	120
Specific stiffness (GPa)	25	27	26	26
Elongation (%)	10	25	45	18
Crystal structure[b]	cph	fcc	bcc	fcc

[a] 1 mm sheet, 2% pre-strain and 32 min at 185°C paint bake.
[b] cph, closed placked hexagonal; fcc, face centered cubic; bcc, body centered cubic.

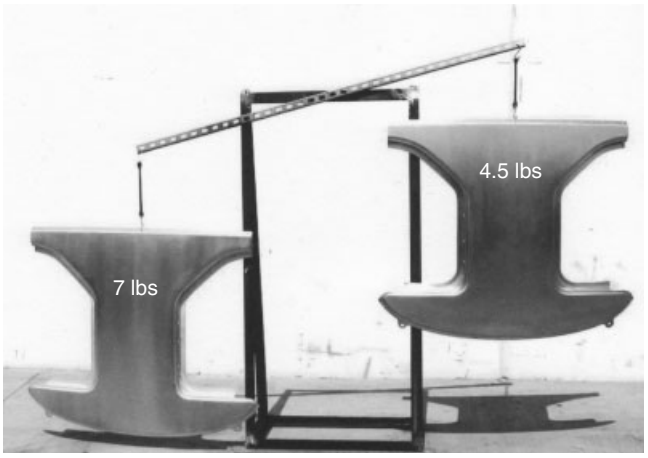


Figure 4. Superplastically formed component. (Superplastically formed component; on left from SPF AA5083, on right from AZ31B. Courtesy Superform Aluminium.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

by superplastic forming or some derivative thereof. Most magnesium alloys appear to be capable of superplastic forming with little, or no, modification to their manufacturing route (Grimes, 2011). For low volume niche vehicle manufacturing, the material condition after TRC is suitable for direct manufacturing by nonconventional routes such as superplastic forming, allowing more freedom in design and significant part consolidation and weight reduction (Figure 4). It is believed that the combination of TRC and superplastic forming of magnesium sheet could be attractive to the high end automotive market and be competitive with carbon fiber processing technology for panels and closures.

5 APPLICATIONS OF CAST MAGNESIUM

The combination of low density and excellent casting behavior makes substitution of steel or aluminium components with magnesium castings an attractive means of making significant mass reduction and it is this that has, over the last decade or so, resulted in the considerable growth in magnesium alloy consumption. This weight saving will improve the performance of the vehicle because of a reduction in the rolling resistance and energy of acceleration, allowing lower fuel consumption and also reduction in the greenhouse gases (Mordike and Ebert, 2001; Blawert, Hort, and Kainer, 2004; Kim and Han, 2008).

Factors inhibiting wider utilization of magnesium castings in various areas of vehicles are chiefly related to the higher cost of primary magnesium and with a lack of knowledge and experience with magnesium. Magnesium alloy castings are currently largely restricted to the top-end, low volume automotive market, where the cost penalty is not significant and high performance is required. However, legislative pressure for low carbon vehicles and also voluntary commitments from manufacturers to reduce the average fuel consumption of their cars seem likely to lead to a significant increase in magnesium-related applications and expansion of magnesium into high volume automotive production (Blawert, Hort, and Kainer, 2004; Robinson, 2011).

Cast magnesium alloys can be classified into those suitable for less demanding applications [body-in-white (BIW), interior, and chassis] and those suitable for the more demanding applications in the powertrain that carry a higher requirement for high temperature stability.

Areas of typical applications of magnesium alloys in motor cars have been suggested as follows (Luo and Sachdev, 2012; Kulekci, 2008):

- Structural applications—interior, chassis, and BIW: requirements for castability, room temperature strength, joining ability, corrosion resistance, damping capacity, and low cost; additional requirements for chassis and BIW components include crash worthiness, fatigue strength, and durability. High surface quality and increased corrosion resistance is necessary for exterior parts.
- Powertrain applications: with additional requirement for mechanical properties at elevated temperatures, creep resistance, fatigue resistance, and corrosion resistance under wear.

5.1 Applications of magnesium alloys for interior, chassis, and body-in-white

Apart from the high specific strength and good rigidity of magnesium alloys, the chief advantage of magnesium compared to aluminium or steel is its low dynamic vibration response together with the ability to absorb energy (Cole, 2011a). Improved corrosion resistance of magnesium alloys can be achieved by maintaining the content of impurities in solid solution to a maximum of 0.002 wt% for Ni and Fe and 0.015 wt% for Cu. AZ91 is the preferred choice for automotive designers because of its high die castability and good room temperature strength. AZ91 can be applied in a wide range of common low-demanding structural applications in the automotive interior where it can substitute for aluminium alloys and compete with polymeric or composite materials. AM50A and AM60B are alternative alloys with better ductility than AZ91; these alloys are generally utilized in applications exposed to possible crash scenarios where the chief design parameter is the energy absorption of the component (Sadayapan and Luo, 2011). Alloys AE42 and AE44 are slightly stronger at elevated temperatures and more creep resistant than AZ and AM alloys; the microstructure of AE alloys is reinforced by a complex precipitation of RE-containing particles.

Examples of interior applications of magnesium alloys include the steering wheel and steering column system, where lighter magnesium is favored to compensate for the increased weight of the integrated airbag assembly. Instrument panel, door handles, pedals, radio and airbag housing, and other smaller interior trims and features are also frequently made of magnesium AZ91 (Abbott, Easton, and Caceres, 2003). Alloys AM20, AM50A, and AM60B are further applied for seat frame and supporting construction, and bracket shells (Schumann and Friedrich, 2006). A summary of common grade alloys is provided in Table 3.

Compared to the interior, current utilization of magnesium alloys in the automotive chassis is limited to engine

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 3. Common grade die-casting alloys; advantages include good die castability, relatively good ductility, improved vibration and energy absorption behavior, and high room temperature strength.

Alloy	AMS/ASTM	Composition wt%[a]			Condition[b]	Mechanical Properties			Remarks
		Al	Mn	Zn		YTS (MPa)	UTS (MPa)	el. (%)	
AZ91	AMS 4437E	8.5–9.5	0.2–0.4	0.5–0.9	F	100	155	2	Sand and chill castings
	AMS 4437D				F	160	230	2	Die casting
	AMS 4446C				T4	80	215	4	Sand casting
	AMS 4452E				T6	140	260	6	Investment casting, high purity grade peak aged
AZ92	AMS 4434M	9.0	0.4	1.9	T6	150	275	3	Higher strength than AZ91, lower ductility
AM20	ATSM AM20	1.6–2.6	0.1	0.2	F	90	214	20	High ductility, impact strength
AM50	ASTM A50A	4.4–5.4	0.3–0.6	0.2	F	120	220	11	High energy absorbtion
AM60	ASTM AM60B	5.5–6.5	0.3–0.6	0.2	F	130	220	9	High energy absorption

[a] Amount of impurities in high purity grades is limited to max 0.01 wt% Cu, 0.002 wt% Ni, and 0.005 wt% Fe.
[b] F, as fabricated; T4, solution heat treated and naturally aged; T5, artificially aged; T6, solution heat treated and artificially aged.

cradles (front or rear) of high performance vehicles (Corvette Z06, Audi R8 GT Spyder) or front-end carrier structures of light trucks (Ford F-150). Alloys AE44 and AM60B have been utilized in these applications. The relatively high price of these components is, however, the main factor restricting wider penetration in the high volume automotive segment (Luo, Sachdev, and Powell, 2010; Luo and Sachdev, 2012). Examples of recent structural components manufactured from magnesium castings for the 2013 Range Rover are shown in Figure 5.

Currently, the most widespread application of Mg alloys in BIW and automotive closures is the instrument panel cross beam, first introduced by Audi as long ago as 1989 (US AMP Magnesium Vision 2020, 2006). To a lesser extent, components for tailgate, hood or door inner assembly reinforcement, bonnet inner parts, and frames for convertible roof tops or "bow top" are also manufactured in magnesium chiefly from AM- and AZ-type magnesium alloys using HPDC process (Blawert, Hort, and Kainer, 2004; Luo and Sachdev, 2012).

The HPDC process is predominantly used for casting of structural applications because of good die castability of common grade AZ- and AM-type magnesium alloys (Schwam, 2011). HPDC technology is able to produce complex thin-walled castings with acceptable surface quality, low porosity or cracking, and minimum need for further machining or joining. For lightweight BIW components or doors, a magnesium inner reinforcement (usually AM50 or AZ31 alloy) is integrated with sheet aluminium outer frame to form optimized hybrid lightweight structure. One of the latest examples of this method is the HPDC window frame for the Porsche Panamera that is manufactured from AM50 alloy and is attached to the cast aluminium door frame (Danisch, 2009).



(a) (b)

Figure 5. Examples of HPDC parts for 2013 Range Rover; (a) front end carrier; (b) instrumental panel cross beam.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

However, due to galvanic corrosion occurring in areas where aluminium skin panels and magnesium reinforcement make contact, additional surface protection of the magnesium is required. Sand casting of AZ91E alloy is a popular option for manufacture of the wheels of racing cars or high performance low volume passenger vehicles. Vacuum-assisted casting processes, originally developed for aluminium castings, are currently considered as an alternative to HPDC process for thick-walled castings; alternatively, the low pressure die casting (LPDC) process is especially attractive due to its ability to produce thick-walled pore-free hollow castings with high strength and improved ductility (Luo, Sachdev, and Powell, 2010). Semisolid casting (SSC) processes such as thixomolding are now accepted commercially and are broadly adopted in manufacturing of smaller parts, thanks to faster cycle times, good surface quality, and net-shaping ability (Carnahan and Decker, 2011).

5.2 Powertrain applications

Magnesium alloys in powertrain applications have recently gained popularity not only because of their good combination of low density, elevated temperature strength, and good die castability but also thanks to the progress in the die casting technology that is allowing manufacture of hybrid Al/Mg parts. AZ- or AM-type alloys could be used for castings in the powertrain system that are not exposed to temperatures above 125°C under load, such as cylinder head covers, intake manifolds, and some other less demanding components (Kulekci, 2008). In this area of application, the chief competitors for magnesium are currently thermoplastic manifolds that provide the cheapest option with similar performance and lightweight characteristic to magnesium or aluminium; however, in the luxury automotive segment, the high noise and vibration reduction combined with lightweight offered by magnesium is preferred. For other powertrain applications, operating at elevated temperatures and high external stresses for extended periods, medium grade Mg–Al-based alloys or highly creep-resistant Mg–RE-based alloys are available as a possible alternative to the currently dominant, creep-resistant Al–Si–Cu die casting grades A380 or A360.

There are two possible ways to improve the creep resistance of magnesium alloys (Pekguleryuz and Celikin, 2010): one is through alloying of the Mg–Al binary system with RE elements or alternatively relatively cheaper additions including Ca, Si, or Sr; another way is to completely eliminate aluminium and adopt complex alloying with Y, Zn, Ag, or RE, and up to 0.75 wt% Zr used for grain refinement. Previously popular creep-resistant alloys containing thorium (H-type alloys applied in military and aerospace industry) are not in automotive use because of the radioactivity of thorium. Alloying with high amounts of RE elements would lead to significantly increased costs and might also reduce die casting ability.

Medium grade Mg alloys based on the Mg–Al system with various modifications are widening the application window for magnesium die casting up to 175°C (Aghion *et al*., 2003). Typical applications of modified Mg–Al alloys are die cast or gravity cast automatic or manual transmission housings and clutch housings, engine front covers, or oil pans. Mercedes Benz 7G-tronic automatic transmission uses die-cast AS31 high purity alloy housings; Porsche Panamera V6 and V8 engines contain two camshaft covers, front cover, and the oil module from AE44 alloys (Figure 6); various VW/Audi, BMW, GM (General Motors), and Ford models are equipped with similar components based on AJ52x/62x or MRI153M/230D HPDC alloys (Kulekci, 2008). Another example of magnesium HPDC is

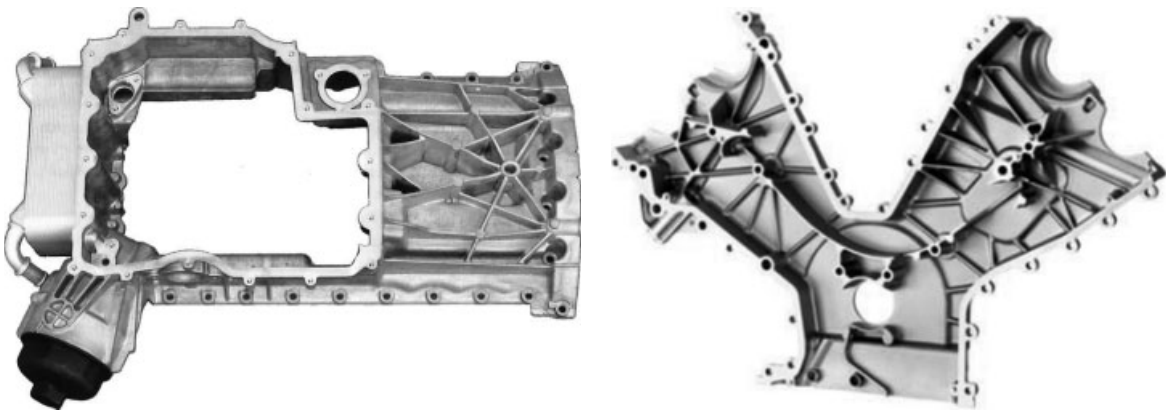


Figure 6. Examples of HPDC parts for V6/V8 engines of the Porsche Panamera; (a) oil module; (b) front cover.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 4. Medium grade die casting alloys with improved elevated temperature mechanical properties, creep resistance and good die castability for applications not exceeding 175°C.

Alloy	AMS/ASTM	Chemical Composition (wt%)				Mechanical Properties[a]			Remarks
		Al	Mn	Zn	Other	YTS (MPa)	UTS (MPa)	el. (%)	
AJ52x	ASTM AJ62	5.0	0.2	0.2	2.0 Sr	135	212	6	Die casting alloy for powertrain
AJ62x	ASTM AJ52	6.0	0.2	0.2	2.0 Sr	145	230	7	Hybrid engine block
AS21X	ASTM AS21	1.7	0.2	0.2	1.0 Si; 0.7 RE	120	210	6	Improved corrosion resistance
AS41	ASTM AS41B	4.0	0.1	0.2	1.0 Si	140	240	9	Structural components
AE42	ASTM AE42	3.5–4.5	0.1	0.2	2.0–3.0 RE	130	237	10	Structural components
AE44	ASTM 44	4.0	0.1	0.2	3.5–4.5 RE	140	245	10	Engine cradle
MRI 153M	—	7.9	0.2	0.1	1.0 Ca; 0.3 Sr	170	250	6	Low cost creep resistance alloy; good castability
MRI 230D	—	6.6	0.4	0.1	2.5 Ca; 0.3 Sr	180	235	5	Creep resistance, automatic transmission housings
ACM522	—	5.3	0.1	0.1	2.0 Ca; 2.5 RE	—	—	—	Oil pans

[a]Mechanical properties for as die-cast condition.

the hybrid cylinder block for the BMW N52 engine, where severely loaded regions of central cylinder parts (cylinder cores, coolant passages, and bolt and stud anchors) are cast in aluminium, whereas outer shell and lower crankcases are cast in the medium grade Mg–Al AJ62 alloy. Apart from up to 24% in weight savings, the Mg/Al hybrid casting technology allows reduced engine noise and vibrations (Hoeschl, Wagener, and Wolf, 2006). For specifications and tensile properties of medium grade magnesium alloys, see Table 4.

Creep-resistant, peak-aged Mg–RE–Y–Zr-based alloys, originally developed for aerospace and military equipment and containing a combination of various expensive RE elements, are currently the only option for automotive engine components working at high external loads at temperatures above 180°C. Regrettably, even though intensive research has been carried out to develop low cost creep-resistant Mg alloys suitable for high volume automotive manufacturing, existing options are still characterized by high price and poor castability. The original duo of military alloys, WE43 and WE54, has in recent years been joined by a variety of slightly cheaper equivalents, such as MRI201S and MRI202S from Dead Sea Magnesium or several grades from Magnesium Elektron: Mg–Zn–RE-based alloys including Elektron 21, ZRE1, and RZ5 and Mg–Ag–RE-based alloys including EQ21 and MSR-B. Finally, Advanced Magnesium Technologies developed Mg–RE-based alloys AM-SC1 and AM-HP2 and successfully demonstrated their applicability for gravity and die-cast engine blocks (Pekguleryuz and Celikin, 2010). Specifications, chemical composition, and tensile properties of creep-resistant magnesium alloys are listed in Table 5; mechanical properties of selected magnesium alloys at elevated temperature are listed in Table 6.

Owing to the poor die-casting ability of creep-resistant magnesium alloys without aluminium, alternative casting methods such as gravity permanent mould casting (GPMC), squeeze casting, or SSC have to be employed. To reduce porosity and improve the creep resistance of alloys suitable for applications in cylinder heads or engine blocks, the squeeze casting process is considered and might gain more interest from automotive manufactures in future, even though squeeze casting is not as productive as HPDC (Kasprzak, Lo, and Jekl, 2011).

It remains to be seen if any of these alloys will be able to find utilization in a volume vehicle market. To the authors' knowledge, apart from aerospace castings for helicopter gearboxes and other high performance military applications, this group of alloys is currently being used only for engine blocks of racing cars.

6 APPLICATIONS OF WROUGHT MAGNESIUM

The development in Germany in about 1925 of flux covers for molten magnesium enabled production of good quality castings and, in turn, allowed development of magnesium alloy wrought products (Brown, 2002). In the period from the 1920s through the late 1950s, magnesium alloy forgings, extrusions, and sheet found widespread use in major structural roles in a large number of aircrafts and performed

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 5. High performance creep resistant alloys for components exposed to long lasting external stresses at temperatures above 180°C; significantly more expensive alloys with reduced castability.

Alloy	AMS/ASTM	Chemical Composition (wt%)				Condition[a]	Mechanical Properties			Remarks
		Zn	RE	Zr	Other		YTS (MPa)	UTS (MPa)	el. (%)	
RZ5	AMS 4439G; ZE41A	4.2	1.3	0.7	—	T5	135	200	4	Good castability
ZC63A	ASTM ZC63A-T6	6.0	—	—	2.7 Cu; 0.5 Mn	T6	170	245	4	Sand casting
ZRE1	EZ33G	2.5	2.5–4.0 rich Nd	0.6	—	T5	94	161	4	Sand casting
						T5	100	155	3	Chill casting
EQ21B	AMS 4417B	—	2	0.6	1.5 Ag	T6	175	240	2	Sand casting
MRS-B	AMS 4418H; QE22A	—	2	0.6	2.5 Ag	F	124	165	4	Sand casting/mold casting
						T6	205	266	2	
Elektron 21	AMS 4429A	0.3	2.8Nd + 1.4Gd	0.6	—	F	155	228	4	Sand casting
						T6	170	280	5	
MRI 202S	—	—	Nd rich RE	0.6	Y	T6	150	250	7	Sand casting
MRI 201S	—	—	Nd rich RE	0.6	Y	T6	170	260	6	Sand casting; cost effective equivalent to WE43
WE43C	AMS 4427C	0.2	2.4–4.4	0.4–1.0	4.3 Y	F	128	176	4	Sand casting
						T6	162	250	2	Sand and squeeze-casting
WE54A	AMS 4426B	0.2	3.5	0.4–1.0	5.5 Y	T6	185	255	2	Sand or squeeze-casting; high strength at el. temperatures

[a] F, as fabricated; T4, solution heat treated and naturally aged; T5, artificially aged; T6, solution heat treated and artificially aged.

Encyclopedia of Automotive Engineering. Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

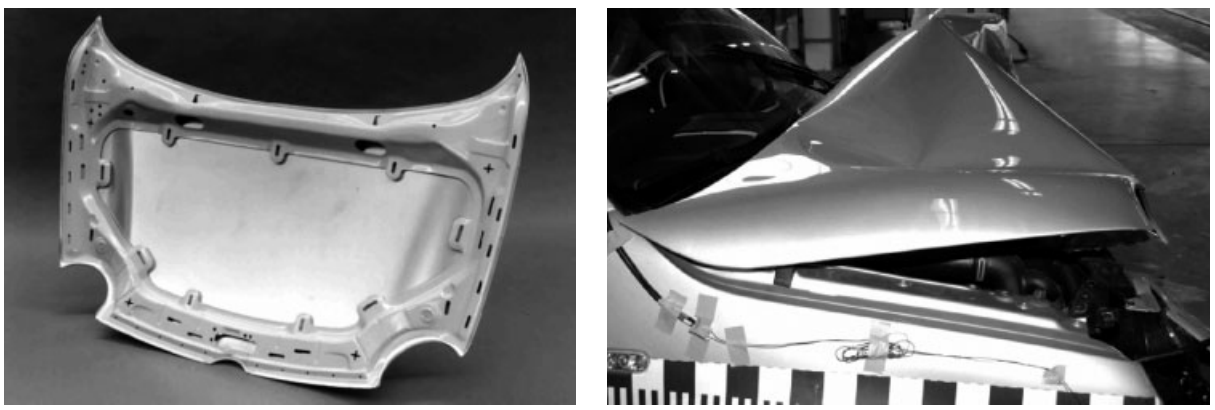
Table 6. Tensile properties of selected magnesium alloys at 175°C.

Alloy	Tensile Properties at 175°C		
	YTS (MPa)	UTS (MPa)	el. (%)
AZ91D	89	138	21
AJ52x	100	141	18
AJ62x	103	143	19
AS21X	78	110	23
AS41	85	127	25
AE42	91	130	23
AE44	110	150	25
MRI 153M	125	172	22
MRI 230D	145	178	18
ACM522	132	152	9
ZE41-T5	120	153	18
Elektron 21-T6	170	250	-
MRI 202S-T6	145	220	13
MRI 201S-T6	170	235	13
WE43-T6	175	210	8
WE54-T6	190	248	6
AM-SC1	114	181	11
AM-HP2	132	136	5

satisfactorily in these roles. Wrought magnesium alloys were also used in several quite mundane land vehicles, such as the body of the Metro-Lite delivery truck that was largely constructed from magnesium sheet and extrusions. However, from the late 1950s until the present, there has been virtually no significant use of wrought magnesium in automobile construction.

The demands on current vehicle builders for performance, safety, and comfort are such that despite its attractions of low density, high specific strength, and good dent resistance, the drawbacks of very limited cold forming capability, galvanic corrosion behavior, availability, price, and price instability have been perceived as too great to warrant its use in volume cars. Use of magnesium sheets in automobile structures has been restricted to a small number of limited-series vehicles. Nevertheless, numerous development programs have been underway for, at least, 10 years. Objectives include developing alloys with improved cold formability and improved corrosion behavior, developing forming and joining techniques, developing protective coatings, and considering recycling procedures. Friedrich and Schumann (2001) (of Volkswagen) suggested that there would be a "New Age of Magnesium" in the automotive industry and that this would include castings, forgings, extrusions, and sheet and they also suggested that the necessary development programs would take more than 10 years to achieve practical implementation. Sheet applications were judged farthest from implementation and in their then state failed to meet either corrosion requirements or the surface quality required for exterior panels. Nevertheless, a bonnet structure had been satisfactorily formed from AZ31B sheet and this had performed in a manner similar to aluminium in an offset deformable barrier crash test (Moll *et al*., 2005) (Figure 7).

Progress toward actual use of magnesium alloy sheet in series production seems to be being made by GM who have recently announced (GM Press Release, 2012) that they have developed a forming process for magnesium sheet that is conducted at 450°C and combined with a proprietary anticorrosion treatment. They have formed a trunk lid inner panel and subjected it to severe tests without problems. It seems probable that the experimental panel was formed by the so-called Quick Plastic Forming process that GM developed for forming aluminium panels (Krajewski and Schroth, 2011). Krajewski is quoted (Magnesium Monthly Review, 2012) as saying that magnesium sheet inner door



(a) (b)

Figure 7. Volkswagen Lupo AZ31B trial bonnet; (a) as warm formed; (b) after crash test. (Reproduced from Moll *et al*. (2005). © Wiley-VCH Verlag GmbH & Co. KGaA.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

and trunk lid panels will be installed on 50 test vehicles by the end of the month (October 2012).

In the United Kingdom, the Morgan Motor Company is managing a collaborative program with funding from the UK Technology Strategy Board to investigate use of wrought magnesium as a chassis construction material. The partners are Magnesium Elektron, Superform Aluminium, Penso Consulting, and Coventry University. Initially, fears centered around flammability but were dispelled by the absence of ignition during severe burning tests and it is also believed that the corrosion behavior is superior to that of steel and no worse than that of high strength aluminium, provided due note is taken of galvanic issues. It is possible that the forthcoming Morgan Eva GT Coupé will have wrought magnesium in its chassis as a consequence of this program (Eureka Magazine, 2012).

Of the potential magnesium sheet manufacturers, POSCO is probably nearest to being able to produce production quality sheet in production quantities. At the conference that POSCO hosted to celebrate the commencement of production of magnesium ingot at the Okgye plant, they reiterated that they plan to deliver magnesium sheet to Renault, Samsung, and Hyundai and Kia Motors Corporation (HKMC) by March 2013. HKMC uses magnesium seat frames fabricated from castings in some of its luxury vehicles and has also been developing seat frames employing magnesium extrusions (Kim and Han, 2008). POSCO have themselves been developing magnesium seat frames that they hope to export to European vehicle builders (Japan Metal Bulletin, 2012).

7 THE FUTURE

7.1 Sourcing of magnesium

Before the year 2000, the great majority of primary magnesium production was conducted in the west and most was extracted using the relatively environmentally friendly electrolytic process. Western production (plus CIS) was about 415,000 tonnes per annum and Chinese production 115,000 tonnes per annum, the Chinese material being largely produced by the labor-intensive and environmentally unfriendly Pidgeon process. Over the next decade, there was huge growth in both the volume produced in China and in China's production capacity (Patzer, 2010) such that by the end of 2009 over 75% of global magnesium production was in China (Figure 8). Western producers could not match Chinese prices and the United States was left with US Magnesium LLC as their sole primary producer. By 2011, it has been estimated that Chinese primary magnesium capacity had risen to almost 2 million tonnes per annum with an actual production of 600,000 tonnes (The CM Group, 2012). Although the great majority of this capacity was for production by the Pidgeon process, the growth was accompanied by significant improvement in the efficiency of the process, by the operation of generally larger plants and by the closure of many small plants. These changes in the location of primary production initially led to a considerable reduction in the metal price, but in the intervening years, increasing labor, energy, and material costs in China have resulted in price increases and price uncertainty. However, the development by Qinghai Salt Lake

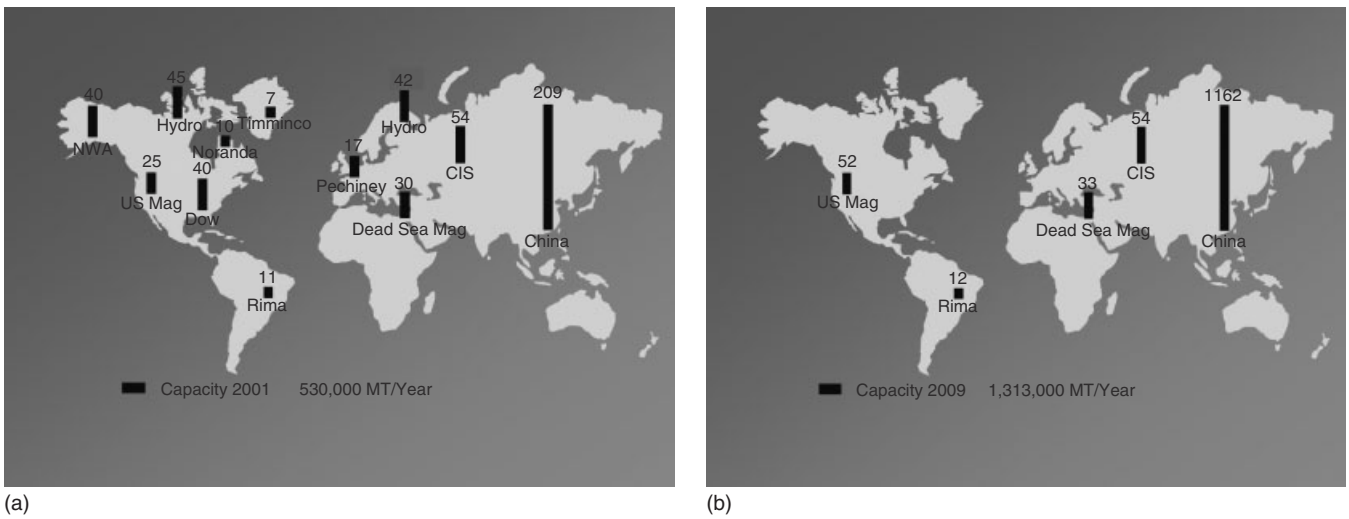


Figure 8. Global magnesium primary production capacity; (a) 2001; (b) 2009. (Reproduced from Patzer, 2010. © John Wiley & Sons, Inc.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

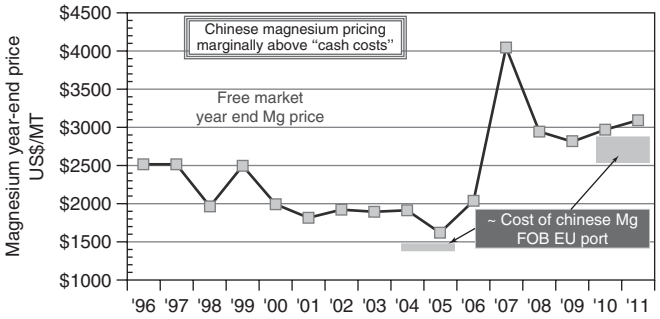


Figure 9. Trends in free market Mg price. From Zuliani and Reeson (2012). (Reproduced by permission of the International Magnesium Association.)

Group of an electrolytic plant employing hydro power to extract magnesium from Qarhan Lake brine, with an eventual capacity of 100,000 tonnes per annum, should result in materials of lower cost (The Hatch Report, 2012). Figure 9, taken from Zuliani and Reeson (2012), shows the changes in magnesium price over the past 15 years and emphasizes its instability.

In 2005, the US International Trade Commission (ITC) imposed punitive tariffs on imports of magnesium to the United States from China and Russia. This met with protests from US magnesium die casters who claimed that their volumes were being reduced as a direct result of the tariffs. In consequence, after review, the ITC revoked the punitive tariffs on Russian imports but retained them for Chinese imports. In turn, this decision was appealed by the US Magnesium LLC with the argument that the ITC should have considered the cumulative effects of imports from China and Russia rather than considering each country separately. Certainly, the result is that the ratio of magnesium price to aluminium price is far more favorable to magnesium in China and in Europe and favorable to aluminium in the United States (The CM Group, 2012) and both the number of magnesium die casters in the United States and the volume of their business have progressively declined since 2005 (Twarog, 2012).

Despite the closure of so many western primary magnesium plants and China's current dominance, new plants have opened outside China and more are under consideration. CVM Minerals in Malaysia set up a Pidgeon plant in Perak to utilize the high grade dolomite that is found there and this started operation in late 2010. When fully operational, this plant is expected to produce 15,000 tonnes per annum (www.cvmminerals.com/ceo.html) but, in due course, a second 15,000 tonnes per annum line is planned.

POSCO, also, has constructed an extraction plant and although delays were experienced because of a dispute over electricity supply cables (Metals Week, 2012b), these have now been resolved and commercial production has recently started (Magnesium Monthly Review, 2012). The plant extracts magnesium from locally mined dolomite and has an initial capacity of 10,000 tonnes per annum but expansion of output to 100,000 tonnes per annum by 2018 is planned (Japan Metal Bulletin, 2011). Extraction employs POSCO's own development of vertical thermal technology with ferrosilicon imported from China.

7.2 Continuing growth

A large proportion of the substantial growth in primary magnesium production has been driven by the growth in the utilization of magnesium alloy castings in motor cars. Nevertheless, the actual percentage of magnesium in a typical current motor car remains very small (<1%), whereas the USAMP study of 2006 suggested that by 2020, 155 kg of magnesium could replace 285 kg of ferrous plus aluminium structure, provided that the significant necessary development programs were conducted (US AMP Magnesium Vision 2020, 2006). In fact, if all of the magnesium components that were in use on a variety of north American cars at the time of the USAMP investigation had been installed on a single vehicle, the weight saving would have been 175 kg, so that the 155 kg target for 2020 does not seem unduly optimistic. Also, the 175 kg would not have included any wrought material. In the intervening years, large development programs have been continuing in Asia, the United States, and Europe and there is considerable collaboration between continents, countries, and companies. The Korean government is allocating 1 trillion won (US$864 million) to its World Premier Materials framework between 2010 and 2018. This consists of 10 consortia involving over 220 enterprises and institutions but, in particular, a "Magnesium Research and Development Project for Ultralight Vehicles" to be led by POSCO but involving Helmholtz-Zentrum Geesthacht, Wrought Magnesium Alloys Group. Major programs to develop sheet continue at MgF in collaboration with the Technical University Freiberg. In the United States, three of the DOE's (Department of Energy) national laboratories have magnesium-related activities. Most of the world's major car builders have magnesium programs. Each year, the annual TMS gathering includes a Magnesium Technology Symposium at which some 100 papers are presented and, in addition, each year there are several other national, or international, conferences devoted to the development of magnesium. Thus, given the huge effort being devoted to magnesium alloy technology, there seems reason to be optimistic that satisfactory solutions to outstanding problems should be found.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In a world where there are growing environmental concerns, however, magnesium's carbon footprint is a cause for concern. Neither the LCA performed by Ehrenberger *et al*. (2008) nor that performed by Das *et al*. (2009) showed magnesium in a particularly good light, although the point was made that the situation would probably have been considerably better had a scrap recycling system been in place. The Pidgeon process invites particularly unfavorable comments for its environmental damage (Ingarao, Di Lorenzo, and Micari, 2011; Cherubinia, Raugei, and Ulgiati, 2008) but, as it is far and away the dominant process currently employed for magnesium extraction and the cheapest, it seems unrealistic to suggest that utilization of magnesium components in motor vehicles should await availability of some, yet to be proved, cheaper and more environmentally friendly process.

It is now well established that a wide range of structural castings can be satisfactorily made from magnesium alloys while conferring significant weight savings. There seems every reason to suppose that both the range of such castings and the number of vehicle builders employing them will increase. As far as wrought alloys are concerned, the technology to produce industrial quality sheet via TRC now exists at POSCO and MgF so that the obstacles to implementation of magnesium sheet in vehicle structures seem, very largely, to be commercial.

REFERENCES

Abbott, T.B., Easton, M.A., and Caceres, C.H. (2003) Designing with magnesium in *Handbook of Mechanical Alloy Design* (eds G.M. Totten, L. Xie, K. Funatani), Marcel Dekker, New York, pp. 487–538.

Aghion, E., Bronfin, B., Von Buch, F., *et al*. (2003) Newly developed magnesium alloys for powertrain applications *JOM*, **55** (11), 30–33.

Avedesian, M.M. and Bake, H. (eds) (1999) *Magnesium and Magnesium Alloys ASM Specialty Handbook*, ASM International, Materials Park, OH.

Blawert, C., Hort, N., and Kainer, K.U. (2004) Automotive applications of magnesium and its alloys. *Transactions of the Indian Institute of Metals*, **57** (4), 397–408.

Brown, R.E. (2002) Magnesium wrought and fabricated products yesterday, today, and tomorrow in *Magnesium Technology 2002* (ed. H.I. Kaplan), TMS, Warrendale, PA.

Carnahan, R.D. and Decker, R.F. (2011) Thixomolding in *Technology for Magnesium Castings: Design, Products & Applications* (eds M. Sahoo and S.P. Thomas), American Foundry Society, Schaumburg, IL, USA, pp. 223–236.

Cherubini, F., Raugei, M., and Ulgiati, S. (2008) LCA of magnesium production: technological overview and worldwide estimation of environmental burdens *Resources, Conservation and Recycling*, **52** (8–9), 1093–1100.

Choo, D., Kim, J., Kim, I., *et al*. (2012) Development of Wide Strip Casting Process of Mg Alloys. *69th Annual World Magnesium Conference*, San Francisco.

Cole, G.S. (2011a) Design and application of cast components in *Technology for Magnesium Castings: Design, Products & Applications* (eds M. Sahoo and S.P. Thomas), Schaumburg, IL, USA, American Foundry Society, pp. 237–262.

Cui, Q. and Ohori, K. (2000) Grain refinement of high purity aluminium by asymmetric rolling *Materials Science and Technology*, **16** (10), 1095–1101.

Danisch, R. (2009) Die technischen Highlights des Porsche Panamera, ATZ online, http://www.atzonline.de/Aktuell/Nachrichten/1/9381/Die-technischen-Highlights-des-Porsche-Panamera.html (accessed 30 October 2013).

Das, S., Dubreuil, A., Bushi, L., and Tharumarajah, A. (2009) A life cycle assessment of a magnesium automotive front end in *Magnesium Technology 2009* (eds E. Nyberg, S.R. Agnew, N.R. Neelameggham, M. Pekguleryuz), John Wiley & Sons, Inc, Hoboken, NJ, pp. 179–184.

Duygulu, O., Ucuncuoglu, S., Oktay, G., *et al*. (2009) Development of rolling technology for twin roll cast 1500 mm wide magnesium AZ31 alloy in *Magnesium Technology 2009* (eds E. Nyberg, S.R. Agnew, N.R. Neelameggham, M. Pekguleryuz), John Wiley & Sons, Inc, Hoboken, NJ, pp. 379–384.

Ehrenberger, S.I., Schmid, S.A., Song, S., and Friedrich, H.E. (2008) Status and potentials of magnesium production in China: life cycle analysis focussing on CO2eq emissions. *65th Annual World Magnesium Conference*, Warsaw, Poland.

United States Environmental Protection Agency, EPA (2008) Alternatives to SF_6 for Magnesium Melt Production, http://www.epa.gov/magnesium-sf6/documents/magbrochure_english.pdf (accessed 30 October 2013).

Eureka Magazine 2012 September, pp. 12–14.

Evans, J.W. (2007) The evolution of technology for light metals over the last 50 years: Al, Mg, and Li *JOM*, **59** (2), 30–38.

Fechner, D., Hort, N., and Kainer, K.U. (2009) Magnesium recycling system prepared by permanent mould- and high pressure die casting in *Magnesium Technology 2009* (eds E. Nyberg, S.R. Agnew, N.R. Neelameggham, M. Pekguleryuz), John Wiley & Sons, Inc, Hoboken, NJ, pp. 111–116.

Friedrich, H. and Schumann, S. (2001) Research for a "new age of magnesium" in the automotive industry *Journal of Materials Processing Technology*, **117** (3), 276–281.

GM Press Release (2012) http://media.gm.com/media/us/en/gm/news.detail.html/content/Pages/news/us/en/2012/Oct/1023_GM_Magnesium.html (accessed 30 October 2013).

Grimes, R. (2011) Superplastic forming of magnesium alloys in *Superplastic Forming of Advanced Metallic Materials* (ed. G. Giuliano), Woodhead Publishing, Cambridge, pp. 304–326.

Hantzsche, K., Bohlen, J., Wendt, J., *et al*. (2010) Effect of rare earth additions on microstructure and texture development of magnesium alloy sheets *Scripta Materialia*, **63** (7), 725–730.

The Hatch Report (2012) www.hatch.ca/News_Publications/Hatch_Report/HR. . . /qslic.html (accessed 30 October 2013).

Herling, D.R. and Carpenter, J.A. (2005) *Cost Assessment of Emerging Magnesium Sheet Production Methods, PNNL-15368*, http://www1.eere.energy.gov/vehiclesandfuels/pdfs/alm_05/2j_herling.pdf (accessed 30 October 2013).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Hoeschl, M., Wagener, W., and Wolf, J. (2006) BMW's magnesium-aluminium composite crankcase, state-of-the-art light metal casting and manufacturing. SAE Technical Paper 2006-01-0069, http://papers.sae.org/2006-01-0069 (accessed 30 October 2013).

Ingarao, G., Di Lorenzo, R., and Micari, F. (2011) Sustainability issues in sheet metal forming processes: an overview *Journal of Cleaner Production*, **19** (4), 337–347.

Japan Metal Bulletin (2011) http://www.japanmetalbulletin.com/?p=18933 (accessed 30 October 2013).

Japan Metal Bulletin (2012) http://www.japanmetalbulletin.com/?p=19501 (accessed 30 October 2013).

Kasprzak, J., Lo, J., and Jekl, J. (2011) Squeeze casting in *Technology for Magnesium Castings: Design, Products & Applications* (eds M. Sahoo and S.P. Thomas), American Foundry Society, Schaumburg, IL, USA, pp. 223–236.

Kawalla, R., Oswald, M., Schmidt, C., *et al*. (2008) New technology for the production of magnesium strips and sheets *Metalurgija*, **47** (3), 195–198.

Kaya, A., Duygulu, O., Ucuncuoglu, S., *et al*. (2008) Production of 150 cm wide AZ31 magnesium sheet by twin roll casting' *Transactions of Nonferrous Metals Society of China*, **18**, 185–188.

Kim, J.J. and Han, D.S. (2008) Recent development and applications of magnesium alloys in the Hyundai and Kia motors corporation *Materials Transactions*, **49** (7), 894–897.

Kim, W.J., Kim, M.J., and Wang, J.Y. (2009) Superplastic behaviour of a fine-grained ZK60 magnesium alloy processed by high-ratio differential speed rolling *Materials Science and Engineering: A*, **527** (1–2), 322–327.

The Korea Herald (2011) POSCO Steps Up Resources Development. http://nwww.koreaherald.com/view.php?ud=20110127000714 (accessed 30 October 2013).

Krajewski, P. and Schroth, J. (2011) Quick plastic forming of aluminium alloys in *Superplastic Forming of Advanced Metallic Materials* (ed. G. Giuliano), Woodhead Publishing, Cambridge, pp. 227–302.

Kulekci, M.K. (2008) Magnesium and its alloys applications in automotive industry *International Journal of Advanced Manufacturing Technology*, **39** (9–10), 851–865.

Liang, D. and Cowley, C. (2004) The twin-roll strip casting of magnesium *JOM*, **56** (5), 26–28.

Luo, A.A., Sachdev, A.K., and Powell, B.R. (2010) Advanced casting technologies for lightweight automotive applications *China Foundry*, **7** (4), 463–469.

Luo, A.A. and Sachdev, A.K. (2012) Application of magnesium alloys in automotive engineering in *Advances in Wrought Magnesium Alloys* (eds C.J. Bettles and M.R. Barnett), Woodhead Publishing, Cambridge, pp. 393–422.

Magnesium Monthly Review (2012) **41** (9).

Metals Week (2012a) vol. 83, June 4, 2012.

Metals Week (2012b) vol. 83, July 23, 2012) .

Moll, F., Mekkaoui, M., Schumann, S., and Friedrich, H. (2005) Application of Mg Sheets in Car Body Structures. *Magnesium: Proceedings of the 6th International Conference Magnesium Alloys and Their Applications* (ed. K.U. Kainer), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG.

Mordike, B.L. and Ebert, T. (2001) Magnesium—properties—applications—potential *Materials Science and Engineering: A*, **302** (1), 37–45.

Oak Ridge National Laboratory (2012) http://www.ornl.gov/info/press_releases/get_press_release.cfm?ReleaseNumber=mr20120620-00 (accessed 30 October 2013).

Park, W.J., Kim, J.J., Kim, I.J., and Choo, D. (2011) Wide strip casting technology of magnesium alloys in *Magnesium Technology 2011* (eds W.H. Sillekens, S.R. Agnew, N.R. Neelameggham, S.N. Mathaudhu), John Wiley & Sons, Inc, Hoboken, NJ, pp. 143–146.

Passoni, R., Romanowski, C., Romano, E., and Cattelino, P.M. (2012) Highlights on the New Fata Hunter Reversing Mill for Magnesium Alloys. *20th Magnesium Automotive and User Seminar*, Dusseldorf, Germany.

Patzer, G. (2010) The magnesium industry today: the global perspective in *Magnesium Technology 2010* (eds S.R. Agnew, N.R. Neelameggham, E. Nyberg, W.H. Sillekens), John Wiley & Sons, Inc, Hoboken, NJ, pp. 85–90.

Pidgeon, L.M. and Alexander, W.A. (1944) Thermal production of magnesium, pilot plant studies on the Retort ferrosilicon process *Transactions of AIME*, **159**, 315–352.

Pekguleryuz, M. and Celikin, M. (2010) Creep resistance in magnesium alloys *International Materials Reviews*, **55** (4), 197–217.

Polmear, I.J. (1999) *Light Alloys*, 4th edn, Butterworth-Heinemann, Oxford.

Prentice, L., Poi, N. and Haque, N. (2010) Life Cycle Assessment of Carbothermal Production of Magnesium in Australia. *67th Annual World Magnesium Conference*, Hong Kong, pp. 78–82.

Prentice, L.H. and Haque, N. (2012) MagSonic™ thermal technology compared with the electrolytic and Pidgeon processes in *Magnesium Technology 2012* (eds S.N. Mathaudhu, W.H. Sillekens, N.R. Neelameggham, N. Hort), John Wiley & Sons, Inc, Hoboken, NJ, pp. 37–41.

Ramakrishnan, S. and Koltun, P. (2004) A comparison of the greenhouse impacts of magnesium produced by electrolytic and Pidgeon processes in *Magnesium Technology 2004* (ed A.A. Luo), John Wiley & Sons, Inc, Hoboken, NJ, pp. 173–178.

Randman, D., Davis, B., Alderman, M.L., *et al*. (2011) The effect of rare earth elements on the texture and formability of shear rolled magnesium sheet in *Magnesium Technology 2011* (eds W.H. Sillekens, S.R. Agnew, N.R. Neelameggham, S.N. Mathaudhu), John Wiley & Sons, Inc, Hoboken, NJ, pp. 187–193.

Robinson, S. (2011) Introduction to the magnesium casting industry in *Technology for Magnesium Castings: Design, Products & Applications* (eds M. Sahoo and S.P. Thomas), American Foundry Society, Schaumburg, IL, USA, pp. 1–7.

Sadayapan, K. and Luo, A.A. (2011) Physical metallurgy in *Technology for Magnesium Castings: Design, Products & Applications* (eds M. Sahoo and S.P. Thomas), American Foundry Society, Schaumburg, IL, USA, pp. 9–27.

Schumann, S. and Friedrich, H.E. (2006) Engineering requirements, strategies and examples in *Magnesium Technology* (eds H.E. Friedrich and B.L. Mordike), Springer, Berlin Heidelberg, pp. 499–632.

Schwam, D. (2011) High pressure die casting in *Technology for Magnesium Castings: Design, Products & Applications* (eds M.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Sahoo and S.P. Thomas), American Foundry Society, Schaumburg, IL, USA, pp. 183–196.

The CM Group (2012) The Global Mg Industry in 2011—The Impact of Chinese Production, Costs and Shipments. *69th Annual World Magnesium Conference*, San Francisco.

Twarog, D. (2012) Magnesium Die Castings—A Solution to Reducing the Carbon Footprint of Automobiles. *IMA's Applications Seminar*, Ann Arbor, MI.

US AMP Magnesium Vision 2020 (2006) www.uscar.org/commands/files_download.php?files_id=99 (accessed 30 October 2013).

Willekens, J.M.A. (2012) Primary Magnesium Production Based on Other Sources than Dolomite. *20th Magnesium Automotive and User Seminar*, Dusseldorf, Germany.

Zuliani, D.J. and Reeson, D. (2012) *Developments in the Zuliani process for Gossan resources' magnesium project*, Magnesium Investing News, http://magnesiuminvestingnews.com/files/2012/04/IMA-2012-Paper-FINAL-Developments-in-the-Zuliani-Process-for-Gossan-Resourcess-Magnesium-Project-1.pdf (accessed 30 October 2013).

FURTHER READING

Bettles, C.J. and Barnett, M.R. (2012) *Advances in Wrought Magnesium Alloys*, Woodhead Publishing, Cambridge.

Blawert, C., Hort, N., and Kainer, K.U. (2004) Automotive applications of magnesium and its alloys. *Transactions of the Indian Institute of Metals*, **57** (4), 397–408.

Cole, G.S. (2011b) Challenges facing the magnesium industry in *Technology for Magnesium Castings: Design, Products & Applications* (eds M. Sahoo and S.P. Thomas), Schaumburg, IL, USA, American Foundry Society, pp. 263–289.

Duan, H., Yan, C., and Wang, F. (2007) Effect of electrolyte additives on performance of plasma electrolytic oxidation films formed on magnesium alloy AZ91D. *Electrochimica Acta*, **52** (11), 3785–3793.

Friedrich, H.E. and Mordike, B.L. (2006) *Magnesium Technology*, Springer, Berlin Heidelberg.

Hort, N., Mathaudhu, S.N., Neelameggham, N.R., and Alderman, M. (eds) (2013) *The Magnesium Technology Symposia from the TMS Annual Meetings, John Wiley & Sons*, Hoboken, New Jersey.

King, J.F. (2006) Technology of magnesium and magnesium alloys in *Magnesium Technology* (eds H.E. Friedrich and B.L. Mordike), Springer, Berlin Heidelberg, pp. 219–430.

Kulekci, M.K. (2008) Magnesium and its alloys applications in automotive industry. *International Journal of Advanced Manufacturing Technology*, **39** (9–10), 851–865.

Luo, A.A. (2004) Recent magnesium alloy development for elevated temperature applications. *International Materials Reviews*, **49** (1), 13–30.

Mathaudhu, S.N., Sillekens, W.H., Neelameggham, N.R., and Hort, N. (eds) (2012) *The Magnesium Technology Symposia from the TMS Annual Meetings, John Wiley & Sons*, Hoboken, New Jersey.

Mordike, B.L. and Ebert, T. (2001) Magnesium—properties—applications—potential. *Materials Science and Engineering A*, **302** (1), 37–45.

Pekguleryuz, M. and Celikin, M. (2010) Creep resistance in magnesium alloys. *International Materials Reviews*, **55** (4), 197–217.

Powell, B.R., Luo, A.A., and Krarajewski, P.E. (2012) *Magnesium alloys for lightweight powertrains and automotive bodies, in* Advanced Materials in Automotive Engineering, Woodhead Publishing Ltd, Cambridge, UK, pp. 150–209.

Sahoo, M. and Thomas, S.P. (2011) *Technology for Magnesium Castings: Design, Products & Applications*, American Foundry Society, Schaumburg, USA.

Sillekens, W.H., Agnew, S.R., Neelameggham, N.R., and Mathaudhu, S.N. (eds) (2011) *The Magnesium Technology Symposia from the TMS Annual Meetings, John Wiley & Sons*, Hoboken, New Jersey.

Song, G. (2005) Recent progress in corrosion and protection of magnesium alloys. *Advanced Engineering Materials*, **7** (7), 563–586.

Vinarcik, E.J. (2002) *High Integrity Die Casting Processes*, John Wiley & Sons, New York.

Vinarcik, E.J. (2002) *High Integrity Die Casting Processes,* John Wiley & Sons, New York.

Westengen, H. and Aune, T.K. (2006) Magnesium casting alloys in *Magnesium Technology* (eds H.E. Friedrich and B.L. Mordike), Springer, Berlin Heidelberg, pp. 145–218.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto273
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Plastic Trim

Kylash Makenji and Ruth Cherrington

University of Warwick, Coventry, UK

1 Introduction 1
2 Plastics Used to Manufacture the Interior Trim 2
3 Performance Testing 2
4 Fitting Plastic Trim to the BIW 2
5 Manufacturing Methods 3
6 Decorating Vehicle Interior Trim 3
7 Future Trends 5
References 5
Further Reading 6

1 INTRODUCTION

The inherent features of plastics such as good mechanical properties, light weight compared to metals, ease of decoration, and ease of fabrication in creating aesthetically pleasing but technically acceptable surfaces that are the major drivers for use in vehicle interiors. Manufacturers have used the materials to attain quality finish and aesthetic styling of the interior trim to set them apart from their competitors. Other considerations include light weight, lower tooling costs for high volumes, and the possibility for them to be fabricated as single complex components. The plastic content of a vehicle is typically 8% of the total weight. The interior comprises ∼48%, the exterior ∼27%, and the under hood ∼14% of the total plastics used (Vidhyaa, 2012). This chapter looks at the plastics that make up the vehicle interior.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto275
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

There are many different components used to make up the interior trim of the cabin and some premium manufactures use traditional materials such as wood, leather, and metal inserts to enhance the aesthetics, as illustrated in Figure 1. Each component is required to pass stringent legal performance and durability standards, be produced at the right cost, and meet the customer's expectations of quality.

The vehicle interior is typically zoned in a number ways to enable the manufacturer to focus the styling around different features; this approach enables them to select the appropriate materials and manufacturing methods in their facilities or their supply chain. This chapter discusses the different zones or areas of the vehicle trim, the materials, and the manufacturing methods employed.

1.1 The zones of the vehicle interior trim

The typical zones of vehicle interior trim identified by manufacturers are outlined below:

1. The roof pillars and body-in-white (BIW). Traditionally, these areas are aesthetically clad to cover the unsightly bodywork.
2. The instrument panel (IP) and central console. The IP houses all of the vital user dials and gages, providing the user with intelligent vehicle data. Systems such as in-car entertainment, climate controls, and safety airbags are also found here. The central console houses driver controls such as gearshifts and emergency handbrake, as well as drink holders and similar fixtures. Plastic versions of these were first introduced into vehicles in the mid-1960s (Maxwell, 1994).
3. The door panels. There is a need to clad these surfaces to make them look attractive, while housing audio speakers,



Figure 1. Vehicle interior.

locking mechanisms, and occupant impact protection beams.

There are others areas of "soft" trim in the vehicle such as the seats, carpets, and headliners. However, while these are vital to the look and feel of the vehicle interior, they will not be considered in this chapter.

2 PLASTICS USED TO MANUFACTURE THE INTERIOR TRIM

There are a number of different types of plastics used in the makeup of the vehicle interior. These include polypropylene (PP), acrylonitrile butadiene styrene (ABS), polycarbonate (PC), polyvinyl chloride (PVC), polyethylene (PE), and thermoplastic poly olefin (TPO). In 2005, approximately 275 million kilograms (kg) of commodity thermoplastic was used in the interior of vehicles, this is predicted to grow by 2.8% per annum (Vidhyaa, 2012).

The drive to improve styling and to meet the durability performance standards and end-of-life recycling targets has led to an increasing popularity in the use of PP in vehicle interiors. PP is a cost-effective commodity semicrystalline thermoplastic that can easily be formed to create the high quality surface and color finish required by the automotive industry. It replaces ABS, which has inferior performance and is more costly. It is also replaces PVC, which was used considerably in vehicle interiors but due to issues with fogging and cracking, its use has gradually declined (Helps, 2001).

The IP is commonly made from TPO. TPO is a plastic blend, usually consisting of fractions of PP, PE, and block copolymer PP. It is applied to a hard or soft PP or ABS substrate to provide a soft-touch feel (Helps, 2001) that is desirable to the customer. Some of the inserted panels fitted to the IP may be manufactured from ABS or PC/ABS blends for styling, decoration, or performance requirements.

3 PERFORMANCE TESTING

Interior automotive components must comply with aesthetic and safety standards. Vehicle manufacturers test the performance of components to ensure that they are able to withstand the user and environmental requirements it is likely to experience during its lifetime.

The materials are subjected to rigorous environmental and substance-resistance testing to establish their suitability for the demands of the everyday motorist. The trim is subjected to sunlight (112–3609 kJ/m^2, depending on location), heat (−40 to 100°C), humidity (5–95% RH), impact (4.5 kg at 610 mm), and chemical substance testing (Ford, 2009). The plastic trim panels will also have to comply with head impact testing, based on Federal Motor Vehicle Safety Standard 201 (FMVSS 201, 1989), regulated by the National Highway Traffic Safety Administration (NHTSA). All areas contactable by the form within the "head impact area" must be tested. The test is used to determine that the trim is constructed using energy-absorbing material that deflects or collapses to within a specified limit of a rigid subsurface without permitting contact with any rigid material.

4 FITTING PLASTIC TRIM TO THE BIW

The trim is designed to consider how it is assembled within the BIW. On the assembly line, the plastic trim has to be fitted in a few seconds, without any damage to the visible exterior surface. It also has to meet a service requirement for removal and refitting such as access, service, or repair, up to five times during the life of a vehicle. Originally, the plastic trim would have been fitted with screws; however, this is unsightly and requires more equipment and time to assemble. Most modern plastic trim is fitted using blade, low or high friction metal clips that are preassembled to the trim on the interior surface, as illustrated in Figure 2. Low friction clips are used where the frequent removal of the trim is required. This system is also used to install an airbag behind the trim. The trim is ejected during airbag deployment; in this arrangement, the trim is tethered to prevent injury to the occupant. These methods of trim retention enable the vehicle interior to have clean styling lines.

Plastic trim linearly expands and contracts ($25–200 \times 10^{-6}$) more than the steel BIW ($10–13 \times 10^{-6}$) that it is attached to. To allow for the greater expansion and the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto275
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

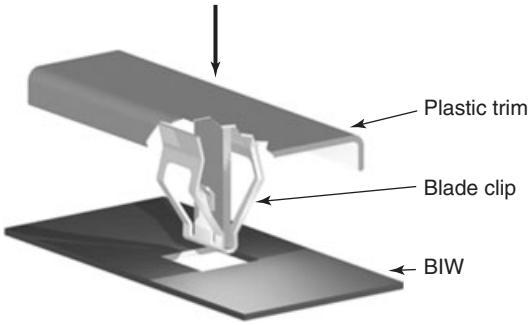


Figure 2. Metal clips used to fix the plastic trim to the BIW. (Reproduced with permission from JETPRESS.com. © JET PRESS.)

varying build tolerances of the BIW structure, the trim is designed with overlapping slip joints where required and positioned depending on the intended method of interior trim assembly on the vehicle assembly line.

5 MANUFACTURING METHODS

Injection molding is a commonly used process to manufacture plastic trim parts for a range of sizes and complexities. Parts can be produced direct from raw plastic materials with minimal further operations. The production time is fast and can be automated to produce high volumes of the same part.

The injection molding machine consists of a clamp and injection unit; these are assembled to the machine frame. The injection unit consists of a reciprocating screw, a heated barrel, and a nonreturn shut-off valve. The function of the injection unit is to plasticize a fixed volume of material ready for injection. The screw flight geometry can be optimized for the type of plastic being processed; however, most machines are installed with general purpose designs.

The clamp unit is used to open and close the mold tool and to create high pressures to hold the mold tool shut during injection molding (DIN 24450, 1987). The moving platen provides the opening and closing movements of the machine; it typically runs along four tie bars on the bed of the machine frame. The main designs of clamp units are toggle and hydraulic, however other types of design also exist.

There are a number of different forms of injection molding where a second or third plastic is injected in the same or a secondary process to create different colors or effects on components. The different techniques of molding are core-back multimaterial, cavity transfer, and co-injection (Goodship and Love, 2002). The secondary plastics need to be compatible to the main one; however, they can be a different type of plastic, grade, or color, or they can be soft.

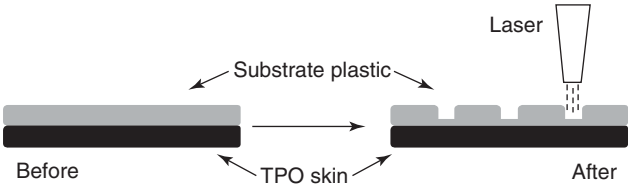


Figure 3. Laser cutter to remove the substrate material.

The IP and door panels typically have a soft-touch TPO skin on the exterior surface of the trim, to provide the driver and passengers with a quality tactile surface. The TPO skin is laid into the mold and the plastic is flowed onto the rear surface of the skin, before the mold is fully closed. This process is known as *injection compression* and it is used to prevent damage to the TPO skin from the high injection pressures of the injection molding machine.

Where this molding system is employed and airbag deployment is required, the rear of the molded article is made weaker to allow the airbag to deploy in a controlled manner. This is completed by using a laser cutter to remove the substrate material without affecting the visible interior trim surface TPO skin, as illustrated in Figure 3.

The injection mold tooling is made from standard components enabling easier manufacture and maintenance. The tooling is typically made from pretoughened steel and the form of the trim is cut directly from the computer-aided design of the part with a compensation factor for plastic shrinkage, cooling channels, and an ejection mechanism. For interior trim products, the surface finish of the tool is designed to the type of decoration required (Menges, Michaeli, Mohren, 2001).

6 DECORATING VEHICLE INTERIOR TRIM

6.1 Color-pigmented plastics

This method is the most common method of decorating plastic trim parts and provides a very durable result. The pigment or dye is uniformly dispersed throughout the plastic using an appropriate mixing technique. Dyes are easy to disperse into the plastic because they dissolve into the resin. The pigments are dispersed through the mechanical energy provided by the mixing technique, so that their particle size is reduced and the plastic can penetrate the surface of the pigment. Color match of trim components that may have been manufactured from different suppliers

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto275
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

using different techniques may also be an issue. This is partially overcome by the use of different colors, gloss levels, and surface texturing, and using a single source of compounded materials. Color and gloss are defined and controlled very stringently by the vehicle manufacturer.

The concentrate pigment is dispersed into a carrier substance and is added to the plastic at 1–5% addition to the natural plastic resin. The natural resin can be purchased in high quantities and is only colored when required at the processing stage of the plastic (Margolis, 1986). The colors of the pigments or dyes can be stock or color-matched to individual requirements; metallic effects can also be produced by the addition of metal flakes (Wheeler, 1999). This method of decorating plastics can result in low gloss and weld and flow lines from the flow into the mold tool.

6.2 Spray painting of plastic parts

Spray painting is commonly used in automotive interior trim where metallic, solid-colored, high gloss, or soft-touch finishes are required. This is especially true where color matching of the component is desirable to a mating part. Paint can be applied to complex geometries without contacting the part and can form thicknesses of 12–25 μm, depending on the application method. The spraying technique is very common where the aesthetic appearance as well as the physical properties, for example, abrasion and scuff resistance, require changes. The surface of the plastic material must be cleaned before to paint application. Some materials such as polyolefins require special treatment to activate the surface. This can include flame treatment or corona discharge. Some plastics may also require a primer to promote the adhesion of the paint to the surface of the plastic.

There are two main categories of spray paint products: solvent-based and water-based. Solvent-based paints contain solvents that evaporate following application onto the part. Depending on the application, these types of paints can be made up of up to 80% by weight of solids (60% of which can be pigments). They typically contain high levels of volatile organic compounds (VOCs), toxins that are limited under the Environmental Protection Act (Tromans, 1991). As a result, the use of solvent-based systems has been in decline and the amount of VOC content released into the atmosphere is reducing (Love, 2002; Sherman, 2004; Smith and Easterlow, 1996).

Water-based paint systems have typically low levels of VOC. They were developed in the 1950s with a view to replace traditional paint systems and to eliminate toxicity and combustion concerns. With these paint systems, water is used as a direct replacement of the traditionally used solvent. Owing to the nature of water, (high boiling point and high latent heat of evaporation), the paint requires more time and heat dry (Margolis, 1986; Stoye, 1998). Paint application can be completed by air atomization, airless atomization, centrifugal spray heads, and electrostatic assists.

6.3 Pad printing, (tampo printing)

Pad printing is a process that enables complex and flat geometry parts to be printed onto the surface. The process is suitable to be used on a variety of surfaces and textures without modification, it is commonly used to decorate or identify automotive decals and instrumentation. A variety of materials can be printed onto surfaces using the process, including thermoplastics, thermoset plastics as well as natural materials such as wood and leather. Some substrate materials may require pretreatment prior to the application of the ink. Cyan, magenta, yellow, and key (CMYK) (black) colors can be printed; different shades enable the whole spectrum of colors to be achieved (Berins, 1991; Gilleo, 1995, Muccio, 1999).

6.4 In-mold decoration (IMD)

In-mold decoration (IMD) is a technique widely used for the application of a film to an injection molded part to create decorative, technical, or wood grain finishes. IMD is one of the most efficient and cost-effective ways to decorate a part. It removes the inconvenience of on-site surfacing with paints or outsourcing to a third party. IMD provides rapid and easy surface changes and also provides the opportunity for significant cost reduction through component integration. This process relies on chemically compatible materials to act as substrates, allowing for remelting to occur during the injection molding process and ensuring material fusion takes place at the IMD film and plastic interface. The use of IMD as a decorative medium has several benefits including

1. Cost—the utilization of IMD techniques assists in the overall reduction of component piece cost (Raphanaud, 2004).
2. Environmental impact—removal of an entire manufacturing step results in reduced energy use and carbon emissions.
3. Styling—the technique provides a high quality finish that can incorporate unique company branding and identity, which improves the salability of new items.

There are a number of different IMD films on the market such as Soliant, Senoplast, 3M, Lakeside Films,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto275
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

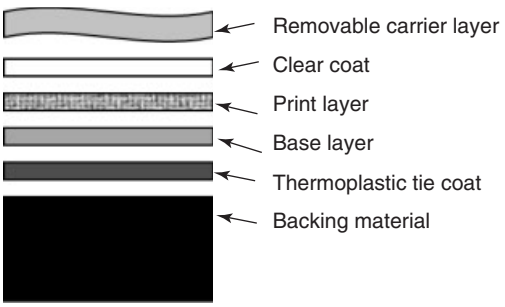


Figure 4. Typical structure of IMD film.

Sabic, Nissha, and Autotype. The films can be applied to a number of substrates such as PP, TPO, ABS, PVC, and PC; however, the compatibility of the injected plastic must be considered (Sherman, 2004).

Films in automotive interiors have been used since the 1970s. The materials have been developed to be durable enough for exterior use, while providing a suitable high quality gloss and color finish. The in-mold films are predominately multilayered structures as shown in Figure 4 (Nastas, 1994; Love, 2002).

The carrier layer is a flexible, formable, and heat-resistant sheet, normally polyethylene terephthalate, with a thickness of 25–75 μm. The clear coat is a cast coating of acrylic poly-vinylidene difluoride, materials such as poly-methyl methacrylate, PVC, and materials based on fluorocarbons. The typical film thickness for the backing material is 25–50 μm (Love, 2002). Data indicate that to make IMD a cost-effective process, the minimum volumes are ~22,000 p/a (Goodship, 2008).

Real wood veneers can also be used in an IMD process; thin sheets of the veneer can be preformed as inserts and placed into the injection mold tool. The injection molding process then follows, enabling a component to be manufactured with a visible real wood surface. Jaguar cars currently use this process to produce interior trim components for their vehicle range (Makenji *et al*., 2006; Makenji *et al*., 2007).

7 FUTURE TRENDS

The interior trim of the vehicle will continue to be manufactured from thermoplastic materials; however there is a trend for increased BIW replacement. Polymers with high temperature and modulus performance will be used further to replace steel materials and improved vehicle light weighting. As end-of-life recycling targets increase and the drive to use more recycled content in the vehicle picks up, the trim will also need to meet these demands; this can be executed by using similar types of plastics throughout. The trend of integration and modularization will escalate so that more assembly, decoration, and installation of subcomponents are completed during the primary manufacturing phase. This impacts the vehicle assembly method, reducing time and labor content. Engel, Austria have developed a one-step process to producing auto interior parts with a pleasing tactile surface known as *Dolphin*. The process uses physical foaming with the Mu-Cell microcellular process from Trexel Inc., Wilmington, Mass., together with a core back process that Engel calls "*reverse compression*." The Dolphin process is not only simpler than other methods of producing soft-surface interior parts but it also allows for complex geometries and undercuts (Anon, 2012a, 2012b). Another area that will increase is the demand for mass customization; in other words, decorating the interior to meet the customer's individual tastes and requirements. Dye sublimation is one such process, which shows fantastic potential; the technology will be developed to meet the tough performance requirements of the automotive industry (Makenji, 2011).

One area that is under the spotlight is the use of printed electronics with interior trim components. The ability to print circuit millimeters thin on thin substrates overcomes traditional wiring and packaging issues with electrical systems. The 2013 model of the Ford Fusion uses T-ink "plywood electronics" replacing the overhead instrument cluster. Different printing methods such as screen, offset, gravure, pad, flexo, rotary, and spray are employed, mainly using carbon and silver inks (Anon, 2012b). There is already a "proliferation of touch screen devices," increasingly printed on the vehicle interior trim. Research at the University of Warwick has enabled the manufacture of electroluminescent surfaces during injection molding. This process allows the vehicle trim to act as direct lighting surfaces or allows the OEM to create secret-till-lit areas in the cabin (Middleton and Goodship, 2012).

REFERENCES

Anon (2012a) First commercial use of 'Dolphin' moulding process. *Plastics Technology*, **58** (6), 6.

Anon, (2012b) First impressions from Printed Electronics USA 2012 California, Printed electronics USA 2012, http://www.printedelectronicsworld.com/articles/first-impressions-from-printed-electronics-usa-2012-california-00004985.asp (accessed May 2013).

Berins, M.L. (1991) *Plastics Engineers Handbook of the Society of the Plastics Industry*, 5th edn, Chapman and Hall, London, UK.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto275
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

DIN 24450 (1987) Machines for processing of plastics and rubber, DIN: Deutsches Institut Fur Normung E.V.

FMVSS 201 (1989) Federal Motor Vehicle Safety Standard 201, U.S. Department of Transportation, national highway traffic safety administration, Laboratory test procedure.

FORD WSSM15P4 F (2009) Engineering Material Specification - Assembly Performance, Hard Mold-in-color Interior Components, Ford Motor Company.

Gilleo, K. (1995) *Plastic Thick Film: Today's Emerging Technology for a Clean Environment Tomorrow*, Chester, UK, Springer.

Goodship, V. (2008) Removing the paint shop process: options for painting and decorating injection mouldings. *International Journal of Environmental Technology and Management*, **8** (4), 339–347.

Goodship, V. and Love, J. C. (2002) *Multi-Material Injection Moulding* Volume 13 of RAPRA review reports: RAPRA Technology Limited,, Shrewsbury, UK, Rapra Technology.

Helps, I. G. (2001) *Plastics in European Cars 2000–2008: A Rapra Industry Analysis, Rapra Industry Analysis Report Series*, iSmithers Rapra Publishing, Shrewsbury, UK.

Love, J. C. (2002) Injection Moulding 2002: Rapra Technology, *Rapra Technology Conference*, Barcelona, Spain.

Makenji, K., Goodship, V., Buckley, S., and Evered, C. (2006) A new in-mould decoration process using real wood veneer. *Progress in Rubber, Plastics and Recycling Technology*, **22** (4), 225–242.

Makenji, K., Goodship, V., Fanning, R., and Buckley, S. (2007) A new in-mould decoration process using real wood veneer. *Progress in Rubber, Plastics and Recycling Technology*, **23** (2), 83–95.

Makenji, K. (2011) Dye sublimation - variation of dye penetration depths with semi-crystalline and amorphous polymers, progress in rubber. *Plastics and Recycling Technology*, **27** (2), 69–84.

Margolis, J. (1986) *Decorating Plastics*, Hanser Publishers, Berlin, Germany.

Maxwell, J. (1994) *Plastics in the Automotive Industry*, Woodhead Publishing, Cambridge, UK.

Menges, G., Michaeli, W., and Mohren, P. (2001) *How to Make Injection Molds*, Hanser Verlag, German, München.

Middleton, B. and Goodship, V. (2012) Injection molding electroluminescent components. *Polymer Engineering & Science*, **53** (7), DOI: 10.1002/pen.23399

Muccio, E. (1999) *Decoration and Assembly of Plastic Parts*, ASM International, Ohio, USA.

Nastas, C. (1994) In Mold Painting and Priming of Plastics Parts Using Paint Film Technology, *Automotive Coating Systems Conference*, p. 8.

Raphanaud, S. (2004) IMD: The view of Peugeot Citroën and examples of applications, Conference Speaker, IMD 2004, Dusseldorf, Germany.

Sherman, L. (2004) NPE news wrap-up: decorating & painting. *Plastic Technology*, **50** (1), 1.

Smith, G. F. and Easterlow, R. (1996) New method for manufacturing painted components. *Plastics and Rubber Composites and Applications*, **25** (3), 116.

Stoye, D. (1998) *Paints Coatings and Solvents*, 2nd edn, Wiley-VCH, Weinheim, Germany.

Tromans, S. (1991) *Environmental Protection Act 1990*, Sweet & Maxwell, Great Britain.

Vidhyaa, S. K. (2012) Automotive Industry Seeks Light-Weight and High-Performance Materials. *ICIS Chemical Business*, **282** (3), 1.

Wheeler, I. (1999) *Metallic Pigments in Plastics*, iSmithers Rapra Publishing, Shrewsbury, UK.

FURTHER READING

Harper, C. (2006) *Handbook of Plastic Processes*, Chichester, UK, John Wiley & Sons.

Whelan, A. (1984) *Injection Moulding Machines*, Elsevier Applied Science Publishers, London, UK.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto275
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Metal Matrix Composites: Automotive Applications

Krishnan K. Chawla[1,2] and Nikhilesh Chawla[2]

[1] *University Alabama at Birmingham, Birmingham, AL, USA*
[2] *Arizona State University, Tempe, AZ, USA*

1 Introduction 1
2 Processing of MMCS 1
3 Automotive Applications of Metal Matrix Composites (MMCS) 2
4 Conclusion 6
Related Articles 6
References 6

1 INTRODUCTION

A composite material is a manufactured material that consists of two or more physically and/or chemically distinct, suitably arranged phases with an interface separating them; the composite has characteristics not shown by any of the components in isolation (Chawla, 2012). In this chapter, we provide a brief account of the processing, characteristics, and use of metal matrix composites (MMCs) in automotive vehicles. The term "*automotive vehicles*" includes passenger cars, sport utility vehicles, vans, trucks, buses, recreational vehicles, and so on. Reinforcements can be in the form of fibers, particles, or flakes; detailed information about fibrous reinforcements can be found in Chawla, 1999.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto279
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

2 PROCESSING OF MMCS

2.1 Liquid-state processing or casting

This is one of the common methods of processing MMCs, particularly, particle-reinforced composites. One of the important variables in the process is the viscosity of the liquid metal. The important fact to take into account is increase in the viscosity of the liquid metal by the addition of the ceramic particles (Chawla, 2012, 1999; Chawla and Chawla, 2006). This fact makes the commonplace gravity casting less attractive. Instead low pressure or squeeze casting is more effective. This involves the preparation of preform of the reinforcement (particle or short fiber plus organic binder material) and then forcing the liquid metal through the preform. Binder burnout is a significant problem. Reinforcement–matrix compatibility is another problem in liquid-state processing of MMCs. This has two components, a chemical component that has to do with an undesirable reaction at the interface, for example, reaction between carbon and aluminum to form Al_4C_3, which is a very brittle and highly hygroscopic compound. The second is a physical component. This involves the problem of poor wettability of the ceramic reinforcement by the liquid metal and the mismatch in the coefficient of thermal expansion of the metal and ceramic. The problem of wettability can be reduced by matrix composition modification or by applying pressure during processing.

2.2 Powder processing

Powder metallurgy (PM) or powder processing is a very versatile route. Ceramic particles are difficult to wet

by molten metal. The PM route obviates that difficulty. Practically, any ceramic particle species can be mixed with any metal or alloy powder to produce a composite.

Metallic powder can be elemental or prealloyed atomized powder (20–40 μm). Rapidly solidified ribbon chopped into flakes can also be used. The metal matrix powder and the ceramic particles are blended into a homogeneous mixture, that is, no agglomerates should be present in the final blend. The size of the powder is perhaps the most critical parameter in this regard. It was observed that in the case of SiC_p/Al, the SiC particle/Al particle ratio of 0.7 gave a more uniform reinforcement distribution than a ratio of 0.3 (Dinwoodie *et al*., 1985). Frequently, metal powders have hydrated oxide films on them. These water molecules must be removed by heating in order to avoid gas evolution and porosity in subsequent processing. The final step is vacuum consolidation to obtain a 95%+ dense composite. Hot pressing can be done below the matrix alloy solidus or in the liquid/solid region. Working in the liquid/solid region can give faster kinetics but there can also lead to an unwanted reaction between the reinforcement and the metal.

Secondary processing can be used. Generally, consolidated composites are subjected to extrusion (extrusion ratio between 15 and 20) to break the oxide film. The oxide film is present between metal powder particles as well as on ceramic particles. The extrusion process results in a plastic flow of the metal matrix and breaks down any clusters of ceramic particle, giving a better distribution of ceramic particles.

Advantages:

1. Any reinforcement/metal combination can be obtained.
2. Reaction between reinforcement/matrix can be minimized in a solid-state process.
3. There is essentially no limit on the volume fraction of ceramic particles that can be incorporated in a metal.
4. Nonequilibrium alloys, for example, those produced by rapid solidification, can be used as the matrix.

Disadvantages:

1. The PM process is generally complex and the product forms are somewhat limited.
2. Highly reactive, potentially explosive powders can be involved.
3. The PM process is quite expensive *vis-à-vis* liquidtate processing methods.

3 AUTOMOTIVE APPLICATIONS OF METAL MATRIX COMPOSITES (MMCS)

MMCs are used in a variety of automotive applications. The main reasons for this are enhanced strength and stiffness, coupled with low density. In addition to these properties, under certain circumstances, we have improved cyclic MMCs, for example, in pistons, cylinder liners, connecting rods, powertrain and suspension components, brake rotors, and calipers. An early and successful engine application was that of selectively reinforced aluminum pistons in the Toyota diesel engine (Donomoto *et al*., 1983). In this application, an alumina–silica chopped fiber preform was incorporated into the ring groove area of the piston during pressure casting of the aluminum. The conventional diesel engine piston has an Al−Si casting alloy with a crown made of a nickel cast iron. The main property requirement was increased wear resistance in this area. The previous approach used a Ni-resist ring that increased weight and differed in coefficient of thermal expansion from the aluminum alloy piston material. Aluminum matrix materials reinforced with SiC particles have also been used in piston applications, primarily in drag racing cars. In this case, the lower coefficient of thermal expansion of the MMC compared to that of conventional aluminum allowed reduced clearances between the piston and cylinder wall leading to improved performance.

Another early MMC application in an automotive engine was a hybrid particulate-reinforced Al matrix composite used as a cylinder liner in the Honda Prelude (Figure 1). The composite consisted of an Al–Si matrix with 12% Al_2O_3 (short fibers for wear resistance, and 9% carbon for lubricity). Both the reinforcements were in the form of short fibers. The composite was integrally cast with the engine block, had improved cooling efficiency, and exhibited improved wear and a 50% weight savings over cast iron, without increasing the engine package size. While this concept was initially implemented in the Honda Prelude 2.3 L engine, it has also been used in the Honda S2000, Toyota Celica, and Porsche Boxster engines (Hunt and Miracle, 2001). Toyota's 2ZZ-GE engine is another example where MMCs have been incorporated in the cylinder walls. These composites are made by high-pressure casting and use ceramic particles and fibers. It is worth recalling that a reduction in the weight of the automobile is a major factor in increasing fuel efficiency. Conventionally, this is done by replacing the heavy cast iron engine block with aluminum. Cast iron liners, however, are still used in aluminum engine blocks because of their superior wear resistance. Piston liners made of MMCs can replace cast

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto279
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

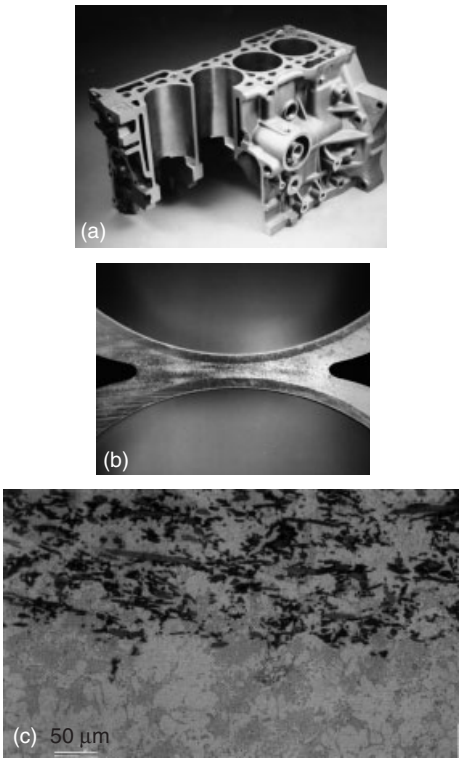


Figure 1. Hybrid particulate-reinforced Al matrix composite used as a cylinder liner in the Honda Prelude (parts (a) and (b) courtesy of D. Miracle). The composite consisted of an Al–Si matrix with 12% Al_2O_3 for wear resistance, and 9% carbon for lubricity. (a) Prelude engine block, (b) magnified view of cylinder liner, and (c) microstructure of composite showing carbon short fibers (black) and Al_2O_3 fibers (dark gray). (Reproduced from *Metal Matrix Composites*, 2006, N. Chawla and K.K. Chawla. With kind permission of Springer Science+Business Media.)

iron liners. The 2ZZ-Ge engine does that. It contains MMC liners that contain 5% alumina–silica fibers (95% alumina, 5% silica) and 10% mullite particles (12 μm) (Fujine *et al*., 2000).



Figure 2. Particulate MMCs for use in brake drums and brake rotors, as a replacement for cast iron (courtesy of D. Miracle). The high wear resistance and thermal conductivity coupled with 50–60% weight savings, make MMCs quite attractive for this application. (Reproduced from *Metal Matrix Composites*, 2006, N. Chawla and K.K. Chawla. With kind permission of Springer Science+Business Media.)

Particulate MMCs, particularly Al-based MMCs, have been used in brake drums and brake rotors as a replacement for cast iron (Figure 2). The high wear resistance and thermal conductivity coupled with 50–60% weight savings, make MMCs quite attractive for this application. An intensive development effort was carried out using cast 359Al/SiC/20_p composite. While the costs for this rotor are somewhat higher than those for cast iron, the benefits were justified in a number of specialty vehicles, such as the Plymouth Prowler, Lotus Elise, and others (Hunt and Miracle, 2001).

One of the disadvantages of MMCs with a ceramic reinforcement is that they are typically more difficult to

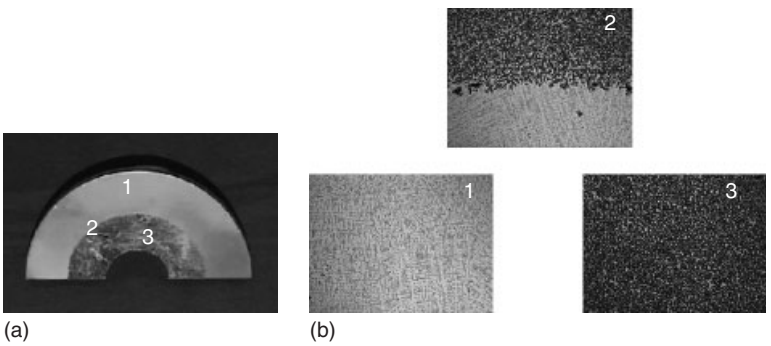


Figure 3. (a) Centrifugally cast brake rotor with selective placement of reinforcement, and (b) regions of the microstructure in the rotor: (1) matrix-rich, (2) interface, (3) reinforcement-rich (courtesy of D. Herling). (Reproduced from *Metal Matrix Composites,* 2006, N. Chawla and K.K. Chawla. With kind permission of Springer Science+Business Media.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto279
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

machine than the unreinforced alloy. In centrifugal casting, optimal placement of the reinforcement can be achieved by using a centrifugal force during casting, which results in a gradient in reinforcement volume fraction (Divecha, Fishman, and Karmarkar, 1981; Tsunekawa *et al*., 1988). In brake rotors, for example, wear resistance is needed on the rotor face, but not in the hub area. Thus, in areas where reinforcement is not as crucial, such as in the hub area, machining would be easier without the reinforcement. Figure 3 shows the microstructure at different points in a centrifugally cast brake rotor, showing a pure aluminum alloy matrix region, interface region, and reinforced region. The process is relatively inexpensive, with a potential for composite materials at as low as \$2/kg. Thus, use of "selective reinforcement," whereby the hard SiC particles are used precisely where high strength and wear resistance are required would appear to be quite attractive.

An important application of MMCs in the automotive sector is the driveshaft. The key limitation in this case arises from the critical rotational speed at which the component becomes dynamically unstable. Use of MMCs reduces the inertia which in turn allows the critical speed of the driveshaft to be increased. The critical rotational speed (N_c) is given by (Hoover, 1991)

$$N_c = \frac{15\pi}{L^2}\left[\frac{E}{\rho}\left(R_o + R_i\right)^2\right]^{\frac{1}{2}}$$

where L is the length of the driveshaft, E is the Young's modulus, ρ is the density, and R_o and R_i are the outer and inner radii of the shaft, respectively. The take away

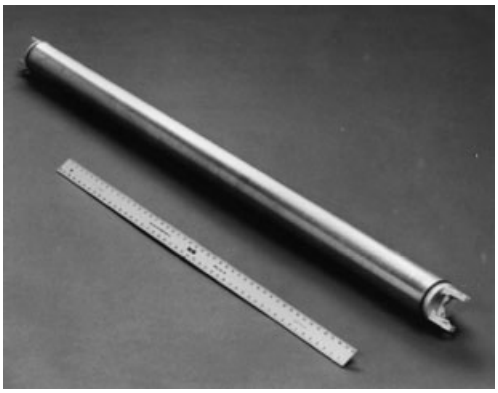


Figure 4. $6061/Al_2O_3/20_p$ composite used as driveshaft in the Corvette (courtesy of D. Miracle). The composite exhibits a 36% increase in specific modulus over steel. (Reproduced from Metal Matrix Composites, 2006, N. Chawla and K.K. Chawla. With kind permission of Springer Science+Business Media.)

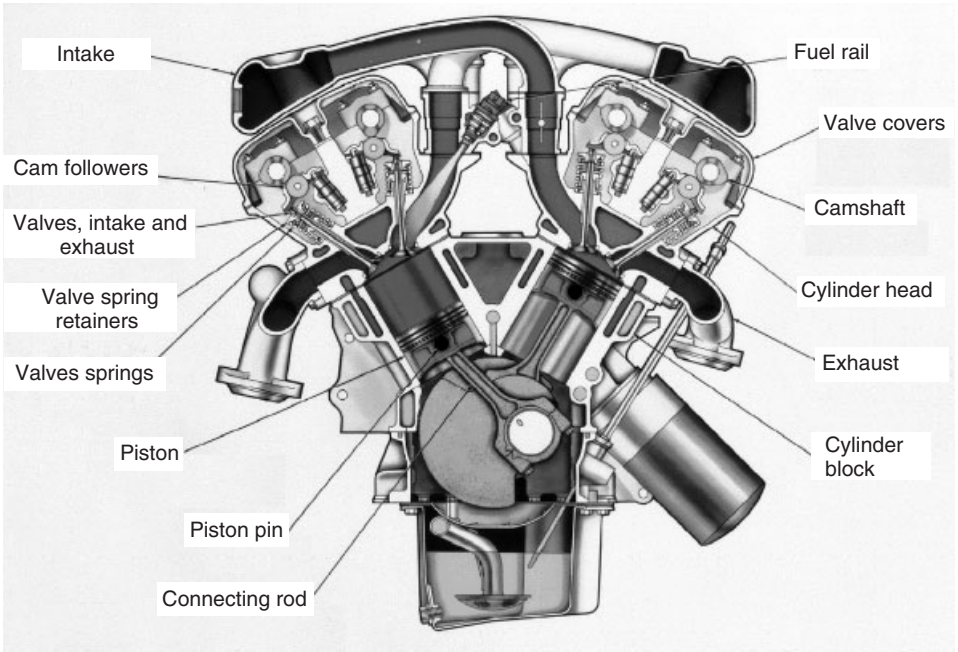


Figure 5. Cross-section of a passenger car engine showing the location of the connecting rod (courtesy of J. Allison). (Reproduced from *Metal Matrix Composites*, 2006, N. Chawla and K.K. Chawla. With kind permission of Springer Science+Business Media.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto279
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

message from this equation is that the material parameter that controls critical speed is the specific modulus, E/ρ. One of the requirements of the driveshaft is that it be welded to a yoke. Thus, SiC reinforcement is precluded from material selection, because of the harmful reaction products formed between Al and SiC in liquid-phase processes. Duralcan has used a $6061/Al_2O_3/20_p$ composite, which exhibits a 36% increase in specific modulus over steel (Figure 4) (Allison, Jones, Davis, 1997). The modulus of the composite increases with the volume fraction of the reinforcement (fiber or particle). As an example, we cite the fact a particle-reinforced MMC can have the same modulus as gray cast iron. Thus, in a modulus-based design, sections can have thicknesses that are the same as that of cast iron and less than that of aluminum sections. This could be a great advantage in terms of E/ρ.

Another important potential replacement of steel by SiC particle-reinforced Al matrix composite is in the connecting rod (Figure 5). The connecting rod requires high fatigue resistance at temperatures as high as 150°C. A lighter connecting rod would result in (a) 12–20% reduction in secondary shaking force, (b) 0.5–1% improvement in fuel economy (with lightweight piston and pin), (c) 15–20% increase in peak RPM, (d) decreased bearing width (package improvement), and (e) increased bearing and crankshaft durability. Initial attempts at developing an MMC connecting rod used hot-pressing followed by extrusion, commonly used to fabricate aerospace components. This technique proved to be too costly in the automotive sector (where production volume is much larger than in the aerospace industry) because of the large amount of wasted material. Near-net-shape sinter forging was used to fabricate MMC connecting rods with tensile and fatigue properties comparable to those of extruded materials (Koczak *et al*., 1993). A prototype model of this rod is shown in Figure 6.

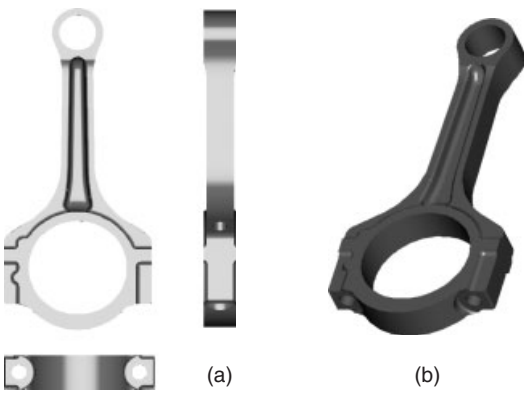


Figure 6. Models of sinter-forged MMC connecting rods: (a) two-dimensional view and (b) three-dimensional view (courtesy of F. Liu). (Reproduced from *Metal Matrix Composites*, 2006, N. Chawla and K.K. Chawla. With kind permission of Springer Science+Business Media.)

Table 1. Weight comparison of MMC and steel connecting rods.

	$2080/SiC/20_p$	Steel
Pin weight (g)	65.2	144.7
Crank weight (g)	184.0	437.7
Total weight (g)	249.2	582.4

In Table 1, we compare the weight of the MMC rod *vis-à-vis* that of the steel rod (Chawla *et al*., 2002). A 57%

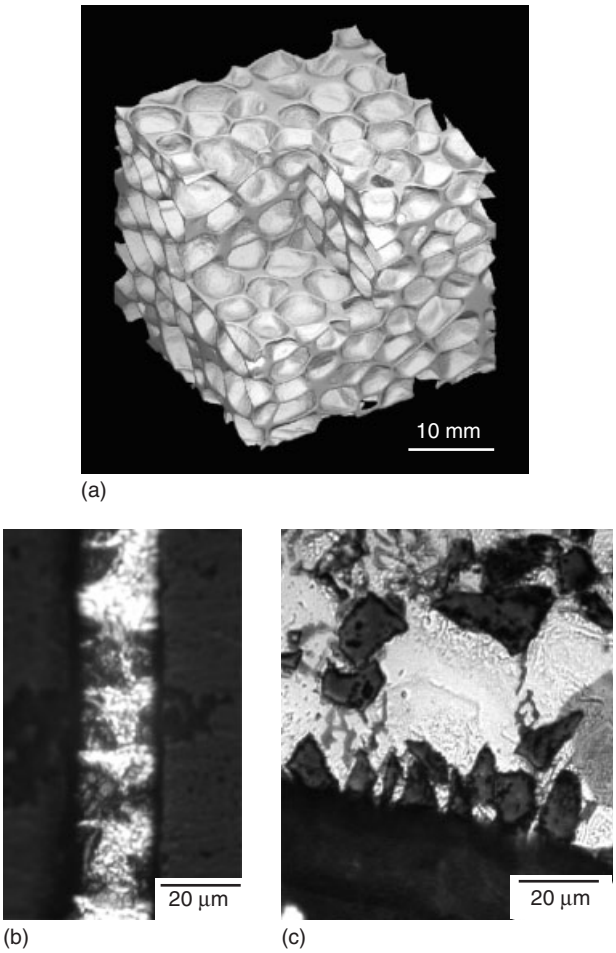


Figure 7. $6061/Al_2O_3/22p$ composite foam used in Ferrari cars (courtesy of H.P. Deigischer): (a) three-dimensional microstructure obtained by computed X-ray tomography, (b) microstructure in the cell wall, which is about twice the SiC particle diameter, and (c) segregation of SiC particles to the cell wall. (Reproduced from *Metal Matrix Composites*, 2006, N. Chawla and K.K. Chawla. With kind permission of Springer Science+Business Media.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto279
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

weight savings was achieved with the MMC rod, with a moderate increase in cost. In addition, it is estimated that for every 1 kg of weight removed from connecting rods, 7 kg of supporting and counterbalancing structure can be eliminated (Hunt and Miracle, 2001). Other demanding powertrain applications are intake and exhaust valves. These components must have good high cycle fatigue performance at elevated temperature, good sliding wear resistance, and creep resistance. Austenitic stainless steel was replaced by a TiB_2-reinforced Ti matrix composite in the Toyota Altezza (Hunt and Miracle, 2001).

MMC foams have also been developed for damping/energy-absorbant applications in automobiles (Leitlmeier, Degischer, Flankl, 2002). A $6061Al/Al_2O_3/22_p$ composite foam was has been used in Ferrari cars. Figure 7a shows the three-dimensional microstructure of the foam obtained by computed X-ray tomography. The pores are a few millimeters in diameter and are homogeneously distributed. Figure 7b shows the microstructure of the composite (Babcsan *et al*., 2004). The SiC particles stabilize the wall thickness (Figure 7b), and segregate to the cell wall (Figure 7c). The thickness of the cell walls is about twice the SiC particle diameter.

4 CONCLUSION

In summary, MMCs find applications in automobiles as niche applications. This stems from their high modulus, high strength, and superior wear characteristics. It has to be admitted that compared to polymer matrix composites, the penetration of MMCs in the automotive sector is not as extensive. The high cost of MMCs would appear to be a major factor.

RELATED ARTICLES

Piston and Ring Development
Material and Process Selection - Cylinder Blocks and Heads
Lightweighting Approach: A Historical Perspective
Wrought Aluminum Alloys: Grades and Properties and Processing
Metal Matrix Composites: Automotive Applications

REFERENCES

Allison, J.E., Davis, L.C., and Jones, J.W. (1997) in *Composites Engineering Handbook*, (ed. P.K. Mallick) Marcel Dekker, New York, pp. 941.

Babcsan, N., Leitlmeier, D., Degischer, H.P., and Banhart, J. (2004) The role of oxidation in blowing particle-stabilised aluminium foams. *Advanced Engineering Materials*, **6**, 421–428.

Chawla, K.K. (1999) *Fibrous Materials*, Cambridge University Press, Cambridge.

Chawla, K.K. (2012) *Composite Materials*, 3rd edn, Springer, New York, p 363.

Chawla, N. and Chawla, K.K. (2006) *Metal Matrix Composites*, Springer, New York.

Chawla, N., Williams, J.J., and Saha, R. (2002) Mechanical behavior and microstructure characterization of Sinter-forged SiC particle reinforced aluminum matrix composites. *Journal of Light Metals*, **2**, 215–227.

Dinwoodie J., Moore E., Langman C.A.J., and Symes W.R., ICCM-V, 1985, pp. 671–685.

Divecha, A.P., Fishman, S.G., and Karmarkar, S.D. (1981) Silicon carbide reinforced aluminum – a formable composite, Sept. *Journal of Metals*, **33** (9), 12.

Donomoto, T., Miura N., Funatani K., and Miyake N., (1983), Ceramic fiber reinforced piston for high performance diesel engine. SAE Tech. Paper no. 83052.

Fujine, M., Kato, S. Takami, T., Hotta, S. (2000) Development of Metal Matrix Composite for Cylinder Block in *Seoul 2000 FISITA World Automotive Congress*, F2000A065, Seoul, Korea.

Hoover, W. (1991) *12th Risø International Symposium* (eds N. Hansen), Risø National Laboratory, Roskilde, Denmark, pp. 387–392.

Hunt, W.H. and Miracle, D.B. (2001) Automotive applications of metal matrix composites in *ASM Handbook – Composites*, vol. 21 (eds D.B. Miracle and S.L. Donaldson), ASM International, Material Park, OH, pp. 1029–1032.

Koczak, M.J., Khatri, S.C., Allison, J.E., and Bader, M. (1993) Metal matrix composites for ground vehicle, aerospace and industrial applications in *Fundamentals of Metal Matrix Composites*, (eds S. Suresh, A. Mortensen, and A. Needleman) Butterworth-Heinemann, Stoneham, MA.

Leitlmeier, D., Degischer, H.P., and Flankl, H. (2002) Development of a foaming process for particulate reinforced aluminium melts. *Advanced Engineering Materials*, **4**, 735–740.

Tsunekawa, Y., Okumiya, M., Niimi, I., and Yoneyama, K. (1988) Centrifugally cast aluminum matrix composites containing segregated alumina fibers. *Journal of Materials Science Letters*, **7**, 830–832.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto279
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Voltage Control and Frequency Control

Christian Rehtanz, Johannes Rolink, and Willi Horenkamp

TU Dortmund University, Dortmund, Germany

1 Introduction 1
2 Voltage Control 1
3 Frequency Control 2
4 Summary 4
Endnotes 5
References 5

1 INTRODUCTION

Electric vehicles (EVs) are nowadays treated as potential candidates for providing ancillary services. Owing to their storage capability, they are able to store energy in times of overproduction, especially from renewable energy sources, and to restore it, for example, when the volatile feed-in is low. Thereby, EVs can help to maintain and improve the quality, reliability, and stability of the supply system.

Ancillary services are services that the grid operator provides for power consumers in addition to electricity transmission and distribution. Ancillary services include (VDN, 2007a):

frequency control;
voltage control;
restoration of supply;
system/operations management.

These services are not usually provided using low voltage power sources. This may change in the future.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto283
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

This chapter looks at *frequency control* and *voltage control* services in connection with EVs. For the analysis, the German perspective is chosen. First of all, voltage control is discussed briefly. This is followed by an overview of the current regulatory framework for providing frequency control in Germany. Eventually, a basic analysis of the technical potential of EVs is presented considering different types of control power. Most of the following results can be generalized for the territory of the former UCTE.[1] Unfortunately, this is not completely possible because of national regulatory aspects.

2 VOLTAGE CONTROL

Low voltage supply systems experience excessively low voltage conditions when they have long lines and high consumption. Excessively high voltage conditions are increasingly being seen in rural areas of Germany in connection with photovoltaic systems. Supply system voltage can be kept within specified tolerances by upgrading the system or reducing the consumption or output of grid-connected power sources.

EVs can help support voltage control in low voltage supply systems. For example, they can supply or consume reactive power. This must be done locally, as reactive power cannot be transported over long distances. Given the primarily resistive nature of low voltage lines (R/X ratio of 2.57 with NAYY^2 $4 \times 150\,\text{mm}^2$ and 50 Hz) and the restricted power factor (VDEW, 2001), power factor correction has only a limited impact (also see Impact of Electric Vehicles on Low Voltage Supply Systems). Power factor correction also requires larger charging converters. The most effective way to regulate voltage in the low voltage supply system is by adjusting overall current; this also changes the active current component. This can be

done, for example, by consuming locally generated power near the point of generation or generating power near the point of consumption. EVs with bidirectional converters can handle both options, allowing them to be used to both raise and lower system voltage. However, owners of consumption and generation equipment in the low voltage supply system are currently under no obligation to actively support voltage control.

3 FREQUENCY CONTROL

Temporary deviations between generation and load are offset using control power, which stabilizes system frequency and returns it to the scheduled value. Deviations occur as a result of generation loss, major grid disturbances, maintenance, incorrect load and generation forecasts, and schedule changes. There are three forms of control power in Germany: primary, secondary, and tertiary controls. They differ in their response times and duty durations. Primary control power is activated automatically and locally. It is supplied mainly by fossil fuel-fired steam power plants. As primary control power must come online quickly, the only alternative to thermal power stations is battery storage (VDE, 2008). Secondary control power is activated automatically and centrally using the network characteristic method (UCTE, 2009). This form of control power is supplied by hydraulic storage power plants and steam power plants. Pumped storage power plants are ideal suppliers of secondary control power, because they have such high rates of change of power. Tertiary control power ("minute reserve") is activated manually, usually by phone. It is supplied by hydroelectric plants, fast-start gas turbines, and thermal power plants operating below full capacity.

3.1 Regulatory environment

Several codes and regulations define the regulatory environment for providing control power. One is the Operation Handbook, which defines technical standards for the territory of the former UCTE (2009). Nationally, the Operation Handbook is supplemented by the Transmission Code (VDN, 2007a) and the prequalification conditions (VDN, 2007b). The procurement of control power is governed by the German Grid Access Regulation (StromNZV) (German Government, 2011), which is interpreted and supplemented by resolutions passed by the German Federal Network Agency.

The transmission system operators (TSOs) require providers of control power to maintain certain availability factors and capacity factors (VDN, 2007b). TSOs in Germany procure control power on a shared internet platform (German Transmission System Operators, 2012). On the platform, the tenders can be either symmetrical or asymmetrical. Symmetrical tenders require bids to cover equal amounts of positive and negative control powers. Asymmetrical tenders, by contrast, treat positive and negative control powers as separate products. Compensation is based on energy and/or demand tariffs. Bidders have to meet minimum lot sizes and bid on specific periods (time slots) during the day (German Federal Network Agency for Electricity, Gas, Telecommunications, Post and Railways, 2011). Table 1 summarizes the various parameters for each type of control power.

If EVs are used to provide control power, a sufficient number of them will have to be pooled together to meet the minimum lot size requirement. In addition, the vehicles will have to be coordinated, so that they comply with the availability and capacity factor requirements. This results in different requirements for the vehicles' communication systems. Primary control power, for example, is automated locally and automatically; therefore, information about how many EVs are contributing to the primary control reserve should be available centrally. The technical units in the secondary control reserve, by contrast, are connected to a closed-loop control circuit with a control cycle of no more than 4 s (VDN, 2007b). The pool operator has to independently assign the set point signal from the secondary

Table 1. Parameters for types of control power.

	Primary	Secondary	Tertiary
Availability factor	100%	95%	100%
Capacity factor	n/a	100%	100%
Tender	Symmetrical	Asymmetrical	Asymmetrical
Compensation	DT	DT + ET	DT + ET
Minimum lot size	±1 MW	±5 MW	±5 MW
Time slots (work day)	1	2	6
Procurement	ACA	ACA	ACA

DT: Demand tariff
ET: Energy tariff
ACA: Across control areas (Germany-wide).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto283
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

controller to the various technical units within the cycle time. The operator also needs to know the availability and capacity factors for the vehicles. The communications link between the pool control center and the secondary controller should be a redundant point-to-point connection (VDN, 2007b). While tertiary control power still tends to be requested by phone, in the future, the EVs will have to be activated automatically because of the expected pool sizes. Here, too, the pool control center and the vehicles will have to exchange measurement and control signals as well as information on availability and capacity factors in order to coordinate the vehicles.

3.2 Contribution to control reserve

EVs can essentially contribute to control reserves in a variety of ways. First, they can be operated as changing loads during the charging process. This can include load shedding. In this case, the maximum power contribution varies depending on the number of vehicles being charged in the course of the day. This approach only provides positive control power. It is, however, also possible to vary charging output within a fixed symmetrical control range for each vehicle. This would provide access to both positive and negative control reserves. As the number of charging vehicles fluctuates in the course of the day, neither approach can provide a constant range of power throughout the day (cf. Figure 1, Impact of Electric Vehicles on Low Voltage Supply Systems).

EVs can also be used outside of the charging process to store power. This allows them to provide both negative and positive control powers. To provide positive control power, they would need a charging converter that can operate in both directions. If each vehicle is required to meet a certain capacity factor, part of the vehicle battery will have to be set aside as an energy reserve for the duration of the time slot while maintaining a constant power range. This operating mode depends on the vehicles' availability factor and is thus unaffected by fluctuations in charging vehicles in the course of the day. Figure 1 clearly illustrates this: it shows the availability factor for EVs charging at 3.7 kW at work or at home on a workday. The two lines show the availability factors when vehicles that are not mobile during the day are classified as not available (−NM) or available (+NM) (Rehtanz and Rolink, 2010).

Vehicles that only serve as a negative control reserve will have to maintain adequate reserve capacity at the start of each time slot. This approach is ideal for EVs that are used every day: when users drive their vehicles, they consume the energy stored in the vehicle batteries, thereby restoring the vehicles' ability to absorb power as part of the negative reserve (i.e., the capacity factor). When providing a positive control reserve, the batteries must be adequately charged at the start of each time slot. This approach is ideal for vehicles that are rarely used.

3.3 Power and energy reserve

3.3.1 Primary control

Primary control is a proportional control. To include electrical vehicles in primary control, each vehicle will need

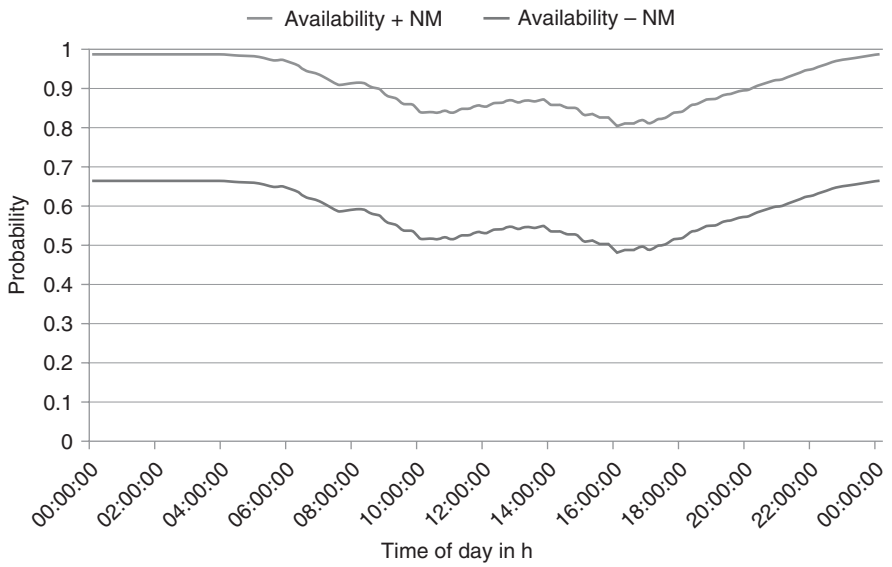


Figure 1. Vehicle availability factor on a workday.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto283
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

to be assigned a certain power frequency characteristic K. This power measure is calculated as follows:

$$K = \frac{\Delta P}{\Delta f} \tag{1}$$

It describes the power difference ΔP of the technical unit divided by the frequency deviation Δf from the target value ($f_t = 50$ Hz). The full primary control reserve must be activated when the frequency deviation is $|\Delta f|_{max} = 200$ mHz (UCTE, 2009).

Assuming a control range of 5% of nominal power per vehicle ($P_N = 3.7$ kW) in bidirectional operation, the symmetrical control range would be $\Delta P_{EV} = \pm 185$ W per vehicle. According to Equation 1, this translates to a power frequency characteristic of $K_{EV} = 925$ W/Hz. With only one million vehicles and a minimum availability factor of 50% (Figure 1), the possible control reserve could reach 92.5 MW, or roughly one-seventh of the primary control reserve maintained in Germany.[3] Given the high availability factor requirements and essentially continuous use of the primary control reserve, the next step is to determine the energy reserve required per vehicle.

Utility frequency is used as a reference signal for timekeeping. As utility frequency fluctuates, there are timekeeping errors. To compensate, utility frequency is regulated in order to keep network phase time synchronized with coordinated universal time (UTC) over the long term. This is done by adjusting the secondary controller. The grid control center in Laufenberg (Switzerland) is responsible for monitoring and corrections. The difference between UTC and synchronized time should not exceed $|\Delta t|_{max} = 30$ s (UCTE, 2009). The phase error $\Delta\theta$ resulting from the time deviation Δt is calculated as follows:

$$\Delta\theta = 2\pi f_t \Delta t \tag{2}$$

The maximum phase error can be calculated as $|\Delta\theta|_{max} = 9424.8$ rad. The relationship between the phase error and the energy E_{PRL} consumed or generated by the vehicle is calculated as follows:

$$E_{PRL} = \frac{K_{EV}}{2\pi}\Delta\theta \tag{3}$$

If we use the power measure as calculated in Equation 3, the necessary energy reserve per vehicle works out to $E_{PRL} = \pm 0.385$ kWh. Analyses of actual frequency curves have confirmed that this magnitude is correct and show that phase error fluctuations span several days, that is, the maximum energy is not reached within a single day. Therefore, a reserve of 0.77 kWh per vehicle is sufficient to provide primary control power under the above conditions.

3.3.2 *Secondary and tertiary control reserves*

To estimate EVs' contribution to secondary and tertiary control reserves, we will assume that the control reserve buffer in the vehicle batteries is sizable enough to provide the full amount of reserve power for at least 4 h.[4] The vehicles are assumed to only participate in one time slot a day. We also assume that the provider pools are homogeneous, that is, they do not simultaneously provide positive and negative secondary and tertiary control powers within a single day. This is done for two reasons: first, providers have to meet different requirements in each pool. Second, combining both options would double the required control reserve buffer.

Drivers' mobility requirements should be the main priority when determining vehicle power and energy reserves. Power set aside for the control reserve should have only a moderate impact on the charging process during the time slot. Similarly, the EV should only have to maintain a moderate amount of energy for positive/negative energy reserve purposes. The amount of power set aside and the energy reserve will have to be chosen based on availability and capacity factors, daily vehicle consumption, battery capacity, and charging power.

In Germany, the average daily consumption of a vehicle that is mobile during the day is assumed to be 6 kWh (Rehtanz and Rolink, 2009). A common battery capacity for modern electric-only vehicles is roughly 30 kWh.[5] Thus, a control reserve buffer of 3 kWh is equal to one-half the assumed daily consumption or 10% of stated battery capacity. If a vehicle with this control reserve buffer supplies or consumes control reserve power during a 4 h time slot and the capacity factor is 100%, the maximum control reserve buffer per vehicle is $P_{max} = 0.75$ kW. Thus, with one million vehicles, the total control reserve would be 750 MW. The maximum energy provided by a vehicle pool of this size is 3000 MWh. As the control reserve must be available during the entire time slot, this might reduce the charging output by 20% during the charging process (Rehtanz and Rolink, 2010).

4 SUMMARY

EVs can be used to provide ancillary services. To ensure greater flexibility and effectiveness, however, the vehicles must be equipped with a bidirectional charging converter. Voltage and frequency controls were investigated. The investigations showed that EVs can make only a limited

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto283
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

contribution to voltage control, because most of the load on low voltage lines is resistive. They can, however, help in integrating additional EVs in certain circumstances. In order to participate in voltage control of low voltage supply systems, the vehicles cannot be allowed to operate autonomously. Instead, they need to be integrated into a management system that coordinates vehicles in response to the electricity supply system's current condition.

At first glance, EVs appear able to make an enormous contribution to provide control power. One million EVs could cover a significant portion of Germany's primary control reserve requirements. Each vehicle would only need to set aside a small portion of its battery for this purpose. However, the actual impact depends largely on vehicle availability. EVs must be coordinated in a pool in order to meet the high availability and capacity factor requirements. The technical requirements for secondary control power are particularly high. However, even primary control, which is activated locally, would require some form of data communications in order to indicate the vehicle's status.

EVs can be involved in primary control and in power factor correction. This requires the development of sound business models for vehicle owners. If none are developed, it may be possible to compel them to provide these services through technical terms of service.

ENDNOTES

1. Union for the Coordination of the Transmission of Electricity.
2. Cable type description according to the German standard VDE 0276-603.
3. At present, approximately 600 MW of primary reserve is provided in Germany.
4. According to VDN (2007b) and Table 1.
5. For example, Think!, BMW Mini E.

REFERENCES

German Federal Network Agency for Electricity, Gas, Telecommunications, Post and Railways (2011) Decisions BK6-10-097, BK6-10-098, and BK6-10-099.

German Government (2011) Stromnetzzugangsverordnung (StromNZV), as of July 28.

German Transmission System Operators (2012) URL: http://www.regelleistung.net/ (accessed 23 July, 2012)

Rehtanz, C. and Rolink, J. (2009) Conditions for the Demand Side Management of PHEVs and EVs. ETG Congress, Düsseldorf, Germany.

Rehtanz, C. and Rolink, J. (2010) Evaluation of the Application of Electric Vehicles for the Provision of Ancillary Services, VDE Congress, Leipzig, Germany.

UCTE (Union for the Coordination of the Transmission of Electricity) (2009) *Operation Handbook*, Policy 1.

VDE (German Association for Electrical, Electronic & Information Technologies) (2008) Energiespeicher in Stromversorgungssystemen mit hohem Anteil erneuerbarer Energieträger.

VDEW (German Association of the Electrical Power Industry) (2001) *Eigenerzeugungsanlagen am Niederspannungsnetz*, 4th edn. VWEW, Frankfurt (Main).

VDN (German Association of the System Operators) (2007a) TransmissionCode 2007.

VDN (German Association of the System Operators) (2007b) Unterlagen zur Präqualifikation zur Erbringung von Regelleistung, TransmissionCode 2007, Annex D1–D3.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto283
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Impact of Electric Vehicles on Low Voltage Supply Systems

Christian Rehtanz, Johannes Rolink, and Willi Horenkamp

TU Dortmund University, Dortmund, Germany

1 Introduction 1
2 Simultaneity Curves of Charging Electric Vehicles 2
3 Selecting Reference Supply Systems 3
4 Supply System Constraints 4
5 Analysis of the Capacity of Low Voltage Supply Systems 5
6 Summary 8
Endnote 9
References 9

1 INTRODUCTION

Electricity supply analyses with focus on electric vehicles (EVs) started in the early 1980s (Heydt, 1983; Boonekamp and Wakkerman, 1993; Rahman and Shrestha, 1993). However, the impact of EVs on power grids has only recently become a significant object of research (Clement-Nyns, Haesen, and Driesen, 2010; Taylor *et al*., 2009; Richardson, Flynn, and Keane, 2010; Babaei *et al*., 2010). EVs are relatively large household loads, and their widespread, mass deployment will have a significant impact. The key questions of this chapter are how many vehicles can the current public power grid handle in Germany and what grid upgrades will be needed over the long term in connection with EVs?

One difficulty is that research findings cannot be fully validated. EVs are few in number and usually deployed locally. While multiple projects have been conducted, the resulting data are too sparse to draw universal conclusions about distribution systems. This makes it even more important to thoroughly describe and discuss what methods are used to estimate grid impacts to gain a more concrete understanding of how EVs will affect the grid in the future.

The research findings depend on many unknowns. For example, it is too early to say whether most EVs will be charged at home, public charging points, or battery swap stations. Another unknown is user behavior. We do not know how most drivers—especially consumers—will behave in the future. Many researchers simply apply the mobility behavior of conventional vehicles to EVs. However, any statements about users' future charging behavior are purely speculative.

Analyses of grid impacts pose a dilemma. If we analyze individual, concrete grid topologies, the results cannot be readily generalized. If we adopt a more general approach using supply system models, the analysis cannot be applied to concrete scenarios, but will only outline general trends. Given the lack of knowledge about EVs' future charging behavior, the general approach seems to be the better choice. Integrating EVs into the low voltage supply system will likely increase the load on the grid even more. Thus, the first logical step is to consider what operating parameters limit connections of EVs to the grid and observe how these limitations come about and where they apply.

This chapter analyzes the impacts of EVs on the low voltage supply system in Germany with respect to its capacity. The analysis focuses on reference supply systems

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto284
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

that are identified using structural data. It first determines the simultaneity behavior of EVs based on uncontrolled charging processes. Next, it explains the methodology used to determine the reference supply systems. Then, it describes the main bottlenecks to connect EVs to the power grid. Finally, it analyzes the capacity of the low voltage supply system. Sections 2–4 are intended for readers keen to learn about the underlying methodology. Readers who are more interested in the results should turn directly to Section 5.

These findings were obtained as part of "e-IKT—Integration of Electric Vehicles into Tomorrow's Power Grids," a project supported by the German Federal Ministry of Economics and Technology.

2 SIMULTANEITY CURVES OF CHARGING ELECTRIC VEHICLES

To analyze the ability of low voltage supply systems to handle EVs, we have to look at the simultaneity distribution during the charging process. The simultaneity distribution is expressed by the simultaneity factor. The expected load on the grid P_{tot} produced by n loads is calculated based on the peak load portion P_{s} of the individual loads and the simultaneity factor $g(n)$ (Kaufmann, 1995).

$$P_{\text{tot}} = g(n) \cdot P_{\text{s}} \cdot n \tag{1}$$

In this chapter, simultaneity curves are generated like household patterns of consumption and used to model the charging load of EVs. As comprehensive information is not available about public charging station use, this analysis only considers EVs that are charged at home. Data on home-charging behavior is derived from conventional car data collected in "Mobility in Germany 2008" (MiG, 2008), a study commissioned by the German Federal Ministry of Transport, Building, and Urban Development (BMVBS). To simplify things, the analysis also assumes that driving patterns on the various days of the week remain the same throughout the year.

Rehtanz and Rolink (2009) described how to estimate charging patterns for uncontrolled EVs based on the MiG study. Figure 1 shows the average charging load curve for EVs that are driven during a workday and only charged at home. The three curves represent different charging outputs.

A comparison of the charging profile of the EVs with the German VDEW standard load profile for households (H0 profile) shows peak loads between 17:00 and 20:00 for both households and EVs. These simultaneous load peaks are essential if we wish to use simultaneity factors for both household loads and EVs.

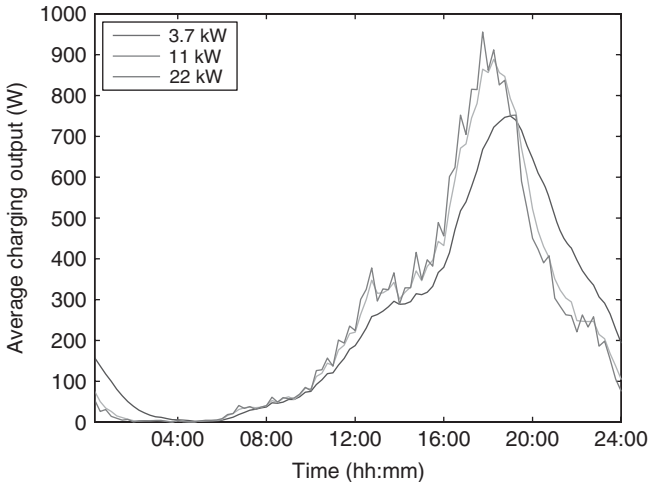


Figure 1. Charging curves for different charging outputs.

To compute the simultaneity curves, we first take 25 random sample populations from all the available "distance traveled" data. The simultaneity curves of these samples are used to determine the simultaneity factor for a given number of vehicles n. Maximum values are used to obtain a conservative estimate for the simultaneity curves.

To simplify matters, the computed simultaneity curves are approximated using a function. Household loads are usually estimated using the function

$$g_1(n) = g_\infty + \frac{(1 - g_\infty)}{n^k} \tag{2}$$

(Approach 1) (Kaufmann, 1995). This result is compared to the result obtained from the exponential function.

$$g_2(n) = g_\infty + (1 - g_\infty) \cdot \mathrm{e}^{-k \cdot n} \tag{3}$$

(Approach 2). The simultaneity factor g_∞ describes the behavior of a very large number of dwelling units or EVs. In both functions, the parameter k is chosen to enable the computed simultaneity curve to be approximated as closely as possible using the least squares method. The total error (distance d) between the computed curve and the approximated curve is minimized using an optimization method. Table 1 shows the resulting exponent k and computed distances for various charging outputs.

The smaller distances show that the exponential approach describes the simultaneity curve much better for all three output levels. Figure 2, for example, shows the computed and exponentially approximated simultaneity factor curve for a charging output of 3.7 kW.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto284
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 1. Parameters and distance of simultaneity functions.

Charging output (kW)	Approach 1			Approach 2		
	g_∞	k	d	g_∞	k	d
3.7	0.367	0.61	0.76	0.367	0.08	0.29
11	0.202	0.69	0.63	0.202	0.11	0.35
22	0.125	0.78	0.44	0.125	0.15	0.35

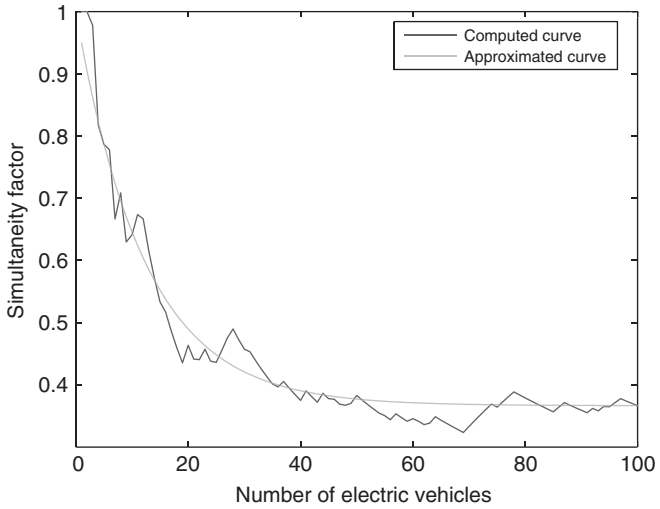


Figure 2. Simultaneity curve for approach 2 and a charging output of 3.7 kW.

3 SELECTING REFERENCE SUPPLY SYSTEMS

This section describes the methodology used to select reference supply systems. Reference supply systems provide the data used in subsequent analyses and should reflect differences in the low voltage supply system as accurately as possible. As the analyses focus on purely residential areas, reference supply systems are selected using residential structural data provided by the Berlin Senate, which contains information on city blocks in Berlin. These blocks are consolidated into groups based on selected parameters. Then, a block is chosen that best represents the set of blocks within each group. Structural data from these representative blocks supply the parameters used in the reference supply systems.

3.1 Choice of methodology

As the data set has a large number of blocks, it makes sense to group them using portioning clustering. This method uses a predetermined number of clusters. Each cluster has a predetermined cluster center. Elements (city blocks) are assigned to whichever cluster has the center that is most similar or least dissimilar to them. An optimization algorithm tries to maximize heterogeneity between the clusters and homogeneity between the elements in each cluster (Backhaus *et al.*, 2006).

Cluster variable selection is extremely important. Variables have a significant impact on cluster composition. Highly correlated variables can overweight certain characteristics because of the redundant information, so variables should be as uncorrelated as possible. These variables can be identified using a correlation analysis. Typical parameters used to describe residential structures and low voltage supply systems include floor space, building footprint, building density, residential density, and the number of dwelling units per service (DPS) entrance (Kaufmann, 1995). A correlation analysis identified floor space and building density as ideal cluster variables.

We can obtain the optimal number of clusters by treating the F-value as a quality measure. The F-value is the quotient of the variance of a variable between clusters and the variance of the variable within the clusters. The higher the F-value is, the greater will be the heterogeneity between clusters and the homogeneity between elements within the clusters. The PRE-value can also be used to determine the number of clusters. It describes the improvement in variability provided by k clusters compared to $(k-1)$ clusters. Using the F-value and PRE-value, the optimal number of clusters was determined to be $k=8$ (Bacher, 1996).

3.2 Selecting representative blocks

After clustering, there are several different ways to identify the best representative for each cluster. One way is to select the block with the most "neighbors." A neighbor is defined as a cluster object whose distance from the reference object lies within a defined range. Obviously, the results vary depending on the limits of the range. Moreover, local accumulations tend to skew the choice of representative (Bacher, 1996). A more global approach is to choose whichever block is the least distant from the cluster center. However, this approach is extremely sensitive to outliers. For this reason, the medoid is used as the representative. The medoid is the block in each cluster with the least average distance (i.e., dissimilarity) from all other objects in the cluster.

The average house distance, average rate of car motorization per household (m), and the average number of DPS entrance are needed to set the reference supply system parameters for subsequent analyses. Table 2 shows the reference supply system parameters.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto284
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Table 2. Reference supply system parameters.

Reference system	House distance (m)	DPS	Motorization rate m
1	21.6	1.28	0.906
2	25.0	13.61	0.357
3	16.8	5.93	0.816
4	20.0	8.28	0.640
5	17.9	10.35	0.817
6	17.0	18.55	0.398
7	8.3	1.39	0.783
8	17.9	10.33	0.391

4 SUPPLY SYSTEM CONSTRAINTS

This section describes the constraints on power distribution systems' ability to support EVs. The main constraints are the load ratings of the *transformers* and the *cables*. This section also examines voltage control and the interrupt condition for feeder fuses.

4.1 Equipment load rating

Load ratings for distribution transformers are mainly determined by thermal operating limits. Exceeding load limits will reduce the transformer's expected service life. Transformer load rating depends on load duration and the temperature when the load begins to rise (Helling *et al*., 1988). Both oil and dry-type transformers are used in Germany to step down voltage from the medium- to the low voltage supply system.

Oil transformers can handle up to 1.3 times their rated load in continuous operation and up to 1.5 times their rated load during peak load periods (Kaufmann, 1995; Helling *et al*., 1988). The large variety of dry-type transformers makes it difficult to generalize their ability to handle above-rating loads (Vosen, 1997). Dry-type transformers are classified into different temperature classes based on their insulation systems. Overtemperatures are defined for each temperature class in the German VDE 0532-76-11 (EN 60076-11) standard. A resin-encapsulated transformer rated above 400 kVA can handle a load of up to approximately 110% in continuous operation and 120% for half an hour at an ambient temperature of 20°C (Kaufmann, 1995).

VDE 0289-8 defines the current-carrying capacity of cables as the maximum current allowed in given operating conditions. Current-carrying capacity is also limited by the maximum permitted operating temperature. Other operating conditions that affect cables' current-carrying capacity include the depth of installation, type of installation, type of soil, soil temperature, and operating mode.

VDE 0276-603 specifies current-carrying capacity for various cable cross sections under reference conditions. Local deviations from the reference conditions (parallel installation of cables) can be accounted for using conversion factors from VDE 0276-1000. Cables can also be overloaded briefly depending on the previous load situation (VDEW, 1986).

If the load rating given in the standards is used for an item of equipment ignoring local operating conditions, it should be noted that the values obtained under reference conditions are only general guidelines. Only a detailed assessment of the actual condition of an item of equipment can determine if it is actually overloaded. As this is normally not done, operators of low voltage supply systems do not know exactly when an item is overloaded. As a result, they do not have any information on remaining equipment capacity in any given load situation.

4.2 Voltage control

IEC 60038 specifies voltage levels for low voltage supply systems. However, devices connected to these supply systems are designed for a certain voltage range. This range must be maintained to preserve device functionality. Under EN 50160, the voltage drop (10 min average) in the medium- and low voltage supply system may not exceed 10% of the rated voltage in 95% of the cases on average for the week. As the 10% encompasses both the low and the medium-voltage supply systems, the maximum permitted voltage drop in the low voltage supply system is usually around 5% of the rated voltage (Heuck, Dettmann, and Schulz, 2007).

4.3 Interrupt condition for feeder fuses

Figure 3 shows the fuses for a secondary substation. A high voltage, high breaking-capacity fuse (HV-HBC fuse) is located on the high voltage side. Low voltage, high breaking-capacity fuses (LV-HBC fuses) are located on the low voltage side. A transformer fuse (LV-gTr) may also be located on the low voltage side. In addition, every feeder on the low voltage side has its own fuse (LV-gL) (Helling *et al*., 1988).

The interrupt condition states that the feeder fuse must be selected to reliably de-energize the line at the smallest expected fault current (single-phase short circuit) (Heuck, Dettmann, and Schulz, 2007), that is, a short circuit between phase and neutral. The maximum line length should thus be chosen to allow the LV-HBC fuse to operate within a certain timeframe at a given current. This value is specified by the conventional fusing current I_f.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto284
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

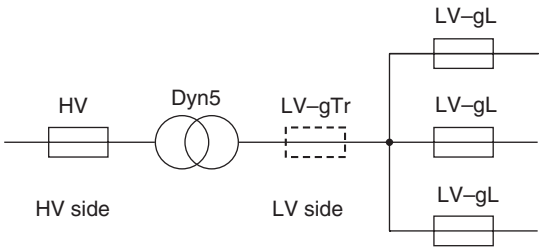


Figure 3. Fuses in a secondary substation.

It is possible to calculate the maximum line length that still maintains the interrupt condition for a minimum fault current. The line length limits the number of connections and affects the maximum permitted voltage drop. For a NAYY[1] $4 \times 150\,\text{mm}^2$ low voltage cable equipped with a 315 A LV-HBC fuse, a transformer with an apparent power rating of $S_N = 100$ kVA and a subtransient short-circuit power of $S_k'' = 50$ kVA at the transformer, the maximum line length is 583 m. This value increases only insignificantly with higher transformer power ratings.

5 ANALYSIS OF THE CAPACITY OF LOW VOLTAGE SUPPLY SYSTEMS

This section examines the capacity of low voltage supply systems to absorb EV charging. In this case, capacity depends on what percentage of vehicles in a particular segment of the supply system can be replaced by EVs, not the absolute number of EVs that the segment can support. This percentage is expressed as the penetration rate p, which is calculated by dividing the number of locally replaceable vehicles by the total quantity of vehicles. The analysis uses the reference supply systems. This section first explains the model assumptions underlying the supply system analysis. Then, it presents the capacity analysis.

5.1 Creating the model

The system capacity analysis relies on several basic assumptions that are required to analyze equipment loading and the voltage drop. A general approach is used to reduce complexity. Several simplifying assumptions are made. Special local circumstances are ignored.

This analysis focuses on home charging. In other words, it generally assumes that EVs will be charged at home and ignores commercial users of EVs. This represents a worst-case scenario from the perspective of a low voltage supply system, which predominantly supplies dwelling units. The vehicles are plugged into an AC or a three-phase power outlet. Charging devices are integrated in the vehicles. DC charging stations are not considered, as the sojourn times at home are supposed to be generally long enough for charging, especially over night. Therefore, the high charging power offered by DC charging stations is not requested. The charging output is assumed to be 3.7, 11, or 22 kW based on VDE 0122-1 (EN 61851-1). The analysis assumes that all vehicles use the same charging output. It does not consider combinations of different charging outputs in order to reduce the complexity of the investigations. The analysis assumes that vehicles being charged on a single phase with 3.7 kW of power are symmetrically distributed among the various phases, so they can be treated like three-phase burdens. Charging rectifiers are assumed to be operated with a power factor of $\cos\varphi_{EV} = 1$, that is, 100% active power, and constant charging current. This assumption is widely valid for many of the currently available EVs.

Exemplary, all the consumers are assumed to be households with an electrification level (EL) of 2. In other words, electric power is used for cooking and other applications, but not for heating or generating hot water. The parameters for the simultaneity function (2) are set to $g_\infty = 0.15$ and $P_s = 8$ kW (Kaufmann, 1995). A conservative power factor of $\cos\varphi_{HH} = 0.9$ (ind.) is used for the household loads (German Government, 2010). The analysis assumes that all the supply systems have a noninterconnected radial topology. This leads to long cables and consequently to high voltage drops. A low voltage cable is assumed to be installed on each side of the street where service entrances are located. Thus, the average distance between the house connections can be assumed according to Table 2. The analysis also assumes that equipment is not loaded past its rated capacity.

The capacity analysis is performed based on expected peak loads. Equipment loading is estimated using a value called *base-level load*. The base-level load is the peak equipment load before the load from the EVs is added. In our model, the base-level load is solely produced by household loads. As this value varies widely depending on local conditions and equipment, it is set up as a variation parameter. The number of households is determined using the simultaneity function (2). The remaining reserve capacity is then incorporated into (3) to calculate the maximum number of EVs that can be connected without violating any operating limits. The number of dwelling units and the number of EVs are used to calculate the maximum penetration p_{max}.

5.2 Analysis of secondary substations

The German VDE 0532 series of standards specifies preferred values for distribution transformer ratings. This

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto284
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

analysis assumes a rated output of 400 kVA. To minimize costs, utilities generally aim for distribution transformers to have an initial load of around 70% (Kaufmann, 1995). As such, the base-level load varies between 60% and 90% of the rated output. The analysis ignores any voltage effect for connected loads, and so the number of low voltage feeders and the line lengths are irrelevant.

Figure 4 charts maximum penetration levels ($p_{\max}$) against base-level loads for the various reference electricity supply systems at a charging output of 11 kW. At a base-level load of 70%, the penetration level ranges from 30% to 80%. This value sinks, as the base-level load rises. Indeed, it is less than 6% in all reference supply systems once the base-level load reaches 90%.

Analyzing the penetration level at different charging outputs shows that only a fraction of the conventional vehicles can be replaced by EVs, given a relatively high base-level transformer load. Even at low loads, it is impossible to completely replace all the vehicles in most cases. The lower the vehicle charging output is, the more vehicles the supply system can support. Supply systems 2, 6, and 8 tend to have higher values because of their lower motorization levels.

Transformers' overload capacity can be used when integrating large, individual loads in the low voltage supply system. If the transformers' overload capacity is exceeded, more powerful transformers can be installed in their stead. If, however, loads are integrated throughout the electricity supply system, it makes more sense to add transformers to the new high load areas and create new supply system segments by relocating the segmentation points. In Germany, the grid was resegmented in this manner in the 1960s when many off-peak storage heaters were connected to low voltage supply systems. Section 5.3 explores the extent to which low voltage power cables limit the ability to connect EVs to the grid.

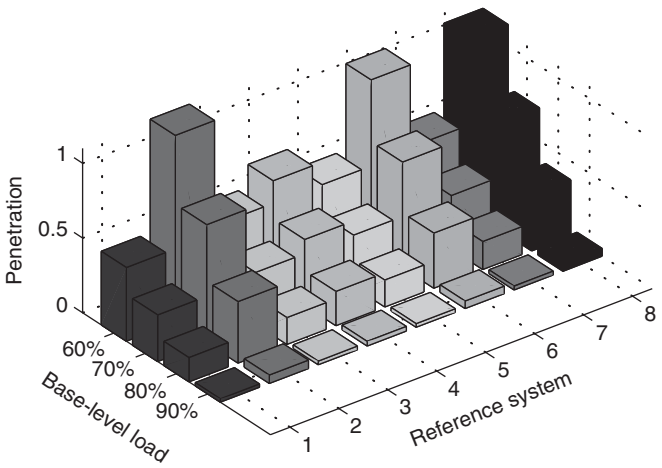


Figure 4. Maximum penetration at a charging output of 11 kW.

5.3 Analysis of low voltage cables

A capacity analysis of low voltage cables requires a different approach, as it needs to consider equipment load, voltage control, and interrupt conditions. This is illustrated by the flow chart in Figure 5. First, the maximum penetration level $p_{\max,I}$ is calculated based on current-carrying capacity. The next step is to determine whether the value is restricted by the voltage range limit. This is done by deriving the line length from the number of households. If the interrupt condition is violated for the computed line length, the algorithm aborts. If the interrupt condition is met and no voltage problems occur, the algorithm ends. In this case, the maximum penetration level is determined by current-carrying capacity. If the interrupt condition is met and an unacceptable voltage drop occurs, the number of vehicles is reduced until the voltage limits are met. The maximum penetration level is then outputted as $p_{\max,U}$.

This analysis uses NAYY $4 \times 150\,\text{mm}^2$ low voltage cables. Their maximum current-carrying capacity is assumed to be $I_z = 275\,\text{A}$. A maximum load utilizes around 50% of the cables' capacity, assuming an optimum grid design (Heuck, Dettmann, and Schulz, 2007). The following analyses vary the base-level load of the main lines between 50% and 80%.

5.3.1 Analysis of current-carrying capacity

Figure 6 shows the maximum penetration level at a charging output of 11 kW. At a base-level load of 50%, it is possible to replace at least 80% of all existing vehicles. The penetration level falls rapidly once the base-level load exceeds 50%. At a base-level load of 80%, this value is only between 6% and 15%. Supply systems 2, 6, and 8, by contrast, have higher penetration levels because of their low levels of motorization.

As long as the base-level load remains below 50%, EVs can replace all the vehicles in the reference electricity supply systems for all three charging outputs. This changes at higher base-level loads. While a charging output of 3.7 kW per car leaves plenty of leeway to integrate EVs, this leeway diminishes as the charging output increases. The supply system's capacity becomes extremely constrained when the base-level load exceeds 60% and the charging output is 22 kW per EV.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto284
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

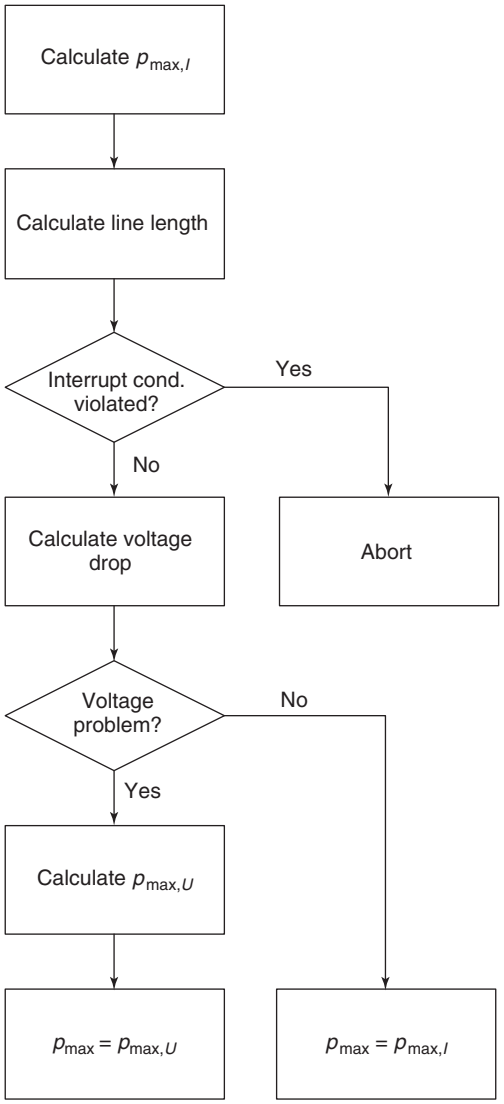


Figure 5. Procedure for analyzing capacity.

5.3.2 *Analysis of voltage control*

This section analyzes voltage control for various load situations. Locally supplying reactive power can increase the grid's transmission capacity (VDEW, 2001). For that reason, the analysis also considers the impacts of having EVs supply capacitive reactive power. First, the general process is explained, starting with the fundamentals. Then, the results are presented.

The voltage drop is estimated mathematically instead of performing a more precise power flow study. The analysis only looks at nonbranching lines—the worst-case scenario in terms of voltage control. In keeping with VDEW (2001), the power factor is assumed to be $\cos\varphi_{EV} = 0.9$ when the EVs supply capacitive reactive power.

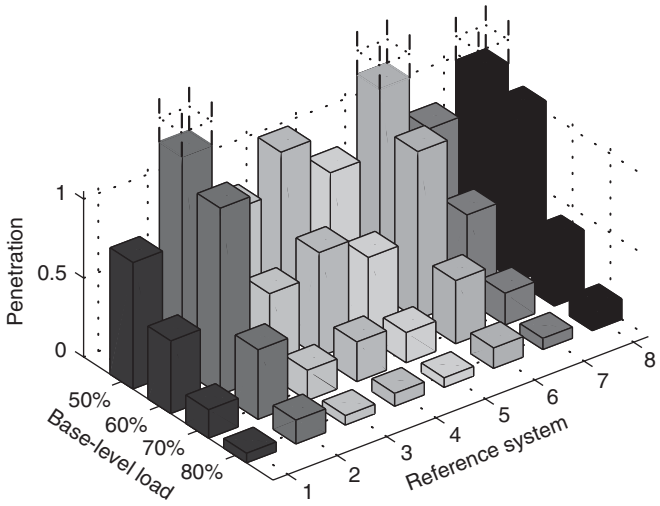


Figure 6. Maximum penetration at 11 kW charging output.

5.3.2.1 Fundamentals. In radial grids, loads spaced along a line can be converted to a lumped load located at the end of a virtual line with a length of $l' = \varepsilon \cdot l$, where it produces the same voltage drop as is observed at the end of the line in the original grid. The conversion factor ε depends on the local load distribution and the size of the individual loads. The maximum allowable line length l_{max} for a given load current I and a maximum voltage drop ΔU_{max} can be approximated by the following function (Nagel, 1994):

$$l_{max} = \frac{\Delta U_{max}}{\sqrt{3} \cdot \varepsilon \cdot I \cdot (r \cdot \cos\varphi + x \cdot \sin\varphi)} \tag{4}$$

In this function, r and x represent the specific line parameters. The load current $\overline{I}$ of the lumped load can be represented as the sum of all individual currents $\overline{I}_i$ at the service entrances:

$$\overline{I} = I \cdot (\cos\varphi + j \cdot \sin\varphi) = \sum_{i=1}^{h} \overline{I}_i \tag{5}$$

The individual currents are approximated based on the complex apparent power consumed by the loads at nominal voltage. The parameter h describes the number of service entrances. The analysis assumes that household and vehicle loads are evenly distributed along the line and that the connected loads are approximately identical. The conversion factor ε can thus be estimated as follows (Nagel, 1994):

$$\varepsilon = \frac{h+1}{2 \cdot h}, \quad \text{for} \quad 0.5 < \varepsilon \leq 1 \tag{6}$$

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto284
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

5.3.2.2 General approach. The following approach was used to analyze voltage control. First, the line length l_{HH} is calculated for the reference electricity supply system without the vehicle loads. This requires the number of dwelling units to be determined based on the base-level load of the line. The line length can then be calculated based on the average number of dwelling units per service entrance and the average distance between houses (Table 2). The analysis ignores scenarios that violate the interrupt condition. The maximum line length l_{EV} is calculated using (4) for the load scenario with EVs. If the inequality 7 is true, there is a voltage control problem.

$$l_{EV} < l_{HH} \tag{7}$$

If a voltage control problem is found to exist, the next step is to determine the maximum penetration level $p_{max,U}$ that breaches the voltage range limit. This is done by reducing the number of EVs until the voltage falls within the allowable range. This is an iterative process, as the voltage situation has to be reexamined after every reduction. Finally, the maximum allowable penetration level p_{max} is defined as:

$$p_{max} = \min\{p_{max,I},\ p_{max,U}\} \tag{8}$$

To ensure comparability, the active component of charging output was retained in the power factor correction, and an additional capacitive reactive component was added to the charging current. This is a more expensive approach, as it requires a larger charging converter. Problems can also occur if equipment limits are reached at a power factor of $\cos\varphi = 1$.

5.3.2.3 Findings. Reference supply systems 2 and 7 have voltage problems. Reference supply system 2 also violates the interrupt condition and so is not included in the analysis. Table 3 shows maximum penetration levels for reference supply system 7 at various base-level loads and charging outputs. Penetration levels are not constrained until the base-level load exceeds 50%. As the table shows, power factor correction improves the maximum penetration level. Given a base-level load of 60% and a charging output of 11 kW, penetration improves from $p_{max,U} = 20\,\%$ to $p_{max,Q} = 39\,\%$. It does not, however, improve penetration indefinitely. The improvements become insignificant once the base-level load reaches 70%.

Table 3. Maximum penetration for reference supply system 7.

Charging output		Cable base-level load			
		50%	60%	70%	80%
3.7	$p_{max,I}$	1.6	1.00	0.61	0.28
3.7	$p_{max,U}$	1.6	0.67	0.03	0.00
3.7	$p_{max,Q}$	1.6	0.85	0.05	0.00
11	$p_{max,I}$	0.93	0.54	0.20	0.07
11	$p_{max,U}$	0.93	0.20	0.00	0.00
11	$p_{max,Q}$	0.93	0.39	0.00	0.00
22	$p_{max,I}$	0.74	0.41	0.06	0.03
22	$p_{max,U}$	0.74	0.07	0.00	0.00
22	$p_{max,Q}$	0.74	0.13	0.00	0.00

6 SUMMARY

This chapter examined the impact of EVs on the low voltage electricity supply system in Germany. It determined simultaneity patterns for charging EVs, identified reference supply systems for analysis and addressed supply system constraints. These findings were then used to identify general tendencies for the ability of low voltage supply systems to support EVs. The analyses focused on purely residential areas. Our findings suggest that the German federal government's goal of one million EVs by 2020 should cause relatively few problems.

Distribution transformers tend to be the limiting factor for EV deployment. Where transformer loads are high, EVs can replace only a fraction of the conventional vehicles. Indeed, replacing all the vehicles will be hard to achieve, even when loads are low. It is, however, usually fairly easy to replace transformers with more powerful units (Meyer, 1982). For that reason, this analysis looked particularly closely at low voltage power lines.

Line capacity tells a more optimistic story. At line loads of up to 50% of total capacity (not counting EV loads), it is possible to replace all the vehicles associated with dwelling units connected to a particular line in every scenario. Higher loads reduce line capacity—in some cases, by a significant margin. At a charging output of 3.7 kW per car, the electricity supply systems were able to handle the replacement of a large number of conventional vehicles with EVs.

Voltage problems caused by EVs are only likely to occur in unusual cases. In most situations covered in this analysis, EVs exceed the lines' current-carrying capacity before breaching the voltage range limits. Some constraints should be expected if long lines are already heavily loaded before EVs are added. If voltage range limits are violated, EVs equipped with charging converters can provide capacitive reactive power to raise supply system voltage. This increases the percentage of vehicles that can be replaced by EVs. However, this kind of power factor correction only offers improvements within a narrow range.

Until now, low voltage supply systems have been designed to handle expected peak loads. Equipment capacity cannot, however, be fully utilized, as the utilities

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto284
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

do not know their equipment's operating conditions in detail. Moreover, reserves freed up in low load periods are rarely used. These reserves can only be used to charge EVs if the supply system's condition is monitored and charging processes are coordinated. Homes and places of work are ideal charging sites, because the vehicles are parked for long periods at a time.

ENDNOTE

1. Cable type description according to the German standard VDE 0276-603.

REFERENCES

Babaei, S., Stehen, D., Tuan, L. A., *et al*. (2010) Effects of Plug-in Electric Vehicles on Distribution Systems: A Real Case of Gothenburg. *IEEE Innovative Smart Grids Conference Europe*, Gothenburg.

Bacher, J. (1996) *Clusteranalyse*, Oldenbourg.

Backhaus, K., Erichson, B., Plinke, W., and Weiber, R. (2006) *Multivariate Analysemethoden*, Springer.

Boonekamp, P. G. M. and Wakkerman, L. G. J. (1993) *The electric vehicle and electricity distribution*. 26th International Symposium on Automotive Technology and Automation, Aachen.

Clement-Nyns, K., Haesen, E., and Driesen, J. (2010) The impact of charging plug-in hybrid electric vehicles on a residential distribution grid. *IEEE Transactions on Power Systems*, **25** (1).

German Government (2010) *Niederspannungsanschlussverordnung* (*NAV*), September 3.

Helling, K., Kaufmann, W., Nagel, H., and Piehl, E. (1988) *Zur Betriebsweise von Ortsnetztransformatoren*.

Heuck, K., Dettmann, K.-D., and Schulz, D. (2007) *Elektrische Energieversorgung*, Vieweg.

Heydt, G.T. (1983) The impact of electric vehicle deployment on load management strategies. *IEEE Transactions on Power Apparatus and Systems*, **PAS-102** (5).

Kaufmann, W. (1995) *Planung öffentlicher Elektrizitätsverteilungs-Systeme*, VWEW.

Meyer, A. (1982) *Analyse von Niederspannungsnetzen in klein- und mittelständischen Versorgungsgebieten*.

Nagel, H. (1994) *Systematische Netzplanung*, VDE-Verlag.

Rahman, S. and Shrestha, G.B. (1993) An investigation into the impact of electric vehicle load on the electric utility distribution system. *IEEE Transactions on Power Delivery*, **8** (2).

Rehtanz, C. and Rolink, J. (2009) Conditions for the demand side management of PHEVs and EVs. ETG Congress, Düsseldorf, Germany.

Richardson, P., Flynn, D., and Keane, A. (2010) Impact assessment of varying penetrations of electric vehicles on low voltage distribution systems. *IEEE Power and Energy Society General Meeting, Minneapolis*.

Taylor, J., Maitra, A., Alexander, M., and Duvall, M. (2009) Evaluation of the impact of plug-in electric vehicle loading on distribution system operations. *IEEE Power and Energy Society General Meeting, Calgary*.

VDEW (1986) *Kabelhandbuch*, VWEW.

VDEW (2001) *Eigenerzeugungsanlagen am Niederspannungsnetz*, 4th edn, VWEW.

Vosen, H. (1997) *Kühlung und Belastbarkeit von Transformatoren*. VDE-Schriftenreihe 72, VDE-Verlag.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto284
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Ceramic Sensors for Power Train ECU Systems

Akira Fujii, Takehiro Watarai, and Akitoshi Mizutani

DENSO Corporation, Kariya, Japan

1 Introduction 1
2 Ceramic Sensor Applications in Gasoline Engines 1
3 Ceramic Sensor Applications in Diesel Engines 10
4 The Future of Ceramic Sensors 14
Related Articles 14
References 14
Further Reading 14

1 INTRODUCTION

Although there is currently no fixed definition of the term *ceramic*, it generally refers to an *inorganic compound* comprising two or more materials such as metallic, metalloid, or nonmetallic elements. Under this definition, it should be possible to precisely control the chemical composition, microstructure, microshape, and manufacturing processes of the ceramic materials to achieve the desired function. However, there are exceptions to this general definition. For example, diamond, which consists only of carbon (C), as well as silicon carbide (SiC) and aluminum nitride (AlN) may be referred to as a *semiconductor* or a ceramic depending on the application (Table 1).

These types of ceramic materials have a wide range of properties and are widely used in vehicles both as structural materials and as sensors. Table 2 shows the general properties and applications of ceramic materials.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Materials used for automotive sensors (particularly powertrain sensors which is installed in engine body and exhaust pipe) commonly use ionic bonded oxides formed between metals and nonmetals. This is because the electrically biased bonds created by ionic bonding have many useful electrical properties. These oxides are regarded as a special genre of materials with high mechanical strength that can also withstand high temperatures. This chapter describes the required functions and detection principles of these ceramic sensors. It also outlines the basic technologies that enable the practical adoption of ceramic materials as sensors, and some of the key products. These issues are addressed focusing on typical powertrain sensors such as the O_2 sensor, air/fuel (A/F) sensor, NO_x sensor, temperature sensors, and knock sensors in gasoline and diesel engines, which represent the two main forms of combustion used in vehicles. In this way, the aim is to clearly identify categories of sensors falling under different combustion methods based on variations in installation position and environment, detection objects and methods, as well as the forms of sensor products.

2 CERAMIC SENSOR APPLICATIONS IN GASOLINE ENGINES

2.1 Overview

Figure 1 shows the installation positions of some typical ceramic sensors in a gasoline engine.

A gasoline engine functions by compressing and combusting intake air and injected fuel in a combustion chamber. The combusted waste gas is then directed to a catalytic converter through the exhaust manifold and exhaust pipes to treat toxic components. Along this route, an A/F sensor or an O_2 sensor is provided immediately

Table 1. Relationship of compound categories and ceramic materials.

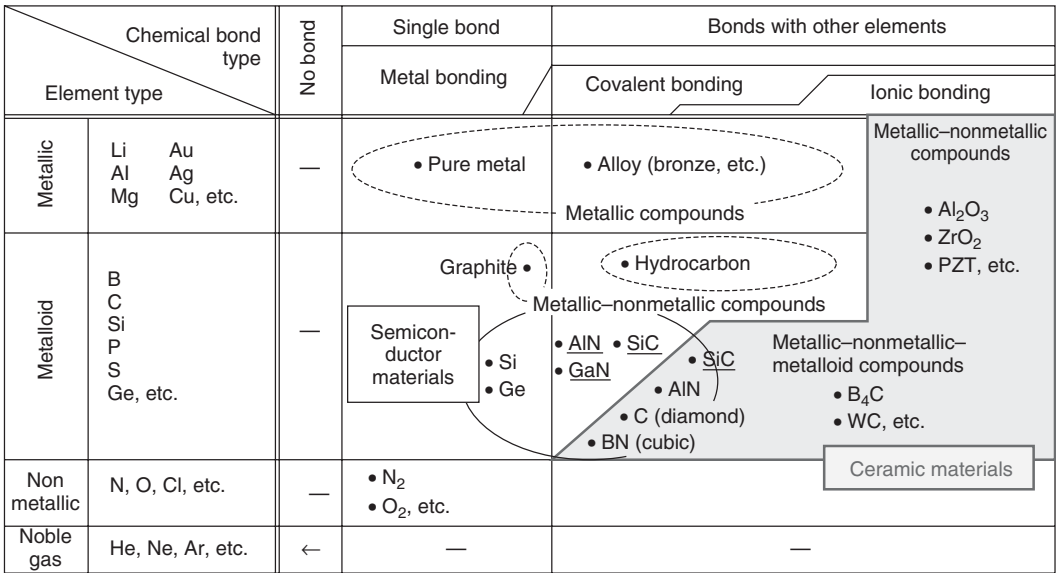


Table 2. Functions of ceramic materials and automotive applications.

Properties		Functional capabilities	Applications
Mechanical		- High-strength - Light - Deformation resistant	- Turbo rotors - Substrates for catalyst carriers - Particulate filters - Heat sink
Thermal		- High melting point - Little thermal expansion - High or low thermal conductivity - Converts heat energy to electrical energy	
Electrical	Ionic conductivity	- Ions conduct as carriers	- O_2 sensors, A/F sensors, NO_x sensors
	Thermistor properties	- Resistance changes due to heat	
	Piezoelectricity	- Mechanical $\Longleftrightarrow$ electrical energy conversion	- Temperature sensors
	Pyroelectric and thermoelectric	- Thermal $\Longleftrightarrow$ electrical energy conversion	- Knock sensors
	Varistor properties	- Resistance changes due to electric field strength	- Varistors
	Electron conductivity	- Conducts or acts as semiconductor	- Semiconductors
	Dielectric properties	- Insulates from electricity - Stores electricity	- Plug insulators, - Capacitors
Magnetic		- Becomes magnetized	- Magnets for motors
Photo		- Transmits and absorbs light - Converts light to electricity	- Ultra-violet ray blocking glass

below the exhaust manifold, and another O_2 sensor is installed after the catalytic converter. The purpose of the A/F sensor or O_2 sensor immediately below the exhaust manifold is to control combustion by detecting and transmitting the concentration of oxygen in the exhaust gas so that toxic components can be efficiently eliminated by the catalytic converter. Either type of sensor is used based on differences in the required format of the oxygen concentration signal.

In contrast, the sensor after the catalytic converter is almost always an O_2 sensor because of the required signal format and the moist environment in this location. This O_2

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

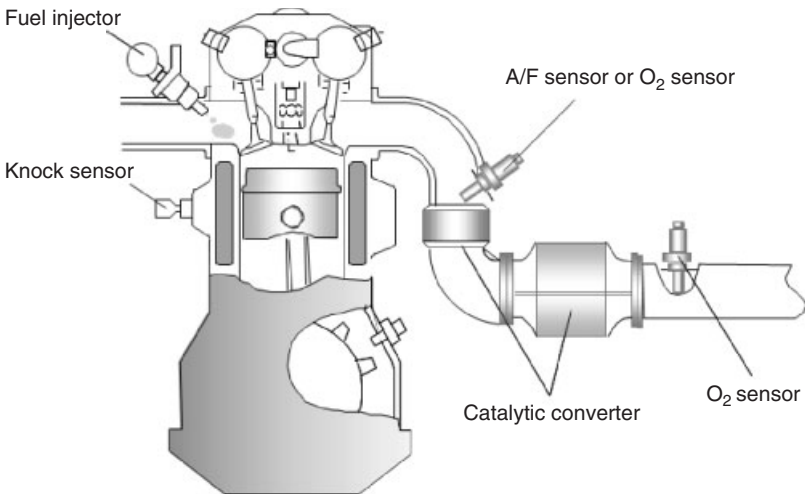


Figure 1. Installation positions of ceramic sensors in gasoline engine.

sensor uses oxygen concentration to detect small quantities of NO_x and hydrocarbons (HCs) that leak from the catalytic converter as the catalyst degrades. However, as a result of continuing advances in research, further optimized sensors are beginning to enter the market (Ito, 2011).

The knock sensor is installed on the engine block close to the combustion chamber. If the combustion frequency is in a range equivalent to that generated by engine knocking, the knock sensor converts the mechanical vibration into a direct electrical charge and transmits this value to the electronic control unit (ECU). The ECU then controls combustion to prevent knocking.

The sensors described earlier are installed in the necessary locations to obtain the signals for controlling combustion or treating the exhaust. These positions are optimized factoring in the restrictions of the environmental conditions. The following sections describe each sensor in more detail.

2.2 O_2 sensor

2.2.1 *Required functions*

In a gasoline engine, O_2 sensors are used to control combustion. The power of a gasoline engine is determined by the amount of fuel supplied with respect to the intake air volume. The mass of air divided by the mass of fuel in the mixture is referred to as the *A/F ratio*. The actual combustion reaction in the cylinder is defined by the following formula:

$$C_xH_y + \left(x + \frac{y}{4}\right) O_2 \rightarrow xCO_2 + \frac{y}{2}H_2O$$

This chemical reaction formula shows that there is an optimum A/F ratio at which the fuel is completely combusted and converted into carbon dioxide and water. For a gasoline engine with 100% octane fuel (C_8H_{18}), 14.7 g of air are necessary to completely combust 1 g of gasoline. This means that the optimum A/F ratio is approximately 14.7. The A/F ratio that satisfies this relationship is known as the *theoretical* A/F ratio. The amount of deviation from the theoretical A/F ratio is called the *excess air ratio*, which is generally depicted using the lambda symbol (λ). This is defined as follows:

$$\lambda = \frac{\text{actual A/F ratio}}{\text{theoretical A/F ratio}}$$

When λ is calculated, a state with a relatively high amount of fuel compared to the theoretical A/F ratio (i.e., $\lambda < 1$) is referred to as the *rich* condition and a state with a relatively low amount of fuel (i.e., $\lambda > 1$) is referred to as the *lean* condition. Under theoretical A/F ratio conditions, $\lambda = 1$.

In a conventional gasoline engine, combustion is controlled so that $\lambda = 1$, which represents the most efficient range for exhaust treatment. To achieve this A/F ratio, a ceramic O_2 sensor with ionic conductivity is installed directly under the exhaust manifold. This sensor examines exhaust gas directly after combustion and transmits the postcombustion oxygen concentration to the ECU.

2.2.2 *Detection principle*

Figure 2 shows an outline of the element used in an O_2 sensor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

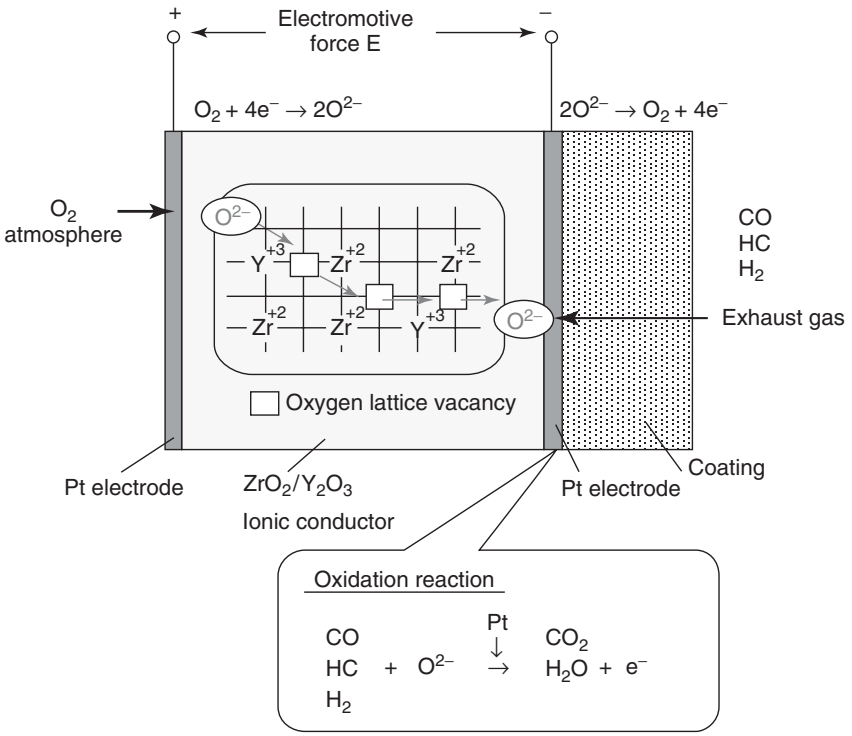


Figure 2. Outline of O_2 sensor element.

In this sensor, the material used to detect the oxygen concentration is a solid solution of zirconia and yttria (ZrO_2/Y_2O_3). This material is placed between platinum (Pt) electrodes to form the element. At high temperatures of approximately 300°C when a difference in oxygen concentration occurs between the electrodes on each side, oxygen ionizes because of the catalytic action of the Pt and transfers through oxygen vacancies in the zirconia crystal lattice. The movement of these ions is detected as a voltage. The resulting electromotive force E is then applied in accordance with the following Nernst equation:

$$E = \frac{RT}{4F} \ln \frac{(P_{O_2})_I}{(P_{O_2})_{II}}$$

F : Faraday constant

R : gas constant

T : absolute temperature

$(P_{O_2})_I$: oxygen partial pressure on atmosphere electrode side

$(P_{O_2})_{II}$: oxygen partial pressure on exhaust electrode side

The electromotive force E exhibits different aspects under lean and rich conditions. Under lean conditions, virtually no electromotive force is generated because there is hardly any difference between the oxygen concentrations at the atmosphere side (approximately 20%) and exhaust side electrodes. In contrast, under rich conditions, unburned HC, CO, and H_2 are combusted at the exhaust side electrode by the oxidative action of the Pt. Consequently, electromotive force E is generated in excess of the Nernst equation. This oxidation reaction continues until the combustion reaches chemical equilibrium and the O_2 concentration in the exhaust gas drops promptly after the reaction is completed. As a result, electromotive force increases dramatically at the rich side when lower than theoretical A/F ratio conditions. This electromotive force acts as the output of the O_2 sensor. Figure 3 shows the relationship between the A/F ratio and the sensor output and catalyst treatment efficiency. When electromotive force is generated, the ECU judges that the state has switched from lean to rich and outputs a fuel injection signal to recover the lean conditions. Therefore, the ECU control uses changes in the signal from the O_2 sensor as a trigger to increase or decrease the fuel injection quantity. In this way, the combustion is constantly controlled to a state close to the theoretical A/F ratio,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

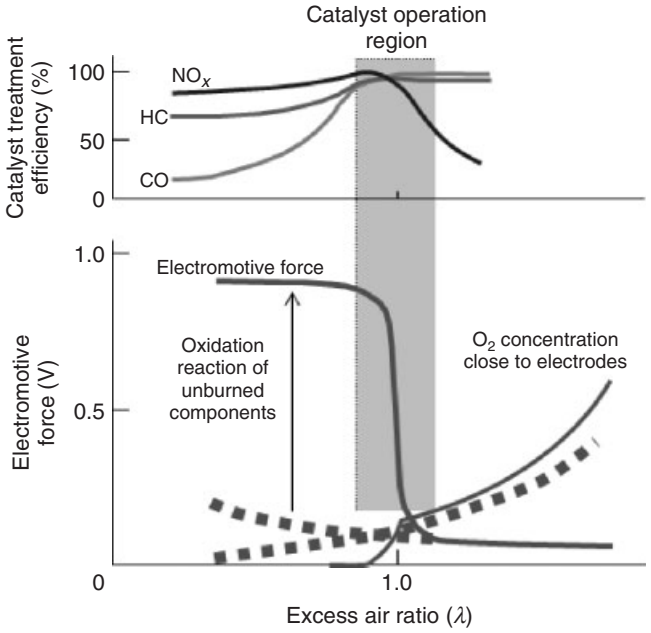


Figure 3. Relationship between A/F ratio and sensor output and catalyst treatment efficiency.

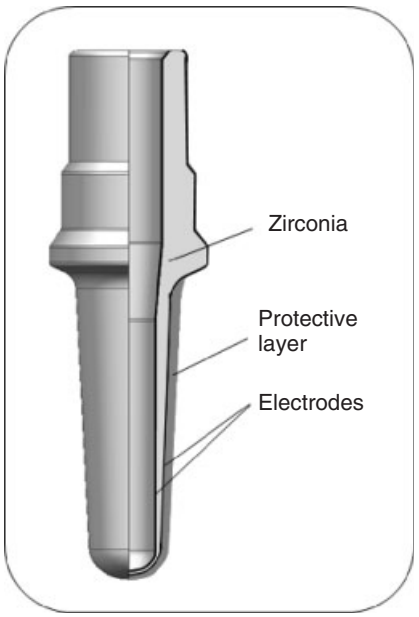


Figure 4. Detection element of cup-type O_2 sensor.

and the exhaust gas treatment rate is kept high at all times.

2.2.3 *Technology and product overviews of O_2 sensor*

Figure 4 shows the structure of the detection element in a cup-type O_2 sensor.

The detection element uses a cup-shaped zirconia/yttria ionic conductor as the base material, enclosed on the inner and outer sides by Pt electrodes. The outer electrode is exposed to the exhaust gas and the inner electrode acts as the atmosphere electrode. The outermost shell of the exhaust side electrode is coated to protect the electrode from the exhaust gas. Figure 5 shows the structure of the O_2 sensor.

An O_2 sensor consists of an element for detecting the A/F ratio, a heater for the zirconia, element covers to protect the sensing portion from exposure to items such as condensation in the exhaust pipe, and a case that outputs the sensor signals and the like. As O_2 sensors operate within the exhaust gas, the operating environment is subject to high temperatures, vibration, and foreign substances. For this reason, material selection and structural design from the element to the case must consider resistance against heat, vibration, and foreign substances. Sealing performance is also a critical design requirement to prevent external leaks

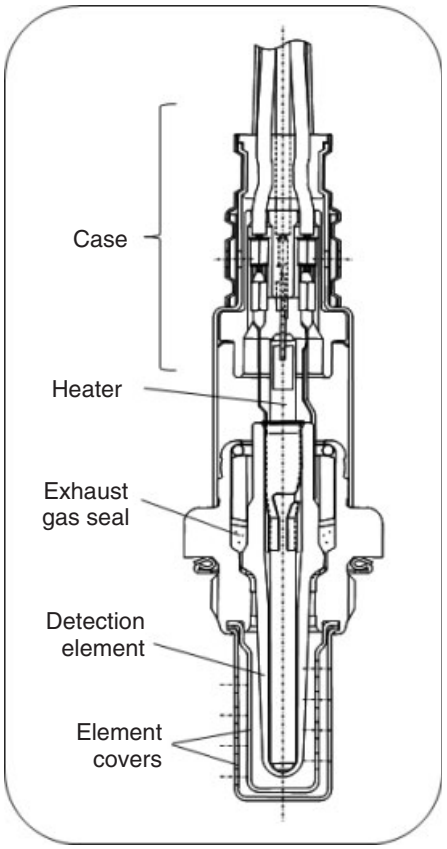


Figure 5. Structure of O_2 sensor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 6. Appearance of O_2 sensor.

of exhaust gases. Figure 6 shows the appearance of an O_2 sensor that satisfies these technological and design requirements.

2.3 A/F sensor

2.3.1 Required functions

In contrast to an O_2 sensor, which outputs a trigger signal using the theoretical A/F ratio ($\lambda = 1$) as the threshold value, an A/F sensor meets strict low emission vehicle (LEV) exhaust regulations by outputting a continuous signal over a wide range of exhaust gas atmospheres and activating immediately after engine start. These two features enable even more precise combustion control. The key points for realizing these two features are how to make the ionic conductor trigger into a continuous action and how to activate the sensor as quickly as possible. Although the material, characteristics, and functions of the ionic conductor are the same as described in Section 2.2.2, the detection principle and product configuration of the A/F sensor to achieve these two features greatly differ from that of an O_2 sensor.

Recent requirements of A/F sensors include the detection of variations in combustion between cylinders and greater robustness with respect to H_2 generated by engine combustion. Research and development of A/F sensors that satisfy these requirements has been published (Su *et al*., 2011; Yamamoto *et al*., 2010; Suzuki *et al*., 2010).

2.3.2 Detection principle

In the same way as O_2 sensors, A/F sensors also use the conduction of oxygen ions for detection. However, A/F sensors utilize a different element structure to detect the A/F ratio continuously under the principle of early activation (Figure 7).

The A/F sensor detection element has a rectangular external shape. The internal structure is layered with cavities and dense structures. The cavity in the center of the element allows the inflow of the atmosphere. The zirconia ionic conductor and two Pt electrodes are located between that cavity and a diffusion resistance layer that allows inflow of the exhaust gas. The diffusion resistance layer performs the same role as the O_2 sensor coating. However, it also has an upper shield layer that provides a certain distance and resistance for the exhaust gas flowing to the exhaust gas side electrode. A heater is embedded in the bottom portion of the element. The heater, which mainly comprises Pt on an alumina-base (Al_2O_3) material, heats the zirconia by conduction. This structure enables the A/F ratio to be detected continuously and immediately after engine start.

In an A/F sensor, under rich conditions, voltage is applied (in a process known as *pumping*) so that the atmospheric oxygen in the internal cavity is transported electrochemically as oxygen ions. This is carried out to calculate the necessary O_2 volume for reacting with the unburned gas under rich conditions. This O_2 volume is called the *pump current* (*IP*). This current is virtually proportional to the concentration of unburned gas, that is, the A/F ratio. In this control, it is necessary to limit the movement of unburned gas using the level of pumping performance. For this reason, a diffusion resistance layer is formed. In contrast, under lean conditions, a voltage is applied to expel excessive O_2 that passes through the diffusion resistance layer. The direction of applied voltage is reversed in either rich or lean conditions using the theoretical A/F ratio as a threshold.

Under rich conditions, IP does not increase in accordance with the applied voltage but reaches saturation because the diffusion resistance layer limits the diffusion of the exhaust gas components. This saturated current is called the limit current (IL). In contrast, under lean conditions, the O_2 volume that passes through the diffusion resistance layer

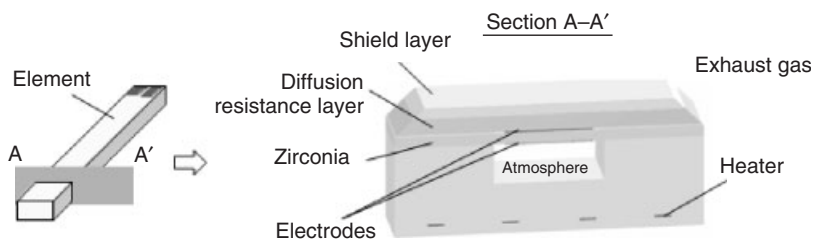


Figure 7. Detection element of A/F sensor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

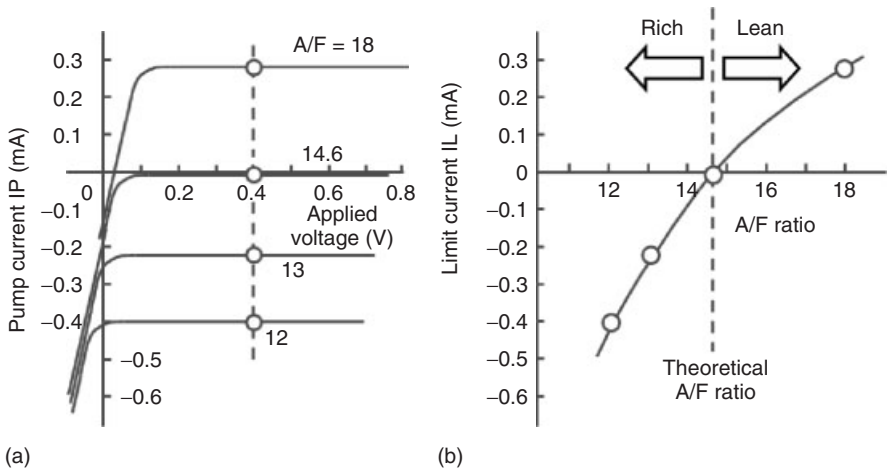


Figure 8. Control and output characteristics of A/F sensor. (a) Relationship between applied voltage, IP, and A/F ratio. (b) Relationship between A/F ratio and IL.

is detected as the emitted current. Therefore, the A/F ratio can be detected continuously in a wide range from A/F = 10 to A/F = atmosphere by measuring IL in both rich and lean conditions. Figure 8 shows the output characteristics of the A/F sensor.

The control and output of the A/F sensor described earlier starts to be obtained when the element temperature is maintained at around 700°C. For this reason, the element is embedded with a conduction heater mainly composed of Pt, which enables early activation of the sensor immediately after engine start. Activation takes only a few seconds.

2.3.3 Technology and product overviews of A/F sensor

Figures 9 and 10 show the structure and appearance of an A/F sensor, respectively. The sensor consists of a rectangular detection element, a cover to protect the sensing portion from exposure to items such as condensation in the exhaust pipe, and a case that outputs the sensor signals and the like.

2.4 Knock sensor

2.4.1 Required functions

In a conventional gasoline engine combustion process, the mixture is ignited by a spark plug, which then spreads throughout the cylinder by flame propagation. However, self-ignition of mixture around the outside of the cylinder before the conventional flame propagation process results in the generation of a shock wave. Cylinder block vibration caused by this shock wave is then transmitted to the atmosphere. This phenomenon is referred to as *knocking*.

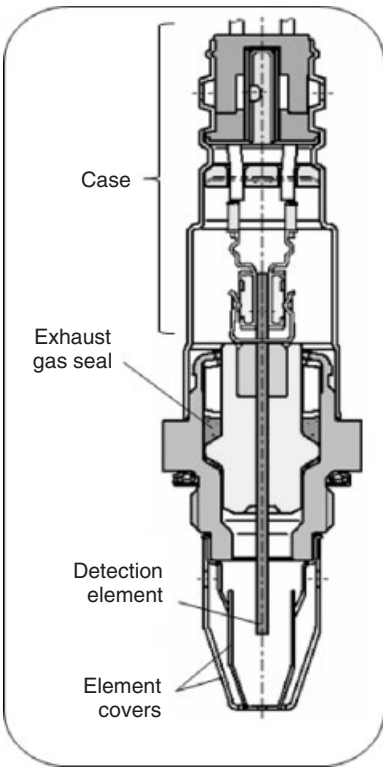


Figure 9. Structure of A/F sensor.



Figure 10. Appearance of A/F sensor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

In a knock control system, the knock sensor is directly attached to the cylinder block. Piezoelectric ceramic material provided in the knock sensor detects the vibration caused by knocking and transmits it to the ECU as a voltage signal. The purpose of the knock control system is to improve both engine power (output torque) and fuel efficiency by controlling the ignition timing to a position extremely close to the engine knock region, without actually entering that region. It accomplishes this control by retarding and advancing the ignition timing in accordance with the degree of knock.

2.4.2 *Detection principle*

The detection element built into the knock sensor is mainly composed of lead zirconate titanate (PZT), a compound that is expressed by the following chemical formula: $Pb(Zr \cdot Ti)O_3$. The sensor uses the piezoelectric properties of this material (Table 2) to convert the applied vibration into electrical charge, which is then output to the ECU. Figure 11 shows the crystalline structure of PZT.

The PZT compound shown in the figure has a cuboid crystal lattice that is referred to as *tetragonal*. Each unit of this crystal possesses a quantity of electricity called *spontaneous polarization* because the position of the central Zr or Ti ion is displaced from the lattice center. However, if this material is left in a bulk-sintered state, the direction of the spontaneous polarization shown by the arrow in the figure becomes random. As a result, the total quantity of electricity possessed by the bulk may cancel itself out. Although the material cannot be used as a detection element in this state, the spontaneous polarization direction can be aligned by applying a high voltage to the bulk so that the bulk possesses a quantity of electricity. If mechanical displacement is applied to the detection element, the total quantity of electricity possessed by the bulk changes, which enables the admission and release of electrical charge. This property is called the *piezoelectric effect*. Figure 12 illustrates a state in which electrical charge is generated by compression.

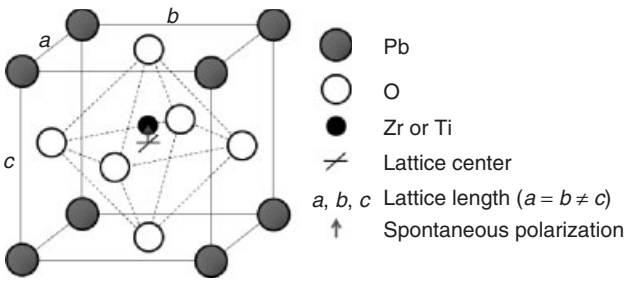


Figure 11. Crystalline structure of PZT.

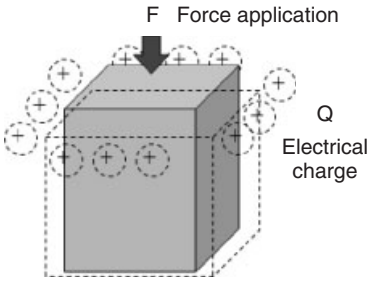


Figure 12. Piezoelectric effect.

Although the electrical charge generated by mechanical displacement is small, a sizeable charge can be obtained if the frequency of the applied vibration matches the resonance frequency as determined by the mass, dimensions, and Young's modulus of the PZT bulk, as resonant vibration causes the mechanical displacement to increase. The knock sensor generates its electrical charge after exposure to displacement caused by the resonance of the assembly. Therefore, the resonance frequency of the detection element (which is determined by the total frequency generated by the components used to fix the knock sensor, the weight used to increase the mechanical displacement, and the physical properties of the surrounding parts) must either be designed to match the frequency of engine knocking (i.e., a resonant knock sensor) or a large enough detection element or additional weight must be provided so that a sufficient electrical charge is generated even without resonance (i.e., a nonresonant knock sensor). Conventional detection elements have an outside diameter of several millimeters and a thickness of approximately 0.5 mm in the case of resonant knock sensors. Nonresonant knock sensors are cylindrical and have an outside diameter of several tens of millimeters and a thickness of approximately 5 mm. Both types have a silver electrode coated on the top and bottom surfaces of the PZT material to extract the charge. A hole is provided in both types to facilitate the fixing of the sensor.

In a resonant knock sensor, the detection element consists of a piezoelectric ceramic material adhered to a plate that acts as a diaphragm. The detection element is embedded in the sensor and generates voltage because of the resonance that occurs when the knock frequency matches the natural frequency of the diaphragm. In contrast, a nonresonant knock sensor combines the piezoelectric ceramic material with a weight to form the detection element. Vibration from the engine changes the force on the weight (i.e., generates acceleration), which compresses the piezoelectric ceramic material and generates the voltage. This type of sensor has flat frequency characteristics without a resonance point in

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

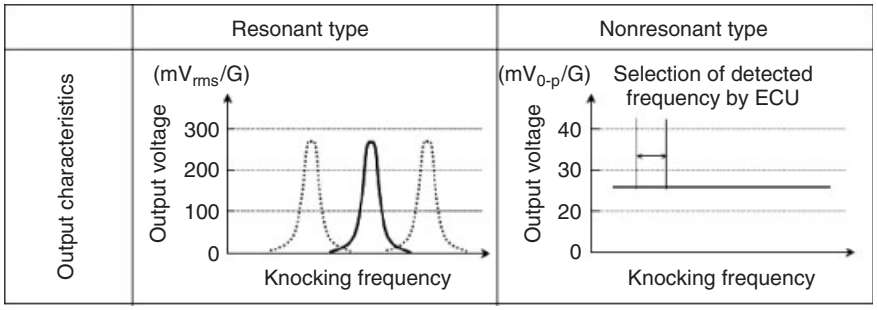


Figure 13. Output characteristics of knock sensors.

the detection frequency range. Figure 13 shows examples of the output from both types of knock sensors.

It should be noted that the output voltage on the vertical axes in Figure 13 differ in accordance with the size of the device and the permittivity (equivalent to the capacitance) of the material, as well as the size of the resonance gain and acceleration. Therefore, the absolute output value of a particular knock sensor may differ from that shown in the figure.

Nonresonant type sensors have several merits. The use of this type of sensor allows the system to electrically select the desired signal frequency and to change the selected frequency in accordance with the engine load. This reduces the work hours required to adapt the sensor to the engine and also improves the accuracy of the system. For these reasons, nonresonant knock sensors are the most commonly adopted type. The output of a nonresonant knock sensor is expressed by the following equation:

$$V = \frac{m \bullet a \bullet g_{33}}{C}$$

V output voltage
M mass of weight
a acceleration
g_{33} piezoelectric constant
C capacitance of PZT.

2.4.3 *Technology and product overviews of knock sensor*

The knock sensor is directly attached to the engine cylinder block to detect knocking. Figure 14 shows the structures of a resonant and nonresonant knock sensor.

A resonant knock sensor contains a thin circular element pierced with a hole that is adhered onto a plate to adjust the resonance frequency to the engine knocking frequency. In contrast, a nonresonant knock sensor has a thick cylindrical element pierced with a hole that is attached to a base along with a weight. Both types must be attached with care as the adhesion and tightening conditions as well as the state of the part itself affect the resonance frequency and output characteristics. Figure 15 shows the appearance of a nonresonant knock sensor and Figure 16 shows its attached state.

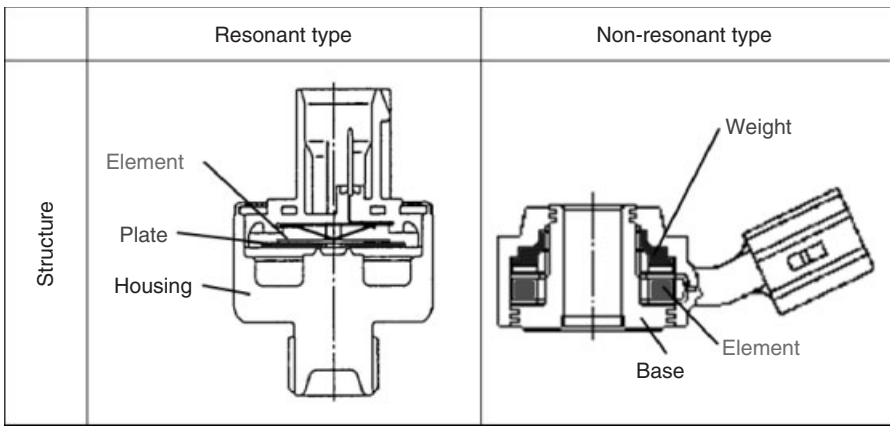


Figure 14. Knock sensor structures.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 15. Appearance of nonresonant knock sensor.

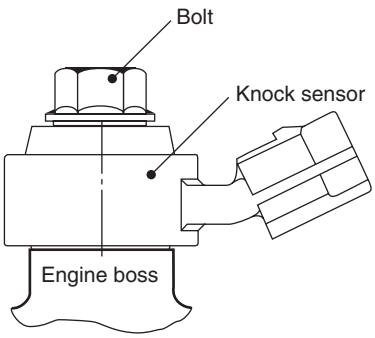


Figure 16. Installed knock sensor.

3 CERAMIC SENSOR APPLICATIONS IN DIESEL ENGINES

3.1 Overview

Diesel engines are more thermally efficient than gasoline engines and, therefore, have lower CO_2 emissions. However, because diesel engines with high power and low emissions are more complex than gasoline engines, a more sophisticated system configuration and more sensors are required. This section describes the NO_x sensor and exhaust gas temperature sensor (EGTS) as examples of ceramic sensors used in a diesel engine. Figure 17 shows the installation position of these sensors.

3.2 NO_x sensor

3.2.1 Required functions

A diesel engine has a larger intake air volume than a gasoline engine. It compresses this air in the cylinder, creating a high temperature that causes directly injected fuel to self-ignite and combust. However, incomplete combustion leads to the generation of soot, and NO_x is generated if the combustion temperature is too high. A NO_x sensor detects the concentration of NO_x to the order of parts per million (ppm) generated by diesel engine combustion. It is used to control and run diagnostics on the exhaust emission control devices.

3.2.2 Detection principle

Figure 18 shows the detection element of a NO_x sensor. The element comprises a pumping cell for expelling oxygen and a sensing cell for detecting NO_x. The electrodes of the former are Pt, whereas the latter is mainly rhodium (Rh). The cells are connected to a heater that enables early activation by electrical heating.

The detection principle is similar to the A/F sensor. O_2 and NO_x enter a chamber through the diffusion resistance layer. O_2, which enters the chamber in the order of parts per hundred (percent) as opposed to NO_x that enters as ppm, undergoes ionic conduction faster than ppm NO_x. As a result, this O_2 blocks NO_x detection and causes noise. Therefore, NO_x is reacted by itself at the detection electrode after the O_2 is expelled by the oxygen expulsion electrode. The detection electrode uses an Rh catalyst to separate NO_x into N_2 and O_2. Ionic conduction is performed within the zirconia, and the O_2 quantity is detected as an electrical current to measure the NO_x concentration.

3.2.3 Technology and product overviews of NO_x sensor

The structure of a NO_x sensor consists of a detection element, covers to protect the sensing portion from exposure to items such as condensation in the exhaust pipe, and a case that outputs the sensor signals and the like. In addition, to enable the detection of NO_x at the ppm level, the sensor is integrated with a circuit to output the detection results. More recently, the size of the control circuit is becoming smaller to improve mountability. Figure 19 shows the structure of a NO_x sensor.

3.3 Exhaust gas temperature sensor (EGTS)

3.3.1 Required functions

Vehicles use a wide variety of sensors to measure the temperature of coolant, oil, intake air, exhaust, and the like. Of these, an EGTS is used in several systems, such as for diesel particulate filter (DPF) control, NO_x catalyst control, urea SCR (selective catalytic reduction) control,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

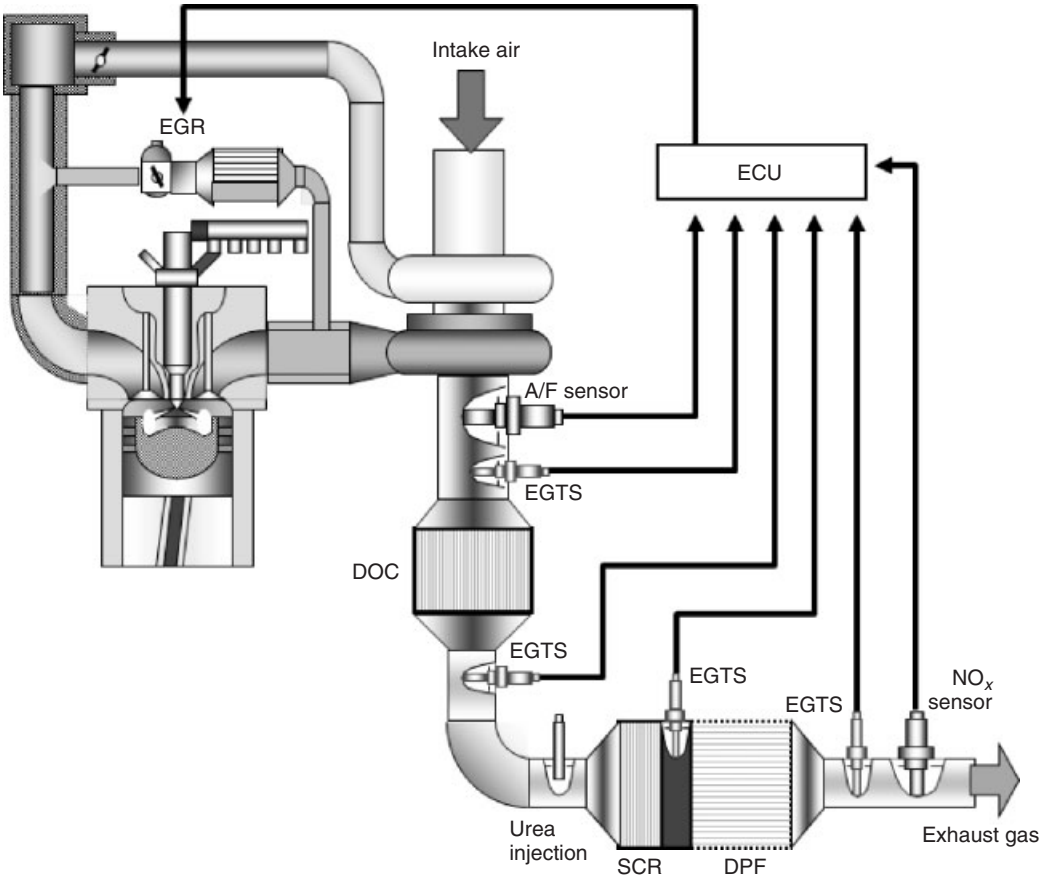


Figure 17. Installation positions of ceramic sensors in diesel engine.

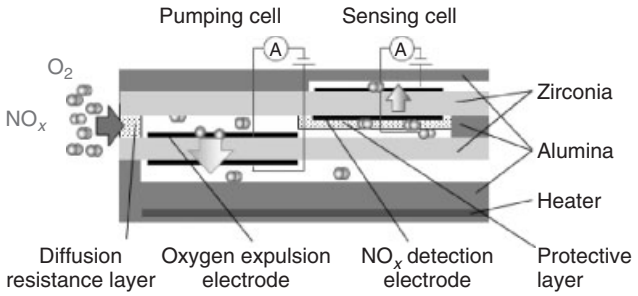


Figure 18. Detection element of NO_x sensor.

turbocharger protection, catalyst on-board diagnostics (OBD), and the like. For this reason, an EGTS is required to measure a much wider temperature range (from low temperatures to approximately 1000°C) than a normal temperature sensor. It is also required to measure temperature highly accurately with a fast response to enable precise exhaust gas temperature feedback control. Another requirement is the durability to resist both high temperatures and vibration. Although some automotive EGTS use the temperature characteristics of metal resistors or a thermocouple, ceramic sensors use the characteristics of a thermistor (Table 2).

3.3.2 Detection principle

A thermistor is a type of resistor with an electrical resistance that varies widely in accordance with changes in temperature. Some thermistor materials are categorized as having a positive temperature coefficient (PTC), in which the resistance increases as the temperature rises. Others have a negative temperature coefficient (NTC) in which the resistance decreases as the temperature rises. NTC thermistors are widely used for automotive EGTS.

The relationship between temperature and resistance in an NTC thermistor is expressed by the following equation:

$$R = R_0 \exp\{B(1/T - 1/T_0)\}$$

where R is the thermistor resistance at the absolute temperature T, R_0 is the thermistor resistance at the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

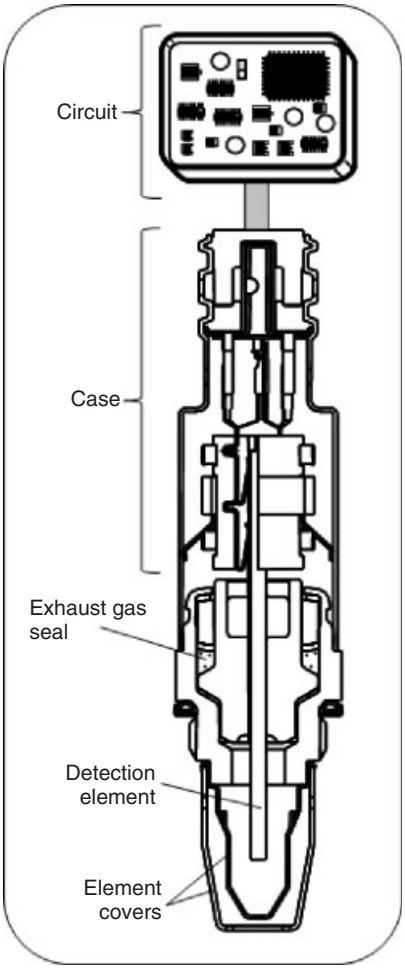


Figure 19. Structure of NO_x sensor.

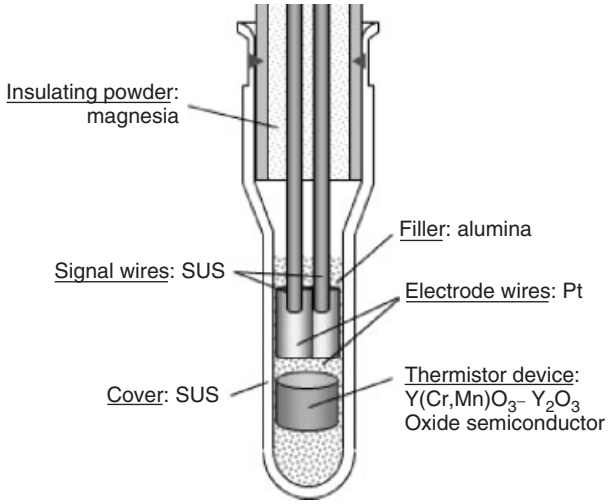


Figure 20. Detection element of EGTS.

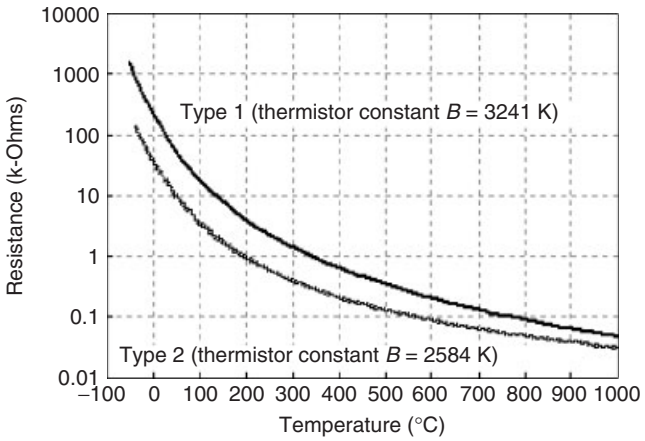


Figure 21. Typical thermistor characteristics.

absolute temperature T_0, and B is the thermistor constant. The thermistor constant B expresses the sensitivity to temperature and differs depending on the thermistor material.

Conventional thermistors satisfy the requirement for high temperature durability by using perovskite oxide semiconductor materials because of their excellent stability at high temperatures. Homogenous mixing and extra-precision molding technology of thermistor materials enable highly accurate measurement as well as the development of smaller thermistors and faster response by reducing the size of the temperature-sensitive portion. Figure 20 shows the configuration of the temperature sensing detection element of an EGTS.

The thermistor material is a mixture consisting of a $Y(Cr \cdot Mn)O_3$ oxide semiconductor material and a Y_2O_3 insulator. The proportion of the mixture can be varied to adjust the thermistor resistance and the thermistor constant B can be adjusted by changing the ratio of Cr and Mn in the semiconductor material. Figure 21 shows the typical characteristics of a thermistor. Type 2, which has a smaller thermistor constant B, is capable of measuring a wider temperature range than type 1.

3.3.3 Technology and product overviews of EGTS

The detection element shown in Figure 20 is provided with a cover to prevent degradation because of exposure to the exhaust gas. It also has a ceramic filler (alumina) inside the metal cover that holds the thermistor device to prevent vibration from fracturing the electrode wires (Pt). These measures ensure that the EGTS satisfies the requirement for high durability.

Figure 22 shows the basic structure of an EGTS. It is screwed to the exhaust pipe by a boss attached to the pipe

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

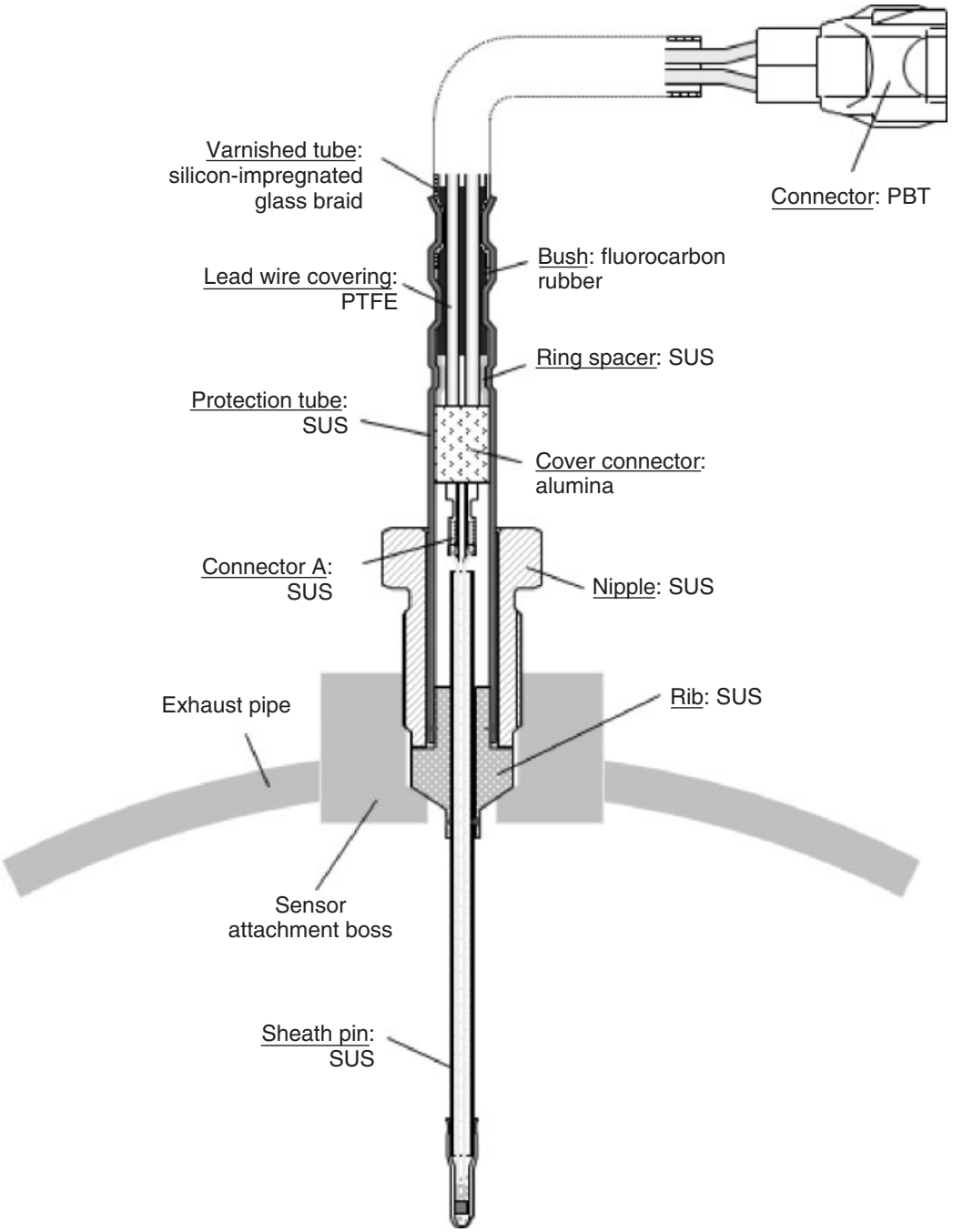


Figure 22. Basic structure of EGTS.

and a nipple. A caulked rubber bush acts a waterproof protection tube. Generally, the EGTS protrudes into the exhaust pipe by 30–50 mm.

Figure 23 shows the appearance of a general EGTS and a long-type EGTS. There is increasing demand for the long-type EGTS because of the recent requirement for more accurate measurement of the temperature in the center of the DPF to improve the efficiency of DPF regeneration. As increasing the length of the EGTS has a major negative effect on durability, a long-type EGTS with the vibration-damping structure shown in the figure has also been developed (Hori and Todo, 2009).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

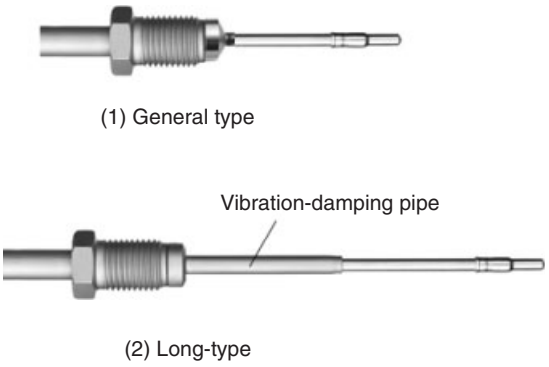


Figure 23. Appearance of EGTS.

4 THE FUTURE OF CERAMIC SENSORS

Engine control is likely to become even more sophisticated and complex as demands for CO_2 reduction and the depletion of fossil fuel resources intensify. Consequently, the number of required parameters will increase, making detection more difficult. The detection of new parameters is another likely requirement in the future. Smart sensor functions that use communication to increase the efficiency of development and adoption work and sensor simulations compatible with model-based development are also beginning to enter the mainstream.

Ceramic sensors are currently being used in locations that best fulfill their capabilities, such as those that require the unique functions of ceramic materials and high durability. However, many ceramics are expensive or use toxic materials, and some are difficult to obtain a stable supply. Other methods will have to be considered unless sensors can be developed that meet the requirements of the customer at a reasonable cost.

Further research is required to enable wider future acceptance of ceramic sensors capable of playing a key role in society and to ensure that ceramics are not left behind as development trends change.

RELATED ARTICLES

Engine Management Systems
Gas Aftertreatment Systems
Exhaust Emission Control Considerations for Diesel Engines
Stoichiometric Exhaust Emission Control
Various Types of Sensors

REFERENCES

Hori, T. and Todo, Y. (2009) High accuracy exhaust gas temperature sensor with anti-resonance structure. *SAE Technical Paper*, 2009-01-0641.

Ito, M., Kobayashi, K., Watarai, T., Kayama, R., and Sasaki, T. (2011) Development of high sensitivity oxygen sensor. *JSAE Journal*, **95** (11), 9–12.

Su, Z. Imamura, H. Kajiyama, N. Nakato, M. (2011) Development of high response A/F sensor. *China Automotive Engineering*, **33** (102011186), 911–913.

Suzuki, Y., Tanaka, A., Itakura, T., Sasaki, T., Nishijima, H., and Tomioka, S. (2010) Development of high accuracy A/F sensor with catalyst layer. *SAE Technical Paper*, 2010-01-0042.

Yamamoto, M., Suzuki, Y., Itakura, T., and Sasaki, Y. (2010) Development of high accuracy A/F sensor with catalyst layer. *Transactions of Society of Automotive Engineers of Japan*, **133** (10), 15–18.

FURTHER READING

Ito, Y. and Yokoi, H. (2012) Overview and trend of automotive ceramic sensors. *Ceramics Japan*, **47**, 431–435.

Kingery, W.D., Bowen, H.K., and Uhlmann, D.R. (xxxx) *Introduction to Ceramics* Wiley Series on the Science and Technology of Materials. ISBN: 0-471-47860-1

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto285
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Body Design, Overview, Targeting a Good Balance between All Vehicle Functionalities

Gerhard F. K. Tecklenburg

Hochschule für Angewandte Wissenschaften (HAW), Hamburg, Germany

1 Introduction 1
2 Master Processes for the PEP 3
3 Packaging/Ergonomics Process and Concept Development 5
4 Technical Support of Aesthetic Design (styling) Exterior and Interior 7
5 Class-A Surfacing/Styling/Techniques Convergence 10
6 Digital Prototyping and Digital Mock-Up (DMU) 13
7 Concept Developments on the Example of Closures 16
8 Concept Competition and Supplier Integration 20
9 Summary 22
Related Articles 22
References 22
Further Reading 23

1 INTRODUCTION

Comprehensive and careful planning is required to meet the numerous conflicting design and manufacturing constraints in the development of a new passenger car that is to be produced in large batches.

Is it necessary to update a vehicle in production (facelift) or to develop the replacement of a successful car? Does it make sense to fill a market niche with a novel vehicle concept or to develop a new market (e.g., emerging markets such as China or India)? Which basic model and which derivates on the basis of which platform should be developed of which should a platform be developed? A number of duties and responsibilities are generated for development, production, and sales from the answers to these questions.

Sales influences the decision-making process for every new vehicle. Therefore, before and parallel to the development process, detailed actions of market and competition analysis are taken up to define and approve vehicle characteristics (e.g., body shape, variants of equipment) and the exact competition positioning (e.g., market segment, customer orientation) in complete lists of objectives for design (styling) and concept development. From these lists of objectives, lists of requirements are elaborated in full detail by project management, advanced design (styling), and concept development for all vehicle areas and assemblies.

Internal and external, national and international driving tests with competitor and company cars in the early concept phase are part of the competition analysis to systematically compare, for instance, full vehicle integration, interior, and access to the vehicles and/or components. Disassembling of competitor vehicles delivers detailed information regarding material, weight, manufacturing sequences, part and tooling costs, and so on.

Furthermore national and international car clinics are organized in the different phases of development to confront potential customers with latest design models, seating mock-ups, concept cars, or prototypes, and to conduct, record, and evaluate detailed interviews.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

Toward the end of the concept development phase, sales defines the ideal variants of equipment together with technical development, design (styling), finance, and production according to technical, creative, and, in particular, financial arguments.

According to market launch strategies, for example, for the exploration of a new market niche, show cars are presented on motor shows more than two years before start of production (SOP) to selectively evaluate the resonances of potential customers, dealers, and specialized press. Sales develops adequate communication/ advertising and market launch strategies with, for instance, extensive press and dealer presentations with pilot series vehicles under support of digital processed material before SOP.

To start the development according to the preparations of sales, all relevant initial strategic plans are chalked out by a product strategy committee. The overall responsibility is assumed by a member of the chief executive committee of the original equipment manufacturer (OEM) or by a management representative appointed for the project. Once the product strategy committee has decided to launch a new automobile project, the product evolution process (PEP) is prepared, comprising product planning, product development, and production planning. For this purpose, the OEM has defined master processes permitting the detailed planning of the project and the reliable validation of the maturity levels. The nature of the development project (e.g., a face-lift or a completely new passenger car) determines the duration and the details of the development and planning scope of the required product development processes.

Independent of a specific automobile project, the OEM and system suppliers continuously monitor the market. This often leads to general investigation and research projects that are independent of specific car or system developments. The PEP comprises analyses, definitions, and coarse and detailed design steps, which are iteratively performed, often jumping back and forth within the steps, and finally leading to optimized solutions.

The development of a new automobile is generally performed independent of the powertrain development. Several other modules and components such as front seats or axles and suspension systems are rarely developed with reference to a specific car project, that is, they remain unchanged in their basic design and are only adapted to the special requirements of the car project.

The tasks of product development of the automotive body and full vehicle integration are shared between different technical disciplines (Figure 1) situated at the OEM, the suppliers, and engineering services working for the OEM. In the discipline of systems layouts, the main engineering task is to develop the electric and electronic concepts for driving, comfort, operating, display, infotainment, air-conditioning systems, and so on, with efficient energy management and most intelligent logic. Important boundary conditions are cost efficiency, design space, ability for diagnoses, and safety. The main tasks of engineers working for the discipline of functional layouts are to lay out or proof and optimize components of sub- and total systems according to loads and strains defined for driving performance, active and passive safety, emissions, climatic and acoustic comfort, and so on, by legislation, institutions for consumer safety assessment, or the OEM itself. The main boundary conditions are to minimize the use of material, design space, energy, weight, and so on, and to maximize the performances. The automotive body engineers working for the discipline of geometrical forms and functions layouts

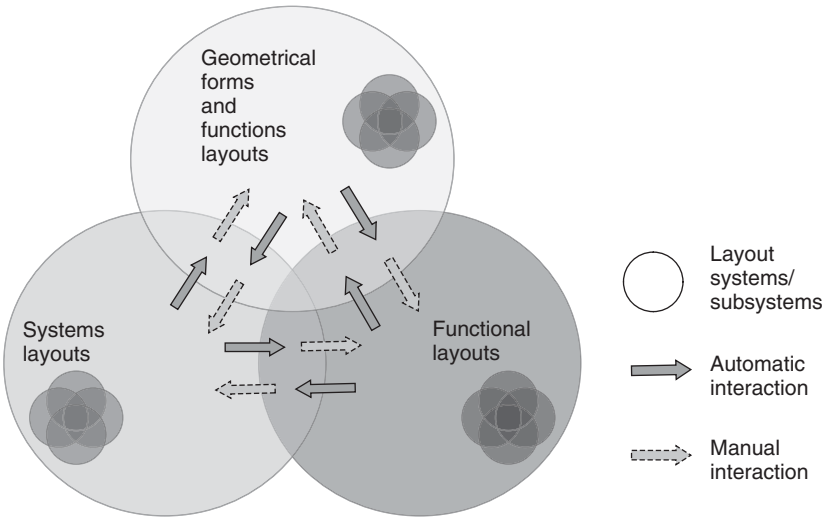


Figure 1. Interaction between layout systems (Albers, 2012). (Reproduced by permission of Thomas Albers.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

are responsible for laying out the shape and safeguard the geometrical functions of parts and assemblies of the automotive body, exterior and interior systems, accounting for ergonomic requirements, styling demands, and existing platform and modularization concepts. Relevant requirements are defined by legislation, the geometric function of the parts and assemblies, as well as tooling and assembly (Albers, 2012).

Besides the layout work, automotive body engineers are responsible for the fulfillment of all system, functional, and geometrical requirements for the components they authorize. Body engineers control the PEP for their components from the early packaging and styling phases to the mass production phase after SOP according to reliability, completeness, and currentness of the components, while the contribution of the other divisions (e.g., powertrain, seating systems) to the PEP often is set as temporary but highly sophisticated auxiliary and supporting work (Albers, 2012).

The majority of development tasks for styling, design, calculation, or simulation, for example, are supported by virtual product development tools implemented in different software and hardware platforms. The advantage of these virtual tools is the early verification of design requirements. Their disadvantage is the rudimentary or nonexistent interaction between the virtual systems. This leads to several parallel modeling efforts of the same geometry or part. The leading system for geometry definition and verification is parametric associative design (PAD). While in the 1950s only a few automotive body engineers were involved in the development of a new automotive body, working on one project, today a large number of engineers from different disciplines are involved in each project, and they are all working on several projects in parallel. The resulting complex interrelations and interdependencies in product development process cannot be controlled by one technical discipline with the aid of its approved discipline-specific methods. Consequently, the automotive industry uses approaches and methods of systems engineering. After elicitation and analysis of all relevant requirements for the system or product to be developed, a reasonable architecture or product structure is conceived. At the same time, well-defined interfaces between subsystems or components belonging to different technical disciplines are identified. The detailed design of the discipline-specific components takes place under consideration of these interfaces and the existing interrelations. It requires continued cooperation of the different technical disciplines. In many aspects, this development principle corresponds with the principles of methodic design. But the principles of systems engineering are not limited to addressing classic questions of mechanical design. They also imply important aspects of project management, reliability, and risk assessment (Abulawi, 2012).

2 MASTER PROCESSES FOR THE PEP

The master processes for the development of the automotive body and the integration of powertrain and chassis describe and control the different maturity processes in the product evolution, for example, the maturity processes for the automobile packaging, the styling, the module definitions, the prototype phase, or the mass production phase based on the analysis of several independent publications (e.g., Braess and Seiffert, 2007a). The master processes of three German OEMs (BMW, Daimler, and Volkswagen) are shown in Figure 2 and can be compared. According to publications, the duration of the master process of a conventional automobile development is approximately 48–60 months. The project timing is structured with the aid of defined phases and milestones (Volkswagen), quality gates (Daimler), or synchronization points (BMW). The milestones are defined control points (often documented in project reports) at which certain results or findings must be available and a certain product maturity level must be achieved. It is the objective of the master processes to define the maturity level validation (to ensure the reliability) in the PEP. Phases and milestones are defined for each new project before the project is kicked off.

Nowadays, most OEMs have adopted platform strategies for deriving different models and production lines from a standardized design basis. Generally, a platform is defined as the reusable design of the floor pan components, which is of indirect influence on the styling and design of the visible parts of the body-in-white (BIW) but has uniform reference points for the automotive body production (e.g., part breakdown, clamping and positioning points, and conveyor adaptation) and for attaching powertrain components such as engine, transmission, and wheel and suspension system. The sophisticated process of defining platform concepts is based on comprehensive analyses and research projects carried out by all OEMs. The consequence of the platform strategy is that the early phases of car body design can and must use a set of carryover parts (COPs) that are available as 3D computer-aided design (CAD) models. The selection of the correct platform components for the product family must occur at the beginning of the design project and must be verified and validated as soon as possible with the highest achievable reliability. BMW subdivide their master PEP, which follows the strategic product planning phase, into the following two major phases: “preparatory phase” and “serial (detailing) phase,” each of which is allowed to take approximately the same number of months. Both major

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

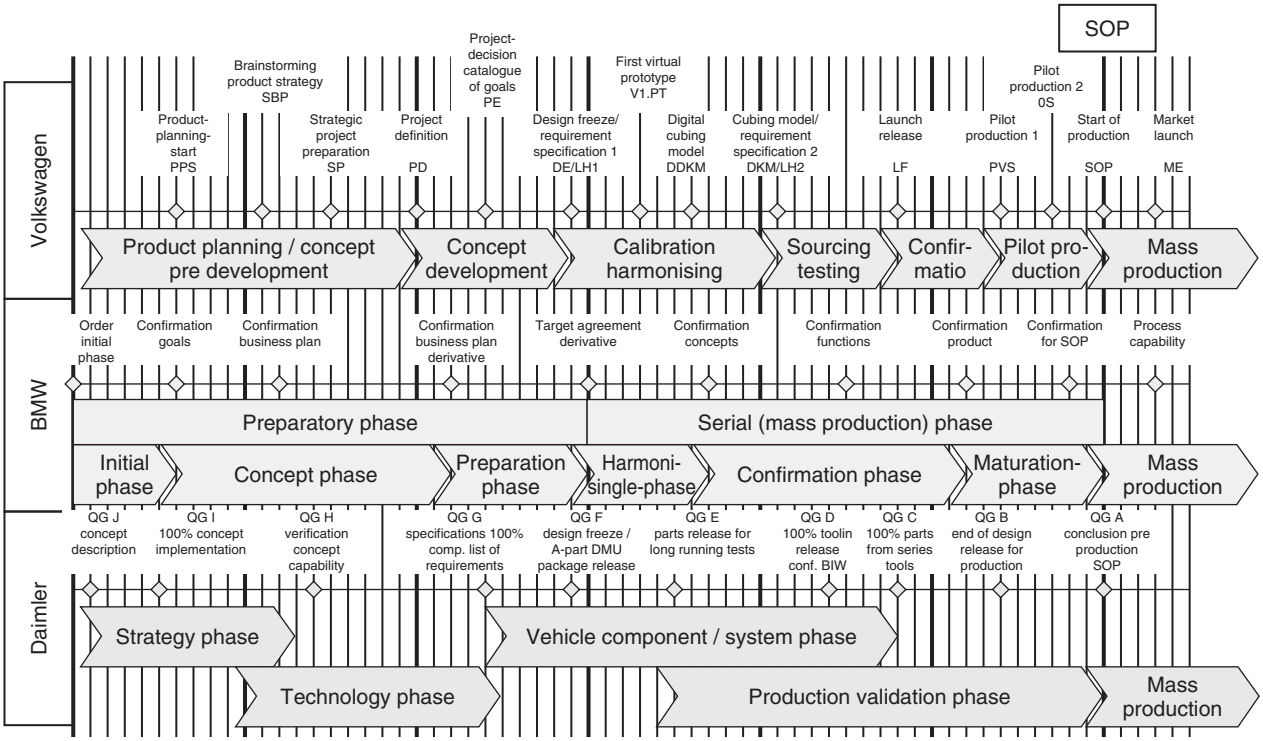


Figure 2. Comparison of master product evolution processes (PEPs).

phases are split up into three subordinate phases. The preparation phase is broken down into "initialization phase," "concept phase," and "design phase," and the detailing phase is broken down into "agreement phase," "confirmation phase," and "maturity phase" (Figure 2).

The objectives of the first major phase, the preparatory phase, are as follows (Figure 3):

- Investigate, define, agree, and verify design, technology, and innovations of the basic model of the new car family.
- Investigate, define, agree, and verify design, technology, and innovations of the cars to be derived from the basic model (this process begins after the basic model has been defined).
- Determine timing and responsibilities for the serial phase.
- Eliminate initial conflicts between design objectives as far as possible.
- Accomplish and freeze package and styling.
- Derive the engineering bill of materials (BOMs).
- Elect development partners.

This phase is responsible for defining about 75% of the product costs, and it has a major impact on the company's profit of the next decade.

In the process of converging packaging, styling, and engineering, which is a part of the preparatory phase, the approach is from interior to exterior (with the focus on technical constraints, i.e., "form following function") as well as from exterior to interior (with the focus on styling, i.e., "form following emotions") (Braess and Seiffert, 2007b). The design freedom of the car body is restricted by the styling-oriented exterior shape of the car and the function-driven package. This leads to numerous conflicting and contradictory objectives that must be resolved in multidisciplinary convergence processes.

The objectives of the second major phase, the serial (mass production, detailing) phase, are listed as follows:

- harmonization of the development output (i.e., the results of the preparatory phase) with current legislative requirements, standards, competitive situations and research conclusions;
- detailed design of individual parts and zones, and their optimization and verification in the context of the complete car design through calculation, simulation, testing, and prototype investigation;
- accomplishment of the homologation legislation and type approval; and
- planning and production of series of production tools and fixtures, and their maturation through sample inspections and pilot production to ensure the capability of all production.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

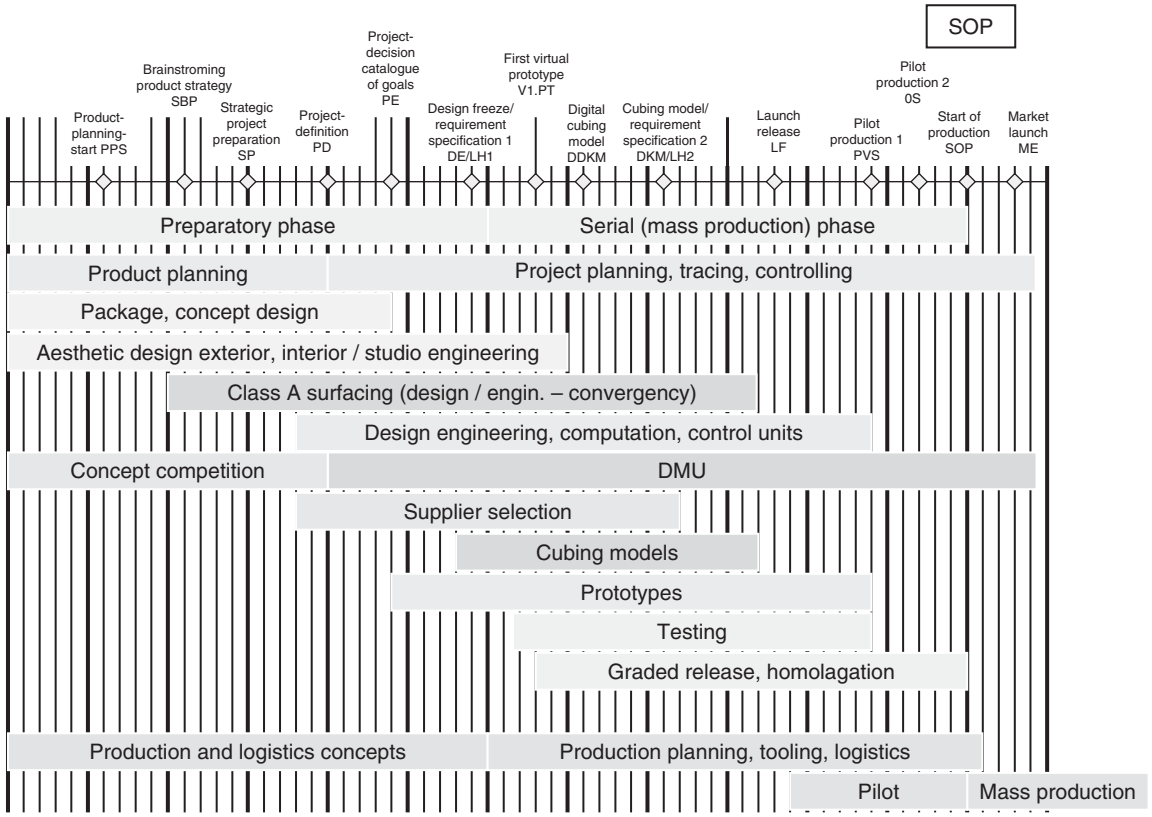


Figure 3. Distributed tasks in PEP.

In the following section, the most significant steps of this second major phase in the evolvement of a new passenger car are described and analyzed from the design engineer's point of view.

3 PACKAGING/ERGONOMICS PROCESS AND CONCEPT DEVELOPMENT

"The packaging process manages and harmonizes the requirements of component locations, ergonomics, and the overall characteristics of a car. It is a multidisciplinary process which accompanies the complete product evolvement cycle, from the first ideas to the end of production. The administration of all geometric data of the car and the control of all associated documents (to safeguard that each document is up to date) also belongs to the responsibilities of the packaging process". (Grabner and Nothhaft, 2006)

It is the objective of the packaging process to control and harmonize the interior of the passenger compartment (e.g., the arrangement of the seats) and luggage compartment, the location of powertrain components, and resulting characteristics such as wheel base, front and tail section, and ground clearance. The basic car concept defined at the beginning of the development phase is continuously substantiated as the development progresses; package plans, styling concepts, and subassembly designs are now getting more and more detailed and refined. In the past, this process of converging package, styling, and design was a two-dimensional (2D) process. Today, the availability of digital 3D tools (3D CAD, computer-aided styling system (CAS), large-scale projections, power wall, virtual reality (VR) caves, five-axis milling machines, rapid prototyping, etc.) has led to a dramatic acceleration and improvement of the development process.

In the first phase of the packaging process, according to 60 months duration, about 60 to 54 months before SOP, initial objectives are defined on the basis of strategic specifications derived from market analyses and predictions as well as project-independent results of technical investigations and research. The resulting concept package is the first 3D package, which initially comprises only about 15 components of powertrain, wheel and suspension system (including their interfaces with the car body), the locations of passenger seats and the driver's fields of vision, and the envelope of the luggage compartment. It is supplemented

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

with components from previous car projects, permitting rapid qualitative visualization of the new car concept. In addition, this concept package documents the resulting basic car dimensions such as overall length, height, width, and the location of the engine and the axles, as well as the anticipated positioning of the driver and the other passengers. On this basis, a dimensional design concept is derived that contains a preliminary course of hard points. In this phase, between five and six competing concepts are developed in parallel, for example, with varying wheel base. Additionally, predevelopment issues are defined, which must be tackled and solved to ensure the realization of the package concepts.

Once the initial design objectives have been determined, the second phase (about 54 to 40 months before SOP) is devoted to the preparation of a correct and marketable overall car concept. All major open issues contained in the initial concept are resolved to obtain the first draft concept already containing about 50 subassemblies of the new car. Conflicts between styling and package are resolved by compromises (convergence), leading to an agreed plan of hard points.

Furthermore, the engineering concept of the automobile project is determined in this second phase, to permit the verification of the package of wheel and suspension system and powertrain. With the aid of up-to-date manikins (virtual anatomical models of the human body), ergonomic investigations are made to validate seating positions, viewing angles, and access to the car (Figure 4). The virtual analyses are verified with the first physical seating mock-up.

On the basis of the initial design objective, the automotive body development process prepares a concept for the automotive body structure. The overall car concept, which is prepared in the concept phase, is aligned with these first automotive body analyses. Initial finite-element method (FEM) calculations are made to verify the overall car concept with respect to structural strength and stiffness. Additionally, concept analyses are performed in all domains of the automotive body development, to harmonize the overall car concept with the structural concept. If the initial design objectives are proved to be feasible, the second development phase leads to the confirmation of these objectives. The final plan of hard points and the first vehicle integration plan are the output of this phase.

The vehicle integration plan, which is documented in engineering drawings (scale 1 : 1), tables, and written descriptions, contains all major interior components such as seats, steering wheel and other cockpit elements, as well as the Society of Automotive and Aeronautical Engineers (SAEs) manikins and seating reference points (SgRPs). The space occupied by the powertrain and wheel and suspension system is visualized in the three main views. Furthermore, all major dimensions, driver viewing angles, nominal loads, slope angles, and so on, are documented in this plan. In the past, the resulting 2D geometry was the basis of the styling tape plans. These tape plans (scale 1 : 1) were used to optimize styling geometries and to initiate the 3D design process.

Today, packaging is a completely 3D process, permitting the ideal integration of the development of initial styling geometries (main curves and guide curves) and the generation of principal sections describing package and component envelopes through boundary surfaces into the overall process. Package functional surfaces are 3D surfaces, describing the space occupied by individual components (e.g., diesel engine with its envelope surface determined for idle vibrations or the front wheel with the envelope surface of permissible steering and suspension movements, arm reach envelopes of the passengers according to legal and ergonomic specifications or viewing pyramids).

50 months before SOP, the principal sections are created in 3D models using standard locations in the car coordinate system, with standard names for the geometric elements. To begin with, cross sections of previous and other cars

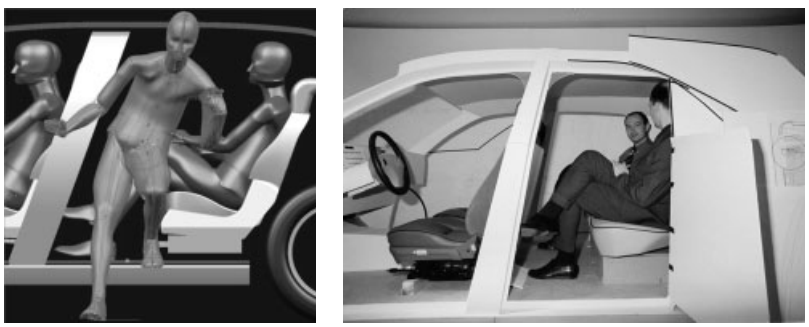


Figure 4. Validation of ergonomics with virtual (RAMSIS) and physical seating mock-ups on the Rolls-Royce Phantom (Lindermaier, 2006). (Reproduced by permission of Lindermaier/BMW AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

from the OEM product portfolio are used to obtain a coarse description of the topology of the zones and compartments to be developed. This coarse description is used to reach agreements and compromises between conflicting objectives with all departments involved in the car evolvement process. Subsequently, the principal sections positioned at standard locations in the car coordinate system are replaced by actual development sections derived from the functional surfaces of the package, styling data, and concept development results. These development sections are used for documenting relevant problems, reaching agreements between the design departments, and creating primary design surfaces.

In the third phase, approximately 40 to 30 months before SOP, the vehicle integration plan is detailed further. In the middle of this phase, the package of the basic car model of the desired product family is officially released (package freeze). In the meantime, about 150 subassemblies of the car would have been entered into the package plan. It is the main goal of the packing process to generate the geometric evidences that all legal and other requirements and specifications including OEM-specific regulations and standards are met and all conflicts resolved, and to keep the number of design variants to a minimum. Examples of such issues are the seating layout, luggage compartment layout, direct and indirect views of the driver, head clearance and head impact zones, accessibility of control elements, access to car, location and layout of restraint systems, pedestrian safety, bumper heights, ground clearance, and slope angles.

With the aid of CAD and digital mock-up (DMU), the packaging team continuously checks virtual design surfaces, module and part designs against hard points and envelope surfaces, and possible collisions of components. This way, the configuration of the car is verified more and more in the 3D models. In this phase, the first physical models of the future subassemblies are produced, for example, milled from hard foam. These physical mock-ups are used for experiments and tests, backing up the virtual validation processes, for example, with ergonomic investigations performed by test persons, and with manufacturability studies. After package freeze, the overall geometric data of the car is prepared according to internal standards and Global Car Manufacturers Information Exchange (GCIE) Group, and simplified overall car plans according to GCIE are prepared for exchanging data with other OEMs.

Even during the detailing phase, design modifications and adjustments accounting for current research results may occur. Consequently, the packaging team remains responsible for the validation and verification of all such changes with respect to legal and other requirements. In the mass production phase (detailing phase), the ergonomics team assumes the responsibility for the fine-tuning of control element surfaces and locations, operating forces, and paths.

At the close of the serial phase, production part homologation processes and type approval are carried out, supported by the packaging team.

4 TECHNICAL SUPPORT OF AESTHETIC DESIGN (STYLING) EXTERIOR AND INTERIOR

The most important function in aesthetic design is to give the product a soul of its own. This soul is not alone created by technical development neither package or with the decoration; the product soul can only be created by a stylist who is able to combine aesthetic, style and emotion.

In the competition for market dominance the manufacturers (OEM) have to satisfy customer's demands in ever increasing market segments. Parallel to the daily complex demands in customer lifestyle and the increasing number of products on the market, styling and "product soul" are growing in importance. You can't just glue on "product character", nor is it achieved by just using the logo symbols and colors of corporate identity. The combination of brand identity and "product soul" is achieved by the interplaying of aesthetic design, interior and exterior, of colors, shapes and materials. Thus aesthetic design defines the brand. (Ostle, 2003)

The aesthetic design process was the last manual layout process that has changed enormously through the developments in the virtual world. With the aid of computers it is possible to serve better process integrations and to have a greater variety of styling layouts in a shorter time span. Years ago, the pure development of design ideas with manual sketches, renderings, tape renderings, clay (Plastilin) modeling, synthetic wood (Epowood, Uriol), or hard foam were in the forefront. Today, the whole design process is developed in a new order with reduced numbers of format and media discontinuations aided by 3D CASs with a large variety of import and design functions supported by large-screen projections and 3D VR visualization (power wall, cave; Figures 12 and 13).

Sketchboards including software for sketching, rendering, and airbrush techniques are available for freehand sketching. It is thus possible to underlay a 3D package while using the 2D sketcher, which leads to the exact and correct proportions and perspectives. It is also possible to scan background photographs or handmade sketches to be used for a 2.5D styling process. Today, the shift from milled and hand-finished design models into the computer (redesign) is possible with the use of modern scan and import functions. High-resolution real-time

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

visualizations and animation of 3D design models enhance the efficiency of the styling process.

Aesthetic designers are still skeptical toward the new process: "Digital models tend to have a more optimistic appearance on the screen and on power walls. Only the physical and tangible model will give you the best impression. It is true that the CAD models save a lot of money. It is however fact that a car factory sells real cars. Because of this fact it is necessary in the phase of deployment to work on real cars." (Kraus, 2007)

4.1 Core areas of aesthetic design (styling)

"The exterior is love at the first sight. The interior is the marriage." (Sielaff, 2004)

The basic form of organization of aesthetic design is advanced design, exterior design, interior design, and color and trim. These areas are supported by modeling, studio engineering, and studio management. Some OEMs assign ergonomics and the Class-A surfacing also to aesthetic design.

Advanced design and concept design follow rules similar to those in early project concept developments of assembly groups within the technical design process. Both are often independent of the core project. The characteristic of concept design and advanced design is to collect ideas and visions without regard to whether they can be integrated into the technical design process. Advanced aesthetic designers track down new design trends based on the information of trend research. The results of their creativity are translated to design models or even show cars for car exhibitions.

Exterior aesthetic design creates the specific visual character of a vehicle by designing proportions of painted panels, glass, and light surfaces. In addition, character lines, highlights, and graphic elements determine the visual character. Thus brand image is created by the typical styling of the front and rear end or the shaping of the C-pillar. Aerodynamics plays an important driving force for composition work of the exterior over the last decades. The recent legislation for pedestrian's protection has influenced the arrangement of the front end considerably during the last years.

Interior aesthetic design must also include haptics, scent, and the auditory sense, besides the visual language of shapes. Not only seating package but also ergonomics and passive safety influence the interior process. Also issues such as seat design and arrangement, accessibility of operating units, and vision of instruments are of great importance. The aim is to offer the customer a comfortable, functional interior and a safe and secure feeling. The selection of materials for the interior components largely influences this point (Figure 5).

In close cooperation with paint, textile, and plastics industries, color and trim develops color and material combinations for the interior and exterior, texture of plastic parts, fabric and leather for seats and trim, and so on, all of which have to correspond to future fashion trends. Careful selection of materials must also meet all requirements for quality and durability.

The roles of designers and modelers often fuse together during the aesthetic design process. In the conventional process, modelers have largely been occupied mostly to convert sketches, renderings, and tape plans into clay models. They optimized the process in close cooperation with the aesthetic designers. Because of their artistic skills and excellent capacity to think in three dimensions, modelers are today occupied with aiding 3D styling models on CAS work stations. Following this, the virtual styling models are milled in clay and are completed and optimized by the modelers. Then the virtual model is updated with the latest design in a redesign process by scanning and surface modeling.

Studio engineers play an important role within the design process. They mediate between the aesthetical and technical design process. Similarly to the modelers, studio engineers must be able to understand and interpret the emotions and motivation of the stylists. Their task is to demonstrate the possibilities and technical limits to the stylist. They are also authorized to represent the stylist's claim to the engineers of the other development departments. Studio engineers continually compile the latest information about working



Figure 5. Rendering and clay models of the interior. (Reproduced by permission of Daimler AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

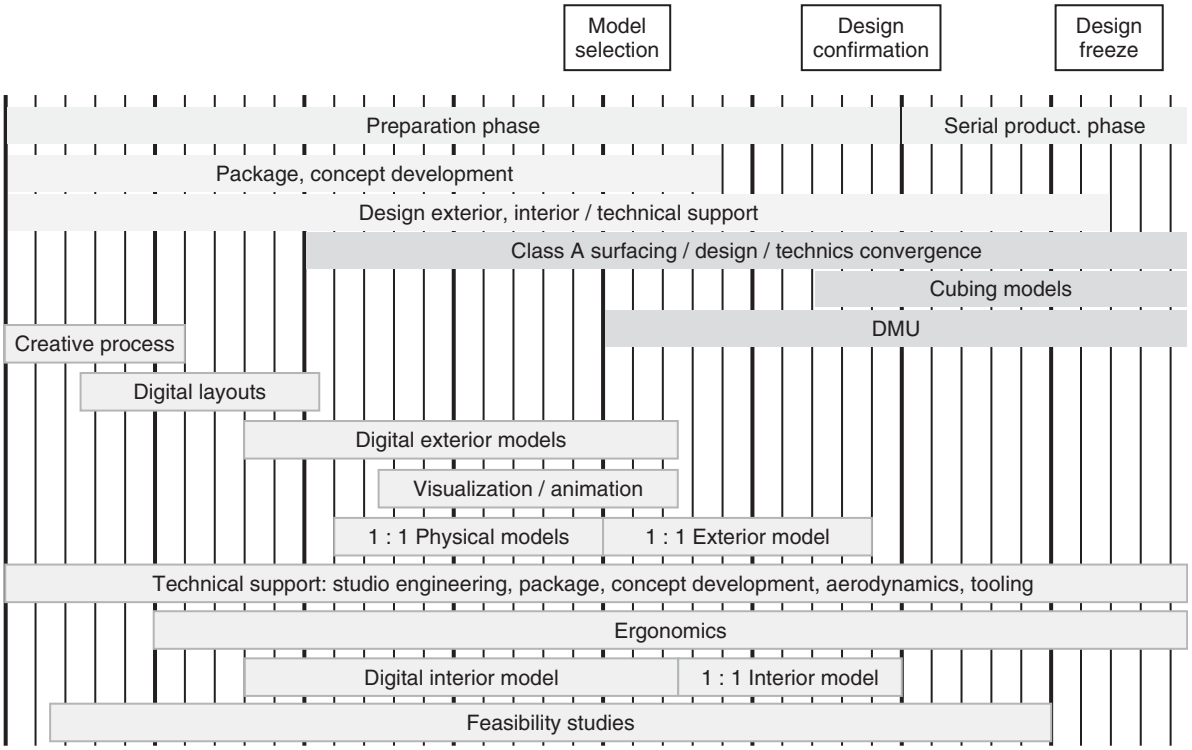


Figure 6. Digital aesthetic design (styling) process.

methods, production processes, materials, legal requirements, and so on, and they develop new technical ideas. The aesthetic designers are continuously informed by them about the possible innovations and consequences for the styling process. Studio engineers are responsible to guarantee the first functional feasibility of a styling by the use of principle sections, layouts, and model building. They conduct aerodynamic tests of design models. Together with method planners, studio engineers take production realization into consideration at this early stage. As "ambassadors" of aesthetic design, studio engineers present the new styling to the other departments of development. They conduct the design and the development from the beginning of the PEP, briefly until the SOP stage.

4.2 The new digital styling process

After a phase of discussion and interpretation of the list of goals of the new project, the styling process begins with a contest of ideas. Every aesthetic designer involved displays his individual interpretation of the project in layouts and sketches. Within the styling process, which consists of creative phases, presentations, discussions, and decisions, several competing aesthetic designs are studied at the same time. At the end of this process, there is one styling concept that combines the essential details of these competing styles (see Figure 6).

At the start of the aesthetic design process, the artistic and creative work comes to the fore. First of all, it is important to let the design ideas ripen. The package/aesthetic design convergence process starts afterward. Thus the artistic process of the tracked topic begins with freehand sketches of free chosen perspectives. Note that until the second step, these handmade sketches are digitally edited with the support of sketch boards and are converted to the proportions of the underlaid 3D package. For that purpose, freehand sketches can be scanned to reshape and to change those on the sketch boards.

In the second phase of the aesthetic design process, the conversion, the refinement of proportions and composition of details are carried out by 3D modeling in CAS. Orthogonal 2D outlines can be used as the basic framework for modeling and designing within the 3D model (sketch mapping). The stylists and modelers develop precisely tailored surfaces on the CAS system. The shapes and proportions of the model can be modified efficiently and the 3D data is available anytime so that a constant surface data flow to the development is possible. In the first instance,

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 7. Samples of painted exterior clay models and the chosen model. (Reproduced by permission of Daimler AG.)

five or more competing concepts are tracked. The concentration on two concepts is carried out shortly thereafter. Physical clay models of these two concepts are manufactured in a scale of 1 : 1. The precise refinement work often takes place at milled partial models, which can be added with assembly components from rapid prototyping or else from mass production.

Within the handcrafted process, renderings and tape plans have been the prestage for the physical model. Alongside the multifarious possibility of inspection and the possibility to animate styling models on a screen or a silver screen for visualization and discussion, renderings (Figure 5) and tape plans are nowadays derived afterward from 3D models of the CAS system.

The redesign process turns out to be demanding. The scanning data of physical models, which have been milled from CAS data and have been changed and optimized, are interpretable and sometimes redesigned surfaces do not reflect the styling idea of the optimized physical models. After closing off the second phase, the aesthetic design of the vehicle is largely detailed, and one final concept is chosen.

The finish of aesthetic design takes place within the third phase. The exterior and interior models merge into one model (Figure 7). Parallel to the second phase, the Class-A surfacing already begins to optimize the surfaces of the design models in CAD. The definition of the gaps in position, width, and producability as well as the quality of the surfaces has to be adjusted between aesthetic design and Class-A surfacing together with the demands of the technical conversion. The designed Class-A surfaces are tuned with the aesthetic design by using physical data control models (DCMs) of the interior, exterior, and the complete vehicle (cubing models) and are finally determined. Stylists accompany the technical design of the assembly components from the concept development until the detailing phase. They ensure the accurate conversion of their aesthetic concept, especially in cases of the often required compromises within the development and the production preparation.

5 CLASS-A SURFACING/STYLING/ TECHNIQUES CONVERGENCE

"Today the Class A surfaces are the geometrical representation of all surfaces that are visible for the customer both interior and exterior under consideration of all technological and shape aesthetical demands." (Lender, 2001)

The former manual Class-A process originally derived from shipbuilding. In the search for flow-enhancing designs of hull shapes, it was possible to design grid sections (length sections = vertical frames, breadth, and height sections) through the hull under support of long, flexible splines made of homogeneous material. These splines are kept in desired form with the use of spline weights (Figure 8). By alternately raising the weights, curves with smooth curvatures developed, which were free of dents, waves, or any kind of inconsistencies. Thus surfaces were clearly defined along the sections; however, the areas between the sections were not defined. Up into the 1990s, automotive Class-A surfaces were designed by this wire frame method of grid sections and 3D curves. Today, however, without exception, the Class-A surfacing is completed with the

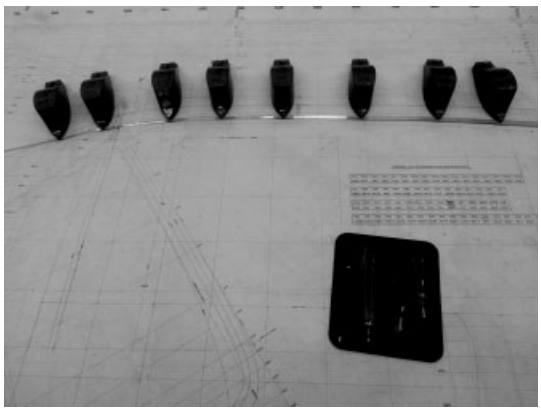
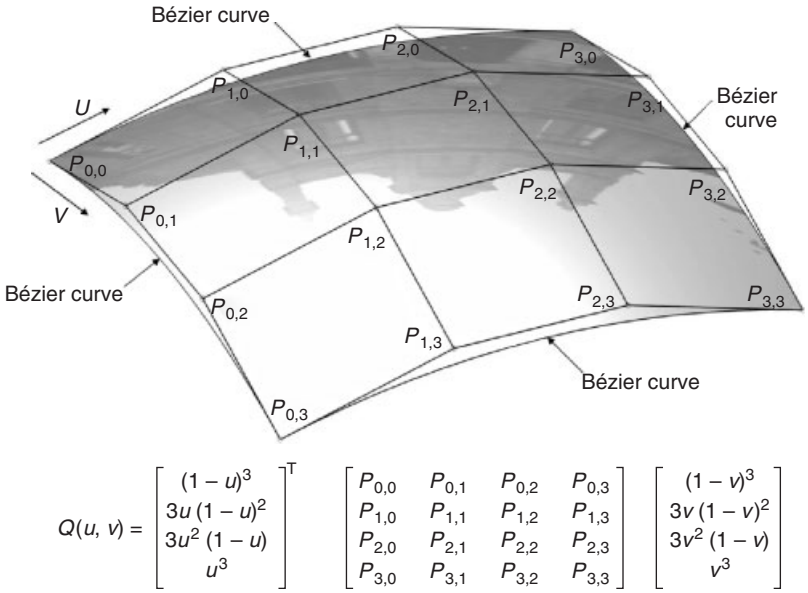


Figure 8. Manual Class-A surfacing under support of spline and spline weights.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



$$Q(u, v) = \begin{bmatrix} (1-u)^3 \\ 3u(1-u)^2 \\ 3u^2(1-u) \\ u^3 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} P_{0,0} & P_{0,1} & P_{0,2} & P_{0,3} \\ P_{1,0} & P_{1,1} & P_{1,2} & P_{1,3} \\ P_{2,0} & P_{2,1} & P_{2,2} & P_{2,3} \\ P_{3,0} & P_{3,1} & P_{3,2} & P_{3,3} \end{bmatrix} \begin{bmatrix} (1-v)^3 \\ 3v(1-v)^2 \\ 3v^2(1-v) \\ v^3 \end{bmatrix}$$

Figure 9. Bézier surface patch of low order.

aid of special CAD programs, for example, ICEMSurf (Dassault) or ALIAS (Autodesk). All these programs are based on the principles of curve and surface algorithms of de Casteljau and mathematics of Bèzier (Figure 9).

The Class-A process is a major component of the PEP. Its primarily importance lies in its early position in the development process and thus the big influence of all the following processes that are continuously dependent upon up-to-date surface data. With the introduction of Class-A surfacing into the early stage of PEP, a dynamic loop process of surface and gap modeling and adjustment under support of all relevant specialists begins to harmonize the demands of styling and technical design. Class-A development is completed with the release of the DCMs, which is followed by the release of all customer-relevant surface data of interior and exterior vehicle surfaces.

The process of surface development is based on scanning data that are generally available as a mesh of polygons originating from a cloud of scanned points. For better orientation, the polygon mesh is added by relevant grid sections. The area that has to be developed is "divided" in the mind into primary surfaces and secondary surfaces. Primary surfaces are those surfaces that are used as the principal or basic surfaces for the adaptation of further geometry, for example, offset surfaces or character curves. Secondary surfaces are those surfaces that are based on primary surfaces or add further details to the primary surfaces, such as bent portions, blend surfaces, or fillets.

As an example, contact points on each of the primary surfaces are used to hustle surface patches of low order (Figure 9) onto the primary surfaces. By careful and equal manipulation of the polygon points, the surface patch will closely resemble the scanning data. The primary surfaces (patches = unbounded surfaces) are larger than the styling surfaces. Character curves can be achieved and smoothened from the scanning data or intersection curves can be gained in combination with other primary surface patches. The patches (unbounded surfaces) are bounded by the curves and become faces (bounded surfaces). Any discrepancies in the scanning data or in the styling model at hand will be interpreted by the Class-A surfacing engineer to harmonize them with the aesthetic design. Surface data that the Class-A surfacing department provides includes all visible surfaces including first bent portions and first radii.

The visible surfaces of the interior and the exterior are interrupted by gaps whereever component parts meet. Gaps are an important stylistic element for an aesthetic designer. While designing the gaps, functionality of movable components and highlights on surfaces have to be taken into account. The unbounded visible surfaces and the curves defined by the aesthetic design and safeguarded by technical design, are the basis for the Class-A design of the gaps. For the development of a gap (Figure 10) the visual constant appearance of the gap, seen from the customer's typical point of view, is to be defined instead of defining a theoretical constant gap. In critical areas, the surface position in the area of the gap is moved within height (displaced) and/or the theoretically arranged gap varies in breadth.

The virtual world offers a multitude of possibilities to evaluate the quality of surfaces. Highlights are a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

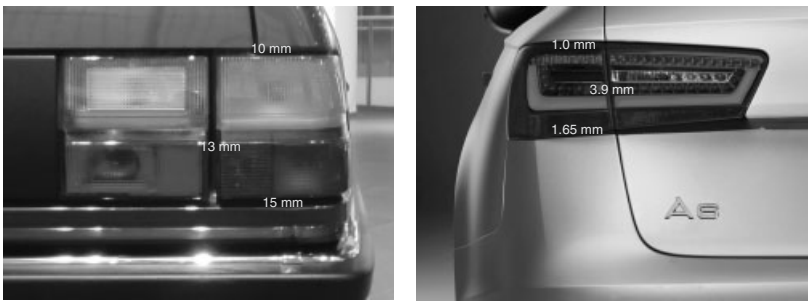


Figure 10. Comparison of surface and gap definition AUDI 100 (1986) and A6 (2012) (Großjohann, 2007). (Reproduced by permission of Daniel Großjohann/AUDI AG.)

common method used. In theory, light sources are reflected under various angles to the free form surface. Lines of constant brightness (highlights, isopodan) are the geometric positions where the angles between light beams and surface normals are identical. A further virtual and even practical method used is the projection of a panel with parallel light sources onto the free form surfaces (Figure 11).

The Class-A department starts with its work as soon as the aesthetic design reduces the number of possible styling concepts to two. The styling models are milled and modeled in precise detail in Plastilin in the scale of 1 : 1 for the first time. To be able to evaluate the models virtually and to make the surface data available to the following departments as quickly as possible, the two models are scanned. Within the Class-A surfacing process, concept surfaces are created out of the 3D cloud of points from the scanning. The concept surfaces do not yet comply with the final requirements because of their quality, curve progression, and continuity. These first CAD models primarily act for judgment and confirmation by the department of aesthetic design, although the departments of technical design already use these concept surfaces as the foundation for the assembly group design.

The criteria for evaluating the Class-A surfaces are

- producibility,
- optical quality,
- compliance with package dimensions, concept coherence, COP concepts, and design variants.

With the choice of the styling model (model selection; Figure 6) by the management, the number of exterior and interior models decreases to one. Shortly thereafter, the design confirmation, including the incorporation of all open points and leadership issues into the model, takes place. The current development status is now harmonized with regard to package, concepts of technical design, aerodynamics, ergonomics, and occupant safety. Hence the technical stage of maturation has proceeded to the point where the development of the final Class-A surfaces begins. With that, a process of virtual and physical DCMs and master gauges begins, which conducts the PEP till the SOP and which safeguards Class-A surfacing and the technical design of components that follows (prototype parts as well as mass production parts). In an automotive project between 600 and 800 interior and exterior components relevant for visible surfaces get processed in Class-A surfacing. Gray zone surfaces, which cannot be seen until the customer

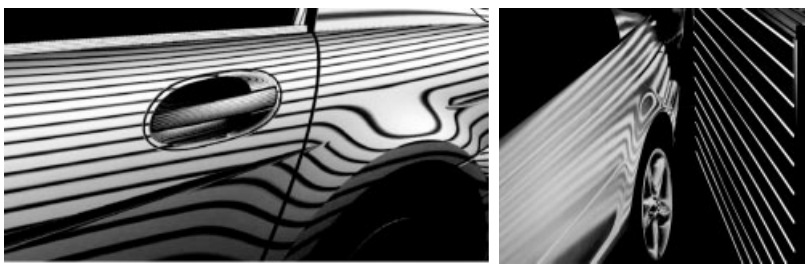


Figure 11. Quality control of surface and gap definition with a panel of parallel light sources (a) in VR (Gegalski, 2003) and (b) in the laboratory (Dehn, 2001). (Reproduced by permission of Günther Weigl, BMW AG.)

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 12. Virtual validation of the interior at the power wall. (Reproduced by permission of Daimler AG.)

opens a door or a lid, and a lot of surfaces of the interior are first designed as functional surfaces within the concept development phase and are at last finished in Class-A surfacing.

The design confirmation (Figure 6) means to "freeze" the current aesthetic design model. From now on, it is no longer possible for the styling department to shape the physical model and the design subject is completely assigned to the follow-up CA processes. Class-A surfacing is the leading department to make modifications and to visualize them. At this stage, all styling subjects are completely displayed. This includes, for instance, all mounting parts of the exterior, for example, handles or trim strips as well as panels or electric components of the interior. Furthermore, a detailed gap plan is defined, which can be adjusted to a minor degree. Even so, the stylists still have the possibility to place change requests within Class-A surfacing. Henceforth, the areas of aesthetic design and Class-A surfacing closely work together with the aid of visualization (Figures 12, 15) and CA tools.

The design freeze (Figure 6) is the next milestone within the Class-A process. At that time, the aesthetic design area finishes a styling adventure model (Figure 7) that contains interior and glass surfaces. All open issues, which were detected at the confirmation of aesthetic design, are incorporated in the Class-A surfaces. This is the base for the digital DCM that assures the design and the visible surfaces within the virtual process. The digital DCM describes the final condition of the visible surfaces in every detail, for example, varnished metal sheet areas, colored and textured interior components, or configuration variants. Thus it can also be used to safeguard the match-up of color and material combinations (Figure 12). Weeks before the completion of the first physical DCM, the virtual evaluation and the certification of all visible components takes place. Parallel to the Class-A surfacing process, a digital controlling of functions continuously takes place in all development areas. Functions, such as kinematics or overpush effects of lids, the air distribution of the air vents, or the head impact zones of the instrument panel are checked in detail (see Digital Prototyping and Digital Mock-Up (DMU)).

The result of the first physical data control model (DCM = Feasibility Cubing) is a complete and full-scale milled model on the basis of CAD surface data. The DCM or else the master pattern controls the Class-A surfaces; it resembles the design model and represents the decision made by the executive committee. Today, the DCM is created by high-performance milling machines that allow varnish-capable surfaces. The assembling takes place on the basis of synthetic woodblock material (Uriol) supported by aluminum beams. Final surface refinements demand accompanying Class-A surface design operations and lead to the point of release for the last milling loop of the final DCM.

The functional cubing models of the exterior and interior complete the safeguarding of all Class-A surface data. The BIW and all mounting and trim components of interior and exterior are displayed as a buildup that is geometrically equivalent to the final product. All exterior and BIW components are made from aluminum. The interior is displayed by Uriol or laminated components. Similarly to a real car, doors and lids can be opened by using realistic kinematics. In addition, the original boot is displayed. Cubing models of mounting parts of the exterior and interior are fixed the same way as in mass production. For example, the functional cubing interior is verified to check if the mounting and assembly components are able to be installed at their fixing points, at the clips holes, and if these fit together. The executive committee affirms the complete field of visible surfaces of the automobile, including secondary surfaces as well as visible gray zone surfaces between interior and exterior.

6 DIGITAL PROTOTYPING AND DIGITAL MOCK-UP (DMU)

The development of a product roughly takes place within the phases of design and simulation, laboratory work on physical models and road testing (Figure 13). Calculation and simulation leads to an early safeguarding of design data. It also allows physical prototypes with a high maturity and an optimized testing. Alongside the early safeguarding by simulation, the testing of physical prototypes within the development process is indispensable (Breitling, 2007). DMU is the virtual replacement of physical mock-ups using 3D computer graphic techniques to support the product engineers validating the designs of the complex assemblies.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

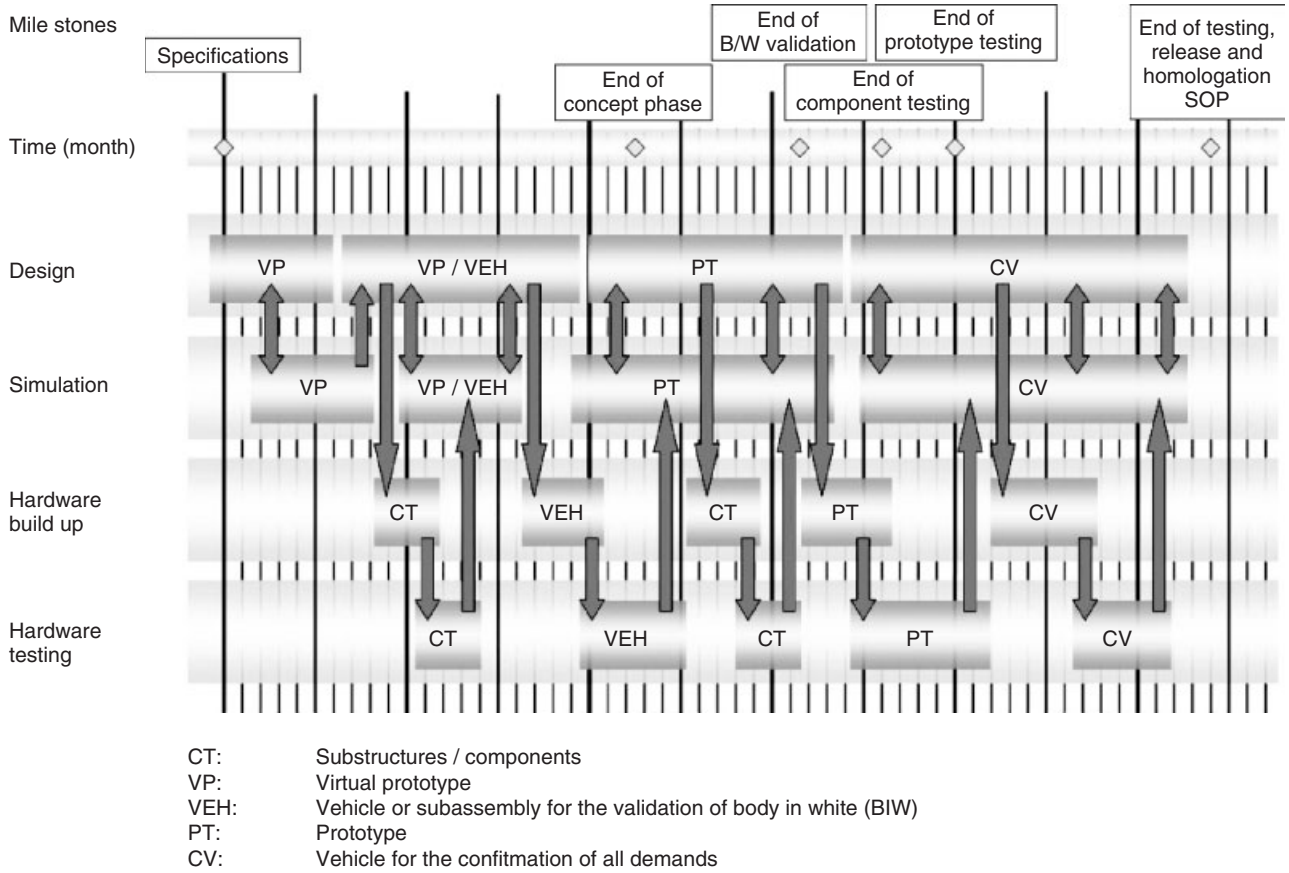


Figure 13. Virtual and physical arrangements for the validation of passive safety. (Reproduced by permission of Rodolfo Schöneburg, Daimler AG.)

The term *DMU*, the digital build of prototypes, no longer describes the full breadth of tasks for the support of visualization and simulation within the PEP. Within the automotive industry, terms such as digital prototyping (DPT) or digital engineering visualization (DEV) are in use.

In automotive body design, several verifications of the 3D geometry take place. In former times, verifications were performed under support of extensive 1 : 1 mock-ups (e.g., seating mock-up for ergonomic examinations or mock-ups built from deep-drawn plastic sheets to verify the assembly of BIW parts). Today we find a mixture of CAD-driven, product data management (PDM)-driven, and special isolated applications for DMU. These DMU-supported technologies evaluate diverse development concepts of the early phase and often can validate different design variants. Today the following fields among the application areas of visualization and simulation are important for the safeguarding of body design:

package layout and homologation (Figure 14);
visualization of design data and Class-A surfacing (Figure 15);

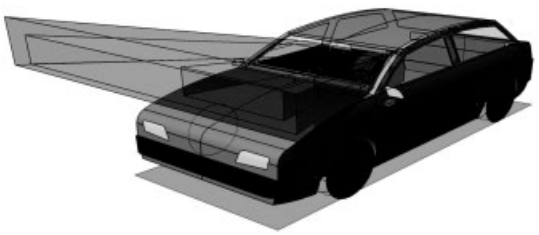


Figure 14. CAVA-software for packaging and homologation.)

aerodynamics;
structural load cases;
kinematic functions of mounting parts;
deformation of parts within production or in common handling;
tolerance management with its effects on surface and gap design as well as within production (Figure 16);
assembly, disassembly, and handling/filling of assemblies;
all crash load cases;
durability of load-bearing components;
noises, vibrations, and harshness (NVH) a passenger feels;

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 15. VR validation of Class-A surfaces, colors, and equipment versions. (Reproduced by permission of Daimler AG.)

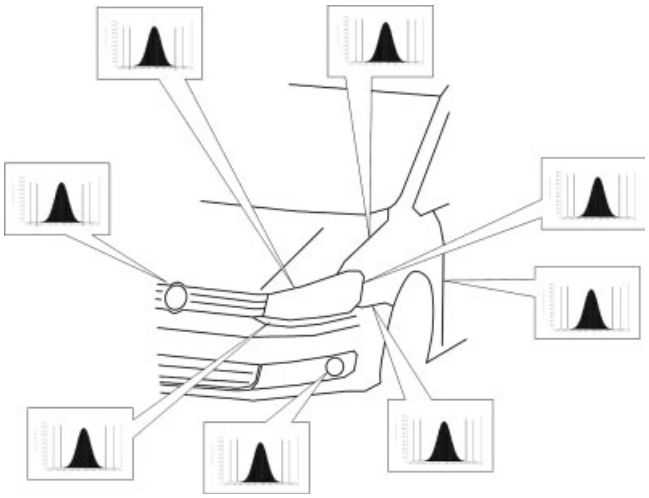


Figure 16. Dimensional variation studies as a confirmation of product dimensional specifications in a car development (Piskun, 2012). (Reproduced by permission of Alexander Piskun, HAW Hamburg.)

thermal functions of the heating/ventilation and air conditioning unit (HVAC);
thermal comfort inside the car;
loads caused by cataphoretic painting process.

Forsen (2003) classifies the activities of DMU into four main areas:

"GI: Geometrical Integration: All CAD-driven and powerless verifications belong to the area of steric geometric applications, such as: Interference Checking, Available Space Analysis, Cinematic Clearance Checks, Tolerance Simulations, and Ergonomic Examinations etc." (Forsen, 2003). Geometrical integration already includes several functional integrations (FIs) such as ergonomics examinations, kinematics of closures and other adjustable body assemblies, lighting, and so on.

FI (Functional Integration): The area of FIs includes all power integrated calculations in the field of DMU, such as: Crash Simulations, Stiffness Calculations, Aerodynamic Simulations, and Driving Dynamics Calculations etc.

PTI (Production Technology Integration): The term of PTI includes all verifications to safeguard manufacturing processes, such as Installation and Dismantling Studies, Deep Draw Simulations, Simulations of Jigs and Fixtures or Spot Welding etc.

EI (Electronic Integration): A strongly growing rate of electronic components in vehicles intensifies the poor design of mechanical components up to the handling of complex functional contexts, controlled by electronic devices. DMU therefore must be enhanced by simulation methods for EIs. (Forsen, 2003)

The design stages (status management) and design variants (variants management), which can only be organized for digital calculation and simulation process under support of intelligent CAD software (PAD) and/or PDM systems, are very important subjects within development and analysis. It is typical to analyze and rate a large number of alternatives within the early project phase. The main part of DMU data is taken over from 3D CAD data. The DMU simulations listed under GI can be verified in the CAD programs themselves. Visualization and rendering of large assemblies or dynamic cross sectioning often is handled within PDM systems. PDM-integrated applications also support verifications in a development process with multi-CAD use. Special solutions of FI often have special data formats that may course problems with feedback of data into the design processes.

At all times, the amount of data confines the actionability of design and simulation. The DMU data format JT (Siemens PLM Solutions) has becoming the new interface on behalf of the digital development process. Besides tessellated data with several levels of detailing, for instance, to visualize design data effectively, the JT format delivers exact geometric data (NURBS), for example, for measurement aspects, after the conversion. It also delivers metadata such as the product structure or attributes such as dimensions, tolerances, reference point system (RPS) points, or welding spots. Thus it is possible, for instance, to represent the complete environmental geometry within JT format while designing an assembly unit in a complex environment. The transmission of design history between OEM and system suppliers or different design departments is not needed.

Analysis of the complete vehicle with configuration and creation of DMU data in real time is not possible nowadays because of the high volume of data and different

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

data formats. For these verifications, DMU data are collected externally and converted overnight into JT or other exchange formats. The preparation of FEM meshes based on this exchange format cannot be done completely automatically, and takes another one to 3 days for an assembly and up to several weeks for a complete vehicle. The results of these verifications are not therefore up to date and often do not deliver all versions (Hagenah and Klar, 2006). The evaluation of constructed space in "Design in Context" applications requires a connection between the DMU and the BOM systems (PDM). Modern, parametric, associative CAD systems with connection to PDM system have taken over the main duties from DMU: to safeguard the continuous consistent management of versions and primarily interference-free development of assemblies.

7 CONCEPT DEVELOPMENTS ON THE EXAMPLE OF CLOSURES

The design of geometry and safeguarding of functions in concept and detail development is performed in several phases and is interrelated in diverse ways. According to the product vision (catalog of goals) declared, design and simulation work is divided from the whole vehicle into the single part of the vehicle body and reassembled several times.

For the concept and package phase, as a general rule, specialists from different areas of the vehicle body development, from package, styling, design, and simulation are concentrated in a project team to develop, organize, and decide about development work for the basic approach within a short period of time and to declare their centers of competence directly. Innovative concepts and new technologies developed independent from a special project are adapted to the new project in this phase. For every goal, several alternative concepts are explored and developed to one optimum concept. While the suppliers and engineering suppliers were involved in the detail development in early years only, they are often integrated from the early phase of concept development now. Project management of design and simulation work is often performed by experienced body designers of the OEM.

Detail development takes place under supervision of the centers of competence responsible for the special areas of the vehicle body. Most part of the engineering work is done by engineering suppliers. During this phase, parts and assemblies are developed in detail and their detail functions (e.g., weight, producibility, comfort, crash) are further optimized. While in the concept phase, simulation for the safeguarding of functions is executed on the basis of unripe CAD models, analog models, or computer-aided engineering (CAE)-modeled geometry in the detail phase simulation safeguards the development on the basis of ripe CAD data. But even in this phase, new cognitions (laws, competition situations, mistakes in early phase, etc.) can lead to a switchback into concept development. As definite project steps are appointed until SOP, expenditures are multiplicated in such situations.

The activities described in the previous chapters will be interpreted in this chapter on the example of side doors. Even here, the explanations can deliver only a small view into the multifarious field of closure development.

Side doors as well as front and rear lids belong within the system "vehicle body" to the subsystem "mounting components exterior" while the structural parts of the vehicle body define the subsystem "body structure." In the organizational structures of some OEMs, the interior trim of closures is also part of closure development.

A closure is a device that allows access to a compartment. Examples are the access to the passenger compartment, the boot, and engine compartment.
A door protects the passenger compartment against climatic influences, noise, unauthorized access, and so on.
Important elements of a swinging door, for instance, are, besides the door structure, the hinges, the latch, and the sealing gaskets.

The side door is one of the most complex assemblies of the vehicle body and must fulfill lots of requirements. In this chapter, requirements are described that can be categorized as follows:

Handling functions
Passive Safety
Quality impressions
Misuse
Durability.

Side doors are designed according to different concepts. During the last years, the frame door concept was established in mass production (Figure 17).

Handling functions are functions that have to be guaranteed for the daily use of the door. Nowadays, handling comfort and safety play an increasing role.

Legal requirements, directives, and specifications (e.g., ECE, FMVSS), the crash test procedures defined by consumer organizations (e.g., Euro NCAP, IIHS), as well as in-house rules and regulations for passive safety are multifaceted and influence the design of side doors and side-wall structures extensively. Side doors must support the safety of passengers at both side and front impacts. In all cases, side doors must guarantee the rescue of the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

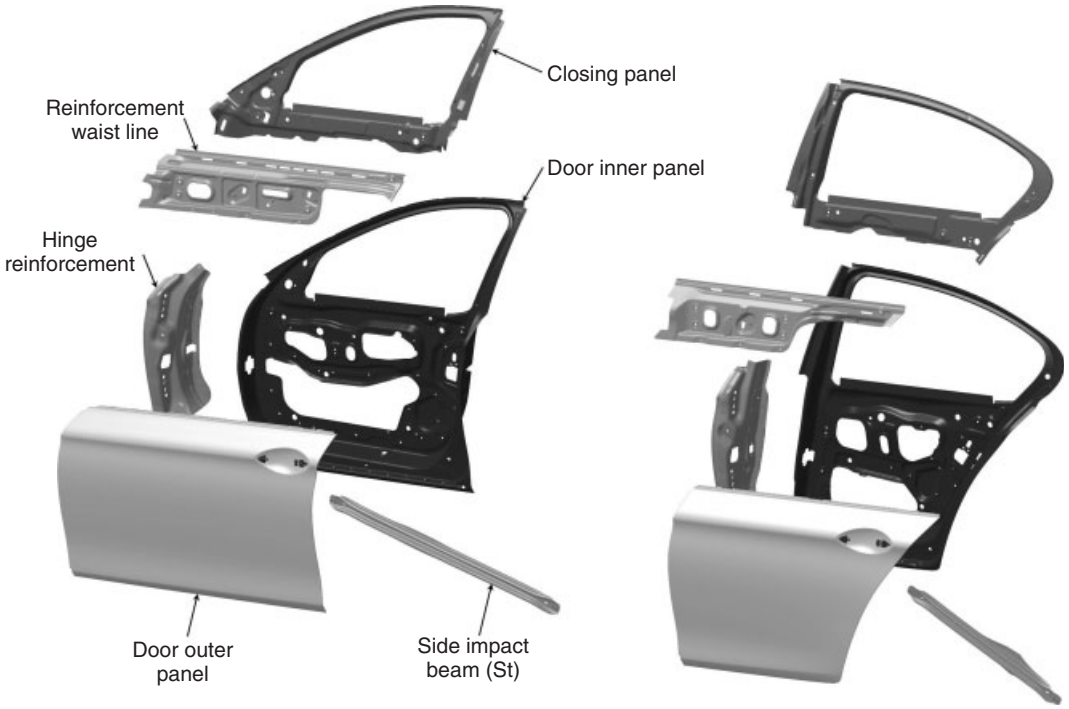


Figure 17. Frame door concept in an aluminum layer-built design. (Reproduced by permission of BMW AG.)

passengers after an accident. As examples from a multitude of designs, the waist rail of the door, which defines a load path for frontal impact, the side impact beam, and the side airbags in the doors, or padding arrangements in the door trim can be appointed. A lack of quality comes up when during polishing, for instance, the elastic deformations of the outer door panel lead to a noisy oil-canning tendency. This effect and plastic deformations are not allowed. Other examples for quality issues are the stiffness of the arm rest and the map case or door-closing noises.

Misuse loads are defined as soon as normal handling loads exceed the expectations. Doors must bear misuse loads up to certain limits. Examples of such conditions are as follows:

When the load case door sags; here, a load is defined that represents a heavy person leaning on the door body.
When a third person tries to get into the car by forced pull on the partially opened door window;
When a door is slammed with partially opened door glass with up to 30*g*.
When the open door is overpulled, this extremely stresses hinges, door stop, door panels, and pillar.

If the occurring loads do not exceed regular working loads, but occur in a great amount of cycles during the vehicle life span durability, load investigations are necessary. Typical examples are the endurance tests "door opening and closing" or "window opening and closing" where wearing parts like sealing gaskets, latch, hinges, door stop etc. are examined.

The following examples show some design requirements in more depth.

7.1 Hinge axis positioning

The hinge axes of passenger cars are positioned within the side wall. This leads to the situation that a part of the door turns into the side wall when the door is opened. The depths of turn in amount depend on the distances between the outer surface and hinge axis as well as front corner of the door and hinge axis (Figure 18).

The hinge axis is inclined marginally to lift the lower rear corner during opening and to allow comfortable opening and closing moments. The position of the hinge axis is defined by the curvature and inclination of the outer surface, the size and distance of the hinges, the general design of the door panels, and connections and the adjusting ranges for tolerance compensation. The hinges used are often COPs. The mounting of the hinges parallel to the planes of the vehicle axis system allows the adjustment within all three directions of the coordinate system. This is necessary to adjust flush outer panels and parallel gaps. Doors are nowadays often reassembled after painting, to enabling separate final assembly. Hence, in most cases, the

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

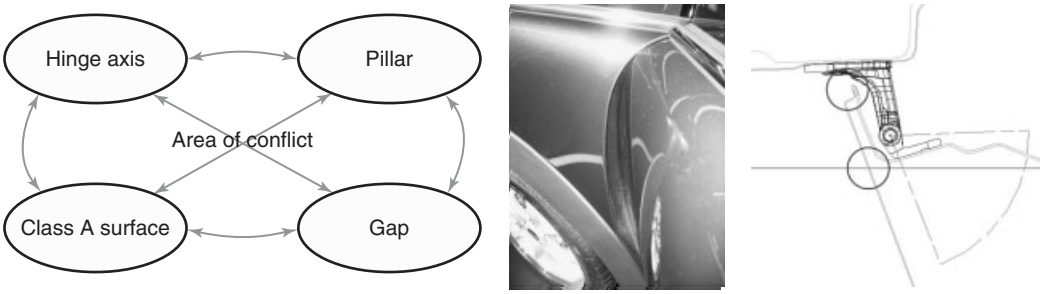


Figure 18. Conflicts during gap and hinge axis definition.

hinges are designed according to a two shear connections to unhinge and hinge the door. The travel needed for the unhinge operation has to be taken into account when the door gap is defined.

The door gap in the hinge axis area has to be safeguarded several times during the development process. As soon as the first 3D styling surfaces have been defined, the studio engineers can check in critical positions on the basis of principle sections, whether the gap defined by the stylist allows opening the door. Or the studio engineer defines limiting points and curves for the stylist to show a bandwidth for the definition of the gap. The first limiting curve and possible front corner of the door is defined when the outer surface of the opened door intersects the outer surface of the closed door at a defined opening angle (Figure 18). The second limiting curve for shaping and positioning of the outer pillar surface is defined by the front corner of the opened door. The more the stylist stays away from the first limiting curve, the more the front corner of the door turns into the side wall and prevents a suitable side-wall structure.

In the concept development, the hinge axis is safeguarded again on the basis of Class-A surfaces and gaps. Under consideration of the parts mentioned above and critical tolerance positions, a 3D corridor for the final position of the hinge axis is defined by three limiting surfaces on basis of the styling gap. From the first concept development stage of the door to the completely detailed door, the designs of the parts will be optimized in several loops. This will lead to several adjustments for the safeguarding of the hinge axis and gap.

7.2 Latch positioning

For low noise emission and wear out of the latch positioning, the latch plane has to be defined normal to the turning circle of the door. The second angle of inclination of the latch plane is defined parallel to the breadth (XZ) sections of the basic slanted surface connecting the inner and outer side walls in the latch area. This is the best way to allow the smallest possible constriction of the pillar, whereas in former designs, the latch often was positioned perpendicular to the plane view.

The basic slanted surface of the door connecting the inner and outer surfaces of the side wall is defined by the travel line of the door glass and the slanting angle necessary to open and close the door. While this functional surface is defined planar at the beginning of concept design, all surfaces of the gray zones except mounting surfaces have to be designed slightly convex at a later stage for manufacturing and aesthetical reasons.

The height of the latch depends on whether the door outer handle fits the latch directly or indirectly. The latch and all its components are positioned within the lower door case. For operation between latch and striker, a slot is necessary in the corner between the inner door surface and basic slanted surface of the inner panel of the door, and in some designs, in the inner door trim as well.

To allow the movement of the door glass, there must be a recess in the basic slanted surface. The size of modern latches with all their comfort functions need large recesses on the doors and corresponding pillar surfaces constricting the pillar structure. Some OEMs define the recess and its slope surfaces locally, only from the waistline down to the sill.

Optimization of the side-wall structure according to passive safety requirements often leads to changes and redesigns in the latch area.

To guarantee the hinge and latch functions, depending on the locking position of the latch, legal requirements define longitudinal and transverse forces. These forces are simulated in designed position in a ridged portal and door environment as soon as the first concept geometry for hinge and latch position is defined.

7.3 Drop glass

When the side door area is defined in styling, it must be ensured that the curvature and inclination of the greenhouse

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

area correspond with the dimensions of the lower door case. On the basis of longitudinal sections (cross sections, ZY) and principal sections of the package, the size, shape, and positioning of the door glass is safeguarded.

As the door glass of former cars were plane or curved cylindrical, nowadays most door glasses are part of a large rotational surface. Measured data of two or more longitudinal sections through the styling model are the basis to define the axis of rotation. As the rotational surface is made more or less elastic, the glass must pass the slot between the waist rails, and limit ranges; and combinations for the curvatures of the glass normal (R 1000 to R 2500) and radial (R 20,000 to R 150,000) to the axis of rotation are defined from experience. Modern CAD programs allow the mathematical optimization of the door glass according to points of interest on the styling model.

The door glasses of the front and rear doors are guided along the B-pillar. The inclination of the front and rear corners of the B-pillar defines a movement of the door glass as a combination of rotation and translation. The daylight area of the door is enlarged in the waist area by supporting and adjustment surfaces. Form and positions of the door glass must match up with the panels of the door, for example, in the sill area and the mounting parts such as door stop movement, side impact beam, and latch. Minimum distances between rigid, flexible, or moving parts have to be considered.

When the first concept design of the door panels is finished, the door can be simulated and optimized with the following load cases:

door glass opening and closing with and without stopper in opened position;
door slam with door glass partially opened with an acceleration of more than 30*g*;
forced entry through the window with a pulling force of 1000 N on the slightly opened door glass.

7.4 Door frame

The door frame defines the upper corner of the door. The nearly flush door glass is guided and sealed on the door frame. The sealing gasket situated on the welding flange of the side-wall structure seals against the seal surface of the door frame. The sealing gasket situated on the door frame seals against the seal surface on the side-wall structure.

The door frame is built by a deep-drawn door inner panel designed from sill to roof rail (Figure 17) and includes the cutout for the door glass. A stiff box section for the door frame combined with a deep-drawn closing panel reaching from roof rail to the upper hinge/latch area is defined..

Contrary to the design concept with, for instance, a constant extruded or rolled frame profile, the deep-drawn inner panel can be designed with variable profile depths. Even here, most OEMs keep the frame constant in the directly visible area of A-pillar and roof rail, and define the variable profile depth in the area of the B-pillar only.

In the front, the door frame ends in the mirror triangle; and in the rear, the door frame ends in a blunt corner with the waist rail. Additional panels may reinforce the changeover from frame and lower door case in the area of the waistline.

The aerodynamic pressure distribution under driving conditions leads to a high vacuum in the joint area of A-pillar and roof rails. The load level can be increased by separate excitations of road and engine. The sealing gaskets and the corresponding surfaces on body or door frame must always overlap each other and only reduced relative deformation of the structures are expected. Otherwise, loud wind noises in the area of the passengers head would occur when the air leaves the car through short-term gaps.

The stiffness of the door frame and its connection to the door case must be defined in a way that preloads of the sealing gaskets, wind loads, and other excitations are over compensated. OEMs therefore overbend the door frame in the joint area of the A-pillar and roof rails. When the door glass lies on the door frame, the door glass is overbent too. The disadvantage of this design is that the closing loads of the door and door glass is increased for the customer as the overbent door frame must be pressed back into its styling position when the door is closed.

"Usually, complex dynamic loads affect the closures and therefore the doors of a vehicle body. Reducing this complex real time behavior to the most important factors and defining appropriate load cases is usually a great challenge for any simulation engineer. The static linear simulation of the window frame stiffness is an example of such degradation. Here, the aerodynamic wind loads, as well as the road and engine excitations, lead to complex vibration during the vehicle operation. These dynamic events can be reduced to a static load case, in which forces are applied to the front and rear window frames. From experience, the resulting stiffness required for satisfactory behavior is known. The method allows fast and reliable identification of information needed for the designer to dimension the door structure (Lauterbach and Dick, 2007)."

7.5 Waist rail

The waist rail covers an important part of the stiffness of a door and safeguards passive safety during frontal, rear, and side impacts (Figure 19). Especially for cabriolets, the waist

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

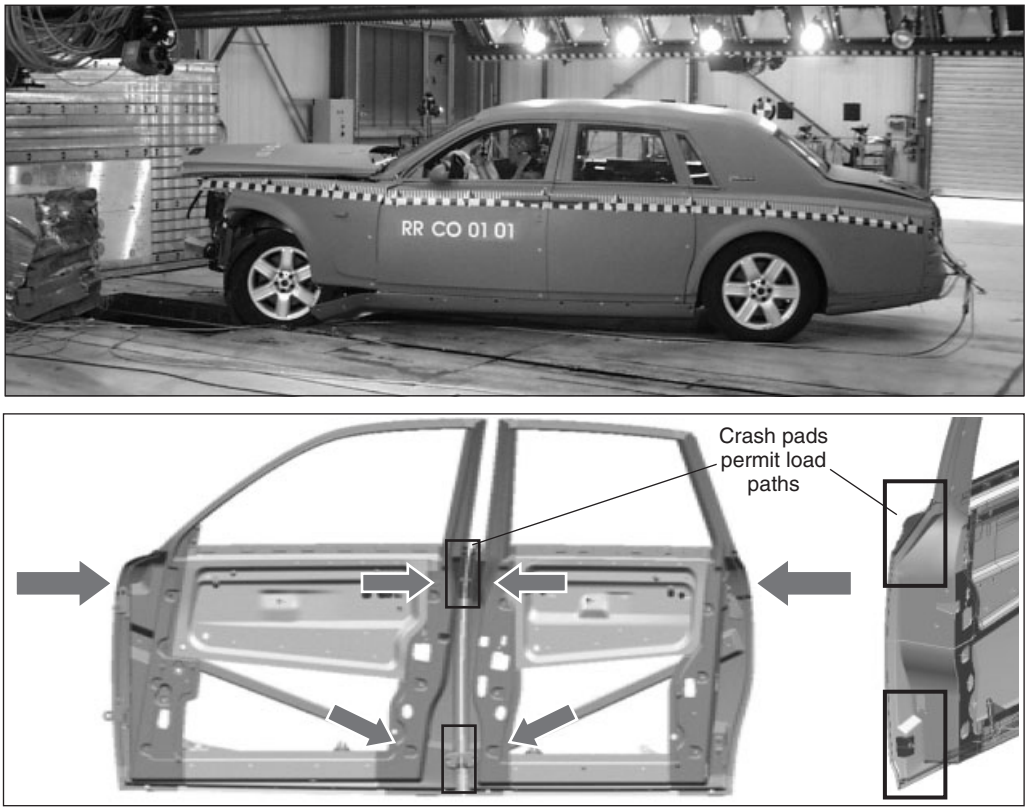


Figure 19. Crash pads permit load paths through side doors (Lindermaier, 2006). (Reproduced by permission of Lindermaier/BMW AG.)

rail defines an important load path for frontal impact. For the local reduction of gaps between doors and pillars in the load paths, beads are designed or crash pads are mounted. The window regulator system (WRS) is pivoted to the inner waist rail to allow the improvement of sealing and guidance by adjustment of the door glass and door frame in the *Y*-direction by adjustment screws between the WRS and inner panel in the sill area. The inner and outer seals are supported by the inner or outer waist rail to protect the inner door from water and dust. The stiffness of the outer waist rail must protect the door from forced opening through the waistline.

The door glass divides the waist rail into an inner and outer profile. The sophisticated design of the front and rear joints under consideration of door functions, producibility, stiffness, and safety is complex. The inner panel and the one-piece inner waist rail panel easily define a closed profile because welding operations are mainly covered by the inner door trim. The outer waist rail must be defined by two separate panels as it is not possible to define a force fit between a single panel and the visible outer panel. So at the outer waist rail assembly, the order and definition of the contacts between rail panels, and inner and outer door panels are a lot more demanding than in the inner side.

Nonlinear dynamic simulation of the whole vehicle is ambitious and time consuming. The goal is to optimize one of certain variants modeled in different ways. So it is not useful to start with the simulation of the whole vehicle. The early layout of the body structure is often carried out under support of partial structures such as the door inner panel combined with the inner and outer waist rails. The geometry of these separate models can be easily defined, meshed, and optimized. According to Lauterbach and Dick (2007) and Hänschke (2007), this approach reduces development time considerably and leads to reliable complete vehicle models. As CAD geometries in the early phase of the product development do not exist or are incomplete, CAE engineers model their partial structures themselves, use body structure libraries of former body structures, or work with parametric structures defined in special CAE tools.

8 CONCEPT COMPETITION AND SUPPLIER INTEGRATION

"After mass production in the 1920s and 'lean production' in the 1980s, the automobile industry is undergoing a

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

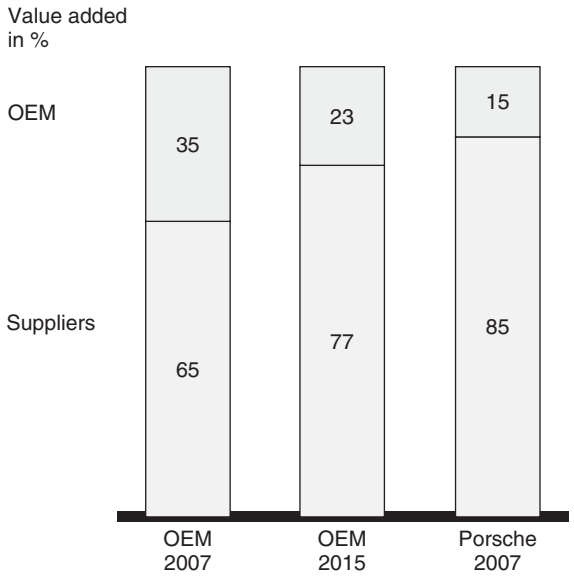


Figure 20. Change of added value between OEMs and suppliers acc. to Wyman (2004) and Binder (2007).

new revolution. By 2015, automotive suppliers will have taken over large parts of R&D and production from the automobile manufacturers, achieving total growth of 70% in this process. During the same period, the auto makers will give up 10% of their current value creation, even though their output will increase by 35%." (Wyman, 2004; see Figure 20)

The cooperation between OEM and suppliers has changed dramatically. The early integration of systems suppliers in the vehicle concept phase, distribution of responsibilities and partnership between all partners involved, concept competitions, target prices set by the OEM, focus of OEMs on core competences, and related reduction of development depth for supplier components have defined a completely new qualification profile for suppliers during the last 15 years.

Hundreds of OEM–supplier relations from concept development to delivery of components for mass production take up subordinate roles in the vehicle development process. Besides OEM and systems suppliers, the engineering suppliers take over most of the design work and a lot of simulation work. According to the master process (PEP) of the OEM, they all follow the same goal and have to be coordinated from the technical and commercial standpoints of the OEM. The basis for most technical communications and simulation processes are 3D CAD models, which always have to be kept up to date and described precisely according to the requirements and maturity defined. Failures in subprocesses endanger the overall goal (SOP) and lead sometimes to superheated catch-up games.

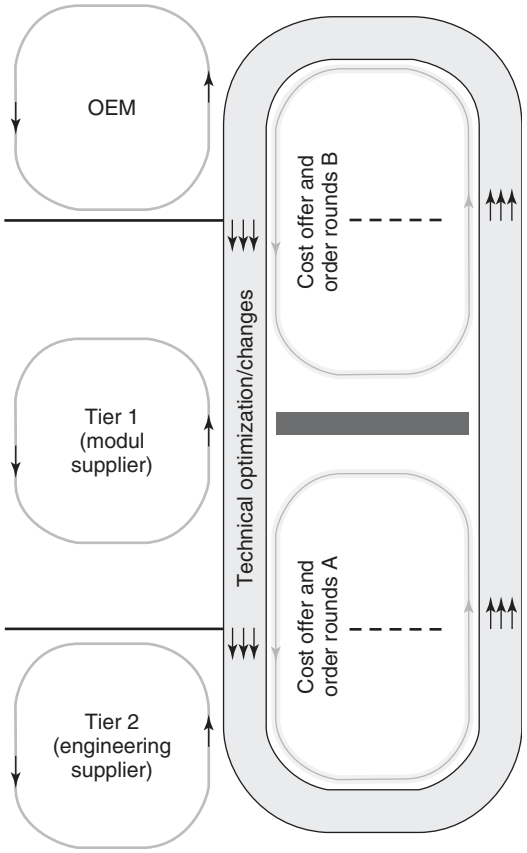


Figure 21. Contradictions in dynamic cooperation between OEM and suppliers.

For the promotion of a number of technical ideas, creativity, and innovation, OEMs define concept competitions at the start of a new project. In every competition, several competitors are invited to apply their technical product proposals. The bases for the proposals are according to the enquired assembly, a list of requirements defined by the OEM, as well as package space restrictions and concept surfaces.

For the OEM, this proceeding requires an extensive supplier management with assessment, development and risk management measures. The suppliers need to be integrated efficiently in suggestion gathering, creativity workshops, and concept competitions and, finally, in the concurrent engineering process.

The technical cooperation between OEM and suppliers is a stringent dynamic process on a high level of outsourcing for engineering work. Otherwise, reduction of time in the development processes would not be possible. The commercial administrative process usually is neglected and often leads to depressive moods in technical cooperation (Figure 21).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

For the systems supplier, the reorientation of the OEMs leads to a new development strategy. While formerly the development of the systems suppliers was realization of the product developed by the OEM, nowadays, besides actual system development innovation management, fundamental research and independent predevelopment of substandard and standard components becomes their responsibility. The independent predevelopment may start 6 or more years before the component goes into mass production. The development of project-defined assemblies starts with the concept competition of an OEM about 3 years before SOP. According to the complexity of the assembly, about 6–12 month are needed to launch and guarantee a safeguarded mass production.

In the context of concurrent engineering of vehicles and their necessary components, an OEM expects from their suppliers at least the same competency and professionalism as their own R&D center. In addition, a high level of communication ability is expected. These abilities not only include the CAD, CAE, and data management systems but also up-to-date testing and analyzing equipment, as well as a high level of technical competence and performance. As long as the capacity of a systems supplier is insufficient, engineering suppliers are assigned to completely or partially develop components in a close coordination with the systems supplier and OEM.

Besides engineering suppliers, the systems supplier engages subsuppliers for the development and delivery of subcomponents. Other suppliers of systems are companies specializing in production equipment and machines. These suppliers play an important role when the product is launched before SOP. The cooperation between all subsuppliers has to be carefully managed and synchronized by the systems supplier in cooperation with the OEM. The systems supplier takes over the entire responsibility for the results of product development, product launch and delivery, and the absolute fulfillment of the list of requirements and target costs defined by the OEM.

9 SUMMARY

This chapter gives a small insight into the various development works of stylists, automotive body designers, CAE engineers, and related partners. Every unit of the automotive body is developed with several variants with the objective of improvement toward one optimum. Every single development step of every variant of every unit needs automotive body design of often complex 3D geometries for its evaluation.

It must be the goal for the future to upgrade product knowledge in the early phase of product development by improvement and systematization of design of 3D geometry supported by PAD and the improvement of the distribution of design and calculation/simulation work especially within the concept phase, and also through the whole development process supported by the approaches of systems engineering.

In the following chapters, several approaches for the styling, interior design, and the validation of the automotive body development will be described in more depth.

RELATED ARTICLES

Innovative Structural Design
The Fascination of Car Body Manufacture: Requirements for Car Body Manufacture from Viewpoint of Production
Styling of Cars, from Sketch to Realization, Main Trends and Milestones
Fundamentals, Basic Principals in Road Vehicle Aerodynamics & Design
Exterior Vehicle Noise Development: Assessment and Control
Vehicle Vibration and Harshness
Vehicle Architecture for Meeting Anthropometric, Posture, Comfort, Health Requirements of Passengers
Human Machine Interface Design in Modern Vehicles
Vehicle Seat Design, Development and Manufacturing
Physics of Car Crashes: Design Concepts for Safer Cars
Reparability and Insurance Ratings in the Development of Cars
Adaptive Restraint Systems: Towards Integral Safety
Pedestrian Protection Overview

REFERENCES

Abulawi, J. (2012) Ansatz zur Beherrschung der Komplexität von vernetzten 3D CAD-Modellen (approach for the control of the complexity of 3D CAD-models). Hamburg. Thesis (Dr.-Ing.). Helmut-Schmidt-Universität Hamburg.

Albers, T. (2012) Das digitale Fahrzeug—Herausforderungen in der virtuellen Produktentwicklung (The digital vehicle—Challenges in the virtual product development). *Mobiles*, **37**, 11–15.

Binder, J. (2007) 'Optimierung durch Integration—Mehr Wertschöpfungseffizienz durch erfolgreiche Lieferantenintegration (Optimisation by Integration—More Edit Value Efficiency by successful Supplier Integration)'. *Fifth International Conference for Supply Chain Management, Logistics and Manufacturing in the Automotive Industry, AKJ Automotive 2007*, Monterrey, México. 09–10 October.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Braess, H.-H. and Seiffert, U. (eds) (2007a) *Handbuch Kraftfahrzeugtechnik (Handbook Automotive Technology)*, 5th edn, Vieweg, Wiesbaden.

Braess, H.-H. and Seiffert, U. (2007b) Design und Technik im Gesamtfahrzeug (Styling and Technologies for the entire Vehicle Integration) in *Automobildesign und Technik—Formgebung, Funktionalität, Technik (Automotive Styling and Technologies—Shaping, Functionalities, Technologies)*, 1st edn (eds H.-H. Braess and U. Seiffert), Vieweg, Wiesbaden, pp. 66–81.

Breitling, T. (2007) *Digital Prototype (DPT)—A Milestone in Vehicle Development*. EDM/CAE Forum. DaimlerChrysler. Stuttgart. 18–19 July.

Dehn, M. (2001) Systematische Definition des Begriffes Designqualität und deren Sicherstellung im technischen Produktentstehungsprozess eines Automobils (systematic definition of the term design-quality and its safeguarding during the technical product evolution (formation) process). Thesis (Dipl.-Ing.). HAW Hamburg.

Forsen, J. (2003) *Ein systemtechnischer Ansatz zur methodisch parametrisch assoziativen Konstruktion am Beispiel von Karosseriebauteilen (A systematic technical Approach for the methodical Parametric Associative Design on the Example of Automotive Body Parts)*, 1st edn, Shaker, Essen.

Gegalski, M. (2003) Parametrische Fugenkonstruktion im Class-A Bereich unter Berücksichtigung der Durchgängigkeit im Entwicklungsprozess (parametric associative design of gaps in class A area under consideration of a consistent development process). Thesis (Dipl.-Ing.). HAW Hamburg.

Grabner, J. and Nothhaft, R. (2006) *Konstruieren von Pkw-Karosserien (Design of Automotive Bodies)*, 3rd edn, Springer, Berlin.

Großjohann, D. (2007) Konzeptentwicklung, Auslegung und Erprobung mehrerer komplexer, teilparametrischer Funktionen zur Fugenaustragung innerhalb von ICEM Shape Design (concept development, lay out und evaluation of several complex, partially parametric methods for the lay out of gaps within ICEM shape design). Thesis (Dipl.-Ing.). HAW Hamburg.

Hagenah, F.; Klar, A. (2006) '*Digital* Mock-Up im "Design in Context" Prozess—Werkzeuge und Methoden (Digital Mock Up in the Design in Context Process—Tools and Methods)'. *Conference "Hamburger Karosseriebautage 2006"*. 23–24 May. Wiesbaden: Vieweg.

Hänschke, A. (2007) 'Parametric Model Knowledgebase: Eine Methode zur Unterstützung der Fahrzeugbewertung schon in der frühen Projektphase (Parametric Model Knowledge Base. A Method to support the Assessment of Vehicle Developments in the early Project Phase)'. *Conference "Parametrisch assoziative Entwicklung von Baugruppen der Fahrzeugkarosserie - Visionen und Erfahrungen für zukünftige Entwicklungsprozesse (Parametric associative Development of Assemblies of Automotive Bodies –Visions and Experiences for Future Development Processes)"*. HdT Essen. 22–23 November. Renningen: Expert

Kraus, W. (2007) Grundsätzliche Aspekte des Automobildesign (Fundamental Aspects of Automotive Styling) in *Automobildesign und Technik—Formgebung, Funktionalität, Technik (Automotive Styling and Technologies—Shaping, Functionalities, Technologies)*, 1st edn (eds H.-H. Braess and U. Seiffert), Vieweg, Wiesbaden, pp. 30–65.

Lauterbach, B.; Dick, J. (2007) Virtual Development of a Vehicle Door. *6th CTI Forum Automotive Doors*. Stuttgart. 05–07 February.

Lender, K. (2001) CAD/CAM Freiformflächengenerierung auf Grundlage weniger empirisch gewonnener Geometriedaten für den Entwurfsprozess im Design (CAD/CAM Generation of free formed Surfaces on Basis of few empiric Geometry Data from Styling). Research Report. IBM Germany, Stuttgart.

Lindermaier, R. (2006) Herausforderungen außergewöhnlicher Einstiegskonzepte am Beispiel der Rolls-Royce Coach Door (Challenges of extraordinary Access Concepts on the Example of the Rolls-Royce Coach Door). *Conference "Hamburger Karosseriebautage 2006"*. 23–24 May. Wiesbaden: Vieweg.

Ostle, A.W. (2003) Design in *Handbuch Kraftfahrzeugtechnik (Handbook Automotive Technology)*, 3rd edn (eds H.-H. Braess and U. Seiffert), Vieweg, Wiesbaden, pp. 95–103.

Piskun, A. (2012) Grundlagen der Karosseriekonstruktion (fundamentals of automotive body design). Hamburg. Lecture handouts. HAW Hamburg.

Sielaff, S. (2004) *Automobil-Produktion* - Sonderausgabe "Innenraum". (special edition), p. 3.

Wyman, O. (2004) *Future Automotive Industry Structure (FAST) 2015—The new Distribution of Work in Automotive Industry: Study by Oliver Wyman, the Fraunhofer Society for Production Technology and Automation (IPA) and the Fraunhofer Society for Materials Management and Logistics (IML)*, VDA, Frankfurt.

FURTHER READING

Gusig, L.-O. and Kruse, A. (2010) Fahrzeugentwicklung im Automobilbau—Aktuelle Werkzeuge für den Praxiseinsatz (Vehicle Development in Automotive Industry—Up-to-date Tools for the Praktical Application), 1st edn, Hanser, Munich.

Leinweber, S. (2007) Entwicklung einer Konstruktionsmethodik für das Konzeptmodel eines Türinnenbleches durch Integration von Funktions- und Stylingflächen mit Unterstützung von CATIA V5 (development of a design methodology for the concept design of a door inner panel by integration of functional and styling surfaces under support of CATIA V5). Thesis (Dipl.-Ing.). HAW Hamburg.

Schreiber, S. (2007) Entwicklung einer templatebasierten Wissensdatenbank zur konzeptionellen und technischen Absicherung im Strakprozess (development of a template based knowledge data base for the conception and technological safeguarding of class A surfacing). Thesis (Dipl.-Ing.) HAW Hamburg.

Tecklenburg, G. (2010) Design of automotive body assemblies with distributed tasks under support of parametric associative design (PAD). Thesis (PhD). University of Hertfordshire.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto286
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Batteries

Jun Furukawa

The Furukawa Battery Co., Ltd., Fukushima, Japan

1 Introduction 1
2 Outline of Automotive Battery 1
3 The Principle of Lead Acid Battery 2
4 Structure, Manufacturing Method, and Types of Automotive Batteries 3
5 Charge and Discharge Characteristics of Automotive Battery 5
6 Degradation Modes of Automotive Batteries and Countermeasures 6
7 The Current State of Automotive Applications 8
8 Future Outlook of Automotive Batteries 10
References 10

1 INTRODUCTION

A power supply system in a vehicle electrical system is composed of three basic functions: power generation function represented by an alternator, power supply–storage function represented by a battery, and power distribution function represented by a wire harness. In this chapter, we introduce the battery as a function of power supply–storage.

2 OUTLINE OF AUTOMOTIVE BATTERY

The lead acid battery was invented by Planté in 1859. Thanks to its excellent practicality, this battery has been widely used even now 150 years later, not only for automobiles application but also in the industrial field, for instance, in emergency power supply, electric vehicle (EV) field such as electric fork-lift truck, and so on. Furthermore, this battery is expected to be used as means for global warming prevention, such as energy storage for load leveling and grid interconnection of the micro grid that utilizes renewable energy such as solar power. Especially, the lead acid battery in automotive application has been in an indispensable position for high durability in the high temperature environment under the hood, cold starting capability, and low cost.

In about 1920, almost 20 years after the introduction of EVs, the lead acid battery started to be used in automobiles equipped with internal combustion engine. Automobiles in those days were equipped with far less electrical devices than today. Electric loads were only the starting system, lighting system, and ignition system. Then, the electrical system voltage of vehicles was half of the current voltage, 7-V (with a 6-V battery), and the automotive battery was called a *starting, lighting, and ignition* (*SLI*) *battery* taking initials of three systems mentioned earlier. In the 1950s, the electrical system voltage was raised to the current of 14-V (with a 12-V battery) because of the increase in electric load. After that, the number of electric equipments and electric load of automobiles increased rapidly because of the pursuit of safety, comfort, convenience, and economy. For instance, electrical equipments for body systems are over 100; various motors are used for opening and closing of windows, mirror adjustment, opening and closing of slide doors, audio-visual equipments, and the navigation system. Chassis and power train equipments, such as electric power steering, electro-hydraulic brake, and even idling stop-start systems, are added to that. In such a trend of electric load increase, lack of electric power supply became an urgent issue mainly in the European luxury vehicles, and

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.

This article is © 2014 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

as a result, a 42-V electrical system (with a 36-V battery) began to be advocated. In 1996, an MIT (Massachusetts Institute of Technology) consortium presided over by MIT began full-scale activity in the United States. Coupled with the mass reduction of the enlarged wire harness, the 42-V electrical system investigation started. However, there were big problems concerning parts, system reliability, and cost in the once in 50 years technical innovation of the 42-V electrical system. Furthermore, significant improvement of electric generation capability, due to the progress in alternator technology, reduced the demand for a 42-V electrical system, which stemmed from the lack of power supply. But, hybrid power systems of 14-V and 42-V were already introduced in some automobiles, for instance, partial voltage step-up using DC/DC converter was applied to electric power steering, keeping electrical system voltage at the conventional 14-V. Hence, 12-V is still the mainstream automotive battery. But it is expected that in near future, the days of two batteries with 36-V and 12-V and one 36-V battery will be introduced.

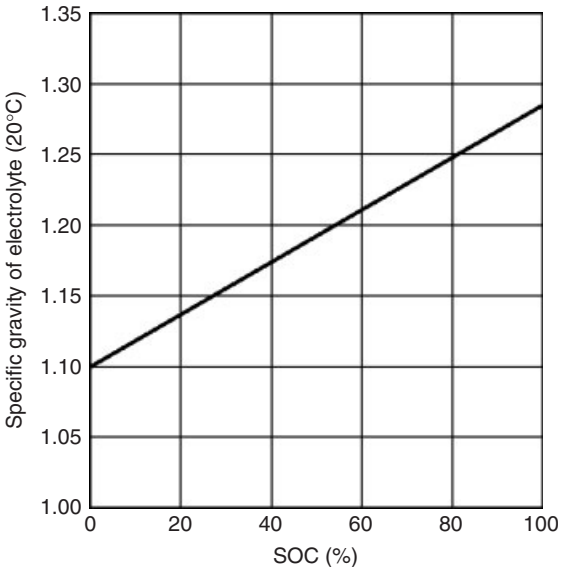


Figure 1. Relation between SOC and specific gravity of electrolyte.

3 THE PRINCIPLE OF LEAD ACID BATTERY

The overview of the lead acid battery is shown in the following sections.

3.1 Discharge reaction of lead acid battery

A lead acid battery has lead dioxide (PbO_2) as positive active material (PAM), spongy lead (Pb) as negative active material (NAM), and dilute sulfuric acid as electrolyte (H_2SO_4).

Equation (1) describes discharge electrochemical reaction of lead acid battery:

$$PbO_2 + Pb + 2HSO_4^- + 2H^+ \rightarrow 2PbSO_4 + 2H_2O \quad (1)$$

Lead dioxide changes to lead sulfate and water at positive electrode, and lead changes to lead sulfate at the negative electrode. In each reaction, active materials generate lead sulfate reacting with sulfuric acid, which is an electrolyte. This means, sulfuric acid density is decreased in the electrolyte with the progress of discharge reaction. Utilizing this behavior, state of charge (SOC) of lead acid battery can be found by measuring the specific gravity of the electrolyte (Figure 1).

On the other hand, as lead sulfate, which is a product due to the discharge, shows little or no electric conductivity, so the crystalline morphology of lead sulfate, at the electrodes in the discharge process, has a big influence on the subsequent charge reaction and the battery characteristic. Especially in the case of the lead acid battery, the situation called *sulfasion* is one of the typical degradation modes. In this situation, lead sulfate accumulates in the electrodes as coarse crystals and after that passivation occurs. This phenomenon makes charging most difficult.

3.2 Charge reaction of lead acid battery

Lead sulfate, a product due to the discharge, is oxidized to lead dioxide at the positive electrode and reduced to lead at the negative electrode. Equation (2) describes this process:

$$2PbSO_4 + 2H_2O \rightarrow Pb + PbO_2 + 2HSO_4^- + 2H^+ \quad (2)$$

After the battery is fully charged, if charging continues, oxygen from the positive electrode and hydrogen from the negative electrode are generated by electrolysis of the water in the electrolyte. So, ignition sources should be kept away from overcharged batteries.

3.3 Open circuit voltage (OCV) of lead acid battery

A lead acid cell has approximately 2-V of OCV (open circuit voltage); this is a very high value for a cell using a water-based electrolyte. This is based on the characteristic that the oxygen overvoltage of lead dioxide at the positive

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

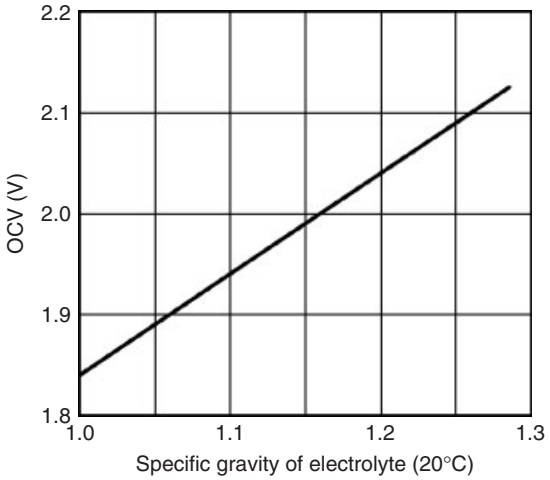


Figure 2. Relation between specific gravity of electrolyte and OCV.

electrode and the hydrogen overvoltage of lead at the negative electrode are high and water electrolysis is more difficult to cause than for other metals and compounds. The OCV of a lead acid battery has a good correlation with the specific gravity of the electrolyte. The relation between specific gravity of electrolyte and OCV is shown in Figure 2. The simplified way to find the OCV is to add 0.84 to the specific gravity of the electrolyte; this gives a nearly equivalent OCV value.

4 STRUCTURE, MANUFACTURING METHOD, AND TYPES OF AUTOMOTIVE BATTERIES

4.1 Structure of automotive battery

Structure, components, and materials of a typical automotive battery are shown in Figure 3.

In an automotive battery, positive plates, inserted into an envelope like polyethylene separator, and negative plates are stacked, and current collectors at the positive and negative plates are welded respectively to plate connectors. This is called *plate group* that configures a 2-V cell as the basic elements of a battery. The plate groups are respectively inserted into six cells in a container that is made of injection-molded polypropylene. Each cell is connected by resistance welding between the direct cell connectors provided on the plate connector. And a polypropylene lid, which is embedded with terminal posts made of lead alloy, is bonded to the container by a hot molding process. Positive and negative poles provided on the plate connector at the first and sixth cells are welded to terminal posts of the lid. Sulfuric acid aqueous solution, with a specific gravity of approximately 1.28, is poured as electrolyte.

4.2 Manufacturing method of automotive battery

An outline of an automobile battery manufacturing method is discussed in the following paragraphs.

Generally, the positive and negative plates of an automotive battery are manufactured by the following procedure, though the type of additive and manufacturing conditions may be different. First, lead oxide powder, the raw material, is manufactured by the ball mill method, which oxidizes lead balls by using the heat generated by the friction between lead balls struck mutually, or by Barton pot method, which oxidizes the pulverized molten lead. Water, sulfuric acid aqueous solution, and additives are added to the lead oxide powder and mixed to a paste form. And lattice-shaped battery grid, which is made of trace elements added lead alloy, is pasted with the paste mentioned earlier, and the surface is dried. Mainly Ca and Sn are used as trace elements for both positive and negative battery grids, even though added amounts are different between the two. Note that, in case of heavy duty application such as trucks, buses, taxis, and construction vehicles, Sb added lead alloy is generally used for the positive battery grid. After that, electrochemical reactivity and mechanical strength are increased by the chemical reaction in high temperature and high moisture atmosphere, called *curing process*, and unformed (uncharged) plates are finished.

The assembly process using the unformed plate is mentioned later (explained in Section 4.1): (i) positive and negative plates stacking, (ii) current collectors welding, (iii) plate group insertion into a container, (iv) welding between the cells, (v) bonding the lid and the container, and (vi) welding between poles and terminal posts.

The electrolyte is poured into the assembled 12-V unformed battery with repeat charge and discharge in the electrochemical conversion process called *formation process*. Thus, charged automotive battery is completed.

As explained earlier, the PAM is made of lead dioxide and the NAM is made of lead. In addition, all the metallic conductor parts, such as battery grids, connecting parts, and terminals, are made of lead alloys. Then, lead acid battery is easy to recycle. Furthermore, the polypropylenes used for the lid and the container and sulfuric acid as electrolyte are recyclable. So, more than 90 mass% of the components of automotive battery can be recycled.

Thus, automotive battery can be ecofriendly products.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

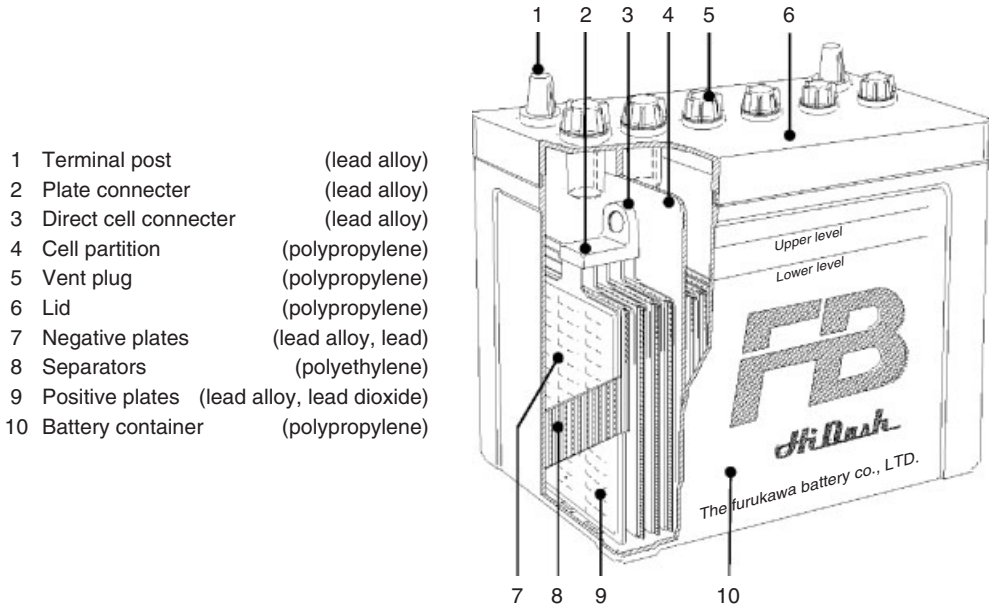


Figure 3. Structure, components, and materials of automotive battery.

4.3 Maintenance-free (MF) battery

At the beginning, lead-antimony alloy, which has excellent mechanical strength, was used as the battery grid alloy for the lead acid battery. But, as antimony significantly lowers the hydrogen overvoltage of lead, so water loss was induced, due to hydrogen gas generation, when used for the negative battery grid. Then, frequent water addition was required and self discharge was significant. In the 1970s, low antimony alloy, which has lower antimony content, was developed as one step toward maintenance-free (MF) batteries. Around the same time, lead-calcium alloy, which had higher hydrogen overvoltage, was developed and MF proceeded largely to today's type. MF batteries are of two different types. One is the calcium type, which uses lead-calcium alloy for both positive and negative battery grids; another is a hybrid-type, which uses lead-low antimony alloy for the positive battery grid and lead-calcium alloy for the negative battery grid. The hybrid type increases hydrogen gas generation while in use, because of the positive battery grid antimony moving to the negative electrode. This type is less MF than the calcium type. But, by using lead-antimony alloy for the positive battery grid, durability of the positive electrode is improved during deep discharge due to the antimony effect on PAM, and longer life is obtained. So the hybrid type is applied to heavy duty applications such as trucks, buses, taxis, and construction vehicles that have repeated deep discharge and recharge. On the other hand, for this heavy duty application, the use of the lead-calcium alloy for the positive battery grid shortens the life. This phenomenon is called the *antimony-free effect*. Nowadays, improvement has been promoted in various ways. Thus, the calcium type and the hybrid type are differently used depending on the applications. Each characteristic of the MF batteries are shown in Table 1.

Table 1. Comparison of maintenance-free lead acid batteries.

Items		Hybrid-Type	Calcium-Type
Grid alloy	Positive grid	Lead-antimony	Lead-calcium
	Negative grid	Lead-calcium	Lead-calcium
Life cycle	Light load life	Excellent	Excellent
	Heavy load life	Excellent	Moderate
Maintenance	Self-discharge	Poor	Excellent
	Maintenance free	Poor	Excellent
Applications		Commercial vehicle Bus	Passenger car

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

4.4 Valve-regulated lead acid (VRLA) battery

In Sections 4.1, 4.2, and 4.3, descriptions are made of the flooded-type lead acid batteries (open ventilated type), which have abundant electrolyte. Recently, the VRLA (valve-regulated lead acid) battery has been widely introduced in luxury vehicles in Europe. In comparison with the sealed type of batteries, such as small size nickel-cadmium battery or nickel-metal hydride battery, the VRLA battery is a type with extremely low valve opening pressure. VRLA batteries are designed to utilize the recombination reaction at the negative electrode for consuming oxygen gas generated from the positive electrode whilst charging. Any extra oxygen gas that cannot be consumed is released to the outside. The recombination reaction means a series of cycles, that is, "Oxygen gas generated at the positive electrode while overcharging, reacts with the spongy lead at the negative electrode and makes lead oxide. The lead oxide reacts with sulfuric acid in the electrolyte and makes lead sulfate. The lead sulfate is reduced to lead by charging again." This cycle continues while in charging, so the sealed condition can be maintained. But an abundant electrolyte, such as in the flooded type of lead acid battery, suppresses the movement of oxygen gas. Thus, a fine glass fiber separator, called *absorbed glass mat* (*AGM*) separator, is applied to the VRLA battery so as to limit the electrolyte amount, which can be absorbed by the separator. With these treatments, VRLA battery facilitates the oxygen gas movement to the negative electrode. Furthermore, this battery has the following features. As the electrolyte amount is limited, electrolyte does not leak even when the battery trips over. As separators like felt compress the plate group, the positive electrode is hardly degraded and the long life span is expected. On the other hand, owing to limiting the electrolyte amount, the VRLA battery has weak points. Its capacity is smaller than that for the same size of flooded-type lead acid battery. As its heat capacity is small, it is easy to cause the battery temperature to rise. By using it for a long term at high temperature, the recombination reaction causes thermal runaway and the battery is overheated. Thus, when a VRLA battery is loaded into an engine bay, sometimes countermeasures such as a thermal barrier, similar to housing in a battery box, may be needed.

5 CHARGE AND DISCHARGE CHARACTERISTICS OF AUTOMOTIVE BATTERY

A charge and discharge characteristic of an automotive battery is shown in Figure 4. Discharging was performed for 5 h at rated current and charged for 10 h rated current.

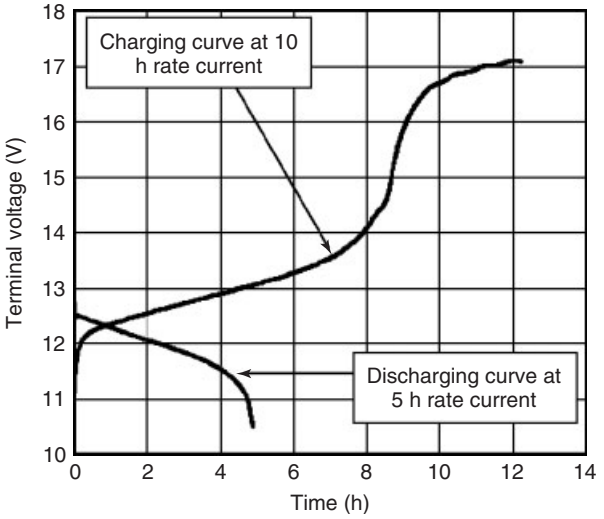


Figure 4. Charge and discharge curve of lead acid battery.

The terminal voltage before discharge began was approximately 12.8 V. A rapid voltage drop caused mainly by internal ohmic resistance occurred right after the start of discharge. After this, the voltage dropped gradually, and rapidly dropped again after 5 h because of depletion of reactants. In the case of charging, immediately after the start of charging, the voltage increased rapidly because of the internal ohmic resistance and polarization. After that, the voltage increased gradually. At the end stage of charging, the voltage increased rapidly and got into the overcharge region, when oxygen and hydrogen gases were generated.

Figure 5 shows the relation between discharge current and time for complete discharge.

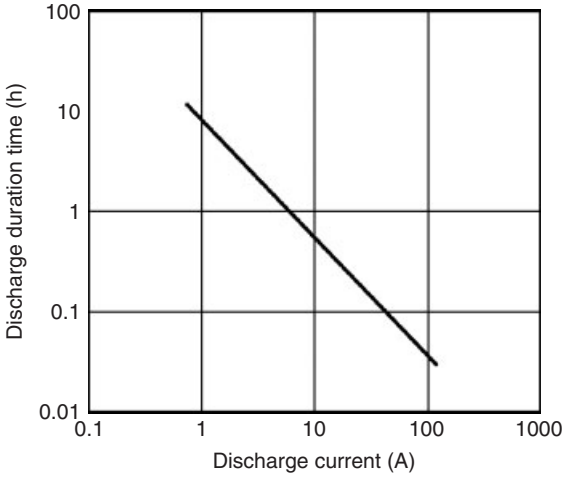


Figure 5. Relation between discharge current and discharge duration.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

From Figure 5, we can find that the discharge duration (capacity) decreases with an increase in discharge current. The relation between discharge current (or discharge rate) and discharge duration can be expressed by the Peukert equation described in Equation (3):

$$t = C \times I^n \tag{3}$$

where I is the discharge current and t is the discharge duration time. C and n are coefficients obtained from plot of I and t in Figure 5.

6 DEGRADATION MODES OF AUTOMOTIVE BATTERIES AND COUNTERMEASURES

A variety of degradation modes of lead acid battery exist depending on operating conditions. Here, we show the typical degradation modes of automotive battery and countermeasures.

6.1 Passivation of the interface between PAM and positive grid (premature capacity loss-1: PCL-1)

In the 1970s, after lead-calcium alloy for the positive battery grid had started to be used instead of lead antimony alloy, a capacity reduction phenomenon frequently occurred because of low conductivity passivation forming in the corrosion layer of the interface between the PAM and the grid. This was improved by adding Sn (tin) to the lead-calcium alloy and forming high conductivity Sn (tin) dioxide layer in the interface between the PAM and the battery grid. On the other hand, the theory mentioned later is also influential. If the adhesiveness of the interface between the PAM and the battery grid is not enough, the electrolyte penetrates into it, and forms a passivation lead sulfate layer or causes a break in the corrosion layer in the interface. Thus, a review of improvement in corrosion resistance of the positive battery grid alloy and improvement of the manufacturing process is now performed. Figure 6 shows scanning electron micrograph (SEM) observation results on the interfaces between the PAM and the positive battery grid after life-cycle tests using a conventional alloy and a new alloy (Furukawa, Nehyo, and Shiga, 2004). The corrosion layer breakage is suppressed in the case of the new alloy shown in Figure 6.

6.2 Softening and shedding of PAM (premature capacity loss-2: PCL-2)

In the charging and discharging reactions of lead acid batteries, PAM and NAM dissolve as Pb^{2+} ion, and this ion precipitates as lead sulfate, lead, or lead dioxide. And repetition of charge and discharge process directly influences the change in the morphology of the active material. Softening of PAM is the phenomenon described later. Owing to above-described dissolution and precipitation reaction and mass transfer of the solid phase, an initial aggregate of fine lead dioxide particles gradually changes to an aggregate of coarse particles. As a result, contact between the particles is reduced and cohesion becomes weaker. This state is called *softening*. Softening is suppressed by adding some kind of metallic ion to PAM. Figure 7 shows SEM observation results of the change in particle size before and after the life-cycle test with and without additives. Coarsening of particles is suppressed by the new additive.

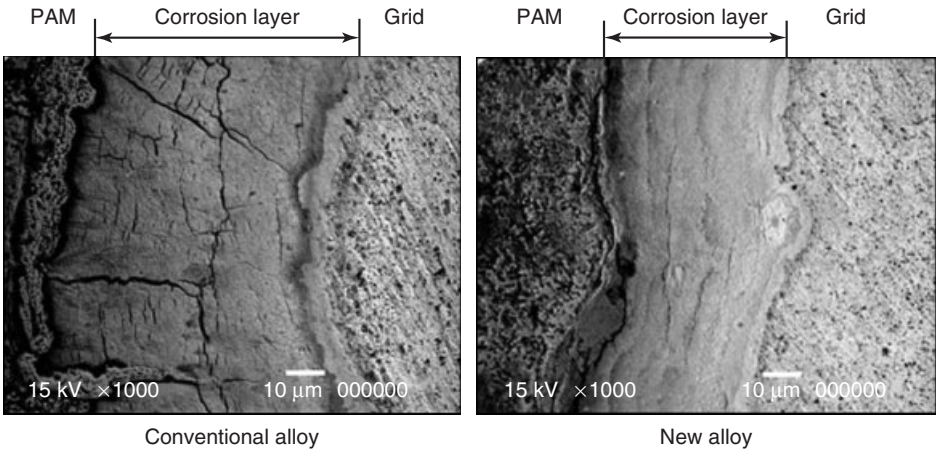


Figure 6. SEM observation results on interfaces between PAM and positive battery grid.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

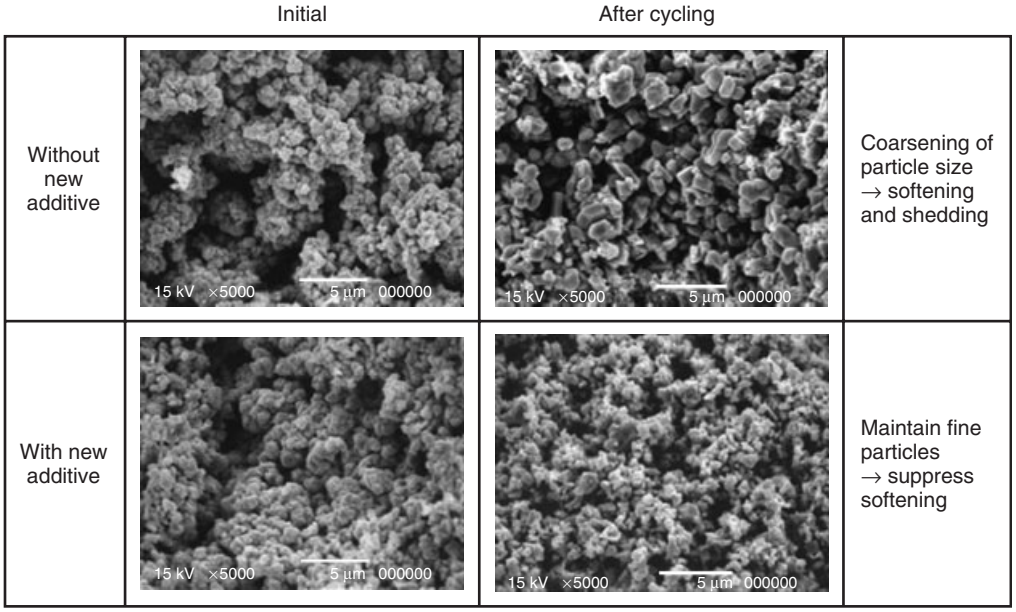


Figure 7. Change in particle size of lead dioxide (PAM) before and after the life-cycle test.

Apart from this, the following theory is also influential. A lot of gel zones, which include hydroxyl group and hydrate water, exist in the initial lead dioxide crystals and act as the glue that connects the crystals. After repeating the charge and discharge process, these crystals change, reducing the glue function and then softening is caused.

6.3 New type of sulfation of negative electrode (premature capacity loss-3: PCL-3)

It is known that when charging, different sulfation occurs because of coarsening of the lead sulfate crystal, a by-product of discharge, in the case of leaving the lead acid battery in a discharged state for a long time. This is caused by the following phenomenon. Owing to a slight lead sulfate dissolution into the electrolyte, initially chargeable fine particles are gradually bound together by the Ostwald ripening effect. Then, the initially fine particles change to more difficult-to-charge coarse particles. Moreover, this new type of negative electrode sulfation can occur within a relatively short time, by repeating rapid charging and discharging in a half way charging state called *high rate–partial state of charge* (*HR-PSOC*). In itself, this HR-PSOC is the battery charge/discharge process of hybrid electric vehicles (HEVs). Similar operating conditions are expected in the case of regenerative braking of enhanced micro-HEVs, which is getting a lot of attention as the next-generation vehicles with idling stop, and mild-HEVs in which the start assist function is added to micro-HEVs. So,

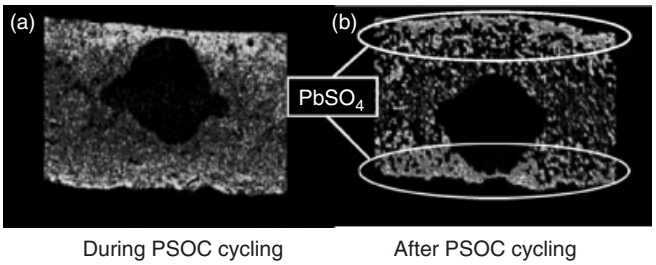


Figure 8. (a,b) EPMA measurements results of lead sulfate density distributions in the negative electrode.

performance improvement of lead acid batteries is needed urgently. In this type of sulfation, stonewall-shaped lead sulfate grows on the surface layer of the negative electrode and impedes the charge and discharge reaction at the negative electrode by blocking the electrolyte movement inside the negative electrode. Figure 8 shows electron proved micro analysis (EPMA) measuring results at the negative electrode cross section during an HR-PSOC life-cycle test and after the test (Takeshima *et al.*, 2004). Bright parts are the high concentration sulfur region in the lead sulfate. This shows that the high concentration of lead sulfate is distributed around the negative electrode surface.

An increase of conductive carbon added in the NAM and the addition of new carbon material has been effective in impeding this sulfation, and more development focused on carbon is now underway.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

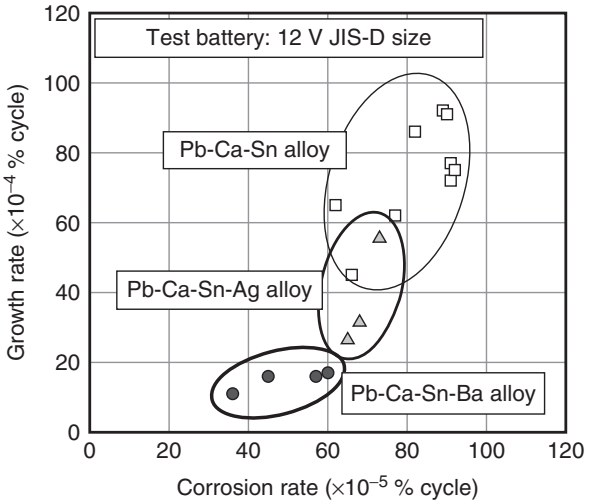


Figure 9. Relation between corrosion rate (weight loss) and growth rate of positive battery grid alloys.

6.4 Corrosion and growth of positive battery grid

Corrosion and growth of the positive battery grid is the most commonly observed degradation mode of automotive batteries that are placed in the high temperature atmosphere under hood and overcharged. Therefore, many lead-calcium alloys to improve corrosion resistivity and growth resistivity have been developed. Lead-calcium-tin alloy with a small amount of barium is a possible alloy that has excellent features in corrosion resistivity and growth resistivity. In addition, lead-calcium-tin alloy with a small amount of silver is used mainly in the United States. The relation between corrosion rate (weight loss) and growth rate of the positive battery grid alloys is shown in Figure 9 (Furukawa, Nehyo, and Shiga, 2004).

7 THE CURRENT STATE OF AUTOMOTIVE APPLICATIONS

7.1 Lead acid battery for idling stop vehicles and micro-HEVs

Vehicles using charging by regenerative braking, in addition to idling stop function, are called *micro-HEVs*. These vehicles have electrical system voltage of 14-V (battery voltage is 12-V). Since 2007, the production of these vehicles has expanded in Europe was more than 25% of the 13.5 million new vehicles manufactured in Europe in 2010. This figure is predicted to reach 60% by 2015 (Fraser-Bell and Prengaman, 2010). Expansion in Japan is predicted as well, but there are some limitations at this time.

As a lead acid battery for idling stop vehicles needs to supply electric power during the no engine idle stop, high durability for deep discharge is required. Also, a large current discharge is necessary for restarting the engine. Thus, low and stable internal resistance is required to ensure reliability on engine restart. Moreover, improved charge acceptance is required for quick recovery of the discharged electricity. For improved deep discharge and internal resistance, the following countermeasures have been developed: increase of PAM density, improved durability for the charge and discharge cycle by the additive described in Section 2.5.2, and optimization of battery grid potential distribution by computer-aided engineering (CAE). Charge acceptance of the lead acid battery is subject to negative electrode performance, so the improvements described in Section 2.5.3 such as an increase of conductive carbon, added to NAM, are performed. As a result, a lead acid battery for idling stop vehicles when tested achieved 60,000 cycles, twice as much as 30,000 cycles target, for idling stop life-cycle test (Takada, Monma, and Furukawa, 2006; Takada and Furukawa, 2008).

Next, it is assumed that the lead acid battery for micro-HEVs is operated under PSOC condition, which means 80% SOC, so as to efficiently receive charge by regenerative braking. For this purpose, suppression of the negative electrode sulfation described in Section 2.5.3 is very important. In addition to this, some kind of additive is needed to suppress crystal growth of the lead sulfate in the negative electrode, and practical application is in progress. Figure 10 shows SEM photographs of observation of the influence on the crystalline morphology of lead sulfate by additive effects (Takada and Furukawa, 2008). It is found that lead sulfate crystal becomes randomized in form and maintains the condition for easy charge.

In addition, although the lead acid battery type may be different depending on the type of vehicle, the flooded-type lead acid battery, which is low cost, tends to be used in compact and standard-sized vehicles, and the VRLA battery, which has higher durability but at a higher cost, tends to be used in high end luxury vehicles. And the VRLA battery tends to be used when improved fuel economy is needed.

7.2 Capacitor hybrid type lead acid battery

The capacitor hybrid-type lead acid battery "UltraBattery" combines an asymmetric capacitor and a lead acid battery in a unit cell, without extra expensive electronic control (Lam and Louey, 2006; Lam *et al*., 2007; Furukawa *et al*., 2010).

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

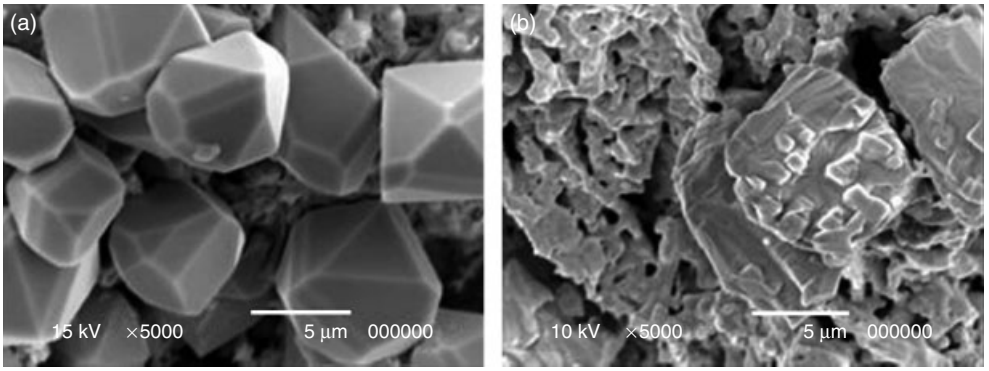


Figure 10. (a,b) Influence to crystal morphology of lead sulfate caused by an additive.

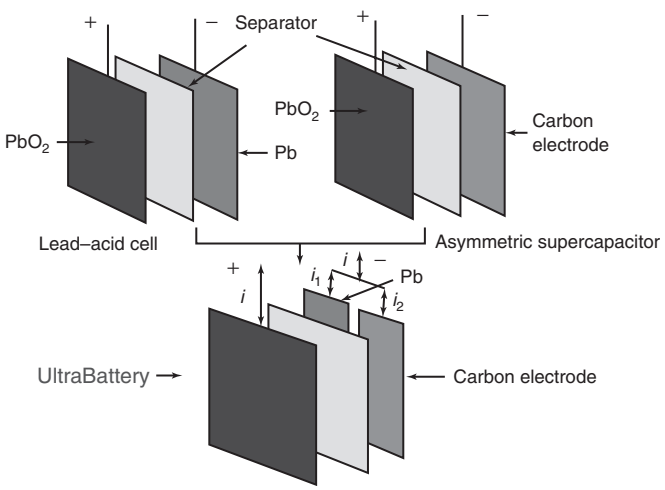


Figure 11. Configuration of UltraBattery.



Figure 12. Prototype of valve-regulated UltraBattery.

Thus, the UltraBattery alone can be used as an electric power source for idling stop vehicles, micro-HEVs, and mild-HEVs in harsh operating conditions, which combines PSOC and large current pulse discharging and charging. The conventional lead acid battery has been difficult to be applied to this condition. The configuration of the UltraBattery is shown in Figure 11.

A normal lead acid cell comprises one lead dioxide positive electrode and one spongy lead negative electrode. In contrast, an asymmetric super-capacitor is composed of one lead dioxide positive electrode and one carbon-based negative electrode (i.e., capacitor electrode). As the positive electrode in the lead acid cell and the asymmetric super-capacitor have a common composition, these two devices can be integrated into one unit cell by internally connecting the capacitor electrode and the lead acid negative electrode in parallel. So, both electrodes can share the electrical load with the lead acid positive electrode. With this design, during discharge and charge, the total current of the combined negative electrode is composed of two components. One is the capacitor current. The other is the lead acid negative electrode current. Accordingly, the capacitor electrode can act as a buffer to share the discharge and charge currents with the lead acid negative electrode and, thus, prevent it by being discharged and charged at the full rates required by HEV duty. Figure 12 shows a prototype valve-regulated ultra battery with the battery sized for motorcycles (5 h rate capacity 8.5 Ah).

For a vehicle demonstration held at the proving ground of GM in Millbrook in the United Kingdom, a Honda Insight HEV was equipped with a 144-V pack, which connected 12 prototype UltraBatteries in series, instead of its usual nickel-metal hydride battery pack, The actual vehicle in the demonstration is shown in Figure 13.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 13. 100 k mile (160 k km) achievement in the UltraBattery demonstration by a Honda Insight HEV.

This was the world's first 100 k mile (160 k km) driving achieved in a vehicle demonstration with lead acid chemistry without charge recovery, which is mandatory for the conventional lead acid battery. In addition, equivalent results to vehicle equipped with nickel-metal hydride battery pack were obtained in driving feeling, fuel consumption, and carbon dioxide emission. As described earlier, the UltraBattery is expected to expand its application to the next generation vehicles such as micro, mild, and high voltage HEVs, for which application the conventional lead acid battery has been considered to be too difficult.

8 FUTURE OUTLOOK OF AUTOMOTIVE BATTERIES

In 1997, Prius—a HEV available from TOYOTA—was launched. It employs an internal combustion engine and an electric motor in combination and successfully achieved less CO_2 emission and fuel consumption than ever before. Until then, "automobiles" meant two types of vehicle, that is, vehicles powered by an internal combustion engine or EVs. The Prius employs an over 200-V high voltage nickel-metal hydride battery pack with a high energy density and power density as a power source for an electric motor, and a 12-V lead-acid battery as an auxiliary power supply. Various hybrid electric vehicles (*x*-HEV) have been introduced after the appearance of Prius. They are called as *strong-HEVs* including Prius, medium-HEVs with an approximate 100-V electrical system, mild-HEVs with a 42-V electrical system (refer to Chapter 1), and micro-HEVs with a 14-V electrical system (refer to Chapter 7), in accordance with hybrid features for reduced CO_2 emission and fuel consumption such as idling stop, regenerative braking, power-assisted driving, and EV driving. Meanwhile, plug-in HEVs with an increased battery capacity for longer EV driving distances appeared as hybrid vehicles closer to EVs. However, the vehicles that require larger amounts of electrical energy involve larger battery mass and volume. So, to meet the need for batteries with a higher energy density and power density, lithium-ion batteries have been developed. Although lithium-ion batteries are expensive, lithium-ion batteries are replacing nickel-metal hydride batteries not only in EVs and plug-in HEVs that require reduction of battery mass and volume but also in strong-HEVs and medium-HEVs that have employed nickel-metal hydride batteries, and it is expected that eventually lithium-ion batteries will replace nickel-metal hydride batteries. Meanwhile, lead-acid batteries are predominantly employed in micro-HEVs with 14-V electrical system (with 12-V battery) because of their limited CO_2 emission and corresponding low fuel consumption. However, a combination of lead-acid batteries with lithium-ion batteries or electric double-layer capacitors is under consideration for reducing cost and improving fuel efficiency. Although, as explained, the predominant main power supplies for automobiles will be classified into lithium-ion battery and lead-acid battery HEVs, in which a 12-V lead-acid battery is mounted as an auxiliary power supply, providing continued use of lead-acid batteries.

REFERENCES

Fraser-Bell, G. and Prengaman, D. (2010) Market and Technology Developments of Extended Cycle Life Flooded Batteries. *International Lead Association, 12th European Lead Battery Conference*, Istanbul, Turkey, 21–24 September 2010. [CD] London; International Lead Association.

Furukawa, J., Nehyo, Y., and Shiga, S. (2004) Development of new positive grid alloy and its application to long-life batteries for automotive industry. *Journal of Power Sources*, **133**, 25–31.

Furukawa, J., Takada, T., Monma, D., and Lam, L.T. (2010) Further demonstration of the VRLA-type UltraBattery under medium-HEV duty and development of the flooded-type UltraBattery for micro-HEV applications. *Journal of Power Sources*, **195**, 1241–1245.

Lam, L.T. and Louey, R. (2006) Development of ultra-battery for hybrid-electric vehicle application. *Journal of Power Sources*, **158**, 1140–1148.

Lam, L.T., Louey, R., Haigh, N.P., *et al*. (2007) VRLA Ultrabattery for high-rate partial-state-of-charge operation. *Journal of Power Sources*, **174**, 16–29.

Takada, T., Monma, D., and Furukawa, J. (2006) Development of lead-acid battery for idling-stop vehicle application. *FB Technical News*, **62**, 15–18.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Takada, T. and Furukawa, J. (2008) Development of lead-acid battery for idling-stop vehicle application. *FB Technical News*, **64**, 43–48.

Takeshima, S., Kourakata, T., Matsumoto, T., *et al*. (2004) Development of compact VRLA "FT7C-HEV" for HEV-type passenger mini-car. *FB Technical News*, **60**, 13–17.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto287
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

Semiconductor Sensors (3): Optical Sensors

Shoji Kanda

DENSO Corporation, Kariya, Japan

1 Optical Sensors 1
2 Applications of Optical Sensors 2
3 Image Sensors 3
4 Conclusion 6
References 6

1 OPTICAL SENSORS

1.1 Principle of the optical sensor (photoelectric effect)

The photoelectric effect phenomenon was first discovered by Hertz and Hallwachs in 1887. Figure 1 shows the principle of a photodiode (PD). When light that has more energy than the silicon (Si) bandgap (Eg) is radiated onto the P–N junction layer, the light produces electron–hole pairs within the Si crystals. The electrons and holes diffuse in accordance with the concentration gradient of the P–N junction. When these reach the depletion layer the accelerated electrons move to the N-layer, while the holes move to the P-layer.

Consequently, the current flows from the N-layer to the P-layer owing to the connection to an external load. In other words, it flows in the opposite direction of the P–N junction. The electrons and holes produced close to the P–N junction are converted efficiently into current, but the electrons and holes that are produced further away from the P–N junction tend to recombine until they reach the depletion layer and are unable to contribute to the production of current.

Light energy is expressed as $E = hv$. If the radiated light energy is considered to be equal to the Si energy band gap (Eg = 1.12 eV), then from $hv = 1.12\,\text{eV}$ and $v = c/\lambda$, the energy becomes $\lambda = hc/1.12$. This equals 1.24/1.12, which is approximately 1100 (nm). Therefore, Si possesses sensitivity up to a wavelength of approximately 1100 nm.

1.2 Relationship between optical sensors and wave length

Figure 2 shows the relationship between the bandgap of each type of semiconductor and light wavelengths.

Owing to the photoelectric effect, the energy of the light in excess of the bandgap is absorbed (i.e., the shorter wavelength light) and the energy below the bandgap passes through the semiconductor (i.e., the longer wavelength light). In the case of a gallium phosphide (GaP) semiconductor, light in the color range from red to yellow passes through as it absorbs light with a wavelength of 564 nm (2.2 eV) or less. For this reason, GaP semiconductors appear to have an orange color. Si and gallium arsenide (GaAs) semiconductors appear black because these materials absorb light up to the infrared light range (however, surface reflection makes the actual crystals appear silver). Zinc sulfide (ZnS) semiconductors allow all light from the visible range to pass and therefore appear transparent. Furthermore, the light absorbed by these materials may also have an effect on the material properties. Light radiated onto the P–N junction of a Si integrated circuit (IC) has an adverse effect as it causes the generation of electron–hole pairs. For this reason, black is selected for the color of molding packages. Si is widely used for optical

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto288
Also published in the *Encyclopedia of Automotive Engineering* (print edition)
ISBN: 978-0-470-97402-5

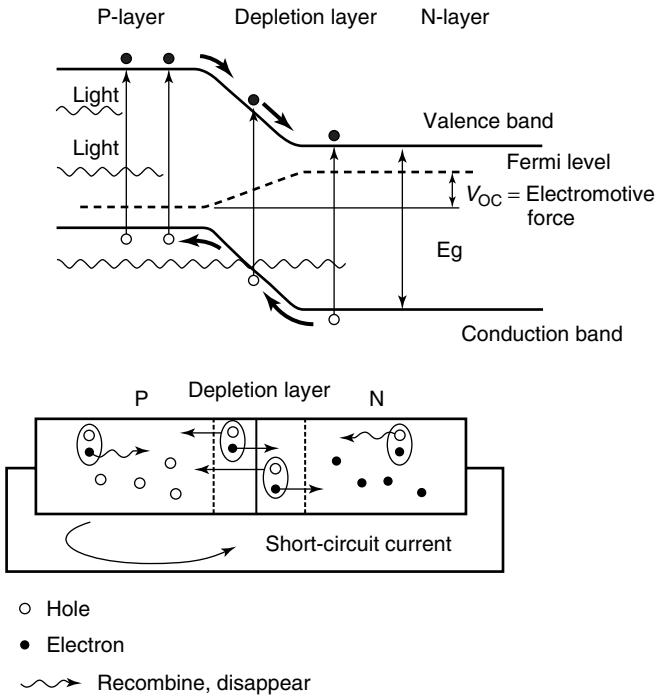


Figure 1. PD principle.

sensors as these sensors are required to detect light in a range extending from visible to infrared rays. Furthermore noncontact temperature sensors that detect infrared light generated from high temperature objects are also required to react to light in the far-infrared range. Ceramic sensors that utilize the pyroelectric effect are generally used for this purpose. However, this type of sensor falls outside the scope of this chapter, which focuses on semiconductor optical sensors.

2 APPLICATIONS OF OPTICAL SENSORS

Examples of optical sensors include automatic headlamp control sensors and sunlight sensors that help to improve air conditioning performance. However, as these sensors are installed close together and share the same basic technology, light sensors that combine both these functions are now the mainstream type of optical sensor. The following sections describe the principles of light sensor operation. Section 2.1 details the structure of a light sensor and Section 2.2 outlines each system used for the headlamps and air conditioning.

2.1 Structure of a light sensor

Figure 3 shows the external appearance and structure of a light sensor and a top view of a sensor chip.

Circular and semicircular PDs are formed on the sensor chip. The aperture is smaller than the PDs on the sensor chip. This structure is capable of detecting the direction and amount of radiated light. Section 2.2 describes the details.

2.2 Characteristics of a light sensor

Figure 4 shows an outline of the systems that use light sensors.

The headlamp control system automatically turns the headlamps and tail lamps on and off in accordance with the brightness of the surrounding environment to improve convenience and safety. However, it also has a function to detect light radiated from the headlamps of oncoming vehicles to prevent the headlamps of the driver's car turning off because of the lights of other cars. Another function of

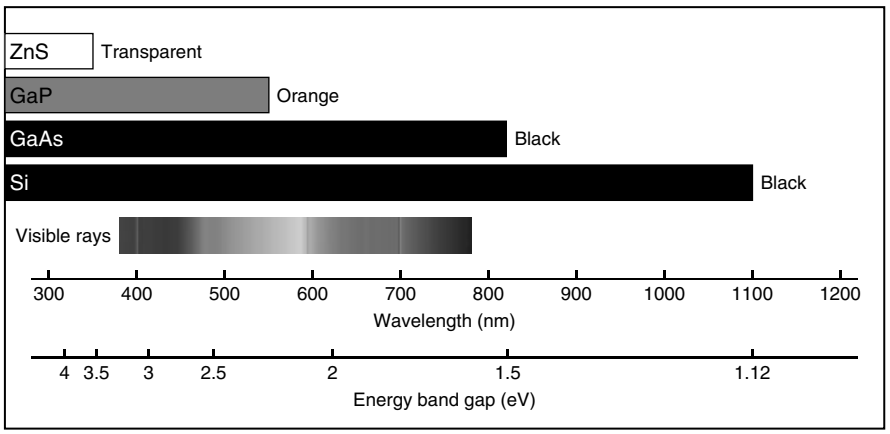


Figure 2. Wavelength and energy bandgap.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto288
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

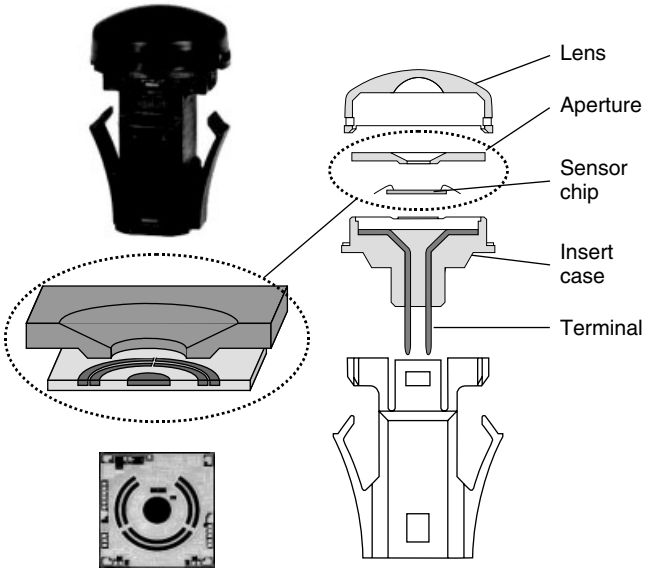


Figure 3. Structure of a light sensor.

this system prevents the driver from forgetting to turn off the headlamps.

The temperature and flow strength of the air conditioner are generally determined on the basis of the difference between the external and internal temperatures. However, as contact with direct sunlight has the effect of making vehicle occupants feel warm, air conditioners may be equipped with light sensors. These detect the direction and amount of sunlight, and improve comfort by independently controlling the air flow strength and temperature at the driver and passenger seats. Figure 5 shows the operating principle of a light sensor system.

Figure 6 illustrates the outputs from the light sensor to the headlamp control and air conditioner.

A two-direction sunlight-sensing function is necessary to simultaneously detect the amount of solar radiation coming from the left and right toward the driver's seat and the front passenger's seat. Sunlight passes through the aperture plate, and the amount of received light that is radiated onto a separate PD is detected to determine whether the light is coming from the left or the right. Figure 5 also shows the circuit that is used to realize the elevation characteristics for the sunlight function. A current is output that is proportional to the amount of light radiated onto the ring-shaped light-receiving section. This current is then subjected to arithmetic processing so that it corresponds to the elevation characteristics of the sunlight sensor and the desired elevation characteristics are obtained.

Different elevation characteristics for automatic headlamps are achieved by dividing the PD pattern on the chip so that the center is for the function for automatic headlamps, while the circumference is for the sunlight sensing function. The output is uniform in all directions owing to the ring-shape of the PD in the same way as it is for the two-direction sunlight-sensing function (Honda, Michiyama, Onoda, 2004).

3 IMAGE SENSORS

Nowadays, cameras are being installed at the front and rear of automobiles, and there has been an increase in automobile systems that use these cameras to monitor the area around the vehicle during driving and backing up. This is being done in consideration of safety around the automobile. The type of sensor used in these cameras is an image sensor.

Charge-coupled device (CCD)-type as well as complementary metal-oxide semiconductor (CMOS)-type image sensors are available now; of these, CMOS sensors are cheaper and consume less power. For these reasons, CMOS-type image sensors are in widespread use in mobile media devices such as cell phones, and they have also become the main type of automotive image sensors. The following sections therefore focus on the principle and applications of CMOS-type image sensors.

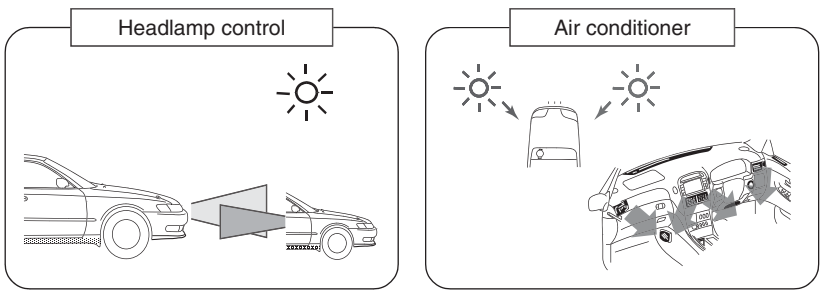


Figure 4. System overview.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto288
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

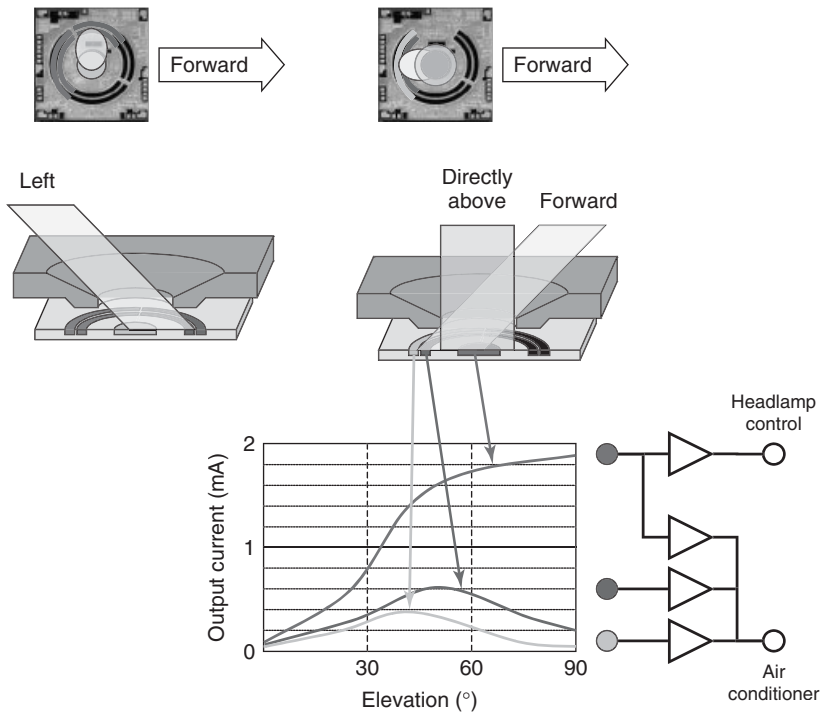


Figure 5. Operating principle of a light sensor.

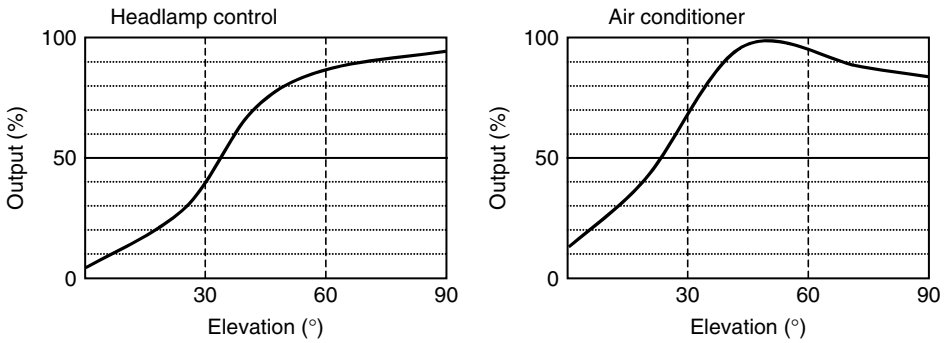


Figure 6. Characteristics of a light sensor.

3.1 Principle of CMOS image sensors

Figure 7 shows the cell composition of a CMOS image sensor, Figure 8 shows its pixel composition, and Figure 9 shows the principle of the sensor.

Since an image sensor obtains images in two dimensions, it uses a matrix layout of cells structured as shown in Figure 8. Each cell is sequentially layered with red, green, and blue filters that allow only the corresponding light to pass. The PD of each cell generates a charge that corresponds to the amount of light radiated onto it in a certain period of time. The image sensor outputs this charge as the light amount of each color. Row and column decoders are provided to acquire the charge from each two-dimensionally laid-out cell without interference. The signal acquisition sequence is as follows. Serial data is outputted by selecting the decoder for each line and each column. This operation is repeated for a single screen. This enables the image signals from the two-dimensional matrix to be outputted as serial data for a single screen. The actual red/green/blue filter layout, sequence, and know-how are proprietary knowledge of each company.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto288
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5

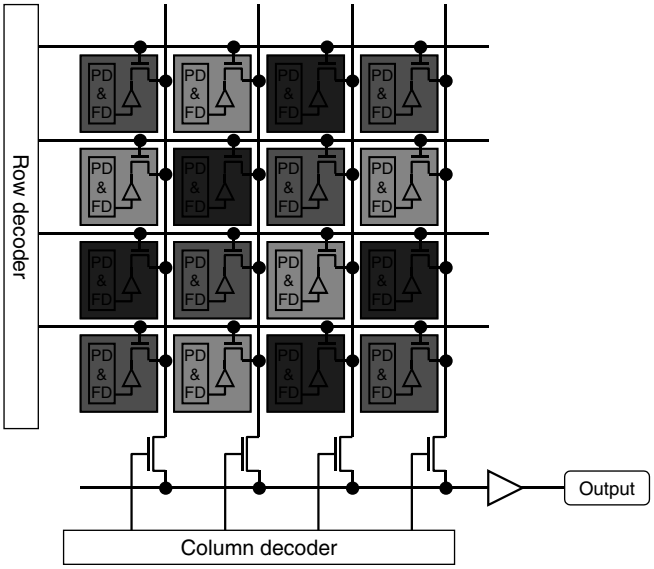


Figure 7. Cell composition of a CMOS image sensor.

Each cell operates as follows. The basic principle is the same as that of a PD, and an electric charge is produced in accordance with the amount of light received by the PD. A transistor (Tr_1) then transmits the electric charge to a floating diode (FD). As shown in Figure 9, voltage sensitivity is increased by transferring the charge from a PD with a large area to a PD with a small area.

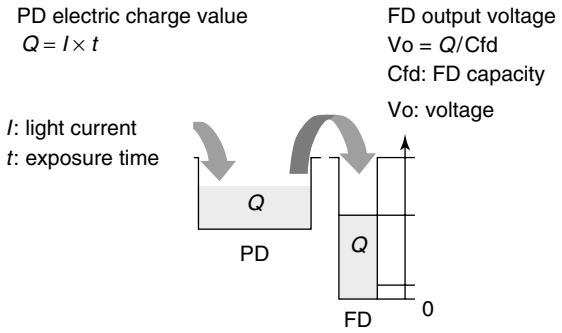


Figure 9. Principle of electric charge transmission.

Next, the charge in the FD is converted into voltage by the electric charge–voltage conversion transistor Tr_3. This voltage signal is then read by the selection transistor Tr_4. Finally, the electric charge in FD is reset by the resetting transistor Tr_2.

Infrared sensors that can detect up to the far-infrared region are classified as thermal sensors that measure infrared rays as heat. These sensors include thermopiles, which utilize the thermoelectric force effect, pyroelectric sensors, which utilize the pyroelectric effect, thermistors, which utilize the thermoelectric effect, and bolometers. There are also quantum sensors that measure infrared rays as light (Tabet, 2002; Fossum, 2008).

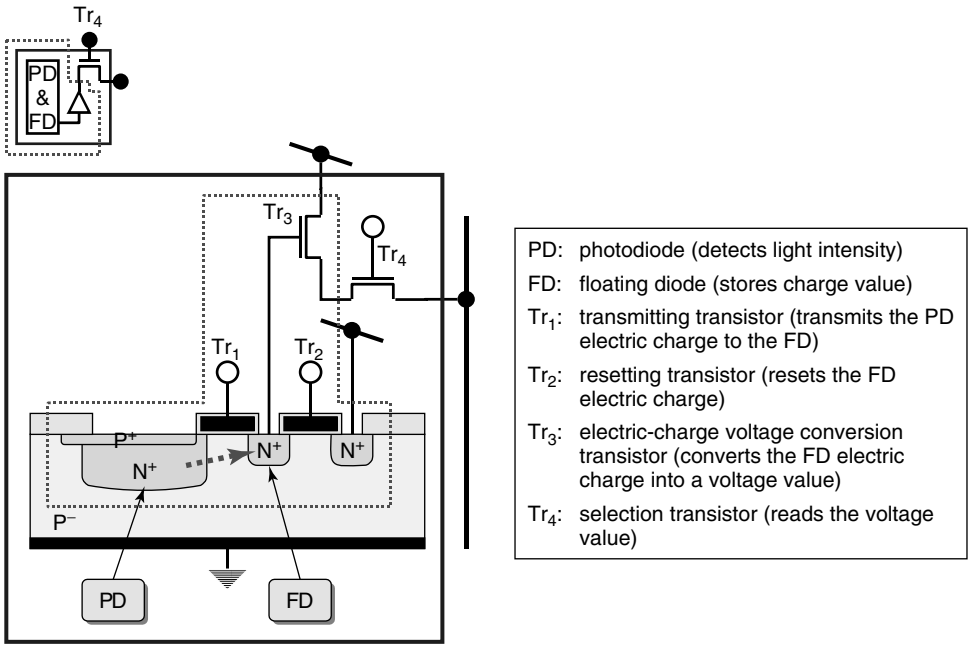


Figure 8. Pixel composition of a CMOS image sensor.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto288
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5



Figure 10. Vision sensor for safety systems.

3.2 Applications of CMOS image sensors

Figure 10 shows an illustration of a vision sensor equipped with a CMOS image sensor. This sensor is installed behind the rear view mirror. It supports safe driving by using an in-built algorithm to judge recognized images in front of the vehicle. Recognition targets include lane markings, road signs, preceding vehicles, objects, pedestrians, and the like. This sensor is currently installed in lane-departure warning (LDW) systems, which help to reduce accidents caused by vehicles leaving their lane, and automatic high beam (AHB) control systems. AHB systems help to improve driver vision at night and prevent the dazzling of drivers in other vehicles by automatically switching between high and low beams based in the presence of a preceding or oncoming vehicle. In the future it is also likely to be adopted for road sign and pedestrian recognition during driving.

Image sensors are installed close to the top edge of the windshield. This is a prominent position that easily attracts the eyes of vehicle occupants and creates a sense of crowding. Therefore, the sensors must be reduced in size to help alleviate these issues. This installation position also tends to become very hot. Therefore, other issues include developing sensors with lower power consumption and packaging with greater heat discharge capabilities.

4 CONCLUSION

This chapter has described how optical sensors are used as PDs to help improve air conditioner performance, control headlamps, and so on. In the future, the number of optical sensors installed in each vehicle will increase because of the rapidly expanding demand for image sensors in the safety field.

As the demand for these sensors increases, technological innovation of hardware is likely to lead to the development of sensors with greater performance (i.e., the improvement of sensitivity, resolution, and compactness). Software innovation is likely to result in the development of improved built-in algorithms. Consequently, sensors installed with image sensors will continue to evolve and new systems will be developed that integrate image sensors with laser radar, millimeter wave radar, sonar, and other peripheral monitoring functions. Ultimately, this technology will help to increase the feasibility of autonomous vehicles in the future.

REFERENCES

Fossum, E.R. (2008) CMOS active pixel image sensors: past, present, and future.

Honda, Y., Michiyama, K., and Onoda, M. (2004) Sensor technology for the air conditioning system. *DENSO Technical Review*, **9** (2), 24–29.

Tabet, M. (2002) Double sampling techniques for CMOS image sensors. Thesis presented to the University of Waterloo.

Encyclopedia of Automotive Engineering, Online © 2014 John Wiley & Sons, Ltd.
This article is © 2014 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118354179.auto288
Also published in the *Encyclopedia of Automotive Engineering* (print edition) ISBN: 978-0-470-97402-5